6. Proximal gradient method

- motivation
- proximal mapping
- proximal gradient method with fixed step size
- proximal gradient method with line search

Proximal mapping

the proximal mapping (prox-operator) of a convex function h is defined as

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

examples

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

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

- $h(x) = \|x\|_1$: prox_h is the 'soft-threshold' (shrinkage) operation

$$\mathrm{prox}_h(x)_i = \begin{cases} x_i - 1 & x_i \geq 1 \\ 0 & |x_i| \leq 1 \\ x_i + 1 & x_i \leq -1 \end{cases}$$

Proximal gradient method

unconstrained optimization with objective split in two components

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

- g convex, differentiable, $\mathbf{dom}\, g = \mathbf{R}^n$
- h convex with inexpensive prox-operator (many examples in lecture 9)

proximal gradient algorithm

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

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

Interpretation

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

from definition of proximal mapping:

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

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

Examples

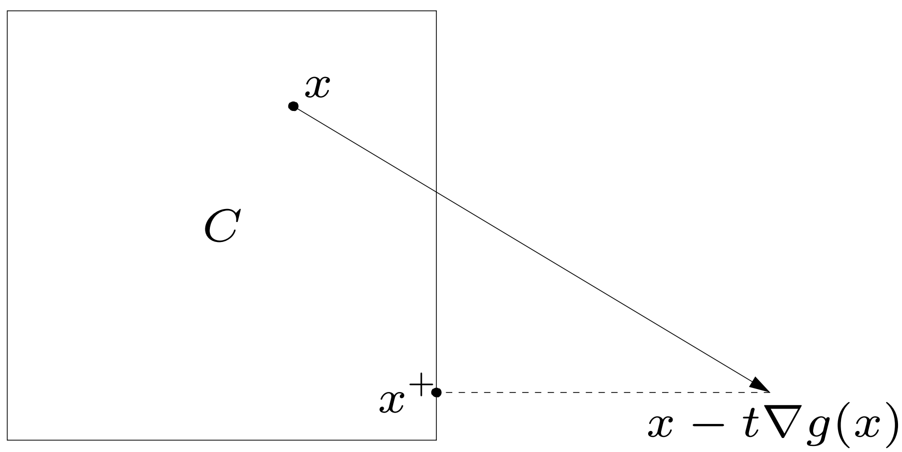
$$\text{minimize} \quad g(x) + h(x)$$

gradient method: special case with $h(x) = 0$

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

gradient projection method: special case with $h(x) = I_C(x)$

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

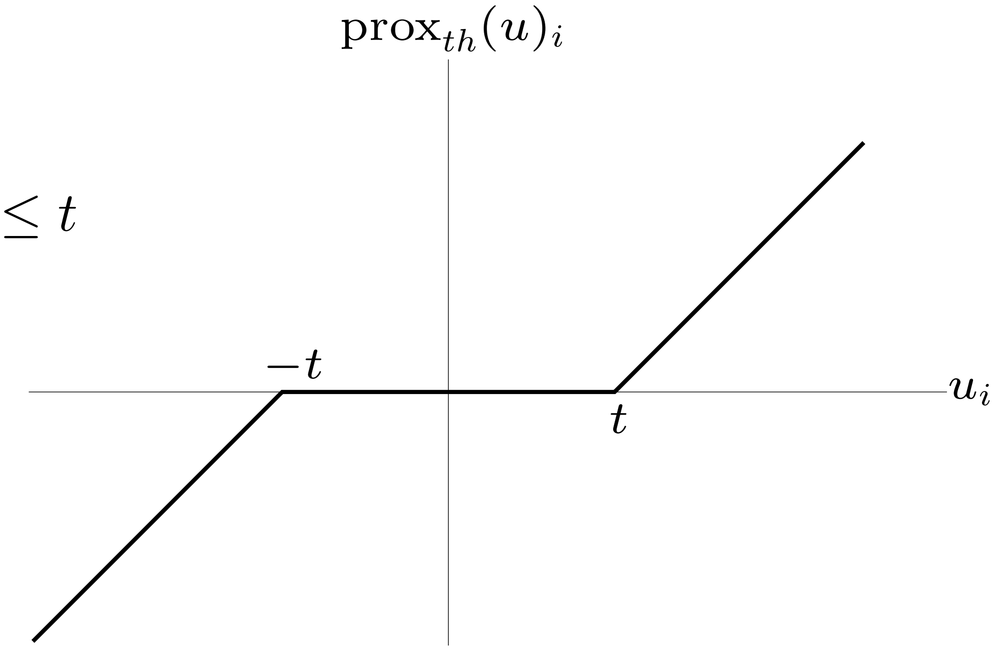


soft-thresholding: special case with $h(x) = \|x\|_1$

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

where

