次梯度和次梯度算法

文再文

北京大学北京国际数学研究中心

教材《最优化：建模、算法与理论》配套电子教案

http://bicmr.pku.edu.cn/~wenzw/optbook.html

致谢：本教案由朱桢源协助准备

提纲

次梯度

- 可微凸函数f 的一阶条件:

$$f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x)$$

- 设f 为适当凸函数，$x \in \mathbf{dom}\, f$. 若向量$g \in \mathbb{R}^n$ 满足

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

则称g 为函数f 在点x 处的一个**次梯度**.

- 进一步地，称集合

$$\partial f(x) = \{g \,|\, g \in \mathbb{R}^n, f(y) \geq f(x) + g^{\mathrm{T}}(y - x), \forall y \in \mathbf{dom}\, f\}$$

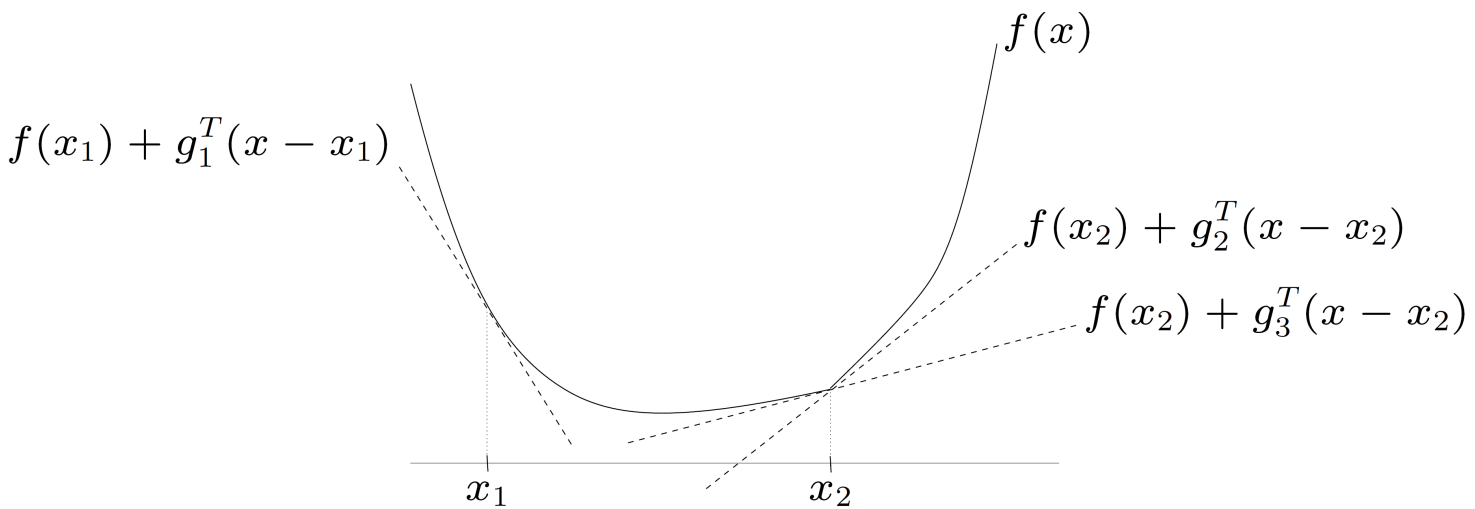
为f 在点x 处的**次微分**.

次梯度

- $f(x)+g^{\mathrm{T}}(y-x)$ 是$f(y)$ 的一个全局下界
- g 可以诱导出上方图**epi** f 在点$(x,f(x))$ 处的一个支撑超平面

$$\begin{bmatrix} g \\ -1 \end{bmatrix}\left(\begin{bmatrix} y \\ t \end{bmatrix}-\begin{bmatrix} x \\ f(x) \end{bmatrix}\right)\leq 0 \quad \forall\,(y,t)\in \mathbf{epi}\, f$$

- 如果f 是可微凸函数, 那么$\nabla f(x)$ 是f 在点x 处的一个次梯度
- 例：g_2, g_3 是点x_2 处的次梯度；g_1 是点x_1 处的次梯度



次梯度存在性

设f为凸函数，$\mathbf{dom}\, f$为其定义域．如果$x \in \mathbf{int}\, \mathbf{dom}\, f$，则$\partial f(x)$是非空的，其中$\mathbf{int}\, \mathbf{dom}\, f$的含义是集合$\mathbf{dom}\, f$的所有内点.

证明：

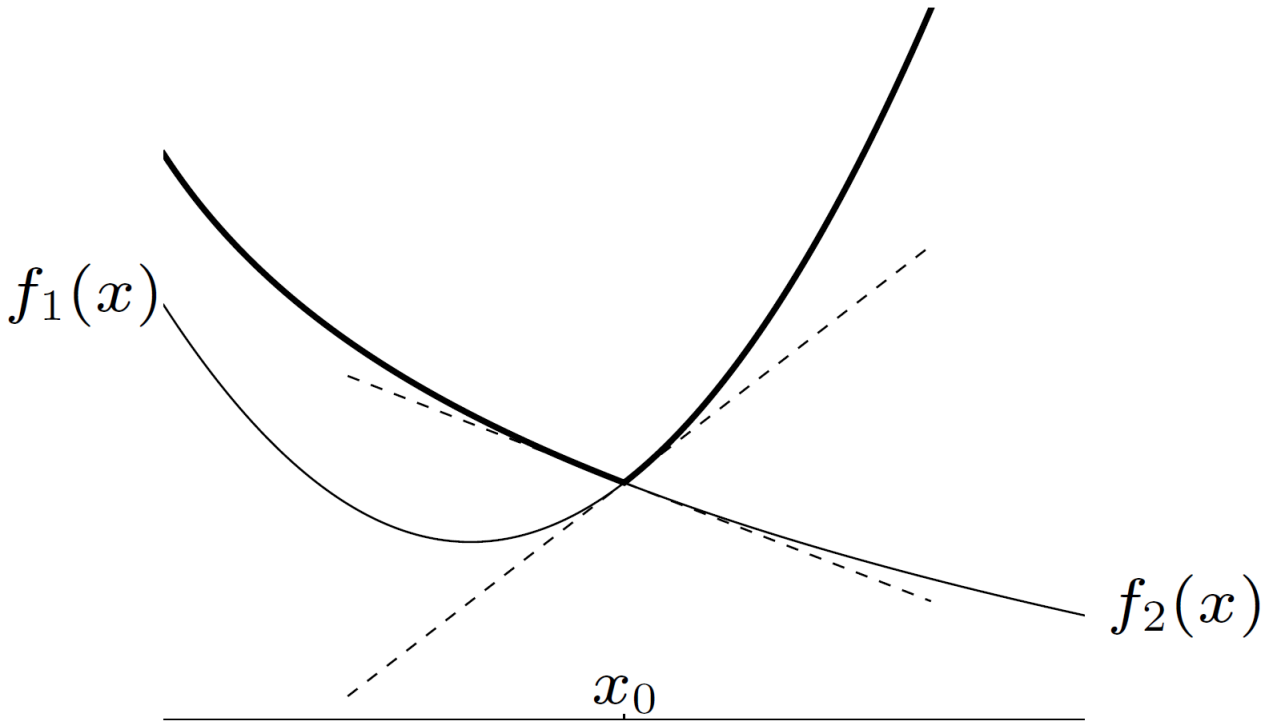
- $(x, f(x))$ 是$\mathbf{epi}\, f$ 边界上的点
- 因此存在$\mathbf{epi}\, f$ 在点$(x, f(x))$ 处的支撑超平面:

$$\exists\, (a,b) \neq 0, \quad \begin{bmatrix} a \\ b \end{bmatrix}^{\mathrm{T}} \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall\, (y,t) \in \mathbf{epi}\, f$$

- 令$t \to +\infty$, 可知$b \leq 0$
- 取$y = x + \epsilon a \in \mathbf{dom}\, f$, $\epsilon > 0$, 可知$b \neq 0$
- 因此$b < 0$ 并且$g = a/|b|$ 是f 在点x 处的次梯度

例

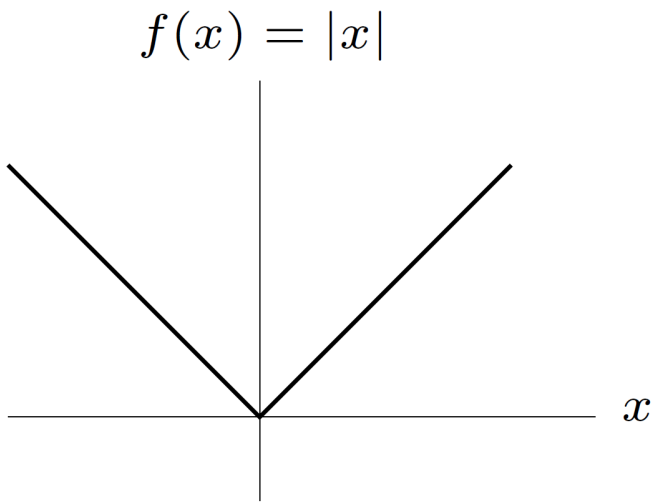
$$f(x) = \max\{f_1(x), f_2(x)\} \quad f_1, f_2 \text{ 是可微凸函数}$$

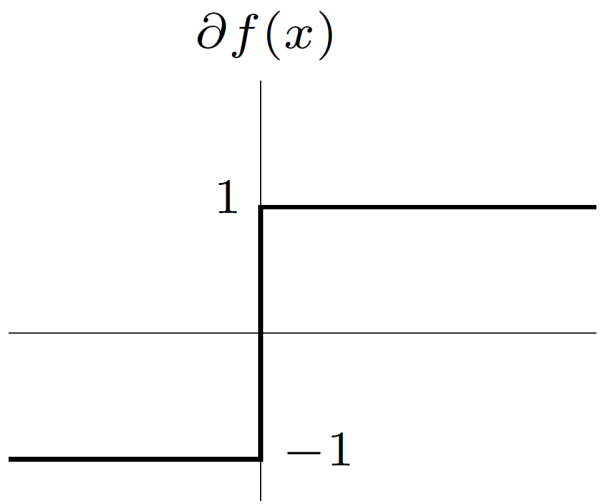


- 点x_0 处的次梯度可取范围$[\nabla f_1(x_0), \nabla f_2(x_0)]$
- 如果$f_1(\hat{x}) > f_2(\hat{x})$, f 在点$\hat{x}$ 处的次梯度等于$\nabla f_1(\hat{x})$
- 如果$f_1(\hat{x}) < f_2(\hat{x})$, f 在点$\hat{x}$ 处的次梯度等于$\nabla f_2(\hat{x})$

例

- 绝对值函数 $f(x) = |x|$





- 欧几里得范数 $f(x) = \|x\|_2$

$$\text{如果 } x \neq 0, \partial f(x) = \frac{1}{\|x\|_2} x, \quad \text{如果 } x = 0, \partial f(x) = \{g | \|g\|_2 \leq 1\}$$

提纲

次微分是闭凸集

对任何$x \in \mathbf{dom}\, f$，$\partial f(x)$ 是一个闭凸集（可能为空集）.

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.

- 令$g_k \in \partial f(x)$为次梯度且$g_k \to g$，则

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

  在上述不等式中取极限，并注意到极限的保号性，最终我们有

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

  这说明$\partial f(x)$为闭集.

内点的次微分非空有界

如果$x \in \mathbf{int}\,\mathbf{dom}\,f$，则$\partial f(x)$ 非空有界集.

证明：

- 非空可由次梯度存在性直接得出
- 取充分小的$r > 0$, 使得

$$B = \{x \pm re_i | i = 1, \cdots, n\} \subset \mathbf{dom}\,f$$

- 对任意非零的$g \in \partial f(x)$, 存在$y \in B$ 满足

$$f(y) \geq f(x) + g^{\mathrm{T}}(y - x) = f(x) + r\|g\|_\infty$$

- 由此得到$\partial f(x)$ 有界:

$$\|g\|_\infty \leq \frac{\max_{y \in B} f(y) - f(x)}{r} < +\infty$$

可微函数的次微分

设凸函数$f(x)$在$x_0 \in \mathbf{int}\,\mathbf{dom}\,f$处可微，则$\partial f(x_0) = \{\nabla f(x_0)\}$.

