8. Fast gradient methods

- fast proximal gradient method (FISTA)
- 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)**
- Nesterov's second method

Fast proximal gradient method

convex problem with composite objective

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

g differentiable with $\mathbf{dom}\, g = \mathbf{R}^n$; h has inexpensive $\mathbf{prox}_{th}$ operator

algorithm: choose $x^{(0)} = y^{(0)} \in \mathbf{dom}\, h$; for $k \geq 1$

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

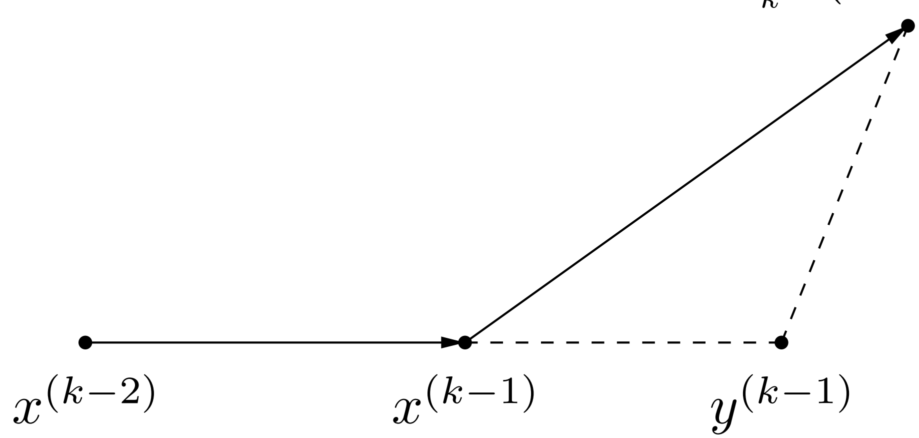
known as FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)

Interpretation

- first iteration $(k = 1)$ is a proximal gradient step at $x^{(0)}$

- next iterations are proximal gradient steps at extrapolated points $y^{(k-1)}$

$$x^{(k)} = \mathbf{prox}_{t_k h}\left(y^{(k-1)} - t_k \nabla g(y^{(k-1)})\right)$$