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



Outline

- introduction
- **proximal mapping**
- proximal gradient method with fixed step size
- proximal gradient method with line search

Proximal mapping

if h is convex and closed (has a closed epigraph), then

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

exists and is unique for all x

- will be studied in more detail in lecture 9
- from optimality conditions of minimization in the definition:

$$\begin{aligned} u = \mathrm{prox}_h(x) \qquad &\Longleftrightarrow \qquad x - u \in \partial h(u) \\ &\Longleftrightarrow \qquad h(z) \geq h(u) + (x - u)^T(z - u) \quad \forall z \end{aligned}$$

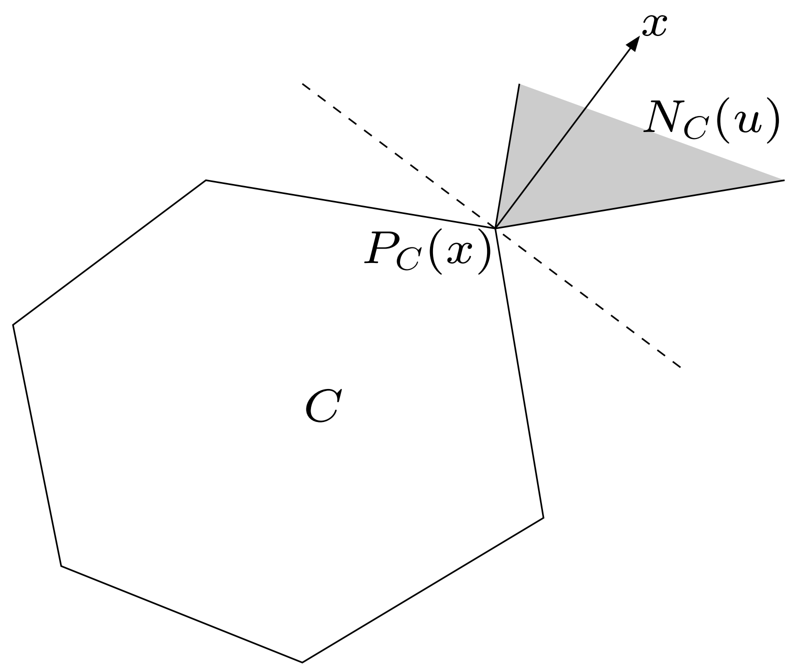
Projection on closed convex set

proximal mapping of indicator function I_C is Euclidean projection on C

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

subgradient characterization

$$u = P_C(x)$$
$$\Updownarrow$$
$$(x-u)^T(z-u) \le 0 \quad \forall z \in C$$



we will see that proximal mappings have many properties of projections

Nonexpansiveness

if $u = \mathrm{prox}_h(x)$, $v = \mathrm{prox}_h(y)$, then

$$(u - v)^T (x - y) \geq \|u - v\|_2^2$$

prox_h is *firmly nonexpansive*, or *co-coercive* with constant 1

- follows from characterization of page 6-7 and monotonicity (page 4-10)

$$x - u \in \partial h(u), \quad y - v \in \partial h(v) \quad \Longrightarrow \quad (x - u - y + v)^T (u - v) \geq 0$$

- implies (from Cauchy-Schwarz inequality)

$$\|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\|_2 \leq \|x - y\|_2$$

  prox_h is *nonexpansive*, or *Lipschitz continuous* with constant 1

Outline

- introduction
- proximal mapping
- **proximal gradient method with fixed step size**
- proximal gradient method with line search

