9. Accelerated proximal gradient methods

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

Proximal gradient method

Results from lecture 6

- each proximal gradient iteration is a descent step:

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

  with $c = 1 - m/L$

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

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

Accelerated proximal gradient methods

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

Assumptions

we consider the same problem and make the same assumptions as in lecture 6:

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

- h is closed and convex (so that prox_{th} is well defined)
- g is differentiable with $\mathbf{dom}\, g = \mathbf{R}^n$
- 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

Algorithm: choose $x^{(0)} = v^{(0)}$ and $\gamma_0 > 0$; for $k \geq 1$, repeat the steps

- define $\gamma_k = \theta_k^2/t_k$ where θ_k is the positive root of the quadratic equation

$$\theta_k^2/t_k = (1-\theta_k)\gamma_{k-1} + m\theta_k$$

- update $x^{(k)}$ and $v^{(k)}$ as follows:

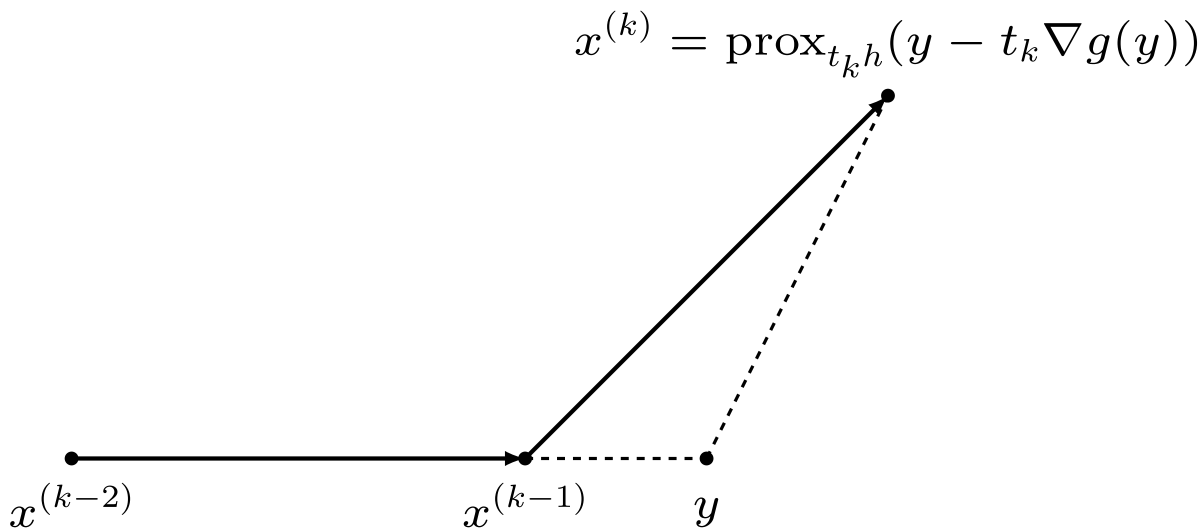
$$\begin{aligned}
y &= x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m\theta_k}(v^{(k-1)} - x^{(k-1)}) \\
x^{(k)} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \\
v^{(k)} &= x^{(k-1)} + \frac{1}{\theta_k}(x^{(k)} - x^{(k-1)})
\end{aligned}$$

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

Momentum interpretation

- first iteration ($k = 1$) is a proximal gradient step at $y = x^{(0)}$
- next iterations are proximal gradient steps at extrapolated points y:

$$
\begin{aligned}
y &= x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m\theta_k}(v^{(k-1)} - x^{(k-1)}) \\
&= x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m\theta_k} \left(\frac{1}{\theta_{k-1}} - 1\right) (x^{(k-1)} - x^{(k-2)})
\end{aligned}
$$

