# 拟牛顿法

拟牛顿法是最重要的非线性算法之一。跟最速下降法一样，它在每次迭代的时候都只用到了优化目标的导数信息，通过利用导数信息达到一个超线性收敛。并且，由于拟牛顿法不适用二阶导数信息，因此它有时比牛顿法更加高效。

拟牛顿法在无约束优化、约束优化及大规模优化问题中有着许许多多的变体。本章我们将主要考虑小规模和中等规模下的拟牛顿法。针对大规模的优化问题，我们将放在第七章讨论。

自动求导的技术（第八章）让牛顿法在不直接调用二阶导数的情况下也可以实现。然而这种技术在很多情况下是不适用的，同时很多时候它的运算量很大。因此，拟牛顿法仍具有极大的价值。

## 目录

# 1 BFGS 法

- 保持正定
- 产生于加权 F 范数最小化
- 权矩阵为平均 Hessian 矩阵
- 秩二校正

BFGS 法是最著名的拟牛顿法，它是由 Broyden, Fletcher, Goldfarb 和 Shanno 共同发明的。本节我们将讨论它和与它很相似的 DFP 算法，并研究他们的理论效果和实现。

首先我们分析下述二次模型

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \tag{1}$$

其中 $B_k$ 是一个 $n \times n$ 的对称正定阵，它会随迭代而更新。注意到 $f_k$ 和 $\nabla f_k$ 的信息是可以使用的，于是我们可以显式写出上述问题的极小子 $p_k$:

$$p_k = -B_k^{-1} \nabla f_k \tag{2}$$

以此作为搜索方向我们有如下迭代

$$x_{k+1} = x_k + \alpha_k p_k \tag{3}$$

其中，$\alpha_k$ 满足 Wolfe 条件。这个迭代和线性搜索的牛顿法很像，区别在于我们使用一个近似 Hessian 阵 $B_k$ 而非 $\nabla^2 f$

然而我们并非每次都完全重新计算 $B_k$，而是根据最近的一些曲率变化对它进行一些简单的更新。假设我们产生了一个新的迭代点 $x_{k+1}$，从而需要去构造一个具有如下形式的新的二次函数

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p$$

那么我们希望 $B_{k+1}$ 有哪些性质呢？有一个很合理的要求是 $m_{k+1}$ 的梯度应该要和优化对象在 $x_k$ 和 $x_{k+1}$ 处梯度相同。于是便有

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k$$

从而我们有

$$B_{k+1} \alpha_k p_k = \nabla f_{k+1} - \nabla f_k \tag{4}$$

为了简化记号, 我们定义:

$$s_k = x_{k+1} - x_k = \alpha_k p_k, \quad y_k = \nabla f_{k+1} - \nabla f_k \tag{5}$$

于是 (4) 化为:

$$B_{k+1} s_k = y_k \tag{6}$$

我们称上式为切方程。由 $B_k$ 的正定性，我们有

$$s_k^T y_k > 0 \tag{7}$$

当 $f$ 是强凸函数时，上式对于任意的 $x_{k+1}$ 和 $x_k$ 都成立。然而这个式子并不是对所有的非凸函数成立的，此时我们需要对线性搜索法增加一些限制以保证 (7) 成立。实际上，只要保证了 Wolfe 条件 (3.6) 或者强 Wolfe 条件 (3.7) 成立，这个

条件便是成立的。为了说明这一点，我们由 (5) 和 (3.6b)$\nabla f_{k+1}^T s_k \geq c_2 \nabla f_k^T s_k$，得

$$y_k^T s_k \geq (c_2 - 1)\alpha_k \nabla f_k^T p_k \tag{8}$$

由于 $c_2 < 1$ 且 $p_k$ 是下降方向，上式右端大于 0，于是曲率条件 (7) 成立。

当曲率条件满足的时候，关于 $B_{k+1}$ 的切方程有解。然而，这个解是不唯一的。因为矩阵的自由度为 $n(n+1)/2$，而切方程只提供了 $n$ 个等式约束。正定性要求又给了 $n$ 个不等式约束——所有顺序主子式大于零——但是这些条件不够控制所有的自由度。

为了确定唯一的 $B_{k+1}$，我们需要提出一些额外的条件。有时我们会选取所有满足切方程的 $B_{k+1}$ 中最接近 $B_k$ 的。即，我们考虑如下问题

$$\min_B \|B - B_k\| \quad s.t. \quad B = B^T, Bs_k = y_k \tag{9}$$

这里每一个不同的范数，都会导出一种不同的拟牛顿算法。其中一种即使得 (9) 容易求解，同时无量纲的方法是使用加权 Frobenius 范数：

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F \tag{10}$$

其中 $W$ 可以取任何满足方程 $Wy_k = s_k$ 的矩阵。为了让它更具象，我们可以假设 $W = \bar{G}_k^{-1}$，其中 $\bar{G}_k$ 是平均 Hessian 阵：

$$\bar{G}_k = \left[\int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) \mathrm{d}\tau\right] \tag{11}$$

下述重要关系

$$y_k = \bar{G}_k \alpha_k p_k = \bar{G}_k s_k \tag{12}$$

可由泰勒定理推出。利用该加权矩阵 $W$，(10) 是无量纲的，于是我们在求解 (9) 时不用考虑问题的单位。

利用这个矩阵范数，(9) 的唯一解为：

$$\text{(DFP)} \quad B_{k+1} = (I - \rho_k y_k s_k^T) B_k (I - \rho_k s_k y_k^T) + \rho_k y_k y_k^T \tag{13}$$