证明：

- 根据可微凸函数的一阶条件可知梯度$\nabla f(x_0)$为次梯度.
- 下证$f(x)$在点x_0处不可能有其他次梯度. 设$g \in \partial f(x_0)$，根据次梯度的定义，对任意的非零$v \in \mathbb{R}^n$且$x_0 + tv \in \mathbf{dom}\,f, t > 0$有

$$f(x_0 + tv) \geq f(x_0) + tg^{\mathrm{T}}v.$$

若$g \neq \nabla f(x_0)$，取$v = g - \nabla f(x_0) \neq 0$，上式变形为

$$\frac{f(x_0 + tv) - f(x_0) - t\nabla f(x_0)^{\mathrm{T}}v}{t\|v\|} \geq \frac{(g - \nabla f(x_0))^{\mathrm{T}}v}{\|v\|} = \|v\|.$$

- 不等式两边令$t \to 0$，根据Fréchet 可微的定义，左边趋于0，而右边是非零正数，可得到矛盾.

次梯度的单调性

设$f : \mathbb{R}^n \to \mathbb{R}$为凸函数，$x, y \in \mathbf{dom}\, f$，则$(u-v)^{\mathrm{T}}(x-y) \geq 0$, 其中$u \in \partial f(x)$，$v \in \partial f(y)$.

证明：

- 由次梯度的定义，

$$f(y) \geq f(x) + u^{\mathrm{T}}(y-x),$$
$$f(x) \geq f(y) + v^{\mathrm{T}}(x-y).$$

- 将以上两个不等式相加即得结论.

次梯度的连续性

设$f(x)$是闭凸函数且∂f在点$\bar{x}$附近存在且非空. 若序列$x^k \to \bar{x}$, $g^k \in \partial f(x^k)$为$f(x)$在点x^k处的次梯度, 且$g^k \to \bar{g}$, 则$\bar{g} \in \partial f(\bar{x})$.

证明:

- 对任意$y \in \mathrm{dom} f$, 根据次梯度的定义,

$$f(y) \geq f(x^k) + \left\langle g^k, y - x^k \right\rangle.$$

- 对上述不等式两边取下极限, 我们有

$$\begin{aligned} f(y) &\geq \liminf_{k\to\infty} [f(x^k) + \left\langle g^k, y - x^k \right\rangle] \\ &\geq f(\bar{x}) + \left\langle \bar{g}, y - \bar{x} \right\rangle, \end{aligned}$$

其中第二个不等式利用了$f(x)$的下半连续性以及$g^k \to \bar{g}$, 由此可推出$\bar{g} \in \partial f(\bar{x})$.

例：非次可微函数

如下函数在点$x=0$处不是次可微的:

- $f:\mathbf{R}\to\mathbf{R}, \mathbf{dom}\, f=\mathbf{R}_+$

$$x=0 \text{ 时}, f(x)=1, x>0 \text{ 时}, f(x)=0$$

- $f:\mathbf{R}\to\mathbf{R}, \mathbf{dom}\, f=\mathbf{R}_+$

$$f(x)=-\sqrt{x}$$

$\mathbf{epi}\, f$ 在点$(0,f(0))$处的唯一支撑超平面是垂直的

方向导数

- 一般函数：设f为适当函数，给定点x_0以及方向$d \in \mathbb{R}^n$，方向导数（若存在）定义为

$$\lim_{t \downarrow 0} \phi(t) = \lim_{t \downarrow 0} \frac{f(x_0 + td) - f(x_0)}{t},$$

其中$t \downarrow 0$表示t单调下降趋于0.

- 凸函数：易知$\phi(t)$ 在$(0, +\infty)$上是单调不减的，上式中的极限号$\lim$可以替换为下确界$\inf$. 上述此时极限总是存在（可以为无穷），进而凸函数总是可以定义方向导数.

- 方向导数的定义：对于凸函数f，给定点$x_0 \in \mathbf{dom}\, f$ 以及方向$d \in \mathbb{R}^n$，其方向导数定义为

$$\partial f(x_0; d) = \inf_{t > 0} \frac{f(x_0 + td) - f(x_0)}{t}.$$

方向导数有限

设$f(x)$为凸函数，$x_0 \in \mathbf{int\ dom}\, f$，则对任意$d \in \mathbb{R}^n$，$\partial f(x_0; d)$有限.

证明：

- 首先$\partial f(x_0; d)$不为正无穷是显然的.
- 由于$x_0 \in \mathbf{int\ dom}\, f$，根据次梯度的存在性定理可知$f(x)$在点$x_0$处存在次梯度$g$.
- 根据方向导数的定义，我们有

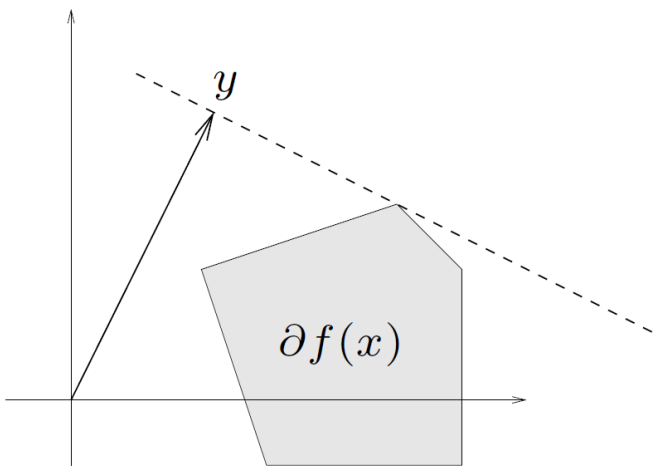
$$\begin{aligned}\partial f(x_0; d) &= \inf_{t>0} \frac{f(x_0 + td) - f(x_0)}{t} \\ &\geq \inf_{t>0} \frac{tg^{\mathrm{T}}d}{t} = g^{\mathrm{T}}d.\end{aligned}$$

  其中的不等式利用了次梯度的定义.
- 这说明$\partial f(x_0; d)$不为负无穷.

