5. Subgradient method

- subgradient method
- convergence analysis
- optimal step size when $f^\star$ is known
- alternating projections
- optimality

Subgradient method

to minimize a nondifferentiable convex function f: choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k g^{(k-1)}, \quad k = 1, 2, \ldots$$

$g^{(k-1)}$ is **any** subgradient of f at $x^{(k-1)}$, and t_k is *step size*; or

$$x^{(k)} = x^{(k-1)} - s_k \frac{g^{(k-1)}}{\|g^{(k-1)}\|_2}, \quad k = 1, 2, \ldots$$

where $s_k = t_k \|g^{(k-1)}\|_2$ has the interpretation of *step length*

step size rules

- fixed step size: t_k constant
- fixed step length: $s_k = \|x^{(k)} - x^{(k-1)}\|_2$ constant
- diminishing: $t_k \to 0$, $\sum_{k=1}^{\infty} t_k = \infty$, similarly for $\{s_k\}$

Assumptions

- f has finite optimal value $f^\star$, minimizer $x^\star$
- f is convex, $\mathbf{dom}\, f = \mathbf{R}^n$
- f is Lipschitz continuous with constant $G > 0$:

$$|f(x) - f(y)| \leq G\|x - y\|_2 \qquad \forall x, y$$

  this is equivalent to $\|g\|_2 \leq G$ for all $g \in \partial f(x)$, all x

Analysis

the subgradient method is not a descent method

the key quantity in the analysis is the distance to the optimal set

with $x^+ = x^{(i)}$, $x = x^{(i-1)}$, $g = g^{(i-1)}$, $t = t_i$:

$$\begin{array}{rcl} \|x^+ - x^\star\|_2^2 & = & \|x - tg - x^\star\|_2^2 \\ & = & \|x - x^\star\|_2^2 - 2tg^T(x - x^\star) + t^2\|g\|_2^2 \\ & \leq & \|x - x^\star\|_2^2 - 2t\left(f(x) - f^\star\right) + t^2\|g\|_2^2 \end{array}$$

combine inequalities for $i = 1, \ldots, k$,

$$\begin{array}{rcl} 2\sum_{i=1}^{k} t_i \left(f(x^{(i-1)}) - f^\star\right) & \leq & \|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 + \sum_{i=1}^{k} t_i^2 \|g^{(i-1)}\|_2^2 \\ & \leq & \|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^{k} t_i^2 \|g^{(i-1)}\|_2^2 \end{array}$$

define $f_{\rm best}^{(k)} = \min_{0\leq i<k} f(x^{(i)})$, then

$$2(\sum_{i=1}^k t_i)\left(f_{\rm best}^{(k)} - f^\star\right) \leq \|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^k t_i^2 \|g^{(i-1)}\|_2^2$$

therefore

$$f_{\rm best}^{(k)} - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^k t_i^2 G^2}{2\sum_{i=1}^k t_i}$$

or, in terms of $s_i = t_i\|g^{(i-1)}\|_2$,

$$f_{\rm best}^{(k)} - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^k s_i^2}{2\sum_{i=1}^k s_i/G}$$

fixed step size $t_i = t$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2 + kt^2G^2}{2kt}$$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $G^2t/2$-suboptimal

fixed step length $s_i = s$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2 + ks^2}{2ks/G}$$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $Gs/2$-suboptimal

diminishing step size $t_i \to 0$, $\sum_{i=1}^{\infty} t_i = \infty$

$$f_{\rm best}^{(k)} - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2 + G^2 \displaystyle\sum_{i=1}^{k} t_i^2}{2 \displaystyle\sum_{i=1}^{k} t_i}$$

can show that $(\sum_{i=1}^{k} t_i^2)/(\sum_{i=1}^{k} t_i) \to 0$; hence, $f_{\rm best}^{(k)}$ converges to $f^\star$

