梯度下降法

- 注意到$\phi(\alpha)=f(x^k+\alpha d^k)$ 有泰勒展开

$$\phi(\alpha)=f(x^k)+\alpha\nabla f(x^k)^{\mathrm{T}}d^k+\mathcal{O}(\alpha^2\|d^k\|^2).$$

- 由柯西不等式，当α 足够小时取$d^k=-\nabla f(x^k)$ 会使函数下降最快.
- 因此梯度法就是选取$d^k=-\nabla f(x^k)$ 的算法，它的迭代格式为

$$x^{k+1}=x^k-\alpha_k\nabla f(x^k).$$

步长α_k 的选取可依赖于线搜索算法，也可直接选取固定的α_k.

- 另一种理解方式：

$$\begin{aligned} x^{k+1} &= \arg\min_x f(x^k)+\nabla f(x^k)^\top(x-x^k)+\frac{1}{\alpha_k}\|x-x^k\|_2^2 \\ &= \arg\min_x \|x-(x^k-\alpha_k\nabla f(x^k))\|_2^2 \\ &= x^k-\alpha_k\nabla f(x^k) \end{aligned}$$

设二次函数$f(x,y)=x^2+10y^2$, 初始点(x^0,y^0) 取为$(10,1)$, 取固定步长$\alpha_k=0.085$. 我们使用梯度法$x^{k+1}=x^k-\alpha_k\nabla f(x^k)$ 进行15 次迭代.

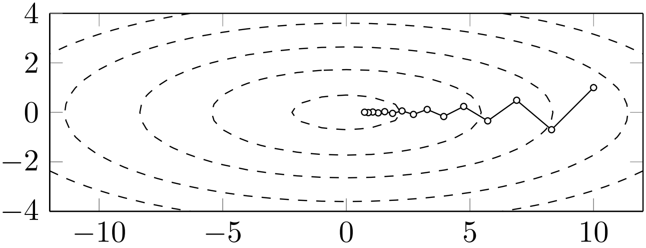


Figure: 梯度法的前15 次迭代

定理 (二次函数的收敛定理)

考虑正定二次函数

$$f(x)=\frac{1}{2}x^{\mathrm{T}}Ax-b^{\mathrm{T}}x,$$

其最优值点为x^*. 若使用梯度法$x^{k+1}=x^k-\alpha_k\nabla f(x^k)$ 并选取α_k 为精确线搜索步长, 即

$$\alpha_k=\frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^{\mathrm{T}}A\nabla f(x^k)},$$

则梯度法关于迭代点列$\{x^k\}$ 是Q-线性收敛的, 即

$$\|x^{k+1}-x^*\|_A^2\le\left(\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}\right)^2\|x^k-x^*\|_A^2,$$

其中λ_1,λ_n 分别为A 的最大、最小特征值, $\|x\|_A\overset{\text{def}}{=\!=}\sqrt{x^{\mathrm{T}}Ax}$ 为由正定矩阵A 诱导的范数.

二次函数的收敛定理

- 定理中线性收敛速度的常数和矩阵A 最大特征值与最小特征值之比有关.
- 从等高线角度来看，这个比例越大则$f(x)$ 的等高线越扁平，迭代路径折返频率会随之变高，梯度法收敛也就越慢.
- 这个结果其实说明了梯度法的一个很重大的缺陷：当目标函数的海瑟矩阵条件数较大时，它的收敛速度会非常缓慢.

梯度利普希茨连续

定义 (梯度利普希茨连续)

给定可微函数f，若存在$L > 0$，对任意的$x, y \in \mathrm{dom}f$有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{3}$$

则称f是**梯度利普希茨连续的**，相应利普希茨常数为L. 有时也简记为**梯度L-利普希茨连续或L-光滑**.

引理 (二次上界)

设可微函数$f(x)$的定义域$\mathrm{dom}f = \mathbb{R}^n$，且为梯度$L$-利普希茨连续的，则函数$f(x)$有二次上界：

$$f(y) \leq f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall\, x, y \in \mathrm{dom}f. \tag{4}$$

证明

可以证明:

$$
\begin{aligned}
& f(y)-f(x)-\nabla f(x)^{\mathrm{T}}(y-x) \\
= & \int_0^1 (\nabla f(x+t(y-x))-\nabla f(x))^{\mathrm{T}}(y-x)\mathrm{d}t \\
\leq & \int_0^1 \|\nabla f(x+t(y-x))-\nabla f(x)\|\|y-x\|\mathrm{d}t \\
\leq & \int_0^1 L\|y-x\|^2 t\mathrm{d}t=\frac{L}{2}\|y-x\|^2,
\end{aligned}
$$

其中最后一行的不等式利用了梯度利普希茨连续的条件(3). 整理可得(4) 式成立.

