拟牛顿法

朱天宇

牛顿法需要求函数的 Hessian 矩阵，对于大规模的问题，这需要很大的代价。**拟牛顿法**使用 Hessian 矩阵或者其逆矩阵来近似，来替代求解 Hessian 矩阵的过程。其近似的矩阵记为 B^k，保留了 Hessian 矩阵的部分性质。

1 割线方程

牛顿法的推导基于函数 $f(x)$ 的泰勒二次展开：

$$\nabla f(x) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x - x^{k+1}) + \mathcal{O}(\|x - x^{k+1}\|^2)$$

这里，令 $x = x^k, s^k = x^{k+1} - x^k, y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$，上式可以写成

$$\nabla^2 f(x^{k+1})s^k + \mathcal{O}(\|s^k\|^2) = y^k$$

忽略高阶项 $\mathcal{O}$，那么我们希望 Hessian 的近似矩阵 B^{k+1} 满足

$$y^k = \mathsf{B}^{k+1}s^k$$

使用 H^{k+1} 表示 B^k 的逆矩阵，就有

$$s^k = \mathsf{H}^{k+1}y^k$$

这两个式子被称为**割线方程**。

割线方程的本质是希望目标函数 $f(x)$ 和其二次近似 $m_{k+1}(d) = f(x^{k+1}) + \nabla f(x^{k+1})^\top d + \frac{1}{2}d^\top \mathsf{B}^{k+1}d$，在 $x = x^k, x = x^k + 1$ 处有相同的梯度。

曲率条件　近似矩阵 B^k 需要正定，对式子 $y^k = \mathsf{B}^{k+1}s^k$ 两边同时左乘 $(s^k)^\top$，就可以得到**必要条件**

$$(s^k)^\top \mathsf{B}^{k+1}s^k = (s^k)^\top y^k > 0$$

这被称为曲率条件。

在**线搜索**的时候，需要使用 Wolfe 准则来使得曲率条件成立。根据 Wolfe 准则的第二个条件 $\varphi'(\alpha) \geq \sigma\varphi'(0)$，写成梯度形式，两侧同时乘以 s^k，就有

$$\nabla f(x^{k+1})^\top s^k \geq \sigma \nabla f(x^k)^\top s^k$$

两边同时减去 $\nabla f(x^k)^\top s^k$，就有

$$(y^k)^\top s^k \geq (\sigma - 1)\nabla f(x^k)^\top s^k > 0$$

因为 $\sigma < 1$，且 $s^k = \alpha_k d^k$，是下降方向。

通常，近似矩阵 $\mathsf{B}^k, \mathsf{H}^k$ 都迭代更新的。有的方法近似 B^k，有的方法近似 H^k

2 秩一更新 SR1

使用**待定系数法**，已知 B^k，满足割线方程，求 B^{k+1}。

设 $\mathsf{B}^{k+1} = \mathsf{B}^k + auu^\top$，带入割线方程

$$\mathsf{B}^{k+1}s^k = (\mathsf{B}^k + auu^\top)s^k = y^k$$

从而有

$$(a \cdot u^\top s^k)u = y^k - \mathsf{B}^k s^k$$

由于 $(a \cdot u^\top s^k)$ 是标量，因此 u 与 $y^k - \mathsf{B}^k s^k$ 方向是相同的。不妨令 $u = y^k - \mathsf{B}^k s^k$，带入上式，有

$$a((y^k - \mathsf{B}^k s^k)^\top s^k)(y^k - \mathsf{B}^k s^k) = y^k - \mathsf{B}^k s^k$$

可得 $a=\frac{1}{(y^k-\mathsf{B}^k s^k)^\top s^k}$，带入最初的式子，就有

$$\mathsf{B}^{k+1}=\mathsf{B}^k+\frac{(y^k-\mathsf{B}^k s^k)(y^k-\mathsf{B}^k s^k)^\top}{(y^k-\mathsf{B}^k s^k)^\top s^k}$$