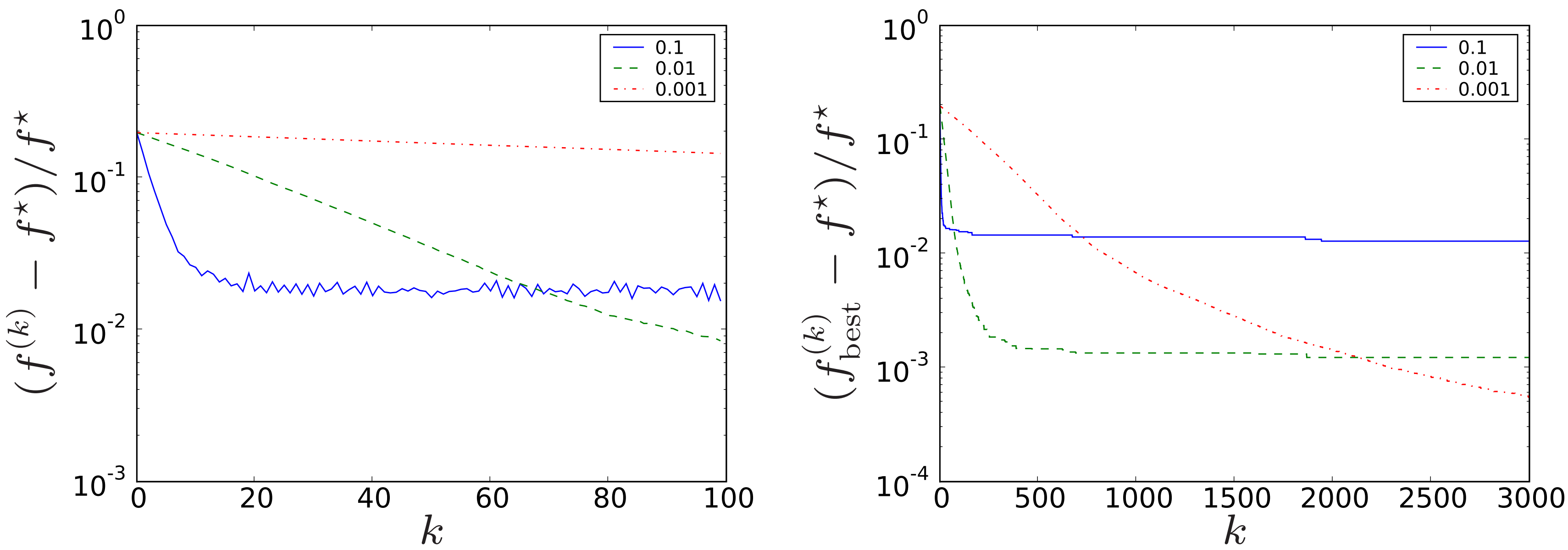
diminishing step length $s_i \to 0$, $\sum_{i=1}^{\infty} s_i = \infty$ works as well

