7. Fast proximal gradient methods

- fast proximal gradient method (FISTA)
- FISTA with line search
- FISTA as descent method
- Nesterov's second method

Fast (proximal) gradient methods

- Nesterov (1983, 1988, 2005): three gradient projection methods with $1/k^2$ convergence rate

- Beck & Teboulle (2008): FISTA, a proximal gradient version of Nesterov's 1983 method

- Nesterov (2004 book), Tseng (2008): overview and unified analysis of fast gradient methods

- several recent variations and extensions

this lecture:

FISTA and Nesterov's 2nd method (1988) as presented by Tseng

Outline

- **fast proximal gradient method (FISTA)**
- FISTA with line search
- FISTA as descent method
- Nesterov's second method

FISTA (basic version)

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

- g convex, differentiable, with $\mathbf{dom}\, g = \mathbf{R}^n$
- h closed, convex, with inexpensive prox_{th} operator

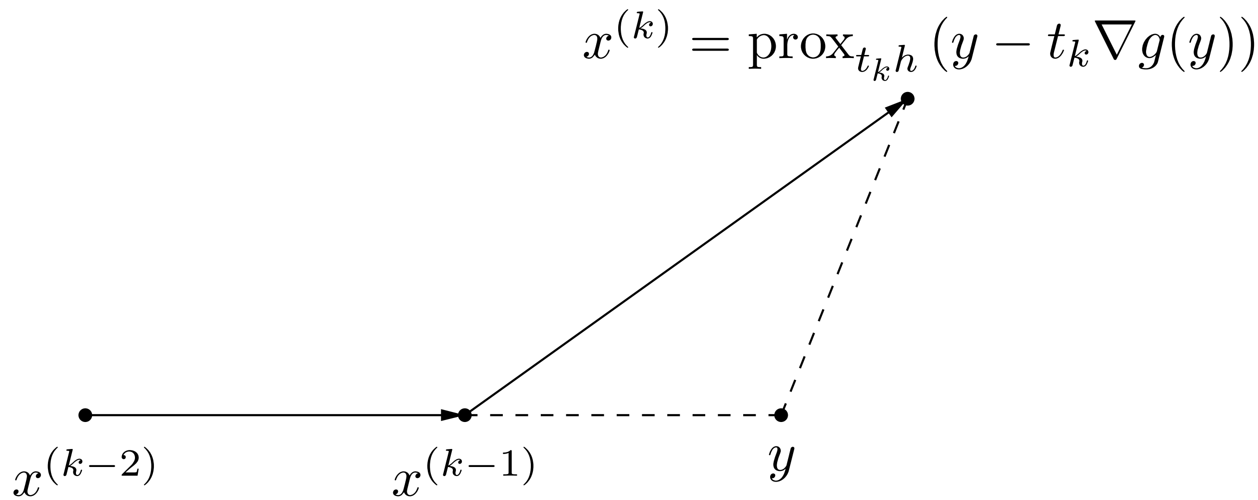
algorithm: choose any $x^{(0)} = x^{(-1)}$; for $k \geq 1$, repeat the steps

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

- step size t_k fixed or determined by line search
- acronym stands for 'Fast Iterative Shrinkage-Thresholding Algorithm'

