3. Proximal gradient method

- introduction
- proximal mapping
- proximal gradient method
- convergence analysis
- accelerated proximal gradient method
- forward-backward method

Proximal mapping

the **proximal mapping** (or proximal operator) of a convex function h is

$$\mathbf{prox}_h(x) = \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

examples

- $h(x) = 0$: $\mathbf{prox}_h(x) = x$
- $h(x) = I_C(x)$ (indicator function of C): $\mathbf{prox}_h$ is projection on C

$$\mathbf{prox}_h(x) = P_C(x) = \underset{u \in C}{\operatorname{argmin}} \|u - x\|_2^2$$

- $h(x) = t\|x\|_1$: $\mathbf{prox}_h$ is shrinkage (soft threshold) operation

$$\mathbf{prox}_h(x)_i = \begin{cases} x_i - t & x_i \geq t \\ 0 & |x_i| \leq t \\ x_i + t & x_i \leq -t \end{cases}$$

Proximal gradient method

unconstrained problem with cost function split in two components

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

- g convex, differentiable, with $\mathbf{dom}\, g = \mathbf{R}^n$
- h closed, convex, possibly nondifferentiable; $\mathbf{prox}_h$ is inexpensive

proximal gradient algorithm

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

$t_k > 0$ is step size, constant or determined by line search

Interpretation

$$x^+ = \mathbf{prox}_{th}\left(x - t\nabla g(x)\right)$$

from definition of proximal operator:

$$\begin{aligned}
x^+ &= \operatorname*{argmin}_u \left(h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2 \right) \\
&= \operatorname*{argmin}_u \left(h(u) + g(x) + \nabla g(x)^T(u - x) + \frac{1}{2t}\|u - x\|_2^2 \right)
\end{aligned}$$

x^+ minimizes $h(u)$ plus a simple quadratic local model of $g(u)$ around x

Examples

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

gradient method: $h(x) = 0$, *i.e.*, minimize $g(x)$

$$x^{(k)} = x^{(k-1)} - t_k \nabla g(x^{(k-1)})$$

gradient projection method: $h(x) = I_C(x)$, *i.e.*, minimize $g(x)$ over C

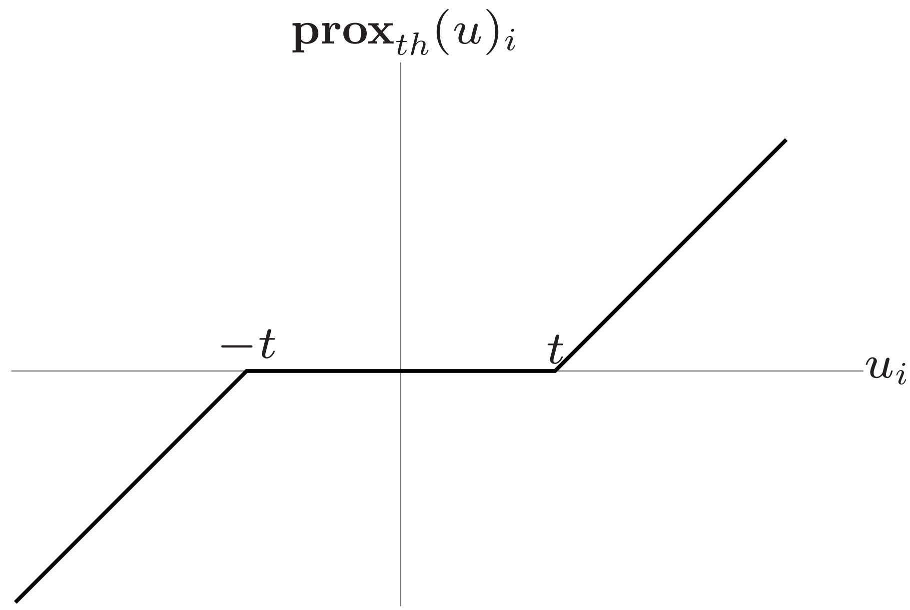
$$x^{(k)} = P_C\left(x^{(k-1)} - t_k \nabla g(x^{(k-1)})\right)$$

iterative soft-thresholding: $h(x) = \|x\|_1$, *i.e.*, minimize $g(x) + \|x\|_1$

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

and

$$\mathbf{prox}_{th}(u)_i = \begin{cases} u_i - t & u_i \geq t \\ 0 & -t \leq u_i \leq t \\ u_i + t & u_i \geq t \end{cases}$$



Outline

- introduction
- **proximal mapping**
- proximal gradient method
- convergence analysis
- accelerated proximal gradient method
- forward-backward method

Definition

proximal mapping associated with closed convex h

$$\mathbf{prox}_h(x) = \operatorname*{argmin}_u \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

it can be shown that $\mathbf{prox}_h(x)$ exists and is unique for all x

subgradient characterization

from optimality conditions of minimization in the definition:

$$u = \mathbf{prox}_h(x) \quad \Longleftrightarrow \quad x - u \in \partial h(u)$$

Projection

recall the definition of **indicator function** of a set C

$$I_C(x) = \begin{cases} 0 & x \in C \\ +\infty & \text{otherwise} \end{cases}$$

I_C is closed and convex if C is a closed convex set

proximal mapping of I_C is the **Euclidean projection** on C

$$\begin{aligned} \mathbf{prox}_{I_C}(x) &= \underset{u \in C}{\operatorname{argmin}} \|u - x\|_2^2 \\ &= P_C(x) \end{aligned}$$

we will see that proximal mappings have many properties of projections

Nonexpansiveness

if $u = \mathbf{prox}_h(x)$, $\hat{u} = \mathbf{prox}_h(\hat{x})$, then

$$(u - \hat{u})^T (x - \hat{x}) \geq \|u - \hat{u}\|_2^2$$

$\mathbf{prox}_h$ is **firmly nonexpansive**, or **co-coercive** with constant 1

- follows from characterization of p.3-7 and monotonicity (p.1-25)

$$x - u \in \partial h(u), \quad \hat{x} - \hat{u} \in \partial h(\hat{u}) \quad \Longrightarrow \quad (x - u - \hat{x} + \hat{u})^T (u - \hat{u}) \geq 0$$

- implies (from Cauchy-Schwarz inequality)

$$\|u - \hat{u}\|_2 \leq \|x - \hat{x}\|_2$$

$\mathbf{prox}_h$ is **nonexpansive**, or **Lipschitz continuous** with constant 1

