Proximal gradient methods

Acknowledgement: slides are based on Prof. Lieven Vandenberghes.

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

Proximal mapping

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

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

Examples

- $h(x) = 0$: $\operatorname{prox}_h(x) = x$
- $h(x)$ is indicator function of closed convex set C: prox_h is projection on C

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

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

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

Proximal gradient method

unconstrained optimization with objective split in two components

$$\text{minimize} \quad f(x) = g(x) + h(x)$$: composite minimization.

- g convex, differentiable, $\operatorname{dom} g = \mathbf{R}^n$
- h convex with inexpensive prox-operator

① GD (×)

② SD (✓) $O\left(\frac{1}{\sqrt{k}}\right)$

Proximal gradient algorithm $\left(O\left(\frac{1}{k}\right)\right)$

$$x_{k+1} = \operatorname{prox}_{t_k h}\left(x_k - t_k \nabla g(x_k)\right)$$

$\nearrow$ GD

$\Rightarrow \underbrace{x_{k+1} = x_k + t_k d_k}$

- $t_k > 0$ is step size, constant or determined by line search
- can start at infeasible x_0 (however $x_k \in \operatorname{dom} f = \operatorname{dom} h$ for $k \geq 1$)

Interpretation

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

from definition of proximal mapping:

$$\begin{aligned} x^+ &= \underset{u}{\operatorname{argmin}} \left(h(u) + \boxed{\frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2} \right) \\ &= \underset{u}{\operatorname{argmin}} \left(h(u) + \underbrace{g(x) + \nabla g(x)^T(u - x) + \boxed{\frac{1}{2t}\|u - x\|_2^2}}_{\tilde g(x;u)} \right) \end{aligned}$$

Hessian.

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

Examples

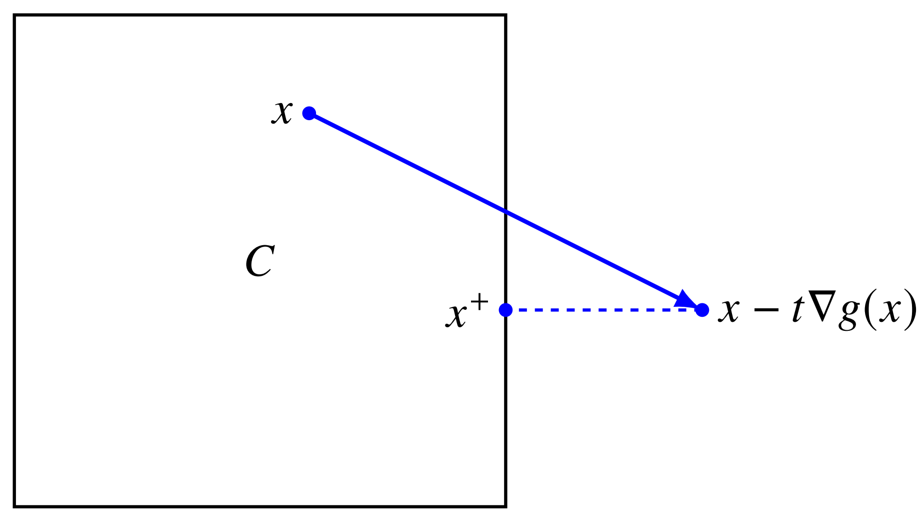
$$\text{minimize} \quad g(x) + h(x)$$

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

$$x^+ = x - t\nabla g(x) \sim O\left(\frac{1}{k}\right)$$

Gradient projection method: special case with $h(x) = \delta_C(x)$ (indicator of C)

