

### 3. Subgradient method

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

3-1

#### Subgradient method

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

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

$g^{(k-1)}$  is **any** subgradient of  $f$  at  $x^{(k-1)}$

#### step size rules

- fixed step:  $t_k$  constant
- fixed length:  $t_k \|g^{(k-1)}\|_2$  constant (i.e.,  $\|x^{(k)} - x^{(k-1)}\|_2$  constant)
- diminishing:  $t_k \rightarrow 0$ ,  $\sum_{k=1}^{\infty} t_k = \infty$

## Assumptions

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

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

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

$$\begin{aligned} \|x^{(i)} - x^*\|_2^2 &= \left\| x^{(i-1)} - t_i g^{(i-1)} - x^* \right\|_2^2 \\ &= \|x^{(i-1)} - x^*\|_2^2 - 2t_i g^{(i-1)T} (x^{(i-1)} - x^*) + t_i^2 \|g^{(i-1)}\|_2^2 \\ &\leq \|x^{(i-1)} - x^*\|_2^2 - 2t_i (f(x^{(i-1)}) - f^*) + t_i^2 \|g^{(i-1)}\|_2^2 \end{aligned}$$

define  $f_{\text{best}}^{(k)} = \min_{0 \leq i < k} f(x^{(i)})$ , and combine inequalities for  $i = 1, \dots, k$ :

$$\begin{aligned} 2\left(\sum_{i=1}^k t_i\right) (f_{\text{best}}^{(k)} - f^*) &\leq \|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 + \sum_{i=1}^k t_i^2 \|g^{(i-1)}\|_2^2 \\ &\leq \|x^{(0)} - x^*\|_2^2 + \sum_{i=1}^k t_i^2 \|g^{(i-1)}\|_2^2 \end{aligned}$$

**fixed step size**  $t_i = t$

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

- 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**  $t_i = s/\|g^{(i-1)}\|_2$

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

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

Subgradient method

3-5

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

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

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

Subgradient method

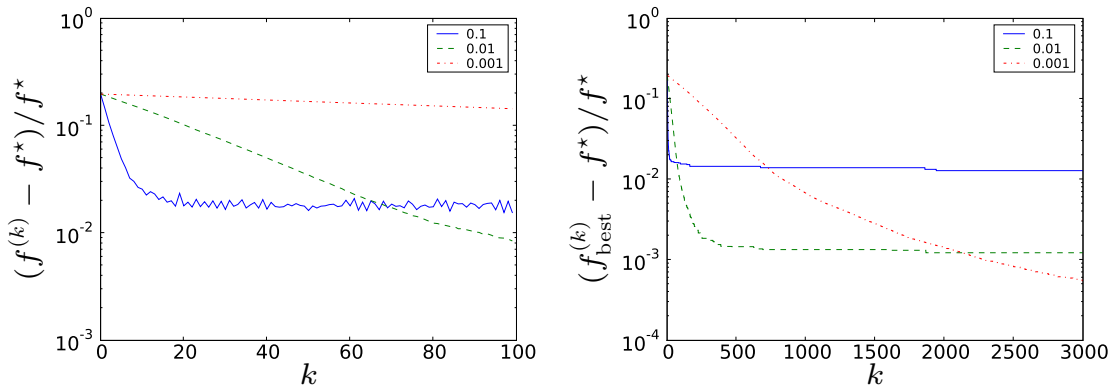
3-6

## Example: 1-norm minimization

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

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

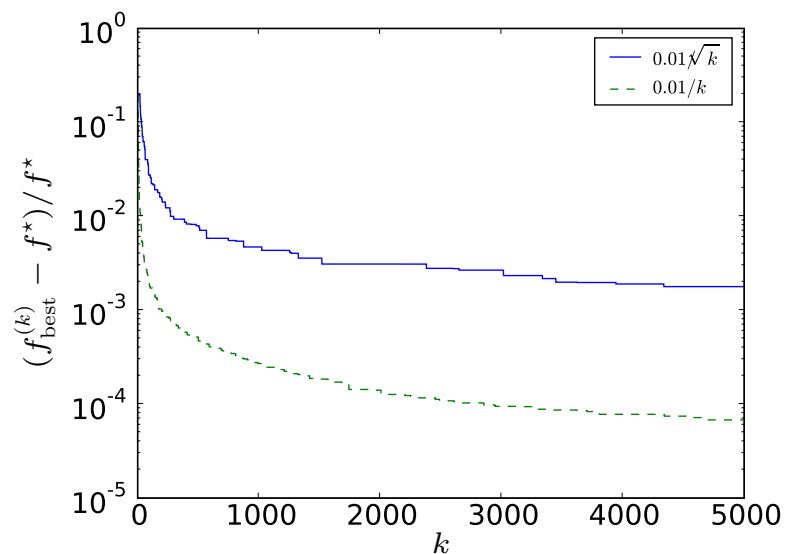
**fixed steplength**  $t_k = s/\|g^{(k-1)}\|_2$ ,  $s = 0.1, 0.01, 0.001$



