

拟牛顿法、最小二乘问题

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：本教案由丁思哲、张轩熙协助准备

- 1 拟牛顿矩阵
- 2 拟牛顿类算法的收敛性和收敛速度
- 3 有限内存BFGS方法
- 4 非线性最小二乘问题

割线方程的推导

设 $f(x)$ 是二阶连续可微函数. 对 $\nabla f(x)$ 在点 x^{k+1} 处一阶泰勒近似, 得

$$\nabla f(x) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x - x^{k+1}) + \mathcal{O}(\|x - x^{k+1}\|^2),$$

令 $x = x^k$, 且 $s^k = x^{k+1} - x^k$ 为点差, $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ 为梯度差, 得

$$\nabla^2 f(x^{k+1})s^k + \mathcal{O}(\|s^k\|^2) = y^k.$$

现忽略高阶项 $\|s^k\|^2$, 只希望近似海瑟矩阵的矩阵 B^{k+1} 满足方程

$$B^{k+1}s^k = y^k,$$

或其逆矩阵 H^{k+1} 满足

$$H^{k+1}y^k = s^k.$$

上述两个方程即称为割线方程.

曲率条件

由于近似矩阵必须保证迭代收敛, 正如牛顿法要求海瑟矩阵正定, B^k 正定也是必须的, 即有必要条件

$$(s^k)^T B^{k+1} s^k > 0 \implies (s^k)^T y^k > 0,$$

定义

曲率条件 在迭代过程中满足 $(s^k)^T y^k > 0, \forall k \in \mathbb{N}^+$.

如果线搜索使用 **Wolfe 准则**:

$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k,$$

其中 $c_2 \in (0, 1)$. 上式即 $\nabla f(x^{k+1})^T s^k \geq c_2 \nabla f(x^k)^T s^k$. 在不等式两边同时减去 $\nabla f(x^k)^T s^k$, 由于 $c_2 - 1 < 0$ 且 s^k 是下降方向, 因此最终有

$$(y^k)^T s^k \geq (c_2 - 1) \nabla f(x^k)^T s^k > 0.$$

拟牛顿算法的基本框架

拟牛顿算法的基本框架为：

算法 1 拟牛顿算法框架

Require: 初始坐标 $x^0 \in \mathbb{R}^n$, 初始矩阵 $B^0 \in \mathbb{R}^{n \times n}$ (或 H^0), $k = 0$.

Ensure: x^K, B^K (或 H^K).

- 1: 检查初始元素.
 - 2: **while** 未达到停机准则 **do**
 - 3: 计算方向 $d^k = -(B^k)^{-1} \nabla f(x^k)$ 或 $d^k = -H^k \nabla f(x^k)$.
 - 4: 通过线搜索(Wolfe)产生步长 $\alpha_k > 0$, 令 $x^{k+1} = x^k + \alpha_k d^k$.
 - 5: 更新海瑟矩阵的近似矩阵 B^{k+1} 或其逆矩阵 H^{k+1} .
 - 6: $k \leftarrow k + 1$.
 - 7: **end while**
-

秩一更新(SR1)

定义

秩一更新 对于拟牛顿矩阵 $B^k \in \mathbb{R}^{n \times n}$, 设 $0 \neq u \in \mathbb{R}^n$ 且 $a \in \mathbb{R}$ 待定, 则 uu^T 是秩一矩阵, 且有秩一更新

$$B^{k+1} = B^k + auu^T.$$

根据割线方程 $B^{k+1}s^k = y^k$, 代入秩一更新的结果, 得到

$$(B^k + auu^T)s^k = y^k,$$

整理得

$$auu^T s^k = (a \cdot u^T s^k)u = y^k - B^k s^k.$$

由于 $a \cdot u^T s^k$ 是标量, 因此上式表明 u 和 $y^k - B^k s^k$ 同向. 简单考虑不妨就令 u 和 $y^k - B^k s^k$ 相等, 即 $u = y^k - B^k s^k$. 代入上式得

$$(a \cdot (y^k - B^k s^k)^T s^k)(y^k - B^k s^k) = y^k - B^k s^k.$$

秩一更新公式

再令 $(a \cdot (y^k - B^k s^k)^T s^k) \neq 0$, 则可以确定 a 为

$$a = \frac{1}{(y^k - B^k s^k)^T s^k}.$$

由同样的过程可以推出基于 H^k 的秩一更新公式.