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



Examples

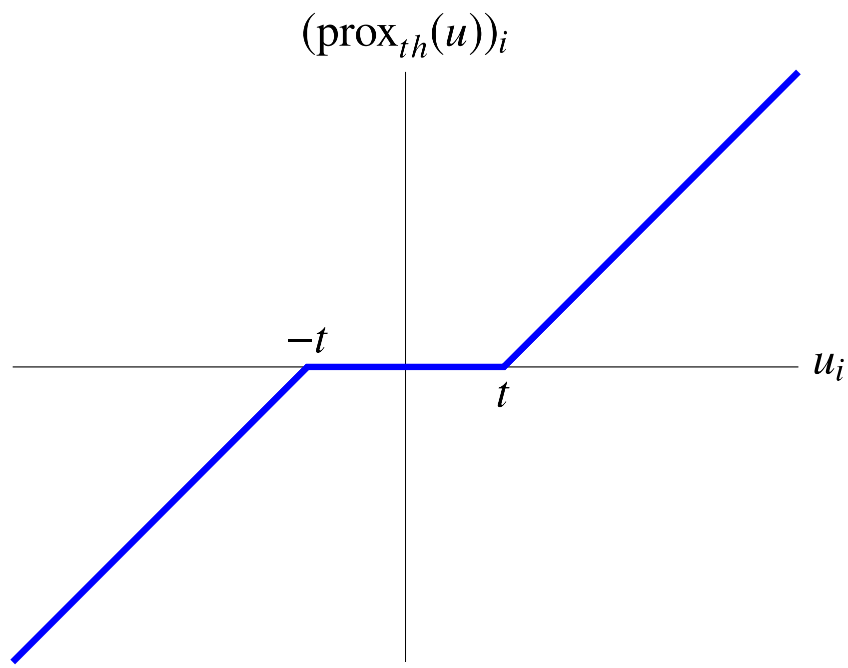
Soft-thresholding: special case with $h(x) = \|x\|_1$

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

$= \operatorname{argmin}_u \{ t\|u\|_1 + \frac{1}{2}\|u - x + t\nabla g(x)\|^2$

where

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



Outline

- motivation
- **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

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

exists and is unique for all x

- will be studied in more detail in one of the next lectures
- from optimality conditions of minimization in the definition:

$$\begin{aligned} u = \operatorname{prox}_h(x) \quad &\Longleftrightarrow \quad x - u \in \partial h(u) \\ &\Longleftrightarrow \quad h(z) \geq h(u) + (x - u)^T(z - u) \quad \text{for all } z \end{aligned}$$

Projection on closed convex set

proximal mapping of indicator function δ_C is Euclidean projection on C

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

$$u = P_C(x)$$

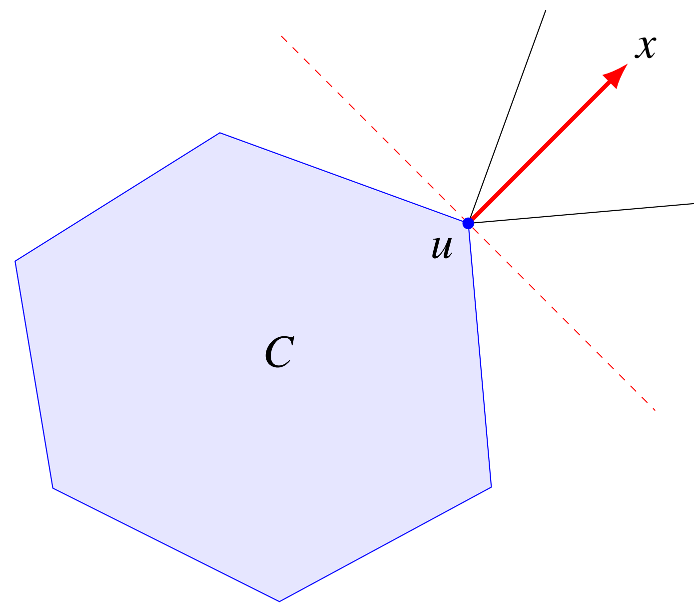
$$\Updownarrow$$

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

$$\Updownarrow$$

$$x - u \in N_C(u).$$

$$\Updownarrow \quad x - u \in \partial \delta_C(u)$$



we will see that proximal mappings have many properties of projections

Firm nonexpansiveness

proximal mappings are **firmly nonexpansive** (co-coercive with constant 1):

$$(\mathrm{prox}_h(x) - \mathrm{prox}_h(y))^T(x - y) \geq \|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\|_2^2$$

