

# Subgradient Method

Acknowledgement: this slides is based on Prof. Lieven Vandenberghes lecture notes

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

# Subgradient method

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

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

$g^{(k-1)}$  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

$$\|g\|_2 \leq G \quad \forall g \in \partial f(x), \forall x$$

(see next page)

*proof*

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

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

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  $y = x + g/\|g\|_2$ :

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

# Analysis

- the subgradient method is not a descent method
- the key quantity in the analysis is the distance to the optimal set with  $x^+ = x^{(i)}$ ,  $x = x^{(i-1)}$ ,  $g = g^{(i-1)}$ ,  $t = t_i$ :

$$\begin{aligned}\|x^+ - x^*\|_2^2 &= \|x - tg - x^*\|_2^2 \\ &= \|x - x^*\|_2^2 - 2tg^\top(x - x^*) + t^2\|g\|_2^2 \\ &\leq \|x - x^*\|_2^2 - 2t(f(x) - f^*) + t^2\|g\|_2^2\end{aligned}$$

combine inequalities for  $i = 1, \dots, k$ , and define

$$f_{\text{best}}^{(k)} = \min_{0 \leq i \leq k} f(x^{(i)}):$$

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

## fixed step size

$$t_i = t$$

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

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

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

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

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

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

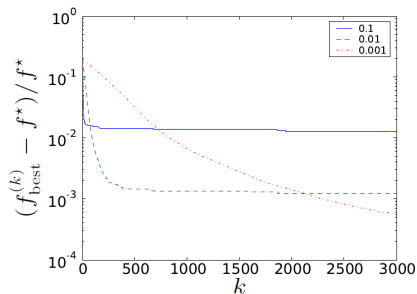
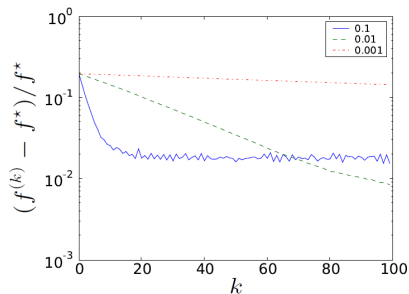
can show that  $(\sum_{i=1}^k t_i^2) / \sum_{i=1}^k t_i \rightarrow 0$ ; hence,  $f_{\text{best}}^{(k)}$  converges to  $f^*$

# Example: 1-norm minimization

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

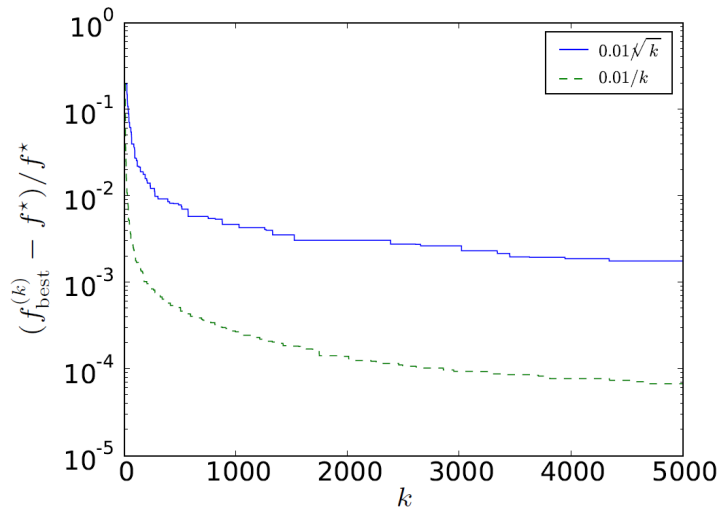
subgradient is given by  $A^\top \mathbf{sign}(Ax - b)$

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





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



# Optimal step size for fixed number of iterations

from page 6: if  $s_i = t_i \|g^{(i-1)}\|_2$  and  $\|x^{(0)} - x^*\|_2 \leq R$ :

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

- for given  $k$ , bound is minimized by fixed step length  $s_i = s = R/\sqrt{k}$
- resulting bound after  $k$  steps is

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

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

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

right-hand side in first inequality of page 6 is minimized by

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

optimized bound is

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

applying recursively (with  $\|x^{(0)} - x^*\|_2 \leq R$  and  $\|g^{(i)}\|_2 \leq G$ ) gives

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

## Exercise: find point in intersection of convex sets

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

$$\min_x f(x) = \max\{d_1(x), \dots, d_m(x)\}$$

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

- $f^* = 0$  if the intersection is nonempty
- $g \in \partial f(\hat{x})$  if  $g \in \partial d_j(\hat{x})$  and  $C_j$  is farthest set from  $\hat{x}$
- subgradient  $g \in \partial d_j(\hat{x})$  from projection  $P_j(\hat{x})$  on  $C_j$ :