定理

拟牛顿算法的秩一更新公式 拟牛顿矩阵 B^k 的秩一更新公式为

$$B^{k+1} = B^k + \frac{uu^T}{u^T s^k}, \quad u = y^k - B^k s^k,$$

拟牛顿矩阵 H^k 的秩一更新公式为

$$H^{k+1} = H^k + \frac{vv^T}{v^T y^k}, \quad v = s^k - H^k y^k.$$

B^k 和 H^k 的公式在形式上互为对偶. 实际上 $H^k = (B^k)^{-1}$, 利用秩一更新的SMW公式即可推出基于 H^k 的公式, 反之亦然.

秩一更新公式的缺陷

即使 B^k 正定，由秩一公式更新的 B^{k+1} 无法保证正定。

定理

秩一更新公式使 B^{k+1} 正定的充分条件 使用秩一更新公式从 B^k 更新 B^{k+1} ， B^{k+1} 正定的充分条件可以是：

- (1) B^k 正定；
- (2) $u^T s^k > 0$ 。

证明：设 $0 \neq w \in \mathbb{R}^n$ ，则

$$w^T B^{k+1} w = w^T B^k w + \frac{w^T u u^T w}{u^T s^k} = w^T B^k w + \frac{(u^T w)^2}{u^T s^k} > 0.$$

同样地，将上述定理中 B 换成 H ， $u^T s^k$ 换成 $v^T y^k$ ，仍然成立。因此，由于无法保证 $u^T s^k$ 或 $v^T y^k$ 恒大于0，上述的秩一更新公式一般不用。

BFGS公式

BFGS公式的核心思想是对 B^k 进行秩二更新.

定义

秩二更新 对于拟牛顿矩阵 $B^k \in \mathbb{R}^{n \times n}$, 设 $0 \neq u, v \in \mathbb{R}^n$ 且 $a, b \in \mathbb{R}$ 待定, 则有秩二更新形式

$$B^{k+1} = B^k + auu^T + bvv^T.$$

根据割线方程, 将秩二更新的待定参量式代入, 得

$$B^{k+1}s^k = (B^k + auu^T + bvv^T)s^k = y^k,$$

整理可得

$$(a \cdot u^T s^k)u + (b \cdot v^T s^k)v = y^k - B^k s^k.$$

简单的取法是令 $(a \cdot u^T s^k)u$ 对应 y^k 相等, $(b \cdot v^T s^k)v$ 对应 $-B^k s^k$ 相等, 即有

$$a \cdot u^T s^k = 1, \quad u = y^k, \quad b \cdot v^T s^k = -1, \quad v = B^k s^k.$$

BFGS公式

将上述参量代入割线方程, 即得**BFGS更新公式**

$$B^{k+1} = B^k + \frac{uu^T}{(s^k)^T u} - \frac{vv^T}{(s^k)^T v}.$$

利用SMW公式以及 $H^k = (B^k)^{-1}$, 可以推出关于 H^k 的BFGS公式.

定义

BFGS公式 在拟牛顿类算法中, 基于 B^k 的BFGS公式为

$$B^{k+1} = B^k + \frac{y^k (y^k)^T}{(s^k)^T y^k} - \frac{B^k s^k (B^k s^k)^T}{(s^k)^T B^k s^k},$$

基于 H^k 的BFGS公式为

$$H^{k+1} = \left(I - \frac{s^k (y^k)^T}{(s^k)^T y^k} \right)^T H^k \left(I - \frac{s^k (y^k)^T}{(s^k)^T y^k} \right) + \frac{s^k (s^k)^T}{(s^k)^T y^k}.$$

推导 H^k 的BFGS公式之提示

对于可逆矩阵 $B \in \mathbb{R}^{n \times n}$ 与矩阵 $U \in \mathbb{R}^{n \times m}$, $V \in \mathbb{R}^{n \times m}$, SMW公式为:

$$(B + UV^T)^{-1} = B^{-1} - B^{-1}U(I + V^TB^{-1}U)^{-1}V^TB^{-1}.$$

在BFGS的推导中, 关于 B^k 的更新公式为:

$$B_{k+1} = B_k + \frac{y_k y_k^T}{s_k^T y_k} - \frac{B_k s_k (B_k s_k)^T}{s_k^T B_k s_k} = B_k + \begin{pmatrix} -\frac{B_k s_k}{s_k^T B_k s_k} & \frac{y_k}{s_k^T y_k} \end{pmatrix} \begin{pmatrix} s_k^T B_k \\ y_k^T \end{pmatrix}.$$