Proximal mapping and conjugate

$$x = \mathbf{prox}_h(x) + \mathbf{prox}_{h^*}(x)$$

proof: define $u = \mathbf{prox}_h(x)$, $v = x - u$

- from subgradient characterization on page 3-7, $v \in \partial h(u)$
- hence (from page 1-38) $u = x - v \in \partial h^*(v)$, *i.e.*, $v = \mathbf{prox}_{h^*}(x)$

example: let L be a subspace of $\mathbf{R}^n$, $L^\perp$ its orthogonal complement

$$h(u) = I_L(u), \qquad h^*(v) = I_{L^\perp}(v)$$

property reduces to orthogonal decomposition

$$x = P_L(x) + P_{L^\perp}(x)$$

Some useful properties

separable sum: $h : \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} \to \mathbf{R}$ with $h(x_1, x_2) = h_1(x_1) + h_2(x_2)$

$$\mathbf{prox}_h(x_1, x_2) = \big(\mathbf{prox}_{h_1}(x_1), \mathbf{prox}_{h_2}(x_2)\big)$$

scaling and translation of argument: $h(x) = f(tx + a)$ with $t \neq 0$

$$\mathbf{prox}_h(x) = \frac{1}{t}\left(\mathbf{prox}_{t^2 f}(tx + a) - a\right)$$

conjugate: from previous page and $(th)^*(y) = th^*(y/t)$

$$\mathbf{prox}_{th^*}(x) = x - t\,\mathbf{prox}_{h/t}(x/t)$$

Examples

quadratic function

$$h(x) = \frac{1}{2}x^T Ax + b^T x + c, \qquad \mathbf{prox}_{th}(x) = (I + tA)^{-1}(x - tb)$$

Euclidean norm: $h(x) = \|x\|_2$

$$\mathbf{prox}_{th}(x) = \begin{cases} (1 - t/\|x\|_2)x & \|x\|_2 \geq t \\ 0 & \text{otherwise} \end{cases}$$

logarithmic barrier

$$h(x) = -\sum_{i=1}^{n} \log x_i, \qquad \mathbf{prox}_{th}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}, \quad i = 1, \ldots, n$$

Norms

prox-operator of general norm: conjugate of $h(x) = \|x\|$ is

$$h^*(y) = \begin{cases} 0 & \|y\|_* \le 1 \\ +\infty & \text{otherwise} \end{cases}$$

$i.e.$, the indicator function of the dual norm ball $B = \{y \mid \|y\|_* \le 1\}$

if projection on dual norm ball is inexpensive, we can therefore use

$$\mathbf{prox}_{th}(x) = x - tP_B(x/t)$$

distance in general norm: $h(x) = \|x - a\|$

$$\mathbf{prox}_{th}(x) = x - tP_B\left(\frac{x-a}{t}\right)$$

for $h(x) = \|x\|_1$, these expressions reduce to soft-threshold operations

Functions associated with convex sets

support function (or conjugate of the indicator function)

$$h(x) = \sup_{y \in C} x^T y, \qquad \mathbf{prox}_{th}(x) = x - tP_C(x/t)$$

squared distance

$$h(x) = \frac{1}{2}\mathbf{dist}(x, C)^2, \qquad \mathbf{prox}_{th}(x) = x + \frac{t}{1+t}(P_C(x) - x)$$

distance: $h(x) = \mathbf{dist}(x, C)$

$$\mathbf{prox}_{th}(x) = \begin{cases} x + \dfrac{t}{\mathbf{dist}(x, C)}(P_C(x) - x) & \mathbf{dist}(x, C) \geq t \\ P_C(x) & \text{otherwise} \end{cases}$$

Outline

- introduction
- proximal mapping
- **proximal gradient method**
- convergence analysis
- accelerated proximal gradient method
- forward-backward method

Gradient map

proximal gradient iteration for minimizing $g(x) + h(x)$

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

can write as $x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})$ where

$$G_t(x) = \frac{1}{t}\left(x - \mathbf{prox}_{th}(x - t\nabla g(x))\right)$$

- from subgradient definition of $\mathbf{prox}$ (page 3-7),

$$G_t(x) \in \nabla g(x) + \partial h\left(x - tG_t(x)\right) \tag{3.1}$$

- $G_t(x) = 0$ if and only if x minimizes $f(x) = g(x) + h(x)$

Line search

to determine step size t in

$$x^+ = x - tG_t(x)$$

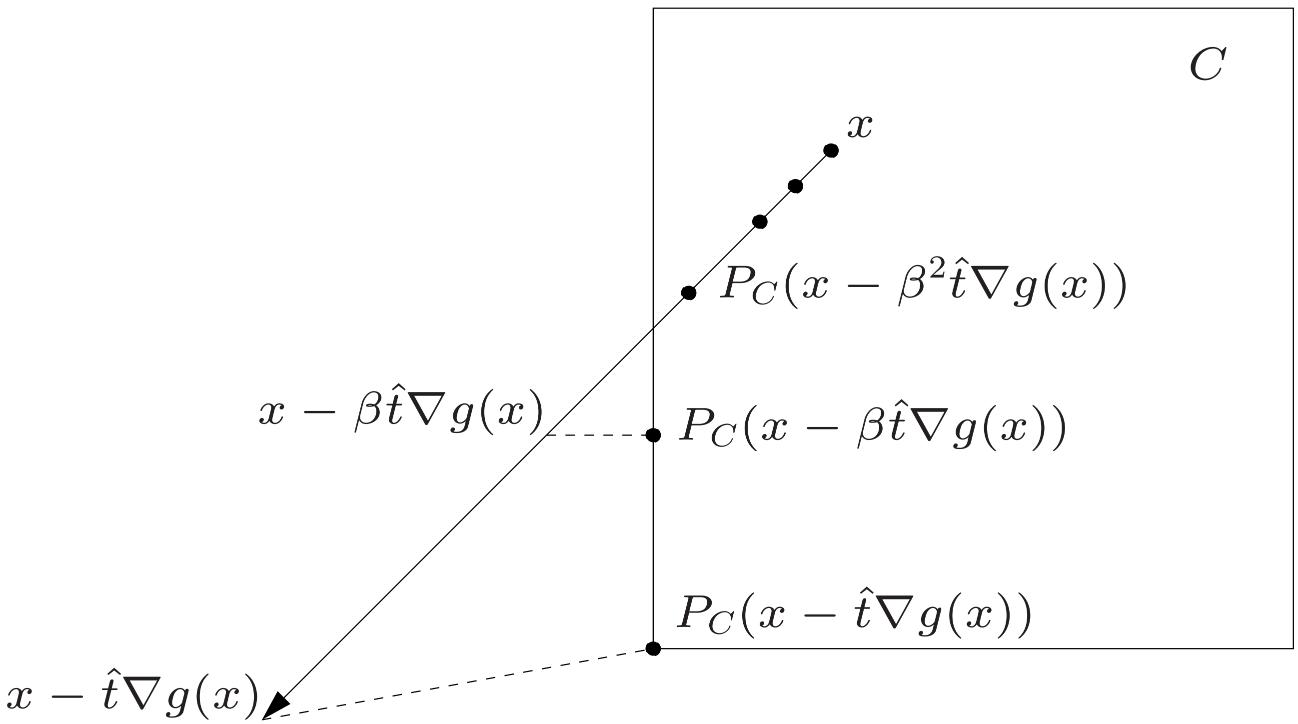
start at some $t := \hat{t}$; repeat $t := \beta t$ (with $0 < \beta < 1$) until

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

- requires one $\mathbf{prox}$ evaluation per line search iteration
- inequality is motivated by convergence analysis (see later)
- many other types of line search work

example: line search for projected gradient method

$$x^+ = x - tG_t(x) = P_C(x - t\nabla g(x))$$



(sometimes called 'arc search')

