Homework 6

陈远洋　23307130322

2025 年 4 月 29 日

Problem 1

In the BFGS method, the approximated Hessian is updated by

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

where $s_k = x_{k+1} - x_k, y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. Show that if B_k is symmetric positive definite, then B_{k+1} is also symmetric positive definite.

Another form of the BFGS update is

$$H_{k+1} = \left(I - \rho_k s_k y_k^T\right) H_k \left(I - \rho_k y_k s_k^T\right) + \rho_k s_k s_k^T,$$

where $H_k = B_k^{-1}$. Apply the Sherman–Morrison–Woodbury formula to verify the second update rule.

证明：

先证一个引理：

Lemma 1. 设 $\langle a, b\rangle_A = a^T A b \, (A \succ 0)$ 表示一个内积, 那么这个内积满足 *Cauchy-Schwarz* 不等式*:*

$$\langle a, b\rangle_A^2 \le \langle a, a\rangle_A \langle b, b\rangle_A$$

证明. 设 $t \in \mathbb{R}$, 则有 $\langle a + tb, a + tb\rangle_A = \langle a, a\rangle_A + 2t\langle a, b\rangle_A + t^2\langle b, b\rangle_A \ge 0$ 当且仅当 a, b 同向时左端项关于 t 的最小值能使等号成立。那么这个关于 t 的二次函数极小值大于等于 0，即 $\langle a, b\rangle_A^2 \le \langle a, a\rangle_A \langle b, b\rangle_A$ 引理得证。 □

回到原题。首先，显然有：$B_{k+1}^T = B_{k+1}$。故 B_{k+1} 对称。由条件 $\nabla f(x_{k+1})^T d_k \ge \nabla f(x_k)^T d_k$

可知 $y_k^T s_k > 0$。对任意 $x \in \mathbb{R}^n$，且 x 与 x_k 不共线时，由 Cauchy-Schwarz 不等式，有：

$$\begin{aligned} x^T B_{k+1} x &= \langle x, x\rangle_{B_k} - \frac{\langle x, s_k\rangle_{B_k}^2}{\langle s_k, s_k\rangle_{B_k}} + \frac{(x^T y_k)^2}{y_k^T s_k} \\ &> \langle x, x\rangle_{B_k} - \frac{\langle x, x\rangle_{B_k}\langle s_k, s_k\rangle_{B_k}}{\langle s_k, s_k\rangle_{B_k}} + \frac{(x^T y_k)^2}{y_k^T s_k} \\ &= \frac{(x^T y_k)^2}{y_k^T s_k} \\ &\geq 0 \end{aligned}$$

在 x 与 s_k 同向时第二步等号成立，但第 4 步不等号严格成立，故在 s_k 与 x 共线时，也有 $x^T B_{k+1} x > 0$。

综上，B_{k+1} 是对称正定矩阵。

接下来证明 H_{k+1} 更新表达式的正确性：令 $U = \begin{pmatrix} y_k & B_k s_k \end{pmatrix}$，$A = \begin{pmatrix} \frac{1}{y_k^T s_k} & 0 \\ 0 & -\frac{1}{s_k^T B_k s_k} \end{pmatrix}$。有：$B_{k+1} = B_k + UAU^T$。由 Shellman-Morrison-Woodbury 公式，有：

