次梯度法

王 尧

西安交通大学智能决策与机器学习中心
(Email: yao.s.wang@gmail.com)

2022. 5

回顾：梯度下降

Consider the problem

$$\min_x \ f(x)$$

for f convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$. Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Step sizes t_k chosen to be fixed and small, or by backtracking line search

If ∇f is Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$. Downsides:

- Requires f differentiable
- Can be slow to converge 注：强凸性可以显著提高收敛速率

Lasso的光滑化求解

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

- LASSO 问题的目标函数$f(x)$ 不光滑，在某些点处无法求出梯度，因此不能直接对原始问题使用梯度法求解
- 不光滑项为$\|x\|_1$，它实际上是x 各个分量绝对值的和，考虑如下一维光滑函数：

$$l_\delta(x) = \begin{cases} \frac{1}{2\delta}x^2, & |x| < \delta, \\ |x| - \frac{\delta}{2}, & \text{其他}. \end{cases}$$

Lasso的光滑化求解

光滑化LASSO 问题为

$$\min \quad f_\delta(x) = \frac{1}{2}\|Ax-b\|^2 + \mu L_\delta(x), \quad \text{其中} \quad L_\delta(x) = \sum_{i=1}^{n} l_\delta(x_i),$$

δ 为给定的光滑化参数.