梯度法在凸函数上的收敛性

考虑梯度法

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

假设：

- 设函数$f(x)$ 为凸的梯度L -利普希茨连续函数
- 极小值$f^* = f(x^*) = \inf_x f(x)$ 存在且可达.
- 如果步长α_k 取为常数α 且满足$0 < \alpha < \frac{1}{L}$

结论：点列$\{x^k\}$的函数值收敛到最优值, 且在函数值的意义下收敛速度为$\mathcal{O}\left(\frac{1}{k}\right)$.

如果函数f 还是m -强凸函数, 则梯度法的收敛速度会进一步提升为Q-线性收敛.

证明

- 因为函数f 是利普希茨可微函数, 对任意的x, 根据二次上界引理,

$$f(x-\alpha\nabla f(x)) \le f(x) - \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2.$$

- 记$\tilde{x} = x - \alpha\nabla f(x)$ 并限制$0 < \alpha < \frac{1}{L}$, 我们有

$$\begin{aligned}
f(\tilde{x}) &\le f(x) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\
&\le f^* + \nabla f(x)^{\mathrm{T}}(x-x^*) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\
&= f^* + \frac{1}{2\alpha}\left(\|x-x^*\|^2 - \|x-x^*-\alpha\nabla f(x)\|^2\right) \\
&= f^* + \frac{1}{2\alpha}(\|x-x^*\|^2 - \|\tilde{x}-x^*\|^2),
\end{aligned}$$

其中第一个不等式是因为$0 < \alpha < \frac{1}{L}$, 第二个不等式为f 的凸性.

证明

- 在上式中取$x=x^{i-1},\tilde{x}=x^i$ 并将不等式对$i=1,2,\cdots,k$ 求和得到

$$
\begin{aligned}
\sum_{i=1}^{k}(f(x^i)-f^*) &\le \frac{1}{2\alpha}\sum_{i=1}^{k}\left(\|x^{i-1}-x^*\|^2-\|x^i-x^*\|^2\right)\\
&=\frac{1}{2\alpha}\left(\|x^0-x^*\|^2-\|x^k-x^*\|^2\right)\\
&\le \frac{1}{2\alpha}\|x^0-x^*\|^2.
\end{aligned}
$$

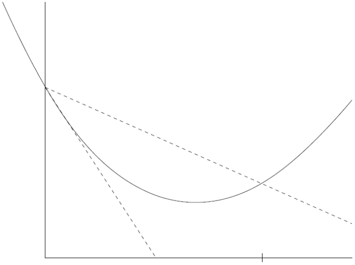
- 由于$f(x^i)$ 是非增的，所以

$$
f(x^k)-f^*\le \frac{1}{k}\sum_{i=1}^{k}(f(x^i)-f^*)\le \frac{1}{2k\alpha}\|x^0-x^*\|^2.
$$

Backtracking line search

initialize t_k at $\hat{t} > 0$(for example, $\hat{t} = 1$); take $t_k := \beta t_k$ until

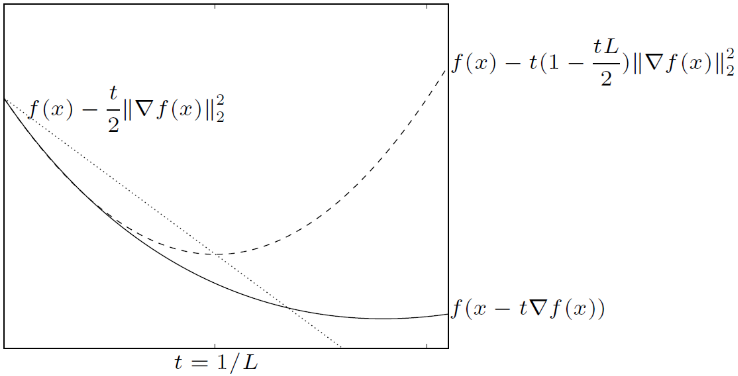
$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k \|\nabla f(x)\|_2^2$$



$0 < \beta < 1$; we will take $\alpha = 1/2$(mostly to simplify proofs)

Analysis for backtracking line search

line search with $\alpha = 1/2$ if f has a Lipschitz continuous gradient



selected step size satisfies $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

Convergence analysis

- from page 37:

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

- add the upper bounds to get

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

conclusion: same $1/k$ bound as with constant step size

凸函数性质

引理

设函数$f(x)$ 是$\mathbb{R}^n$ 上的凸可微函数，则以下结论等价：

1. f 的梯度为L -利普希茨连续的；
2. 函数$g(x) \stackrel{\text{def}}{=\!=} \frac{L}{2}x^{\mathrm{T}}x - f(x)$ 是凸函数；
3. $\nabla f(x)$ 有**余强制性**，即对任意的$x, y \in \mathbb{R}^n$，有

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

$(1) \implies (2)$ 即证$g(x)$ 的单调性. 对任意$x, y \in \mathbb{R}^n$,

$$\begin{aligned}(\nabla g(x) - \nabla g(y))^{\mathrm{T}}(x - y) &= L\|x - y\|^2 - (\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \\ &\geq L\|x - y\|^2 - \|x - y\|\|\nabla f(x) - \nabla f(y)\| \geq 0.\end{aligned}$$

因此$g(x)$ 为凸函数.

凸函数性质

引理 (梯度L-利普希茨函数的性质)

设可微函数$f(x)$ 的定义域为$\mathbb{R}^n$ 且存在一个全局极小点x^*, 若$f(x)$ 为梯度L-利普希茨连续的, 则对任意的x 有

$$\frac{1}{2L}\|\nabla f(x)\|^2 \le f(x) - f(x^*).$$

(2) $\Longrightarrow$ (3)

- 构造辅助函数

$$f_x(z) = f(z) - \nabla f(x)^{\mathrm{T}} z,$$
$$f_y(z) = f(z) - \nabla f(y)^{\mathrm{T}} z,$$

  容易验证f_x 和f_y 均为凸函数.

- $g_x(z) = \frac{L}{2} z^{\mathrm{T}} z - f_x(z)$ 关于z 是凸函数. 根据凸函数的性质, 我们有

$$g_x(z_2) \ge g_x(z_1) + \nabla g_x(z_1)^{\mathrm{T}}(z_2 - z_1), \quad \forall\, z_1, z_2 \in \mathbb{R}^n.$$

  整理可推出$f_x(z)$ 有二次上界, 且对应的系数也为L.

- 注意到$\nabla f_x(x)=0$, 这说明x 是$f_x(z)$ 的最小值点. 由上页引理,

$$
\begin{aligned}
f_x(y)-f_x(x) &= f(y)-f(x)-\nabla f(x)^{\mathrm{T}}(y-x) \\
&\geq \frac{1}{2L}\|\nabla f_x(y)\|^2=\frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|^2.
\end{aligned}
$$

- 同理, 对$f_y(z)$ 进行类似的分析可得

$$
f(x)-f(y)-\nabla f(y)^{\mathrm{T}}(x-y)\geq \frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|^2.
$$

将以上两式不等号左右分别相加, 可得余强制性.

$(3)\implies(1)$ 由余强制性和柯西不等式,

$$
\begin{aligned}
\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2 &\leq (\nabla f(x)-\nabla f(y))^{\mathrm{T}}(x-y) \\
&\leq \|\nabla f(x)-\nabla f(y)\|\|x-y\|,
\end{aligned}
$$

整理后即可得到$f(x)$ 是梯度L-利普希茨连续的.

梯度法在强凸函数上的收敛性

定理 (梯度法在强凸函数上的收敛性)

设函数$f(x)$ 为m -强凸的梯度L -利普希茨连续函数, $f(x^*) = \inf_x f(x)$ 存在且可达. 如果步长α 满足$0 < \alpha < \frac{2}{m+L}$, 那么由梯度下降法迭代得到的点列$\{x^k\}$ 收敛到x^*, 且为Q-线性收敛.

- 首先根据f 强凸且∇f 利普希茨连续, 可得

$$g(x) = f(x) - \frac{m}{2} x^{\mathrm{T}} x$$

为凸函数且$\frac{L-m}{2} x^{\mathrm{T}} x - g(x)$ 为凸函数.

- 由引理知函数$g(x)$ 是梯度$(L-m)$ -利普希茨连续的. 再次利用引理可得关于$g(x)$ 的余强制性

$$(\nabla g(x) - \nabla g(y))^{\mathrm{T}} (x - y) \geq \frac{1}{L-m} \|\nabla g(x) - \nabla g(y)\|^2.$$

梯度法在强凸函数上的收敛性

- 代入$g(x)$ 的表达式, 可得

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

- 再估计固定步长下梯度法的收敛速度. 设步长$\alpha \in \left(0, \frac{2}{m+L}\right)$, 对$x^k, x^*$ 应用上式并注意到$\nabla f(x^*) = 0$ 得

