拟牛顿法

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：本教案由丁思哲协助准备

提纲

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

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

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

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

现忽略高阶项$\left\|s^k\right\|^2$, 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

$$\left(s^k\right)^{\mathrm{T}} B^{k+1} s^k > 0 \Longrightarrow \left(s^k\right)^{\mathrm{T}} y^k > 0,$$

定义

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

如果线搜索使用Wolfe准则:

$$\nabla f(x^k + \alpha d^k)^{\mathrm{T}} d^k \geqslant c_2 \nabla f(x^k)^{\mathrm{T}} d^k,$$

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

$$(y^k)^{\mathrm{T}} s^k \geqslant (c_2 - 1) \nabla f(x^k)^{\mathrm{T}} s^k > 0.$$

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$uu^{\mathrm{T}}$是秩一矩阵, 且有秩一更新

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

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

$$(B^k + auu^{\mathrm{T}})s^k = y^k,$$

整理得

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

由于$a \cdot u^{\mathrm{T}}s^k$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

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

秩一更新公式

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

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

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

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

拟牛顿矩阵H^k的秩一更新公式为

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

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) $u^{\mathrm{T}}s^k > 0$.

证明: 设$0 \neq w \in \mathbb{R}^n$, 则

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

同样地, 将上述定理中B换成H, $u^{\mathrm{T}}s^k$换成$v^{\mathrm{T}}y^k$, 仍然成立. 因此, 由于无法保证$u^{\mathrm{T}}s^k$或$v^{\mathrm{T}}y^k$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

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

根据割线方程, 将秩二更新的待定参量式代入, 得

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

整理可得

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

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

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

BFGS公式

将上述参量代入割线方程, 即得BFGS更新公式

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

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

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

基于H^k的*BFGS*公式为

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

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：

$$\left(B + UV^{\mathrm{T}}\right)^{-1} = B^{-1} - B^{-1}U\left(I + V^{\mathrm{T}}B^{-1}U\right)^{-1}V^{\mathrm{T}}B^{-1}.$$

在BFGS的推导中，关于B^k的更新公式为：

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

对照SMW公式，令式中$B = B_k$，且

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

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

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

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$\left(s^k\right)^{\mathrm{T}} y^k > 0, \forall k \in \mathbb{N}^+$.

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

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

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$Ws^k = y^k$.
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

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

基于B^k的*DFP*更新公式为

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

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

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

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

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

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

λ \ ϵ	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

λ \ ϵ	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

提纲

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

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

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

$$m \|z\|^2 \leqslant z^{\mathrm{T}} \nabla^2 f(x) z \leqslant M \|z\|^2$$

(即$z^{\mathrm{T}} \nabla^2 f(x) z$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

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

通过这一公式, 我们可以说明:

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

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, 先证明一个结论.

定理

设$x, y, u, v \in \mathbb{R}^n$, 则

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

Proof: 只需注意到利用降阶公式成立

$$\begin{aligned}\det\left(I_{n\times n} + xy^{\mathrm{T}} + uv^{\mathrm{T}}\right) &= \det\left(I_{n\times n} + \begin{pmatrix} x & u \end{pmatrix}\begin{pmatrix} y^{\mathrm{T}} \\ v^{\mathrm{T}} \end{pmatrix}\right) \\ &= \det\left(I_{2\times 2} + \begin{pmatrix} x^T \\ u^T \end{pmatrix}\begin{pmatrix} y & v \end{pmatrix}\right).\end{aligned}$$

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

$$s^k = x^{k+1} - x^k = -\alpha_k \left(B^k\right)^{-1} \nabla f\left(x^k\right),$$
$$B^k s^k = -\alpha_k \nabla f\left(x^k\right).$$

再设

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

将上述定义代入(b)式以及余弦式, 得到

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

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

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$\cos\theta_k > 0, k \to \infty$. 设