- follows from page 4.7: if $u = \mathrm{prox}_h(x)$, $v = \mathrm{prox}_h(y)$, then

$$x - u \in \partial h(u), \qquad y - v \in \partial h(v)$$

combining this with monotonicity of subdifferential (page 2.9) gives

$$(x - u - y + v)^T(u - v) \geq 0 \quad (\checkmark)$$

- a weaker property is **nonexpansiveness** (Lipschitz continuity with constant 1):

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

follows from firm nonexpansiveness and Cauchy–Schwarz inequality

Outline

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

Assumptions

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

- h is closed and convex (so that prox_{th} is well defined) $\triangle$
- g is differentiable with $\mathrm{dom}\, g = \mathbf{R}^n$, and L-smooth for the Euclidean norm, *i.e.*,

$$\star \quad \frac{L}{2}x^T x - g(x) \quad \text{is convex} \iff \nabla g \text{ is } L\text{-Lip.}$$

- there exists a constant $m \geq 0$ such that

$$g(x) - \frac{m}{2}x^T x \quad \text{is convex} \quad \Rightarrow g \text{ is strongly convex.}$$

  when $m > 0$ this is m-strong convexity for the Euclidean norm

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

Implications of assumptions on g

Lower bound

- convexity of the the function $g(x) - (m/2)x^T x$ implies (page 1.19):

Strong convex

$$\Rightarrow g(y) \geq g(x) + \nabla g(x)^T (y - x) + \frac{m}{2}\|y - x\|_2^2 \qquad \text{for all } x, y \tag{1}$$

- if $m = 0$, this means g is convex; if $m > 0$, strongly convex (lecture 1)

Upper bound

- convexity of the function $(L/2)x^T x - g(x)$ implies (page 1.12):

L- smooth

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

- this is equivalent to Lipschitz continuity and co-coercivity of gradient (lecture 1)

$x^+ = \underset{u}{\operatorname{argmin}} \{ th(u) + \frac{1}{2}\|u - x + t\nabla g(x)\|^2 \}$

Gradient map

$0 \in t\,\partial h(x^+) + (x^+ - x + t\nabla g(x)) \Rightarrow \boxed{\frac{x - x^+}{t}} \in \partial h(x^+) + \nabla g(x)$

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

$G_t(x) = 0$

$\Leftrightarrow x^+ = x$

$\Leftrightarrow 0 \in \partial h(x) + \nabla g(x)$

$\Leftrightarrow x \in \underset{x}{\operatorname{argmin}}\, f(x)$

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

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

$\Rightarrow$ approximate gradient.

step size

- $G_t(x)$ is not a gradient or subgradient of $f = g + h$
- from subgradient definition of prox-operator (page 4.7),

$$G_t(x) \in \nabla g(x) + \partial h(x - tG_t(x)) = \partial h(x^+)$$

$= \nabla g(x) + \partial h(x)$

$+ \underbrace{\partial h(x^+) - \partial h(x)}$.

Continuous.

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

$\Rightarrow$ Stopping Creterion: $\|G_t(x)\| \le \varepsilon$

$\Leftrightarrow \left\|\frac{x^+ - x}{t}\right\| \le \varepsilon$

Consequences of quadratic bounds on g

substitute $y = x - tG_t(x)$ in the bounds (1) and (2): for all t,

$L \geq m.$

$t < \frac{1}{L} \leq m.$

$$\underbrace{\frac{mt^2}{2}\|G_t(x)\|_2^2}_{\text{lower bound.}} \leq g\,(\underbrace{x - tG_t(x)}_{x^+}) - g(x) + t\nabla g(x)^T G_t(x) \leq \underbrace{\frac{Lt^2}{2}\|G_t(x)\|_2^2}_{\text{upper bound.}}$$

