

### 3. 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

$$x_{k+1} = x_k - t_k g_k, \quad k = 0, 1, \dots$$

$g_k$  is any subgradient of  $f$  at  $x_k$

## Step size rules

- fixed step:  $t_k$  constant
- fixed length:  $t_k \|g_k\|_2 = \|x_{k+1} - x_k\|_2$  is constant
- diminishing:  $t_k \rightarrow 0$  and  $\sum_{k=0}^{\infty} t_k = \infty$

# Assumptions

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

$$|f(x) - f(y)| \leq G\|x - y\|_2 \quad \text{for all } x, y$$

this is equivalent to  $\|g\|_2 \leq G$  for all  $x$  and  $g \in \partial f(x)$  (see next page)

*Proof.*

- assume  $\|g\|_2 \leq G$  for all subgradients; choose  $g_y \in \partial f(y)$ ,  $g_x \in \partial f(x)$ :

$$g_x^T(x - y) \geq f(x) - f(y) \geq g_y^T(x - y)$$

by the Cauchy–Schwarz inequality

$$G\|x - y\|_2 \geq f(x) - f(y) \geq -G\|x - y\|_2$$

- assume  $\|g\|_2 > G$  for some  $g \in \partial f(x)$ ; take  $y = x + g/\|g\|_2$ :

$$\begin{aligned} f(y) &\geq f(x) + g^T(y - x) \\ &= f(x) + \|g\|_2 \\ &> f(x) + G \end{aligned}$$

# Analysis

- the subgradient method is not a descent method
- therefore  $f_{\text{best},k} = \min_{i=0,\dots,k} f(x_i)$  can be less than  $f(x_k)$
- the key quantity in the analysis is the distance to the optimal set

## Progress in one iteration

- distance to  $x^\star$ :

$$\begin{aligned}\|x_{i+1} - x^\star\|_2^2 &= \|x_i - t_i g_i - x^\star\|_2^2 \\ &= \|x_i - x^\star\|_2^2 - 2t_i g_i^T (x_i - x^\star) + t_i^2 \|g_i\|_2^2 \\ &\leq \|x_i - x^\star\|_2^2 - 2t_i (f(x_i) - f^\star) + t_i^2 \|g_i\|_2^2\end{aligned}$$

- best function value: combine inequalities for  $i = 0, \dots, k$ :

$$\begin{aligned}2\left(\sum_{i=0}^k t_i\right)(f_{\text{best},k} - f^\star) &\leq \|x_0 - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2 + \sum_{i=0}^k t_i^2 \|g_i\|_2^2 \\ &\leq \|x_0 - x^\star\|_2^2 + \sum_{i=0}^k t_i^2 \|g_i\|_2^2\end{aligned}$$

# Fixed step size and fixed step length

**Fixed step size:**  $t_i = t$  with  $t$  constant

$$f_{\text{best},k} - f^\star \leq \frac{\|x_0 - x^\star\|_2^2}{2(k+1)t} + \frac{G^2 t}{2}$$

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

**Fixed step length:**  $t_i = s/\|g_i\|_2$  with  $s$  constant

$$f_{\text{best},k} - f^\star \leq \frac{G\|x_0 - x^\star\|_2^2}{2(k+1)s} + \frac{Gs}{2}$$

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

## Diminishing step size

$$t_i \rightarrow 0, \quad \sum_{i=0}^{\infty} t_i = \infty$$

- bound on function value:

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

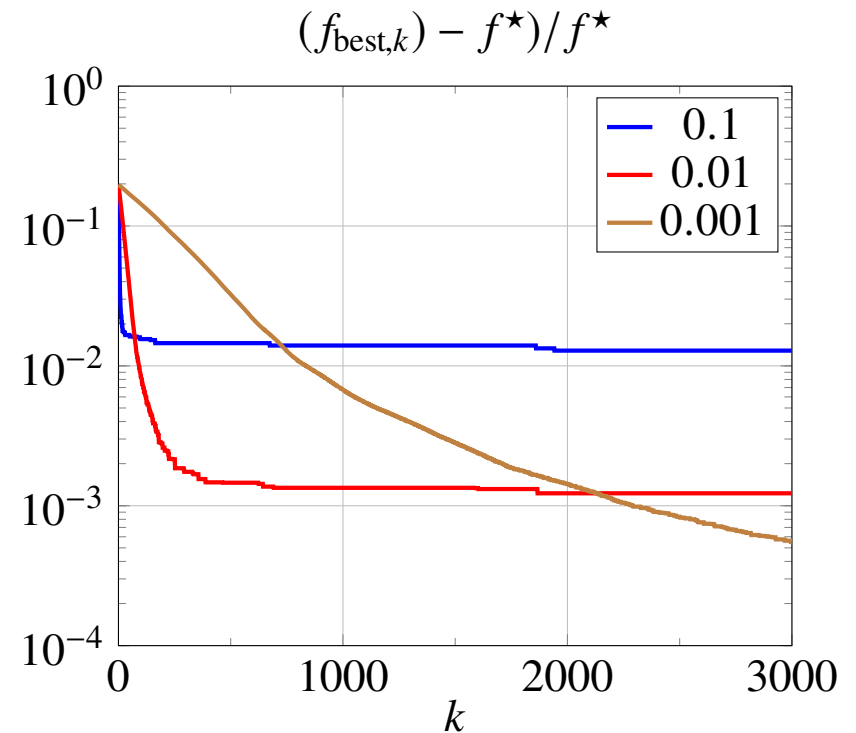
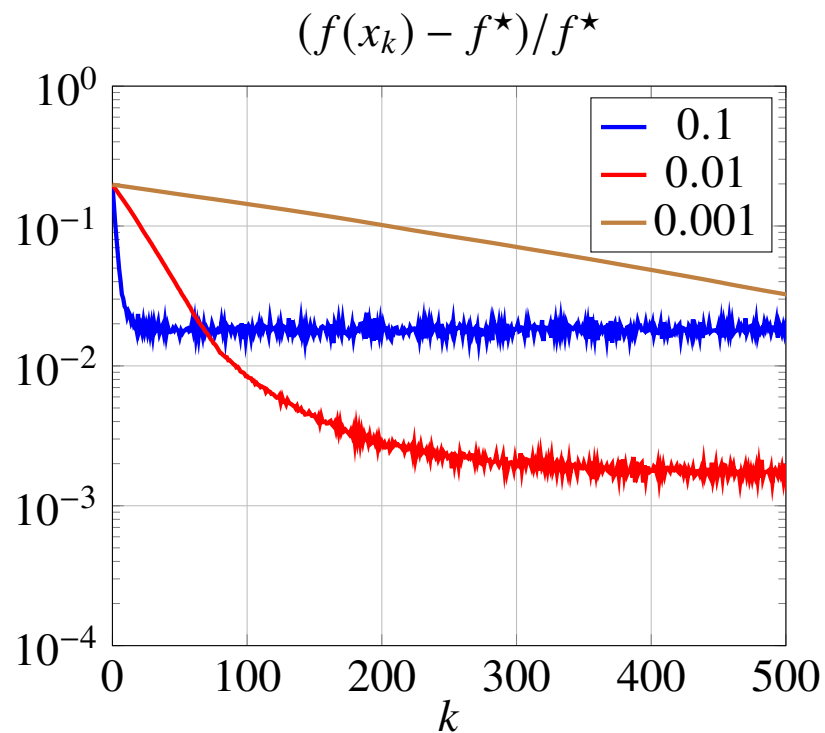
- can show that  $(\sum_{i=0}^k t_i^2)/(\sum_{i=0}^k t_i) \rightarrow 0$ ; hence,  $f_{\text{best},k}$  converges to  $f^{\star}$
- examples:  $t_i = \tau/(i+1)$  or  $t_i = \tau/\sqrt{i+1}$