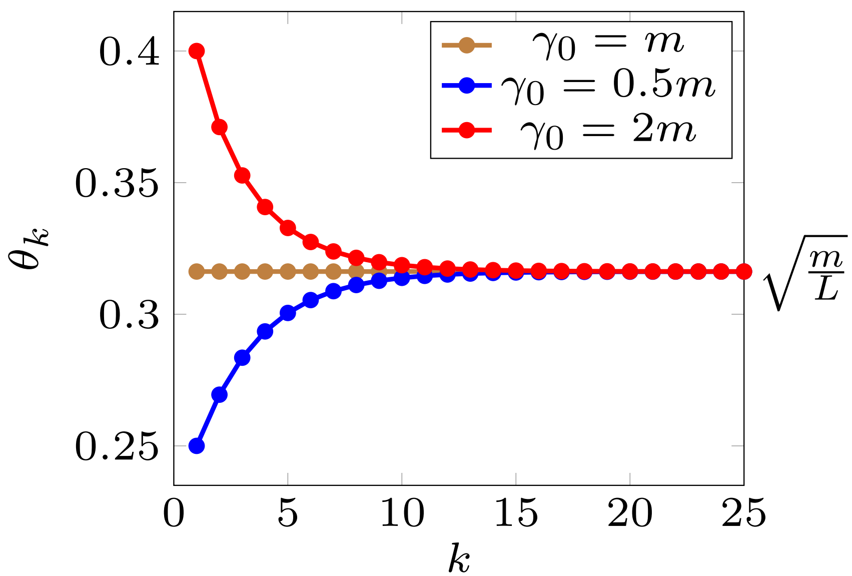


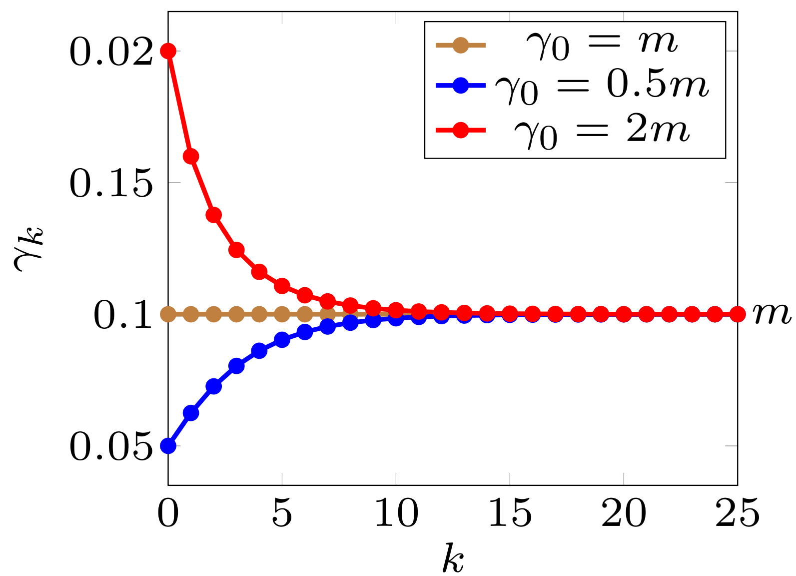
Algorithm parameters

$$\frac{\theta_k^2}{t_k} = (1-\theta_k)\gamma_{k-1} + m\theta_k, \qquad \gamma_k = \frac{\theta_k^2}{t_k}$$

- θ_k is positive root of the quadratic equation
- $\theta_k < 1$ if $mt_k < 1$
- if t_k is constant, sequence θ_k is completely determined by starting value γ_0

Example: $L = 1$, $m = 0.1$, $t_k = 1/L$





FISTA

if we take $m = 0$ on page 9-4, the expression for y simplifies:

$$\begin{aligned} y &= x^{(k-1)} + \theta_k \, (v^{(k-1)} - x^{(k-1)}) \\ x^{(k)} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \\ v^{(k)} &= x^{(k-1)} + \frac{1}{\theta_k}(x^{(k)} - x^{(k-1)}) \end{aligned}$$

eliminating the variables $v^{(k)}$ gives the equivalent iteration (for $k \geq 2$)

