LASSO 问题的光滑化策略及正则化参数的选取：

本小节利用梯度法来求解 LASSO 问题．这个问题的原始形式为

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

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．考虑到目标函数的不光滑项为 $\|x\|_1$，它实际上是 x 各个分量绝对值的和，如果能找到一个光滑函数来近似绝对值函数，那么梯度法就可以被用在 LASSO 问题的求解上．在实际应用中，我们可以考虑如下一维光滑函数：

$$l_\delta(x)=\begin{cases}\dfrac{1}{2\delta}x^2, & |x|<\delta,\\ |x|-\dfrac{\delta}{2}, & \text{其他}.\end{cases} \tag{6.2.11}$$

定义(6.2.11)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．图6.6展示了当 δ 取不同值时 $l_\delta(x)$ 的图形．

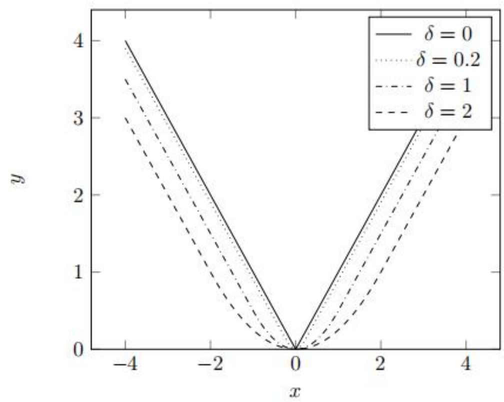

图 6.6　当 δ 取不同值时 $l_\delta(x)$ 的图形

因此，我们构造光滑化 LASSO 问题为

$$\min \quad f_\delta(x)=\frac{1}{2}\|Ax-b\|^2+\mu L_\delta(x), \tag{6.2.12}$$

其中 δ 为给定的光滑化参数，在这里

$$L_\delta(x)=\sum_{i=1}^{n} l_\delta(x_i),$$

即对 x 的每个分量作用光滑函数 (6.2.11) 再整体求和. 容易计算出 $f_\delta(x)$ 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:

考虑 LASSO 问题

$$\min_x \quad \frac{1}{2}\|Ax-b\|^2 + \mu\|x\|_1,$$

其中 $\mu > 0$ 是正则化参数. 我们知道求解 LASSO 问题的最终目标是为了解决如下基追踪 (BP) 问题:

$$\begin{aligned} \min \quad & \|x\|_1, \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

在这里 $Ax = b$ 是一个欠定方程组. 注意到 BP 问题是一个等式约束的非光滑优化问题, 我们使用二次罚函数作用于等式约束 $Ax = b$, 可得

$$(\nabla L_\delta(x))_i = \begin{cases} \operatorname{sign}(x_i), & |x_i| > \delta, \\ \dfrac{x_i}{\delta}, & |x_i| \leqslant \delta. \end{cases}$$

显然 $f_\delta(x)$ 的梯度是利普希茨连续的, 且相应常数为 $L = \|A^{\mathrm{T}}A\|_2 + \dfrac{\mu}{\delta}$. 根据定理6.3, 固定步长需不超过 $\dfrac{1}{L}$ 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.

$$\min_x \quad \|x\|_1 + \frac{\sigma}{2}\|Ax-b\|^2.$$

令 $\mu = \dfrac{1}{\sigma}$, 则容易看出使用 $\dfrac{1}{\mu}$ 作为二次罚因子时, BP 问题的罚函数子问题就等价于 LASSO 问题. 这一观察至少说明了以下两点: 第一, LASSO 问题的解和 BP 问题的解本身不等价, 当 μ 趋于 0 时, LASSO 问题的解收敛到 BP 问题的解; 第二, 当 μ 比较小时, 根据之前的讨论, 此时 BP 问题罚函数比较病态, 若直接求解则收敛速度会很慢. 根据罚函数的思想, 罚因子应该逐渐增加到无穷, 这等价于在 LASSO 问题中先取一个较大的 μ, 之后再不断缩小 μ 直至达到我们所要求解的值. 具体算法在算法7.2中给出.

牛顿共轭梯度方法：

该算法利用非精确牛顿法（牛顿-共轭梯度法）求解无约束优化问题

$$\min_x f(x)$$

在第 k 步迭代，下降方向 d^k 通过求解下面的牛顿方程 $(\nabla^2 f(x^k))d^k = -\nabla f(x^k)$ 得到。选取合适的步长 α_k，牛顿法的迭代格式为 $x^{k+1} = x^k + \alpha_k d_k$.

对于规模较大的问题，精确求解牛顿方程组的代价比较高。事实上，牛顿方程求解等价于无约束二次优化问题：

$$\min_{d^k} \frac{1}{2}(d^k)^\top \nabla^2 f(x^k) d^k + (\nabla f(x^k))^\top d^k,$$

其可以通过共轭梯度法来进行求解。

共轭梯度法：

对于二次极小化问题

$$\min_d \quad q(s) \stackrel{\text{def}}{=\!=} g^\mathrm{T} s + \frac{1}{2} s^\mathrm{T} B s,$$

给定初值 $s^0 = 0, r^0 = g, p^0 = -g$，共轭梯度法的迭代过程为

$$\begin{aligned}
\alpha_{k+1} &= \frac{\|r^k\|^2}{(p^k)^\mathrm{T} B p^k}, \\
s^{k+1} &= s^k + \alpha_k p^k, \\
r^{k+1} &= r^k + \alpha_k B p^k, \\
\beta_k &= \frac{\|r^{k+1}\|^2}{\|r^k\|^2}, \\
p^{k+1} &= -r^{k+1} + \beta_k p^k,
\end{aligned}$$

其中迭代序列 $\{s^k\}$ 最终的输出即为二次极小化问题的解，算法的终止准则是判断 $\|r^k\|$ 是否足够小.

