From Newton to BFGS

张海鹏

University of JiangNan

zhpmatrix@gmail.com

2017 年 4 月 21 日

Outline

- Newton Method
- Quasi-Newton Condition
- DFP
- BFGS
- L-BFGS

$y = f(x)$ 的零点?

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k) \tag{1}$$

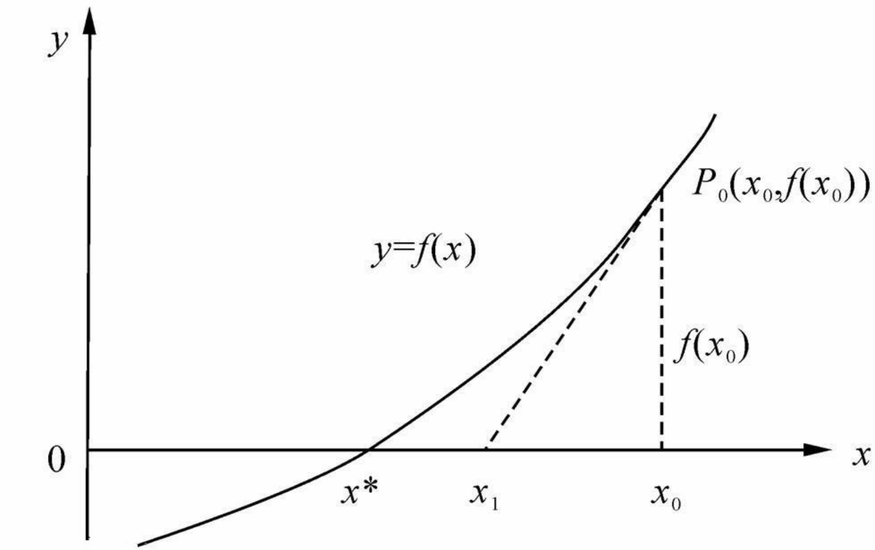
由 $f(x) = 0$ 得:

$$x = x_k - \frac{f(x_k)}{f'(x_k)} \tag{2}$$

迭代过程:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \tag{3}$$

Newton Method



Newton Method

二阶 taylor 展开:

$$f(x) \approx f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) \tag{4}$$

由 $\nabla f(x) = 0$ 得:

$$g_k + H_k(x - x_k) = 0 \tag{5}$$

其中，$g_k = \nabla f(x_k), H_k = \nabla^2 f(x_k)$
迭代过程:

$$x_{k+1} = x_k - H_k^{-1} g_k \tag{6}$$

Newton Method

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix} \tag{7}$$

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \frac{\partial^2 f}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix} \tag{8}$$

令 $d_k = H_k^{(-1)} g_k$, 通过精确线搜找步长

$$\lambda_k = arg\min_{\lambda \in R} f(x_k + \lambda d_k) \tag{9}$$

Pros: 二阶收敛，二次终止

Cons: 二阶可导，Hessian 正定 (可逆)，计算多和存储多

$$f(x) \approx f(x_{k+1}) + \nabla f(x_{k+1})(x - x_{k+1}) + \frac{1}{2}(x - x_{k+1})^T \nabla^2 f(x_{k+1})(x - x_{k+1}) \tag{10}$$

用梯度算子 ∇ 作用两边，有:

$$\nabla f(x) \approx \nabla f(x_{k+1}) + H_{k+1}(x - x_{k+1}) \tag{11}$$

令 $x = x_k$，有:

$$g_{k+1} - g_k \approx H_{k+1}(x_{k+1} - x_k) \tag{12}$$

令 $s_k = x_{k+1} - x_k, y_k = g_{k+1} - g_k$，有:

$$s_k \approx H_{k+1}^{-1} y_k \tag{13}$$

找到 $\boldsymbol{D}_{k+1}$ 替代 $\boldsymbol{H}_{k+1}^{-1}$, 有:

$$\boldsymbol{s}_k \approx \boldsymbol{D}_{k+1}\boldsymbol{y}_k \tag{14}$$

迭代过程:

$$\boldsymbol{D}_{k+1} = \boldsymbol{D}_k + \Delta \boldsymbol{D}_k, \boldsymbol{k} = 0, 1, 2, \ldots \tag{15}$$

构造过程:

$$\Delta \boldsymbol{D}_k = \alpha\mu\mu^{\boldsymbol{T}} + \beta\nu\nu^{\boldsymbol{T}} \tag{16}$$

带入 (14)，有:

$$\boldsymbol{s}_k = \boldsymbol{D}_k\boldsymbol{y}_k + \alpha\mu\mu^{\boldsymbol{T}}\boldsymbol{y}_k + \beta\nu\nu^{\boldsymbol{T}}\boldsymbol{y}_k \tag{17}$$