Outline

- introduction
- proximal mapping
- proximal gradient method
- **convergence analysis**
- accelerated proximal gradient method
- forward-backward method

Convergence of proximal gradient method

assumptions

- ∇g is Lipschitz continuous with constant $L > 0$

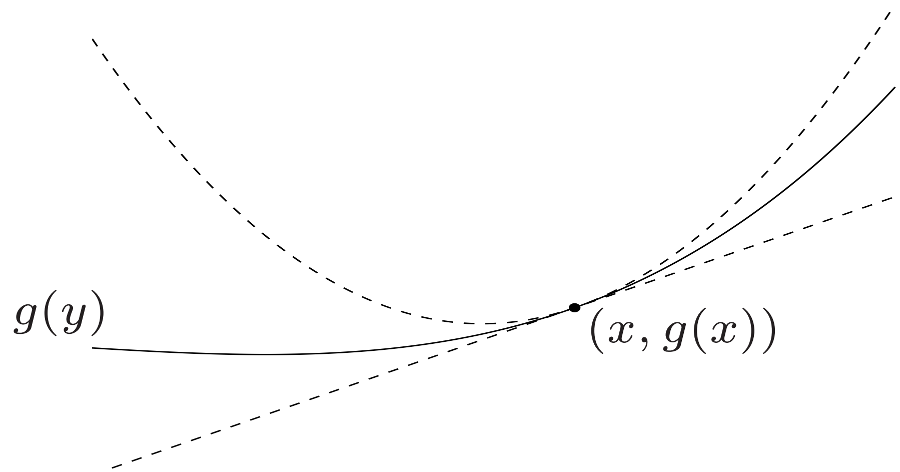
$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

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

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

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

Quadratic upper bound from Lipschitz property



- affine lower bound from convexity

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \qquad \forall x, y$$

- quadratic upper bound from Lipschitz property

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

proof of upper bound (define $v = y - x$)

$$\begin{aligned}
g(y) &= g(x) + \nabla g(x)^T v + \int_0^1 (\nabla g(x+tv) - \nabla g(x))^T v \, dt \\
&\leq g(x) + \nabla g(x)^T v + \int_0^1 \|\nabla g(x+tv) - \nabla g(x)\|_2 \|v\|_2 \, dt \\
&\leq g(x) + \nabla g(x)^T v + \int_0^1 Lt\|v\|_2^2 \, dt \\
&= g(x) + \nabla g(x)^T v + \frac{L}{2}\|v\|_2^2
\end{aligned}$$

Consequences of Lipschitz assumption

- from page 3-19 with $y = x - tG_t(x)$,

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

- therefore, the line search inequality

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (3.2)$$

  is satisfied for $0 \leq t \leq 1/L$

- backtracking line search starting at $t = \hat{t}$ terminates with

$$t \geq t_{\min} \stackrel{\Delta}{=} \min\{\hat{t}, \beta/L\}$$

A global inequality

if the line search inequality (3.2) holds, then for all z,

$$f(x - tG_t(x)) \leq f(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|_2^2 \qquad (3.3)$$

proof (with $v = G_t(x) - \nabla g(x)$)

$$\begin{aligned}
f(x - tG_t(x)) &\leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\
&\leq g(z) + \nabla g(x)^T(x - z) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + h(z) + v^T(x - z - tG_t(x)) \\
&= g(z) + h(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|_2^2
\end{aligned}$$

line 2 follows from convexity of g and h, and $v \in \partial h(x - tG_t(x))$

Progress in one iteration

$$x^+ = x - tG_t(x)$$

- inequality (3.3) with $z = x$ shows the algorithm is a descent method:

$$f(x^+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

- inequality (3.3) with $z = x^\star$:

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

(hence, $\|x^+ - x^\star\|_2 \leq \|x - x^\star\|_2$, $i.e.$, distance to optimal set decreases)

Analysis for fixed step size

add inequalities for $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t = 1/L$

$$\begin{aligned}\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) &\leq \frac{1}{2t}\sum_{i=1}^{k}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2\right) \\ &\leq \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2\end{aligned}$$

since $f(x^{(i)})$ is nonincreasing,

$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2kt}\|x^{(0)} - x^\star\|_2^2$$

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

Analysis with line search

add inequalities for $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t = t_i \geq t_{\min}$

$$\begin{aligned}
\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) &\leq \sum_{i=1}^{k} \frac{1}{2t_i}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \\
&\leq \frac{1}{2t_{\min}} \sum_{i=1}^{k}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \\
&= \frac{1}{2t_{\min}}\left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2\right)
\end{aligned}$$

since $f(x^{(i)})$ is nonincreasing,

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

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

Outline

- introduction
- proximal mapping
- proximal gradient method
- convergence analysis
- **accelerated proximal gradient method**
- forward-backward method

Accelerated proximal gradient method

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

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

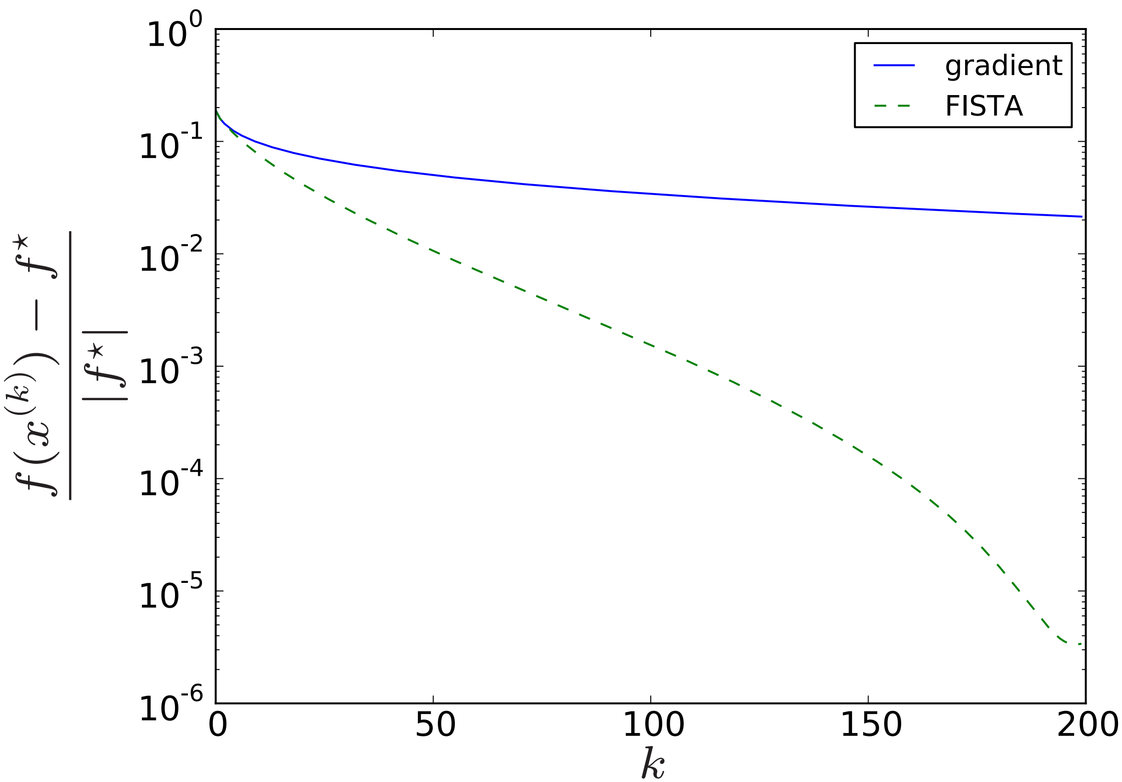
- t_k is fixed or determined by line search
- same complexity per iteration as basic proximal gradient method
- also known as proximal gradient method with extrapolation, FISTA

