Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5435-5443
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118572

Hans汉斯

Gauss-Newton BFGS方法所产生的迭代矩阵序列的收敛性

陈初阳

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2022年7月8日；录用日期：2022年8月2日；发布日期：2022年8月11日

摘　要

收敛速度的快慢是决定一个算法好坏的重要因素。在拟牛顿算法中，算法的收敛性在某种程度上等价于Dennis-Moré条件，但这并不意味着算法所产生的迭代矩阵就会收敛到Hessian矩阵。本文证明了由求解对称非线性方程组的Gauss-Newton BFGS方法所产生的迭代矩阵序列的收敛性，并通过数值实验对结论进行验证。

关键词

BFGS方法，收敛性

The Convergence of Iterate Matrices Sequences Generated by Gauss-Newton BFGS Methods

Chuyang Chen

School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha Hunan

Received: Jul. 8th, 2022; accepted: Aug. 2nd, 2022; published: Aug. 11th, 2022

Abstract

The speed of convergence is an important factor that determines the quality of an algorithm. In the quasi-Newton algorithm, the convergence of the algorithm is equivalent to the Dennis-Moré condition to some extent, but this does not mean that the iterative matrix generated by the algo-

文章引用: 陈初阳. Gauss-Newton BFGS 方法所产生的迭代矩阵序列的收敛性[J]. 应用数学进展, 2022, 11(8): 5435-5443.
DOI: 10.12677/aam.2022.118572

rithm will converge to the Hessian matrix. This paper proves the convergence of the iterative matrix sequence generated by the Gauss-Newton BFGS method for solving symmetric nonlinear equations, and validates the conclusion by numerical experiments.

Keywords

BFGS Method, Convergence

1. 引言

我考虑如下非线性方程组：

$$g(x)=0,\quad x\in R^n.$$

其中 g 是 R^n 到 R^n 上的可微映射且 g 的雅可比阵 $\nabla g(x)$ 是对称的。在这种情况下我把 g 看作某个函数 f 从 R^n 到 R 的梯度映射，从而 $g(x)=\nabla f(x)$，$\nabla g(x)=\nabla^2 f(x)$，$g(x)=0$ 就是无约束优化问题：

$$\min_{x\in R^n} f(x).$$

的一阶必要条件。

DONGHUI LI and MASAO FUKUSHIMA 在[1]中提出了处理此类问题的 Gauss-Newton BFGS 方法，其迭代格式为：

$$x_{k+1}=x_k+\lambda_k p_k,$$

$$p_k:=-\frac{B_k^{-1}\left(g\left(x_k+\lambda_{k-1}g_k\right)-g_k\right)}{\lambda_{k-1}},\quad g_k:=g(x_k).$$

其中 p_k 是搜索方向，步长 $\lambda_k>0$ 且使得序列 $\{g_k\}$ 在总体上具有近似范数下降性。矩阵 B_k 由 BFGS 公式进行更新来确保它的正定性从而 p_k 就是映射 g 在迭代点 x_k 处的下降方向，因此 B_k 满足割线方程：

$$B_{k+1}s_k=y_k,$$

其中

$$s_k:=x_{k+1}-x_k,\quad y_k:=g\left(x_k+\delta_k\right)-g_k,\quad \delta_k=g_{k+1}-g_k.$$

和常用的 BFGS 公式不同的是我经常用 y_k 来表示梯度差而这里用 δ_k 表示，在这种情况下我有以下近似关系：

$$y_k\approx\nabla g_{k+1}\delta_k\approx\nabla g_{k+1}\nabla g_{k+1}s_k.$$

又因为 B_{k+1} 满足割线方程且雅可比阵 ∇g_k 是对称的，我又有以下近似关系：

$$B_{k+1}s_k\approx\nabla g_{k+1}\nabla g_{k+1}s_k\approx\nabla g_{k+1}\nabla g_{k+1}^{\mathrm{T}}s_k.$$

这就意味着 B_{k+1} 沿 s_k 方向近似于 $\nabla g_{k+1}\nabla g_{k+1}^{\mathrm{T}}$，又因为

$$\lambda_{k-1}^{-1}\left(g\left(x_k+\lambda_{k-1}g_k\right)-g_k\right)\approx\nabla g_k g_k,$$