对照SMW公式, 令式中 $B = B_k$, 且

$$U_k = \begin{pmatrix} -\frac{B_k s_k}{s_k^T B_k s_k} & \frac{y_k}{s_k^T y_k} \end{pmatrix}, \quad V_k = (B_k s_k \quad y_k),$$

此时公式的左端就等于 B_{k+1}^{-1} , 且右端只需计算一个2阶矩阵的逆. 假设 $B_k^{-1} = H_k$, 由SMW公式就得到

$$H_{k+1} = (B_k + U_k V_k^T)^{-1} = \left(I - \frac{s_k y_k^T}{s_k^T y_k} \right) H_k \left(I - \frac{y_k s_k^T}{s_k^T y_k} \right) + \frac{s_k s_k^T}{s_k^T y_k}.$$

BFGS公式的有效性

BFGS公式产生的 B^{k+1} 或 H^{k+1} 是否正定呢？

定理

BFGS公式使拟牛顿矩阵正定的充分条件 使用秩二更新公式从 B^k 或 H^k 更新 B^{k+1} 或 H^{k+1} , 拟牛顿矩阵正定的充分条件可以是:

- (1) B^k 或 H^k 正定;
- (2) 满足曲率条件 $(s^k)^T y^k > 0, \forall k \in \mathbb{N}^+$.

证明上述定理, 只需要从基于 H^k 的BFGS公式分析即可, 从而得到 H^{k+1} 和其逆 B^{k+1} 均正定.

因为在确定步长时使用某一Wolfe准则线搜索即可满足曲率条件, 因此BFGS公式产生的拟牛顿矩阵有望保持正定, 是有效算法.

从优化意义理解BFGS格式

基于 H^k 的BFGS格式恰好是优化问题

$$\begin{aligned} \min_H \quad & \mathbf{OPT} = \|H - H^k\|_W, \\ \text{s.t.} \quad & H = H^T, \\ & Hy^k = s^k. \end{aligned}$$

的解. 上式中 $\|\cdot\|_W$ 是加权范数, 定义为

$$\|H\|_W = \left\| W^{1/2} H W^{1/2} \right\|_F,$$

且 W 满足割线方程, 即 $W s^k = y^k$.

注意 $Hy^k = s^k$ 是割线方程, 因此优化问题的意义是在满足割线方程的**对称矩阵**中找到距离 H^k 最近的矩阵 H 作为 H^{k+1} . 因此我们可以进一步认知, BFGS格式更新的拟牛顿矩阵是正定对称的, 且在满足割线方程的条件下采取的是最佳逼近策略.

DFP公式

DFP公式利用与BFGS公式类似的推导方法,不同的是其以割线方程 $H^{k+1}y^k = s^k$ 为基础进行对 H^k 的秩二更新.

基于 H^k 满足的DFP公式,利用SMW公式以及 $B^k = (H^k)^{-1}$,可以推出关于 B^k 的DFP公式. (关键的推导步骤仍然可以参考推导BFGS公式时给出的提示)

定义

DFP公式 基于 H^k 的DFP更新公式为

$$H^{k+1} = H^k - \frac{H^k y^k (H^k y^k)^T}{(y^k)^T H^k y^k} + \frac{s^k (s^k)^T}{(y^k)^T s^k},$$

基于 B^k 的DFP更新公式为

$$B^{k+1} = \left(I - \frac{y^k (s^k)^T}{(s^k)^T y^k}\right)^T B^k \left(I - \frac{y^k (s^k)^T}{(s^k)^T y^k}\right) + \frac{y^k (y^k)^T}{(s^k)^T y^k}.$$

从优化意义上理解DFP公式

有了BFGS公式的优化意义做铺垫, 讨论DFP公式的优化意义显得十分简单. 利用对偶性质, 基于 B^k 的DFP格式将是优化问题

$$\begin{aligned} \min_B \quad & \mathbf{OPT} = \|B - B^k\|_W, \\ \text{s.t.} \quad & B = B^T, \\ & Bs^k = y^k. \end{aligned}$$

的解. 上式中 $\|\cdot\|_W$ 是加权范数, 定义为

$$\|B\|_W = \left\| W^{1/2} B W^{1/2} \right\|_F,$$

且 W 满足另一割线方程, 即 $Wy^k = s^k$.