Nesterov (1983, 2004), Beck and Teboulle (2009), Tseng (2008)

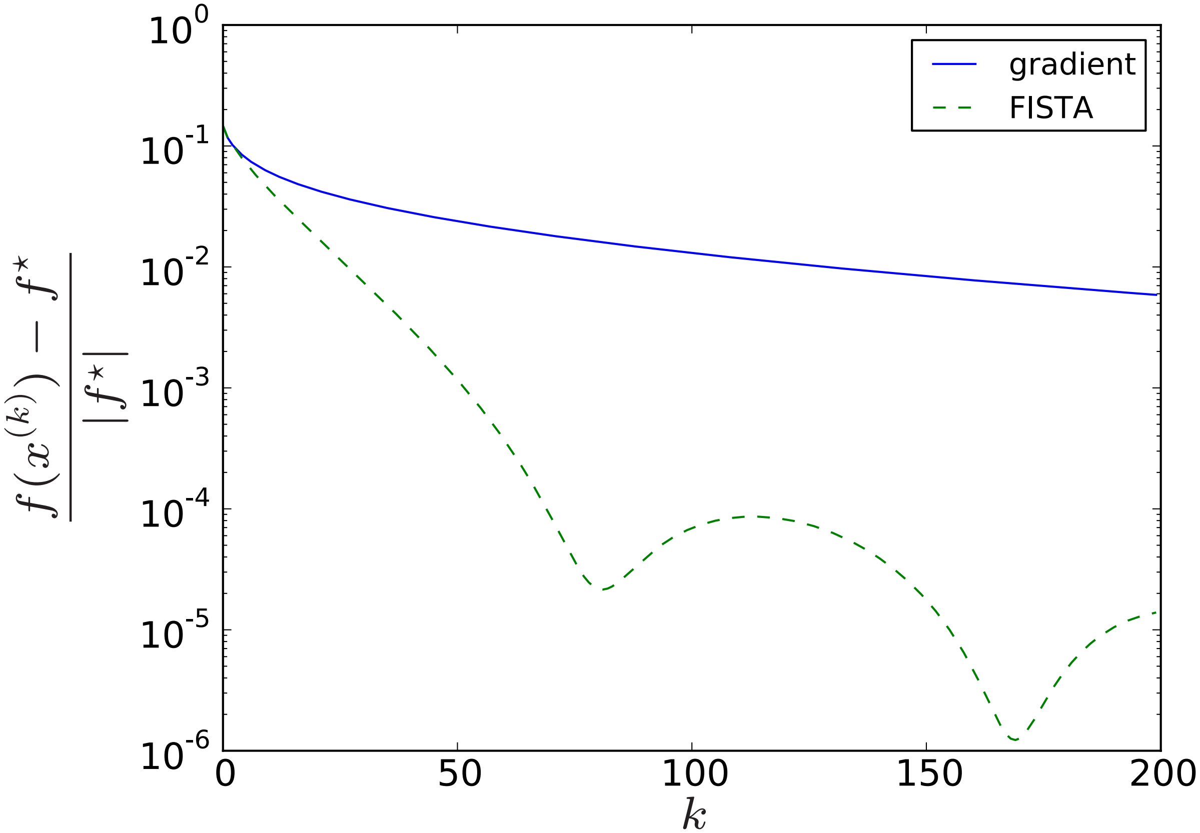
Example

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

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



another instance



Line search

purpose: determine step size t_k in

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

algorithm: start at $t := t_{k-1}$ and repeat $t := \beta t$ until

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

(where $y = y^{(k-1)}$)

- for t_0, can choose any positive value $t_0 = \hat{t}$
- this line search method implies $t_k \leq t_{k-1}$

Convergence of accelerated proximal gradient method

assumptions

- ∇g is Lipschitz continuous with constant $L > 0$

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

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

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

Consequences of Lipschitz assumption

from page 3-21 and 3-22

- the line search inequality

$$g(y - tG_t(y)) \leq g(y) - t\nabla g(y)^T G_t(y) + \frac{t}{2}\|G_t(y)\|_2^2 \qquad (3.4)$$

  holds for $0 \leq t \leq 1/L$

- backtracking line search terminates with $t \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

- if t satisfies the line search inequality, then, for all z,

$$f(y - tG_t(y)) \leq f(z) + G_t(y)^T(y - z) - \frac{t}{2}\|G_t(y)\|_2^2 \qquad (3.5)$$

Notation

define $v^{(0)} = x^{(0)}$ and, for $k \geq 1$,

$$\theta_k = \frac{2}{k+1}, \qquad v^{(k)} = x^{(k-1)} + \frac{1}{\theta_k}(x^{(k)} - x^{(k-1)})$$

- update of $y^{(k)}$ can be written as

$$y^{(k)} = (1 - \theta_{k+1})x^{(k)} + \theta_{k+1}v^{(k)}$$

- $v^{(k)}$ satisfies

$$\begin{aligned} v^{(k)} &= x^{(k-1)} + \frac{1}{\theta_k}\left(y^{(k-1)} - t_k G_t(y^{(k-1)}) - x^{(k-1)}\right) \\ &= v^{(k-1)} - \frac{t_k}{\theta_k} G_{t_k}(y^{(k-1)}) \end{aligned}$$

- θ_k satisfies $(1 - \theta_k)/\theta_k^2 \leq 1/\theta_{k-1}^2$

Progress in one iteration

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

use inequality (3.5) with $z = x$ and $z = x^\star$, and make convex combination:

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

$$
\frac{1}{\theta_i^2}(f(x^{(i)}) - f^\star) + \frac{1}{2t_i}\|v^{(i)} - x^\star\|_2^2 \leq \frac{1-\theta_i}{\theta_i^2}(f(x^{(i-1)}) - f^\star) + \frac{1}{2t_i}\|v^{(i-1)} - x^\star\|_2^2
$$

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,

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

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

Analysis for backtracking line search

recall that step sizes satisfy $t_{i-1} \geq t_i \geq t_{\min}$

apply inequality on page 3-33 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: #iterations to reach $f(x^{(k)}) - f^\star \leq \epsilon$ is $O(1/\sqrt{\epsilon})$