方向导数和次梯度

设$f:\mathbb{R}^n \to (-\infty,+\infty]$为凸函数，点$x_0 \in \mathbf{int\ dom}\, f$，$d$为$\mathbb{R}^n$中任一方向，则

$$\partial f(x_0;d) = \max_{g\in\partial f(x_0)} g^{\mathrm{T}}d.$$



$\partial f(x;y)$是$\partial f(x)$的支撑函数

- 对于可微函数, $\partial f(x_0;d) = \nabla f(x_0)^{\mathrm{T}}d$
- 这也说明$\partial f(x_0;d)$对所有的$x_0 \in \mathbf{int\ dom}\, f$, 以及所有的$d$都存在

提纲

次梯度的计算规则

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

下面我们假设$x \in \mathbf{int}\ \mathbf{dom}\ f$

基本规则

- **可微凸函数：** 若凸函数f在点x处可微，则$\partial f(x) = \{\nabla f(x)\}$.
- **凸函数的非负线性组合：** 设凸函数f_1, f_2满足$\mathbf{int\ dom}\, f_1 \cap \mathbf{dom}\, f_2 \neq \emptyset$，而$x \in \mathbf{dom}\, f_1 \cap \mathbf{dom}\, f_2$. 若

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x), \quad \alpha_1, \alpha_2 \geq 0,$$

则$f(x)$的次微分

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x).$$

- **线性变量替换：** 设h为适当凸函数，f满足$f(x) = h(Ax+b)$. 若存在$x^\sharp \in \mathbb{R}^m$，使得$Ax^\sharp + b \in \mathbf{int\ dom}\, h$，则

$$\partial f(x) = A^{\mathrm{T}} \partial h(Ax+b), \quad \forall\, x \in \mathbf{int\ dom}\, f.$$

两个函数之和的次梯度

设$f_1, f_2 : \mathbb{R}^n \to (-\infty, +\infty]$是两个凸函数，则对任意的$x_0 \in \mathbb{R}^n$，

$$\partial f_1(x_0) + \partial f_2(x_0) \subseteq \partial(f_1 + f_2)(x_0).$$

进一步地，若$\mathbf{int}\ \mathbf{dom}\, f_1 \cap \mathbf{dom}\, f_2 \neq \emptyset$，则对任意的$x_0 \in \mathbb{R}^n$，

$$\partial(f_1 + f_2)(x_0) = \partial f_1(x_0) + \partial f_2(x_0).$$

证明:

- 第一个结论由次梯度的定义是显然的
- 第二个结论证明参考教材

函数族的上确界

设$f_1, f_2, \cdots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ 均为凸函数，令

$$f(x) = \max\{f_1(x), f_2(x), \cdots, f_m(x)\}, \quad \forall x \in \mathbb{R}^n.$$

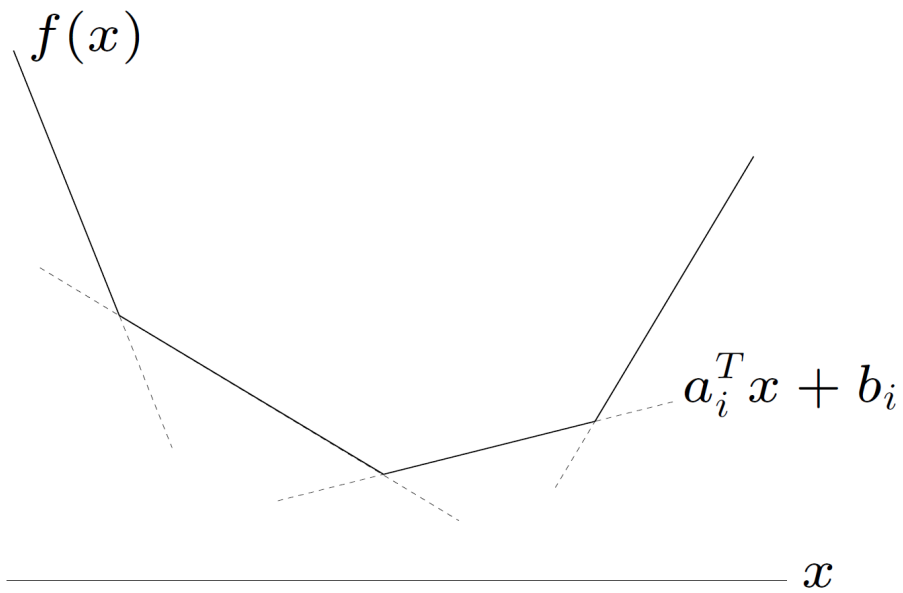
对$x_0 \in \bigcap_{i=1}^{m} \mathbf{int}\,\mathbf{dom}\, f_i$，定义$I(x_0) = \{i \mid f_i(x_0) = f(x_0)\}$，则

$$\partial f(x_0) = \mathbf{conv} \bigcup_{i \in I(x_0)} \partial f_i(x_0).$$

- $I(x_0)$表示点x_0 处“有效”函数的指标
- $\partial f(x_0)$ 是点x_0 处“有效”函数的次微分并集的凸包
- 如果f_i 可微, $\partial f(x_0) = \mathbf{conv}\{\nabla f_i(x_0) \mid i \in I(x_0)\}$

例：分段线性函数

$$f(x) = \max_{i=1,2,\cdots,m} \{a_i^{\mathrm{T}} x + b_i\}$$



- 点x处的次微分是一个多面体

$$\partial f(x) = \mathbf{conv}\{a_i \mid i \in I(x)\}$$

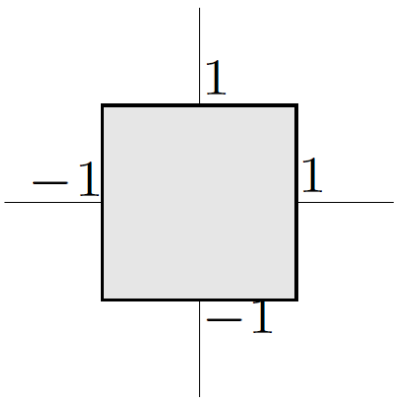
其中$I(x) = \{i \mid a_i^{\mathrm{T}} x + b_i = f(x)\}$

例: ℓ_1-范数

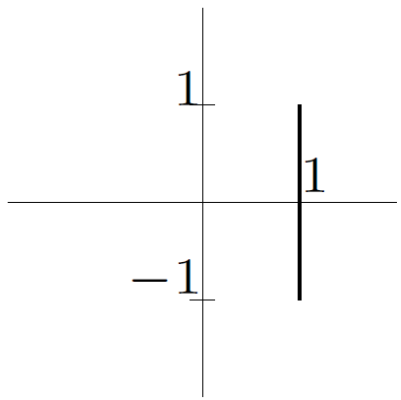
$$f(x) = \|x\|_1 = \max_{s \in \{-1,1\}^n} s^{\mathrm{T}} x$$