注意 $Bs^k = y^k$ 是另一割线方程, 因此优化问题的意义是在满足割线方程的**对称矩阵**中找到距离 B^k 最近的矩阵 B 作为 B^{k+1} .

DFP公式的缺陷

尽管DFP格式与BFGS对偶,但从实际效果而言,DFP格式的求解效率整体上不如BFGS格式. M.J.D. Powell曾求解问题

$$\min_{x \in \mathbb{R}^2} f(x) = \frac{1}{2} \|x\|_2^2.$$

设置初始值

$$B^0 = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}, \quad x_1 = \begin{pmatrix} \cos \psi \\ \sin \psi \end{pmatrix},$$

其中 $\tan^2 \psi = \lambda$. 当误差阈 $\epsilon = 10^{-4}$ 时,分别取 λ 为不同的值,使用BFGS算法与DFP算法所产生的迭代步数分别如下表(见下页)所示. 由此看出,在本问题中,BFGS算法的求解效率要远高于DFP算法.

(参考文献: Powell M J D. How bad are the BFGS and DFP methods when the objective function is quadratic?[J]. Mathematical Programming, 1986, 34(1): 34-47.)

DFP公式的缺陷

Table: BFGS方法的迭代次数

$\lambda \backslash \epsilon$	0.1	0.01	10^{-4}	10^{-8}
10	5	6	8	10
100	7	8	10	12
10^4	12	13	15	17
10^6	17	18	20	22
10^9	24	25	27	29

Table: DFP方法的迭代次数

$\lambda \backslash \epsilon$	0.1	0.01	10^{-4}	10^{-8}
10	10	13	16	19
30	25	32	37	40
100	80	99	107	111
300	237	290	307	313
10^3	787	958	1006	1014

- 1 拟牛顿矩阵
- 2 拟牛顿类算法的收敛性和收敛速度
- 3 有限内存BFGS方法
- 4 非线性最小二乘问题

BFGS全局收敛性

我们利用Zoutendijk条件得到基本收敛性. 需要复习的读者可参看纸质本Page 213的定理6.1.

根据对BFGS格式有效性的分析, 我们先确保初始矩阵 B^0 是对称正定的.

定理

BFGS全局收敛性 设初始矩阵 B^0 是对称正定矩阵, 目标函数 $f(x)$ 是二阶连续可微函数, 下水平集

$$\mathcal{L} = \{x \in \mathbb{R}^n | f(x) \leq f(x^0)\}$$

凸, 且存在 $m, M \in \mathbb{R}^+$ 使得对 $\forall z \in \mathbb{R}^n, x \in \mathcal{L}$ 满足

$$m \|z\|^2 \leq z^T \nabla^2 f(x) z \leq M \|z\|^2$$

(即 $z^T \nabla^2 f(x) z$ 被 $\|z\|$ 控制), 那么BFGS格式结合Wolfe线搜索的拟牛顿算法全局收敛到 $f(x)$ 的极小值点 x^* .

BFGS全局收敛性的证明

通过Zoutendijk条件来证明收敛性. 因为BFGS拟牛顿矩阵在定理条件下对称正定, 每个拟牛顿方向是下降方向. 因此, 只需要证明**搜索方向与负梯度的夹角不太差**.

基于 B^k 的BFGS格式为

$$B^{k+1} = B^k + \frac{y^k (y^k)^T}{(s^k)^T y^k} - \frac{B^k s^k (B^k s^k)^T}{(s^k)^T B^k s^k},$$

通过这一公式, 我们可以说明:

- (a) $\text{Tr}(B^{k+1}) = \text{Tr}(B^k) - \frac{\|B^k s^k\|^2}{(s^k)^T B^k s^k} + \frac{\|y^k\|^2}{(s^k)^T y^k},$
- (b) $\det(B^{k+1}) = \det(B^k) \frac{(s^k)^T y^k}{(s^k)^T B^k s^k}.$

关于迹的公式是显然的, 因为迹的运算保和. 我们主要说明为何关于行列式的结论成立(习题6.11).