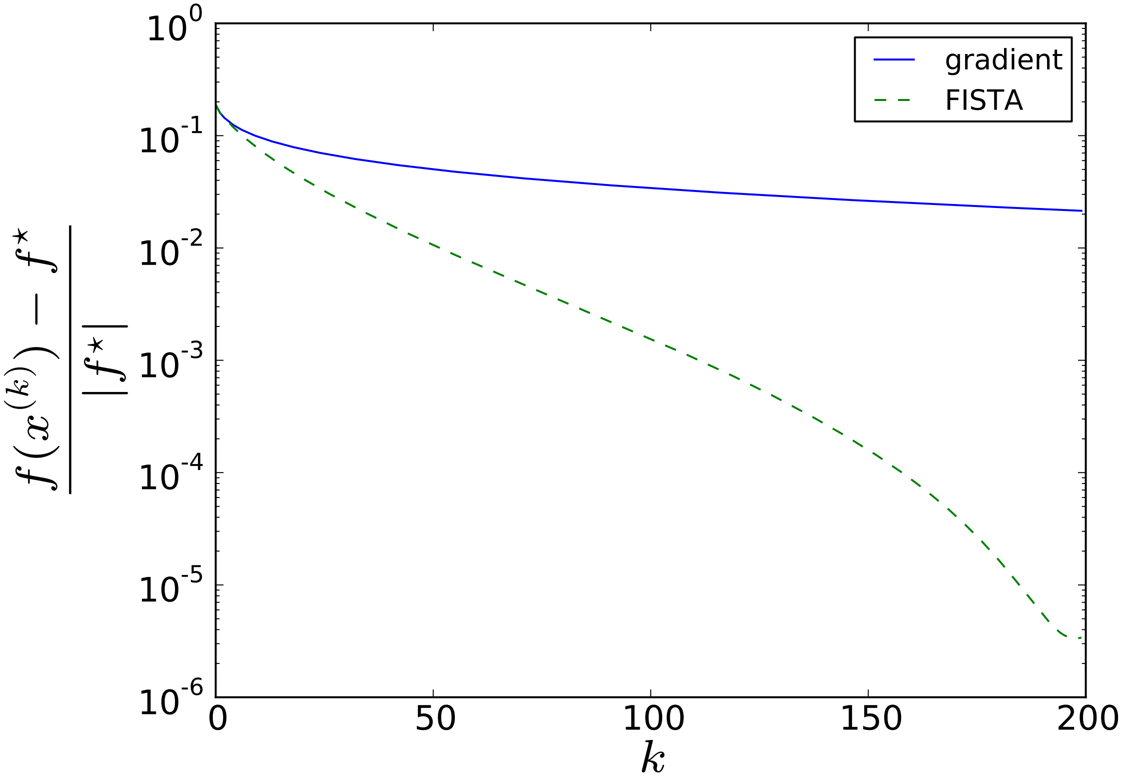


sequence $x^{(k)}$ remains feasible (in $\mathbf{dom}\, h$); sequence $y^{(k)}$ not necessarily

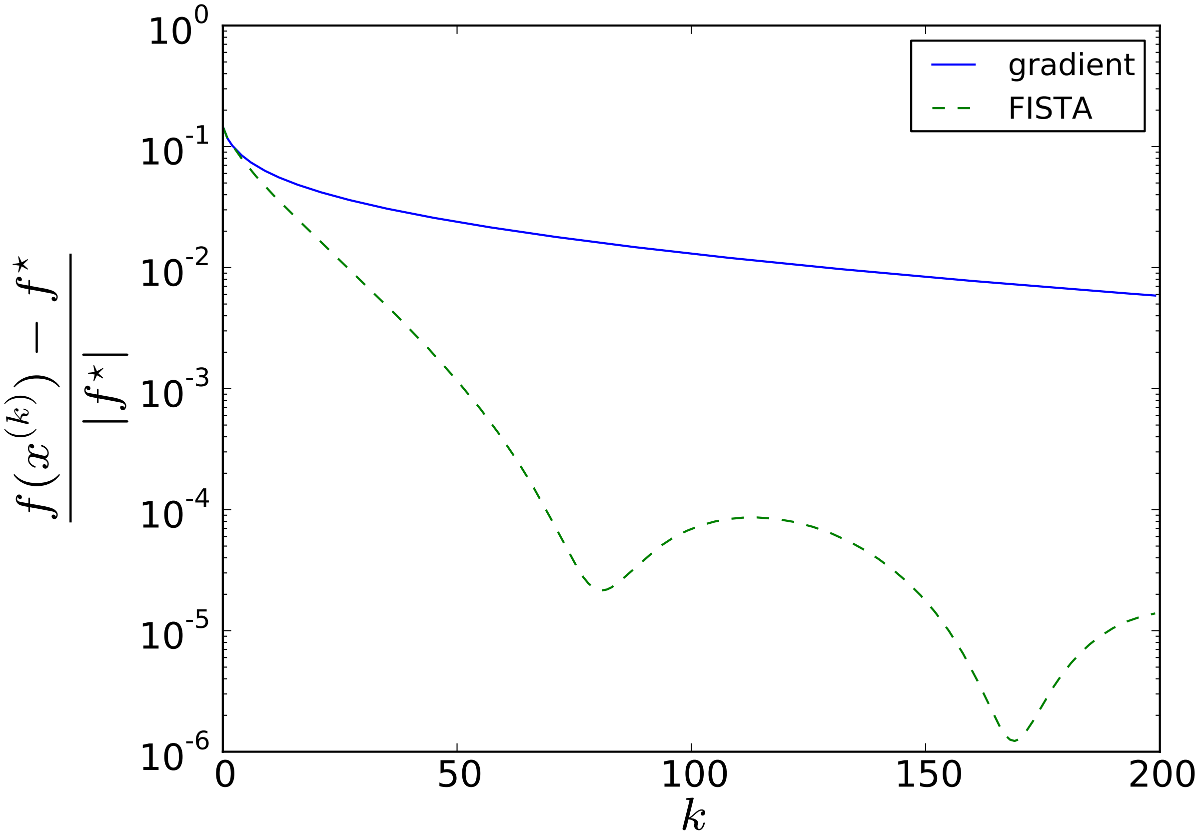
Example

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

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



another instance



FISTA is not a descent method

Convergence of FISTA

assumptions

- optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)
- $\mathbf{dom}\, g = \mathbf{R}^n$ and ∇g is Lipschitz continuous with constant $L > 0$:

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

- h is closed and convex (hence $\mathbf{prox}_{th}(u)$ exists and is unique for all u)

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

- if fixed step size $t_k = 1/L$ is used
- if backtracking line search is used

Reformulation of FISTA

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

algorithm: choose $x^{(0)} = y^{(0)} = v^{(0)} \in \mathbf{dom}\, h$; for $k \geq 1$

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

- substituting expression for $v^{(k)}$ in step 3 gives algorithm on page 8-3
- $\theta_k = 2/(k+1)$ satisfies

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

Key inequalities

upper bound from Lipschitz property

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

property of proximal operators: if $u = \mathbf{prox}_{th}(w)$,

$$h(u) \leq h(z) + \frac{1}{t}(w - u)^T (u - z) \quad \forall z$$

this follows from subgradient characterization of prox-operator (page 4-15)

$$u = \mathbf{prox}_{th}(w) \quad \Longleftrightarrow \quad w - u \in t\partial h(u)$$

Progress in one iteration

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

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

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

- from property 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 $z = x^\star$

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

conclusion: if the inequality (1) holds (for example, if $0 < t \le 1/L$), then

$$
\frac{1}{\theta^2}\left(f(x^+) - f^\star\right) + \frac{1}{2t}\|v^+ - x^\star\|_2^2 \le \frac{1-\theta}{\theta^2}\left(f(x) - f^\star\right) + \frac{1}{2t}\|v - x^\star\|_2^2
$$

Analysis for fixed step size

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

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

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 $O(\sqrt{L/\epsilon})$ iterations

Line search

purpose: determine step size $t = t_k$ in

$$x^+ = \mathbf{prox}_{th}(y - t\nabla g(y)) \qquad \text{(with } x^+ = x^{(k)}, y = y^{(k-1)}\text{)}$$

backtracking line search: start at $t := t_{k-1}$; repeat $t := \beta t$ until

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

for t_0, can choose any positive value $t_0 = \hat{t}$

- from Lipschitz property, $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
- guarantees that inequality (1) on page 8-10 holds
- initialization implies $t_k \leq t_{k-1}$, *i.e.*, step sizes are nonincreasing

Analysis for backtracking line search

apply inequality on page 8-11 recursively to get

$$
\begin{aligned}
\frac{t_{\min}}{\theta_k^2}(f(x^{(k)}) - f^\star) &\leq \frac{t_k}{\theta_k^2}(f(x^{(k)}) - f^\star) + \frac{1}{2}\|v^{(k)} - x^\star\|_2^2 \\
&\leq \frac{t_1(1-\theta_1)}{\theta_1^2}(f(x^{(0)}) - f^\star) + \frac{1}{2}\|v^{(0)} - x^\star\|_2^2 \\
&= \frac{1}{2}\|x^{(0)} - x^\star\|_2^2
\end{aligned}
$$

