李健宁
机器学习中的优化问题
2025 年 4 月 23 日

## Homework 9

### Question 1.

**梯度与 Hessian.**

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} x_1^3 - x_2 + 1 \\ x_2 - x_1 - 1 \end{pmatrix}, \qquad \nabla^2 f(x) = \begin{pmatrix} 3x_1^2 & -1 \\ -1 & 1 \end{pmatrix}.$$

**DFP 算法迭代.** 取 $H_0 = I_2$，分别对两组初始值做迭代：

表 1. DFP 算法在不同初始点的迭代过程比较

| 初始点 $(0,0)^T$ | | | | | | 初始点 $(1.5,1.0)^T$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $x_1^{(k)}$ | $x_2^{(k)}$ | $f(x^{(k)})$ | $\|\nabla f(x^{(k)})\|$ | $\alpha_k$ | $k$ | $x_1^{(k)}$ | $x_2^{(k)}$ | $f(x^{(k)})$ | $\|\nabla f(x^{(k)})\|$ | $\alpha_k$ |
| 0 | 0.000000 | 0.000000 | 0.000000 | 1.414214 | 0.596072 | 0 | 1.500000 | 1.000000 | 0.765625 | 3.693322 | 0.254111 |
| 1 | -0.596072 | 0.596072 | -0.627631 | 0.271732 | 1.586866 | 1 | 0.642375 | 1.381167 | -0.629639 | 0.285845 | 2.235703 |
| 2 | -0.946621 | 0.358522 | -0.700745 | 0.368605 | 2.422839 | 2 | 0.855450 | 1.981643 | -0.724054 | 0.377354 | 1.747929 |
| 3 | -1.050298 | -0.007492 | -0.746425 | 0.157065 | 0.644399 | 3 | 1.012417 | 2.078697 | -0.747647 | 0.077927 | 1.070668 |
| 4 | -0.999043 | 0.003383 | -0.749996 | 0.002480 | 0.999313 | 4 | 0.997143 | 1.998481 | -0.749991 | 0.007154 | 0.834262 |
| 5 | -0.999946 | 0.000014 | -0.750000 | 0.000152 | 1.059407 | 5 | 0.999995 | 1.999972 | -0.750000 | 0.000025 | 1.001478 |
| 6 | -1.000000 | -0.000000 | -0.750000 | 0.000000 | — | 6 | 1.000000 | 2.000000 | -0.750000 | 0.000000 | — |

两组不同初始点分别落入了函数的两个局部最小点 $(-1,0)$ 和 $(1,2)$。故 DFP 算法并未收敛到同一个点，而是陷入了不同的局部最优解。

### Question 2.

两种算法都顺利求出了最优解 $x = (3,9,84)^\top$，且迭代过程的所有 $H$ 都是正定的。图1展示了两种算法下梯度范数随迭代次数的变化。具体的算法代码请见附件 `HW9-code.ipynb`。

### Question 3.

使用 L-BFGS 算法和 Wolfe 条件。具体的参数选择上，$f(x)$ 的参数 $\alpha = 100$，L-BFGS 算法的 $m \in [1,3,5,10,20,30]$，停止条件 $\nabla f(x) \le 10^{-6}$，最大迭代次数 1000 次，Wolfe 条件中初始 $\alpha = 1.0, c_1 = 10^{-4}, c_2 = 0.9$，采用二分法最多迭代 20 次，寻找满足条件的 $\alpha_k$。

最终结果如表2，$m$ 的增大增加了对内存的占用，但也加速了迭代。当 $m$ 超过实际迭代次数时，L-BFGS 算法实际上就是 BFGS 算法，因此达到稳定。

### Question 4.

**Lipschitz gradient majorant.** 设 $f(x) = \frac{1}{2}\|Ax-b\|_2^2$，其梯度 Lipschitz 常数为 $L = \lambda\max(A^TA)$，则有：

$$f(x) \le f(x^{(k)}) + \nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{L}{2}\|x - x^{(k)}\|_2^2$$

也即

$$\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \le f(x^{(k)}) + \nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{L}{2}\|x - x^{(k)}\|_2^2 + \lambda\|x\|_1 := g_k(x)$$