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

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

## subgradient method with optimal step size

- optimal step size for  $f^* = 0$  and  $\|g^{(i-1)}\|_2 = 1$  is  $t_i = f(x^{(i-1)})$ .
- at iteration  $k$ , find farthest set  $C_j$  (with  $f(x^{(k-1)}) = d_j(x^{(k-1)})$ ); take

$$\begin{aligned}x^{(k)} &= x^{(k-1)} - \frac{f(x^{(k-1)})}{d_j(x^{(k-1)})} (x^{(k-1)} - P_j(x^{(k-1)})) \\ &= P_j(x^{(k-1)})\end{aligned}$$

- a version of the *alternating projections* algorithm
- at each step, project the current point onto the farthest set
- for  $m = 2$ , projections alternate onto one set, then the other

# LASSO 问题求解

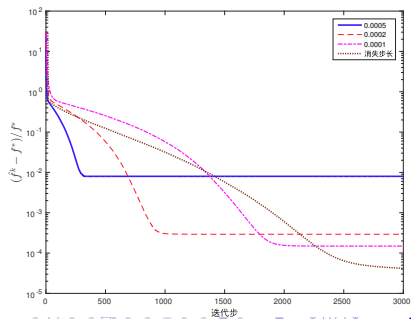
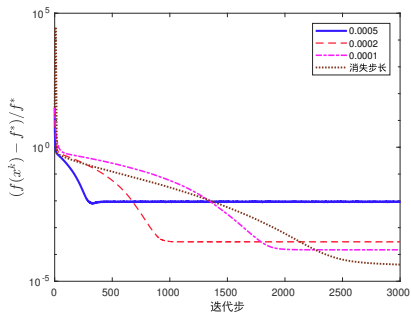
考虑LASSO 问题

$$\min f(x) = \frac{1}{2} \|Ax - b\|^2 + \mu \|x\|_1, \quad (1)$$

$f(x)$ 的一个次梯度为  $g = A^T(Ax - b) + \mu \text{sign}(x)$ , 其中  $\text{sign}(x)$  是关于  $x$  逐分量的符号函数. 因此的次梯度算法为

$$x^{k+1} = x^k - \alpha_k (A^T(Ax^k - b) + \mu \text{sign}(x^k)),$$

步长  $\alpha_k$  可选为固定步长或消失步长.



# LASSO 问题求解

对于  $\mu = 10^{-2}, 10^{-3}$ , 采用连续化次梯度算法进行求解. 若  $\mu_t > \mu$ , 则取固定步长  $\frac{1}{\lambda_{\max}(A^T A)}$ ; 若  $\mu_t = \mu$ , 则取步长

$$\frac{1}{\lambda_{\max}(A^T A) \cdot (\max\{k, 100\} - 99)},$$

其中  $k$  为迭代步数

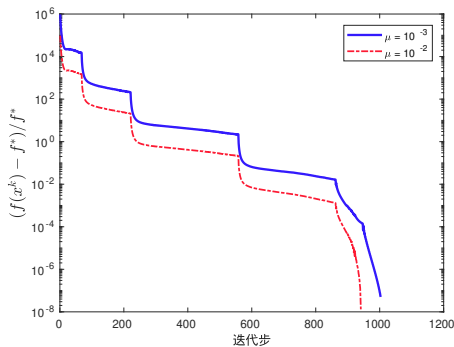


Figure: LASSO 问题在不同正则化参数下的求解结果

## Example: Positive semidefinite matrix completion

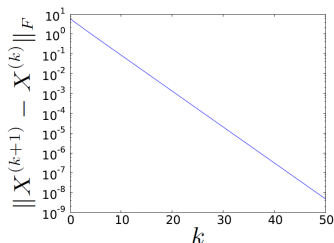
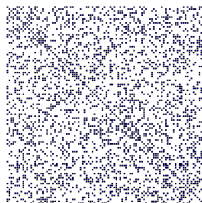
some entries of  $X \in \mathbf{S}^n$  fixed; find values for others so  $X \succeq 0$

- $C_1 = \mathbf{S}_+^n$ ,  $C_2$  is (affine) set in  $\mathbf{S}^n$  with specified fixed entries
- projection onto  $C_1$  by eigenvalue decomposition, truncation