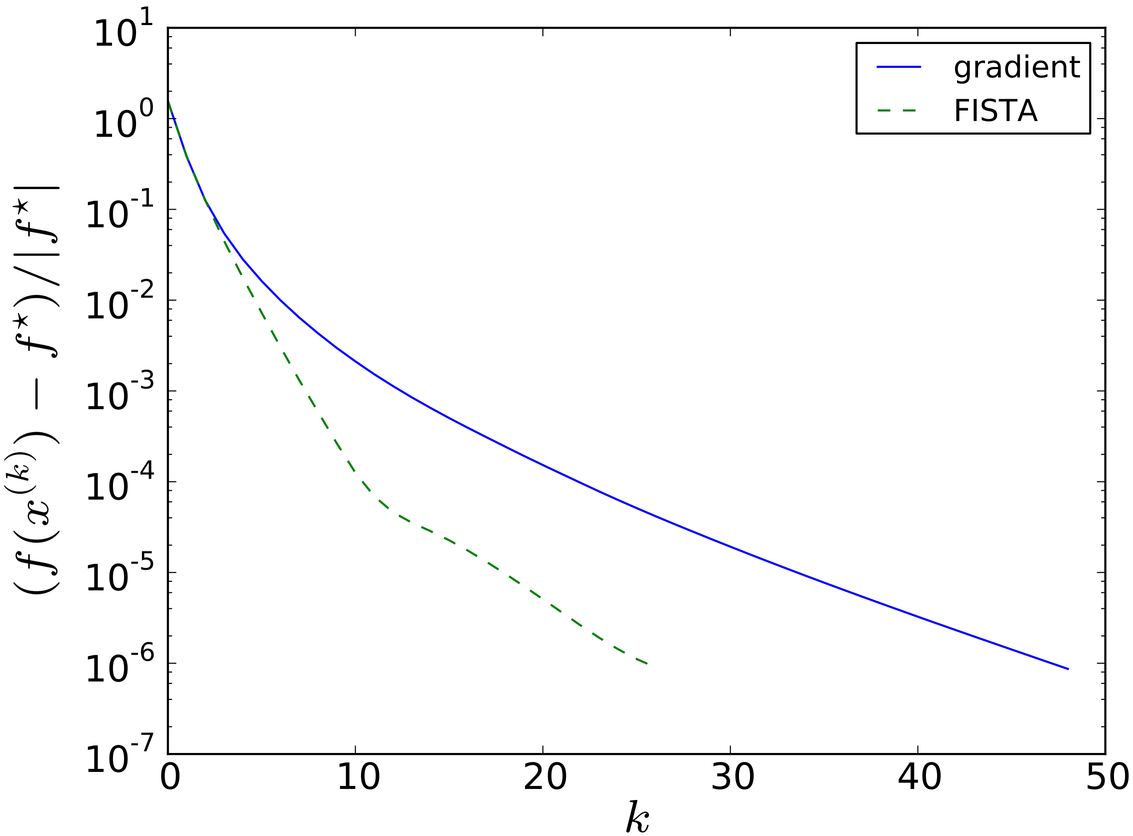
therefore

$$
f(x^{(k)}) - f^\star \leq \frac{2}{(k+1)^2 t_{\min}}\|x^{(0)} - x^\star\|_2^2
$$