Interpretation

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

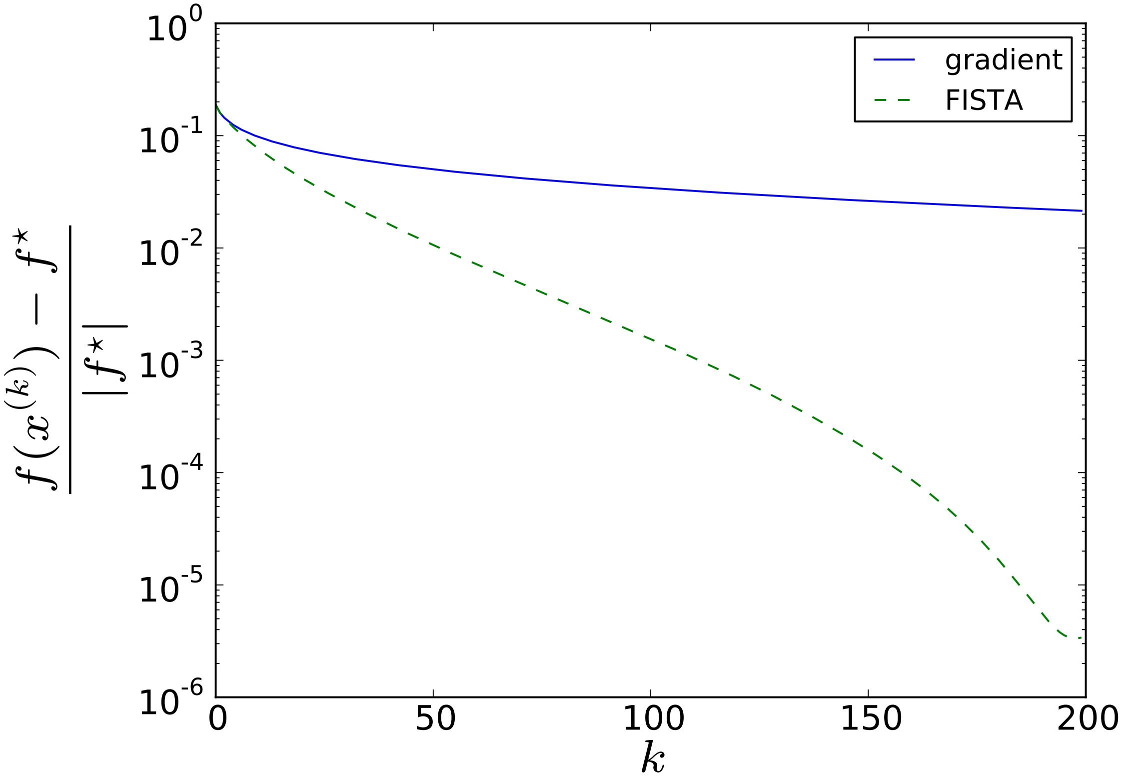


note: $x^{(k)}$ is feasible (in $\mathbf{dom}\, h$); y may be outside $\mathbf{dom}\, h$