- 次微分是区间的乘积

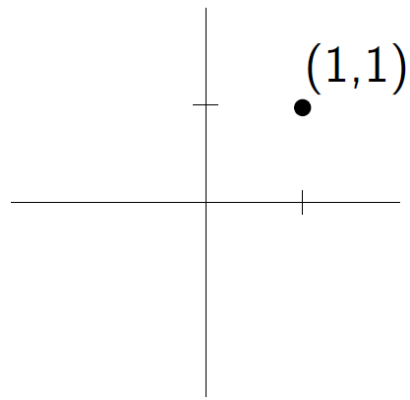
$$\partial f(x) = J_1 \times \cdots \times J_n, \quad J_k = \begin{cases} [-1,1], & x_k = 0 \\ \{1\}, & x_k > 0 \\ \{-1\}, & x_k < 0 \end{cases}$$



$\partial f(0,0) = [-1,1] \times [-1,1]$



$\partial f(1,0) = \{1\} \times [-1,1]$



$\partial f(1,1) = \{(1,1)\}$

逐点上确界函数

设$\{f_\alpha \mid \mathbb{R}^n \to (-\infty, +\infty]\}_{\alpha\in\mathcal{A}}$是一族凸函数，令

$$f(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x).$$

- 对$x_0 \in \cap_{\alpha\in\mathcal{A}}\mathbf{int\ dom}\, f_\alpha$，定义$I(x_0) = \{\alpha \in \mathcal{A} \mid f_\alpha(x_0) = f(x_0)\}$，则

$$\mathbf{conv} \bigcup_{\alpha\in I(x_0)} \partial f_\alpha(x_0) \subseteq \partial f(x_0).$$

- 如果还有$\mathcal{A}$是紧集且f_α关于α连续，则

$$\mathbf{conv} \bigcup_{\alpha\in I(x_0)} \partial f_\alpha(x_0) = \partial f(x_0).$$

例：最大特征值函数

$A(x) = A_0 + x_1A_1 + \cdots + x_nA_n$ 并且系数A_i 对称，令

$$f(x) = \lambda_{\max}(A(x)) = \sup_{\|y\|_2=1} y^{\mathrm{T}}A(x)y$$

计算点$\hat{x}$ 处的一个次梯度：

- 选择特征值$\lambda_{\max}(A(\hat{x}))$ 对应的任一单位特征向量y
- $y^{\mathrm{T}}A(x)y$ 在点$\hat{x}$ 处的梯度是f 的一个次梯度：

$$(y^{\mathrm{T}}A_1y, \cdots, y^{\mathrm{T}}A_ny) \in \partial f(\hat{x})$$

固定分量的函数极小值

$$f(x) = \inf_y h(x,y), \quad h \text{ 关于 } (x,y) \text{ 联合凸}$$

计算点$\hat{x}$处的一个次梯度：

- 设$\hat{y} \in \mathbb{R}^m$ 满足$h(\hat{x},\hat{y}) = f(\hat{x})$
- 存在$g \in \mathbb{R}^n$ 使得$(g,0) \in \partial h(\hat{x},\hat{y})$，则$g \in \partial f(\hat{x})$

证明：对任意$x \in \mathbb{R}^n, y \in \mathbb{R}^m$

$$\begin{aligned} h(x,y) &\geq h(\hat{x},\hat{y}) + g^{\mathrm{T}}(x-\hat{x}) + 0^{\mathrm{T}}(y-\hat{y}) \\ &= f(\hat{x}) + g^{\mathrm{T}}(x-\hat{x}) \end{aligned}$$

于是

$$f(x) = \inf_y h(x,y) \geq f(\hat{x}) + g^{\mathrm{T}}(x-\hat{x})$$

例: 距离函数

设C 是$\mathbb{R}^n$ 中一闭凸集，令

$$f(x) = \inf_{y \in C} \|x - y\|_2$$

计算点$\hat{x}$ 处的一个次梯度：

- 若$f(\hat{x}) = 0$，则容易验证$g = 0 \in \partial f(\hat{x})$；
- 若$f(\hat{x}) > 0$，取$\hat{y}$ 为$\hat{x}$ 在C 上的投影，即$\hat{y} = \mathcal{P}_c(\hat{x})$，计算

$$g = \frac{1}{\|\hat{x} - \hat{y}\|_2}(\hat{x} - \hat{y}) = \frac{1}{\|\hat{x} - \mathcal{P}_c(\hat{x})\|_2}(\hat{x} - \mathcal{P}_c(\hat{x}))$$

复合函数

设$f_1, f_2, \cdots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ 为m 个凸函数，$h : \mathbb{R}^m \to (-\infty, +\infty]$ 为关于各分量单调递增的凸函数，令

$$f(x) = h(f_1(x), f_2(x), \cdots, f_m(x)).$$

计算点$\hat{x}$ 处的一个次梯度:

- $z = (z_1, z_2, \cdots, z_m) \in \partial h(f_1(\hat{x}), f_2(\hat{x}), \cdots, f_m(\hat{x}))$ 以及$g_i \in \partial f_i(\hat{x})$
- $g \overset{\text{def}}{=\!=} z_1 g_1 + z_2 g_2 + \cdots + z_m g_m \in \partial f(\hat{x})$

证明:

$$\begin{aligned} f(x) &\geq h\left(f_1(\hat{x}) + g_1^{\mathrm{T}}(x - \hat{x}), f_2(\hat{x}) + g_2^{\mathrm{T}}(x - \hat{x}), \cdots, f_m(\hat{x}) + g_m^{\mathrm{T}}(x - \hat{x})\right) \\ &\geq h(f_1(\hat{x}), f_2(\hat{x}), \cdots, f_m(\hat{x})) + \sum_{i=1}^{m} z_i g_i^T (x - \hat{x}) \\ &= f(\hat{x}) + g^{\mathrm{T}}(x - \hat{x}), \end{aligned}$$

最优值函数

设函数f_i是凸函数，定义$h(u,v)$为如下凸问题的最优值

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le u_i,\ i = 1, \cdots, m \\ & Ax = b + v \end{aligned}$$