$$\begin{aligned} H_{k+1} &= H_k - H_k U (A^{-1} + U^T H_k U)^{-1} U^T H_k \\ &= H_k - \begin{pmatrix} H_k y_k & s_k \end{pmatrix} \left(\begin{pmatrix} y_k^T s_k & 0 \\ 0 & -s_k^T B_k s_k \end{pmatrix} + \begin{pmatrix} y_k^T H_k y_k & y_k^T H_k B_k s_k \\ s_k^T B_k H_k y_k & s_k^T B_k H_k B_k s_k \end{pmatrix} \right)^{-1} \begin{pmatrix} y_k^T H_k \\ s_k^T \end{pmatrix} \\ &= H_k - \begin{pmatrix} H_k y_k & s_k \end{pmatrix} \begin{pmatrix} y_k^T s_k + y_k^T H_k y_k & y_k^T s_k \\ s_k^T y_k & 0 \end{pmatrix}^{-1} \begin{pmatrix} y_k^T H_k \\ s_k^T \end{pmatrix} \\ &= H_k + \frac{1}{(y_k^T s_k)^2} \begin{pmatrix} H_k y_k & s_k \end{pmatrix} \begin{pmatrix} 0 & -y_k^T s_k \\ -s_k^T y_k & y_k^T s_k + y_k^T H_k y_k \end{pmatrix} \begin{pmatrix} y_k^T H_k \\ s_k^T \end{pmatrix} \\ &= H_k + \frac{1}{(y_k^T s_k)^2} \begin{pmatrix} H_k y_k & s_k \end{pmatrix} \begin{pmatrix} -y_k^T s_k s_k^T \\ y_k^T H_k y_k s_k^T + y_k^T s_k s_k^T - s_k^T y_k y_k^T H_k \end{pmatrix} \\ &= H_k + \frac{1}{(y_k^T s_k)^2} \left(-H_k y_k y_k^T s_k s_k^T + s_k y_k^T H_k y_k s_k^T - s_k s_k^T y_k y_k^T H_k + s_k y_k^T s_k s_k^T \right) \\ &= H_k + \frac{s_k y_k^T H_k y_k s_k^T}{(y_k^T s_k)^2} - \frac{H_k y_k s_k^T}{y_k^T s_k} - \frac{s_k y_k^T H_k}{y_k^T s_k} + \frac{s_k s_k^T}{y_k^T s_k} \\ &= \left(I - \rho_k s_k y_k^T\right) H_k \left(I - \rho_k y_k s_k^T\right) + \rho_k s_k s_k^T \end{aligned}$$

其中 $\rho_k = \frac{1}{y_k^T s_k}$.

综上，H_{k+1} 更新表达式的正确性证毕。

Problem 2

There are many other quasi-Newton updating formulae apart from BFGS, DFP, and SR1 method. Of particular interest is the Broyden class, a family of updates specified by the following general formula:

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

where ϕ_k is a scalar and v_k is defined as

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

Show that the formula (1) is a "linear combination" of DFP and BFGS method as

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

where B_{k+1}^{BFGS} and B_{k+1}^{DFP} denotes B_{k+1} in BFGS and DFP method. Also verify the secant equation holds for B_{k+1} in (1).

证明：

首先，我们知道，BFGS 方法的更新公式为：

$$B_{k+1}^{\text{BFGS}} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \tag{2}$$

而 DFP 方法的更新公式为：

$$H_{k+1}^{\text{DFP}} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{s_k y_k^T} \tag{3}$$

由 Problem1 的证明可知，我们可以进一步将DFP 方法的更新公式写成关于 B_k 的表达式：

$$B_{k+1}^{\text{DFP}} = B_k + \frac{y_k y_k^T}{s_k^T y_k} + \frac{s_k^T B_k s_k}{(s_k^T y_k)^2} y_k y_k^T - \frac{B_k s_k y_k^T + y_k s_k^T B_k}{s_k^T y_k} \tag{4}$$

而对 ϕ_k 引导的尾项，我们有：

$$\begin{aligned}(s_k^T B_k s_k) v_k v_k^T &= (s_k^T B_k s_k)\left(\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}\right)\left(\frac{y_k^T}{s_k^T y_k} - \frac{s_k^T B_k}{s_k^T B_k s_k}\right) \\ &= \frac{s_k^T B_k s_k}{(s_k^T y_k)^2} y_k y_k^T - \frac{B_k s_k y_k^T + y_k s_k^T B_k}{s_k^T y_k} + \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}\end{aligned}$$

所以，我们可以将关于 B_{k+1} 的 DFP 方法的更新公式可以写成：

$$B_{k+1}^{\text{DFP}} = B_{k+1}^{\text{BFGS}} + \left(s_k^T B_k s_k\right) v_k v_k^T \tag{5}$$