BFGS全局收敛性的证明

为说明(b)式成立, 先证明结论: 设 $x, y, u, v \in \mathbb{R}^n$, 则

$$\det(I_{n \times n} + xy^T + uv^T) = (1 + y^T x)(1 + v^T u) - (x^T v)(y^T u).$$

Proof: 令 $U = (x, u)$, $V = (y^T; v^T)$ 。构造分块矩阵 $A = (I_{n \times n}, U; V, I_{2 \times 2})$ 。利用Schur补的性质成立

$$\det(I_{n \times n} + xy^T + uv^T) = \det(I_{n \times n} + UV) = \det(I_{2 \times 2} + U^T V^T).$$

利用上述定理, 将BFGS公式的左右两边乘以 $(B^k)^{-1}$ (B^k 正定), 并

设 $x = \frac{-s^k}{\sqrt{(s^k)^T B^k s^k}}$, $y = \frac{B^k s^k}{\sqrt{(s^k)^T B^k s^k}}$, $u = \frac{(B^k)^{-1} y^k}{\sqrt{(y^k)^T s^k}}$, $v = \frac{y^k}{\sqrt{(y^k)^T s^k}}$, 代入即可证明结论成立。

BFGS全局收敛性的证明

定义 $\cos \theta_k = \frac{(s^k)^T B^k s^k}{\|s^k\| \|B^k s^k\|}$ 是欲求夹角的余弦. 这是因为

$$\begin{aligned} s^k &= x^{k+1} - x^k = -\alpha_k (B^k)^{-1} \nabla f(x^k), \\ B^k s^k &= -\alpha_k \nabla f(x^k). \end{aligned}$$

再设

$$q_k = \frac{(s^k)^T B^k s^k}{(s^k)^T s^k}, \quad m_k = \frac{(y^k)^T s^k}{(s^k)^T s^k}, \quad M_k = \frac{(y^k)^T y^k}{(y^k)^T s^k},$$

将上述定义代入**(b)**式以及余弦式, 得到

$$\text{(c)} \quad \det(B^{k+1}) = \det(B^k) \frac{m_k}{q_k}.$$

$$\text{(d)} \quad \frac{\|B^k s^k\|^2}{(s^k)^T B^k s^k} = \frac{q_k}{\cos^2 \theta_k}.$$

上述**(a)**,**(b)**,**(c)**,**(d)**均是准备公式.

BFGS全局收敛性的证明

我们目标是构造一个不等式, 使得 $\cos \theta_k > 0, k \rightarrow \infty$. 设

$$\Psi(B) = \text{Tr}(B) - \ln(\det(B)),$$

注意上式成立 $\Psi(B) > 0$.

在上式中代入 $B = B^{k+1}$ 以及上述准备公式, 成立

$$\begin{aligned}\Psi(B^{k+1}) &= \text{Tr}(B^{k+1}) - \ln(\det(B^{k+1})) \\ &\leq \Psi(B^k) + (M_k - \ln(m_k) - 1) + 2 \ln(\cos \theta_k).\end{aligned}$$

同时注意到 $m_k \geq m, M_k \leq M$, 所以又有

$$\begin{aligned}\Psi(B^{k+1}) &\leq \Psi(B^k) + (M_k - \ln(m_k) - 1) + 2 \ln(\cos \theta_k) \\ &\leq \Psi(B^0) + (k+1)(M - \ln(m) - 1) + 2 \sum_{j=0}^k \ln(\cos \theta_j).\end{aligned}$$

如果我们假设迭代将不会停止, 则右式若无界, 将导出矛盾.

BFGS全局收敛性的证明

不妨设 $\cos\theta_k \rightarrow 0$, 因此有 $\ln(\cos^2\theta_k) \rightarrow -\infty$. 这意味着, 存在 $K \in \mathbb{N}^+$, 使得对 $\forall j \geq K$, 均成立

$$\ln(\cos^2\theta_j) < -2(M - \ln(m) - 1) = -2C < 0,$$

联立上述两个最近的控制式, 注意 $\Psi(B^{k+1}) > 0$, 则有 $k \rightarrow \infty$ 时

$$0 < \Psi(B^0) + (k+1)C + 2 \sum_{j=0}^K \ln(\cos\theta_j) + \sum_{j=K+1}^k (-2C) \rightarrow -\infty,$$