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

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

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

同时注意到$m_k \geqslant m, M_k \leqslant M$, 所以又有

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

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$cos\theta_k \to 0$，因此有$\ln(\cos^2\theta_k) \to -\infty$. 这意味着, 存在 $K \in \mathbb{N}^+$, 使得对$\forall j \geqslant K$, 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$\Psi\left(B^{k+1}\right) > 0$, 则有$k \to \infty$时

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

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

$$\lim_{k\to\infty}\mathbf{inf}\left\|\nabla f\left(x^k\right)\right\| \to 0.$$

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

BFGS局部收敛性与R-收敛速度

上述全局定理说明在一定假设下, 使用BFGS格式确定下降方向, 搭配Wolfe线搜索确定步长后是全局收敛的. 下面这个定理从局部收敛性给出了其收敛速度.

定理

BFGS局部收敛性 设$f(x)$二阶连续可微, 点列$\{x_k\}$是由*BFGS*格式产生的, 并收敛于x^*, $\nabla^2 f(x^*)$是对称正定矩阵. 设初始矩阵B^0为任意的对称正定矩阵, 那么存在$0 \leqslant c < 1$, $K \in \mathbb{N}^+$, 使得对$\forall k > N$, 成立

$$f\left(x^{k+1}\right) - f\left(x^*\right) \leqslant c^{k-K+1}\left(f\left(x^K\right) - f\left(x^*\right)\right),$$

$$\sum_{k=0}^{\infty} \left\|x^k - x^*\right\| < \infty.$$

上述定理表明, 序列$\left\{f\left(x^k\right) - f\left(x^*\right)\right\}_k$是压缩的, 收敛将具有R-超线性收敛速度.

BFGS格式的Q-收敛速度

由局部收敛性定理, 可以进一步推出BFGS的Q-收敛速度.

定理

BFGS的Q-超线性收敛速度 除要求*BFGS*局部收敛性的假设外, 再要求f的海瑟矩阵在x^*处*Lip*-连续, 则有

$$\lim_{k\to\infty}\frac{\left\|x^{k+1}-x^*\right\|}{\left\|x^k-x^*\right\|}=0.$$

即$\{x^k\}$为*Q*-超线性收敛到x^*.

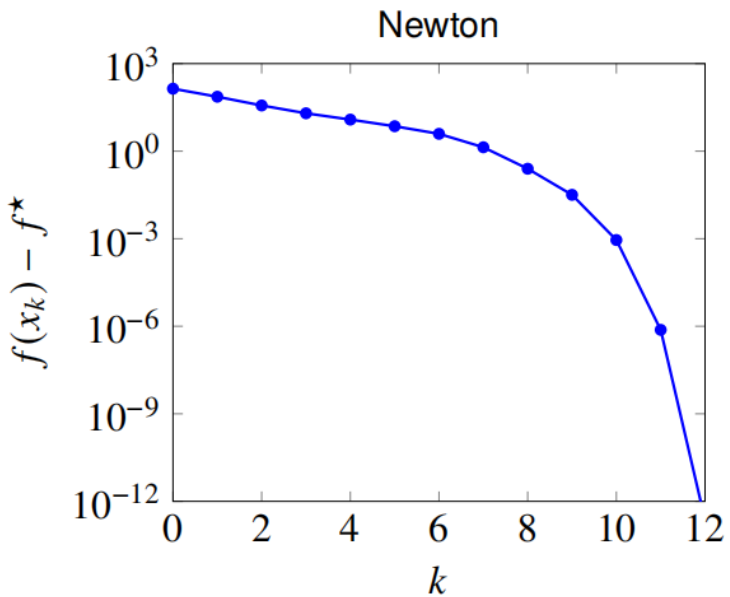
以BFGS格式为代表的拟牛顿类算法由于仅仅使用了海瑟矩阵的近似, 因此很难达到二阶收敛速度, 最多只能达到Q-超线性收敛速度. 但是, 由于拟牛顿方法对近似矩阵的更新代价可能远小于牛顿方法计算海瑟矩阵的代价, 因此它在大规模问题中的开销可能远小于牛顿算法, 较为实用.