- if $\underbrace{\boxed{0 < t \leq 1/L}}_{\text{step size.}}$, then the upper bound implies

$$(\star)\quad g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \tag{3}$$

- if the inequality (3) is satisfied and $tG_t(x) \neq 0$, then $mt \leq 1$
- if the inequality (3) is satisfied, then for all z,

$$(\star)\quad f(x - tG_t(x)) \leq f(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - z\|_2^2 \tag{4}$$

(proof on next page)

① $z = x \Rightarrow f(x^+) \leq f(x) - \frac{t}{2}\|G_t(x)\|^2$

$\Rightarrow f(x^+) \leq f(x)$

② $z = x^*$

Proof of (4):

$$
\begin{aligned}
f(x - tG_t(x)) &= g(x^+) + h(x^+) \\
&\le g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\
&\le \boxed{g(z) - \nabla g(x)^T(z - x) - \frac{m}{2}\|z - x\|_2^2} - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + h(x - tG_t(x)) \qquad G_t(x) \in \nabla g(x) + \partial h(x^+) \\
&\le g(z) - \nabla g(x)^T(z - x) - \frac{m}{2}\|z - x\|_2^2 - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + h(z) - (G_t(x) - \nabla g(x))^T(z - x + tG_t(x)) \\
&= g(z) + h(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - z\|_2^2 \qquad \star
\end{aligned}
$$

Lower bound.

- in the first step we add $h(x - tG_t(x))$ to both sides of the inequality (3)
- in the next step we use the lower bound on $g(z)$ from (1)
- in step 3, we use $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$ (see page 4.12)

Progress in one iteration

for a step size t that satisfies the inequality (3), define

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

- inequality (4) with $z = x$ shows that the algorithm is a descent method:

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

$\{f(x^k)\}\downarrow$

Lower bonded

- inequality (4) with $z = x^\star$ shows that

$\Rightarrow \exists\, \bar{f}$ s.t. $f(x^k) \to \bar{f}$.

$$\begin{aligned}
f(x^+) - f^\star &\le G_t(x)^T(x - x^\star) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - x^\star\|_2^2 \\
&= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star - tG_t(x)\|_2^2\right) - \frac{m}{2}\|x - x^\star\|_2^2 \\
&= \frac{1}{2t}\left((1 - mt)\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \qquad (5) \\
&\le \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \qquad (6)
\end{aligned}$$

Analysis for fixed step size

add inequalities (6) with $x = x_i$, $x^+ = x_{i+1}$, $t = t_i = 1/L$ from $i = 0$ to $i = k - 1$

$$
\begin{aligned}
\sum_{i=1}^{k}(f(x_i) - f^\star) &\leq \frac{1}{2t}\sum_{i=0}^{k-1}\left(\|x_i - x^\star\|_2^2 - \|x_{i+1} - x^\star\|_2^2\right) \\
&= \frac{1}{2t}\left(\|x_0 - x^\star\|_2^2 - \|x_k - x^\star\|_2^2\right) \\
&\leq \frac{1}{2t}\|x_0 - x^\star\|_2^2
\end{aligned}
$$

since $f(x_i)$ is nonincreasing,

$$
f(x_k) - f^\star \leq \frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f^\star) \leq \frac{1}{2kt}\|x_0 - x^\star\|_2^2 \qquad O\left(\frac{1}{k}\right)
$$

Distance to optimal set

- from (5) and $f(x^+) \geq f^\star$, the distance to the optimal set does not increase:

$$
\begin{aligned}
\|x^+ - x^\star\|_2^2 &\leq (1 - mt)\|x - x^\star\|_2^2 \\
&\leq \|x - x^\star\|_2^2
\end{aligned}
$$