即:

$$s_k = D_k y_k + (\alpha \nu^T y_k)\mu + (\beta \nu^T y_k)\nu \tag{18}$$

令 $\alpha\mu^T y_k = 1, \beta\nu^T y_k = -1$，有

$\mu - \nu = s_k - D_k y_k$, 取 $\mu = s_k, \nu = D_k y_k$, 得到:

$$\alpha = \frac{1}{s_k^T y_k} \tag{19}$$

$$\beta = -\frac{1}{(D_k y_k)^T y_k} \tag{20}$$

$$\Delta D_k = \frac{s_k s_k^T}{s_k^T y_k} - \frac{D_k y_k y_k^T D_k}{y_k^T D_k y_k} \tag{21}$$

找到 B_{k+1}替代H_{k+1}, 有:

$$\Delta B_k = \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T D_k}{s_k^T B_k s_k} \tag{22}$$

BFGS

B_{k+1} 的迭代公式:

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T D_k}{s_k^T B_k s_k} \tag{23}$$

应用 Sherman-Morrison 公式得到:

$$B_{k+1}^{-1} = (I - \frac{s_k y_k^T}{y_k^T s_k}) B_k^{-1} (I - \frac{y_k s_k^T}{y_k^T s_k}) + \frac{s_k s_k^T}{y_k^T s_k} \tag{24}$$

设 $\boldsymbol{A} \in \boldsymbol{R}^n$ 为非奇异方阵，$\boldsymbol{\mu}, \boldsymbol{\nu} \in \boldsymbol{R}^n$, 若 $1 + \boldsymbol{\nu}^T \boldsymbol{A}^{-1} \boldsymbol{\mu} \neq 0$, 则有:

$$(\boldsymbol{A} + \boldsymbol{\mu}\boldsymbol{\nu}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{\mu}\boldsymbol{\nu}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{\nu}^T\boldsymbol{A}^{-1}\boldsymbol{\mu}} \tag{25}$$

证明见附件。

为了方便表示，用 D_k 来代替 B_k^{-1}，则 D_k 的存储开销过大，解决的思路是存储计算过程中的最新的 m 个向量序列 $\{s_i\},\{y_i\}$。
BFGS 的迭代过程为:

$$D_{k+1} = (I - \frac{s_k y_k^T}{y_k^T s_k}) D_k (I - \frac{y_k s_k^T}{y_k^T s_k}) + \frac{s_k s_k^T}{y_k^T s_k} \tag{26}$$

记 $\rho_k = \frac{1}{y_k^T s_k}, V_k = I - \rho_k y_k s_k^T$, 则上式可写成:

$$D_{k+1} = V_k^T D_k V_k + \rho_k s_k s_k^T \tag{27}$$

记 $\hat{m} = min\{k, m-1\}$ 则有，

$$
\begin{aligned}
D_{k+1} &= (V_k^T V_{k-1}^T \ldots V_{k-\hat{m}+1}^T V_{k-\hat{m}}^T) D_0 (V_{k-\hat{m}} V_{k-\hat{m}+1} \ldots V_{k-1} V_k) \\
&+ (V_k^T V_{k-1}^T \ldots V_{k-\hat{m}+2}^T V_{k-\hat{m}+1}^T)(\rho_0 s_0 s_0^T)(V_{k-\hat{m}+1} V_{k-\hat{m}+2} \ldots V_{k-1} V_k) \\
&+ (V_k^T V_{k-1}^T \ldots V_{k-\hat{m}+3}^T V_{k-\hat{m}+2}^T)(\rho_1 s_1 s_1^T)(V_{k-\hat{m}+2} V_{k-\hat{m}+3} \ldots V_{k-1} V_k) \\
&+ \ldots \\
&+ (V_k^T k_{k-1}^T)(\rho_{k-2} s_{k-2} s_{k-2}^T)(V_{k-1} V_k) \\
&+ V_k^T (\rho_{k-1} s_{k-1} s_{k-1}^T) V_k \\
&+ \rho_k s_k s_k^T
\end{aligned}
$$

$D_k g_k$ **Algs**:

1) 初始化

$$\delta = \begin{cases} 0, \text{若} k \leq m \\ k - m, \text{若} k > m \end{cases} \qquad L = \begin{cases} k, \text{若} k \leq m \\ m, \text{若} k > m \end{cases} \qquad q_L = g_k$$

2) 后向循环:

$FOR \quad i = L-1, L-2, \ldots, 1, 0 \quad DO \quad \{j = i + \delta; \alpha_i = \rho_j s_j^T q_{i+1}; q_i = q_{i+1} - \alpha_i y_j\}$

3) $r_0 = D_0 q_0$;

4) 前向循环:

$FOR \quad i = 0, 1, \ldots, L-2, L-1 \quad DO \quad \{j = i + \delta; \beta_j = \rho_j y_j^T r_i; r_{i+1} = r_i + (\alpha_i - \beta_i) s_j\}$

5) $r_L = D_k g_k$

L-BFGS(4/4)

Broyden, Fletcher, Goldfarb, Shanno

Summary

- Newton Method(Pros and Cons)
- Quasi-Newton Condition($s_k \approx H_{k+1}^{-1} y_k$)
- DFP($D_{k+1} = H_{k+1}^{-1}$)
- BFGS($B_{k+1} = H_{k+1}$,Sherman-Morrision)
- L-BFGS(Limited-storage,Two Loop Recursion)

TKS(Q&R)