那么(1) 式可改写为：

$$\begin{aligned} B_{k+1} &= B_{k+1}^{\mathrm{BFGS}} + \phi_k \left(s_k^T B_k s_k\right) v_k v_k^T \\ &= B_{k+1}^{\mathrm{BFGS}} + \phi_k (B_{k+1}^{\mathrm{DFP}} - B_{k+1}^{\mathrm{BFGS}}) \\ &= (1-\phi_k) B_{k+1}^{\mathrm{BFGS}} + \phi_k B_{k+1}^{\mathrm{DFP}} \end{aligned}$$

接下来验证所谓的 secant equation。直观上，我们有 $H_{k+1}^{\mathrm{DFP}} y_k = s_k, B_{k+1}^{\mathrm{BFGS}} s_k = y_k$，所以有：$B_{k+1}^{\mathrm{DFP}} s_k = y_k$。可以推出 $((1-\phi_k) B_{k+1}^{\mathrm{BFGS}} + \phi_k B_{k+1}^{\mathrm{DFP}}) s_k = (1-\phi_k) y_k + \phi_k y_k = y_k$。

而更进一步地，我们也只需验证 $v_k^T s_k = 0$ 即可。而：

$$\begin{aligned} v_k^T s_k &= \left(\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}\right)^T s_k \\ &= \frac{y_k^T s_k}{y_k^T s_k} - \frac{s_k^T B_k s_k}{s_k^T B_k s_k} \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

所以(1) 式引导的更新表达式满足所谓的 secant equation。

Problem 3

(**Coding problem**) Consider the logistic regression problem for binary classification:

$$\min_x \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp(-b_i a_i^T x)\right) + \mu \|x\|_2^2,$$

where $x \in \mathbb{R}^n, a_i \in \mathbb{R}^n$, $b_i \in \{\pm 1\}$ denotes the class that a_i belongs to, n is the dimension of the variable, m is the number of data. μ is the regularization parameter.

Please use a quasi-Newton method to solve above problem, using the **a1a** data given in LIBSVM datasets, or data generated by yourself. A detailed description of the data can be found in `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#a1a`. Please write a report to illustrate the method you used and present your experiment results.

Note:

- The gradient of the objective function is

$$\nabla f(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 + \exp\left(-b_i a_i^T x\right)} \cdot \exp\left(-b_i a_i^T x\right) \cdot (-b_i a_i) + 2\mu x$$

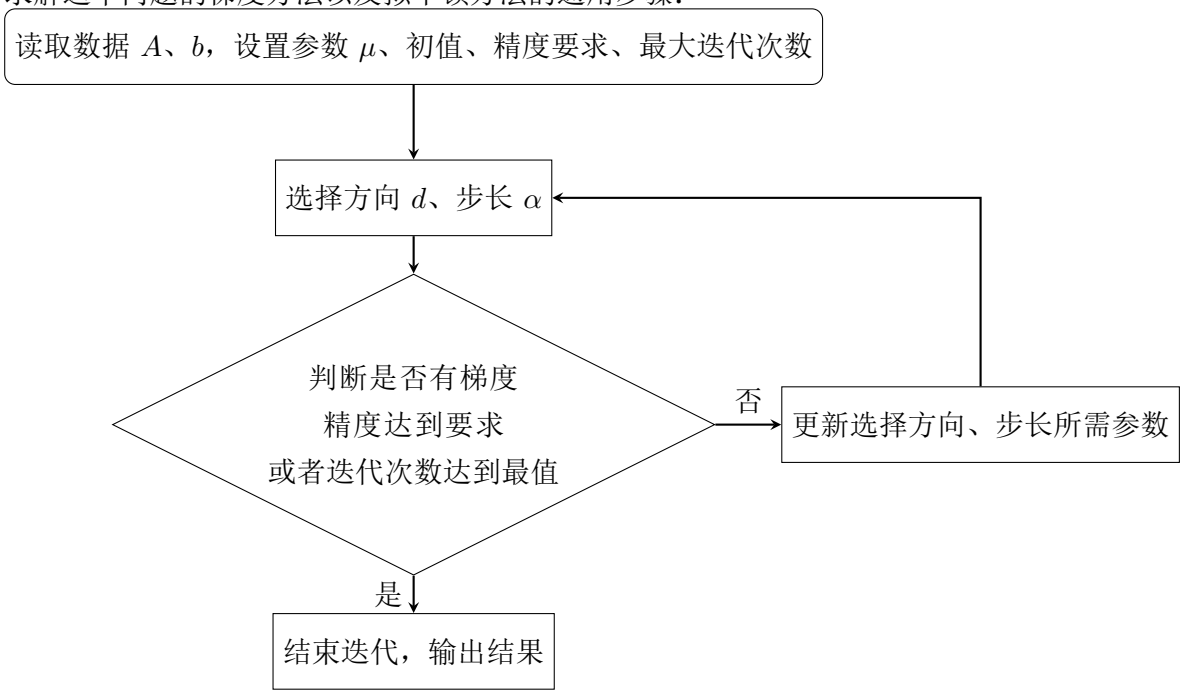
- The regularization parameter μ can be set as $\mu = 1e-3$.
- If you use 'a1a' dataset, you only need to use the 'a1a' file, where $m = 1605$, and $n = 123$. In the file, the first column of each row denotes the label (b_i) of each data, and the following rows indicate elements of which coordinates of a_i are non-zero.
- If you generate the data by yourself, you can set $m \geq 100, n \geq 50$.