Subgradient method

3-7

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



Subgradient method

3-8

## Optimal step size for fixed number of iterations

from page 3–4: if  $\|x^{(0)} - x^*\|_2 \leq R$  and  $s_i = t_i \|g^{(i-1)}\|_2$

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

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

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

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

## Optimal step size when $f^*$ is known

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

$t_i$  minimizes r.h.s. in first inequality of page 3–4; optimized bound is

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

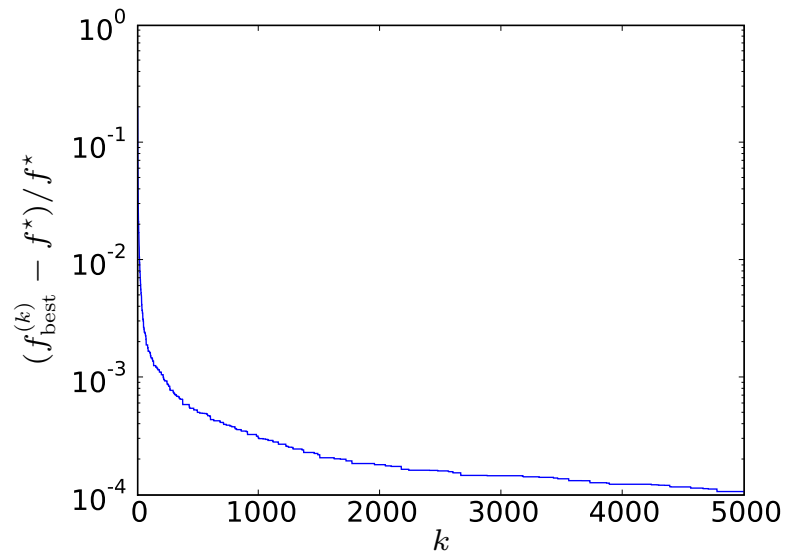
applying recursively gives

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

if  $\|x^{(0)} - x^*\|_2 \leq R$ ,

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

## 1-norm example with optimal step size



## Finding a point in the intersection of convex sets

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

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

where

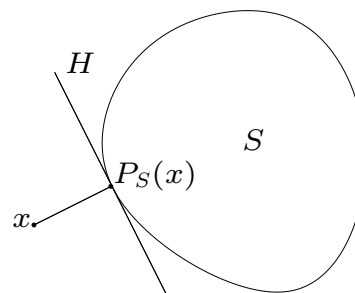
$$\mathbf{dist}(x, C_j) = \inf_{z \in C_j} \|x - z\|_2 = \|x - P_{C_j}(x)\|_2$$

( $P_{C_j}$  is projection on  $C_j$ )

- $\mathbf{dist}(x, C_j)$  is a convex function if  $C_j$  is convex
- $f^* = 0$  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  $S$

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



therefore

$$\mathbf{dist}(y, S) \geq \frac{(x - P_S(x))^T(y - P_S(x))}{\|x - P_S(x)\|_2}$$

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

hence,

$$\mathbf{dist}(y, S) \geq \|x - P_S(x)\|_2 + \frac{(x - P_S(x))^T(y - x)}{\|x - P_S(x)\|_2}$$

conclusion:  $(x - P_S(x)) / \mathbf{dist}(x, S)$  is a subgradient at  $x \notin S$

**subgradient method** with optimal step size for

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

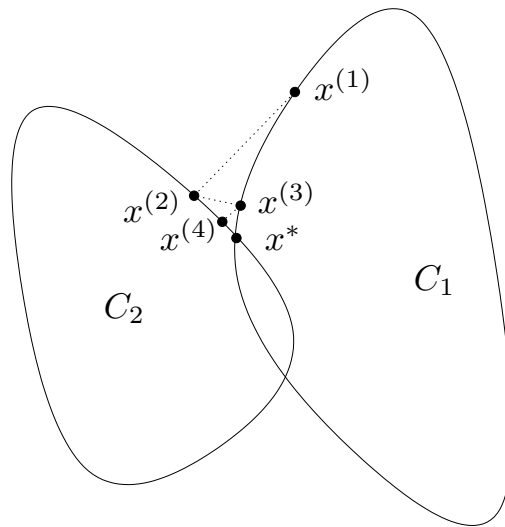
if  $C_j$  is the farthest set at iteration  $k$  (i.e.,  $\mathbf{dist}(x^{(k-1)}, C_j) = f(x^{(k-1)})$ ):