其中

$$\rho_k = \frac{1}{y_k^T s_k} \tag{14}$$

这个式子叫做 DFP 迭代公式，它由 Davidon 在 1959 年提出，并由 Fletcher 和 Powell 推广。

$B_k$ 的逆通常记作 $H_k = B_k^{-1}$，它在使用该算法时很有用。它让我们在计算搜索方向 (2) 时，只需要使用一个矩阵与向量的乘积。使用 Sherman-Morrison-Woodbury 公式，我们可以得到 $H_k$ 的迭代公式：

$$\text{(DFP)} \quad H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k^T s_k}{y_k^T s_k} \tag{15}$$

注意到上式右端的后两项都是秩一矩阵，因此对 $H_k$ 的更新其实是一个秩二校正。不难看出 (13) 也是一个秩二校正。这便是拟牛顿法的基本思想：并不在每

一步重新估计 Hessian 矩阵，而是基于我们对目标函数的观察对现有的估计进行校正。

DFP 其实已经很高效了，然而 BFGS 的效果更好。BFGS 法被认为是最高效的拟牛顿法迭代公式。只需对 (13) 进行简单的更改，便可以得到 BFGS 迭代公式。这里，我们不考虑 $B_k$ 的条件，而去考虑 $H_k$ 需要满足的条件。首先 $H_{k+1}$ 必须是对称正定的，必须满足切方程 (6):

$$H_{k+1}y_k = s_k$$

于是求解 $H_k$ 我们需要考虑如下问题:

$$\min_H \|H - H_k\| \quad s.t. \quad H = H^T, Hy_k = s_k \tag{16}$$

这里，我们还是使用前述加权 Frobenius 范数，其权矩阵 $W$ 满足 $Ws_k = y_k$。(为了直观，我们假设 $W = \bar{G}_k$。) 于是此时 (16) 的唯一解为

$$\text{(BFGS)} \quad H_{k+1} = (I - \rho_k s_k y_k^T)H_k(I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T \tag{17}$$

其中 $\rho_k$ 定义同 (14)。

至此，我们距实现 BFGS 算法只剩最后一个问题: 如何去选取初始估计 $H_0$? 然而不幸的是，并没有一个很好的方法在任何情况下都很有效。我们需要用的问题的具体信息，如使用 $x_0$ 处 Hessian 矩阵的有限差分解的逆。有时，我们也会直接使用单位阵，或者数乘后的单位阵并借此表现变量的尺度。

**Algorithm 1** BFGS 法

1: 取初值 $x_0$ 和终止条件 $\epsilon > 0$, 计算 Hessian 的估计 $H_0$;
2: $k$ <- 0;
3: **while** $\|\nabla f_k\| > \epsilon$ **do**
4: 计算搜索方向

$$p_k = -H_k \nabla f_k \tag{18}$$

5: 令 $x_{k+1} = x_k + \alpha_k p_k$，其中 $\alpha_k$ 由满足 Wolfe 条件的线性搜索法得到
6: 令 $s_k = x_{k+1} - x_k, y_k = \nabla f_{k+1} - \nabla f_k$
7: 根据 (17) 计算 $H_{k+1}$
8: $k < -k + 1$
9: **end while**

每次迭代的复杂度是 $O(n^2)$ 个方程，且没有任何需要 $O(n^3)$ 的操作 (如线性系统求解和矩阵乘法)。这个算法是稳定的，而且超线性收敛，能够满足大部分的使用需求。虽然牛顿法的收敛更快，但是每次迭代的时候的运算量往往会更大，因为它要计算二阶导数、求解线性系统。

我们同样可以推导基于 $B_k$ 的 BFGS 算法，其更新公式如下:

$$\text{(BFGS)} \quad B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \tag{19}$$

如果单纯地使用这个式子和 $B_k p_k = -\nabla f_k$ 是很没有效率的。然而后续，我们会基于 $B_k$ 的 Cholesky 分解给出一个相对高效的实现方法。

## 1.1 性质

BFGS 的超线性收敛性质是很容易观察到的。对于 Rosenbrock 方程，我们从 (-1.2,1) 出发，为使 $\|\nabla f_k\| \leq 10^{-5}$，最速下降法需要迭代 5264 次，而 BFGS 和牛顿法分别只需要 34 次和 21 次。

首先注意到在求解极小化问题 (16)，即求解 $H_{k+1}$ 时，我们并没有显式地保证 $H_{k+1}$ 正定。然而，我们可以说明：如果 $H_k$ 正定，那么 $H_{k+1}$ 正定。首先，由 (8) 有，$y_k^T s_k > 0$，于是 (17) 和 (14) 是可用的。对于非零向量 $z$，我们有

$$z^T H_{k+1} z = w^T H_k w + \rho_k (z^T s_k)^2 \geq 0$$

其中 $w = z - \rho_k y_k (s_k^T z)$，上式第二项当且仅当 $s_k^T z = 0$ 时非正，然而此时第一项 $w = z \neq 0$，于是 $H_{k+1}$ 的正定性得证。

为了让拟牛顿法无量纲化，我们需要让 (9) 和 (14) 通过同样的变化后无量纲化。而我们在 (9) 和 (14) 中使用的权重矩阵 $W$ 则可以满足这个要求。当然，$W$ 还有其他的很多选择，但是除了某些特定的问题，目前还没发现显著优于 BFGS 法的公式。