这就导出了矛盾. 因此 $\cos\theta_k \rightarrow 0$ 是不成立的. 换句话说, 存在子列 $\{j_k\}_{k=1,2,\dots}$, 使得 $\cos\theta_{j_k} \geq \delta > 0$. 根据Zoutendijk条件, 又可以得到

$$\liminf_{k \rightarrow \infty} \|\nabla f(x^k)\| \rightarrow 0.$$

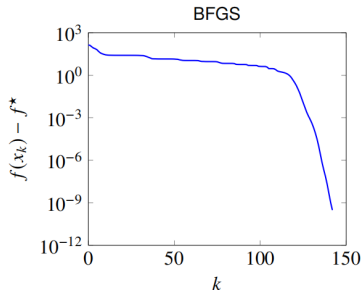
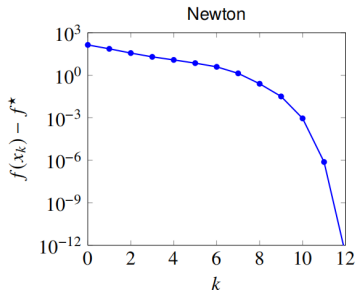
结合上述问题对 $x \in \mathcal{L}$ 是强凸的, 所以导出 $x^k \rightarrow x^*$.

BFGS方法的收敛速度之例

例 考虑极小化问题

$$\min_{x \in \mathbb{R}^{100}} c^T x - \sum_{i=1}^{500} \ln(b_i - a_i^T x),$$

下图展示了误差 $f(x_k) - f^*$ 与迭代次数 k 之间的关系(k 是迭代次数). 虽然BFGS方法的迭代次数显著得多, 但由于牛顿法每次迭代的计算代价为 $O(n^3)$ 加上**计算海瑟矩阵的代价**, 而BFGS方法的每步计算代价仅为 $O(n^2)$, 因此BFGS算法可能更快取得优势解.



- 1 拟牛顿矩阵
- 2 拟牛顿类算法的收敛性和收敛速度
- 3 有限内存BFGS方法
- 4 非线性最小二乘问题

有限内存方法的基本思路

基本思路 标准的拟牛顿近似矩阵的更新公式可以记为

$$B^{k+1} = g(B^k, s^k, y^k), \quad s^k = x^{k+1} - x^k, \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k).$$

如果只保存最近的 m 组数据, 那么迭代公式可以写成

$$B^{k+1} = g(g(\dots g(B^{k-m+1}, s^{k-m+1}, y^{k-m+1}))).$$

考虑BFGS方法:

$$d^k = -(B^k)^{-1} \nabla f(x^k) = -H^k \nabla f(x^k).$$

重写BFGS更新公式为

$$H^{k+1} = (V^k)^T H^k V^k + \rho_k s^k (s^k)^T,$$

其中

$$\rho_k = \frac{1}{(y^k)^T s^k}, \quad V^k = I_{n \times n} - \rho_k y^k (s^k)^T.$$

有限内存BFGS方法

将上式递归地展开 m 次, 即

$$\begin{aligned} H^k &= \left(\prod_{j=k-m}^{k-1} V^j \right)^T H^{k-m} \left(\prod_{j=k-m}^{k-1} V^j \right) + \\ &\rho_{k-m} \left(\prod_{j=k-m+1}^{k-1} V^j \right)^T s^{k-m} (s^{k-m})^T \left(\prod_{j=k-m}^{k-1} V^j \right) + \dots + \\ &\rho_{k-1} s^{k-1} (s^{k-1})^T. \end{aligned}$$

为了节省内存, 我们只展开 m 次, 利用 H^{k-m} 进行计算, 即可求出 H^{k+1} .

下面介绍一种不计算 H^k , 只利用展开式计算 $d^k = -H^k \nabla f(x^k)$ 的巧妙算法: **双循环递归算法**. 它利用迭代式的结构尽量节省计算 d^k 的开销.