Descent version of accelerated proximal gradient method

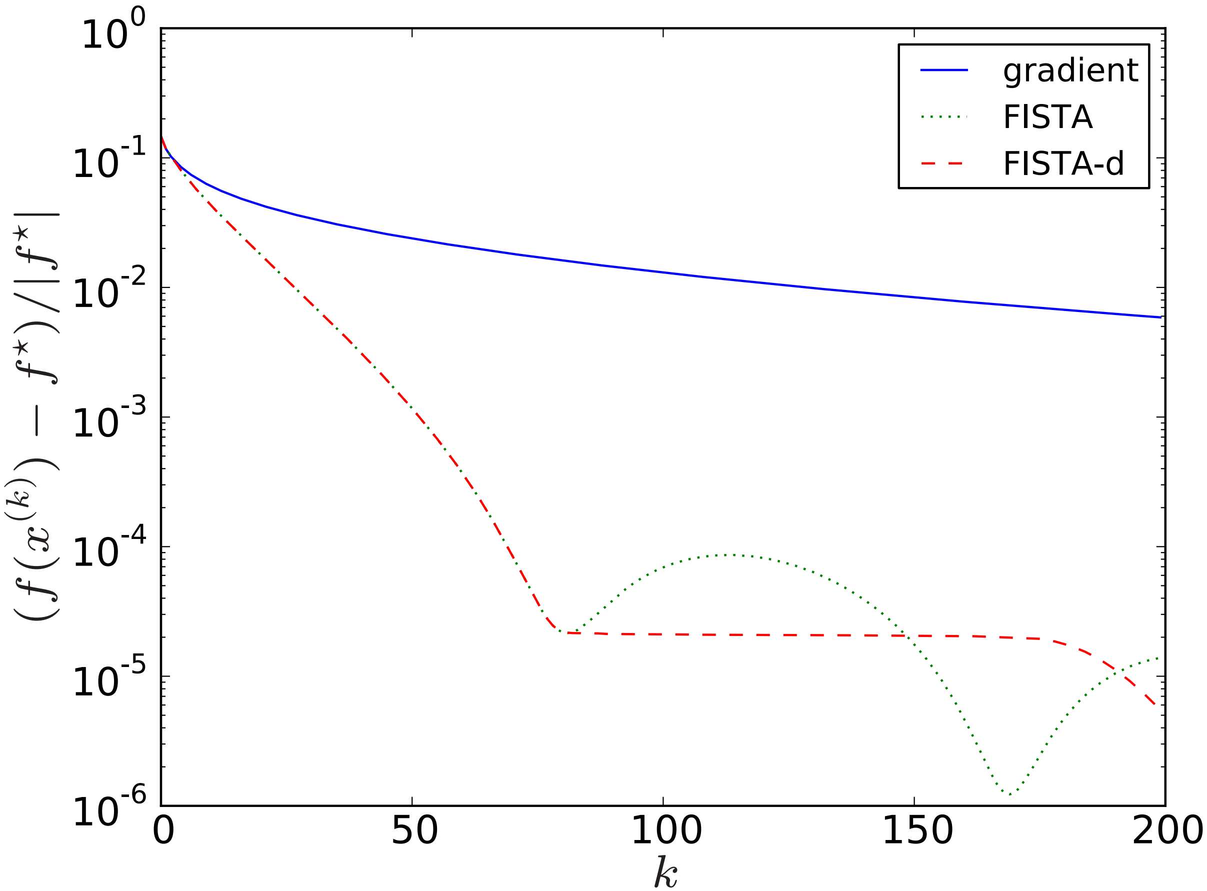
a modification that guarantees $f(x^{(k)}) \leq f(x^{(k-1)})$

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

same complexity; in the analysis of page 3-33, replace first line with

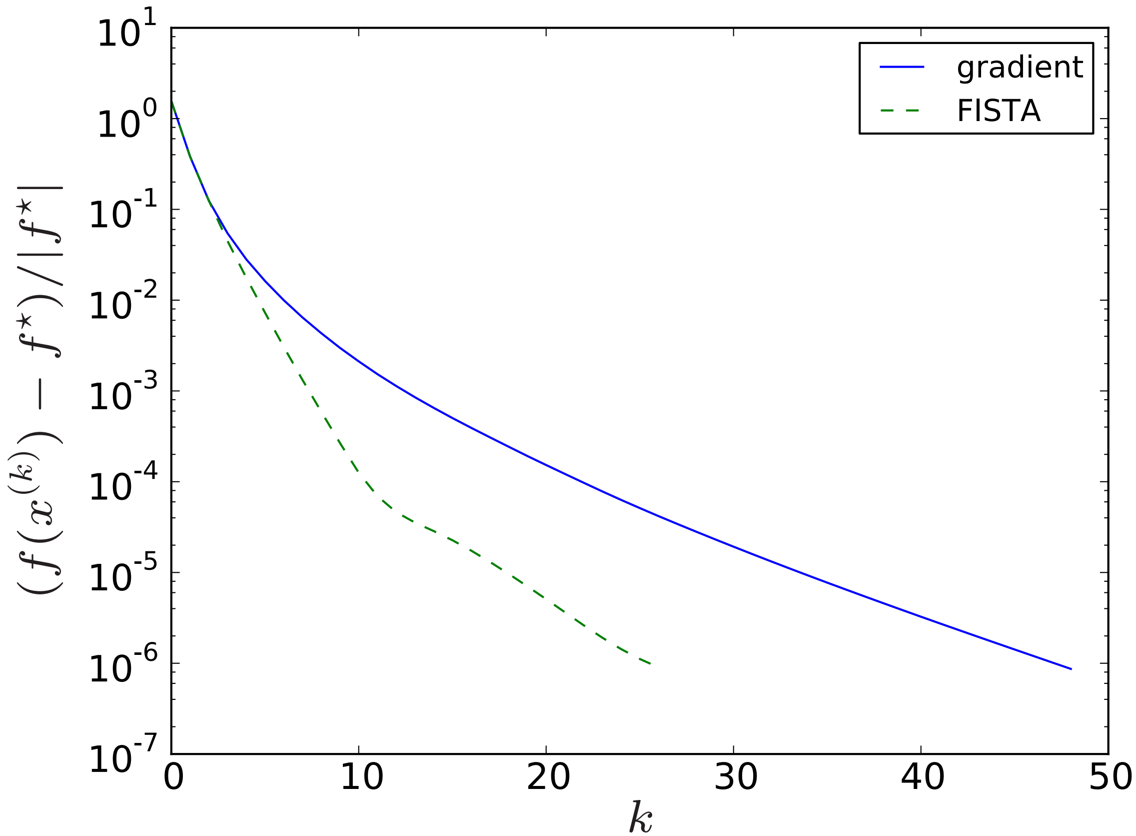
$$
\begin{aligned}
f(x^+) &\leq f(z^+) \\
&\leq (1-\theta)f(x) + \theta f^\star + G_t(y)^T(y - (1-\theta)x - \theta x^\star) - \frac{t}{2}\|G_t(y)\|_2^2
\end{aligned}
$$

example (from page 3-28)