BFGS 法在优化二次函数的时候还有许多有趣的性质，这些性质我们将在后续 Broyden 族的分析中讨论，BFGS 法是其中的一个特例。

当然我们也会去考虑 (17) 会不会有时产生很差的结果。如果 $H_k$ 与真实 Hessian 差别很大，有没有校正的方法。举例来说，当 $y_k^T s_k$ 很小的时候，$H_{k+1}$ 会包含一些很大的元素，这种行为合理么？这个问题其实跟舍入误差是息息相关的。前述问题就变成在拟牛顿法中，舍入误差会抹去所有有用的信息么？

这些问题都已经得到了很好的研究和实验，目前已知 BFGS 法有非常高效自我校正能力。而 DFP 算法的自我校正能力则比较弱，这也被认为是其实际效果不那么好的原因。当使用了合适的线性搜索法时，BFGS 算法才具备自我校正能力。特别的，Wolfe 线性搜索条件可以保证模型 (1) 捕捉到合适的曲率信息。

× BFGS 比 DFP 好的性质：

- BFGS 有着非常高效的自我校正能力，而 DFP 的自我校正能力比较弱

还有一个很有趣的性质是，DFP 算法和 BFGS 算法的迭代公式是相互对偶的。也就是说我们可以通过令 $s \leftrightarrow y$，$B \leftrightarrow H$ 进行相互转化。当然从算法的产生过程来看，这种对称性是比较自然的。

## 1.2 实现

为了得到一个高效的优化算法，我们还要对算法 1做一些改良。首先，线性搜索必须满足 Wolfe 条件或强 Wolfe 条件，并且初值取 1。实践结果表明，使用一个不太精确的的线性搜索法会让算法开销更小。另外，实践中常常使用 $c_1 = 10^{-4}, c_2 = 0.9$。

跟前面提到的一样，很多时候 $H_0$ 会直接取成 $\beta I$，但是 $\beta$ 的选取其实也是没有一个较好的通用准则的。如果 $\beta$ 太大，那么初始方向的长度会很长，于是在计算步长的时候，可能会需要比较大的计算量。有些软件会要求用户定义一个初始步长的长度 $\delta$，并取 $H_0 = \delta \|g_0\|^{-1} I$。

于是，我们有这么一种启发式的方法：首先取 $H_0 = I$，然后在进行 BFGS 迭代之前，我们令

$$H_0 = \frac{y_k^T s_k}{y_k^T y_k} I \tag{20}$$

这种做法让 $H_0$ 的规模与 $\nabla^2 f(x_0)^{-1}$ 在下述意义下相似。假设 (11) 定义的评价 Hessian 矩阵是镇定的，于是存在一个平方根 $\bar{G}^{1/2}$ 使得 $\bar{G}_k = \bar{G}_k^{1/2}\bar{G}_k^{1/2}$。于是，通过定义 $z_k = \bar{G}_k^{1/2}s_k$ 并利用 (12)，我们有

$$\frac{y_k^T s_k}{y_k^T y_k} = \frac{(\bar{G}_k^{1/2}s_k)^T\bar{G}_k^{1/2}s_k}{(\bar{G}_k^{1/2})^T\bar{G}_k\bar{G}_k^{1/2}s_k} = \frac{z_k^T z_k}{z_k^T\bar{G}z_k} \tag{21}$$

上式的倒数是对 $\bar{G}_k$ 的特征值的估计，实际上也与 $\nabla^2 f(x_k)$ 的特征值很接近。其他的方法也有，但是这种方法是实践中最成功的。

我们在 (19) 中给了一种使用 $B_k$ 而非 $H_k$ 的 BFGS 算法。有一种高效的做法是不储存 $B_k$ 而是储存它的 Cholesky 分解 $L_kD_kL_k^T$。于是我们可以从 (19) 中推导出 $L_k, D_k$ 的时间复杂度是 $O(n^2)$ 更新公式，同时，对于这种方法，我们求解线性方程的时候时间复杂度也是 $O(n^2)$。因此这种做法的复杂度和算法1差不多，但是这么做有一个好处：我们可以适当地增大 $D_k$，当他们不够大的时候，从而保证算法的稳定性。然而，实践经验告诉我们这样做并没有实际的好处，所以我们还是更倾向于使用算法 1。

在线性搜索法不满足 Wolfe 条件的情况下，BFGS 法的效果可能并不是很好。比如只使用 Armijo 条件，此时 $y_k^T s_k > 0$ 的条件是不能被保证的。为了克服这个缺陷，这些算法有时候会选择令 $H_{k+1} = H_k$，当 $y_k^T s_k$ 很小或者非正的时候。但是这种方法其实是不推荐使用的，因为这样做可能会跳过很多迭代步，曲率信息无法得到更新。18 章我们会讨论一种阻尼 BFGS 更新法，用来处理那些曲率条件 (7) 不满足的情况。

# 2 SR1 法

- 秩一矫正
- 不保证正定
- 能较好估计 Hessian 矩阵
- 适用于 Hessian 矩阵不正定问题

BFGS 法和 DFP 法都是秩 2 校正的方法，而下面要介绍的 SR1 方法是秩 1 校正的方法。与前两种方法相比，秩 1 校正的方法不能保证 $B_{k+1}$ 的正定性，然而，基于 SR1 法的算法在实际应用中表现出了很好的性能，所以我们需要对它了解一下。