Convergence of proximal gradient method

to minimize $g + h$, choose $x^{(0)}$ and repeat

$$x^{(k)} = \mathrm{prox}_{t_k h}\left(x^{(k-1)} - t\nabla g(x^{(k-1)})\right), \qquad k \geq 1$$

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 prox_{th} is well defined)
- optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)

convergence result: $1/k$ rate convergence with fixed step size $t_k = 1/L$

Gradient map

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

$G_t(x)$ is the negative 'step' in the proximal gradient update

$$\begin{aligned} x^+ &= \mathrm{prox}_{th}\left(x - t\nabla g(x)\right) \\ &= x - tG_t(x) \end{aligned}$$

- $G_t(x)$ is not a gradient or subgradient of $f = g + h$
- from subgradient definition of prox-operator (page 6-7),

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

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

Consequences of Lipschitz assumption

recall upper bound (p.1-12) for convex g with Lipschitz continuous gradient

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

- substitute $y = x - tG_t(x)$:

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

- if $0 < t \leq 1/L$, then

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

A global inequality

if the inequality (1) holds, then for all z,

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

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

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

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 (2) with $z = x$ shows the algorithm is a descent method:

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

- inequality (2) with $z = x^\star$:

$$\begin{aligned} f(x^+) - f^\star &\leq G_t(x)^T(x - x^\star) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star - tG_t(x)\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \left\|x^+ - x^\star\right\|_2^2\right) \qquad (3) \end{aligned}$$

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

Analysis for fixed step size

add inequalities (3) for $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t = t_i = 1/L$