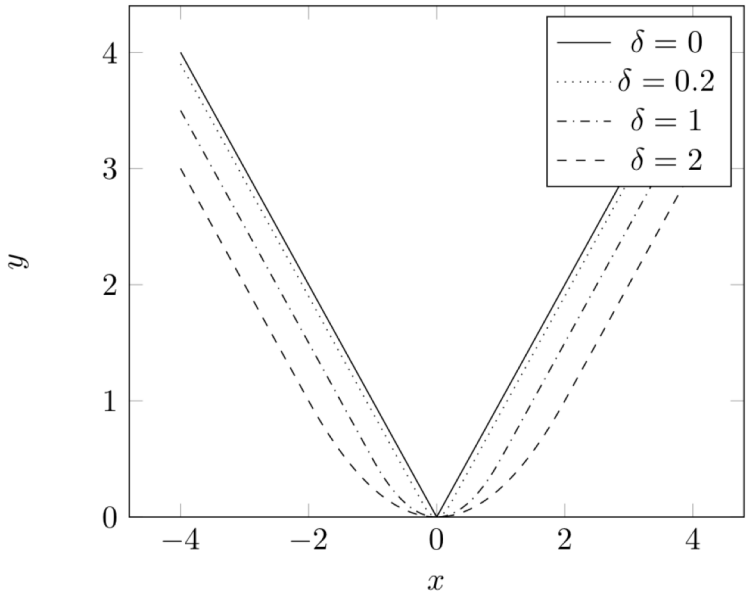


Figure: 当δ 取不同值时$l_\delta(x)$ 的图形

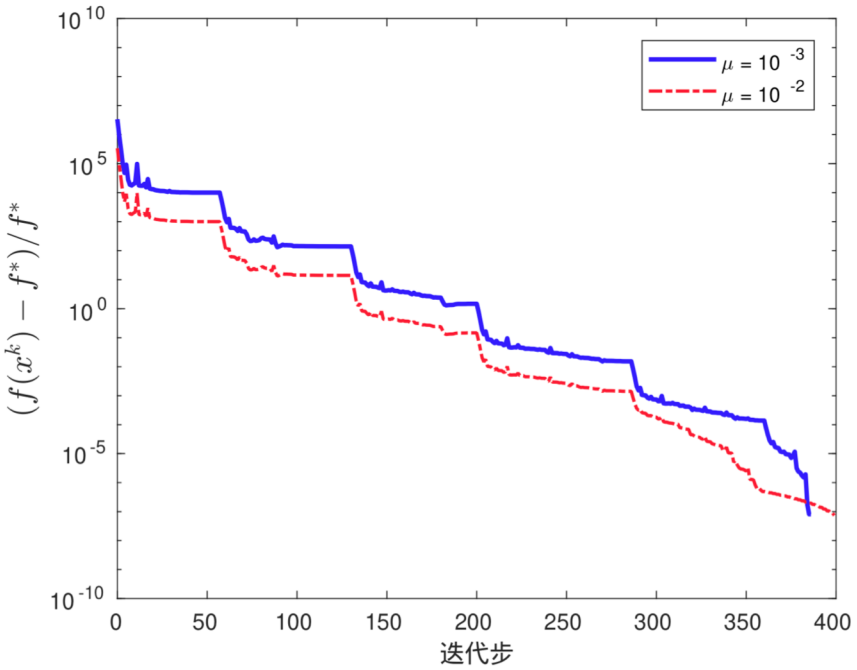
Lasso的光滑化求解

- $f_\delta(x)$ 的梯度为

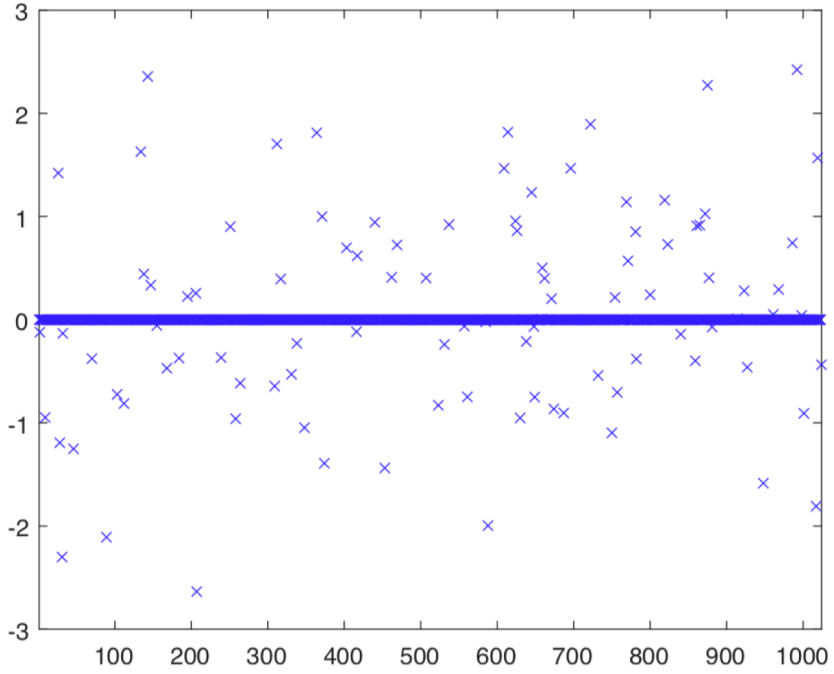
$$\nabla f_\delta(x) = A^{\mathrm{T}}(Ax - b) + \mu \nabla L_\delta(x),$$

其中$\nabla L_\delta(x)$ 是逐个分量定义的：

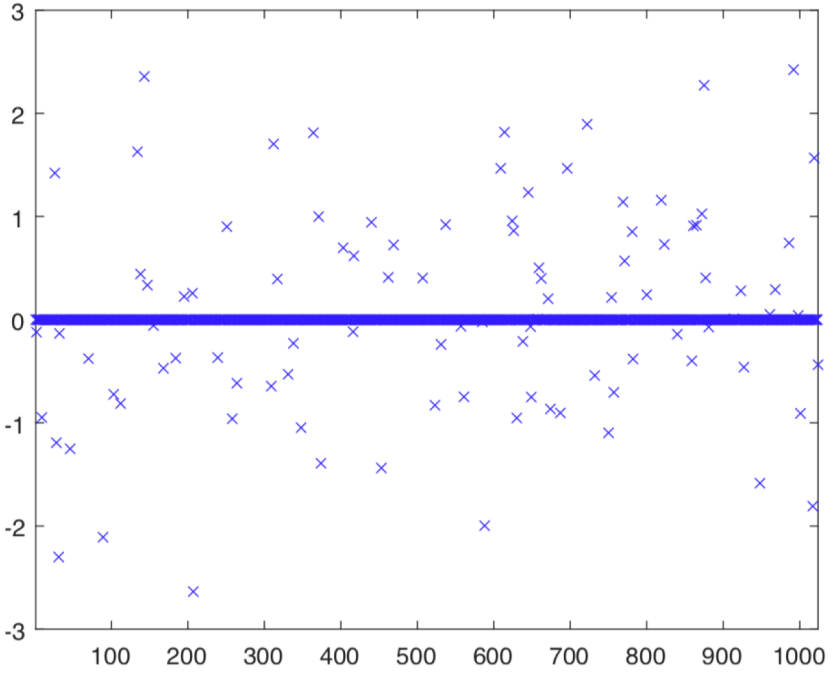
$$(\nabla L_\delta(x))_i = \begin{cases} \mathrm{sign}(x_i), & |x_i| > \delta, \\ \frac{x_i}{\delta}, & |x_i| \le \delta. \end{cases}$$