## SHERMAN–MORRISON–WOODBURY 公式及助记推导

对与秩一矫正 $\bar{A} = A + uv^T$，我们有下述重要公式

$$\bar{A}^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

这是很容易验证的，下面我们给出一个推导过程：$\bar{A}^{-1} = A^{-1} + X$，则有

$$\begin{aligned}
I = \bar{A}X \Rightarrow & AX + uv^TX + uv^TA^{-1} = 0 \\
\Rightarrow & X + A^{-1}uv^TX + A^{-1}uv^TA^{-1} = 0 \\
& \text{[根据方程形式，我们取} X = kA^{-1}uv^TA^{-1}\text{]} \\
\Rightarrow & kA^{-1}uv^TA^{-1} + kA^{-1}uv^TA^{-1}uv^TA^{-1} = -A^{-1}uv^TA^{-1} \\
& k(1 + v^TA^{-1}u)A^{-1}uv^TA^{-1} = -A^{-1}uv^TA^{-1} \\
& k = -\frac{1}{1 + v^TA^{-1}u} \quad X = -\frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}
\end{aligned}$$

对称秩一矫正公式形式如下：

$$B_{k+1} = B_k + \sigma v v^T$$

由于我们要求迭代满足切方程 (6)$y_k = B_{k+1}s_k$，因此，可推出

$$(SR1) \quad B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k} \tag{22}$$

利用 Sherman-Morrison 公式，有

$$(SR1) \quad H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k} \tag{23}$$

如同我们前面所说的，$B_k$ 的正定性并不能保证 $B_{k+1}$ 具有正定性，这也是大家最初认为的基于 SR1 算法的线性搜索法的缺点。然而随着洗那里与方法的出现，SR1 算法变得非常得实用，同时，它能够推广到 Hessian 矩阵非正定的情况的性质成为了它的一大重要优点。

SR1 算法的主要缺点在于它的迭代方程可能不成立。注意到前述的迭代算法似乎没什么问题，然而，它是基于 $(y_k - B_k s_k)^T s_k \neq 0$ 这个条件推导出来的。当这个条件不成立时，有如下两种情况：

1. $y_k = B_k s_K$，于是此时我们只要让 $B_{k+1} = B_k$ 即可
2. $y_k \neq B_k s_k$ 但 $(y_k - B_k s_k)^T s_k = 0$，此时不存在满足切方程的对称秩一迭代。

上述情况说明了秩一迭代不能给算法提供足够的自由度。简单的 SR1 算法是数值不稳定甚至会崩溃的。

尽管如此，基与下述考虑，SR1 公式还是很有应用价值的:

i) 我们可以用一个简单的保障算法防止崩溃和数值不稳定的情况发生

ii) SR1 对 Hessian 矩阵的估计效果很好——有时候甚至比 BFGS 法还要好

iii) 对于约束优化问题或局部分离的函数，曲率条件 $y_k^T s_k > 0$ 可能无法保证，此时，是不建议使用 BFGS 算法的。事实上，在这些问题中我们希望产生一个非正定的 Hessian 估计，因为很多时候真实的 Hessian 就是非正定的

下面我们介绍一个防止 SR1 算法崩溃的策略：我们只在下述情况更新 (22):

$$|s_k^T(y_k - B_k s_k)| \geq r\|s_k\|\|y_k - B_k s_k\| \tag{24}$$

期中，$r \in (0,1)$ 通常是一个很小的数，如 $r = 10^{-8}$。若上式不满足，则取 $B_{k+1} = B_k$。大部分 SR1 算法都会使用类似的跳过策略。

那么为什么对于 SR1 方法可以用这样的跳过策略，而之前的 BFGS 方法就不可以用类似的手段呢？这是因为这两种情况其实是很不一样的。条件 $s_k^T(y_k - B_k s_k) \approx 0$ 是很不容易发生的，同时又因为它暗示了 $s_k^T \bar{G} s_k \approx s_k^T B_k s_k$，期中 $\bar{G}$ 是平均 Hessian 矩阵，也就是说当前的估计 $B_k$ 已经很准确，因此我们可以采用跳过的策略。而在 BFGS 方法中，如果我们不基于 Wolfe 条件寻找步

**Algorithm 2** SR1 信赖域算法

给定初值 $x_0$，和初始 Hessian 估计 $B_0$，
　　信赖域半径 $\Delta_0$，终止条件 $\epsilon > 0$，
　　参数 $\eta \in (0, 10^{-3}), r \in (0, 1)$
$k \leftarrow 0$
**while** $\|\nabla f_k\| > \epsilon$ **do**
　　通过求解下述子问题计算 $s_k$：

$$\min_s \nabla f_k^T s + \frac{1}{2} s^T B_k s \quad \text{s.t. } \|s\| \le \Delta_k \tag{25}$$

　　计算

$$\begin{aligned} y_k &= \nabla f(x_k + x_k) - \nabla f_k \\ ared &= f_k - f(x_k + s_k) \quad \text{actual reduction} \\ pred &= -(\nabla f_k^T s_k + \frac{1}{2} s_k^T B_k s_k) \quad \text{predicted reduction} \end{aligned}$$

　　**if** $ared/pred > \eta$ **then**
　　　　$x_{k+1} = x_k + \eta$
　　**else**
　　　　$x_{k+1} = x_k$
　　**end if**
　　**if** $ared/pred > 0.75$ **then**
　　　　**if** $\|s_k\| \le 0.8\Delta$ **then**
　　　　　　$\Delta_{k+1} = \Delta_k$
　　　　**else**
　　　　　　$\Delta_{k+1} = 2\Delta_k$
　　　　**end if**
　　**else if** $0.1 \le ared/pred \le 0.75$ **then**
　　　　$\Delta_{k+1} = \Delta_k$
　　**else**
　　　　$\Delta_{k+1} = 0.5\Delta_k$
　　**end if**
　　**if** (24) 成立 **then**
　　　　使用 (22) 更新 $B_{k+1}$　　▷ 即使 $x_{k+1} = x_k$
　　**else**
　　　　$B_{k+1} \leftarrow B_k$
　　**end if**
　　$k \leftarrow k + 1$
**end while**