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

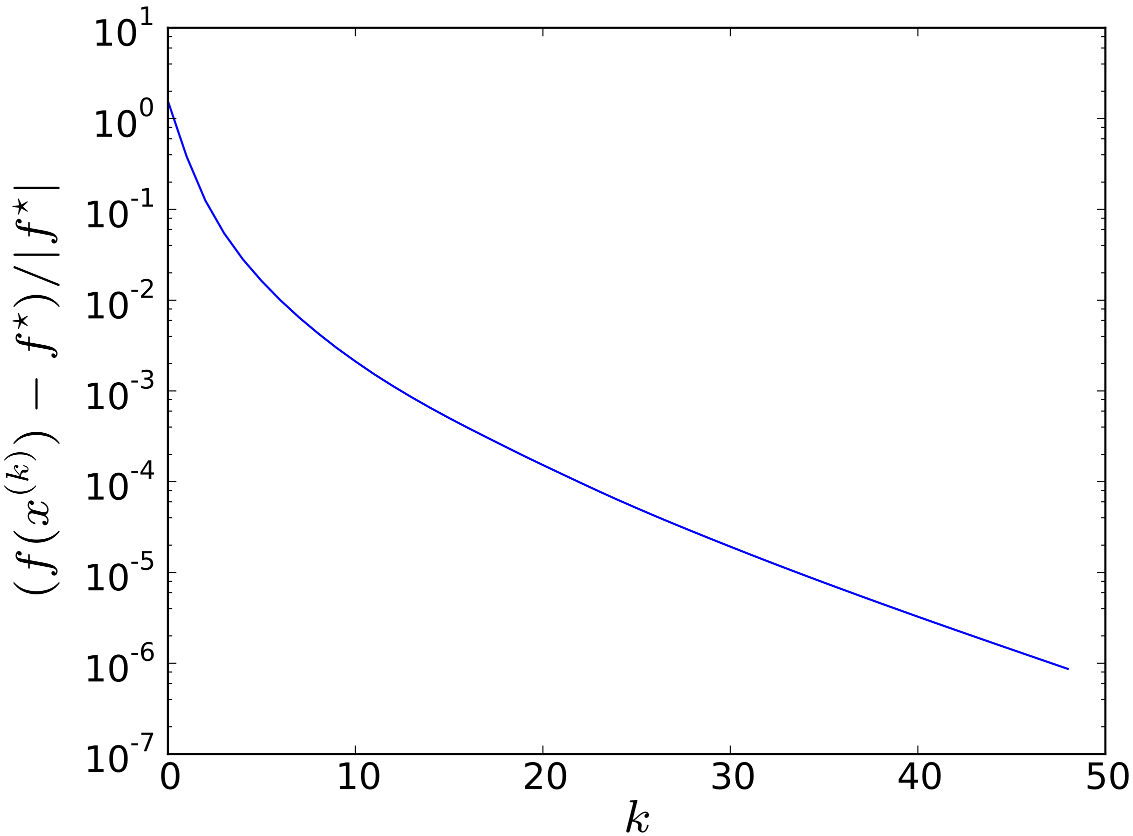
since $f(x^{(i)})$ is nonincreasing,

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

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

Quadratic program with box constraints

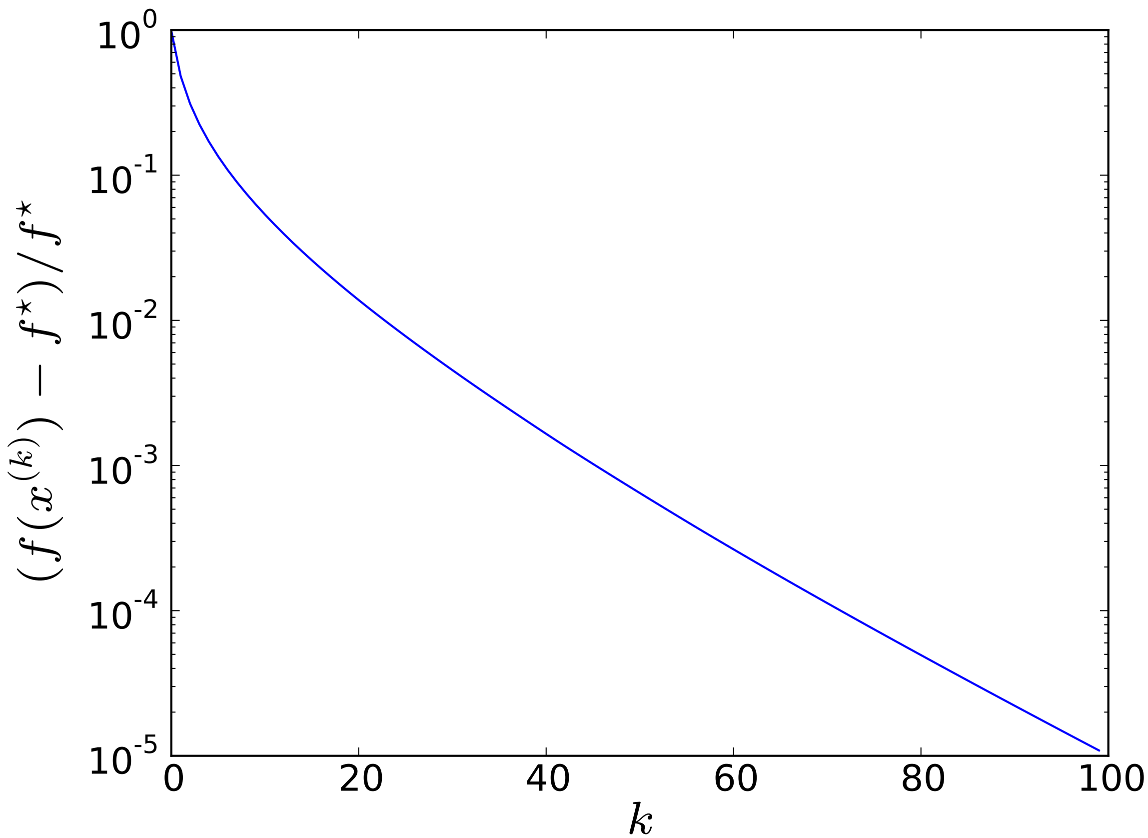
$$\begin{array}{ll} \text{minimize} & (1/2)x^TAx + b^Tx \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1} \end{array}$$