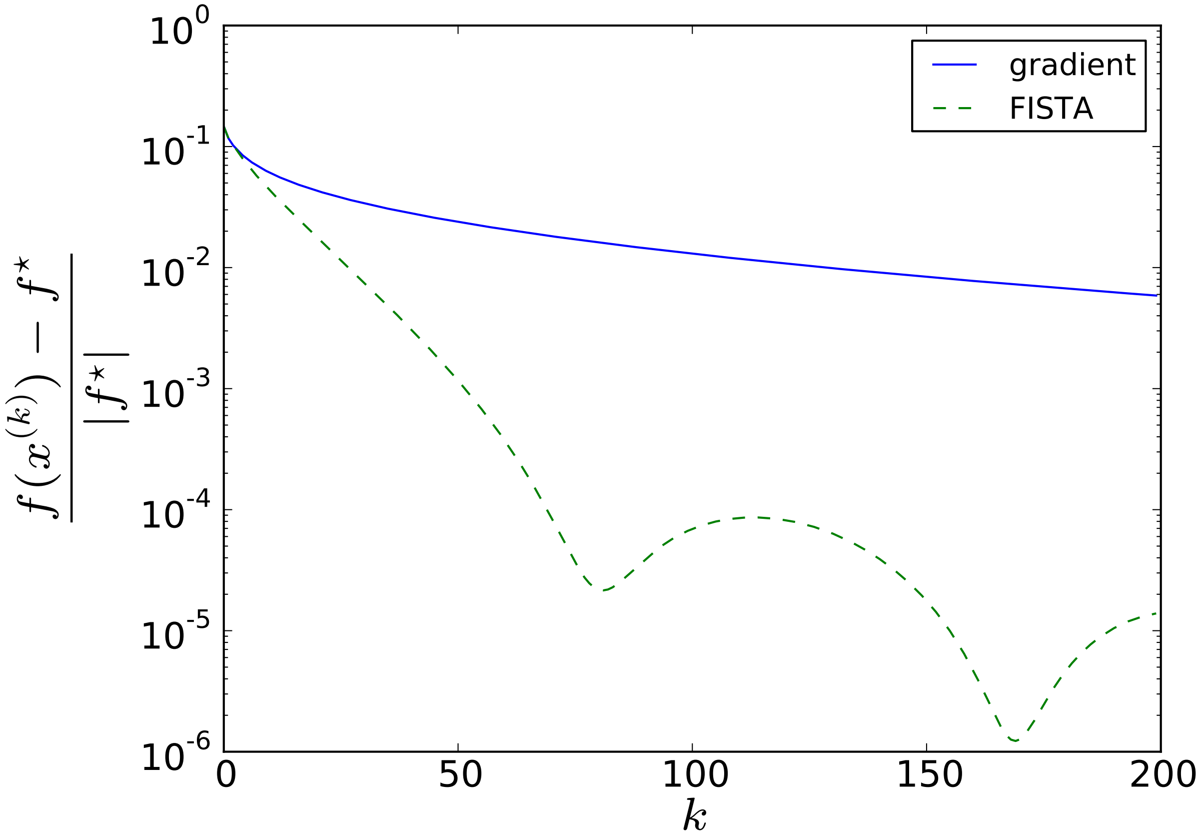
Example

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

randomly generated data with $m = 2000$, $n = 1000$, same fixed step size



another instance



FISTA is not a descent method

Convergence of FISTA

assumptions

- g convex with $\mathbf{dom}\, g = \mathbf{R}^n$; ∇g Lipschitz continuous with constant L:

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \qquad \forall x, y$$

- h is closed and convex (so that $\mathrm{prox}_{th}(u)$ is well defined)
- optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)

convergence result: $f(x^{(k)}) - f^\star$ decreases at least as fast as $1/k^2$

- with fixed step size $t_k = 1/L$
- with suitable line search

Reformulation of FISTA

define $\theta_k = 2/(k+1)$ and introduce an intermediate variable $v^{(k)}$

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

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

substituting expression for $v^{(k)}$ in formula for y gives FISTA of page 7-3

Important inequalities

choice of θ_k: the sequence $\theta_k = 2/(k+1)$ satisfies $\theta_1 = 1$ and

$$\frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2}, \qquad k \geq 2$$

upper bound on g from Lipschitz property (page 1-12)

$$g(u) \leq g(z) + \nabla g(z)^T(u-z) + \frac{L}{2}\|u-z\|_2^2 \qquad \forall u, z$$

upper bound on h from definition of prox-operator (page 6-7)

$$h(u) \leq h(z) + \frac{1}{t}(w-u)^T(u-z) \qquad \forall w,\ u = \mathrm{prox}_{th}(w),\ z$$

Progress in one iteration

define $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $v = v^{(i-1)}$, $v^+ = v^{(i)}$, $t = t_i$, $\theta = \theta_i$

- upper bound from Lipschitz property: if $0 < t \leq 1/L$,

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t}\|x^+ - y\|_2^2 \qquad (1)$$

- upper bound from definition of prox-operator:

$$h(x^+) \leq h(z) + \nabla g(y)^T(z - x^+) + \frac{1}{t}(x^+ - y)^T(z - x^+) \quad \forall z$$

- add the upper bounds and use convexity of g

$$f(x^+) \leq f(z) + \frac{1}{t}(x^+ - y)^T(z - x^+) + \frac{1}{2t}\|x^+ - y\|_2^2 \quad \forall z$$

- make convex combination of upper bounds for $z = x$ and $z = x^\star$

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

conclusion: if the inequality (1) holds at iteration i, then

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

Analysis for fixed step size

take $t_i = t = 1/L$ and apply (2) recursively, using $(1-\theta_i)/\theta_i^2 \leq 1/\theta_{i-1}^2$:

$$\begin{aligned} & \frac{t}{\theta_k^2}\left(f(x^{(k)}) - f^\star\right) + \frac{1}{2}\|v^{(k)} - x^\star\|_2^2 \\ \leq \quad & \frac{(1-\theta_1)t}{\theta_1^2}\left(f(x^{(0)}) - f^\star\right) + \frac{1}{2}\|v^{(0)} - x^\star\|_2^2 \\ = \quad & \frac{1}{2}\|x^{(0)} - x^\star\|_2^2 \end{aligned}$$

