非线性最小二乘问题

宋晓良

大连理工大学数学科学学院

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

提纲

最小二乘问题

$$\min_{x} \quad f(x) = \frac{1}{2}\sum_{j=1}^{m} r_j^2(x) \tag{1}$$

- 其中$r_j : \mathbb{R}^n \rightarrow \mathbb{R}$ 是光滑函数，并且假设$m \geq n$. 称r_j 为残差.
- 记$r : \mathbb{R}^n \rightarrow \mathbb{R}^m$

$$r(x) = (r_1(x), r_2(x), ..., r_m(x))^T.$$

  问题可以表述为$\min \quad f(x) = \frac{1}{2}\|r(x)\|_2^2$
- 一般情况下不是凸问题
- 问题1是无约束优化问题，可以直接使用线搜索或拟牛顿法求解

- 记$J(x) \in \mathbb{R}^{m\times n}$是向量值函数$r(x)$ 在点x 处的雅可比矩阵：

$$J(x) = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}.$$

- $f(x)$的梯度和海瑟矩阵：

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^{\mathrm{T}} r(x), \tag{2a}$$

$$\nabla^2 f(x) = \sum_{j=1}^{m} \nabla r_j(x) \nabla r_j(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x) \tag{2b}$$

$$= J(x)^{\mathrm{T}} J(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x), \tag{2c}$$

$$\nabla^2 f(x) = J(x)^{\mathrm{T}} J(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x)$$

- $\nabla^2 f(x)$在形式上分为两部分
- 在计算$f(x)$导数时已经求出$J(x)$，第一项可自然得到。第二项需要额外计算。
- 如果在最优解附近，残量值较小或残量函数接近线性函数，第二项可以忽略，可以用$J(x)^T J(x)$近似海瑟矩阵，基于牛顿法，结合线搜索或信赖域方法，可设计出高斯-牛顿方法和Levenberg-Marquardt 方法
- 如果第二项不可忽略，则需要引入带结构的拟牛顿方法。

最小二乘问题

- 若$r(x) = Ax - b$，则是线性最小二乘问题，
- $\nabla f(x) = A^T(Ax - b)$. 最优解满足正则化方程：

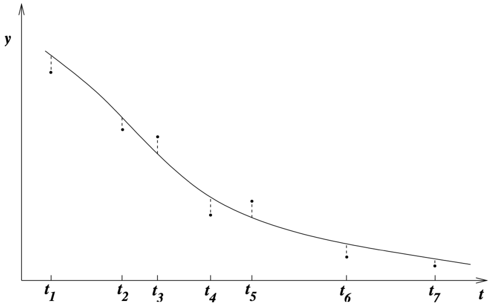
$$A^T Ax = A^T b$$

- 线性问题的求解是非线性问题求解的基础. 求解线性最小二乘问题的方法有：
  - 对$A^T A$直接做choloskey分解，简单便捷但受A的条件数影响大。
  - 对A做QR分解，较稳定，相对误差小。
  - 对A做SVD分解，可以获得更精确的敏感性信息。
  - 当问题规模较大时，迭代法更有效，如共轭梯度法。

实例:模型拟合

$$\min_x \quad \frac{1}{2}\sum_{j=1}^{n}(\phi(t_j;x)-y_j)^2$$

- 模型$\phi(t;x)=x_1+tx_2+t_3^x+x_4e^{-x_5t}$ 依赖于参数向量x。
- $(t_1,y_1),(t_2,y_2),\cdots,(t_n,y_n)$是数据点
- 目标为寻找合适的模型参数x



提纲

高斯-牛顿方法

- 可被看作牛顿法+线搜索
- 在迭代点x_k，标准牛顿法为，计算更新方向d_k：

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

然后做一步更新 $x_{k+1} = x_k + d_k$.

- 高斯-牛顿法的迭代方向 d_k^{GN} 满足：

$$J_k^T J_k d_k^{GN} = -J_k^T r_k. \tag{3}$$

其中J_k、r_k分别是$J(x_k)$、$r(x_k)$的简写

- 使用了近似

$$\nabla^2 f_k \approx J_k^T J_k$$

省略了对 $\nabla^2 r_j$ 的计算，极大的减少了计算量。

高斯-牛顿方法

- 方程(3)与线性最小二乘问题的正则化方程类似，迭代方向是如下问题的解
$$\min_d \ \frac{1}{2}\|J_k d + r_k\|^2$$
- 求解该问题时，可以直接对J_k做QR分解或SVD分解，无需计算出$J_k^T J_k$。
- 若使用共轭梯度法求解，需要计算向量和矩阵$J_k^T J_k$的乘法，可以依次乘J_k和J_k^T，无需计算$J_k^T J_k$。
- 另一种理解高斯-牛顿方法的方式为，在点x_k处，考虑下一步更新$x_k + d$，做近似$r(x_k + d) \approx r_k + J_k d$，原问题近似为：

$$\min_d \quad f(x_k + d) = \frac{1}{2}\|r(x_k + d)\|^2 \approx \frac{1}{2}\|J_k d + r_k\|^2$$