计算点$(\hat{u}, \hat{v})$处的一个次梯度:

- 假设$h(\hat{u}, \hat{v})$有限, 强对偶成立

$$\begin{aligned} \max \quad & \inf_x \left(f_0(x) + \sum_i \lambda_i (f_i(x) - \hat{u}_i) + w^{\mathrm{T}} (Ax - b - \hat{v}) \right) \\ \text{s.t.} \quad & \lambda \ge 0 \end{aligned}$$

- 如果$\hat{\lambda}, \hat{w}$是最优对偶变量，那么$(-\hat{\lambda}, -\hat{w}) \in \partial h(\hat{u}, \hat{v})$

证明:由弱对偶原理可得

$$
\begin{aligned}
h(u,v) &\geq \inf_x \left(f_0(x) + \sum_i \hat{\lambda}_i (f_i(x) - u_i) + \hat{w}^{\mathrm{T}}(Ax - b - v) \right) \\
&= \inf_x \left(f_0(x) + \sum_i \hat{\lambda}_i (f_i(x) - \hat{u}_i) + \hat{w}^{\mathrm{T}}(Ax - b - \hat{v}) \right) \\
&- \hat{\lambda}^{\mathrm{T}}(u - \hat{u}) - \hat{w}^{\mathrm{T}}(v - \hat{v}) \\
&= h(\hat{u}, \hat{v}) - \hat{\lambda}^{\mathrm{T}}(u - \hat{u}) - \hat{w}^{\mathrm{T}}(v - \hat{v})
\end{aligned}
$$

期望函数

u是一个随机变量，h是关于x的凸函数，令

$$f(x) = \mathbb{E}h(x,u)$$

计算点$\hat{x}$处的一个次梯度：

- 选择一个函数g满足$g(u) \in \partial_x h(\hat{x},u)$
- $g = \mathbb{E}_u\, g(u) \in \partial f(\hat{x})$

证明：由h的凸性和$g(u)$的定义，

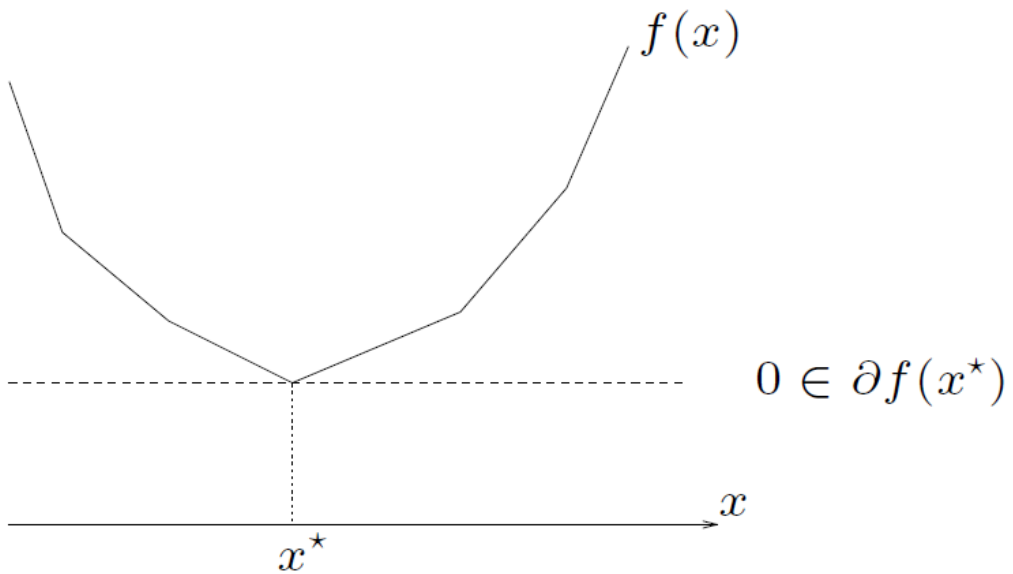
$$\begin{aligned} f(x) &= \mathbb{E}h(x,u) \\ &\geq \mathbb{E}\left(h(\hat{x},u) + g(u)^{\mathrm{T}}(x-\hat{x})\right) \\ &= f(\hat{x}) + g^{\mathrm{T}}(x-\hat{x}) \end{aligned}$$

提纲

最优性条件：无约束问题

x^* 是$f(x)$ 的极小点当且仅当

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



证明：根据定义

$$f(y) \geq f(x^*) + 0^{\mathrm{T}}(y - x^*),\ \forall y \quad \Leftrightarrow \quad 0 \in \partial f(x^*)$$

例: 分片线性极小

$$f(x) = \max_{i=1,\cdots,m} (a_i^{\mathrm{T}} x + b_i)$$

- **最优性条件**

$$0 \in \mathbf{conv}\{a_i \mid i \in I(x^*)\}, \quad \text{其中 } I(x) = \{i \mid a_i^{\mathrm{T}} x + b_i = f(x)\}$$

- 也就是说, x^* 是最优解当且仅当存在λ 使得

$$\lambda \geq 0, \quad \mathbf{1}^{\mathrm{T}}\lambda = 1, \quad \sum_{i=1}^{m} \lambda_i a_i = 0, \quad \lambda_i = 0 \text{ for } i \notin I(x^*)$$

- 这是等价线性规划问题的最优性条件: $A = [a_1^{\top}; \ldots; a_m^{\top}]$

$$\begin{array}{ll} \min & t \\ \text{s.t.} & Ax + b \leq t\mathbf{1} \end{array} \qquad \begin{array}{ll} \max & b^{\mathrm{T}}\lambda \\ \text{s.t.} & A^{\mathrm{T}}\lambda = 0 \\ & \lambda \geq 0, \quad \mathbf{1}^{\mathrm{T}}\lambda = 1 \end{array}$$

最优性条件：约束问题

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = 1, \cdots, m \end{aligned}$$

如果强对偶成立，那么x^*, λ^* 是最优原始、对偶变量当且仅当

1. x^* 是可行的
2. $\lambda^* \ge 0$
3. $\lambda_i^* f_i(x^*) = 0$, $i = 1, \cdots, m$
4. x^* 是下式的一个极小值点

$$x^* = \arg\min_x L(x, \lambda^*) = f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x)$$

即：

$$0 \in \partial L_x(x^*, \lambda^*) = \partial f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \partial f_i(x^*).$$

提纲

次梯度算法结构

为了极小化一个不可微的凸函数f，可类似梯度法构造如下次梯度算法的迭代格式：

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k),$$

其中$\alpha_k > 0$为步长. 它通常有如下四种选择：

1. 固定步长$\alpha_k = \alpha$；
2. 固定$\|x^{k+1} - x^k\|$，即$\alpha_k \|g^k\|$为常数；
3. 消失步长$\alpha_k \to 0$且$\sum_{k=0}^{\infty} \alpha_k = +\infty$；
4. 选取α_k使其满足某种线搜索准则.