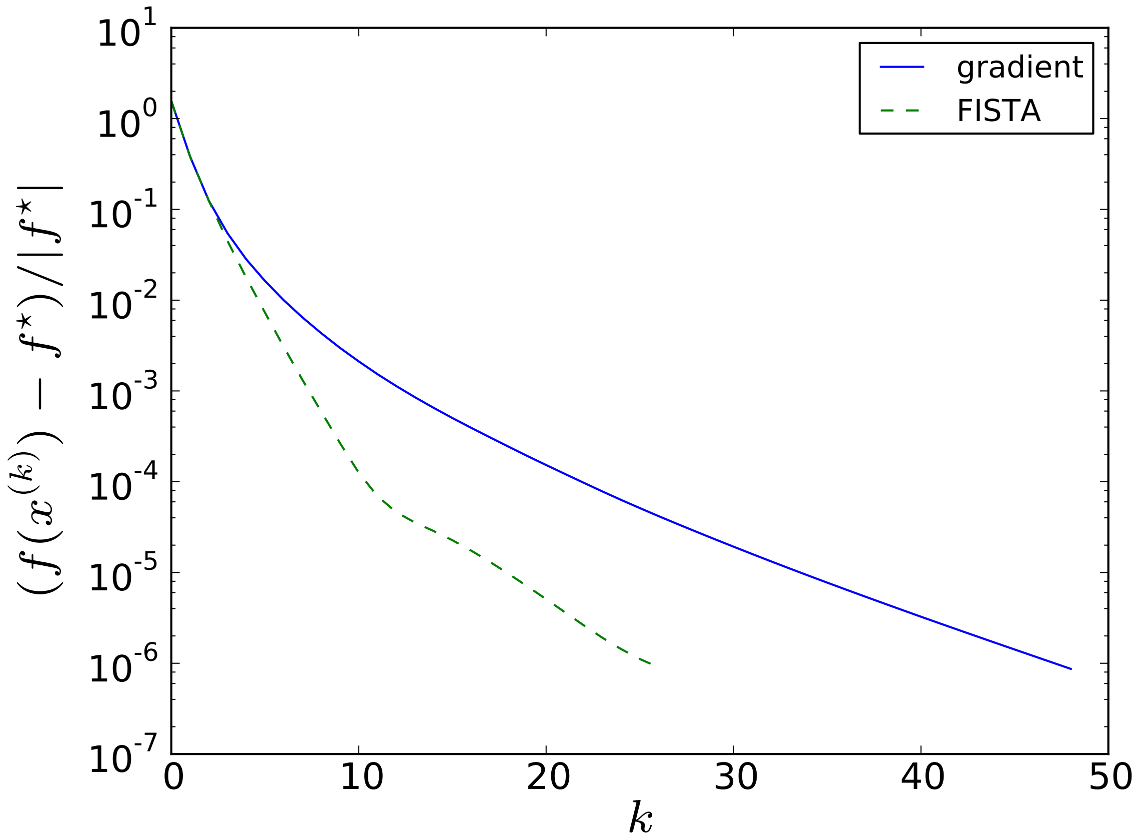
therefore,

$$f(x^{(k)}) - f^\star \leq \frac{\theta_k^2}{2t}\|x^{(0)} - x^\star\|_2^2 = \frac{2L}{(k+1)^2}\|x^{(0)} - x^\star\|_2^2$$

conclusion: reaches $f(x^{(k)}) - f^\star \leq \epsilon$ after $O(1/\sqrt{\epsilon})$ iterations