$m > 0$

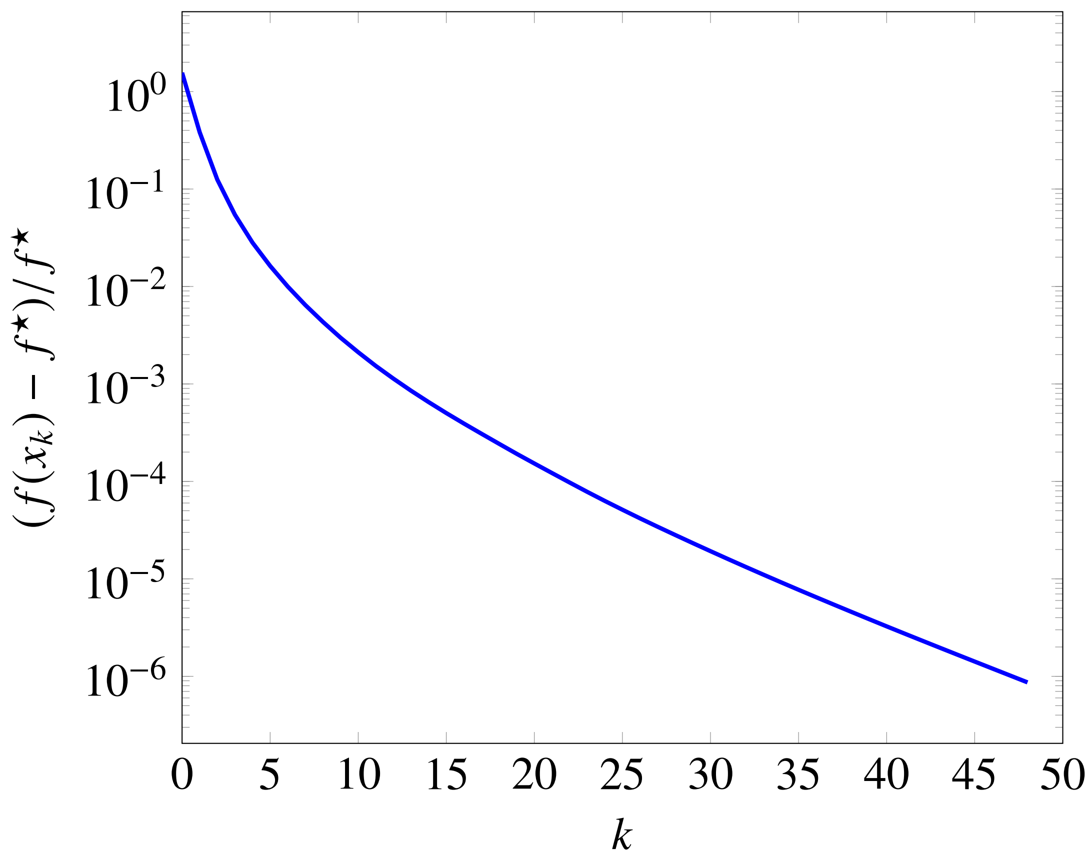
- for fixed step size $t_k = 1/L$

$$
\|x_k - x^\star\|_2^2 \leq c^k \|x_0 - x^\star\|_2^2, \qquad c = 1 - \frac{m}{L}
$$

i.e., linear convergence if g is strongly convex ($m > 0$)