长，曲率条件 $s_k^T y_k \geq 0$ 是很容易不成立的，那么如果我们采用跳过的策略，会导致这种情况发生很频繁，从而严重影响估计的质量。

因为相对于线性搜索法，信赖域方法能更好地处理非正定 Hessian 估计，所以下面我们给出一个信赖域方法框架下的 SR1 算法 2

这个算法是一个典型的信赖域算法，其中在处理信赖域半径的时候我们采取了一个比较特殊的启发式策略。

为了让算法快速收敛，很重要的一点就是即使对于一个失败的方向 $d_k$，我们也要更新我们的矩阵。实际上，迭代效果不好说明了 $B_k$ 对 Hessian 矩阵的估计不够好。除非我们改善了对 Hessian 矩阵的估计，否则在后续的迭代中我们还是会产生类似的迭代，因此，拒绝这些更新反而会影响超线性收敛。

## 2.1 SR1 迭代的性质

SR1 的一个主要的优点就在于它能够产生很好的 Hessian 矩阵的估计。为了说明这一点，我们先对二次函数进行研究。对于二次函数，步长并不影响迭代，所以我们可以设步长为 1，于是

$$p_k = -H_k \nabla f_k, \quad x_{k+1} = x_k + p_k \tag{26}$$

于是 $p_k = s_k$

**定理 1.** *假设 $f$ 是强凸二次函数 $f(x) = b^T x + \frac{1}{2} x^T A x$，其中 $A$ 对称正定。那么无论取什么初值 $x_0$ 和初始矩阵 $H_0$，只要 $(s_k - H_k y_k)^T y_k \neq 0, \forall k$，SR1 方法产生的 $\{x_k\}$ 都能在至多 $n$ 步内收敛到最小值。并且，如果 $n$ 步迭代后，$p_i$ 都是线性无关的，那么 $H_n = A^{-1}$*

证明思路:

首先利用归纳法证明:

$$H_k y_j = s_j, \quad j = 0, 1, \ldots, k-1 \tag{27}$$

于是有 $s_j = H_n y_j = H_n A s_j, \quad j = 0, 1, \ldots, n-1$, 迭代步线性无关时, $H_n = A^{-1}$

于是第 $n$ 步是牛顿步，自然终止了。

假如这些迭代步线性相关，由 (27) 可推出 $H_k y_k = s_k$，于是

$$H_k y_k = H_k(\nabla f_{k+1} - \nabla f_k) = s_k = -H_k \nabla f_k \Rightarrow \nabla f_{k+1} = 0$$

于是 $x_{k+1}$ 是最优解。 □

(27) 说明当 $f$ 是二次函数时，无论使用怎样的线性搜索法，切方程对于所有之前的方向都成立。下一节我们会看到，当线性搜索法是精确搜索时，BFGS 也有类似的结果。

对于一般的非线性函数，SR1 仍然能够在一定条件下对 Hessian 矩阵产生好的估计。

**定理 2.** *假设 $f$ 二次连续可谓，且 Hessian 矩阵在 $x^*$ 的邻域内有界且 Lipschitz 连续。记 $\{x_k\}$ 是逼近 $x^*$ 的任意一个迭代列。假设取 $r \in (0, 1)$，(24) 对于所有的 $k$ 都成立，且 $s_k$ 一致线性无关。那么 SR1 法产生的矩阵 $B_k$ 满足*

$$\lim_{k \to \infty} \|B_k - \nabla^2 f(x^*)\| = 0$$

此处的一致线性无关是说，每次迭代步不会掉入一个维数小于 $n$ 的子空间中。这个假设常常但不总是能在应用中满足。

# 3 Broyden 方法

Broyden 方法包含了一大类如下形式的拟牛顿算法:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi_k (s_k^T B_k s_k) v_k v_k^T \tag{28}$$

其中，$\phi_k$ 是一个收缩子，$v_k$ 定义如下:

$$v_k = \left[ \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k} \right] \tag{29}$$

不难发现，BFGS 法和 DFP 法都是 Broyden 方法，且 $\phi_k = 0$ 时为 BFGS 法，$\phi_k = 1$ 时为 DFP 法。于是 Broyden 方法满足切方程 (6)，我们可以把前述定义改写为:

$$B_{k+1} = (1 = \phi_k) B_{k+1}^{BFGS} + \phi_k B_{k+1}^{DFP}$$

因此，$s_k^T y_k > 0$ 时，该方法能保证对 Hessian 矩阵的估计是正定的。

很多分析是基于约束 Broyden 方法进行的，即要求 $\phi_k \in [0, 1]$。对于二次函数，这种方法有下述许多性质。由于分析都是与步长无关的，因此我们假设

$$p_k = -B_k^{-1} \nabla f_k, \quad x_{k+1} = x_k + p_k \tag{30}$$

**定理 3.** *假设 $f$ 是强凸二次函数 $f(x) = b^T x + \frac{1}{2} x^T A x$，其中 $A$ 对称正定。记 $x_0$ 是任意的初始点，且 $B_0$ 是任意的正定初始矩阵，假设 $B_k$ 是由 Broden 公式得到的，且 $\phi_k \in [0, 1]$。记 $\lambda_1^k \le \lambda_2^k \le \cdots \le \lambda_n^k$ 是*

$$A^{\frac{1}{2}} B_k^{-1} A^{\frac{1}{2}} \tag{31}$$

*的特征值。于是对于所有的 $k$，有*

$$\min\{\lambda_i^k, 1\} \le \lambda_i^{k+1} \le \max\{\lambda_i^k, 1\}, \quad i = 1, 2, \ldots, n \tag{32}$$