Example: 1-norm minimization

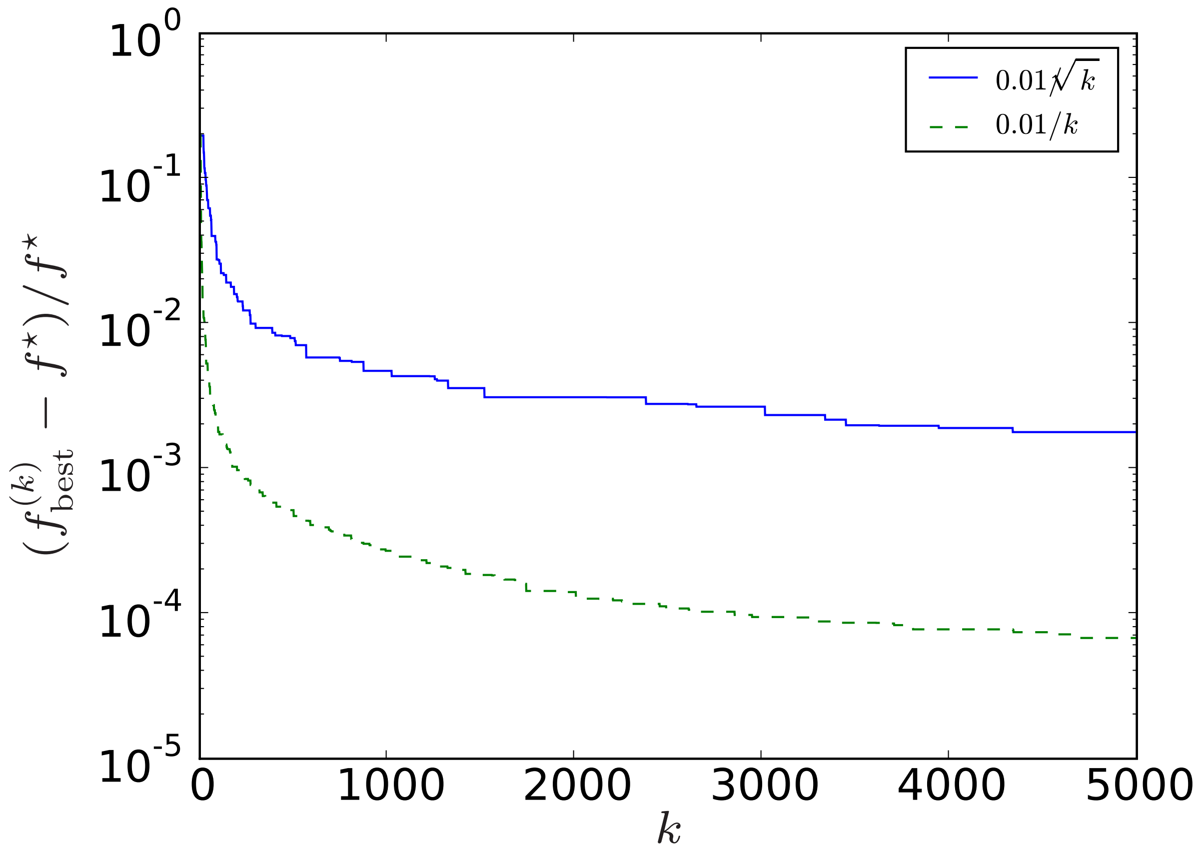
$$\text{minimize} \quad \|Ax - b\|_1 \qquad\qquad (A \in \mathbf{R}^{500\times 100}, b \in \mathbf{R}^{500})$$

subgradient is given by $A^T \mathbf{sign}(Ax - b)$

fixed steplength $s = 0.1,\ 0.01,\ 0.001$



diminishing step size $t_k = 0.01/\sqrt{k}$, $t_k = 0.01/k$



Optimal step size for fixed number of iterations

suppose N is fixed, and assume $\|x^{(0)} - x^\star\|_2 \leq R$, then from page 5–5:

$$f_{\text{best}}^{(N)} - f^\star \leq \frac{R^2 + \displaystyle\sum_{i=1}^{N} s_i^2}{2 \displaystyle\sum_{i=1}^{N} s_i / G}$$

- upper bound is minimized by step length $s_i = R/\sqrt{N}$, $i = 1, \ldots, N$
- resulting bound after N steps is

$$f_{\text{best}}^{(N)} - f^\star \leq \frac{GR}{\sqrt{N}}$$

#iterations to reach $f_{\text{best}}^{(N)} - f^\star \leq \epsilon$ is $O(1/\epsilon^2)$

Note about diminishing step length

if we use $s_i = R/\sqrt{i}$ for $i = 1, 2, \ldots$, then

$$f_{\mathrm{best}}^{(k)} - f^\star \;\leq\; \frac{R^2 + \sum_{i=1}^{k} s_i^2}{2\sum_{i=1}^{k} s_i/G} \;=\; GR\frac{1 + \sum_{i=1}^{k} \frac{1}{i}}{2\sum_{i=1}^{k} \frac{1}{\sqrt{i}}} \;\approx\; \frac{GR}{\sqrt{k}}\left(\frac{1+\gamma+\ln k}{4}\right)$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant

however, we can use the sum of last $\lfloor k/2 \rfloor$ terms on page 5–4 to obtain

$$f_{\mathrm{best}}^{(k)} - f^\star \;\leq\; \frac{R^2 + \sum_{i=\lfloor k/2 \rfloor}^{k} s_i^2}{2\sum_{i=\lfloor k/2 \rfloor}^{k} s_i/G} \;=\; GR\frac{1 + \sum_{i=\lfloor k/2 \rfloor}^{k} \frac{1}{i}}{2\sum_{i=\lfloor k/2 \rfloor}^{k} \frac{1}{\sqrt{i}}} \;\approx\; \frac{GR}{\sqrt{k}}\left(\frac{1+\ln 2}{4-2\sqrt{2}}\right)$$

