3.5 拟牛顿法

牛顿法的主要优点是收敛速度快, 因而倍受欢迎. 但是, 牛顿法也有一个大的缺陷: 需要计算目标函数的Hesse阵的逆矩阵，其计算量比较大，并且有时很难计算，甚至不能计算，这就导致了一个想法：能否利用目标函数值和一阶导数的信息，来逼近Hesse阵的逆矩阵，并且使得方法具有类似牛顿法收敛速度快的优点. 拟牛顿法就是这样的一类算法，由于它不需要计算目标函数的二阶导数，计算量小，适合于求解大规模问题，且算法的收敛速度快，通常比牛顿法更有效.

3.5 拟牛顿法 — 拟牛顿条件

§3.5.1 拟牛顿条件

类似于牛顿法，给出以下迭代公式：

$$x^{k+1} = x^k - \lambda_k H_k g_k, \tag{3.5.1}$$

其中，λ_k为迭代步长. 在(3.5.1)式中, 若令$H_k = G_k^{-1}$, 则(3.5.1) 式为牛顿迭代公式. 拟牛顿法就是利用目标函数值和一阶导数的信息，构造合适的H_k来逼近G_k^{-1}，使得既不需要计算G_k^{-1}，算法又收敛得快. 为此，H_k的选取应满足以下的条件.

(i) H_k是对称正定矩阵. 显然，当H_k是对称正定矩阵时，若$g_k \neq 0$, 则$g_k^{\top}(-H_k g_k) = -g_k^{\top} H_k g_k < 0$, 此时，$d^k = -H_k g_k$为下降方向. 因此，若对任意的$k$，$H_k$是对称正定矩阵，由(3.5.1)式定义的算法为下降算法.

(ii) H_{k+1}由H_k经简单校正而得，$H_{k+1} = H_k + E_k$, 其中，E_k称为校正矩阵，上式称为校正公式. E_k不同, 即可得到不同的算法.

(iii) H_k满足拟牛顿方程: $H_{k+1} y_k = s_k$, 其中$s_k = x^{k+1} - x^k$且$y_k = g_{k+1} - g_k$.

3.5 拟牛顿法 — 拟牛顿条件

由Taylor公式有$g_k \approx g_{k+1} + G_{k+1}(x^k - x^{k+1})$. 当$G_{k+1}$非奇异时，有$G_{k+1}^{-1} y_k \approx s_k$. 因为目标函数在极小值点附近的形态与二次函数近似，所以一个合理的想法就是，如果选择B_{k+1}使其满足

$$B_{k+1} s_k = y_k, \tag{3.5.2}$$

或选取$H_{k+1} = B_{k+1}^{-1}$使其满足

$$H_{k+1} y_k = s_k, \tag{3.5.3}$$

那么，B_{k+1}就可以较好地近似G_{k+1}，H_{k+1}也就可以较好地近似G_{k+1}^{-1}. 关系式(3.5.2)式和(3.5.3)式均称为拟牛顿方程.

容易看到: 拟牛顿方程(3.5.2)（或(3.5.3)）中含有$(n^2 + n)/2$个未知数，但只有n个方程，所以方程组(3.5.2)（或(3.5.3)）一般有无穷多个解. 因而，由拟牛顿方程确定的是一簇算法，称之为拟牛顿法. 值得提到的是, 在拟牛顿算法的设计中，有些是基于拟牛顿方程(3.5.2)（在此情况下，校正公式为$B_{k+1} = B_k + E_k$）; 有些是基于拟牛顿方程(3.5.3). 在本书中，选择后一种方式. 由于在每次迭代中，尺度矩阵H_k(或B_k)总是变化的，所以，拟牛顿法也称为变尺度法.

3.5 拟牛顿法 — 拟牛顿条件

正如牛顿法是在椭球范数$\|\cdot\|_{G_k}$意义下的最速下降法一样，拟牛顿法是在椭球范数$\|\cdot\|_{B_k}$意义下的最速下降法. 事实上，由$f(x^k+d) = f(x^k) + g_k^\top d + o(\|d\|)$可知, $g_k^\top d$越小，目标函数f下降的越快. 所以极小化问题

$$\begin{aligned} \min \quad & g_k^\top d \\ \text{s.t.} \quad & \|d\|_{B_k} = 1 \end{aligned} \tag{3.5.4}$$

的最优解就是目标函数f在椭球范数$\|\cdot\|_{B_k}$意义下x^k处的最速下降方向. 由广义Cauchy-Schwartz 不等式, 于是有

$$(g_k^\top d)^2 \le (g_k^\top B_k^{-1} g_k)(d^\top B_k d),$$

且当$d = -B_k^{-1}g_k$时等式成立，此时$g_k^\top d$最小. 所以，$d = -\frac{B_k^{-1}g_k}{\|B_k^{-1}g_k\|_{B_k}}$是优化问题(3.5.4)的最优解. 故拟牛顿方向$d = -B_k^{-1}g_k$ 是目标函数f在椭球范数$\|\cdot\|_{B_k}$意义下x^k处的最速下降方向.