Example: quadratic program with box constraints

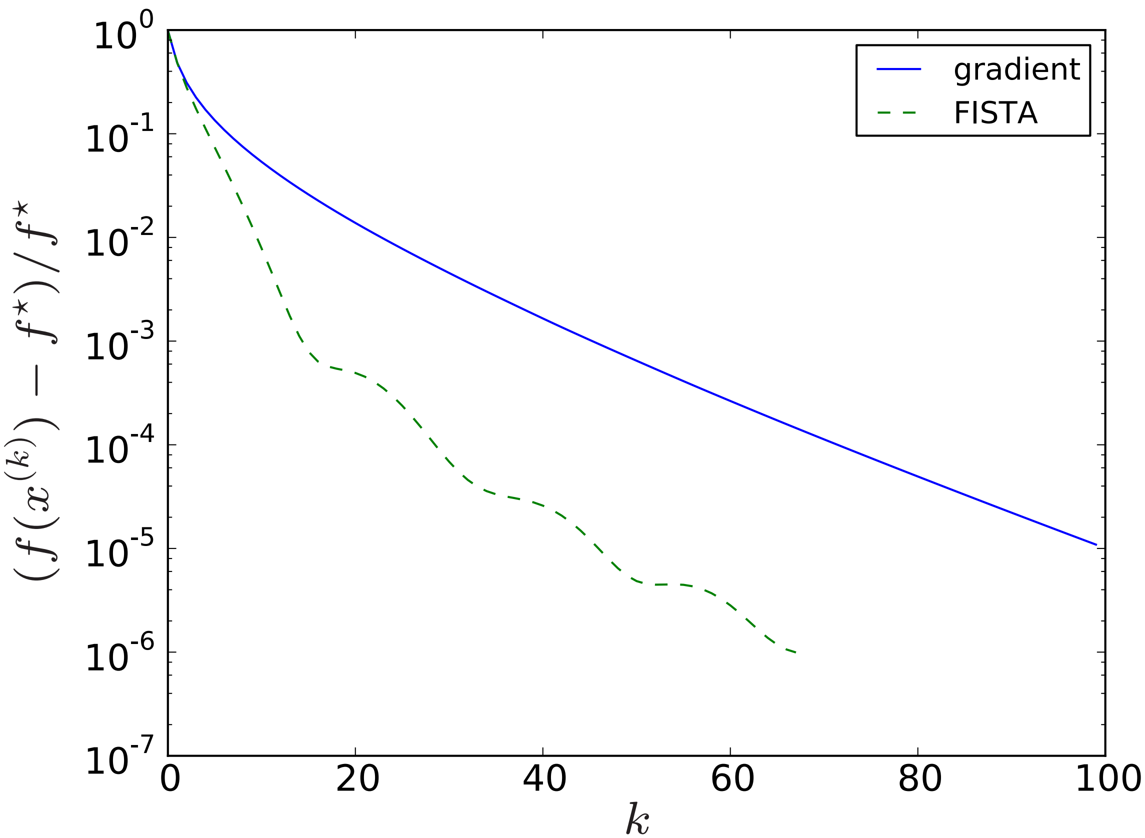
$$\begin{array}{ll} \text{minimize} & (1/2)x^T Ax + b^T x \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1} \end{array}$$



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

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 $h(X) = \|X\|_*$

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

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

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

where

$$\hat{\sigma}_k = \begin{cases} \sigma_k - t & \sigma_k \geq t \\ 0 & -t \leq \sigma_k \leq t \\ \sigma_k + t & \sigma_k \leq -t \end{cases}$$

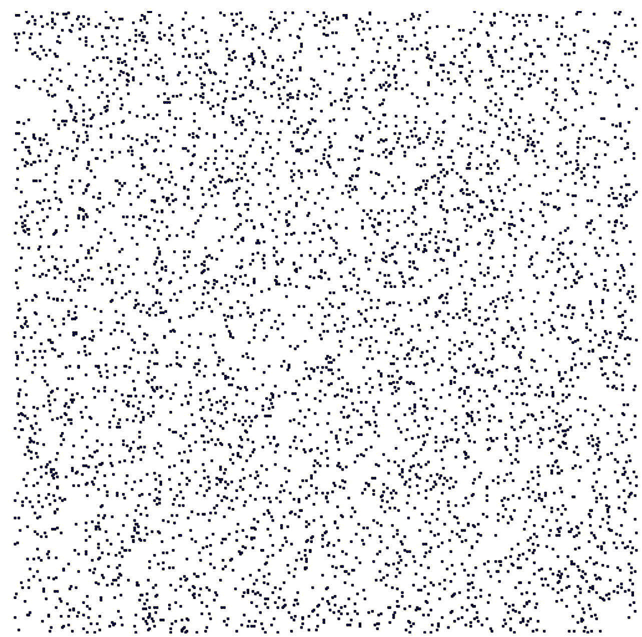
Approximate low-rank completion

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

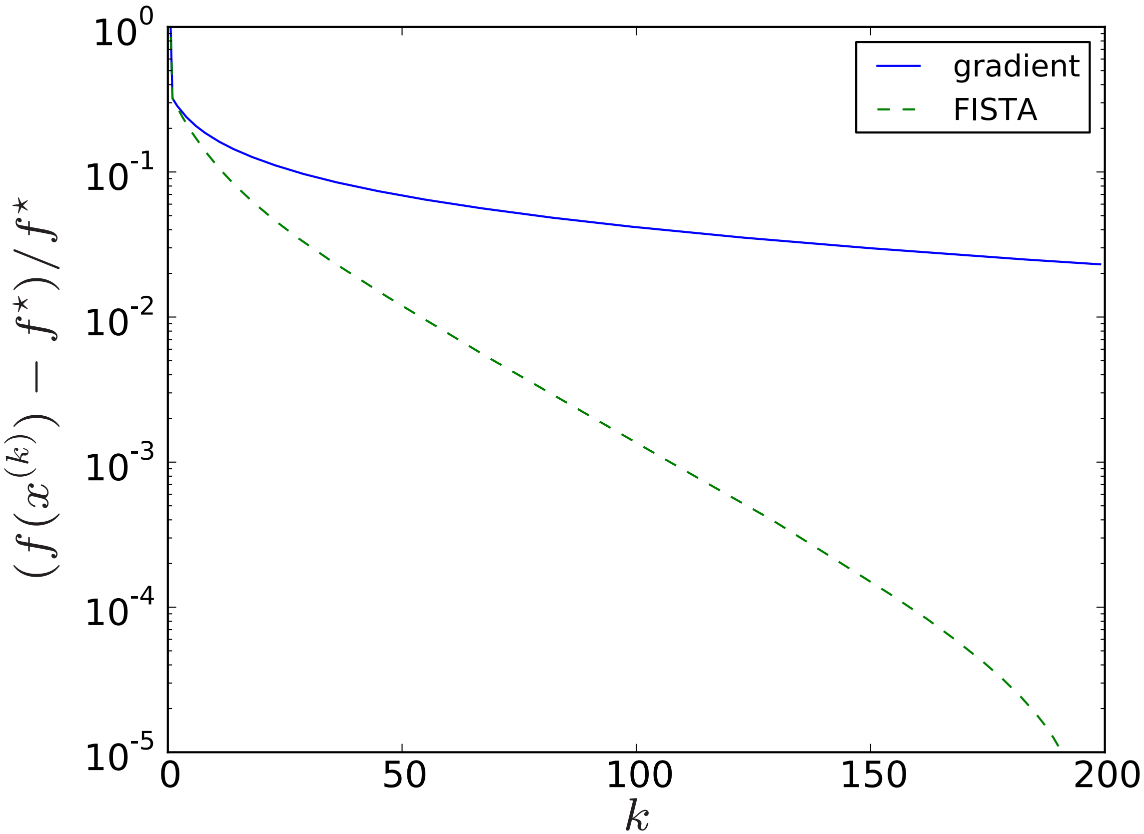
- entries $(i, j) \in N$ are approximately specified ($X_{ij} \approx A_{ij}$); rest is free
- nuclear norm regularization added to obtain low rank X