解存在性分析

首先，该问题一定是有解并且唯一的。前面的求和部分与后面正则项在一定范围内是有界的，并且前面部分是恒正的。不妨设有 0 点附近的领域 $D = \{x|\mu\|x\|_2^2 \leq f(0)\}$。由 weirstrass 定理，我们知道在 D 内存在原函数最小值 $f(x_0) \leq f(0)$。而在 D 之外，$f(x) > \mu\|x\|_2^2 > f(0) \geq f(x_0)$。所以 x_0 也是全局最小值。另外，正则化项 $\mu\|x\|_2^2$ 是严格凸函数。而单看前面的求和中的每一项，对于函数 $h(z) = \log(1 + \exp(-z))$，其导数为 $h'(z) = \frac{-exp(-z)}{1+\exp(-z)}$，而 $h''(z) = \frac{exp(-z)}{(1+\exp(-z))^2} > 0$。所以这个函数也是严格凸的。所以 x_0 也是全局唯一的最小值。

代码构成分析

求解这个问题的梯度方法以及拟牛顿方法的通用步骤：



在主函数中我主要实现初始化数据阶段，以及效果比较。对拟牛顿法的实现上，我写了 BFGS 和 DFP 两个方法。而迭代的过程则可以通过调用不同接口而对应不同方法。下面来看各个辅助函

数以及其对应的作用：

- `logisticRegression`: 对应优化的目标函数，输入参数为 μ、A、b、x。在后续使用过程中，可以通过在句柄中设定参数将其变为关于 x 的单变量函数。
- `grad_LR`: 顾名思义，上一个函数关于 x 的梯度。
- `LineSearch`: 输入参数 f、df、d 、x 、a、c 、β。输出在 d 方向下满足 $armijo$ 条件的以 a 为初始值，每次递减 β 的最大步长。但是实际实现中，为了确保拟牛顿算法 H 的正定性，我们要求 $\nabla f(x+ad)^T d < \nabla f(x)^T d$ （相当于加了一个 c2=1 的 wolfe 准则）。但是值得注意的是，这个条件非常宽松，以至于在本次实验中我没有遇到满足前一个条件而不满足这个条件的情况，所以 LineSearch 的过程可以认为是 $armijo$ 的。我设定 $a = 1.5$,$c = 0.01$, $\beta = 0.99$。在后续使用中，可以将已经调试好的参数通过固定在句柄中的方式调用。
- `BFGS`: 对应 BFGS 的解决方法。输入参数 f、df、x_0、$maxIter$、tol。输出 x 的 BFGS 迭代值。值得说明的是我们知道 BFGS 关于 H 的迭代表达式:

  $$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T = H_k - (t + t^T) + (\rho_k + \rho_k^2 y_k^T H_k y_k) s_k s_k^T$$

  其中 $\rho_k = \frac{1}{s_k^T y_k}, t = \rho_k H_k y_k s_k^T$，通过预先计算 t 跟 ρ_k，我们可以避免矩阵与矩阵之间的乘法，把单次更新的时间复杂度降为 $O(n^2)$。
- `DFP`: 对应 DFP 的解决方法。输入参数 f、df、x_0、$maxIter$、tol。输出 x 的 DFP 迭代值。DFP 的迭代过程与 BFGS 类似，总之注意加上括号改变运算顺序，避免矩阵与矩阵之间的直接乘法。保证单次更新时间复杂度在 $O(n^2)$ 级别。