所以我得到

$$B_k p_k + \nabla g_k g_k \approx 0.$$

因此我把 p_k 看作 Gauss-Newton 的近似方向，把这种方法称为 Gauss-Newton-based BFGS 方法。该方法的优点在于不需要计算雅可比阵并且在合适的条件下具有全局收敛性和超线性收敛性。

我让函数 f 满足以下假设：

1) 函数 f 在 R^n 上二阶连续可微。

2) $\exists M > 0$，对 $\forall x \in \Omega_1$，$\left\|\nabla^2 f(x)\right\| \le M$。

3) $\nabla^2 f(x)$ 在 Ω_1 上一致非奇异。

其中 Ω_1 的定义见[1]中假设 A。

在以上假设条件下，由[1]我可以得到以下结论：

4) $\exists x^*$ 使得 $g(x^*) = 0$。

5) $\sum_k \left\|x_k - x^*\right\| < \infty$，$\sum_k \|s_k\| < \infty$。

6) 当 $k \to \infty$ 时，B_{k+1} 总是由 BFGS 公式更新。

我进一步假设：

7) $\nabla^2 f(x^*)$ 是正定的且 $\nabla^2 f(x)$ 在 x^* 处 Lipschitz 连续，即 $\left\|\nabla^2 f(x) - \nabla^2 f(x^*)\right\| \le L\left\|x - x^*\right\|$，其中 x 属于 x^* 的一个邻域。

同样我从[1]中得到以下结论：

8) $\sup_k \|B_k\| < \infty$，$\sup_k \left\|B_k^{-1}\right\| < \infty$。

9) 当 $k \to \infty$ 时，λ_k 恒等于 1。

类似于常用的 BFGS 方法，超线性收敛性在某种程度上等价于 Dennis-Moré 条件[2]，但由于 y_k 的取不同，条件的表现形式也有所改变，具体如下：

$$\lim_{k\to\infty} \frac{\left\|\left(B_k - \nabla g_* \nabla^{\mathrm{T}} g_*\right) s_k\right\|}{\|s_k\|} = 0.$$

但这并不意味着 B_k 就一定最终等于 $\nabla g_* \nabla^{\mathrm{T}} g_*$ (由[2]中的反例我可以得到这个结论)。最早证明相关结论的是 Ge Ren-Pu and Powell [3]中证明了 DFP 方法和 BFGS 方法取恒定步长为一所产生的迭代矩阵序列的收敛性，该证明不要求 B_k 最终等于 $\nabla^2 f(x^*)$，之后 Stoer [4]把该结论推广到 Broyden 族方法上且步长最终收敛到 1 即可。由于 1)到 9)的结论满足[4]的假设 B，所以本文的证明参照[4]进行了一点点的改动，以下我统称为假设最后我给出 Gauss-Newton-based BFGS 算法的框架。

Gauss-Newton-based BFGS 算法

Step 0. Choose an initial point $x_0 \in R^n$, an initial symmetric positive definite matrix $B_0 \in R^{n\times n}$, a positive sequence $\{\omega_k\}$ satisfying $\sum_{k=0}^{\infty} \omega_k < \infty$, and constants $r, \rho \in (0,1)$, $\sigma_1, \sigma_2 > 0, \lambda_{-1} > 0$. Let $k := 0$.

Step 1. Stop if $g_k = 0$. Otherwise, solve the following linear equation to get p_k:

$$B_k p + \lambda_{k-1}^{-1}\left(g\left(x_k + \lambda_{k-1} g_k\right) - g_k\right) = 0.$$

Step 2. If

$$\left\|g\left(x_k+p_k\right)\right\| \le \rho\left\|g_k\right\|,$$

then take $\lambda_k=1$ and go to Step 4. Otherwise go to Step 3.

Step 3. Let i_k be the smallest nonnegative integer i such that:

$\left\|g\left(x_k+\lambda p_k\right)\right\|^2-\left\|g_k\right\|^2 \le -\sigma_1\left\|\lambda g_k\right\|^2-\sigma_2\left\|\lambda p_k\right\|^2+\omega_k\left\|g_k\right\|^2$ holds for $\lambda=r^i$.