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

Example: quadratic program with box constraints

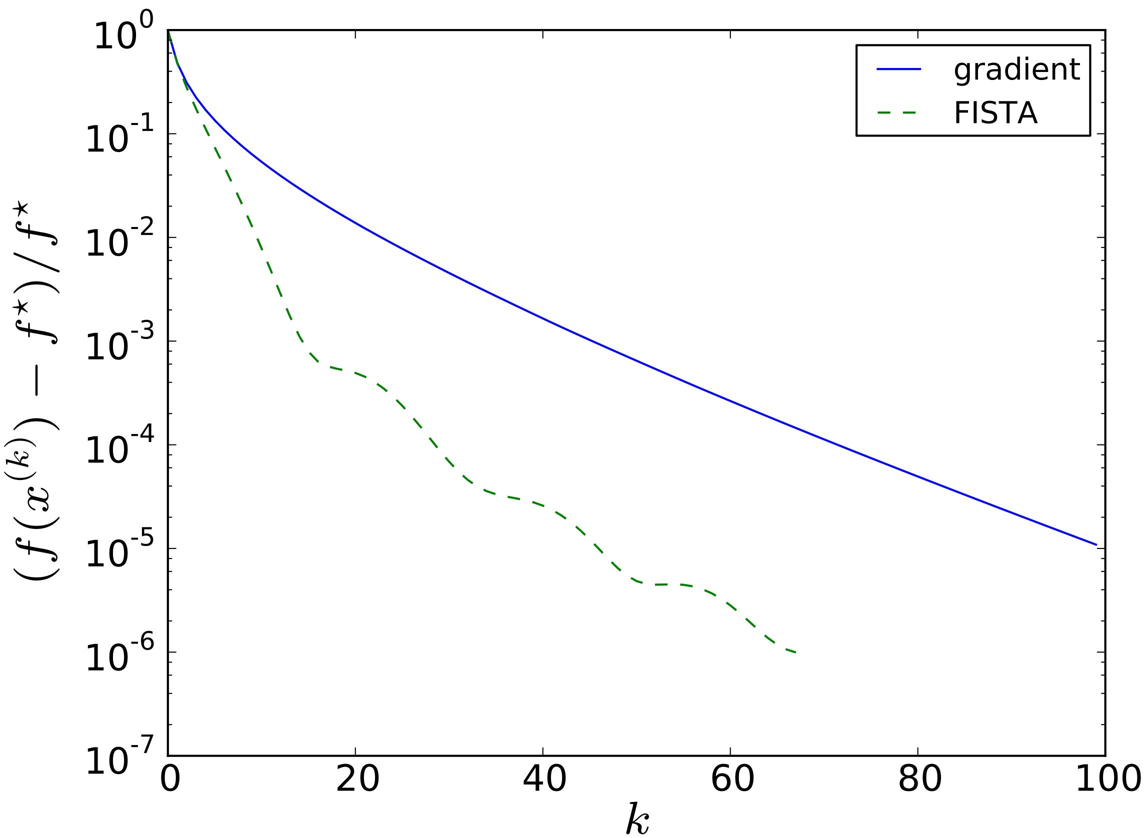
$$\begin{array}{ll} \text{minimize} & (1/2)x^TAx + b^Tx \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1} \end{array}$$



$n = 3000$; 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 $L = \lambda_{\max}(A^TA)$

Example: nuclear norm regularization

$$\text{minimize} \quad g(X) + \|X\|_*$$

g is smooth and convex; variable $X \in \mathbf{R}^{m \times n}$ (with $m \geq n$)

nuclear norm

$$\|X\|_* = \sum_i \sigma_i(X)$$

- $\sigma_1(X) \geq \sigma_2(X) \geq \cdots$ are the singular values of X
- the dual norm of the matrix norm $\|\cdot\|$ (maximum singular value)
- for diagonal X, reduces to the 1-norm of $\mathbf{diag}(X)$
- popular as penalty function that promotes low rank

prox operator of $\mathbf{prox}_{th}(X)$ for $h(X) = \|X\|_*$

$$\mathbf{prox}_{th}(X) = \underset{U}{\operatorname{argmin}} \left(\|U\|_* + \frac{1}{2t} \|U - X\|_F^2 \right)$$

- take singular value decomposition $X = P\, \mathbf{diag}(\sigma_1, \ldots, \sigma_n) Q^T$
- apply soft thresholding to singular values:

$$\mathbf{prox}_{th}(Y) = P\, \mathbf{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n) Q^T$$

where

$$\hat{\sigma}_k = \sigma_k - t \quad (\sigma_k \geq t), \qquad \hat{\sigma}_k = 0 \quad (\sigma_k \leq t)$$