参数说明

关于为什么步长初值取成 1.5，是出于选择大步长以及减小每次迭代的抖动的综合考量。以 DFP 为例，当步长初值较小时，可能每次向前“迈的步子”太小，导致离当前方向下的最优解很远。而这也会导致梯度模长的“抖动”。当步长初值较大时，可能每次向前“迈的步子”太大，导致迭代过程过于不稳定。出于这样的综合考量，在实验中，我试出来以 1.5 为初值步长的效果最好。详见下图：

实验结果

最优解采取 BFGS 得到的结果，放在 x1.mat 文件中。当设置停机条件为梯度模长小于 1×10^{-8} 时，最优解对应的函数值为 0.33691514。达到终点时，BFGS 迭代次数为 144，DFP 迭代次数为 513。二者终点处达到的梯度模长分别为 8.95×10^{-9} 跟 7.35×10^{-9}。

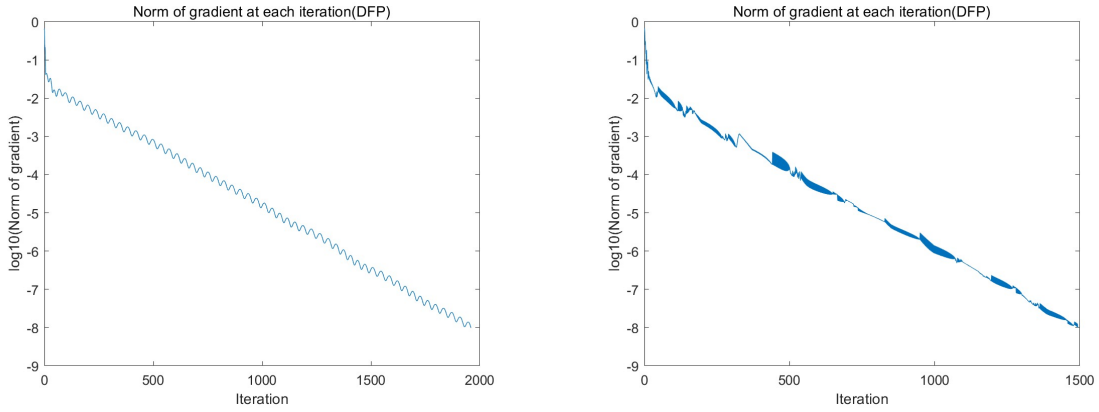


图 1: 初始步长过大或者过小导致的抖动（左图初始步长为 1，右图初始步长为 2）

```
>> QuassiNewton
Total 144 iterations to get 1e-08 accuracy by my self-implemented BFGS
Total 513 iterations to get 1e-08 accuracy by my self-implemented DFP

Norm of optimal point difference between BFGS and DFP: 9.6326663e-07
Gradient norm: 8.95398e-09(BFGS), 7.35302e-09(DFP)
F-value at optimal solution: 0.33691514(BFGS), 0.33691514(DFP)

with accuracy 1e-08, Number of iterations: 3761 (SD)
f value at Optimal solution: 0.33691514 (SD)
norm of gradient at Optimal solution: 9.97214e-09 (SD)
>>
```