Let $\lambda_k=r^{i_k}$.

Step 4. Let the next iterate be $x_{k+1}=x_k+\lambda_k p_k$.

Step 5. Put $s_k=x_{k+1}-x_k=\lambda_k p_k$, $\delta_k=g_{k+1}-g_k$, and $y_k=g\left(x_k+\delta_k\right)-g\left(x_k\right)$. If $y_k^{\mathrm{T}}s_k \le 0$, then $B_{k+1}=B_k$ and go to Step 6. Otherwise, update B_k by the BFGS formula:

$$B_{k+1}=B_k-\frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k}+\frac{y_k^{\mathrm{T}} y_k}{s_k^{\mathrm{T}} y_k}.$$

Step 6. Let $k:=k+1$. Go to Step 1.

2. 收敛性的证明

由 BFGS 方法的不变性，我假设 $\nabla^2 f\left(x^*\right)\nabla^2 f\left(x^*\right)^{\mathrm{T}}=I$。如果函数 $f(x)$ 是二次函数，那么 $g(x)$ 就是线性的，那么此时

$$p_k=-B_k^{-1}g,\quad y_k=g_{k+1}-g_k=\delta_k,$$

我直接由[3]得到结论，对于一般情况我有

$$\begin{aligned} y_k &= g\left(x_k+\delta_k\right)-g_k=\int_0^1 \nabla^2 f\left(x_k+\theta\delta_k\right)\mathrm{d}\theta\delta_k \\ &= \int_0^1 \nabla^2 f\left(x_k+\theta\delta_k\right)\mathrm{d}\theta\int_0^1 \nabla^2 f\left(x_k+\theta s_k\right)\mathrm{d}\theta s_k=Q_k P_k s_k \end{aligned}$$

其中 $Q_k=\int_0^1 \nabla^2 f\left(x_k+\theta\delta_k\right)\mathrm{d}\theta$，$P_k=\int_0^1 \nabla^2 f\left(x_k+\theta s_k\right)\mathrm{d}\theta$。

有假设得知当 k 充分大时，P_k 是正定矩阵且

$$P_k=\mathrm{I}+O\left(\left\|s_k\right\|\right),$$

同理当 k 充分大时

$$\delta_k=g_{k+1}-g_k=\int_0^1 \nabla^2 f\left(x_k+ts_k\right)\mathrm{d}\theta s_k=s_k+O\left(\left\|s_k\right\|^2\right),$$

$$y_k=\left[\mathrm{I}+O\left(\left\|\delta_k\right\|\right)\right]\left[\mathrm{I}+O\left(\left\|s_k\right\|\right)\right]\mathrm{d}\theta s_k=s_k+O\left(\left\|s_k\right\|^2\right).$$

因此由 BFGS 公式我有

$$\begin{aligned} B_{k+1} &= B_k-\frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k}+\frac{y_k^{\mathrm{T}} y_k}{s_k^{\mathrm{T}} y_k} \\ &= B_k-\frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k}+\frac{s_k^{\mathrm{T}} s_k}{s_k^{\mathrm{T}} s_k}+O\left(\left\|s_k\right\|\right) \\ &= B_{k+1}'+O\left(\left\|s_k\right\|\right) \end{aligned} \tag{1}$$

我定义 $E_k:=B_k-\mathrm{I}$，$\lambda_{ki}, i=1,2,\cdots,n$ 是它的特征值且 $\left|\lambda_{k1}\right| \le \left|\lambda_{k2}\right| \le \cdots \le \left|\lambda_{kn}\right|$，$\upsilon_{ki}, i=1,2,\cdots,n$ 是特征值对应的正交特征向量，$E_k\upsilon_{ki}=\lambda_{ki}\upsilon_{ki}$。由[5]中的结论我得到对于 $E_{k+1}:=B_{k+1}-\mathrm{I}$ 的特征值 $\left\{\lambda_{k+1,i}\right\}$ 按照同样