Example: quadratic program with box constraints

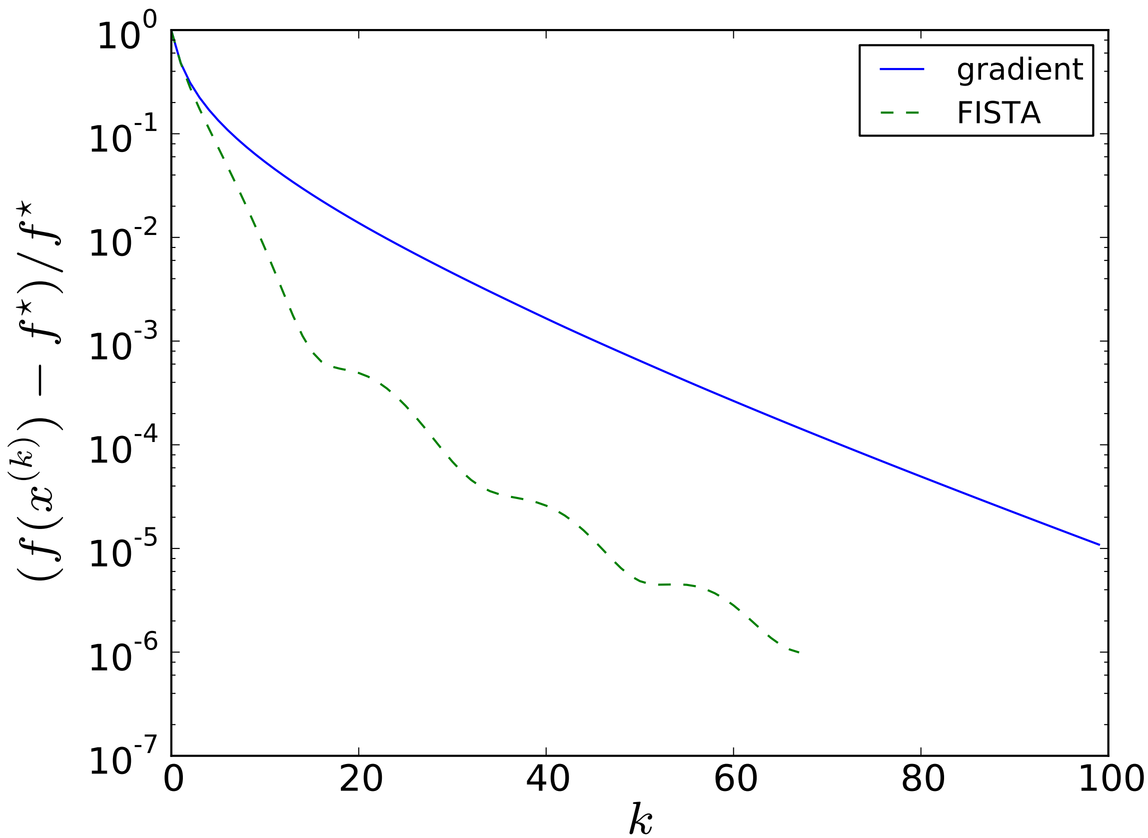
$$\begin{array}{ll} \text{minimize} & (1/2)x^T Ax + b^T x \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1} \end{array}$$



$n = 3000$; fixed step size $t = 1/\lambda_{\max}(A)$

1-norm regularized least-squares

$$\text{minimize} \quad \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_1$$



randomly generated $A \in \mathbf{R}^{2000\times 1000}$; step $t_k = 1/L$ with $L = \lambda_{\max}(A^TA)$

Outline

- fast proximal gradient method (FISTA)
- **FISTA with line search**
- FISTA as descent method
- Nesterov's second method

Key steps in the analysis of FISTA

- the starting point (page 7-10) is the inequality

$$g(x^+) \le g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t}\|x^+ - y\|_2^2 \qquad (1)$$

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

- if (1) holds, then the progress made in iteration i is bounded by

$$\begin{aligned} &\frac{t_i}{\theta_i^2}\left(f(x^{(i)}) - f^\star\right) + \frac{1}{2}\|v^{(i)} - x^\star\|_2^2 \\ &\quad \le \quad \frac{(1-\theta_i)t_i}{\theta_i^2}\left(f(x^{(i-1)}) - f^\star\right) + \frac{1}{2}\|v^{(i-1)} - x^\star\|_2^2 \qquad (2) \end{aligned}$$