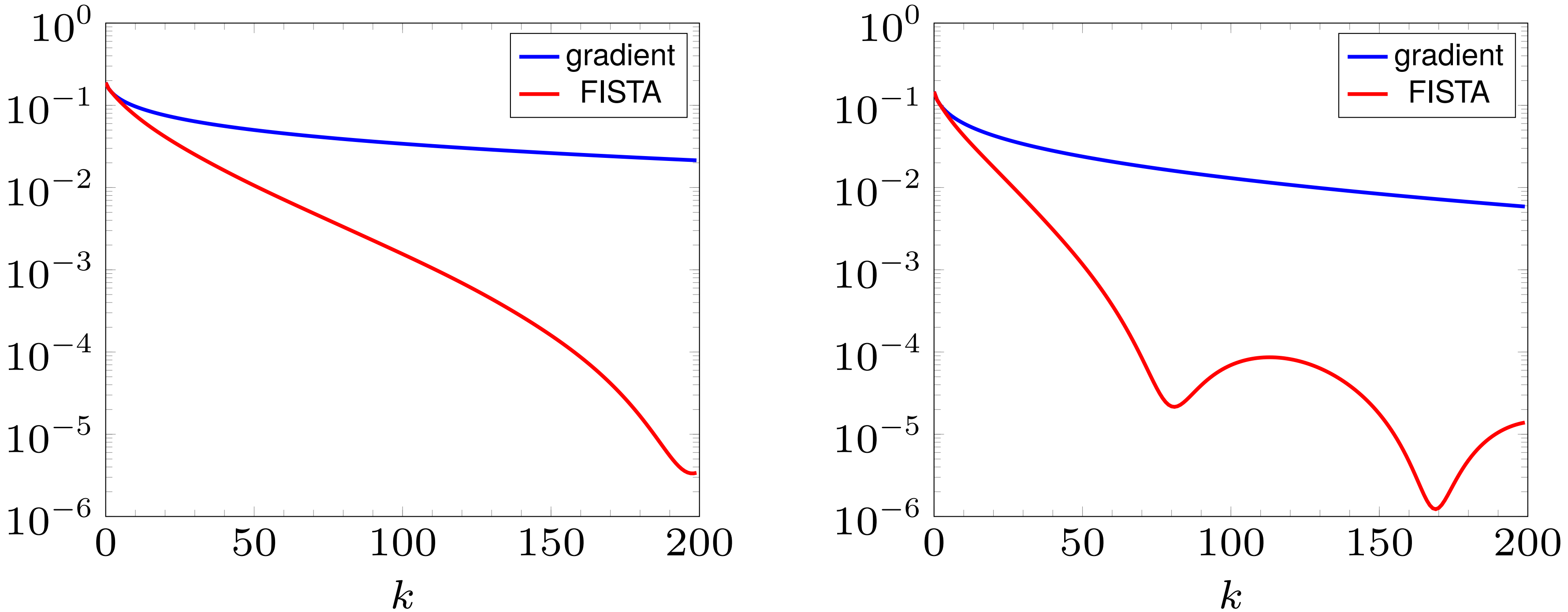
$$\begin{aligned} y &= x^{(k-1)} + \theta_k \, (\frac{1}{\theta_{k-1}} - 1)(x^{(k-1)} - x^{(k-2)}) \\ x^{(k)} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \end{aligned}$$

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

Example

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

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



Nesterov's simplest method

- if $m > 0$ and we choose $\gamma_0 = m$, then

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

- the algorithm on p. 9-4 and p. 9-5 simplifies:

$$\begin{aligned} y &= x^{(k-1)} + \frac{\sqrt{t_k}}{\sqrt{t_{k-1}}} \frac{1 - \sqrt{m t_{k-1}}}{1 + \sqrt{m t_k}} \left(x^{(k-1)} - x^{(k-2)}\right) \\ x^{(k)} &= \operatorname{prox}_{t_k h}(y - t_k \nabla g(y)) \end{aligned}$$

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

$$y = x^{(k-1)} + \frac{1 - \sqrt{m/L}}{1 + \sqrt{m/L}} \left(x^{(k-1)} - x^{(k-2)}\right)$$

Outline

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

Overview

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

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

- therefore the rate of convergence is determined by $\lambda_k = \prod_{i=1}^{k}(1-\theta_i)$:

$$\begin{aligned} f(x^{(k)}) - f^\star \quad & \leq \quad f(x^{(k)}) - f^\star + \frac{\gamma_k}{2}\|v^{(k)} - x^\star\|_2^2 \\ & \leq \quad \lambda_k \left(f(x^{(0)}) - f^\star + \frac{\gamma_0}{2}\|x^{(0)} - x^\star\|_2^2\right) \end{aligned}$$

(here we assume that $x^{(0)} \in \operatorname{dom} h = \operatorname{dom} f$)

Notation for one iteration

quantities in iteration i of the algorithm on page 9-4

- define $t = t_i$, $\theta = \theta_i$, $\gamma = \gamma_i$, and $\gamma^+ = \gamma_i$:

$$\gamma^+ = (1-\theta)\gamma + m\theta, \qquad \gamma^+ = \theta^2/t$$

- define $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $v = v^{(i-1)}$, and $v^+ = v^{(i)}$:

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

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

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

Proof (last identity):

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

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

- multiply with $\gamma^+ = \gamma + m\theta - \theta\gamma$:

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

Bounds on objective function

recall the results on the proximal gradient update (page 6-13):