的方式排列$\left|\lambda_{k+1,1}\right| \le \left|\lambda_{k+1,2}\right| \le \cdots \le \left|\lambda_{k+1,n}\right|$，它满足

$$\left|\lambda_{k+1,i}\right| \le \left|\lambda_{k,i}\right|,$$

$$sign\lambda_{k+1,i} = sign\lambda_{k,i},\ \ i = 1,2,\cdots,n.$$

由(1)以及$\sum_k \|s_k\|$收敛，根据文献[3]中引理 4 可知对每个$i = 1,2,\cdots,n$，极限$\lim_{k\to\infty}\lambda_{ki}$存在。我假设极限趋于 0 的特征值个数为$m$，其余的都大于$\beta > 0$，即

$$\lim_k \lambda_{ki} = 0,\ \ i = 1,2,\cdots,n$$

$$\left|\lambda_{ki}\right| \ge \beta,\ \ i = m+1,\cdots,n.$$

又因为

$$\frac{s_k^{\mathrm{T}} E_{k+1} s_k}{s_k^{\mathrm{T}} s_k} = \frac{s_k^{\mathrm{T}}\left(y_k - s_k\right)}{s_k^{\mathrm{T}} s_k} = \frac{s_k^{\mathrm{T}}\left(Q_k P_k - \mathrm{I}\right)s_k}{s_k^{\mathrm{T}} s_k} = O\left(\|s_k\|\right)$$

而$\lim_k \|s_k\| = 0$，所以$m \ge 1$。

由[3]我对E_k进行如下分解

$$E_k := \Delta_k + H_k$$

其中

$$\Delta_k := \sum_{i=1}^{m} \lambda_{ki}\upsilon_{ki}\upsilon_{ki}^{\mathrm{T}},\ \ H_k := \sum_{i=m+1}^{n} \lambda_{ki}\upsilon_{ki}\upsilon_{ki}^{\mathrm{T}},$$

$$S_k := span\left\{\upsilon_{k1},\upsilon_{k2},\cdots,\upsilon_{km}\right\}.$$

那么$S_k^{\perp} = span\left\{\upsilon_{k,m+1},\cdots,\upsilon_{k,n}\right\}$。

由以上定义我有$\Delta_k S_k \subset S_k$，$H_k S_k^{\perp} \subset S_k^{\perp}$，$\Delta_k S_k^{\perp} = H_k S_k = 0$，根据$m$的定义有$\lim_k \Delta_k = 0$。

通过以上讨论，我将通过证明$\sum_k \|H_{k+1} - H_k\| < \infty$来证明序列$\{B_k\}$的收敛性。我将证明

$$\|H_{k+1} - H_k\|_F \le c_1 \|\eta_k\| / \|s_k\| + c_2 \|s_k\| \tag{2}$$

对充分大的k都成立。这里η_k取自s_k的正交分解

$$s_k = \gamma_k + \eta_k, \gamma_k \in S_k, \eta_k \in S_k^{\perp}.$$

(2)式的证明就是[3]中对 DFP 方法的证明，这里不再赘述。

因为$\sum_k \|s_k\| < \infty$，由(2)我只需证明$\sum_k \|\eta_k\| / \|s_k\| < \infty$。首先证明一个重要不等式

$$\sum_k \frac{\|E_k s_k\|^2}{\|s_k\|^2} < \infty \tag{3}$$

$$\begin{aligned} E_{k+1} &= B_{k+1} - \mathrm{I} = B'_{k+1} - \mathrm{I} + O\left(\|s_k\|\right) \\ &= E_k - \frac{\left(\mathrm{I} + E_k\right)s_k s_k^{\mathrm{T}}\left(\mathrm{I} + E_k\right)}{\|s_k\|^2 + s_k^{\mathrm{T}} E_k s_k} + \frac{s_k s_k^{\mathrm{T}}}{\|s_k\|^2} + O\left(\|s_k\|\right) \end{aligned}$$

由文献[6]的定理 2 我有

$$\left\|E_{k+1}\right\|_F^2 \le \left\|E_k\right\|_F^2 - \frac{\left(E_k s_k\right)^{\mathrm{T}} B_k \left(E_k s_k\right)}{s_k^{\mathrm{T}} B_k s_k} + O\left(\left\|s_k\right\|\right),$$

不等式两边对 k 进行累加，又因为 $\{B_k\}$ 和 $\{B_k^{-1}\}$ 都是正定有界的，(3)式得证。从文献[4]我立马又得到 $\sum_k \left\|B_{k+1} - B_k\right\| < \infty$，证明没有任何改变。