Lasso的光滑化求解



(a) 精确解



(b) 梯度法解

更多细节参见文再文版6.2.3

次梯度的定义

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

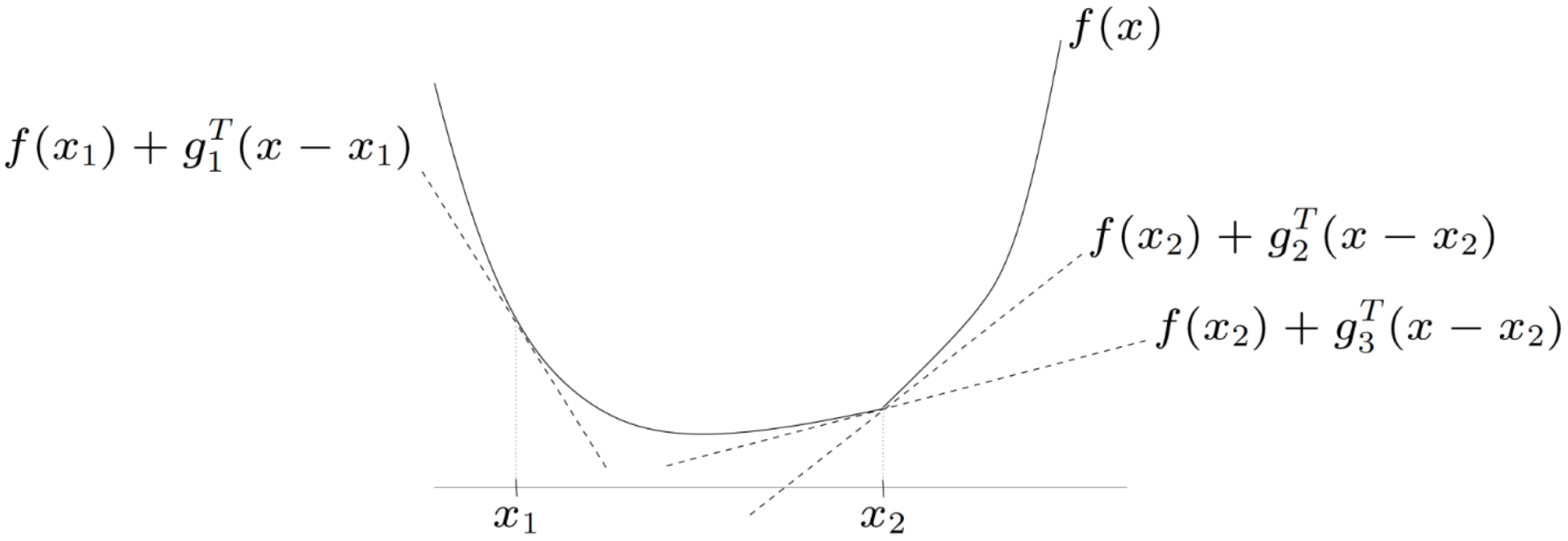
That is, linear approximation always underestimates f