L-BFGS 求解优化问题：

针对无约束优化问题

$$\min_x f(x),$$

L-BFGS 在拟牛顿法 BFGS 迭代格式的基础上进行修改， 用以解决大规模问题的存储和计算困难。

对于拟牛顿法中的迭代方向 $d^k = H^k \nabla f(x^k)$ 考虑利用递归展开的方式进行求解。首先，对于

BFGS 迭代格式，

$$H^{k+1} = (V^k)^\top H^k V^k + \rho_k s^k (s^k)^\top, \text{ 其中 } \rho_k = \frac{1}{(y^k)^\top s^k}, V^k = I - \rho_k y^k (s^k)^\top$$

将其递归地展开得到

$$
\begin{aligned}
-H^k\nabla f(x^k) = \quad & -(V^{k-m}\cdots V^{k-1})^\top H^{k-m}(V^{k-m}\cdots V^{k-1})\nabla f(x^k) \\
& -\rho_{k-m}(V^{k-m+1}\cdots V^{k-1})^\top s^{k-m}(s^{k-m})^\top(V^{k-m+1}\cdots V^{k-1})\nabla f(x^k) \\
& -\rho_{k-m+1}(V^{k-m+2}\cdots V^{k-1})^\top s^{k-m+1}(s^{k-m+1})^\top(V^{k-m+2}\cdots V^{k-1})\nabla f(x^k) \\
& -\cdots \\
& -\rho_{k-1}s^{k-1}(s^{k-1^\top})\nabla f(x^k).
\end{aligned}
$$

我们只需对其中的 H^{k-m} 进行某种估计，即可在展开深度为 m 的情况对 $d^k = H^k\nabla f(x^k)$ 进行近似求解。当用数量矩阵来近似时，即

$$
\hat{H}^{k-m} = \gamma_k I\text{，其中 } \gamma_k = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}}
$$

对应 BB 方法的第二个步长。

LASSO 问题的 PPA 算法：

近似点算法（PPA）对于 LASSO 问题，考虑其等价形式

$$
\min_{x,y} f(x,y) = \mu\|x\|_1 + \tfrac{1}{2}\|y\|_2^2, \quad \text{s.t.} \quad Ax - y - b = 0.
$$

近似点算法的一个迭代步为

$$
(x^{k+1}, y^{k+1}) \approx \arg\min_{(x,y)\in\mathcal{D}}\left\{f(x,y) + \tfrac{1}{2t_k}(\|x - x^k\|_2^2 + \|y - y^k\|_2^2)\right\}
$$

其中 $\mathcal{D} = \{(x,y)\big|Ax - y = b\}$。对于子问题考虑其对偶问题，引入拉格朗日乘子 z，并令

$$
\Phi_k(z) = \inf_x\left\{\mu\|x\|_1 + z^\top Ax + \tfrac{1}{2t_k}\|x - x^k\|_2^2\right\} + \inf_y\left\{\tfrac{1}{2}\|y\|_2^2 - z^\top y\right.
$$

$$
\left. + \tfrac{1}{2t_k}\|y - y^k\|_2^2\right\} - b^\top z,
$$

则其迭代格式满足

$$
z^{k+1} = \arg\max_z \Phi_k(z).
$$

低秩矩阵恢复的 PGA 算法

考虑低秩矩阵恢复模型 (1.3.3)：

$$\min_{X\in\mathbb{R}^{m\times n}} \quad \mu\|X\|_* + \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij}-M_{ij})^2,$$

其中 M 是想要恢复的低秩矩阵，但是只知道其在下标集 Ω 上的值．令

$$f(X)=\frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij}-M_{ij})^2, \quad h(X)=\mu\|X\|_*,$$

定义矩阵 $P\in\mathbb{R}^{m\times n}$：

$$P_{ij}=\begin{cases}1, & (i,j)\in\Omega,\\ 0, & \text{其他},\end{cases}$$

则

$$\begin{aligned} f(X) &= \frac{1}{2}\|P\odot(X-M)\|_F^2,\\ \nabla f(X) &= P\odot(X-M),\\ \mathrm{prox}_{t_k h}(X) &= U\mathrm{Diag}\left(\max\{|d|-t_k\mu,0\}\right)V^{\mathrm{T}}, \end{aligned}$$

其中 $X=U\mathrm{Diag}(d)V^{\mathrm{T}}$ 为矩阵 X 的约化的奇异值分解．近似点梯度法的迭代格式为

$$\begin{aligned} Y^k &= X^k - t_k P\odot(X^k-M),\\ X^{k+1} &= \mathrm{prox}_{t_k h}(Y^k). \end{aligned}$$