接下来我证明 $\sum_k \left\|\eta_k\right\| / \left\|s_k\right\| < \infty$，由 $k \to \infty$ 时 $\lambda_k \equiv 1$ 得，存在常数 k_1，当 $k \ge k_1$ 时，$\lambda_k = 1$。下面我证明几个不等式，其中 $\bar{E} = B_k - Q_k P_k$。

a) $\sum_{k_1} \frac{\left\|E_k \bar{E}_k s_k\right\|^2}{\left\|\bar{E}_k s_k\right\|^2} < \infty$,

b) $\sum_{k_1} \frac{\left\|\bar{E}_k s_k\right\|^2}{\left\|s_k\right\|^2} < \infty$,

c) $\sum_{k_1} \frac{\left\|E_k \bar{E} s_k\right\|}{\left\|s_k\right\|} < \infty$,

d) $\sum_{k_1} \frac{\left\|E_k^2 s_k\right\|}{\left\|s_k\right\|} < \infty$.

证明和[4]中类似但是有两个小变化。一个是当 $k \ge k_1$ 时步长 $\lambda_k = 1$，这时

$$p_k = s_k = -B_k^{-1}\left(g\left(x_k + g_k\right) - g_k\right) = -B_k^{-1}\int_0^1 \nabla^2 f\left(x_k + l g_k\right) \mathrm{d}l g_k = -B_k^{-1} G_k g_k$$

另一个是

$$\begin{aligned} \bar{E}_k s_k &= B_k s_k - y_k = -G_k g_k - Q_k \delta_k \\ &= \left(Q_k - G_k\right) g_k - Q_k g_{k+1} \\ &= \left(Q_k - G_k\right) g_k + Q_k G_{k+1}^{-1} B_{k+1} s_{k+1} \end{aligned}$$

当 k 充分大时，由于函数 f 二阶可微可得 $Q_k - G_k \to 0$，所以上式最后一个等式的第一部分趋于 0，由假设 2)，3)以及 B_k 的有界性得证。

现在我的结论显而易见，由 $E_k := \Delta_k + H_k$，$s_k = \gamma_k + \eta_k$ 以及 $\left|\lambda_{ki}\right| \ge \beta, i = m+1, \cdots, n$ 我有

$$\frac{\left\|E_k^2 s_k\right\|}{\left\|s_k\right\|} = \frac{\left\|\Delta_k^2 \gamma_k + H_k^2 \eta_k\right\|}{\left\|s_k\right\|} \ge \frac{\left\|H_k^2 \eta_k\right\|}{\left\|s_k\right\|} \ge \beta^2 \frac{\left\|\eta_k\right\|}{\left\|s_k\right\|}$$

所以 $\sum_k \left\|\eta_k\right\| / \left\|s_k\right\| < \infty$，得证。

3. 数值实验

在这部分我通过画图来观察 $\left\|B_k - \nabla g_* \nabla^{\mathrm{T}} g_*\right\|$ 的收敛性，我在三个不同的方面进行比较：初始点、维度、和精度。我把 ∇g_* 记作 j，这里所取的问题就是[1]中的原问题，是为了验证自己的编程是否正确。

问题 1

$$g(x) \triangleq Ax + \frac{1}{(n+1)^2} F(x) = 0,$$

其中 $A \in R^{n\times n}$

$$A=\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix},$$

$$F(x)=\left(F_1(x),F_2(x),\cdots,F_n(x)\right)^{\mathrm{T}}$$

$$F_i(x)=\sin x_i-1,\quad i=1,2,\cdots,n.$$

以下是不同初始点(见表 1)的影响(见图 1)，维度 $n=19$。

Table 1. Different initial points and their manifestations
表 1. 不同初始点及其表现形式

x_0	$(1,1,\cdots,1)$	$(0,0,\cdots,0)$	$(1,0,1,0,\cdots)$	$(0,1,0,1,\cdots)$	$(1,2,3,\cdots)$
表现形式	−b+	−go	−r*	−mx	−yd