图 2: 最终结果（加上 SD）预期输出

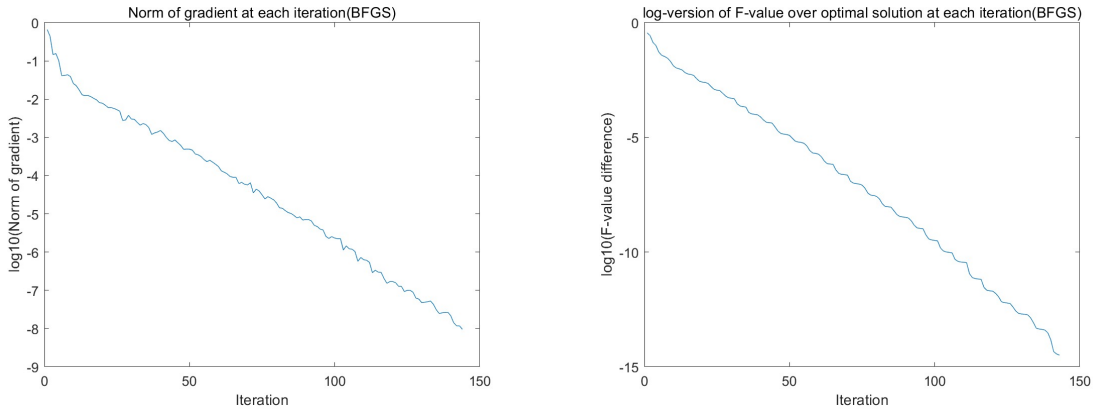
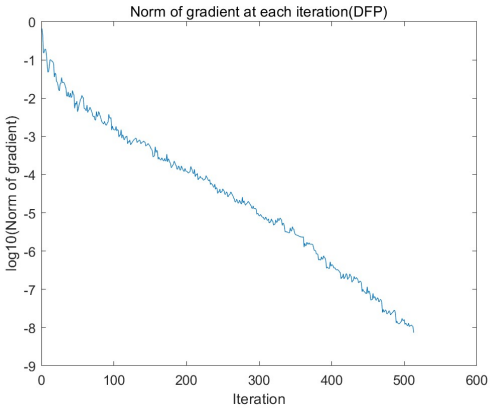


图 3: BFGS 的收敛情况

每步迭代后函数值减去最小值取对数的图像以及梯度模长变化如图3、4、5所示：

可以看到总体上大致是一个线性的过程，整体上呈现微弱的超线性收敛趋势，这似乎与我们之前的预期不太相符。但为了验证拟牛顿法的优越性，我们加入最速下降的梯度法作为对比。达到相同精度时，最速下降法的迭代次数为 3761。所以在这个问题上我们的拟牛顿方法远远快于梯度方法。



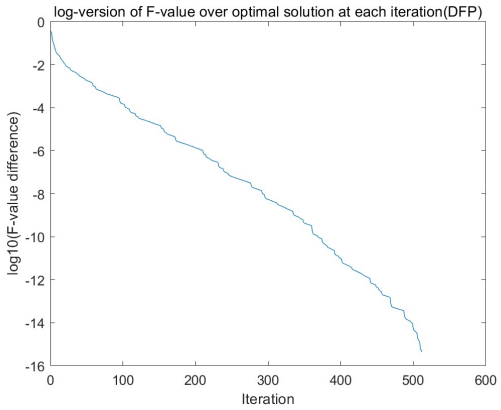
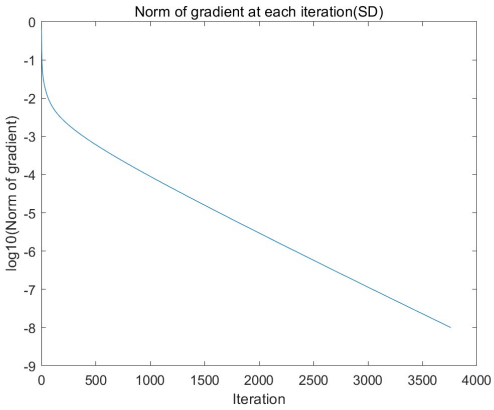


图 4: DFP 的收敛情况



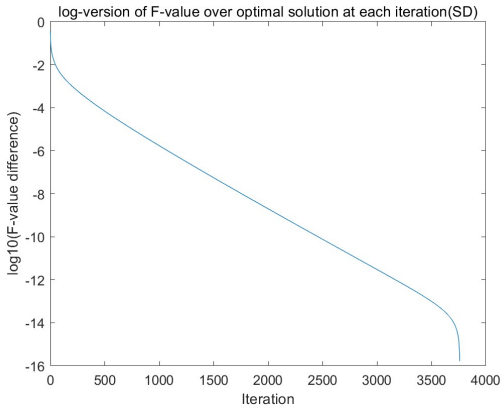


图 5: 最速下降法（SD）的收敛情况