§3.5.2 DFP算法

DFP算法是最先被研究的一类拟牛顿法，是由Davidon于1959年最先提出的，后来分别由Flecher 和Powell于1963年加以改进而成. 为了得到其校正公式，考虑对称秩二校正：

$$H_{k+1} = H_k + \alpha uu^\top + \beta vv^\top.$$

结合拟牛顿方程(3.5.3)式，进一步可得

$$H_{k+1}y_k = H_k y_k + \alpha uu^\top y_k + \beta vv^\top y_k = s_k.$$

在此公式中，选择$u = s_k$及$v = H_k y_k$，并且α和β分别由$\alpha u^\top y_k = 1$及$\beta v^\top y_k = -1$来确定，即

$$\alpha = \frac{1}{u^\top y_k} = \frac{1}{s_k^\top y_k}, \quad \beta = -\frac{1}{v^\top y_k} = -\frac{1}{y_k^\top H_k y_k}.$$

于是，得到的校正公式为:

$$H_{k+1} = H_k - \frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k} + \frac{s_k s_k^\top}{y_k^\top s_k}. \tag{3.5.5}$$

称此公式为DFP校正公式.

3.5 拟牛顿法 — DFP算法

算法 3.5.1 *(DFP算法)* 设函数f由问题*(3.0.1)*给出. 选取初始点$x^0 \in \mathbb{R}^n$, 若$\|g_0\| = 0$, 算法终止. 否则, 选取初始矩阵H_0(通常取为单位矩阵I), 置$k := 0$.

步*1* 置$d^k := -H_k g_k$.

步*2* 由精确一维线搜索$f(x^k + \lambda_k d^k) = \min_{\lambda \geq 0} f(x^k + \lambda d^k)$, 计算步长$\lambda_k$.

步*3* 置$x^{k+1} := x^k + \lambda_k d^k$, 若$\|g_{k+1}\| = 0$, 算法终止. 否则, 转步*4*.

步*4* 置$s_k := x^{k+1} - x^k$, $y_k := g_{k+1} - g_k$. 由*DFP*校正公式*(3.5.5)*式得H_{k+1}.

步*5* 置$k := k + 1$, 转步*1*.

注. 在步2中，迭代步长λ_k也可由非精确一维线搜索来确定，由于对应算法的理论分析较为复杂，所以本书中只考虑步长λ_k是由精确一维线搜索来确定的情况.

3.5 拟牛顿法 — DFP算法

定理 3.5.1 *(DFP校正公式的正定继承性)* 在DFP算法中，如果初始矩阵H_0对称正定，那么整个矩阵序列$\{H_k\}$都是对称正定的.

证明 使用数学归纳法. 显然，当$k = 0$时, 结论成立. 假设当$k = i$时结论成立，即矩阵H_i是对称正定的且$g_i \neq 0$(否则迭代终止). 为了使用DFP校正公式计算H_{i+1}，需要说明计算H_{i+1}的校正公式(3.5.5)是有意义的，为此，需证明$y_i^{\top} s_i > 0$. 由y_i, s_i的定义及迭代公式，于是有

$$y_i^{\top} s_i = (g_{i+1}-g_i)^{\top}(x^{i+1}-x^i) = (g_{i+1}-g_i)^{\top}(-\lambda_i H_i g_i) = -\lambda_i g_{i+1}^{\top} H_i g_i + \lambda_i g_i^{\top} H_i g_i.$$

利用精确一维搜索可得$g_{i+1}^{\top} H_k g_i = 0$. 又因为$H_i$正定，所以$-H_i g_i$为下降方向. 又由$\lambda_i > 0$，故

$$y_i^{\top} s_i = \lambda_i g_i^{\top} H_i g_i > 0.$$

所以, 计算H_{i+1}的校正公式(3.5.5)是有意义的.