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

Outline

- introduction
- proximal mapping
- proximal gradient method with fixed step size
- **proximal gradient method with line search**

Line search

- the analysis for fixed step size (page 6-12) starts with the inequality

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

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

- if L is not known, we can satisfy (1) by a backtracking line search:

  start at some $t := \hat{t} > 0$ and backtrack ($t := \beta t$) until (1) holds

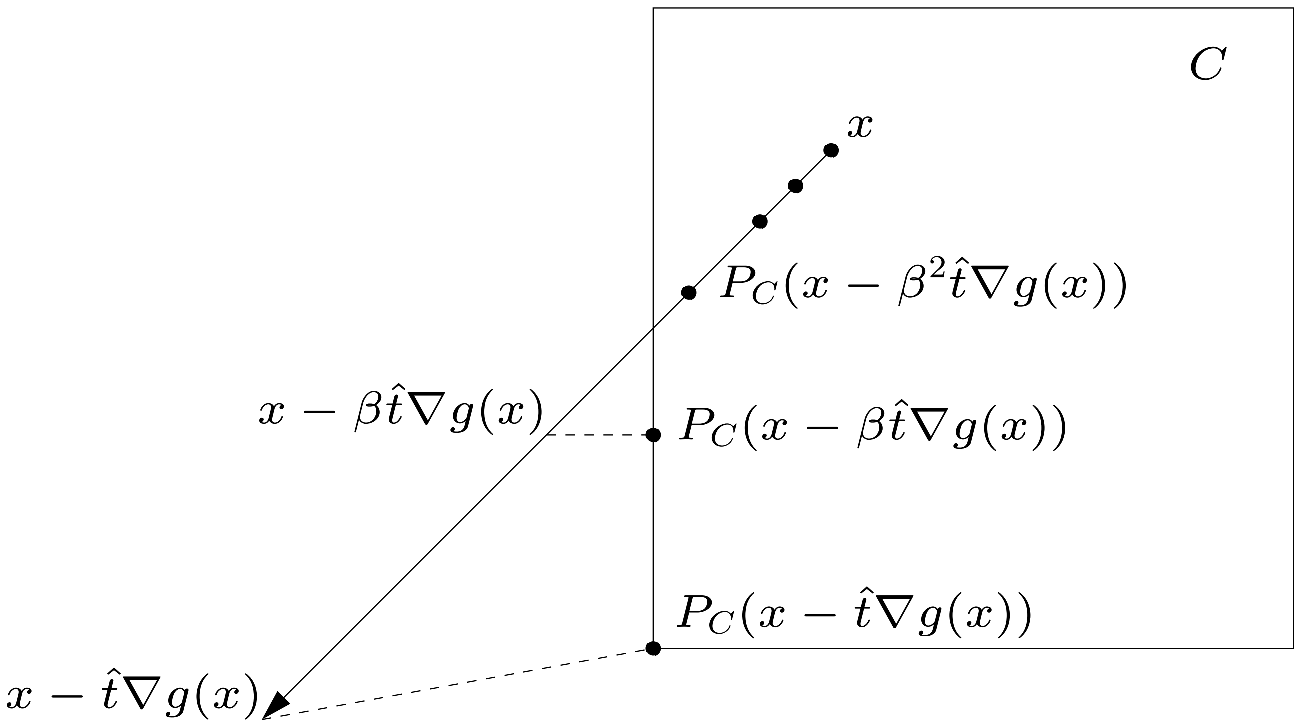
- step size t selected by the line search satisfies $t \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

- requires one evaluation of g and prox_{th} per line search iteration

several other types of line search work

example: line search for projected gradient method

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



backtrack until $x - tG_t(x)$ satisfies 'sufficient decrease' inequality (1)

Analysis with line search

from p. 6-14, if (1) holds in iteration i, then $f(x^{(i)}) < f(x^{(i-1)})$ and

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

- adding inequalities for $i = 1$ to $i = k$ gives

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

- since $f(x^{(i)})$ is nonincreasing, obtain similar $1/k$ bound as for fixed t_i:

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

References

convergence analysis of proximal gradient method

- 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)