高斯-牛顿方法

算法可总结如下

Algorithm 1 高斯–牛顿法

给定初始值x_0，$k \leftarrow 0$．
while 未达到收敛准则 **do**
    计算残差向量r_k，雅可比矩阵J_k．
    求解线性最小二乘问题 $\min_d \frac{1}{2}\|J_k d + r_k\|^2$ 确定下降方向d_k．
    使用线搜索准则计算步长α_k．
    更新：$x_{k+1} = x_k + \alpha_k d_k$．
    $k \leftarrow k + 1$．
end while

- 线搜索条件可选择Armijo或Wolfe.
- 若J_k满秩且∇f_k非零，则d_k^{GN}是一个下降方向：

$$\begin{aligned}(d_k)^{\mathrm{T}}\nabla f(x_k) &= d_k^{\mathrm{T}} J_k^{\mathrm{T}} r_k = -d_k^{\mathrm{T}} J_k^{\mathrm{T}} J_k d_k \\ &= -\|J_k d_k\|^2 \leq 0.\end{aligned}$$

- 那么d^k是一个合适的线搜索方向，全局收敛性的证明可以套用线搜索的证明。

全局收敛性分析

- 注意到，雅可比矩阵J_k 的非奇异性很关键，在这个条件下建立收敛性.
- 具体为：假设雅可比矩阵$J(x)$的奇异值一致地大于0，即存在$\gamma > 0$使得

$$\|J(x)z\| \geq \gamma\|z\|, \quad \forall\, x \in \mathcal{N}, \tag{4}$$

其中$\mathcal{N}$是下水平集

$$\mathcal{L} = \{x \mid f(x) \leq f(x_0)\} \tag{5}$$

的一个邻域，x_0是算法的初始点，且假设$\mathcal{L}$是有界的.

全局收敛性

Theorem

全局收敛性 如果每个残差函数r_j在有界下水平集(5)的一个邻域$\mathcal{N}$内是利普希茨连续可微的，并且雅可比矩阵$J(x)$在$\mathcal{N}$内满足一致满秩条件(4)，而步长满足*Wolfe*准则，则对高斯–牛顿法得到的序列$\{x_k\}$有

$$\lim_{k\to\infty}(J_k)^{\mathrm{T}}r_k=0.$$

- 在最优值点x^*附近，当$\sum_{i=1}^m r_i(x)\nabla^2 r_i(x)$较小时（$r(x^*)$很小或在$x^*$附近$r$接近仿射函数），$J_k^T J_k$占主导位置，高斯-牛顿方法有类似牛顿法的收敛速度。

Theorem (局部收敛性)

设$r_i(x)$二阶连续可微，x^*是最小二乘问题(1) 的最优解，海瑟矩阵$\nabla^2 f(x)$和其近似矩阵$J(x)^\mathrm{T} J(x)$均在点x^*的一个邻域内利普希茨连续，则当高斯–牛顿算法步长α_k 恒为1 时，

$$\|x_{k+1} - x^*\| \le C\|((J^*)^\mathrm{T} J^*)^{-1} H^*\| \|x_k - x^*\| + \mathcal{O}(\|x_k - x^*\|^2), \tag{6}$$

其中$H^* = \sum_{i=1}^m r(x^*)\nabla^2 r(x^*)$为海瑟矩阵$\nabla^2 f(x^*)$去掉$J(x^*)^\mathrm{T} J(x^*)$的部分，$C > 0$为常数.

提纲

Levenberg-Marquardt 方法

- 当J_k不满秩时，(3)有很多个解，应该怎么更新?
- LM方法本质为信赖域方法，更新方向为如下问题的解

$$\min_{d} \quad \frac{1}{2}\|J^k d + r^k\|^2, \quad \text{s.t.} \quad \|d\| \le \Delta_k. \tag{7}$$

- LM 方法将如下近似当作信赖域方法中的m_k:

$$m_k(d) = \frac{1}{2}\|r^k\|^2 + d^{\mathrm{T}}(J^k)^{\mathrm{T}} r^k + \frac{1}{2} d^{\mathrm{T}}(J^k)^{\mathrm{T}} J^k d. \tag{8}$$

- 同样使用$(J^k)^{\mathrm{T}}J^k$来近似海瑟矩阵.