Optimal step size when $f^\star$ is known

$$t_i = \frac{f(x^{(i-1)}) - f^\star}{\|g^{(i-1)}\|_2^2}$$

t_i minimizes r.h.s. in first inequality of page 5–4; optimized bound is

$$\|x^{(i)} - x^\star\|_2^2 \le \|x^{(i-1)} - x^\star\|_2^2 - \frac{(f(x^{(i-1)}) - f^\star)^2}{\|g^{(i-1)}\|_2^2}$$

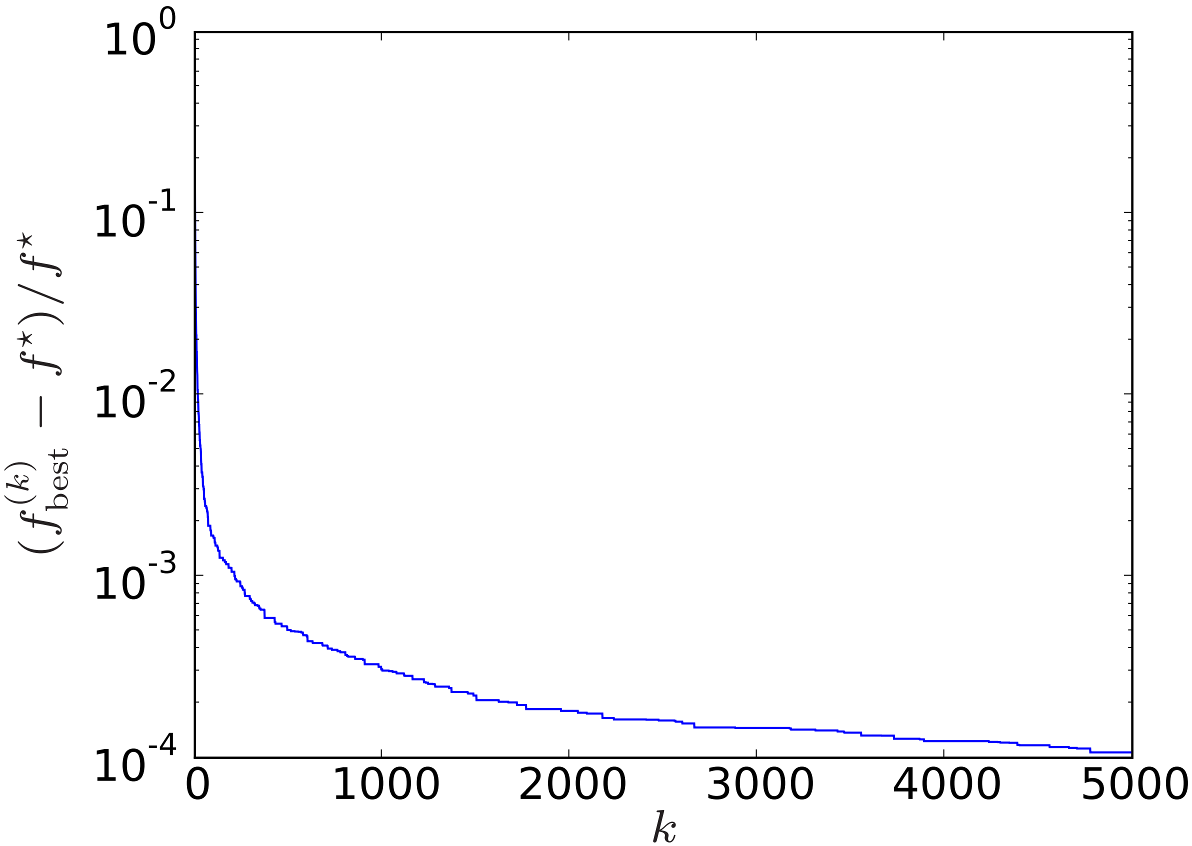
applying recursively gives

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

if $\|x^{(0)} - x^\star\|_2 \le R$,

$$\sum_{i=1}^{k} (f(x^{(i-1)}) - f^\star)^2 \le R^2 G^2, \qquad f_{\text{best}}^{(k)} - f^\star \le \frac{GR}{\sqrt{k}}$$