## Example: 1-norm minimization

$$\text{minimize } \|Ax - b\|_1$$

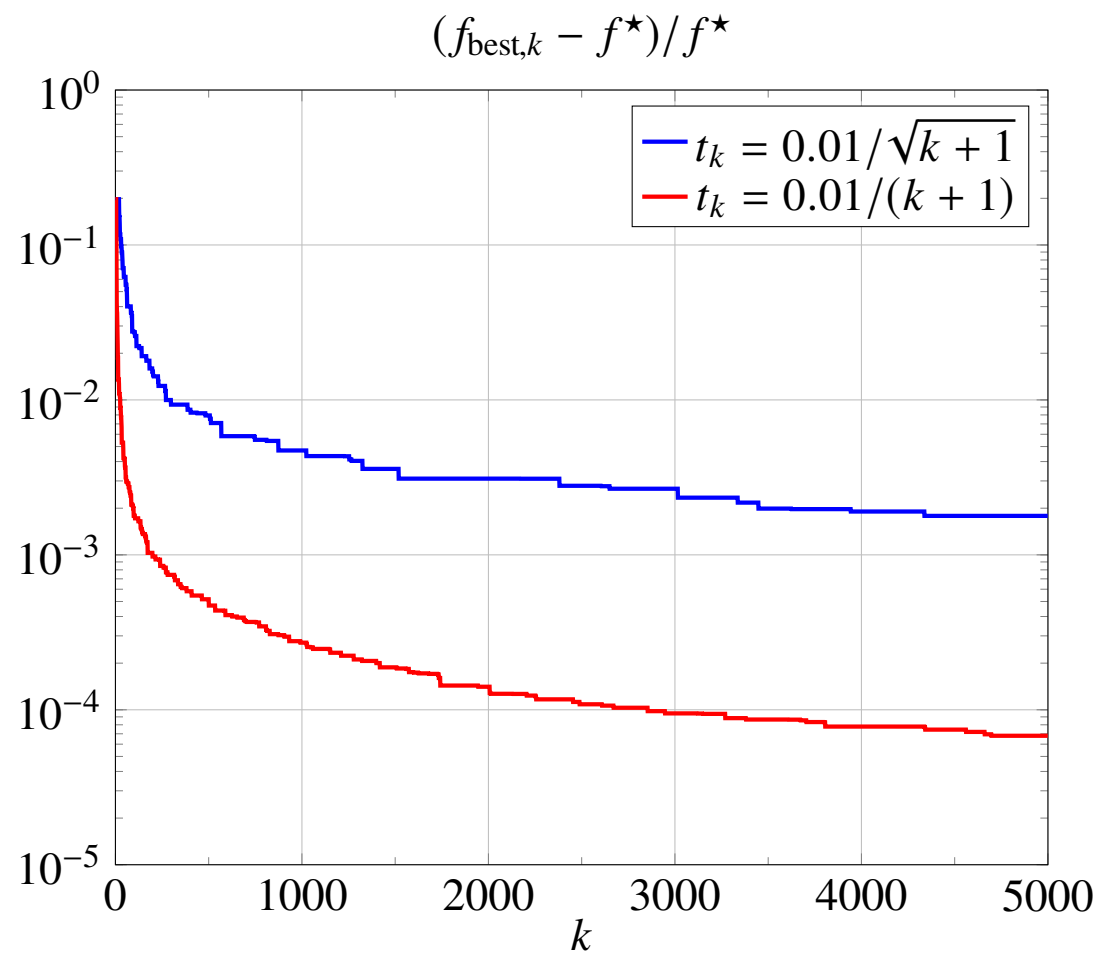
- subgradient is given by  $A^T \text{sign}(Ax - b)$
- example with  $A \in \mathbf{R}^{500 \times 100}$ ,  $b \in \mathbf{R}^{500}$

**Fixed steplength**  $t_k = s/\|g_k\|_2$  for  $s = 0.1, 0.01, 0.001$





**Diminishing step size:**  $t_k = 0.01/\sqrt{k+1}$  and  $t_k = 0.01/(k+1)$



## Optimal step size for fixed number of iterations

from page 3.5: if  $s_i = t_i \|g_i\|_2$  and  $\|x_0 - x^\star\|_2 \leq R$ , then

$$f_{\text{best},k} - f^\star \leq \frac{R^2 + \sum_{i=0}^k s_i^2}{2 \sum_{i=0}^k s_i / G}$$

- for given  $k$ , the right-hand side is minimized by the fixed step length

$$s_i = s = \frac{R}{\sqrt{k+1}}$$

- the resulting bound after  $k$  steps is

$$f_{\text{best},k} - f^\star \leq \frac{GR}{\sqrt{k+1}}$$

- this guarantees an accuracy  $f_{\text{best},k} - f^\star \leq \epsilon$  in  $k = O(1/\epsilon^2)$  iterations

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

- the right-hand side in the first inequality of page 3.5 is minimized by

$$t_i = \frac{f(x_i) - f^\star}{\|g_i\|_2^2}$$

- the optimized bound is

$$\frac{(f(x_i) - f^\star)^2}{\|g_i\|_2^2} \leq \|x_i - x^\star\|_2^2 - \|x_{i+1} - x^\star\|_2^2$$