相同方法可得

$$\mathsf{H}^{k+1}=\mathsf{H}^k+\frac{(s^k-\mathsf{H}^k y^k)(s^k-\mathsf{H}^k y^k)^\top}{(s^k-\mathsf{H}^k y^k)^\top s^k}$$

一个有趣的观察是，将公式 1 做如下替换则可获得公式 2：

$$\mathsf{B}^k\to\mathsf{H}^k,\quad s^k\to y^k$$

SR1 公式结构简单，但无法保证矩阵在迭代中保持正定。之前的条件只是充分条件。

3 秩二更新 BFGS

同样使用待定系数法，设

$$\mathsf{B}^{k+1}=\mathsf{B}^k+auu^\top+bvv^\top$$

可推出基于 B^k 的更新公式：

$$\mathsf{B}^{k+1}=\mathsf{B}^k+\frac{y^k(y^k)^\top}{(s^k)^\top y^k}-\frac{\mathsf{B}^k s^k(\mathsf{B}^k s^k)^\top}{(s^k)^\top\mathsf{B}^k s^k}$$

带入 SMW 公式可以获得 H^k 的更新公式：

$$\mathsf{H}^{k+1}=(\mathsf{I}-\rho_k s^k(y^k)^\top)^\top\mathsf{H}^k(\mathsf{I}-\rho_k s^k(y^k)^\top)+\rho_k s^k(s^k)^\top$$

其中，$\rho_k=\frac{1}{(s^k)^\top y^k}$

BFGS 公式产生的矩阵 H^{k+1} 正定的条件是满足不等式 $(s^k)^\top y^k>0$，这可以通过线搜索保证。

4 有限内存的 BFGS

大规模问题需要存储拟牛顿矩阵 B^k 或者 H^k，需要消耗 $\mathcal{O}(n^2)$ 的内存，不现实。LBFGS 可以解决这个问题。

首先，基于 H^k 的 BFGS，引入新符号后的表示为

$$\mathsf{H}^{k+1}=(\mathsf{V}^k)^\top\mathsf{H}^k\mathsf{V}^k+\rho_k s^k(s^k)^\top$$

其中，$\rho_k=\frac{1}{(s^k)^\top y^k},\mathsf{V}^k=\mathsf{I}-\rho_k y^k(s^k)^\top$

这个公式是一个 类似递推的形式 ，因此尝试展开 m 次：

$$\begin{aligned}
&\mathsf{H}^k\\
&=(\mathsf{V}^{k-m}...\mathsf{V}^{k-1})^\top\mathsf{H}^{k-m}(\mathsf{V}^{k-m}...\mathsf{V}^{k-1})+\\
&\rho_{k-m}(\mathsf{V}^{k-m+1}...\mathsf{V}^{k-1})^\top s^{k-m}(s^{k-m})^\top(\mathsf{V}^{k-m+1}...\mathsf{V}^{k-1})+\\
&\rho_{k-m+1}(\mathsf{V}^{k-m+2}...\mathsf{V}^{k-1})^\top s^{k-m+1}(s^{k-m+1})^\top(\mathsf{V}^{k-m+2}...\mathsf{V}^{k-1})+\\
&...+\\
&\rho_{k-1}s^{k-1}(s^{k-1})^\top
\end{aligned}$$

这样，通过 H^{k-m} 就可以获得 H^k。但 H^{k-m} 无法显示求出，所以要找**近似矩阵**$\hat{\mathsf{H}}^{k-m}$，一般来说这个矩阵的结构要比较简单。

实际上，H^k 的显式形式根本无需计算。只要计算 $\mathsf{H}^k\nabla f(x^k)$ 即可。可以通过这个算法巧妙地计算 $\mathsf{H}^k\nabla f(x^k)$。

Algorithm 1 L-BFGS 双循环递归算法

Require: $n \geq 0 \vee x \neq 0$

Ensure: $y = x^n$ 较看脸

```
y ← 1
if n < 0 then
    X ← 1/x
    N ← −n jlk 就
else
    X ← x
    N ← n
end if
while N ≠ 0 do
    if N is even then
        X ← X × X
        N ← N/2
    else if N is odd then
        y ← y × X
        N ← N − 1
    end if
end while
```