Example: quadratic program with box constraints

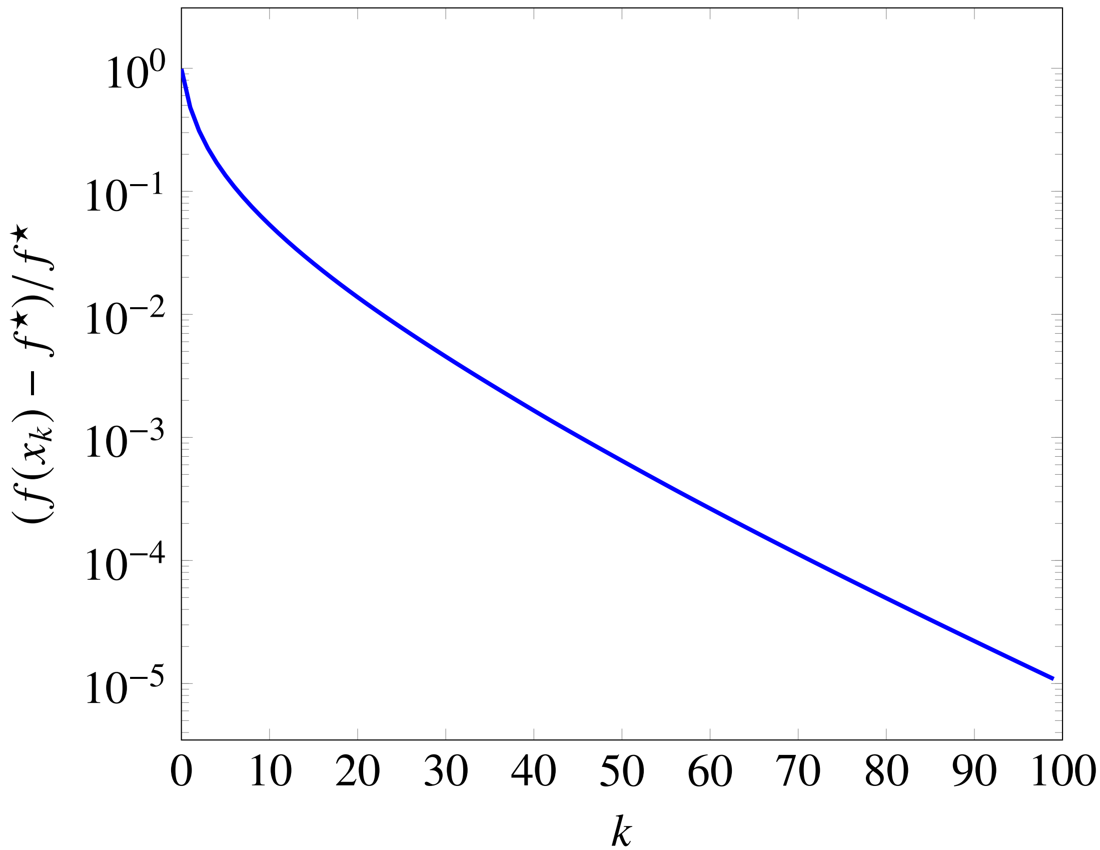
$$\begin{array}{ll} \text{minimize} & (1/2)x^T A x + b^T x \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1} \end{array}$$



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

Example: 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

$t \in (0, \frac{1}{L}]$ ✓

- the analysis for fixed step size (page 4.13) starts with the inequality

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (3)$$

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

  t: BB step w.r.t. g

- if L is not known, we can satisfy (3) by a backtracking line search:

  start at some $t := \hat{t} > 0$ and backtrack ($t := \beta t$) until (3) 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