- 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,

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

- combine the inequalities for $z = x$ and $z = x^\star$:

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

Progress in one iteration

- the definition of γ^+ and (1) imply that

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

- combining this with the last inequality on page 9-13 gives

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

- the last term on the right-hand side is

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

last step uses definitions of γ^+ and y (chosen so that $\theta\gamma(y-v) = \gamma^+(x-y)$)

- substituting this in the last inequality on page 9-14 gives the result on page 9-10

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

Analysis for fixed step size

the product $\lambda_k = \prod_{i=1}^{k}(1-\theta_i)$ determines the rate of convergence (page 9-10)

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

$$\lambda_k \leq \frac{4}{(2 + \sqrt{\gamma_0}\sum_{i=1}^{k}\sqrt{t_i})^2}$$

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

$$\lambda_k \leq \frac{4}{(2 + k\sqrt{\gamma_0/L})^2}$$

- combined with the inequality on p. 9-10, this shows the $1/k^2$ convergence rate:

$$f(x^{(k)}) - f^\star \leq \frac{4}{(2 + k\sqrt{\gamma_0/L})^2}\left(f(x^{(0)}) - f^\star + \frac{\gamma_0}{2}\|x^{(0)} - x^\star\|_2^2\right)$$

Proof.

- recall that γ_k and θ_k are defined by $\gamma_k = (1-\theta_k)\gamma_{k-1} + \theta_k m$ and $\gamma_k = \theta_k^2/t_k$
- we first note that $\lambda_k \leq \gamma_k/\gamma_0$; this follows from

$$\lambda_k = (1-\theta_k)\lambda_{k-1} = \frac{\gamma_k - \theta_k m}{\gamma_{k-1}}\lambda_{k-1} \leq \frac{\gamma_k}{\gamma_{k-1}}\lambda_{k-1}$$

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

$$\begin{aligned} \frac{1}{\sqrt{\lambda_i}} - \frac{1}{\sqrt{\lambda_{i-1}}} &\geq \frac{\lambda_{i-1} - \lambda_i}{2\lambda_{i-1}\sqrt{\lambda_i}} \qquad \text{(because } \lambda_i \leq \lambda_{i-1}\text{)} \\ &= \frac{\theta_i}{2\sqrt{\lambda_i}} \\ &\geq \frac{\theta_i}{2\sqrt{\gamma_i/\gamma_0}} \\ &= \frac{1}{2}\sqrt{\gamma_0 t_i} \end{aligned}$$

Strongly convex functions

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

- if $\gamma_0 \geq m$ then $\gamma_k \geq m$ for all k and

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

(proof on next page)

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

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

- combined with the inequality on p. 9-10, this shows

$$f(x^{(k)}) - f^\star \leq \left(1 - \sqrt{\frac{m}{L}}\right)^k \left(f(x^{(0)}) - f^\star) + \frac{\gamma_0}{2}\|x^{(0)} - x^\star\|_2^2\right)$$

Proof.

- if $\gamma_{k-1} \geq m$, then

$$\begin{aligned} \gamma_k &= (1-\theta_k)\gamma_{k-1} + \theta_k m \\ &\geq m \end{aligned}$$

- since $\gamma_0 \geq m$, we have $\gamma_k \geq m$ for all k

- it follows that $\theta_i = \sqrt{\gamma_i t_i} \geq \sqrt{m t_i}$ and

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

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, $x^{(k)}$ in the algorithm on p. 9-4

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

Analysis with line search

- from page 9-16:

$$\lambda_k \leq \frac{4}{(2 + \sqrt{\gamma_0} \sum_{i=1}^{k} \sqrt{t_i})^2} \leq \frac{4}{\left(2 + k\sqrt{\gamma_0 t_{\min}}\right)^2}$$

- from page 9-18, if $\gamma_0 \geq m$:

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

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

References

Accelerated gradient methods

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004).
  The material in the lecture is from §2.2 of this book.
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008).
- S. Bubeck, *Convex Optimization: Algorithms and Complexity*, Foundations and Trends in Machine Learning (2015), §3.7.

FISTA

- 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).

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).

Interpretation and insight

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), §2.2.
- W. Su, S. Boyd, E. Candès, *A differentiable equation for modeling Nesterov's accelerated gradient method: theory and insight*, NIPS (2014).
- H. Lin, J. Mairal, Z. Harchaoui, *A universal catalyst for first-order optimization,* arXiv:1506.02186 (2015).

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).