3.5 拟牛顿法 — DFP算法

由H_i的对称性及校正公式(3.5.5)知: 矩阵H_{i+1}是对称的. 下面证明矩阵H_{i+1}是正定的. 任取$x \in \mathbb{R}^n \setminus \{0\}$, 由DFP校正公式可得

$$\begin{aligned} x^\top H_{i+1} x &= x^\top H_i x - \frac{x^\top H_i y_i y_i^\top H_i x}{y_i^\top H_i y_i} + \frac{(s_i^\top x)^2}{y_i^\top s_i} \\ &= \frac{x^\top H_i x y_i^\top H_i y_i - x^\top H_i y_i y_i^\top H_i x}{y_i^\top H_i y_i} + \frac{(s_i^\top x)^2}{y_i^\top s_i}. \end{aligned}$$

因为H_i对称正定，所以存在对称正定矩阵D_i使得$H_i = D_i^2$. 记$u_i = D_i x, v_i = D_i y_i$，则上式变为

$$x^\top H_{i+1} x = \frac{u_i^\top u_i v_i^\top v_i - (u_i^\top v_i)^2}{y_i^\top H_i y_i} + \frac{(s_i^\top x)^2}{y_i^\top s_i}. \tag{3.5.6}$$

由Chauchy-Schwartz不等式知, $(u_i^\top v_i)^2 \le u_i^\top u_i v_i^\top v_i$且等号成立当且仅当$u_i = \gamma v_i$ $(\gamma \neq 0)$ 时成立. 因此，(3.5.6)式右端的第一项非负，且仅当$u_i = \gamma v_i$ $(\gamma \neq 0)$时等于零; 而此时(3.5.6)式右端的第二项

$$\frac{(s_i^\top x)^2}{y_i^\top s_i} = \frac{(\gamma s_i^\top y_i)^2}{y_i^\top s_i} = \gamma^2 s_i^\top y_i > 0.$$

又因为(3.5.6)式右端的第二项大于零, 所以，总有$x^\top H_{i+1} x > 0$, 即矩阵H_{i+1}是正定的. 由数学归纳法，定理得证. □

3.5 拟牛顿法 — DFP算法

引理 3.5.1 将*DFP*算法用于求解由*(3.4.1)*式所定义严格凸二次函数的极小化问题. 如果初始矩阵H_0对称正定，产生的迭代点互异，且产生的迭代方向分别记为$d^0, d^1, \ldots, d^k (k \le n)$, 那么,

(i) $H_k y_i = s_i, \quad \forall i \in \{0, 1, \ldots, k-1\};$

(ii) $(d^i)^\top G d^j = 0, \quad \forall i, j \in \{0, 1, \ldots, k\}$且$i < j.$

证明 对k使用数学归纳法. 注意到

$$s_i = x^{i+1} - x^i = \lambda_i d^i, \quad y_i = g_{i+1} - g_i = G(x^{i+1} - x^i) = G s_i.$$

当$k = 1$时，由拟牛顿方程易验证: $H_1 y_0 = s_0$成立. 另外，利用$Gd^0 = \frac{1}{\lambda} G s_0$, $d^1 = -H_1 g_1$, $G s_0 = y_0$, $H_1 y_0 = s_0$, $\frac{1}{\lambda_0} s_0 = d^0$以及$g_1^\top d^0 = 0$可得

$$\begin{aligned} (d^0)^\top G d^1 &= (G d^0)^\top d^1 = -\frac{1}{\lambda_0} (G s_0)^\top H_1 g_1 \\ &= -\frac{1}{\lambda_0} y_0^\top H_1 g_1 = -\frac{1}{\lambda_0} g_1^\top s_0 = -g_1^\top d^0 = 0. \end{aligned}$$

所以，当$k = 1$时, 引理的结论成立.

3.5 拟牛顿法 — DFP算法

假设当$k = l$时, 引理的结论成立. 需证当$k = l + 1$时, 引理的结论也成立. 由归纳假设，则

$$\begin{aligned} &H_l y_i = s_i, \quad \forall i \in \{0, 1, \ldots, l-1\}, \\ &(d^i)^\top G d^j = 0, \quad \forall i, j \in \{0, 1, \ldots, l\} \text{且} i < j. \end{aligned} \tag{3.5.7}$$