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \alpha \nabla f(x^k) - x^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\alpha \nabla f(x^k)^{\mathrm{T}}(x^k - x^*) + \alpha^2 \|\nabla f(x^k)\|^2 \\
&\leq \left(1 - \alpha \frac{2mL}{m+L}\right)\|x^k - x^*\|^2 + \alpha\left(\alpha - \frac{2}{m+L}\right)\|\nabla f(x^k)\|^2 \\
&\leq \left(1 - \alpha \frac{2mL}{m+L}\right)\|x^k - x^*\|^2
\end{aligned}$$

$$\Rightarrow \|x^k - x^*\|^2 \leq c^k \|x^0 - x^*\|^2, \quad c = 1 - \alpha \frac{2mL}{m+L} < 1.$$

函数值收敛

强凸函数假设下

- 迭代点列$\{x^k\}$Q-线性收敛
- 如果取$t = \frac{2}{m+L}$，则有$c = \left(\frac{(\gamma-1)}{(\gamma+1)}\right)^2$ 且$\gamma = L/m$

如果$\mathbf{dom}\, f = \mathbf{R}^n$且$f$ 有极小点x^*, 则

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \le f(x) - f(x^*) \le \frac{L}{2}\|x - x^*\|_2^2 \quad \forall x$$

因此：

$$f(x^k) - f^* \le \frac{L}{2}\|x^k - x^*\|_2^2 \le \frac{c^k L}{2}\|x^0 - x^*\|_2^2$$

函数值的估计： 达到$f(x^k) - f^* \le \epsilon$ 的迭代步数是$O(\log(1/\epsilon))$

Barzilar-Borwein 方法

- Barzilar-Borwein (BB) 方法是一种特殊的梯度法，经常比一般的梯度法有着更好的效果.
- BB 方法的下降方向仍是点x^k 处的负梯度方向$-\nabla f(x^k)$，但步长α_k 并不是直接由线搜索算法给出的.
- 考虑梯度下降法的格式：

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \Longleftrightarrow x^{k+1} = x^k - D^k \nabla f(x^k),$$

其中$D^k = \alpha_k I$.

- BB 方法选取的α_k 是如下两个最优问题之一的解：

$$\min_{\alpha} \quad \|\alpha y^{k-1} - s^{k-1}\|^2,$$

$$\min_{\alpha} \quad \|y^{k-1} - \alpha^{-1} s^{k-1}\|^2,$$

其中引入记号$s^{k-1} \overset{\text{def}}{=\!=} x^k - x^{k-1}$ 以及$y^{k-1} \overset{\text{def}}{=\!=} \nabla f(x^k) - \nabla f(x^{k-1})$.

Barzilar-Borwein 方法

- 容易验证问题的解分别为

$$\alpha_{\mathrm{BB1}}^k \stackrel{\mathrm{def}}{=\!=} \frac{(s^{k-1})^{\mathrm{T}} y^{k-1}}{(y^{k-1})^{\mathrm{T}} y^{k-1}} \text{ 和 } \alpha_{\mathrm{BB2}}^k \stackrel{\mathrm{def}}{=\!=} \frac{(s^{k-1})^{\mathrm{T}} s^{k-1}}{(s^{k-1})^{\mathrm{T}} y^{k-1}},$$

- 因此可以得到BB 方法的两种迭代格式：

$$x^{k+1} = x^k - \alpha_{\mathrm{BB1}}^k \nabla f(x^k) \text{ 和 } x^{k+1} = x^k - \alpha_{\mathrm{BB2}}^k \nabla f(x^k).$$

- 计算两种BB 步长的任何一种仅仅需要函数相邻两步的梯度信息和迭代点信息，不需要任何线搜索算法即可选取算法步长.
- BB方法计算出的步长可能过大或过小，因此我们还需要将步长做上界和下界的截断，即选取$0 < \alpha_m < \alpha_M$ 使得

$$\alpha_m \le \alpha_k \le \alpha_M.$$

- BB 方法本身是非单调方法，有时也配合非单调收敛准则使用以获得更好的实际效果.

非单调线搜索的BB方法

Algorithm 2 非单调线搜索的BB方法

给定x^0, 选取初值$\alpha > 0$, 整数$M \geq 0$, $c_1, \beta, \varepsilon \in (0,1)$, $k = 0$.
while $\|\nabla f(x^k)\| > \varepsilon$ **do**
  while $f(x^k - \alpha \nabla f(x^k)) \geq \max_{0 \leq j \leq \min(k,M)} f(x^{k-j}) - c_1 \alpha \|\nabla f(x^k)\|^2$ **do**
    令$\alpha \leftarrow \beta\alpha$.
  end while
  令$x^{k+1} = x^k - \alpha \nabla f(x^k)$.
  根据BB步长公式之一计算α, 并做截断使得$\alpha \in [\alpha_m, \alpha_M]$.
  $k \leftarrow k + 1$.
end while

二次函数的BB 方法

- 设二次函数$f(x,y) = x^2 + 10y^2$, 并使用*BB* 方法进行迭代, 初始点为$(-10,-1)$.
- *BB* 方法的收敛速度较快, 在经历*15* 次迭代后已经接近最优值点. 从等高线也可观察到*BB* 方法是非单调方法.
- 实际上, 对于正定二次函数, BB方法有R-线性收敛速度.