有限内存BFGS方法

将等式两边同右乘 $\nabla f(x^k)$, 则等式左侧为 $-d^k$. 观察等式右侧需要计算

$$V^{k-1}\nabla f(x^k), \dots, V^{k-m} \dots V^{k-1}\nabla f(x^k).$$

这些计算可以递归地进行. 同时在计算 $V^{k-l} \dots V^{k-1}\nabla f(x^k)$ 的过程中, 可以计算上一步的 $\rho_{k-l}(s^{k-l})^T[V^{k-l+1} \dots V^{k-1}\nabla f(x^k)]$, 这是一个标量. 记

$$q = V^{k-m} \dots V^{k-1}\nabla f(x^k),$$

$$\alpha_{k-l} = \rho_{k-l}(s^{k-l})^T[V^{k-l+1} \dots V^{k-1}\nabla f(x^k)],$$

因此递归公式可化为如下的形式:

$$H^k \nabla f(x^k) = \left(\prod_{j=k-m}^{k-1} V^j \right)^T H^{k-m} q + \left(\prod_{j=k-m+1}^{k-1} V^j \right)^T s^{k-m} \alpha_{k-m} + \dots + s^{k-1} \alpha_{k-1}$$

有限内存BFGS方法

在双循环递归算法中,除了上述第一个循环递归过程(自下而上)外,还有以下第二个循环递归过程.我们需要在公式中自上而下合并每一项.以前两项为例,它们有公共的因子 $(V^{k-m+1} \dots V^{k-1})^T$,提取后可以将前两项写为(注意将 V^{k-m} 的定义回代)

$$\begin{aligned} & (V^{k-m+1} \dots V^{k-1})^T \left[(V^{k-m})^T r + \alpha_{k-m} s^{k-m} \right] \\ &= (V^{k-m+1} \dots V^{k-1})^T (r + (\alpha_{k-m} - \beta) s^{k-m}), \end{aligned}$$

这正是第二个循环的迭代格式.注意合并后原递归式的结构仍不变,因此可以递归地计算下去.最后,变量 r 就是我们期望的结果 $H^k \nabla f(x^k)$.

L-BFGS双循环递归算法

拟牛顿算法的基本框架为：

算法 2 L-BFGS双循环递归

Require: 初始化 $q \leftarrow \nabla f(x^k)$.

Ensure: r , 即 $H^k \nabla f(x^k)$.

- 1: 检查初始元素.
 - 2: **for** $i = k - 1, \dots, k - m$ **do**
 - 3: 计算并保存 $\alpha_i \leftarrow \rho_i (s^i)^\top q$.
 - 4: 更新 $q \leftarrow q - \alpha_i y^i$.
 - 5: **end for**
 - 6: 初始化 $r \leftarrow \hat{H}^{k-m} q$, 其中 \hat{H}^{k-m} 是 H^{k-m} 的近似矩阵.
 - 7: **for** $i = k - m, \dots, k - 1$ **do**
 - 8: 计算 $\beta \leftarrow \rho_i (y^i)^\top r$.
 - 9: 更新 $r \leftarrow r + (\alpha_i - \beta) s^i$.
 - 10: **end for**
-

算法分析

L-BFGS双循环递归算法约需要 $4mn$ 次乘法运算, $2mn$ 次加法运算; 若近似矩阵 \hat{H}^{k-m} 是对角矩阵, 则额外需要 n 次乘法运算. 由于 m 不会很大, 因此算法的复杂度是 $\mathcal{O}(mn)$. 算法需要的额外存储为临时变量 α_i , 其大小是 $\mathcal{O}(m)$.

\hat{H}^{k-m} 的一种取法可以是取对角矩阵

$$\hat{H}^{k-m} = \gamma_k I_{n \times n} \triangleq \frac{(s^{k-1})^T y^{k-1}}{(y^{k-1})^T y^{k-1}} I_{n \times n}.$$

这恰好是BB方法的第一个步长.

- 1 拟牛顿矩阵
- 2 拟牛顿类算法的收敛性和收敛速度
- 3 有限内存BFGS方法
- 4 非线性最小二乘问题

最小二乘问题

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

- 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 $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$

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

问题可以表述为 $\min f(x) = \frac{1}{2} \|r(x)\|_2^2$

- 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信赖域法求解

最小二乘问题

- 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：

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

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

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

高斯-牛顿方法

- 使用近似 $\nabla^2 f_k \approx J_k^T J_k$. 省略 $\nabla^2 r_j$ 的计算, 减少了计算量。其中 J_k 、 r_k 分别是 $J(x_k)$ 、 $r(x_k)$ 的简写。
- 高斯-牛顿法的迭代方向 d_k^{GN} 满足:

$$J_k^T J_k d_k^{GN} = -J_k^T r_k. \quad (3)$$

- 另一种理解: 在点 x_k 处, 考虑近似 $r(x_k + d) \approx r_k + J_k d$ 得到:

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

求解该问题时也可对 J_k 做QR分解或SVD分解, 无需计算出 $J_k^T J_k$ 。

- 然后更新 $x_{k+1} = x_k + \alpha_k d_k$.