- 类似信赖域方法，引入如下定义来衡量$m_k(d)$近似程度的好坏：

$$\rho_k=\frac{f\left(x^k\right)-f\left(x^k+d^k\right)}{m_k\left(0\right)-m_k\left(d^k\right)} \tag{9}$$

为函数值实际下降量与预估下降量（即二阶近似模型下降量）的比值．

- 如果ρ_k接近1，说明$m_k(d)$来近似$f(x)$是比较成功的，则应该扩大Δ_k；如果ρ_k非常小甚至为负，就说明我们过分地相信了二阶模型$m_k(d)$，此时应该缩小Δ_k．
- 只有当ρ_k足够大，也就是对模型拟合较好时，才进行一步更新，否则不更新。

Levenberg-Marquardt 方法

Algorithm 2 Levenberg-Marquardt 方法

给定最大半径$\Delta_{\max}$，初始半径Δ_0，初始点x^0，$k \leftarrow 0$.

给定参数$0 \le \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1$，$\gamma_1 < 1 < \gamma_2$.

while 未达到收敛准则 **do**

  计算子问题(7)得到迭代方向d^k.

  根据(9) 计算下降率ρ_k.

  更新信赖域半径：

$$\Delta_{k+1} = \begin{cases} \gamma_1 \Delta_k, & \rho_k < \bar{\rho}_1, \\ \min\{\gamma_2 \Delta_k, \Delta_{\max}\}, & \rho_k > \bar{\rho}_2 \text{ 以及 } \|d^k\| = \Delta_k, \\ \Delta_k, & \text{其他}. \end{cases}$$

  更新自变量：

$$x^{k+1} = \begin{cases} x^k + d^k, & \rho_k > \eta, \\ x^k, & \text{其他}. \end{cases} \qquad /\star \text{ 只有下降比例足够大才更新} \star/$$

  $k \leftarrow k+1$.

end while

子问题求解

Corollary

向量d^*是信赖域子问题

$$\min_d \quad \frac{1}{2}\|Jd+r\|^2, \quad \text{s.t.} \quad \|d\| \le \Delta$$

的解当且仅当d^*是可行解并且存在数$\lambda \ge 0$使得

$$(J^\mathrm{T}J+\lambda I)d^* = -J^\mathrm{T}r, \tag{10}$$
$$\lambda(\Delta - \|d^*\|) = 0. \tag{11}$$

- 问题(10)等价于线性最小二乘问题，具体实现时可利用系数矩阵的结构

$$\min_d \quad \frac{1}{2}\left\|\begin{bmatrix} J \\ \sqrt{\lambda}I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix}\right\|^2.$$

子问题求解

$$\min_{d} \quad \frac{1}{2}\left\|\begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix}\right\|^2.$$

- 在试探λ的值时，J的块不变，设$J = QR$，则

$$\begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} QR \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} R \\ \sqrt{\lambda} I \end{bmatrix}.$$

- 矩阵$\begin{bmatrix} R \\ \sqrt{\lambda} I \end{bmatrix}$有较多的零元素，可以使用Household变换或Givens变换完成QR分解。
- 如果矩阵J没有显式形式，只能提供矩阵乘法，则仍然可以用截断共轭梯度法。

收敛性分析

Theorem

假设常数$\eta \in (0, \frac{1}{4})$，下水平集$\mathcal{L}$是有界的且每个$r_i(x)$在下水平集$\mathcal{L}$的一个邻域$\mathcal{N}$内是利普希茨连续可微的．假设对于任意的$k$，子问题(7)的近似解$d_k$满足

$$m_k(0) - m_k(d_k) \geq c_1 \|(J_k)^{\mathrm{T}} r_k\| \min\left\{\Delta_k, \frac{\|(J_k)^{\mathrm{T}} r_k\|}{\|(J_k)^{\mathrm{T}} J_k\|}\right\},$$

其中$c_1 > 0$且$\|d_k\| \leq \gamma \Delta_k, \gamma \geq 1$，则

$$\lim_{k\to\infty} \nabla f(x_k) = \lim_{k\to\infty} (J_k)^{\mathrm{T}} r_k = 0.$$

LMF

- 信赖域型LM 方法本质上是固定信赖域半径Δ, 通过迭代寻找满足条件的乘子λ, 每一步迭代需要求解线性方程组

$$\left(J^{\mathrm{T}}J + \lambda I\right) d = -J^{\mathrm{T}} r$$

  该步计算代价较大。

- 注意到在LM方法中，由于$J_k^T J_k \succ 0$，那么有$-\lambda_1 < 0$，此时有$\lambda > -\lambda_1$，因此若λ越大，d的模长就越小。