1-norm example with optimal step size



Exercise: Finding a point in the intersection of convex sets

to find point $x \in C = C_1 \cap \cdots \cap C_m$ (m closed convex sets):

$$\text{minimize} \quad f(x) = \max\{\mathbf{dist}(x, C_1), \ldots, \mathbf{dist}(x, C_m)\}$$

where

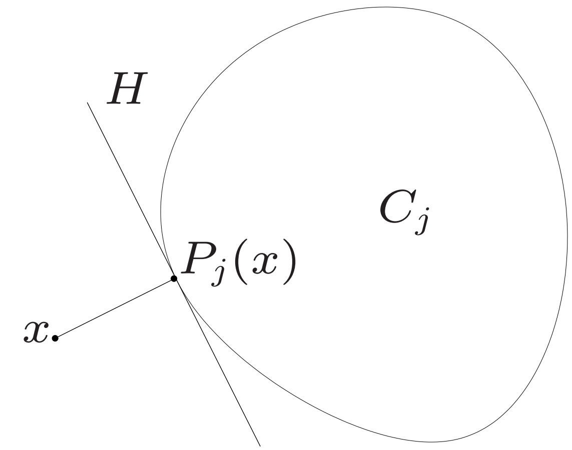
$$\mathbf{dist}(x, C_j) = \inf_{z \in C_j} \|x - z\|_2 = \|x - P_j(x)\|_2$$

(P_j is projection on C_j)

- $\mathbf{dist}(x, C_j)$ is a convex function if C_j is convex
- $f^\star = 0$ if the intersection is nonempty
- to find subgradient of f, need subgradient of distance to farthest set C_j

subgradient of distance to closed convex set C_j

$$C_j \subseteq H = \{z \mid (x - P_j(x))^T (z - P_j(x)) \leq 0\}$$



therefore

$$\textbf{dist}(y, C_j) \geq \frac{(x - P_j(x))^T (y - P_j(x))}{\|x - P_j(x)\|_2}$$

(for $y \notin H$, r.h.s. is distance to H; for $y \in H$, r.h.s. is nonpositive)

hence,

$$\textbf{dist}(y, C_j) \geq \|x - P_j(x)\|_2 + \frac{(x - P_j(x))^T (y - x)}{\|x - P_j(x)\|_2}$$

conclusion: $(x - P_j(x))/\,\textbf{dist}(x, C_j)$ is a subgradient at $x \notin C_j$

subgradient method with optimal step size for

$$\text{minimize} \quad f(x) = \max\{\mathbf{dist}(x, C_1), \ldots, \mathbf{dist}(x, C_m)\}$$

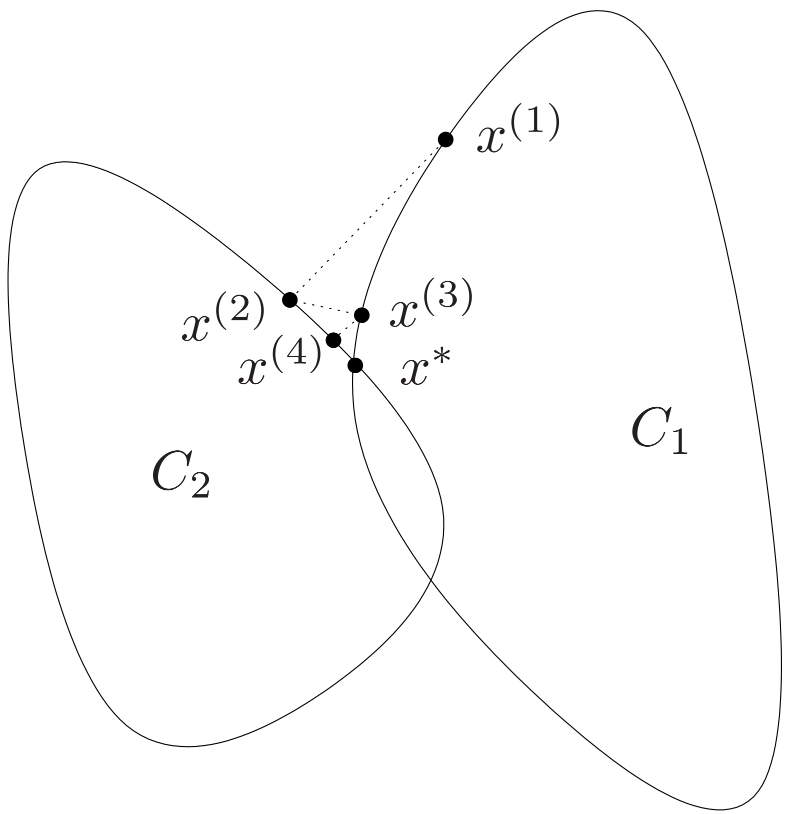
if C_j is the farthest set at iteration k ($i.e.$, $\mathbf{dist}(x^{(k-1)}, C_j) = f(x^{(k-1)})$):