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

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

100 × 100  
matrix missing  
71% entries





# 算法的渐进收敛速度

设 $\{x^k\}$ 为算法产生的迭代点列且收敛于 $x^*$

- 算法（点列）**Q-线性收敛**：对充分大的 $k$ 有

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq a, \quad a \in (0, 1)$$

- 算法（点列）**Q-超线性收敛**：对充分大的 $k$ 有

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0,$$

- 算法（点列）**Q-次线性收敛**：对充分大的 $k$ 有

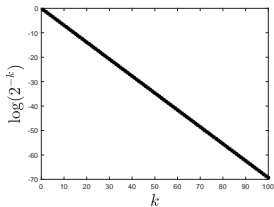
$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 1,$$

- 算法（点列）**Q-二次收敛**：对充分大的 $k$ 有

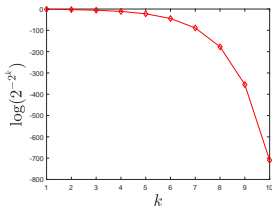
$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} \leq a, \quad a > 0,$$

# 算法的渐进收敛速度

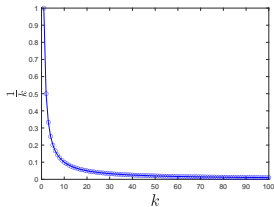
我们举例来更直观地展示不同的Q-收敛速度，参见下图（图中对所考虑的点列作了适当的变换）。点列 $\{2^{-k}\}$ 是Q-线性收敛的，点列 $\{2^{-2^k}\}$ 是Q-二次收敛的（也是Q-超线性收敛的），点列 $\{\frac{1}{k}\}$ 是Q-次线性收敛的。一般来说，具有Q-超线性收敛速度和Q-二次收敛速度的算法是收敛较快的



(a) Q-线性收敛



(b) Q-二次收敛



(c) Q-次线性收敛

Figure: 不同Q-收敛速度比较

# 算法的渐进收敛速度

- 算法（点列）**R-线性收敛**：设 $\{x^k\}$ 为算法产生的迭代点且收敛于 $x^*$ ，若存在**Q-线性收敛**于0的非负序列 $t_k$ 并且

$$\|x^k - x^*\| \leq t_k$$

对任意的 $k$ 成立。类似地，可定义**R-超线性收敛**和**R-二次收敛**等收敛速度。从**R-收敛速度**的定义可以看出序列 $\{\|x^k - x^*\|\}$ 被另一趋于0的序列 $\{t_k\}$ 控制。当知道 $t_k$ 的形式时，我们也称算法（点列）的收敛速度为 $\mathcal{O}(t_k)$ 。

- **算法复杂度**。设 $x^*$ 为全局极小点，某一算法产生的迭代序列 $\{x^k\}$ 满足

$$f(x^k) - f(x^*) \leq \frac{c}{\sqrt{k}}, \quad \forall k > 0,$$

其中 $c > 0$ 为常数。如果需要计算算法满足精度 $f(x^k) - f(x^*) \leq \varepsilon$ 所需的迭代次数，只需令 $\frac{c}{\sqrt{k}} \leq \varepsilon$ 则得到 $k \geq \frac{c^2}{\varepsilon^2}$ ，因此该优化算法对

应的（迭代次数）复杂度为 $N(\varepsilon) = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ 。

# Optimality of the subgradient method

can the  $f_{\text{best}}^{(k)} - f^* \leq \frac{GR}{\sqrt{k}}$  bound on page 11 be improved?

## problem class

- $f$  is convex, with a minimizer  $x^*$
- we know a starting point  $x^{(0)}$  with  $\|x^{(0)} - x^*\|_2 \leq R$
- we know the Lipschitz constant  $G$  of  $f$  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:**  $k$  iterations of any method that chooses  $x^{(i)}$  in

$$x^{(0)} + \text{span}\{g^{(0)}, g^{(1)}, \dots, g^{(i-1)}\}$$

# test problem and oracle

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

- solution:  $x^* = -\frac{1}{k} (\underbrace{1, \dots, 1}_k, \underbrace{0, \dots, 0}_{n-k})$  and  $f^* = -\frac{1}{2k}$
- $R = \|x^{(0)} - x^*\|_2 = \frac{1}{\sqrt{k}}$  and  $G = 1 + \frac{1}{\sqrt{k}}$
- oracle returns subgradient  $e_{\hat{j}} + x$  where  $\hat{j} = \min\{j \mid x_j = \max_{i=1, \dots, k} x_i\}$

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



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

**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  $\epsilon$ -suboptimal point
- an 'optimal' 1st-order method:  $O(1/\epsilon^2)$  bound cannot be improved

# References

-  S. Boyd, lecture notes and slides for EE364b, Convex Optimization II
-  B. T. Polyak, *Introduction to Optimization* (1987), section 1.4  $\mathcal{S}$  3.2.1 with the example on page 16 of this lecture