下面我们讨论在不同步长取法下次梯度算法的收敛性质.

假设条件

(1) f为凸函数；

(2) f至少存在一个有限的极小值点x^*，且$f(x^*) > -\infty$；

(3) f为利普希茨连续的，即

$$|f(x) - f(y)| \leq G\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

其中$G > 0$为利普希茨常数.

我们下面证明这等价于$f(x)$的次梯度是有界的，即

$$\|g\| \leq G, \quad \forall\, g \in \partial f(x), x \in \mathbb{R}^n.$$

证明：

- 充分性：假设$\|g\|_2 \leq G, \forall\, g \in \partial f(x)$；取$g_y \in \partial f(y), g_x \in \partial f(x)$:

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

再由柯西不等式

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

- 必要性：反设存在x和$g \in \partial f(x)$，使得$\|g\|_2 > G$；取$y = x + \frac{g}{\|g\|_2}$

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

这与$f(x)$是G-利普希茨连续的矛盾.

收敛性分析

- 次梯度方法不是一个下降方法，即无法保证$f(x^{k+1}) < f(x^k)$；
- 收敛性分析的关键是分析$f(x)$历史迭代的最优点所满足的性质.
- 设x^*是$f(x)$的一个全局极小值点，$f^* = f(x^*)$，根据迭代格式，

$$\begin{aligned}\left\|x^{i+1} - x^*\right\|^2 &= \left\|x^i - \alpha_i g^i - x^*\right\|^2 \\ &= \left\|x^i - x^*\right\|^2 - 2\alpha_i \left\langle g^i, x^i - x^*\right\rangle + \alpha_i^2 \left\|g^i\right\|^2 \\ &\leqslant \left\|x^i - x^*\right\|^2 - 2\alpha_i \left(f\left(x^i\right) - f^*\right) + \alpha_i^2 G^2\end{aligned}$$

- 结合$i = 0, \cdots, k$ 时相应的不等式，并定义$\hat{f}^k = \min_{0 \leqslant i \leqslant k} f\left(x^i\right)$：

$$\begin{aligned}2\left(\sum_{i=0}^{k} \alpha_i\right)\left(\hat{f}^k - f^*\right) &\leqslant \left\|x^0 - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2 \\ &\leqslant \left\|x^0 - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2\end{aligned}$$

不同步长下的收敛性

(1) 取$\alpha_i = t$为固定步长，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时,$\hat{f}^k$ 近似为$G^2t/2$-次优的

(2) 取α_i使得$\|x^{i+1} - x^i\|$固定，即$\alpha_i\|g^i\| = s$为常数，则

$$\hat{f}^k - f^* \leq \frac{G\|x^0 - x^*\|^2}{2ks} + \frac{Gs}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时,$\hat{f}^k$ 近似为$Gs/2$-次优的

不同步长下的收敛性

(3) 取α_i为消失步长，即$\alpha_i \to 0$且$\sum_{i=0}^{\infty} \alpha_i = +\infty$，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i};$$

进一步可得$\hat{f}^k$收敛到f^*.

- 和梯度法不同，只有当α_k取消失步长时$\hat{f}^k$才具有收敛性.
- 一个常用的步长取法是$\alpha_k = \frac{1}{k}$.

固定迭代步数下的最优步长

- 假设$\left\|x^0 - x^*\right\| \leqslant R$，并且总迭代步数$k$是给定的，在固定步长下，

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2} \leq \frac{R^2}{2kt} + \frac{G^2 t}{2}.$$

- 由平均值不等式知当t满足$\frac{R^2}{2kt} = \frac{G^2 t}{2}$, 即 $t = \frac{R}{G\sqrt{k}}$时，右端达到最小.
- k 步后得到的上界是

$$\hat{f}^k - f^* \leqslant \frac{GR}{\sqrt{k}}$$

- 这表明在$k = O(1/\epsilon^2)$ 步迭代后可以得到$\hat{f}^k - f^* \leq \epsilon$ 的精度
- 类似地可证明第二类步长选取策略下，取$s = \frac{R}{\sqrt{k}}$，可得到估计

$$\hat{f}^k - f^* \leq \frac{GR}{\sqrt{k}}.$$

f^* 已知时的最优步长

- 第41页第一个不等式右端在

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

时取到极小.

- 这等价于

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

- 递归地利用上式并结合$\|x^0 - x^*\| \leq R$和$\|g^i\| \leq G$，可以得到

$$\hat{f}^k - f^* \leq \frac{GR}{\sqrt{k}}.$$

练习：计算交集中的一点

为了寻找一个落在m个闭凸集$C_1, ..., C_m$的交集中的点：

$$\min \quad f(x) = \max\{d_1(x), \cdots, d_m(x)\}$$

其中$d_j(x) = \inf_{y \in C_j} \|x - y\|_2$表示点$x$到集合$C_j$的欧几里得距离

- 如果交集非空，$f^* = 0$
- 如果$g \in \partial d_j(\hat{x})$并且$C_j$是距离$\hat{x}$最远的集合，那么$g \in \partial f(\hat{x})$
- 次梯度$g \in \partial d_j(\hat{x})$：

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

其中$P_j(\hat{x})$是到C_j的投影；注意当$\hat{x} \notin C_j$时，$\|g\|_2 = 1$

最优步长选取下的次梯度法：

- 当$f^* = 0$ 且$\|g^{i-1}\|_2 = 1$ 时，最优步长是$\alpha_i = f(x^{i-1})$.
- 在第k 步迭代, 找到距离最远的集合C_j (此时$f(x^{k-1}) = d_j(x^{k-1})$)；取