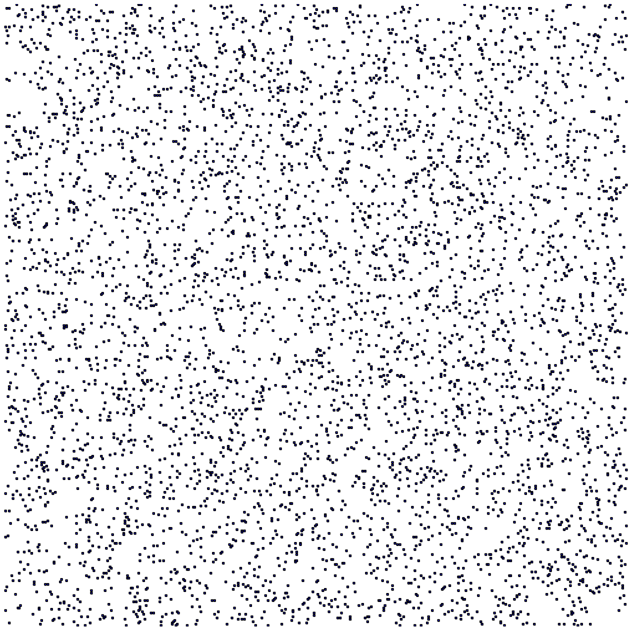
Approximate low-rank completion

$$\text{minimize} \quad \sum_{(i,j)\in N} (X_{ij} - A_{ij})^2 + \gamma \|X\|_*$$

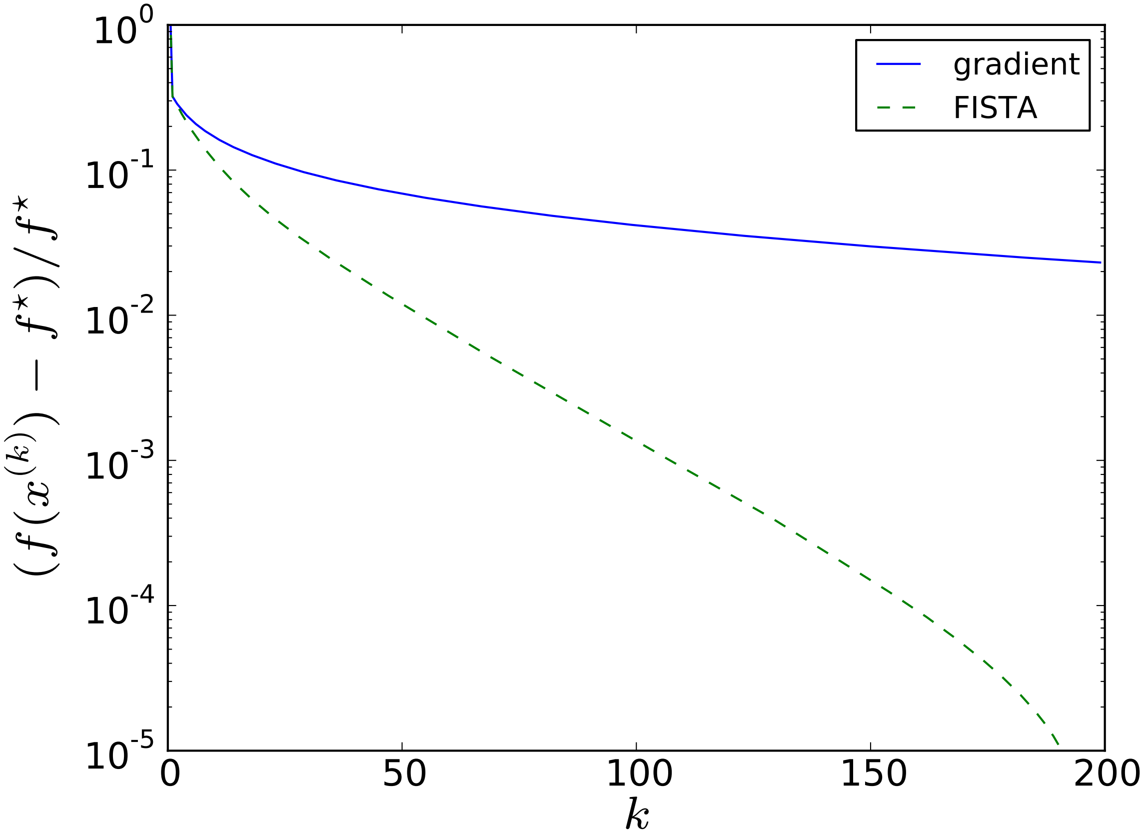
- entries $(i,j) \in N$ are approximately specified $(X_{ij} \approx A_{ij})$; rest is free
- nuclear norm regularization added to obtain low rank X