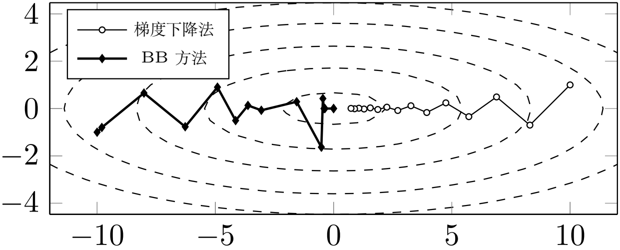


Figure: 梯度法与BB 方法的前15 次迭代

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

- 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：

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

- 上述定义实际上是Huber 损失函数的一种变形，当$\delta \to 0$ 时，光滑函数$l_\delta(x)$ 和绝对值函数$|x|$ 会越来越接近.

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.

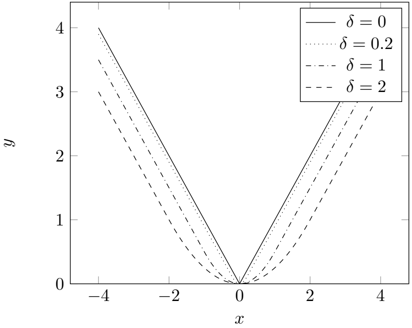


Figure: 当 δ 取不同值时 $l_\delta(x)$ 的图形

- $f_\delta(x)$ 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：

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

- $f_\delta(x)$ 的梯度是利普希茨连续的，且相应常数为$L = \|A^\mathrm{T}A\|_2 + \frac{\mu}{\delta}$.
- 根据梯度法在凸函数上的收敛性定理，固定步长需不超过$\frac{1}{L}$ 才能保证算法收敛，如果δ 过小，那么我们需要选取充分小的步长α_k 使得梯度法收敛.

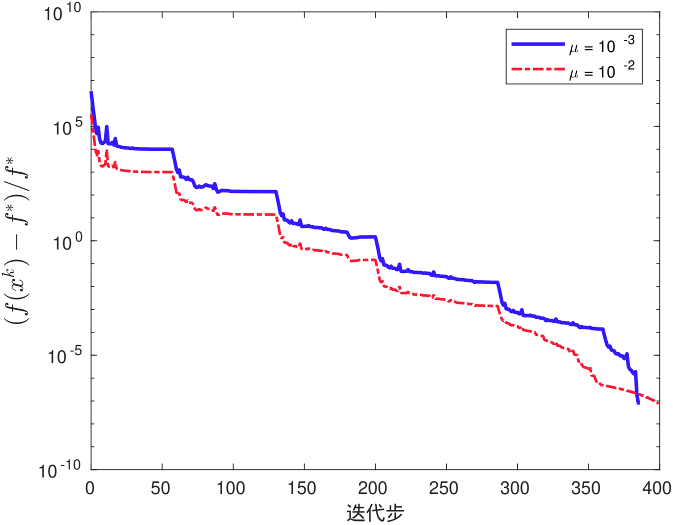
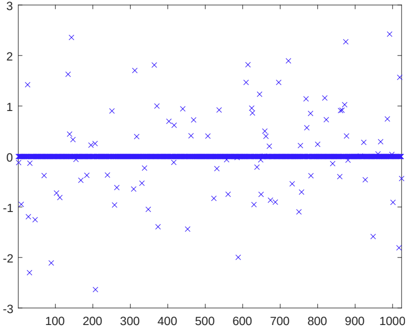
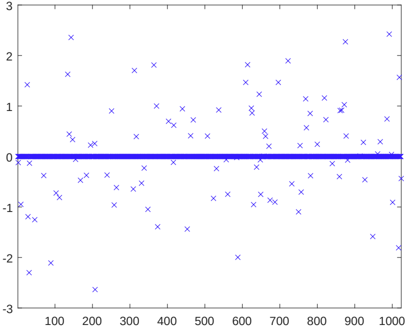


Figure: 光滑化LASSO 问题求解迭代过程

LASSO问题求解



(a) 精确解



(b) 梯度法解

Figure: 光滑化LASSO 问题求解结果