- 调整λ的大小等价于调整信赖域半径的大小，这意味着，Δ被λ隐式决定。

LMF

- LM的更新基于Δ，LMF的更新直接基于λ，每一步求解子问题：

$$\min_{d} \quad \|Jd + r\|_2^2 + \lambda\|d\|_2^2.$$

- 调整λ的原则可以参考信赖域半径的调整原则
- 考虑参数ρ_k

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)} \tag{12}$$

较大可以减小下一步的λ，较小可以增大下一步的λ。

Algorithm 3 LMF 方法

1: 给定初始点x_0，初始乘子λ_0，$k \leftarrow 0$.
2: 给定参数$0 \le \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1$，$\gamma_1 < 1 < \gamma_2$.
3: **while** 未达到收敛准则 **do**
4: 求解LM 方程$((J_k)^{\mathrm{T}} J_k + \lambda I)d = -(J_k)^{\mathrm{T}} r_k$得到迭代方向$d_k$.
5: 根据(12)式计算下降率ρ_k.
6: 更新乘子：

$$\lambda_{k+1} = \begin{cases} \gamma_2 \lambda_k, & \rho_k < \bar{\rho}_1, \quad \text{/* 扩大乘子（缩小信赖域半径）*/} \\ \gamma_1 \lambda_k, & \rho_k > \bar{\rho}_2, \quad \text{/* 缩小乘子（扩大信赖域半径）*/} \\ \lambda_k, & \text{其他}. \quad \text{/* 乘子不变 */} \end{cases}$$

7: 更新自变量：

$$x_{k+1} = \begin{cases} x_k + d_k, & \rho_k > \eta, \quad \text{/* 只有下降比例足够大才更新 */} \\ x_k, & \text{其他}. \end{cases}$$

8: $k \leftarrow k + 1$.
9: **end while**

提纲

1. 非线性最小二乘问题
2. 高斯-牛顿方法
3. Levenberg-Marquardt 方法
4. 大残量问题的拟牛顿算法

大残量问题的拟牛顿算法

- 大残量问题中，海瑟矩阵的第二部分不可忽视，此时高斯– 牛顿法和LM 方法可能只有线性的收敛速度。
- 此时如果直接使用牛顿法，则开销太大；直接使用拟牛顿法，又似乎忽略了问题的特殊结构。
- 重新写出海瑟矩阵：

$$\nabla^2 f(x) = J(x)^{\mathrm{T}} J(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x)$$

第一项是容易求解，可以保留。第二项不易求解但不可忽略，用拟牛顿法进行近似。

大残量问题的拟牛顿算法

- 使用B_k来表示$\nabla^2 f(x_k)$的近似矩阵，T_k表示$\sum_{j=1}^m r_j(x_k)\nabla^2 r_j(x_k)$的近似，即

$$B_k = (J_k)^{\mathrm{T}} J_k + T_k,$$

- 目标为

$$T_{k+1} \approx \sum_{j=1}^{m} r_j\left(x_{k+1}\right) \nabla^2 r_j\left(x_{k+1}\right)$$

- 记$s_k = x_{k+1} - x_k$，T_{k+1} 应该尽量保留原海瑟矩阵的性质

$$\begin{aligned} T_{k+1} s_k &\approx \sum_{j=1}^{m} r_j(x_{k+1}) \left(\nabla^2 r_j(x_{k+1})\right) s_k \\ &\approx \sum_{j=1}^{m} r_j(x_{k+1})(\nabla r_j(x_{k+1}) - \nabla r_j(x_k)) \\ &= (J_{k+1})^{\mathrm{T}} r_{k+1} - (J_k)^{\mathrm{T}} r_{k+1}. \end{aligned}$$

大残量问题的拟牛顿算法

- 拟牛顿条件为：

$$T_{k+1}s_k = (J_{k+1})^{\mathrm{T}} r_{k+1} - (J_k)^{\mathrm{T}} r_{k+1}$$

- Dennis, Gay, 和Welsch给出的一种更新格式为：

$$T_{k+1} = T_k + \frac{\left(y^{\#} - T_k s_k\right) y^T + y\left(y^{\#} - T_k s_k\right)^T}{y^T s_k} - \frac{\left(y^{\#} - T_k s_k\right)^T s_k}{\left(y^T s\right)^2} y y^T$$

其中

$$\begin{aligned} s_k &= x_{k+1} - x_k \\ y &= J_{k+1}^T r_{k+1} - J_k^T r_k \\ y^{\#} &= J_{k+1}^T r_{k+1} - J_k^T r_{k+1} \end{aligned}$$