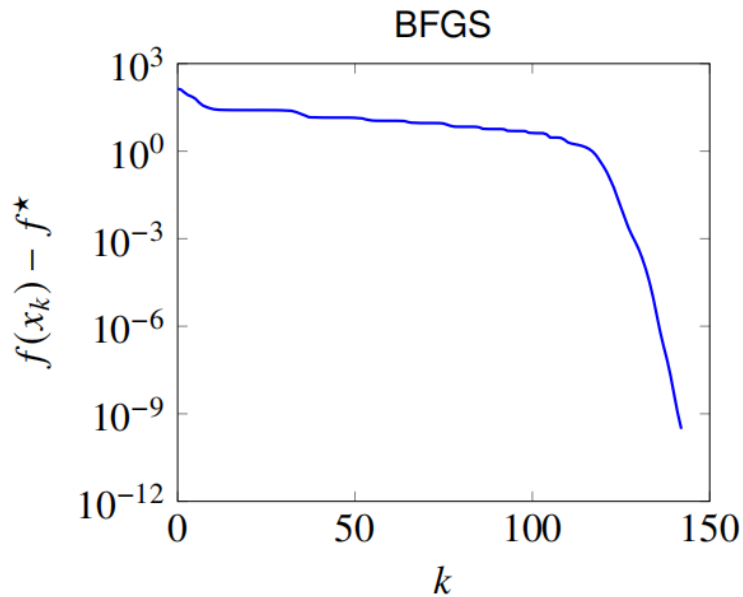
BFGS方法的收敛速度之例

例　考虑极小化问题

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

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$\mathcal{O}(n^3)$加上计算海瑟矩阵的代价, 而BFGS方法的每步计算代价仅为$\mathcal{O}(n^2)$, 因此BFGS算法可能更快取得优势解.





提纲

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

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

如果只保存最近的m组数据, 那么迭代公式可以写成

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

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

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

其中

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

将上式递归地展开m次, 即

$$H^k = \left(\prod_{j=k-m}^{k-1} V^j\right)^{\mathrm{T}} H^{k-m} \left(\prod_{j=k-m}^{k-1} V^j\right) +$$
$$\rho_{k-m} \left(\prod_{j=k-m+1}^{k-1} V^j\right)^{\mathrm{T}} s^{k-m} \left(s^{k-m}\right)^{\mathrm{T}} \left(\prod_{j=k-m}^{k-1} V^j\right) + \cdots +$$
$$\rho_{k-1} s^{k-1} \left(s^{k-1}\right)^{\mathrm{T}}.$$

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

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

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

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

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

因此递归公式可化为如下的形式:

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

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$(V^{k-m+1}\cdots V^{k-1})^{\mathrm{T}}$，提取后可以将前两项写为(注意将$V^{k-m}$的定义回代)

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

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$H^k \nabla f(x^k)$.

检查初始元素.

for $i = k-1, \cdots, k-m$ **do**

  计算并保存$\alpha_i \leftarrow \rho_i (s^i)^{\mathrm{T}} q$.

  更新$q \leftarrow q - \alpha_i y^i$.

end for

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.

for $i = k-m, \cdots, k-1$ **do**

  计算$\beta \leftarrow \rho_i (y^i)^{\mathrm{T}} r$.

  更新$r \leftarrow r + (\alpha_i - \beta) s^i$.

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

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

这恰好是BB方法的第一个步长.

L-BFGS方法

Algorithm 3 L-BFGS方法

Input: 选择初始点x^0, 参数$m > 0$, $k \leftarrow 0$.

Output: 达到收敛准则后的x^{k+1}.