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

- a version of the famous *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:  $\mathbf{dist}(x^{(k)}, C) \rightarrow 0$  as  $k \rightarrow \infty$

# 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

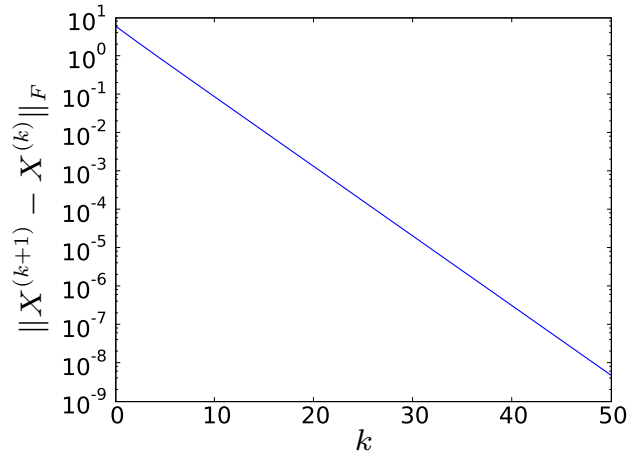
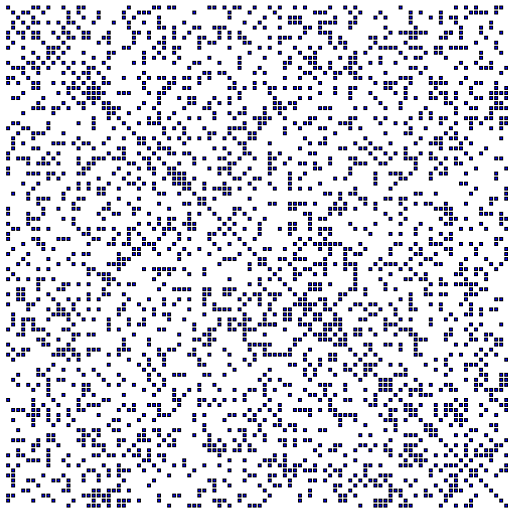
$$P_{C_1}(X) = \sum_{i=1}^n \max\{0, \lambda_i\} q_i q_i^T \quad \text{if } X = \sum_{i=1}^n \lambda_i q_i q_i^T$$

- $C_2$  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:**  $100 \times 100$  matrix missing about 71% of its entries

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



## Optimality of the subgradient method

can the  $f_{\text{best}}^{(k)} - f^* \leq GR/\sqrt{k}$  bound on page 3-9 be improved?

### problem class

- $f$  is convex, with a minimizer  $x^*$
- we know a starting point  $x^{(0)}$  with  $\|x^{(0)} - x^*\|_2 \leq R$
- $f$  is Lipschitz continuous with constant  $G$  on  $\{x \mid \|x - x^{(0)}\|_2 \leq R\}$

**algorithm class:** any subgradient method that

- obtains function information via an oracle (black box): for given  $x$ , the oracle returns  $f(x)$  and a subgradient at  $x$
- chooses the iterate  $x^{(k)}$  in the set  $x^{(0)} + \text{span}\{g^{(0)}, g^{(1)}, \dots, g^{(k-1)}\}$

## test problem

$$f(x) = \max_{i=1,\dots,n} x_i + \frac{1}{2}\|x\|_2^2, \quad x^{(0)} = 0$$

- solution:  $x^* = -(1/n)\mathbf{1}$ ,  $f^* = -1/(2n)$
- Lipschitz continuous on  $\{x \mid \|x\|_2 \leq R = 1/\sqrt{n}\}$  with  $G = 1 + 1/\sqrt{n}$

**oracle:** returns subgradient  $e_j + x$  where  $j = \min\{i \mid x_i = \max_j x_j\}$

**accuracy** after  $k - 1$  iterations of any method in the algorithm class

- $x_{i+1}^{(i)} = \dots = x_n^{(i)} = 0$  for  $i = 0, \dots, n - 1$ ; hence  $f_{\text{best}}^{(k)} \geq 0$  for  $k \leq n$
- if  $n = k$ ,

$$f_{\text{best}}^{(k)} - f^* \geq \frac{1}{2k} = \frac{GR}{2(1 + \sqrt{k})}$$

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

## Summary

### subgradient method

- often very slow
- no good stopping criterion
- theoretical complexity:  $O(1/\epsilon^2)$  iterations to find  $\epsilon$ -suboptimal point

## References and sources

- 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 3–18 of this lecture