example

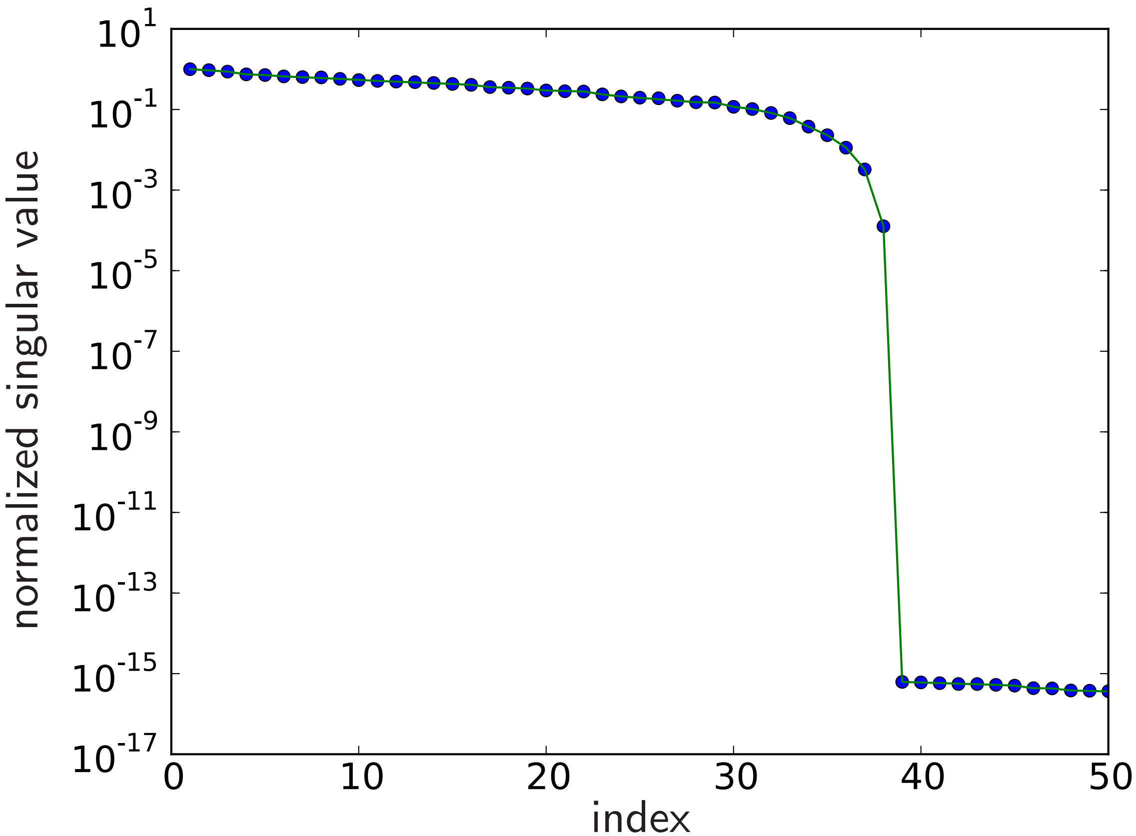
$m = n = 500$
5000 specified entries



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



result



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

Outline

- introduction
- proximal mapping
- proximal gradient method
- convergence analysis
- accelerated proximal gradient method
- **forward-backward method**

Monotone inclusion problems

a multivalued mapping F ($i.e.$, mapping x to a set $F(x)$) is **monotone** if

$$(u-v)^T(x-y) \geq 0 \qquad \forall x, y, u \in F(x), v \in F(y)$$

monotone inclusion problem: find x with

$$0 \in F(x)$$

examples

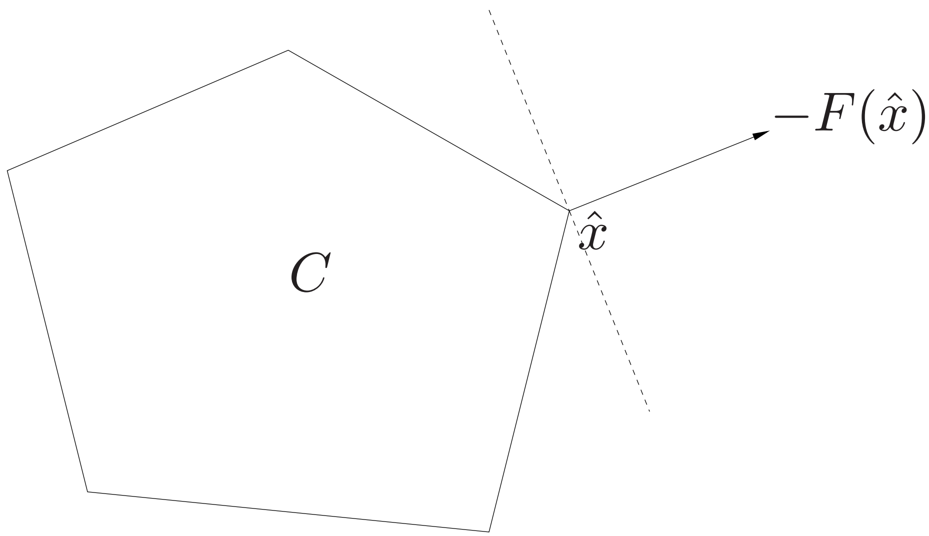
- unconstrained convex optimization: $0 \in \partial f(x)$
- saddle point of convex-concave function $f(x,y)$

$$0 \in \partial_x f(x,y) \times \partial_y(-f)(x,y)$$

Monotone variational inequality

given continuous monotone F, closed convex set C, find $\hat{x} \in C$ such that

$$F(\hat{x})^T(x - \hat{x}) \geq 0 \quad \forall x \in C$$

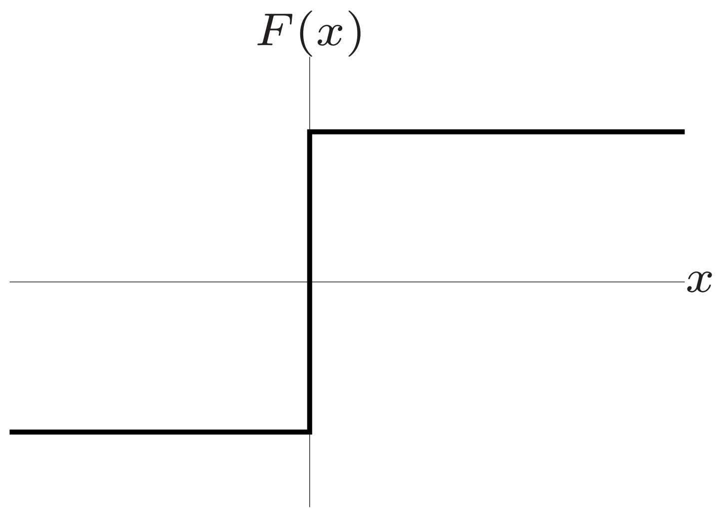


- with $F(x) = \nabla f(x)$, gives optimality condition for convex optimization
- includes as special cases various types of equilibrium problems
- a monotone inclusion: $0 \in N_C(x) + F(x)$ ($N_C(x)$ is normal cone at x)