*此外，该性质在 $\phi_k$ 不在 $[0, 1]$ 区间内时不成立。*

如何理解这个定理的意义呢？首先，如果 (31) 的特征值全是 1，那么 $B_k$ 等于 $A$。因此我们希望这些特征值尽可能接近 1。而 (32) 告诉我们，特征值单调收敛到 1。虽然我们并不能保证它们收敛到 1，但是我们可以希望它们有这样的性质。然而假设我们允许 $\phi_k$ 不在 $[0, 1]$ 内，特征值可能会远离 1。特别地，即使线性搜索法不精确时，上述定理的结果仍成立。

虽然从上述定理看来，约束 Broyden 法是比较好的。然而有些分析和实验结果也表明，在允许 $\phi_k$ 取负数的情况下，算法性能可能比 BFGS 法更好。SR1 法就是一个典型的例子，它是下述 Broyden 法:

$$\phi_k = \frac{s_k^T y_k}{s_k^T y_k - s_k^T B_k s_k}$$

显然它不是约束 Broyden 法。

后续我们将更具体地讨论如何选取 $\phi_k$ 以保证正定性。

(28) 实质上是一个秩一校正，由交错特征值定理可知，它会在 $\phi_k > 0$ 时增大特征值。于是 $\phi_k \geq 0$ 时，$B_{k+1}$ 正定。另一方面，由同一个定理知，它会在 $\phi_k < 0$ 时减小特征值。于是可能会让矩阵变得非奇异非正定。可以通过计算知，当 $\phi_k$ 取下述值，$B_{k+1}$ 将变为奇异阵

$$\phi_k^C = \frac{1}{1-\mu_k} \tag{33}$$

其中，

$$\mu_k = \frac{(y_k^T B^{-1} y_k)(s_k^T B_k s_k)}{(y_k^T s_k)^2} \tag{34}$$

由柯西不等式知，$\mu_k \geq 0$。

当线性搜索法精确，那么所有选取 $\phi_k \geq \phi_k^C$ 的 Broyden 法都会产生相同的迭代。从实验结果来看，在使用精确线性搜索法时，这个结果对于更一般的非线性函数也是成立的，不同 Broyden 法的不同仅仅是其产生的搜索方向的长度不同。

对于二次函数问题，在使用精确线性搜索法时，Broyden 法有一些非常引人注意的性质。下面我们将不加证明地去叙述这些性质。

**定理 4.** *假设我们用 Broyden 法去优化一个强凸二次函数 $f(x) = b^T x + \frac{1}{2} x^T A x$，初始点 $x_0$ 和初始对称正定矩阵 $B_0$ 都是任意的。假设 $\alpha_k$ 是精确步长且对于所有 $k$，都有 $\phi_k \geq \phi_k^C$，则我们有：*

(i) 迭代与 $\phi_k$ 的选取无关，并且至多 $n$ 步迭代收敛到解

(ii) 切方程对于所有之前的搜索方向都成立，即

$$B_k s_j = y_j, \quad j = k-1, k-2, \ldots, 1$$

(iii) 假设初始矩阵是 $B_0 = I$，那么迭代与共轭梯度法相同。特别地，搜索方向是共轭的，即

$$s_i^T A s_j = 0, \quad \text{for } i \neq j$$

(iv) $n$ 次迭代后有 $B_n = A$

我们可以对上述结果做一些简单的拓展：它在 Hessian 估计非奇异但非正定时仍然是成立的。也即我们允许 $\phi_k$ 小于 $\phi_k^C$，只要它产生的矩阵非奇异。对于 (iii)，我们有：如果 $B_0$ 不是单位阵，那么该方法等价于一个以 $B_0$ 作为预条件子的预条件共轭梯度法。

然而前述定理主要贡献是它的理论价值，因为在实际应用中，我们常常使用非精确搜索，此时，算法的性能会大大不同。但是这样的分析对拟牛顿法的发展还是做了很多贡献。

# 4 收敛性分析

本节我们将讨论 BFGS 法和 SR1 法在实践中的全局和局部收敛性。Hessian 矩阵估计采用迭代公式的均值让拟牛顿法的分析比最速下降法和牛顿法复杂很多。

虽然这两个方法在实践中都表现出了很好的鲁棒性，然而我们还没法给出它们对于一般非线性函数的全局收敛性。也就是说我们不能保证对于任意的初值和矩阵估计，算法都能收敛到一个稳定点。我们在分析中要么假设函数是凸的要么就假设迭代满足某些性质。而另一方面，在一些合理假设下局部收敛有超线性收敛的结果。

本节我们用 $\|\cdot\|$ 表示欧几里得范数，用 $G(x)$ 表示 $\nabla^2 f(x)$。

## 4.1 BFGS 法的全局收敛性

我们首先研究一下 BFGS 法在使用实际的线性搜索法时的收敛性，这里我们假设初始点任意，初始矩阵为任意对称正定阵。更具体的假设如下：

**假设 1.** *(i)* $f$ 二次可微

*(ii)* 水平集 $\mathcal{L} = \{x \in R^n | f(x) \le f(x_0)\}$ 是凸集，且存在正常数 $m$ 和 $M$ *s.t.*

$$m\|z\|^2 \le z^T G(x) z \le M\|z\|^2, \quad \forall z \in R^n, x \in \mathcal{L} \tag{35}$$