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

- Always exists
- If f differentiable at x, then $g = \nabla f(x)$ uniquely

次梯度的定义



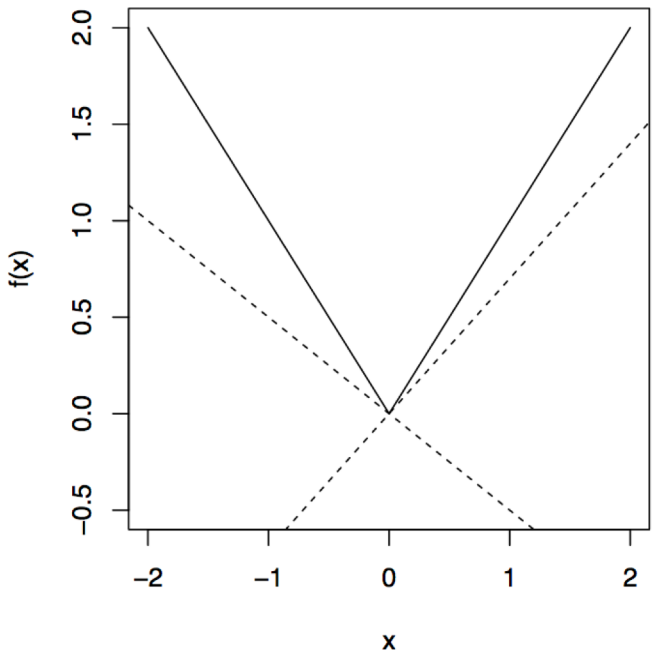
g_2, g_3 are subgradients at x_2; g_1 is a subgradient at x_1

the **subdifferential** $\partial f(x)$ of f at x is the set of all subgradients:

$$\partial f(x) = \{g | g^\top (y - x) \leq f(y) - f(x)\} \quad \forall\, y \in \mathbf{dom}\, f$$

实例I

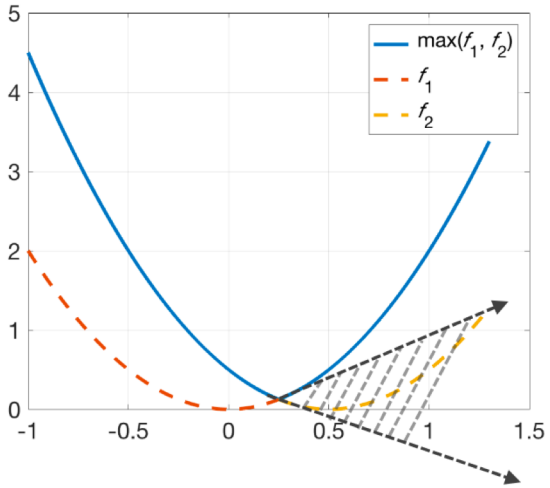
Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$



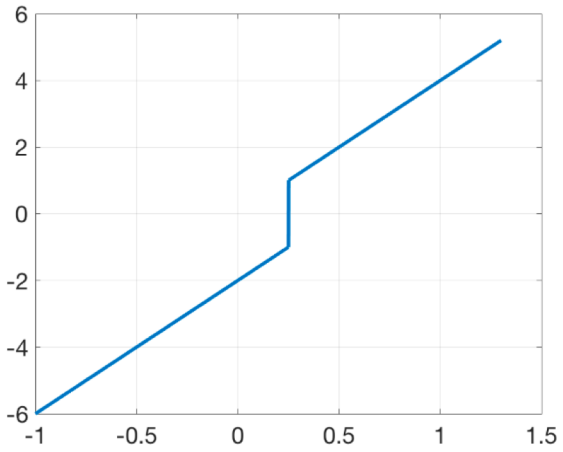
- For $x \neq 0$, unique subgradient $g = \text{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

实例11

$f(x) = \max\{f_1(x), f_2(x)\}$



$\partial f(x)$



$f(x) = \max\{f_1(x), f_2(x)\}$ where f_1 and f_2 are differentiable

$$\partial f(x) = \begin{cases} \{f_1'(x)\}, & \text{if } f_1(x) > f_2(x) \\ [f_1'(x), f_2'(x)], & \text{if } f_1(x) = f_2(x) \\ \{f_2'(x)\}, & \text{if } f_1(x) < f_2(x) \end{cases}$$