检查初始元素 **while** 未达到收敛准则 **do**

    选取近似矩阵$\hat{H}^{k-m}$;

    使用算法2(L-BFGS双循环递归算法)计算$d^k = -H^k \nabla f(x^k)$;

    使用满足Wolfe准则的线搜索算法确定步长α_k;

    更新$x^{k+1} = x^k + \alpha_k d^k$.

    if $k > m$ **then**

        从内存空间中删除s^{k-m}, y^{k-m}.

    计算并保存$s^k = x^{k+1} - x^k, y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.

    $k \leftarrow k + 1$.

BFGS块迭代格式

实际上, L-BFGS格式下的拟牛顿矩阵的形式可以直接给出, 这为我们分析L-BFGS方法提供了很大的便利. 我们称这类方法为块迭代.
引入记号

$$S^k = \left[s^0, \cdots, s^{k-1}\right], \quad Y^k = \left[y^0, \cdots, y^{k-1}\right],$$

则成立如下定理.

定理

块迭代格式　设H^0为*BFGS*格式的初始矩阵, 它对称正定; 向量对$\left\{s^i, y^i\right\}_{i=0}^{k-1}$满足$(s^i)^{\mathrm{T}} y^i > 0$, H^k是由*BFGS*格式产生的拟牛顿矩阵, 则

$$H^k = H^0 + \begin{bmatrix} S^k & H^0 Y^k \end{bmatrix} \begin{bmatrix} W^k & -\left(\left(R^k\right)^{-1}\right)^{\mathrm{T}} \\ -\left(R^k\right)^{-1} & 0 \end{bmatrix} \begin{bmatrix} \left(S^k\right)^{\mathrm{T}} \\ \left(Y^k\right)^{\mathrm{T}} H^0 \end{bmatrix},$$

BFGS块迭代格式

在上述定理中,

$$W^k = \left(\left(R^k\right)^{-1}\right)^{\mathrm{T}} \left(D^k + \left(Y^k\right)^{\mathrm{T}} H^0 Y^k\right) \left(R^k\right)^{-1},$$

$$\left(R^k\right)_{ij} = \begin{cases} \left(s^{i-1}\right)^{\mathrm{T}} y^{i-1}, & i \leqslant j, \\ 0, & i > j, \end{cases}$$

$$D^k = \mathrm{Diag}\left(\left(s^0\right)^{\mathrm{T}} y^0, \left(s^1\right)^{\mathrm{T}} y^1, \cdots, \left(s^{k-1}\right)^{\mathrm{T}} y^{k-1}\right).$$

上述公式的重要意义在于, 如果给定了H^0, S^k, Y^k, 可以直接计算出BFGS迭代矩阵H^k. 如果H^0是一个近似矩阵, 那么以上定理将给出L-BFGS迭代矩阵的显式格式, 进而求出d^k.

BFGS块迭代格式

使用块迭代格式还有一个优势，那就是可以利用SMW公式(几乎是直接代入即可得)，由H^k所满足的迭代格式直接推出B^k所满足的迭代格式，依据的还是割线方程以及$B^k = (H^k)^{-1}$.

定理

块迭代格式 设B^0为*BFGS*格式的初始矩阵，它对称正定；向量对$\{s^i, y^i\}_{i=0}^{k-1}$满足$(s^i)^{\mathrm{T}} y^i > 0$，$B^k$是由*BFGS*格式产生的拟牛顿矩阵，则

$$B^k = B^0 + \begin{bmatrix} B^0 S^k & Y^k \end{bmatrix} \begin{bmatrix} (S^k)^{\mathrm{T}} B^0 S^k & L^k \\ (L^k)^{\mathrm{T}} & -D^k \end{bmatrix} \begin{bmatrix} (S^k)^{\mathrm{T}} B^0 \\ (Y^k)^{\mathrm{T}} \end{bmatrix},$$

$$(L^k)_{ij} = \begin{cases} (s^{i-1})^{\mathrm{T}} y^{i-1}, & i > j, \\ 0, & i \leqslant j, \end{cases}$$

其他符号的含义同基于H^k的块迭代格式.