- to combine these inequalities recursively, we need

$$\frac{(1-\theta_i)t_i}{\theta_i^2} \le \frac{t_{i-1}}{\theta_{i-1}^2} \qquad (i \ge 2) \qquad (3)$$

- if $\theta_1 = 1$, combining the inequalities (2) from $i = 1$ to k gives the bound

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

conclusion: rate $1/k^2$ convergence if (1) and (3) hold with

$$\frac{\theta_k^2}{t_k} = O(\frac{1}{k^2})$$

FISTA with fixed step size

$$t_k = \frac{1}{L}, \qquad \theta_k = \frac{2}{k+1}$$

these values satisfy (1) and (3) with

$$\frac{\theta_k^2}{t_k} = \frac{4L}{(k+1)^2}$$

FISTA with line search (method 1)

replace update of x in iteration k (page 7-8) with

$$
\begin{array}{l}
t := t_{k-1} \quad \text{(define } t_0 = \hat{t} > 0\text{)} \\
x := \mathrm{prox}_{th}(y - t\nabla g(y)) \\
\text{while } g(x) > g(y) + \nabla g(y)^T(x-y) + \frac{1}{2t}\|x-y\|_2^2 \\
\qquad t := \beta t \\
\qquad x := \mathrm{prox}_{th}(y - t\nabla g(y)) \\
\text{end}
\end{array}
$$

- inequality (1) holds trivially, by the backtracking exit condition
- inequality (3) holds with $\theta_k = 2/(k+1)$ because $t_k \leq t_{k-1}$
- Lipschitz continuity of ∇g guarantees $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
- preserves $1/k^2$ convergence rate because $\theta_k^2/t_k = O(1/k^2)$:

$$
\frac{\theta_k^2}{t_k} \leq \frac{4}{(k+1)^2 t_{\min}}
$$

FISTA with line search (method 2)

replace update of y and x in iteration k (page 7-8) with

$$
\begin{array}{l}
t := \hat{t} > 0 \\
\theta := \text{positive root of } t_{k-1}\theta^2 = t\theta_{k-1}^2(1-\theta) \\
y := (1-\theta)x^{(k-1)} + \theta v^{(k-1)} \\
x := \mathrm{prox}_{th}(y - t\nabla g(y)) \\
\text{while } g(x) > g(y) + \nabla g(y)^T(x-y) + \frac{1}{2t}\|x-y\|_2^2 \\
\qquad t := \beta t \\
\qquad \theta := \text{positive root of } t_{k-1}\theta^2 = t\theta_{k-1}^2(1-\theta) \\
\qquad y := (1-\theta)x^{(k-1)} + \theta v^{(k-1)} \\
\qquad x := \mathrm{prox}_{th}(y - t\nabla g(y)) \\
\text{end}
\end{array}
$$

assume $t_0 = 0$ in the first iteration ($k = 1$), $i.e.$, take $\theta_1 = 1$, $y = x^{(0)}$

discussion

- inequality (1) holds trivially, by the backtracking exit condition
- inequality (3) holds trivially, by construction of θ_k
- Lipschitz continuity of ∇g guarantees $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
- θ_i is defined as the positive root of $\theta_i^2/t_i = (1-\theta_i)\theta_{i-1}^2/t_{i-1}$; hence

$$\frac{\sqrt{t_{i-1}}}{\theta_{i-1}} = \frac{\sqrt{(1-\theta_i)t_i}}{\theta_i} \leq \frac{\sqrt{t_i}}{\theta_i} - \frac{\sqrt{t_i}}{2}$$

  combine inequalities from $i=2$ to k to get $\sqrt{t_1} \leq \dfrac{\sqrt{t_k}}{\theta_k} - \dfrac{1}{2}\displaystyle\sum_{i=2}^{k}\sqrt{t_i}$

- rearranging shows that $\theta_k^2/t_k = O(1/k^2)$:

$$\frac{\theta_k^2}{t_k} \leq \frac{1}{\left(\sqrt{t_1} + \frac{1}{2}\sum\limits_{i=2}^{k}\sqrt{t_i}\right)^2} \leq \frac{4}{(k+1)^2 t_{\min}}$$