$$\begin{aligned} x^{(k)} &= x^{(k-1)} - \frac{f(x^{(k-1)})}{\mathbf{dist}(x^{(k-1)}, C_j)}(x^{(k-1)} - P_j(x^{(k-1)})) \\ &= P_j(x^{(k-1)}) \end{aligned}$$

- a version of the *alternating projections* algorithm
- at each step, project the current point onto the farthest set
- for $m = 2$ sets, projections alternate onto one set, then the other
- convergence: $\mathbf{dist}(x^{(k)}, C) \to 0$ as $k \to \infty$

Alternating projections

first few iterations:



. . . $x^{(k)}$ eventually converges to a point $x^* \in C_1 \cap C_2$

Example: Positive semidefinite matrix completion

some entries of $X \in \mathbf{S}^n$ fixed; find values for others so $X \succeq 0$

- $C_1 = \mathbf{S}^n_+$

  projection onto C_1 by eigenvalue decomposition, truncation

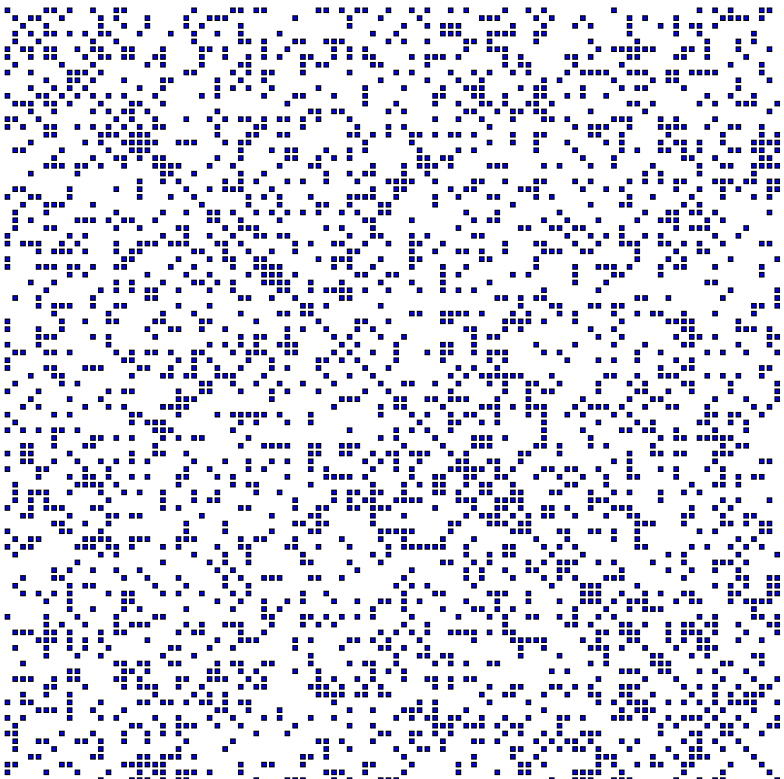
$$P_1(X) = \sum_{i=1}^{n} \max\{0, \lambda_i\} q_i q_i^T \qquad \text{if } X = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