- applying this recursively from  $i = 0$  to  $i = k$  (and using  $\|g_i\|_2 \leq G$ ) gives

$$f_{\text{best},k} - f^\star \leq \frac{G\|x_0 - x^\star\|_2}{\sqrt{k+1}}$$

## Exercise: find point in intersection of convex sets

find a point in the intersection of  $m$  closed convex sets  $C_1, \dots, C_m$ :

$$\text{minimize } f(x) = \max \{f_1(x), \dots, f_m(x)\}$$

where  $f_j(x) = \inf_{y \in C_j} \|x - y\|_2$  is Euclidean distance of  $x$  to  $C_j$

- $f^\star = 0$  if the intersection is nonempty
- (from page 2.14)  $g \in \partial f(\hat{x})$  if  $g \in \partial f_j(\hat{x})$  and  $C_j$  is farthest set from  $\hat{x}$
- (from page 2.20) subgradient  $g \in \partial f_j(\hat{x})$  follows from projection  $P_j(\hat{x})$  on  $C_j$ :

$$g = 0 \quad \text{if } \hat{x} \in C_j, \quad g = \frac{1}{\|\hat{x} - P_j(\hat{x})\|_2} (\hat{x} - P_j(\hat{x})) \quad \text{if } \hat{x} \notin C_j$$

note that  $\|g\|_2 = 1$  if  $\hat{x} \notin C_j$

## Subgradient method

- optimal step size (page 3.11) for  $f^\star = 0$  and  $\|g_i\|_2 = 1$  is  $t_i = f(x_i)$
- at iteration  $k$ , find farthest set  $C_j$  (with  $f(x_k) = f_j(x_k)$ ), and take

$$\begin{aligned}x_{k+1} &= x_k - \frac{f(x_k)}{f_j(x_k)}(x_k - P_j(x_k)) \\ &= P_j(x_k)\end{aligned}$$

at each step, we project the current point onto the farthest set

- a version of the *alternating projections* algorithm
- for  $m = 2$ , projections alternate onto one set, then the other
- later, we will see faster versions of this that are almost as simple

# Optimality of the subgradient method

can the  $f_{\text{best},k} - f^\star \leq GR/\sqrt{k+1}$  bound on page 3.10 be improved?

## Problem class

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

## Algorithm class

- algorithm can choose any  $x^{(i+1)}$  from the set  $x^{(0)} + \text{span}\{g^{(0)}, g^{(1)}, \dots, g^{(i)}\}$
- we stop after a fixed number  $k$  of iterations

## Test problem and oracle

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

- subdifferential  $\partial f(x) = \{x\} + \text{conv}\{e_j \mid 1 \leq j \leq k+1, x_j = \max_{i=1,\dots,k+1} x_i\}$
- solution and optimal value

$$x^\star = -(\underbrace{\frac{1}{k+1}, \dots, \frac{1}{k+1}}_{k+1 \text{ times}}, 0, \dots, 0), \quad f^\star = -\frac{1}{2(k+1)}$$

- distance of starting point to solution is  $R = \|x^{(0)} - x^\star\|_2 = 1/\sqrt{k+1}$
- Lipschitz constant on  $\{x \mid \|x - x^\star\|_2 \leq R\}$ :

$$G = \sup_{g \in \partial f(x), \|x - x^\star\|_2 \leq R} \|g\|_2 \leq \frac{2}{\sqrt{k+1}} + 1$$

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

## Iteration

- after  $i \leq k$  iterations of any algorithm in the algorithm class,

$$x^{(i)} = (x_1^{(i)}, \dots, x_i^{(i)}, 0, \dots, 0), \quad f(x^{(i)}) \geq \|x^{(i)}\|_2 \geq 0$$

- suboptimality after  $k$  iterations

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

## Conclusion

- example shows that  $O(GR/\sqrt{k})$  bound cannot be improved
- subgradient method is “optimal” (for this problem and algorithm class)



## 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  $\epsilon$ -suboptimal point
- an “optimal” first-order method:  $O(1/\epsilon^2)$  bound cannot be improved

# References

- S. Boyd, *Lecture slides and notes for EE364b, Convex Optimization II*.
- Yu. Nesterov, *Lectures on Convex Optimization* (2018), section 3.2.3. The example on page 3.15 is in §3.2.1.
- B. T. Polyak, *Introduction to Optimization* (1987), section 5.3.