当$k = l + 1$时，对于任意的$i \in \{0, 1, \ldots, l-1\}$, 有

$$\begin{aligned} H_{l+1} y_i &= \left(H_l - \frac{H_l y_l y_l^\top H_l}{y_l^\top H_l y_l} + \frac{s_l s_l^\top}{y_l^\top s_l} \right) y_i = H_l y_i - \frac{H_l y_l y_l^\top H_l y_i}{y_l^\top H_l y_l} + \frac{s_l s_l^\top y_i}{y_l^\top s_l} \\ &= s_i - \frac{H_l y_l (y_l^\top s_i)}{y_l^\top H_l y_l} + \frac{s_l s_l^\top G s_i}{y_l^\top s_l} = s_i - \frac{H_l y_l (s_l^\top G s_i)}{y_l^\top H_l y_l} + \frac{s_l (s_l^\top G s_i)}{y_l^\top s_l}. \end{aligned}$$

再由(3.5.7)式及$s_i = \lambda_i d^i$知: 上式右端后两项为零. 所以对于任意的$i \in \{0, 1, \ldots, l-1\}$有$H_{l+1} y_i = s_i$成立; 又根据拟牛顿方程可知: $H_{l+1} y_l = s_l$. 所以，当$k = l + 1$时引理的结论(i)成立.

3.5 拟牛顿法 — DFP算法

另外，为了证明当$k = l + 1$时引理的结论(ii)成立, 由(3.5.7)式可知，只需证明

$$(d^i)^\top G d^{l+1} = 0, \quad \forall i \in \{0, 1, \ldots, l\}.$$

事实上，

$$\begin{aligned}(d^i)^\top G d^{l+1} &= (Gd^i)^\top d^{l+1} = \frac{1}{\lambda_i} y_i^\top (-H_{l+1} g_{l+1}) = -\frac{1}{\lambda_i} s_i^\top g_{l+1} = -g_{l+1}^\top d^i \\ &= -\left(g_{i+1} + \sum_{j=i+1}^{l} y_j\right)^\top d^i = -g_{i+1}^\top d^i - \sum_{j=i+1}^{l} (s_j^\top G d^i) = 0.\end{aligned}$$

所以，当$k = l + 1$时引理的结论(ii)成立.

由数学归纳法，定理得证. □

3.5 拟牛顿法 — DFP算法

定理 3.5.2 *(DFP算法的二次终止性)* 假设引理*3.5.1*的条件成立. 则*DFP*算法至多迭代n次就可达到函数的极小值点, 即存在$k_0 \in \{0, 1, \ldots, n\}$使得$x^{k_0} = x^*$. 特别地, 如果对任意的$k \in \{0, 1, \ldots, n-1\}$有$x^k \neq x^*$, 那么$H_n = G^{-1}$.

证明 由引理3.5.1知: DFP算法是一种共轭方向法, 因而定理的第一个结论成立. 另外, 若对任意的$k \in \{0, 1, \ldots, n-1\}$有$x^k \neq x^*$, 那么DFP算法将产生$n$个共轭方向, 不妨记为$d^0, d^1, \ldots, d^{n-1}$, 于是方向$d^0, d^1, \ldots, d^{n-1}$线性无关, 而对任意的$i \in \{0, 1, \ldots, n-1\}$有$s_i = \lambda_i d^i$且$\lambda_i > 0$, 所以$s_0, s_1, \ldots, s_{n-1}$线性无关. 又根据引理3.5.1知: 对任意的$i \in \{0, 1, \ldots, n-1\}$, 有$H_n y_i = H_n G s_i = s_i$成立, 即

$$H_n G(s_0 \ s_1 \ \cdots \ s_{n-1}) = (s_0 \ s_1 \ \cdots \ s_{n-1}), \tag{3.5.8}$$

其中$(s_0 \ s_1 \ \cdots \ s_{n-1})$表示以$s_0, s_1, \ldots, s_{n-1}$为列的矩阵. 因为$s_0, s_1, \ldots, s_{n-1}$线性无关, 所以矩阵$(s_0 \ s_1 \ \cdots \ s_{n-1})$非奇异. 因此, 由(3.5.8)式可得: $H_n G = I$, 即$H_n = G^{-1}$. 定理得证. □

另外, 可以证明: 在一定的条件下, DFP算法是全局收敛的. 也可证明: 在一定的条件下, DFP算法是局部超线性收敛的. 由于收敛性分析较为复杂, 本书从略.

§3.5.3 BFGS算法

BFGS算法是目前求解无约束优化问题的最流行的也是最有效的拟牛顿算法，它是分别由Broyden, Flecher, Goldfarb和Shanno于1970年独立提出的拟牛顿法.