Comparison of line search methods

method 1

- uses nonincreasing step sizes (enforces $t_k \leq t_{k-1}$)
- one evaluation of $g(x)$, one prox_{th} evaluation per line search iteration

method 2

- allows non-monotonic step sizes
- one evaluation of $g(x)$, one evaluation of $g(y)$, $\nabla g(y)$, one evaluation of prox_{th} per line search iteration

the two strategies can be combined and extended in various ways

Outline

- fast proximal gradient method (FISTA)
- FISTA with line search
- **FISTA as descent method**
- Nesterov's second method

Descent version of FISTA

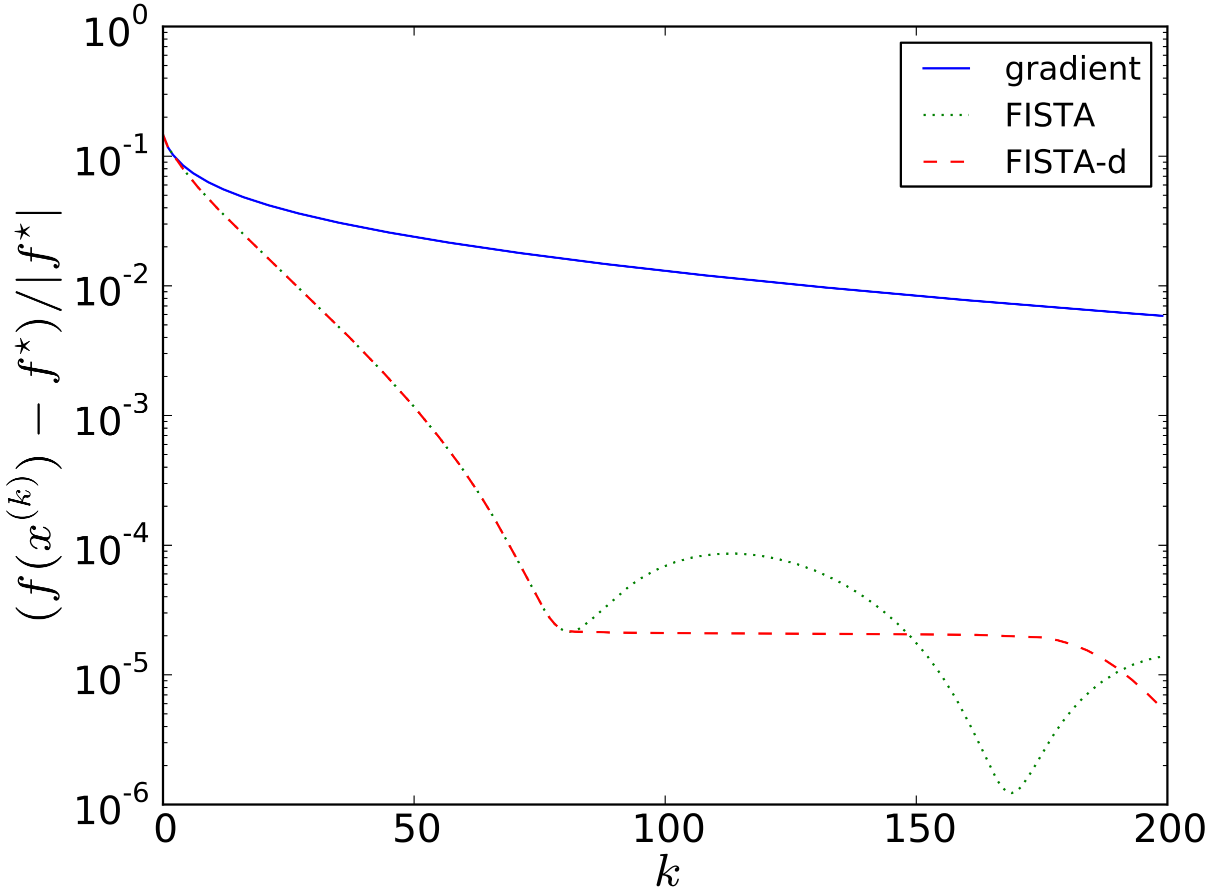
choose $x^{(0)} = v^{(0)}$; for $k \geq 1$, repeat the steps