第二条假设说明 $G(x)$ 在 $\mathcal{L}$ 上正定，于是 $f$ 在 $\mathcal{L}$ 上有唯一极值 $x^*$。

利用 (12) 和 (35) 有

$$\frac{y_k^T s_k}{s_k^T s_k} = \frac{s_k^T \bar{G}_k s_k}{s_k^T s_k} \ge m \tag{36}$$

其中 $\bar{G}_k$ 是 (11) 定义的平均 Hessian 矩阵。利用假设，我们有 $\bar{G}$ 也是正定的，于是它是存住平方根矩阵的。因此，如同 (21) 一样，我们定义 $z_k = \bar{G}_k^{1/2}$，于是有

$$\frac{y_k^T y_k}{y_k^T s_k} = \frac{s_k^T \bar{G}_k^2 s_k}{s_k^T \bar{G}_k s_k} = \frac{z_k^T \bar{G}_k z_k}{z_k^T z_k} \le M \tag{37}$$

至此我们已经可以给出 BFGS 法收敛性的分析了。由于很难直接去分析 $B_k$ 的条件数，因此我们需要引入一些新的技巧。这里我们将从迹和行列式入手，对 $B_k$ 的最大最小的特征值进行估计。矩阵的迹是所有特征值的和，矩阵的行列式是所有特征值的积。

**定理 5.** 取 $B_0$ 为任意的对称正定矩阵，$x_0$ 是满足假设 *1*的初值。那么由算法 *1*产生的迭代序列 $\{x_k\}$ 收敛到 $f$ 的极小子 $x^*$，其中 $\epsilon = 0$。

**证明**

首先定义：

$$m_k = \frac{y_k^T s_k}{s_k^T s_k}, \quad M_k = \frac{y_k^T y_k}{y_k^T s_k} \tag{38}$$

由前述分析有

$$m_k \ge m, \quad M_k \le M \tag{39}$$

利用 (19) 知，BFGS 中迹的更新公式如下：

$$\operatorname{trace}(B_{k+1}) = \operatorname{trace}(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} \tag{40}$$

行列式的更新公式如下：

$$\det(B_{k+1}) = \det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k} \tag{41}$$

下面我们定义：

$$\cos\theta_k = \frac{s_k^T B_k s_k}{\|s_k\|\|B_k s_k\|}, \quad q_k = \frac{s_k^T B_k s_k}{s_k^T s_k} \tag{42}$$

即 $\theta_k$ 表示 $s_k$ 和 $B_k s_k$ 之间的家教，我们有

$$\frac{\|B_k s_k\|^2}{s_k^T B_k s_k} = \frac{\|B_k s_k\|^2\|s_k\|^2}{(s_k^T B_k s_k)^2}\frac{s_k^T B_k s_k}{\|s_k\|^2} = \frac{q_k}{\cos^2\theta_k} \tag{43}$$

此外，

$$\det(B_{k+1}) = \det(B_k)\frac{y_k^T s_k}{s_k^T s_k}\frac{s_k^T s_k}{s_k^T B_k s_k} = \det(B_k)\frac{m_k}{q_k} \tag{44}$$

我们将用下面这个函数将迹和行列式结合起来：

$$\psi(B) = \operatorname{trace}(B) - \ln(\det(B)) \tag{45}$$

其中，$\ln(\cdot)$ 表示自然对数，不难证明 $\psi(B) > 0$。于是由前述关系，我们有

$$\begin{aligned}\psi(B_{k+1}) =& \operatorname{trace}(B_k) + M_k - \frac{q_k}{\cos^2\theta_k} - \ln(\det(B_k)) - \ln(m_k) + \ln(q_k)\\ =& \psi(B_k) + (M_k - \ln(m_k) - 1)\\ &+ \left[1 - \frac{q_k}{\cos^2\theta_k} + \ln\frac{q_k}{\cos^2\theta_k}\right] + \ln\cos^2\theta_k\end{aligned} \tag{46}$$

由于 $h(t) = 1 - t + \ln t$ 对于所有的 $t > 0$ 都非负，中括号里的项是非负的，于是有

$$0 < \psi(B_{k+1}) \le \psi(B_k) + c(k+1) + \sum_{j=0}^{k}\ln\cos^2\theta_j \tag{47}$$

不失一般性，我们可以取 $c = M - \ln m - 1$。

下面我们将利用 3.2 节中的结果完成论证。注意到 (42) 定义的角度其实就是搜索方向与最速下降法方向的夹角，于是我们可以利用 3.2 中的结果去分析全局收敛性，即当且仅当 $\cos\theta_j \to 0$ 时，$\|\nabla f_k\|$ 会有界远离 0。

下面我们假设 $\cos\theta_j \to 0$，于是存在 $k_1 > 0$s.t.

$$\ln\cos^2\theta_j < -2c, \quad \forall j > k$$

带入 (47) 有

$$0 < \psi(B_0) + c(k+1) + \sum_{j=0}^{k_1} \ln\cos^2\theta_j + \sum_{j=k_1+1}^{k} (-2c)$$
$$= \psi(B_9) + \sum_{j=0}^{k_1} \ln\cos^2\theta_j + 2ck_1 + c - ck$$

然而，右式在 $k$ 很大的时候显然是负的，矛盾，从而下极限收敛到 0，特别的，对于凸函数，我们便得到了全局收敛性。

上述定理可以拓展到除了 DFP 法外的所有约束 Broyden。换句话说，该命题在 $\phi_k \in [0,1)$ 时均成立，但在 $\phi_k$ 接近 1 时可能会由于自我校正能力的缺失而出现问题。