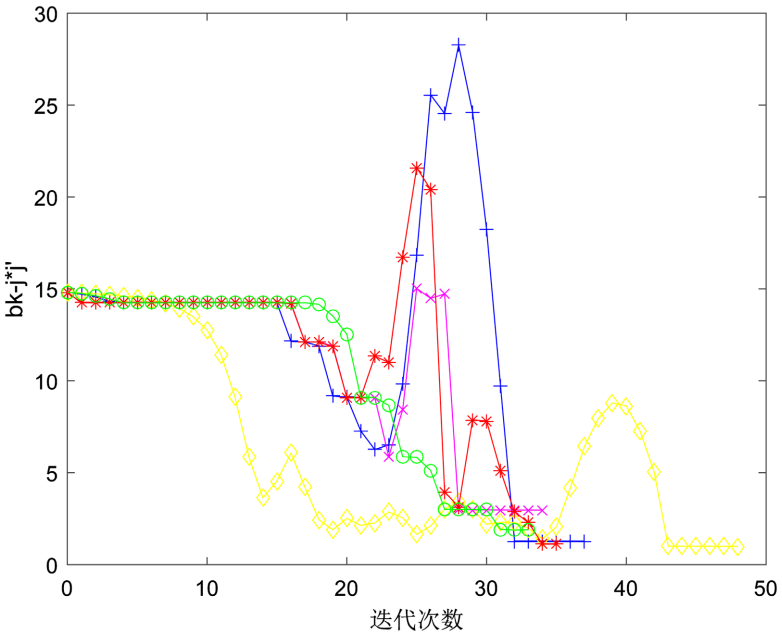

Figure 1. Performance of different starting points
图 1. 不同初始点的表现

以下是不同维度(见表 2)的影响(见图 2)，初始点为全 1 向量。

Table 2. Different dimensions and their manifestations
表 2. 不同维度及其表现形式

维度	$n=19$	$n=39$	$n=59$	$n=79$	$n=99$
表现形式	−b+	−go	−r*	−mx	−yd

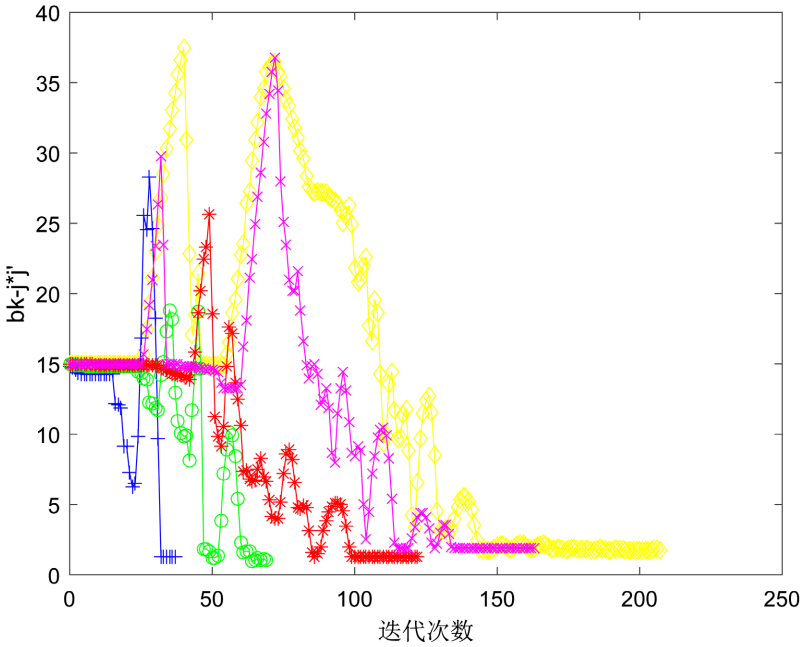

Figure 2. Performance of different dimensions
图 2. 不同维度的表现

以下是不同精度(见表 3)的影响(见图 3)，维度 $n = 19$，初始点设置为全 1 向量。

Table 3. Precision
表 3. 精度

精度	10^{-7}	10^{-9}	10^{-11}	10^{-13}	10^{-15}

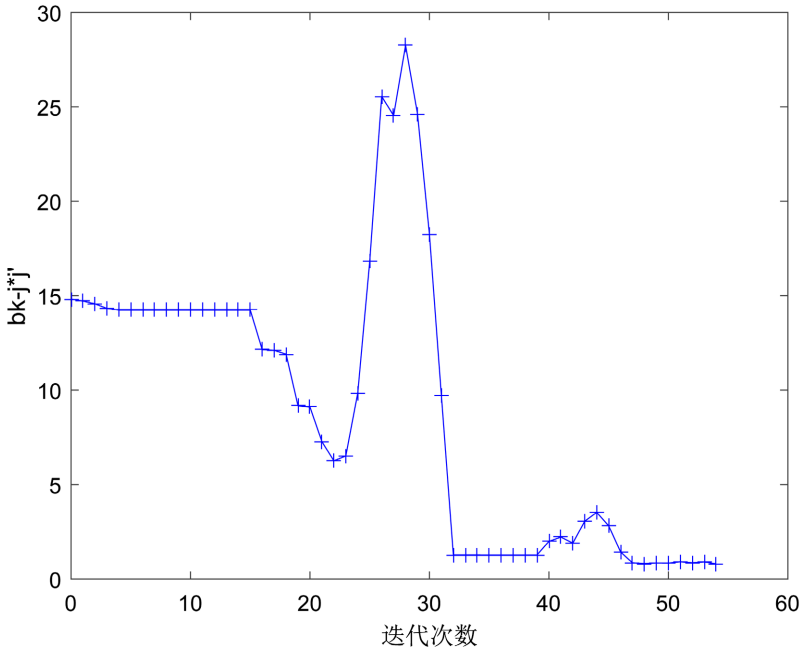

Figure 3. Performance of different precision
图 3. 不同精度的表现

在实验中对应的参数为 $r=0.1$，$\rho=\sqrt{0.9}$，$\sigma_1=\sigma_2=10^{-5}$，$\lambda_{-1}=0.01$，and $\omega_k=k^{-2}$，初始矩阵 B_0 为单位阵。在前两个实验中精度固定为 10^{-5}，第二个实验我选取全 1 向量的原因是因为问题是对称问题，在这种情况下初始点的运动轨迹是一样的。而在最后一个实验中由于精度的提高我们将 MATLAB 程序显示的有效数字调整为 15 位来得到更好的准确性。从这三张图我们可以清晰的看出当迭代点趋近最小点时函数图像基本保持水平，所以结论是有效的。

4. 总结

本文证明了处理特殊问题的 Gauss-Newton-based BFGS 方法所产生的迭代矩阵序列的收敛性，由此可以猜测这种收敛性是公式本身的性质，这也启发我如果在应用相应公式做算法时，如果算法不具有超线性收敛性，是否可以通过修正 B_k 使他倾向于 Hessian 阵从而具有超线性收敛性。

参考文献