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

- 是一种交替投影算法
- 每步迭代都将当前迭代点投影到最远的集合上
- 当$m = 2$ 时, 交替投影到两个集合上

例：LASSO 问题求解

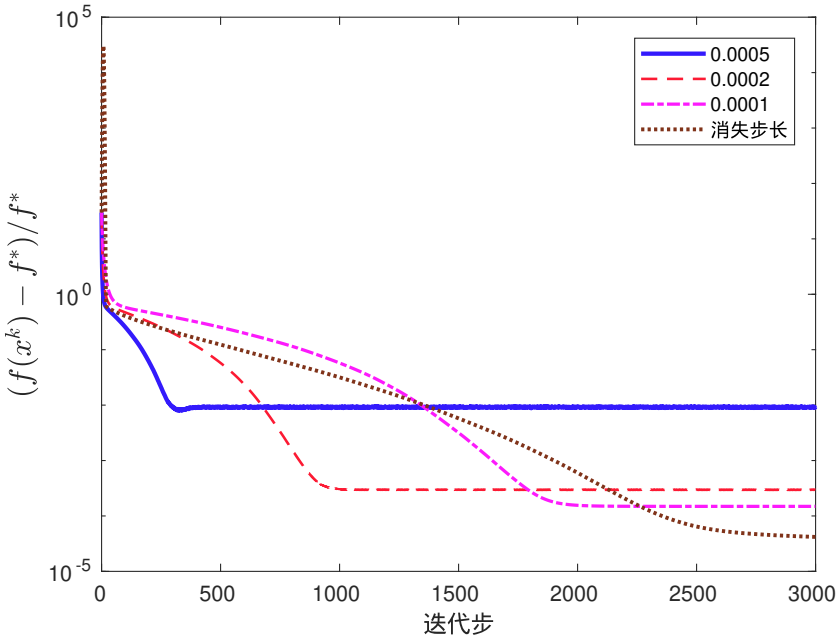
考虑LASSO 问题

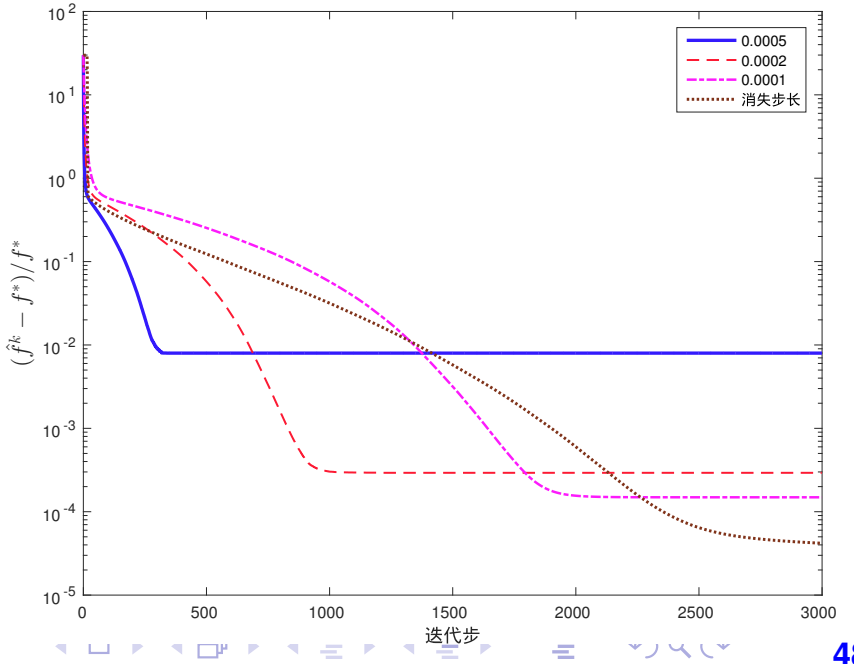
$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1,$$

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

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

步长α_k可选为固定步长或消失步长.





例：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$\frac{1}{\lambda_{\max}(A^{\mathrm{T}}A)}$；若$\mu_t = \mu$，则取步长

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

其中k为迭代步数

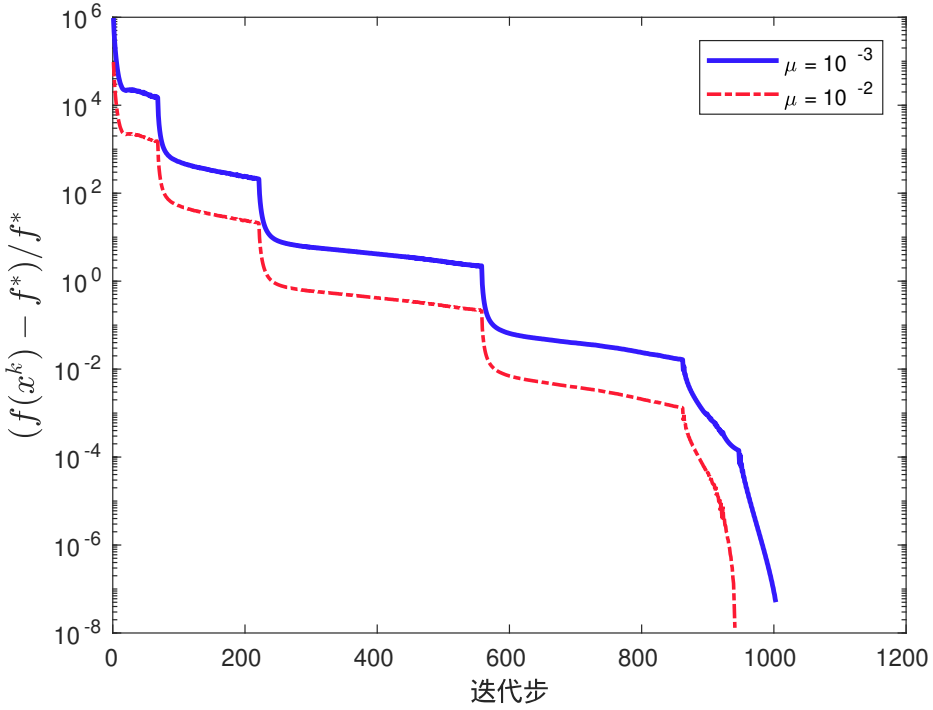


Figure: LASSO 问题在不同正则化参数下的求解结果