从前述分析，我们可以得到迭代的收敛速率至少是线性的。然而我们还可以证明 $\|x_k - x^*\|$ 收敛到 0 的速率足够快 s.t.

$$\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty \tag{48}$$

我们不会去证明这个命题，但是我们将基于这个命题说明收敛速率其实是超线性的。

## 4.2 BFGS 法的超线性收敛性

这里的分析用了 Dennis 和 Moré 提出的关于超线性收敛的特征 (3.36)。它适用于一般的非线性函数（不仅仅是凸函数）。但是我们需要增加下述假设：

**假设 2.** *Hessian 矩阵 $G$ 在 $x^*$ 处 Lipschitz 连续，即*

$$\|G(x) - G(x^*)\| \le L\|x - x^*\|, \quad \forall x \ near \ x^*$$

b

首先我们引入下述记号：

$$\tilde{s}_k = G_*^{1/2} s_k, \quad \tilde{y}_k = G_*^{-1/2} y_k, \quad \tilde{B}_k = G_*^{-1/2} B_k G_*^{-1/2}$$

其中，$G_* = G(x^*)$，且 $x^*$ 是 $f$ 的极小子。与 (42) 相似，我们定义

$$\cos\tilde{\theta}_k = \frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\|\|\tilde{B}_k \tilde{s}_k\|}, \quad \tilde{q}_k = \frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\|^2}$$

类比于 (38) 和 (39) 有

$$\tilde{M}_k = \frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T \tilde{s}_k}, \quad \tilde{m}_k = \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{s}_k}$$

通过对 (19) 预乘以 $G_*^{-1/2}$，我们有

$$\tilde{B}_{k+1} = \tilde{B}_k - \frac{\tilde{B}_k \tilde{s_k} \tilde{s_k}^T \tilde{B}_k}{\tilde{s}_k^T \tilde{B}_k \tilde{s_k}} + \frac{\tilde{y}_k \tilde{y}_k^T}{\tilde{y}_k^T \tilde{s}_k}$$

因为这个表达式和原来的表达式形式完全相同，所以类似于 (46)，有

$$\begin{aligned} \psi(\tilde{B}_{k+1}) ==& \psi(\tilde{B}_k) + (\tilde{M}_k - \ln(\tilde{m}_k) - 1) \\ &+ \left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right] + \ln \cos^2 \tilde{\theta}_k \end{aligned} \tag{49}$$

回忆 (12)，有

$$y_k - G_* s_k = (\bar{G}_k - G_*) s_k$$

因此

$$\tilde{y}_k - \tilde{s}_k = G_*^{-1/2} (\bar{G}_k - G_*) G_*^{-1/2} \tilde{s}_k$$

利用假设 2，并回忆 (11)，我们有

$$\|\tilde{y}_k - \tilde{s}_k\| \leq \|G_*^{-1/2}\|^2 \|\tilde{s}_k\| \|\bar{G}_k - G_*\| \leq \|G_*^{-1/2}\|^2 \|\tilde{s}_k\| L \epsilon_k$$

其中 $\epsilon_k = \max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$

于是对于某些正常数 $\bar{c}$，有

$$\frac{\|\tilde{y}_k - \tilde{s}_k\|}{\|\tilde{s}_k\|} \leq \bar{c} \epsilon_k \tag{50}$$

不等式 (48) 在证明超线性收敛的过程中起着重要的作用。

**定理 6.** *假设 $f$ 二次连续可微，则当假设 2成立时，BFGS 算法产生的迭代点收敛到极小子 $x^*$。假设 (48) 成立，则收敛速率是超线性的。*

## 4.3 SR1 法的收敛性分析

SR1 方法的收敛性质其实还没有像 BFGS 法研究得那样透彻。我们既没有如同前述的全局收敛结果，也没有如同前述的超线性收敛结果。但是这里我们将叙述信赖域 SR1 方法的一个有趣的结果。

**定理 7.** *记 $x_k$ 为算法 2产生的迭代结果，且下述条件满足：*

*c1* 迭代序列不终止，但是停留在一个闭有界凸集 $D$ 中，$f$ 在 $D$ 上二次连续可微，且在 $D$ 中有唯一稳定点 $x^*$

*c2* Hessian 矩阵 $f(x^*)$ 正定，且 $\nabla^2 f(x)$ 在 $x^*$ 附近 Lipschitz 连续

*c3* $\{B_k\}$ 的范数有界

*c4* 条件 (24) 在每次迭代的时候都成立，其中 $r$ 是一个 (0,1) 范围内的常数

那么 $\lim_{k\to\infty} x_k = x^*$，并且

$$\lim_{k\to\infty} \frac{\|x_{k+n+1} - x^*\|}{\|x_k - x^*\|} = 0$$

注意到 BFGS 法并不要求有界性条件。也如同我们之前提到的，SR1 迭代并不能保证矩阵 $B_k$ 的正定性。在实际使用中，$B_k$ 可能是非正定的，也就是说，信赖域边界可能始终是起作用的。然而有趣的是，我们可以说明 SR1 算法在大部分时间里矩阵都是正定的。更具体的结果如下：在定理 7的假设成立的条件下，

$$\lim_{k\to\infty} \frac{\text{number of indices } j = 1, 2, \ldots, k \text{ for which } B_j \text{ is positive semidefinite}}{k} = 1$$

而且这个结果跟初值矩阵是否正定无关。