example

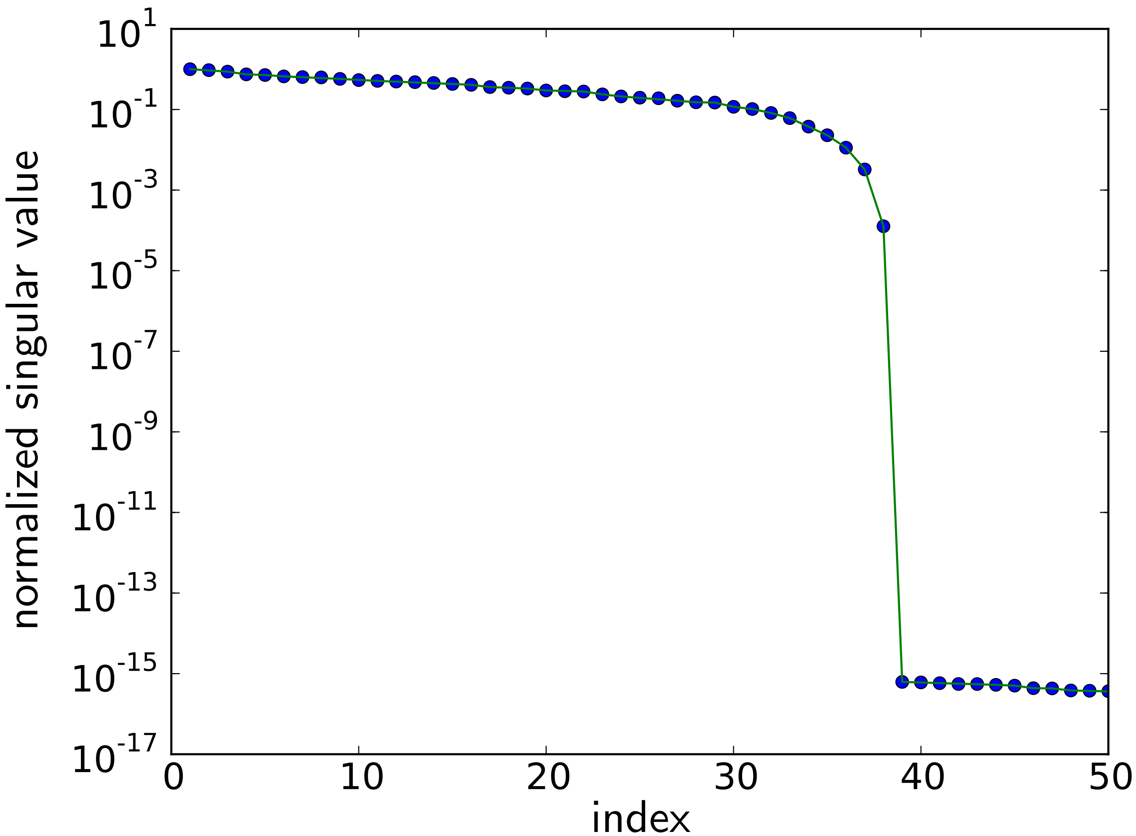
$m = n = 500$
5000 specified entries



convergence (fixed step size $t = 1/L$)