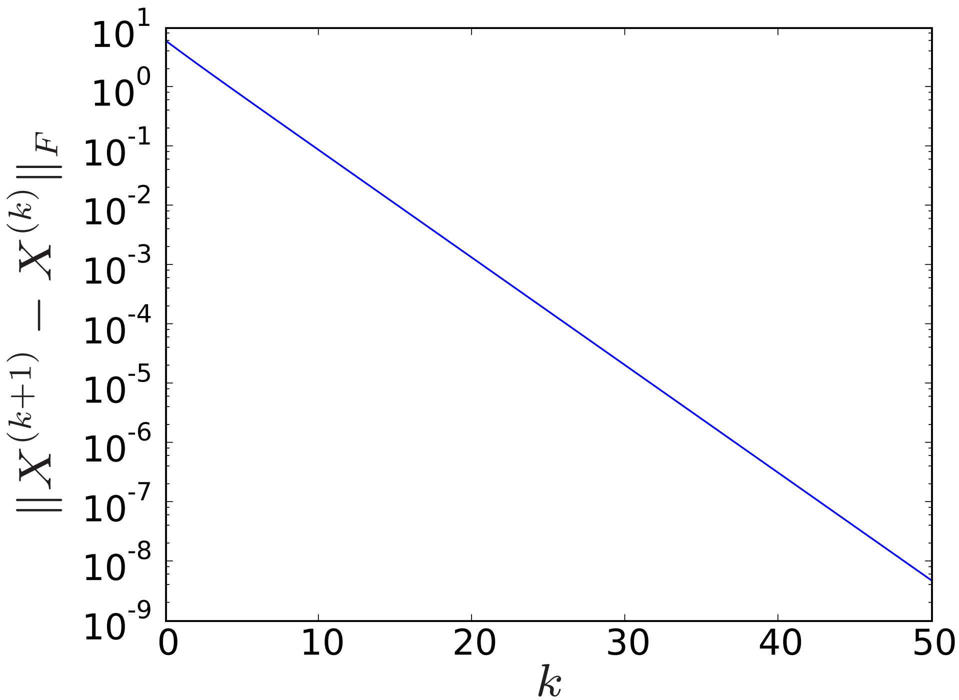
- C_2 is (affine) set in $\mathbf{S}^n$ with specified fixed entries

  projection of X onto C_2 by re-setting specified entries to fixed values

example: 100×100 matrix missing about 71% of its entries

initialize $X^{(0)}$ with unknown entries set to 0





Optimality of the subgradient method

can the $f_{\text{best}}^{(k)} - f^\star \leq GR/\sqrt{k}$ bound on page 5–10 be improved?

problem class

- f is convex, with a minimizer $x^\star$
- we know a starting point $x^{(0)}$ with $\|x^{(0)} - x^\star\|_2 \leq R$
- f is Lipschitz continuous with constant G on $\{x \mid \|x - x^{(0)}\|_2 \leq R\}$
- f is defined by an oracle: given x, oracle returns $f(x)$ and a subgradient

algorithm class: any subgradient method that

- k iterations of any method that chooses the iterate $x^{(i)}$ in the set $x^{(0)} + \mathrm{span}\{g^{(0)}, g^{(1)}, \ldots, g^{(i-1)}\}$

test problem

$$f(x) = \max_{i=1,\ldots,k} x_i + \frac{1}{2}\|x\|_2^2, \qquad x^{(0)} = 0$$

- solution: $x^\star = -\frac{1}{k}(1,\ldots,1,0,\ldots,0)$, $f^\star = -\frac{1}{2k}$
- Lipschitz continuous on $\{x \mid \|x\|_2 \leq R = 1/\sqrt{k}\}$ with $G = 1 + 1/\sqrt{k}$

oracle: returns subgradient $e_{\hat{j}} + x$ where
$\hat{j} = \min\{j \mid x_j = \max_{i=1,\ldots,k}\ x_i\}$

iteration: for $i = 0,\ldots,k-1$ entries $x_{i+1}^{(i)},\ldots,x_k^{(i)}$ are zero

$$f_{\mathrm{best}}^{(k)} - f^\star = \min_{i<k} f(x^{(i)}) - f^\star \geq -f^\star = \frac{GR}{2(1+\sqrt{k})}$$

conclusion: $O(1/\sqrt{k})$ bound cannot be improved

Summary

subgradient method

- handles general nondifferentiable convex problem
- often leads to very simple algorithms
- convergence can be very slow
- no good stopping criterion
- theoretical complexity: $O(1/\epsilon^2)$ iterations to find ϵ-suboptimal point
- an 'optimal' 1st-order method: $O(1/\epsilon^2)$ bound cannot be improved

References

- L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2012), UCLA.
- S. Boyd, lecture notes and slides for EE364b, Convex Optimization II
- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)

  §3.2.1 with the example on page 5–20 of this lecture