范数在零点的次梯度

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for any norm $\|\cdot\|$, then for any $\boldsymbol{g}$ obeying $\|\boldsymbol{g}\|_* \leq 1$,

$$\boldsymbol{g} \in \partial f(\mathbf{0})$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ (i.e. $\|\boldsymbol{x}\|_* := \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|\leq 1} \langle \boldsymbol{z}, \boldsymbol{x} \rangle$)

Proof: To see this, it suffices to prove that

$$f(\boldsymbol{z}) \geq f(\mathbf{0}) + \langle \boldsymbol{g}, \boldsymbol{z} - \mathbf{0} \rangle, \qquad \forall \boldsymbol{z}$$

$$\iff \quad \langle \boldsymbol{g}, \boldsymbol{z} \rangle \leq \|\boldsymbol{z}\|, \qquad \forall \boldsymbol{z}$$

This follows from generalized Cauchy-Schwarz, i.e.

$$\langle \boldsymbol{g}, \boldsymbol{z} \rangle \leq \|\boldsymbol{g}\|_* \|\boldsymbol{z}\| \leq \|\boldsymbol{z}\|$$

次微分的凸性

证明：

- 设$g_1, g_2 \in \partial f(x)$，并设$\lambda \in (0,1)$，由次梯度的定义

$$f(y) \geq f(x) + g_1^{\mathrm{T}}(y-x), \quad \forall y \in \mathrm{dom} f,$$
$$f(y) \geq f(x) + g_2^{\mathrm{T}}(y-x), \quad \forall y \in \mathrm{dom} f.$$

  由上面第一式的λ倍加上第二式的$(1-\lambda)$倍，我们可以得到$\lambda g_1 + (1-\lambda) g_2 \in \partial f(x)$，从而$\partial f(x)$是凸集.

次微分的单调性

the subdifferential of a convex function is a *monotone operator:*

$$(u - v)^T (x - y) \geq 0 \quad \text{for all } x, y, u \in \partial f(x), v \in \partial f(y)$$

Proof: by definition

$$f(y) \geq f(x) + u^T (y - x), \qquad f(x) \geq f(y) + v^T (x - y)$$

combining the two inequalities shows monotonicity

次梯度的计算

弱次梯度计算：得到一个次梯度

- 足以满足大多数不可微凸函数优化算法
- 如果可以获得任意一点处$f(x)$的值，那么总可以计算一个次梯度

强次梯度计算：得到$\partial f(x)$，即所有次梯度

- 一些算法、最优性条件等，需要完整的次微分
- 计算可能相当复杂

基本计算规则

- **scaling:** $\partial(\alpha f) = \alpha \partial f$ (for $\alpha > 0$)
- **summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- **affine transformation:** if $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, then

$$\partial h(\boldsymbol{x}) = \boldsymbol{A}^\top \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

- **chain rule:** suppose f is convex, and g is differentiable, *nondecreasing*, and *convex*. Let $h = g \circ f$, then

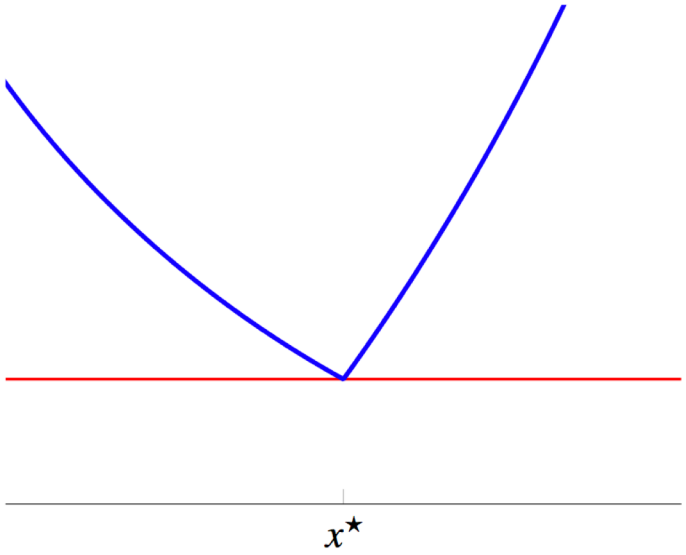
$$\partial h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\partial f(\boldsymbol{x})$$