result



optimal X has rank 38; relative error in specified entries is 9%

Descent version of FISTA

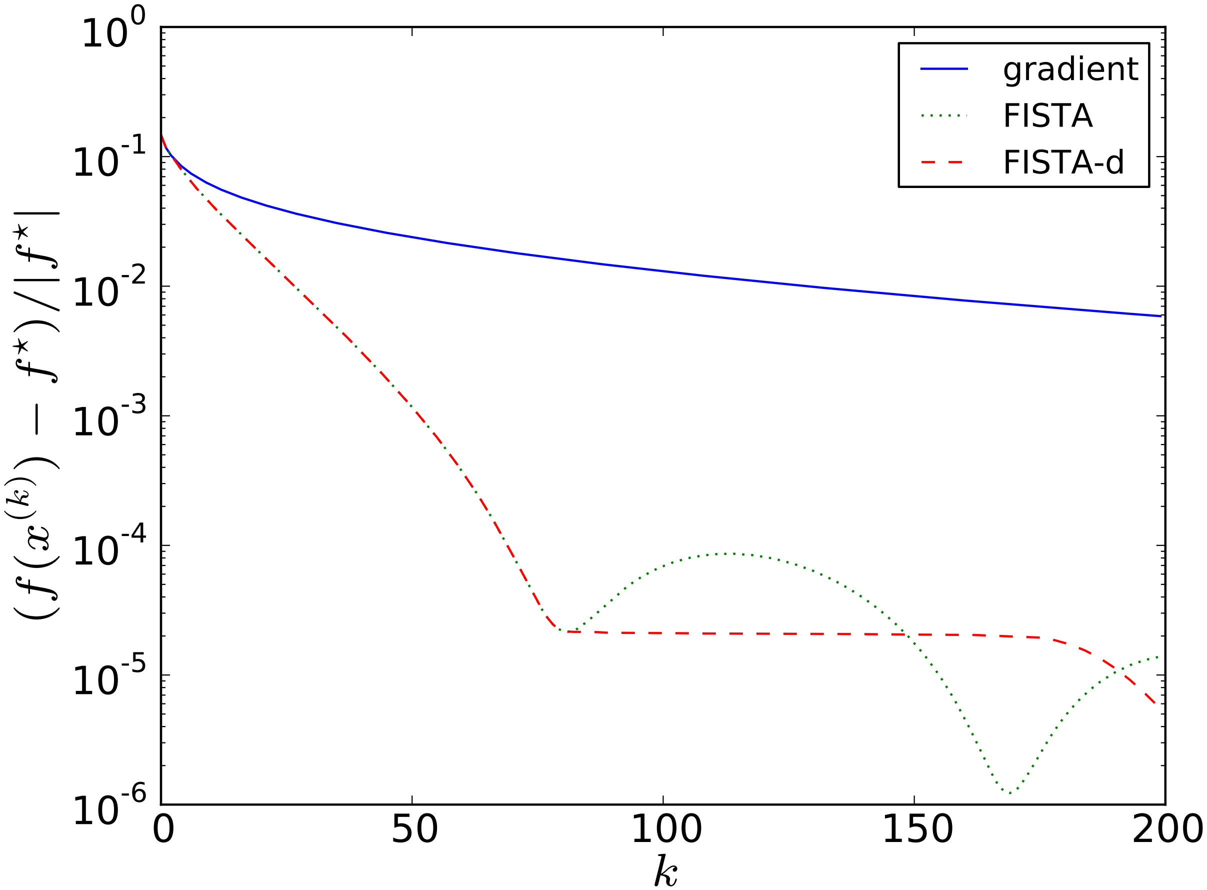
choose $x^{(0)} \in \mathbf{dom}\, h$ and $y^{(0)} = x^{(0)}$; for $k \geq 1$

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

- step 2 implies $f(x^{(k)}) \leq f(x^{(k-1)})$
- same iteration complexity as original FISTA
- changes on p. 8-10: replace x^+ with $u^+ = u^{(i)}$ and use $f(x^+) \leq f(u^+)$

Example

(from page 8-6)



Outline

- fast proximal gradient method (FISTA)
- **Nesterov's second method**

Nesterov's second method

algorithm: choose $x^{(0)} = y^{(0)} = v^{(0)} \in \mathbf{dom}\, h$; for $k \geq 1$

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

- $\theta_k = 2/(k+1)$
- can be shown to be identical to FISTA if $h(x) = 0$
- unlike in FISTA, $y^{(k)}$ remains feasible ($i.e.$, in $\mathbf{dom}\, h$)

Convergence of Nesterov's second method

assumptions

- optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)
- ∇g is Lipschitz continuous on $\mathbf{dom}\, h \subseteq \mathbf{dom}\, g$ with constant $L > 0$:

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

- h is closed and convex

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

- if fixed step size $t_k = 1/L$ is used
- if backtracking line search is used

Analysis of one iteration

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

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

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

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

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

- from convexity of g, h

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

- from property of prox-operator on page 8-9

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

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

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

the same final inequality as in the analysis of FISTA on page 8-11

conclusion: same $1/k^2$ complexity as FISTA

- for fixed step size $t_i = 1/L$
- backtracking line search that ensures (2) and $t_i \le t_{i-1}$

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)

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)