[1] Li, D. and Fukushima, M. (1999) A Globally and Superlinearly Convergent Gauss—Newton-Based BFGS Method for Symmetric Nonlinear Equations. *SIAM Journal on numerical Analysis*, **37**, 152-172. https://doi.org/10.1137/S0036142998335704

[2] Dennis, J.E. and Moré, J.J. (1974) A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods. *Mathematics of Computation*, **28**, 549-560. https://doi.org/10.1090/S0025-5718-1974-0343581-1

[3] Ren-Pu, G. and Powell, M.J. (1983) The Convergence of Variable Metric Matrices in Unconstrained Optimization. *Mathematical Programming*, **27**, 123-143. https://doi.org/10.1007/BF02591941

[4] Stoer, J. (1984) The Convergence of Matrices Generated by Rank-2 Methods from the Restricted β-Class of Broyden. *Numerische Mathematik*, **44**, 37-52. https://doi.org/10.1007/BF01389753

[5] Fletcher, R. (1970) A New Approach to Variable Metric Algorithms. *The Computer Journal*, **13**, 317-322. https://doi.org/10.1093/comjnl/13.3.317

[6] Powell, M.J. (1978) The Convergence of Variable Metric Methods for Nonlinearly Constrained Optimization Calculations. In: *Nonlinear Programming* 3, Academic Press, 27-63. https://doi.org/10.1016/B978-0-12-468660-1.50007-4