有兴趣的同学可完成上述规则的证明(选做)

无约束优化的最优性条件

$x^\star$ minimizes $f(x)$ if and only

$$0 \in \partial f(x^\star)$$



this follows directly from the definition of subgradient:

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) \quad \text{for all } y \qquad \Longleftrightarrow \qquad 0 \in \partial f(x^\star)$$

Lasso的最优性条件

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, lasso problem can be parametrized as

$$\min_{\beta} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality:

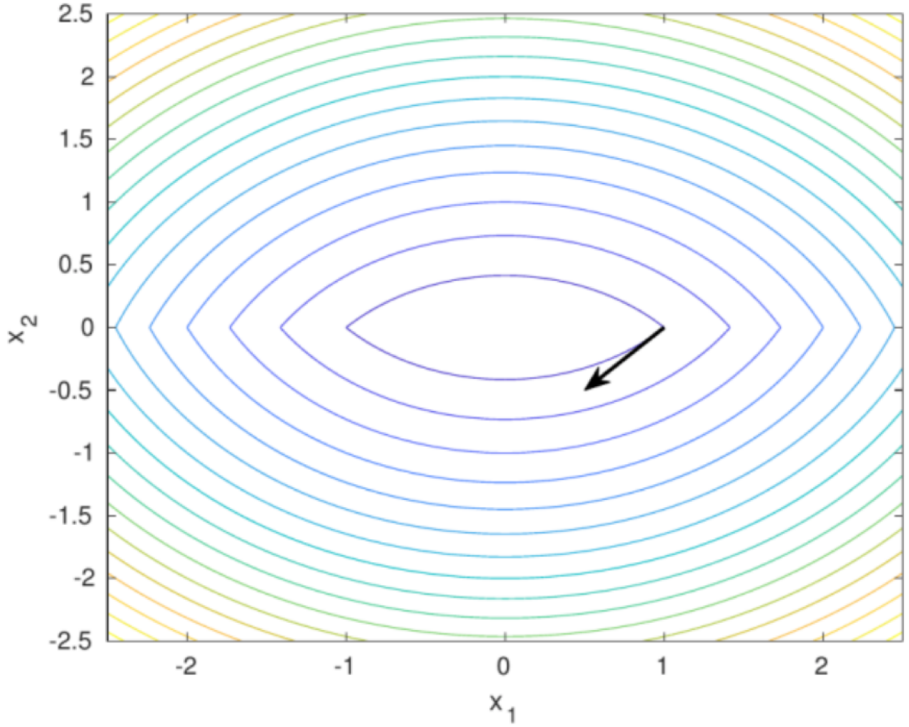
$$\begin{aligned} 0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) & \\ \iff \ 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1 & \\ \iff \ X^T(y - X\beta) = \lambda v & \end{aligned}$$

for some $v \in \partial\|\beta\|_1$, i.e.,

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots, p$$

次梯度与下降方向

- the negative gradient of a differentiable f is a descent direction (if $\nabla f(x) \neq 0$)
- negative subgradient is **not** always a descent direction



$$f(\boldsymbol{x}) = \max[x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2]$$

次梯度下降算法

Now consider f convex, having $\mathrm{dom}(f) = \mathbb{R}^n$, but not necessarily differentiable

Subgradient method: like gradient descent, but replacing gradients with subgradients. Initialize $x^{(0)}$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot g^{(k-1)}, \quad k = 1, 2, 3, \ldots$$