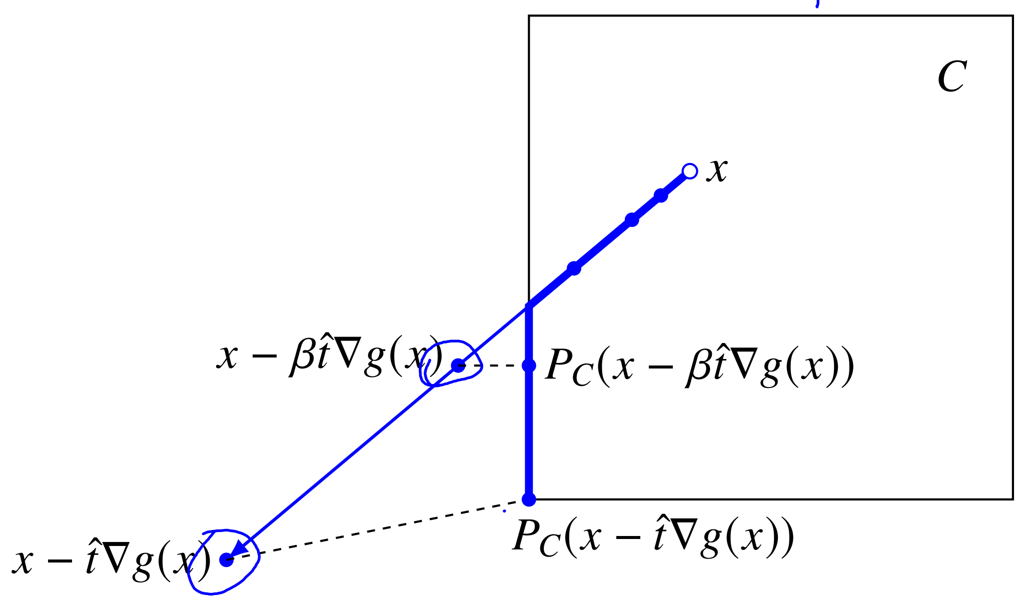
several other types of line search work

Example

line search for gradient projection method

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

$$f(x) - f(x^+) \geq \eta \,\|x - x^+\|^2 \quad (\checkmark)$$



backtrack until $P_C(x - t\nabla g(x))$ satisfies the "sufficient decrease" inequality (3)

Analysis with line search

from page 4.15, if (3) holds in iteration i, then $f(x_{i+1}) < f(x_i)$ and

$$t_i(f(x_{i+1}) - f^\star) \le \frac{1}{2}\left(\|x_i - x^\star\|_2^2 - \|x_{i+1} - x^\star\|_2^2\right)$$

- adding inequalities for $i = 0$ to $i = k - 1$ gives

$$kt_{\min} \le \left(\sum_{i=0}^{k-1} t_i\right)(f(x_k) - f^\star) \le \sum_{i=0}^{k-1} t_i(f(x_{i+1}) - f^\star) \le \frac{1}{2}\|x_0 - x^\star\|_2^2$$

  first inequality holds because $f(x_i)$ is nonincreasing

- since $t_i \ge t_{\min}$, we obtain a similar $1/k$ bound as for fixed step size

$$f(x_k) - f^\star \le \frac{1}{2\sum_{i=0}^{k-1} t_i}\|x_0 - x^\star\|_2^2 \le \frac{1}{2kt_{\min}}\|x_0 - x^\star\|_2^2 \qquad O\left(\frac{1}{k}\right)$$

Distance to optimal set

from page 4.15, if (3) holds in iteration i, then

$$\begin{aligned}
\|x_{i+1} - x^\star\|_2^2 &\leq (1 - mt_i)\|x_i - x^\star\|_2^2 \quad \checkmark \\
&\leq (1 - mt_{\min})\,\|x_i - x^\star\|_2^2 \\
&= c\,\|x_i - x^\star\|_2^2 \\
\|x_k - x^\star\|_2^2 &\leq c^k\,\|x_0 - x^\star\|_2^2
\end{aligned}$$

with

$$c = 1 - mt_{\min} = \max\{1 - \frac{\beta m}{L}, 1 - m\hat{t}\}$$

hence linear convergence if $m > 0$

Summary: proximal gradient method

- minimizes sums of differentiable and non-differentiable convex functions

$$f(x) = g(x) + h(x) \qquad O\left(\frac{1}{k}\right)$$

- useful when nondifferentiable term h is simple (has inexpensive prox-operator)
- convergence properties are similar to standard gradient method ($h(x) = 0$)
- less general but faster than subgradient method

· pratical stopping creterion.

Best complexity : $O\left(\frac{1}{k^2}\right)$. ?

References

- A. Beck, *First-Order Methods in Optimization* (2017), §10.4 and §10.6.
- 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).
- Yu. Nesterov, *Lectures on Convex Optimization* (2018), §2.2.3–2.2.4.
- B. T. Polyak, *Introduction to Optimization* (1987), §7.2.1.