$$
\begin{aligned}
y &= (1-\theta_k)x^{(k-1)} + \theta_k v^{(k-1)} \\
u &= \mathrm{prox}_{t_k h}\left(y - t_k \nabla g(y)\right) \\
x^{(k)} &= \begin{cases} u & f(u) \leq f(x^{(k-1)}) \\ x^{(k-1)} & \text{otherwise} \end{cases} \\
v^{(k)} &= x^{(k-1)} + \frac{1}{\theta_k}(u - x^{(k-1)})
\end{aligned}
$$

- step 3 implies $f(x^{(k)}) \leq f(x^{(k-1)})$
- use $\theta_k = 2/(k+1)$ and $t_k = 1/L$, or one of the line search methods
- same iteration complexity as original FISTA
- changes on page 7-10: replace x^+ with u and use $f(x^+) \leq f(u)$

Example

(from page 7-6)



Outline

- fast proximal gradient method (FISTA)
- line search strategies
- enforcing descent
- **Nesterov's second method**

Nesterov's second method

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

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

- use $\theta_k = 2/(k+1)$ and $t_k = 1/L$, or one of the line search methods
- identical to FISTA if $h(x) = 0$
- unlike in FISTA, y is feasible (in $\mathbf{dom}\, h$) if we take $x^{(0)} \in \mathbf{dom}\, h$

Convergence of Nesterov's second method

assumptions

- g convex; ∇g is Lipschitz continuous on $\mathbf{dom}\, h \subseteq \mathbf{dom}\, g$

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \qquad \forall x, y \in \mathbf{dom}\, h$$

- h is closed and convex (so that $\mathrm{prox}_{th}(u)$ is well defined)
- optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)

convergence result: $f(x^{(k)}) - f^\star$ decreases at least as fast as $1/k^2$

- with fixed step size $t_k = 1/L$
- with suitable line search

Analysis of one iteration

define $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $v = v^{(i-1)}$, $v^+ = v^{(i)}$, $t = t_i$, $\theta = \theta_i$

- from Lipschitz property if $0 < t \leq 1/L$

$$g(x^+) \leq g(y) + \nabla g(y)^T (x^+ - y) + \frac{1}{2t}\|x^+ - y\|_2^2$$

- plug in $x^+ = (1-\theta)x + \theta v^+$ and $x^+ - y = \theta(v^+ - v)$

$$g(x^+) \leq g(y) + \nabla g(y)^T \left((1-\theta)x + \theta v^+ - y\right) + \frac{\theta^2}{2t}\|v^+ - v\|_2^2$$

- from convexity of g, h

$$\begin{aligned} g(x^+) &\leq (1-\theta)g(x) + \theta\left(g(y) + \nabla g(y)^T (v^+ - y)\right) + \frac{\theta^2}{2t}\|v^+ - v\|_2^2 \\ h(x^+) &\leq (1-\theta)h(x) + \theta h(v^+) \end{aligned}$$

- upper bound on h from p. 7-9 (with $u = v^+$, $w = v - (t/\theta)\nabla g(y)$)

$$h(v^+) \leq h(z) + \nabla g(y)^T(z - v^+) - \frac{\theta}{t}(v^+ - v)^T(v^+ - z) \quad \forall z$$

- combine the upper bounds on $g(x^+)$, $h(x^+)$, $h(v^+)$ with $z = x^\star$

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

this is identical to the final inequality (2) in the analysis of FISTA on p.7-11

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

References

surveys of fast gradient methods

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008)

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)
- Yu. Nesterov, *Gradient methods for minimizing composite objective function* (2007)
- O. Güler, *New proximal point algorithms for convex minimization*, SIOPT (1992)

Nesterov's third method (not covered in this lecture)

- Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming (2005)
- S. Becker, J. Bobin, E.J. Candès, *NESTA: a fast and accurate first-order method for sparse recovery*, SIAM J. Imaging Sciences (2011)