Levenberg-Marquardt (LM) 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

$$\min_d \frac{1}{2} \|J^k d + r^k\|^2, \quad \text{s.t.} \quad \|d\| \leq \Delta_k. \quad (4)$$

- LM 方法将如下近似当作信赖域方法中的 m_k :

$$m_k(d) = \frac{1}{2} \|r^k\|^2 + d^T (J^k)^T r^k + \frac{1}{2} d^T (J^k)^T J^k d. \quad (5)$$

- 同样使用 $(J^k)^T J^k$ 来近似海瑟矩阵.

Levenberg-Marquardt 方法

- 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：

$$\rho_k = \frac{f(x^k) - f(x^k + d^k)}{m_k(0) - m_k(d^k)} \quad (6)$$

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 3 Levenberg-Marquardt 方法

- 1: 给定最大半径 Δ_{\max} , 初始半径 Δ_0 , 初始点 x^0 , $k \leftarrow 0$.
- 2: 给定参数 $0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1$, $\gamma_1 < 1 < \gamma_2$.
- 3: **while** 未达到收敛准则 **do**
- 4: 计算子问题(4)得到迭代方向 d^k .
- 5: 根据(6) 计算下降率 ρ_k .
- 6: 更新信赖域半径:

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

- 7: 更新自变量:

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

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

Corollary

向量 d^* 是信赖域子问题

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

的解当且仅当 d^* 是可行解并且存在数 $\lambda \geq 0$ 使得

$$(J^T J + \lambda I)d^* = -J^T r, \quad (7)$$

$$\lambda(\Delta - \|d^*\|) = 0. \quad (8)$$

- 问题(7)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

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

子问题求解

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

- 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(4)的近似解 d_k 满足

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

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

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

- 根据 $r_j(x)$ 的连续性，存在 $M > 0$ ，使得 $\|J_k^T J_k\| \leq M$ 对任意的 k 成立。由于 f 有下界，可以直接套用信赖域算法全局收敛性的证明。

- 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

$$(J^T J + \lambda I) d = -J^T r$$

- 调整 λ 的大小等价于调整信赖域半径的大小， Δ 被 λ 隐式决定。
- LM的更新基于 Δ ，LMF的更新直接基于 λ ，每一步求解子问题：

$$\min_d \|Jd + r\|_2^2 + \lambda \|d\|_2^2.$$

- 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

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)} \quad (9)$$

较大可以减小下一步的 λ ，较小可以增大下一步的 λ 。

Algorithm 4 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 $0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1$, $\gamma_1 < 1 < \gamma_2$.
- 3: **while** 未达到收敛准则 **do**
- 4: 求解LM方程 $((J_k)^T J_k + \lambda I)d = -(J_k)^T r_k$ 得到迭代方向 d_k .
- 5: 根据(9)式计算下降率 ρ_k .
- 6: 更新乘子:

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

- 7: 更新自变量:

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

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

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：

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

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

$$B_k = J_k^T J_k + T_k,$$

- 目标为

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

- 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

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

大残量问题的拟牛顿算法

- 拟牛顿条件为：

$$T_{k+1}s_k = J_{k+1}^T r_{k+1} - J_k^T r_{k+1}$$

- Dennis, Gay, 和Welsch给出的一种更新格式为：

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

其中

$$s_k = x_{k+1} - x_k$$

$$y = J_{k+1}^T r_{k+1} - J_k^T r_k$$

$$y^\# = J_{k+1}^T r_{k+1} - J_k^T r_{k+1}$$

大残量问题的拟牛顿算法2

- 原始问题的拟牛顿公式：

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

- 改写上式得到

$$T_{k+1}s_k = \tilde{y}_k,$$

其中

$$\tilde{y}_k = J_{k+1}^T r_{k+1} - J_k^T r_k - J_{k+1}^T J_{k+1} s_k.$$

- 再利用拟牛顿公式更新