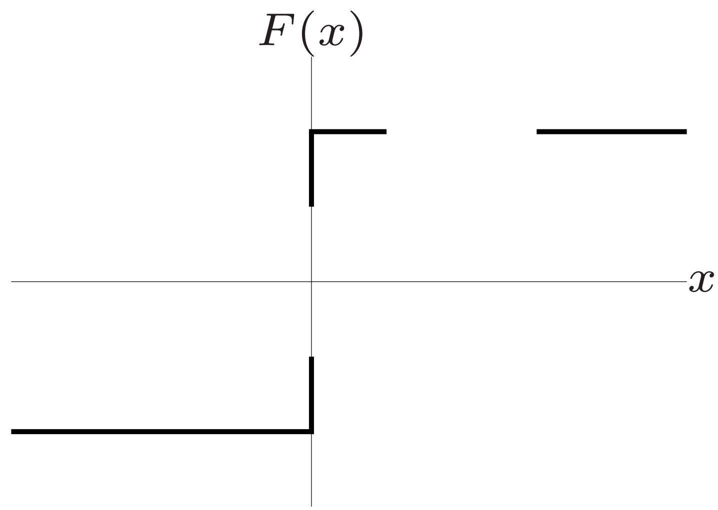
Maximal monotone operator

the **graph** of F is the set $\mathbf{gr}(F) = \{(x, y) \mid y \in F(x)\}$

monotone F is **maximal monotone** if $\mathbf{gr}(F)$ is not contained in the graph of another monotone mapping



maximal monotone



not maximal monotone

example: the subdifferential ∂f of a closed convex function f

Resolvent

the **resolvent** of a multivalued mapping A is the mapping

$$R_t = (I + tA)^{-1}$$

(with $t > 0$), *i.e.*, $\mathbf{gr}(R_t) = \{(y + tz, y) \mid z \in A(y)\}$

- if A is monotone than R_t is firmly nonexpansive:

$$y \in R_t(x),\ \hat{y} \in R_t(\hat{x}) \quad \Longrightarrow \quad (y - \hat{y})^T (x - \hat{x}) \geq \|y - \hat{y}\|_2^2$$

  hence $R_t(x)$ is single valued and Lipschitz continuous on $\mathbf{dom}\, R_t$:

$$\|R_t(x) - R_t(\hat{x})\|_2 \leq \|x - \hat{x}\|_2$$

- if A is maximal monotone, then $\mathbf{dom}\, R_t = \mathbf{R}^n$

Resolvent of subdifferential

the resolvent of ∂h is the proximal mapping:

$$\begin{aligned}(I + t\partial h)^{-1}(x) &= \mathbf{prox}_{th}(x) \\ &= \underset{y}{\operatorname{argmin}} \left(h(y) + \frac{1}{2t}\|y - x\|_2^2 \right)\end{aligned}$$

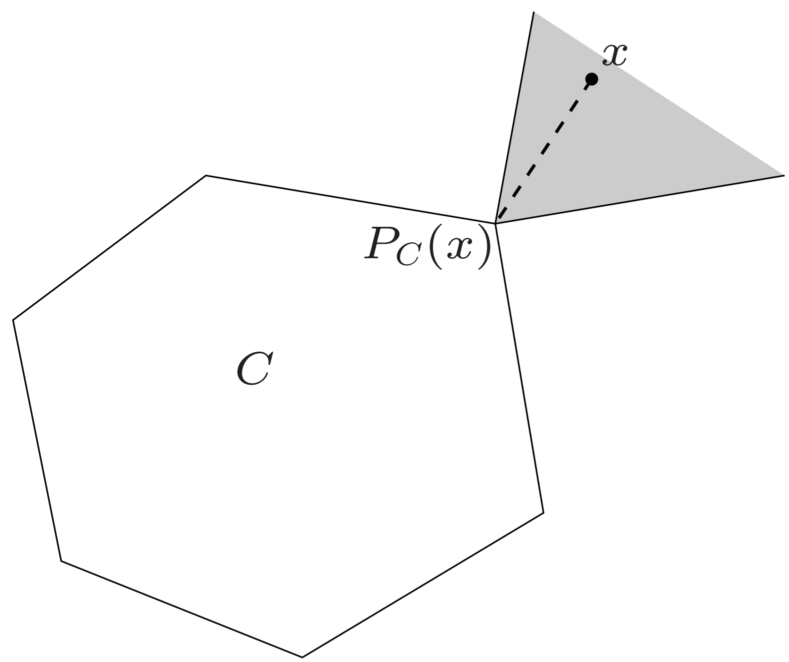
from optimality conditions in the definition of $\mathbf{prox}_{th}$:

$$\begin{aligned}y = \mathbf{prox}_{th}(x) \quad &\Longleftrightarrow \quad 0 \in \partial h(y) + \frac{1}{t}(y - x) \\ &\Longleftrightarrow \quad x \in (I + t\partial h)(y)\end{aligned}$$

Resolvent of normal cone

the resolvent of the normal cone operator N_C is the projection on C:

$$(I + tN_C)^{-1}(x) = P_C(x)$$



$$\begin{aligned} y = (I + tN_C)^{-1}(x) \quad &\Longleftrightarrow \quad x \in y + tN_C(y) \\ &\Longleftrightarrow \quad y = P_C(x) \end{aligned}$$

Forward-backward method

monotone inclusion $0 \in F(x)$

operator splitting: write F as $F(x) = A(x) + B(x)$

- A, B monotone
- $A(x)$ single valued
- B has easily computed resolvent

forward backward algorithm

$$x^{(k)} = (I + t_k B)^{-1}(I - t_k A)(x^{(k-1)})$$

- 'forward operator' $I - t_k A$ followed by 'backward operator' $(I + t_k B)^{-1}$
- step size rules depend on monotonicity properties of A or A^{-1}

Applications

proximal gradient method for minimizing $g(x) + h(x)$

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

this is the forward-backward method with $A(x) = \nabla g(x)$, $B(x) = \partial h(x)$

projection method for variational inequality defined by F, C

$$x^{(k)} = P_C\left(x^{(k-1)} - t_k F(x^{(k-1)})\right)$$

this is the forward-backward method with $A(x) = F(x)$, $B(x) = N_C(x)$

References

Proximal mappings

- P. L. Combettes and V.-R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Modeling and Simulation (2005)
- P. L. Combettes and J.-Ch. Pesquet, *Proximal splitting methods in signal processsing* `arxiv.org/abs/0912.3522v4`

Accelerated proximal gradient method

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008)
- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal 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)
- A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Transactions on Image Processing (2009)