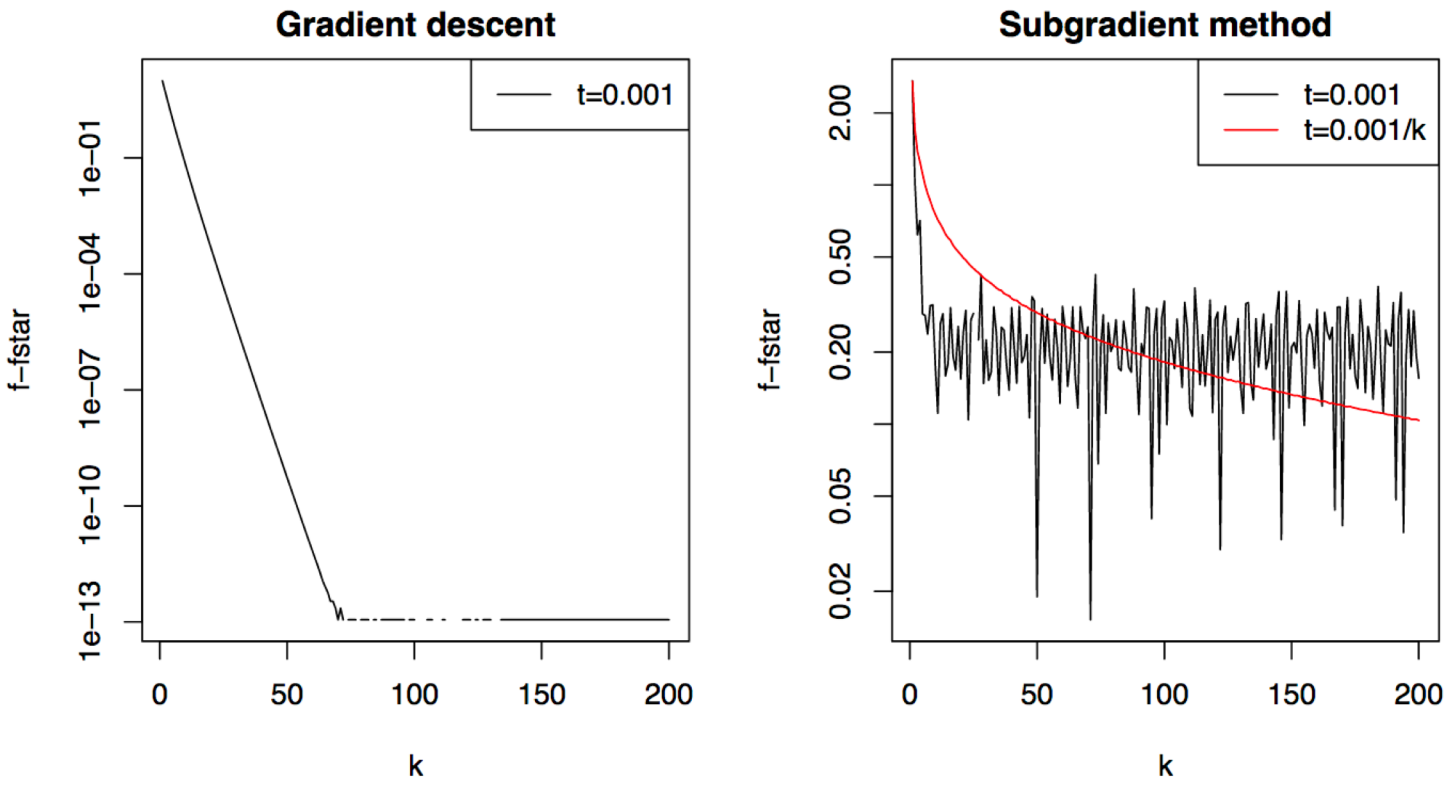
where $g^{(k-1)} \in \partial f(x^{(k-1)})$, any subgradient of f at $x^{(k-1)}$

Subgradient method is not necessarily a descent method, thus we keep track of best iterate $x^{(k)}_{\text{best}}$ among $x^{(0)}, \ldots, x^{(k)}$ so far, i.e.,

$$f(x^{(k)}_{\text{best}}) = \min_{i=0,\ldots,k} f(x^{(i)})$$

对比

Ridge: use gradients; lasso: use subgradients. Example here has $n = 1000$, $p = 20$:



Thank you for your attentions !