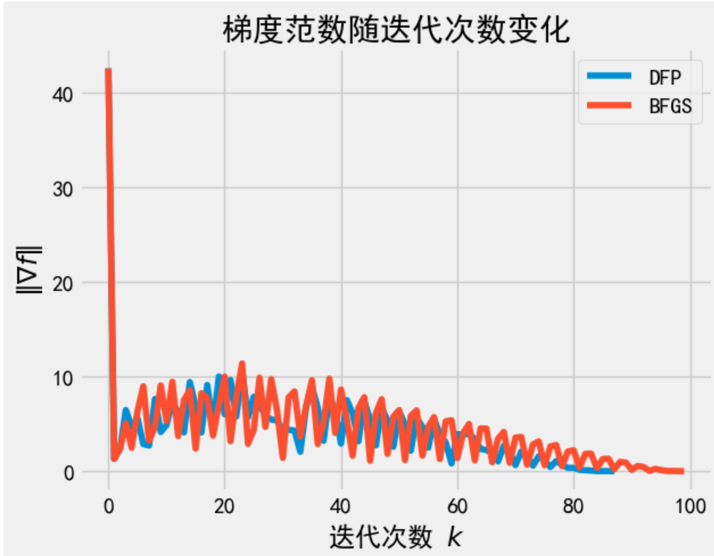


图 1. 两种算法下梯度范数随迭代次数的变化

表 2. 不同 $m$ 对 L-BFGS 算法的影响

| **m** | **迭代次数** | **最终函数值** | **时间 (s)** |
|---|---|---|---|
| 1 | 70 | $3.87 \times 10^{-15}$ | 0.18 |
| 3 | 27 | $9.04 \times 10^{-17}$ | 0.11 |
| 5 | 26 | $2.51 \times 10^{-20}$ | 0.06 |
| 10 | 25 | $1.92 \times 10^{-16}$ | 0.06 |
| 20 | 25 | $8.55 \times 10^{-17}$ | 0.06 |
| 30 | 25 | $8.55 \times 10^{-17}$ | 0.06 |

子问题 $\min g_k(x)$ 解析解为

$$x^{(k+1)} = \text{soft}_{\frac{\lambda}{L}}(x^{(k)} - \frac{1}{L}A^T(Ax^{(k)} - b))$$

其中 $\text{soft}_\tau(u) = \text{sign}(u)\max\{|u| - \tau, 0\}$。迭代 1000 次，得到结果。

**Variational majorant function.** 利用恒等式

$$|x_i| = \min_{d_i>0} \frac{1}{2}(d_i x_i^2 + d_i^{-1}),$$

有

$$\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \leq \frac{1}{2}\|Ax - b\|_2^2 + \lambda[\frac{1}{2}(x^\top \mathrm{D}x + 1^\top \mathrm{D}^{-1}1)] := h_k(x, d)$$

易求得

$$\arg\min_d h_k(x^{(k)}, d) = \text{diag}(\frac{1}{|x_i^{(k)}|}).$$

为保证数值稳定，取 $d_i^{(k)} = \frac{1}{|x_i^{(k)}|+\varepsilon}$。那么

$$g_k(x) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{\lambda}{2}\sum_i d_i^{(k)} x_i^2 + C,\ x^{(k+1)} = \arg\min_x\ g_k(x),$$

($C$ 为与 $x$ 无关常数)，则 $x^{(k+1)}$ 是线性方程组

$$(A^T A + \lambda\,\text{diag}(d^{(k)}))\,x = A^T b$$

的解。

**数据实验.** 数据实验中，我们取 $A$ 为 $m \times n$ 的取值在 $[0,1]$ 的随机矩阵，生成仅有前 10 个分量不为 0 的随机稀疏解向量 $x_{true}$，根据 $x_{true}$，计算 $b = Ax_{true} + 0.01s$，$s$ 为随机生成的每个分量取值在 $[0,1]$ 的 $m$ 维向量。

取 $m = 100, n = 200, \lambda = 0.1, \varepsilon = 10^{-6}$，迭代 5000 次，得到的结果如图2和表3。需要说明的是 $x_{true}$ 并非原问题的真实最优解，两种方法求出的最优解与 $x_{true}$ 的差的 L2 范数和我们为 $b$ 引入的随机扰动的尺度（每个分量方差为 0.01）相吻合。

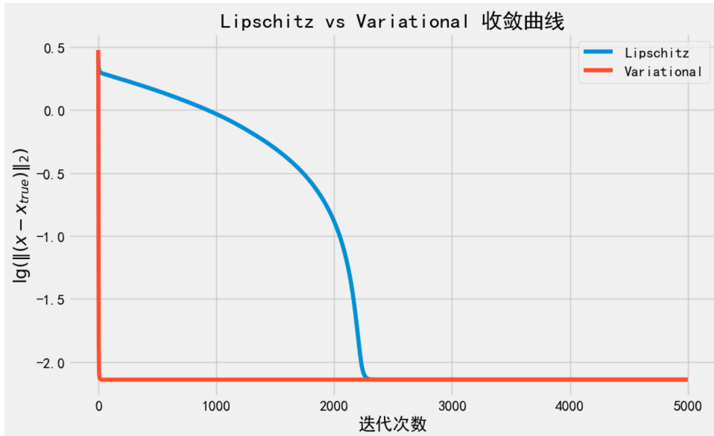


图 2. 两种算法下 L2 范数随迭代次数的变化

表 3. 两种算法下的最终函数值和 L2 范数

| **Majorant** | **最终解 $x$ 与 $x_{true}$ 的 L2 范数** $\|x - x_{true}\|_2$ | **最终函数值** |
|---|---|---|
| Lipschitz $x = x_L$ | $7.2321 \times 10^{-3}$ | $6.6589 \times 10^{-1}$ |
| Variational $x = x_V$ | $7.2209 \times 10^{-3}$ | $6.6590 \times 10^{-1}$ |
| 相对误差 $\|x_L - x_V\|_2$ | $1.1226 \times 10^{-4}$ | |