在DFP校正公式的推导中，使用拟牛顿方程(3.5.2)式代替拟牛顿方程(3.5.3)式，可得

$$B_{k+1}s_k = B_k s_k + \alpha u u^\top s_k + \beta v v^\top s_k = y_k.$$

类似于DFP校正公式的推导，由此可得相对于B_k的BFGS校正公式为:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{y_k^\top s_k}. \tag{3.5.9}$$

3.5 拟牛顿法 — BFGS算法

两次使用逆的秩一校正的Sherman-Morrison公式，即

$$(A+uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}, \quad \text{其中} A \in \mathbb{R}^{n\times n}\text{可逆}, u, v \in \mathbb{R}^n, 1+v^\top A^{-1}u \neq 0,$$

进而, 可得相对于H_k的BFGS校正公式:

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)s_k^\top + s_k(s_k - H_k y_k)^\top}{s_k^\top y_k} - \frac{(s_k - H_k y_k)^\top y_k}{(s_k^\top y_k)^2} s_k s_k^\top. \tag{3.5.10}$$

记

$$w_k := \sqrt{y_k^\top H_k y_k}\left(\frac{s_k}{y_k^\top s_k} - \frac{H_k y_k}{y_k^\top H_k y_k}\right), \tag{3.5.11}$$

那么，容易验证：(3.5.10)式等价于

$$H_{k+1} = H_k - \frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k} + \frac{s_k s_k^\top}{y_k^\top s_k} + w_k w_k^\top, \tag{3.5.12}$$

其中w_k由(3.5.11)式定义.

3.5 拟牛顿法 — BFGS算法

因此，在DFP算法(即算法3.5.1)中，用“由BFGS校正公式(3.5.12)式得到的矩阵H_{k+1}”代替“由DFP 校正公式(3.5.5)式得到的矩阵H_{k+1}”，进而可得BFGS算法.

下面介绍Broyden簇拟牛顿法，其中校正公式为

$$H_{k+1}^{\phi} = H_k - \frac{H_k y_k y_k^{\top} H_k}{y_k^{\top} H_k y_k} + \frac{s_k s_k^{\top}}{y_k^{\top} s_k} + \phi w_k w_k^{\top}, \tag{3.5.13}$$

其中ϕ为一个实参数，w_k由(3.5.11)定义. 显然，当$\phi = 0$时，Broyden簇校正公式(3.5.13) 正好是DFP校正公式(3.5.5); 当$\phi = 1$时，Broyden簇校正公式(3.5.13) 正好是BFGS校正公式(3.5.12). 在DFP算法(算法3.5.1)中，使用“由Broyden簇校正公式(3.5.13)式得到的矩阵H_{k+1}”来代替“由DFP校正公式(3.5.5)式得到的矩阵H_{k+1}”，于是可得到Broyden簇拟牛顿法. 因此，DFP 算法和BFGS算法是Broyden簇拟牛顿法中两个特殊的算法.

3.5 拟牛顿法 — BFGS算法

1972年，Dixon证明了下面的定理(本书证明从略).

定理 3.5.3 假设函数f在$\mathbb{R}^n$上是连续可微的. 则由*Broyden*簇拟牛顿法产生的迭代点列$\{x^k\}$与参数ϕ无关.

利用定理3.5.3, 可以把DFP算法的性质推广到所有的Broyden簇拟牛顿法.

在实际计算中，由于舍入误差等因素，DFP算法的效率会受到很大的影响，但BFGS算法受到的影响要小得多. 特别是采用非精确一维线搜索时，DFP算法的效率很低，而BFGS算法仍然十分有效. 另外，本书中只是介绍了基本的BFGS算法，有很多改进的BFGS算法非常有效，如：L-BFGS算法(见文献[18]). 因此, BFGS算法是目前被公认为最好的拟牛顿算法.

3.5 拟牛顿法 — 作业

3.15 用精确一维线搜索的DFP算法求解下列无约束优化问题.

(1) $\min \ f(x_1, x_2) = 2x_1^2 + x_2^2 - 4x_1 + 2$, 其中初始点取为$x^0 = (2, 1)^\top$, 初始矩阵取为单位矩阵.

(2) $\min \ f(x_1, x_2) = x_1^2 - x_1x_2 + x_2^2 + 2x_1 - 4$, 其中初始点取为$x^0 = (2, 2)^\top$, 初始矩阵取为单位矩阵.

3.16 设H_k是一个奇异的半正定矩阵, 试证明: 用精确一维线搜索的DFP修正公式得到的矩阵H_{k+1}也是奇异矩阵.