优化理论与算法

第九章

牛顿与拟牛顿法

郭加熠|助理教授

目录

讲　　员

郭加熠，江波，刘慧康

牛顿法

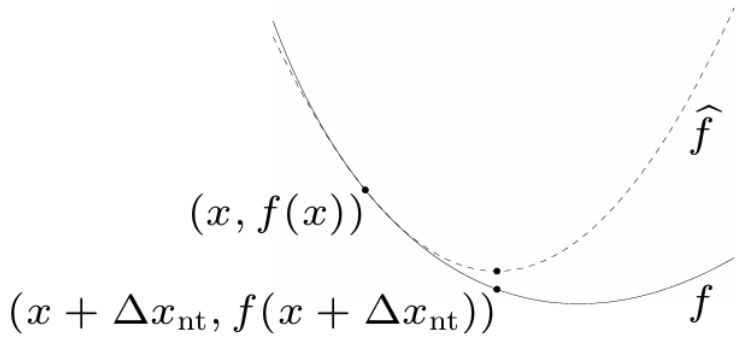
考虑无约束最小化

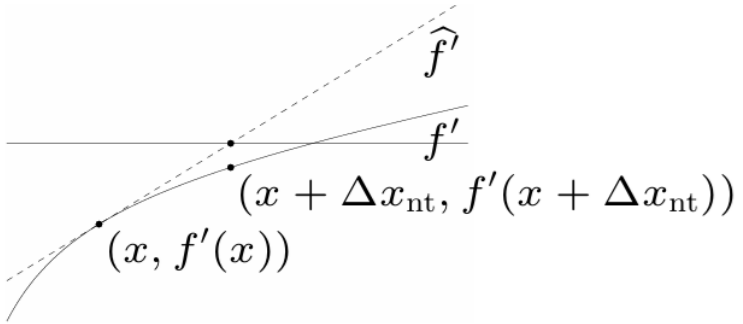
$$\min \quad f(x)$$

- f 凸，二次连续可微

牛顿法

$$x^+ = x - t\nabla^2 f(x)^{-1}\nabla f(x)$$





拟牛顿方法

牛顿法

$$x^+ = x - t\nabla^2 f(x)^{-1}\nabla f(x)$$

- **优点**：快速收敛，仿射不变性
- **缺点**：需要二阶导数，求解线性方程，对于大规模问题过于昂贵

拟牛顿方法

$$x^+ = x - tH^{-1}\nabla f(x)$$

$H \succ 0$ 是在 x 处的 Hessian 矩阵的近似，选择它是为了：

- 避免计算二阶导数
- 简化搜索方向的计算
- 仍然满足仿射不变性

算法构建

给定起始点 $x^{(0)} \in \mathrm{dom}\ f$, $H_0 \succ 0$

1. 计算拟牛顿方向 $\Delta x = -H_{k-1}^{-1} \nabla f(x^{(k-1)})$
2. 确定步长 t （例如，通过回溯线搜索）
3. 计算 $x^{(k)} = x^{(k-1)} + t\Delta x$
4. 计算 H_k（或 H_k^{-1}）

- 不同的方法在步骤 4 中使用不同的规则来更新 H
- 也可以直接更新 H_k^{-1} 以简化 Δx 的计算

割线条件

拟牛顿更新满足割线条件 $H_k s = y$，即，

$$H_k(x^{(k)} - x^{(k-1)}) = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$$

解释： 在 $x^{(k)}$ 处定义二阶近似

$$f_{\text{quad}}(z) = f(x^{(k)}) + \nabla f(x^{(k)})^T (z - x^{(k)}) + \frac{1}{2}(z - x^{(k)})^T H_k (z - x^{(k)})$$

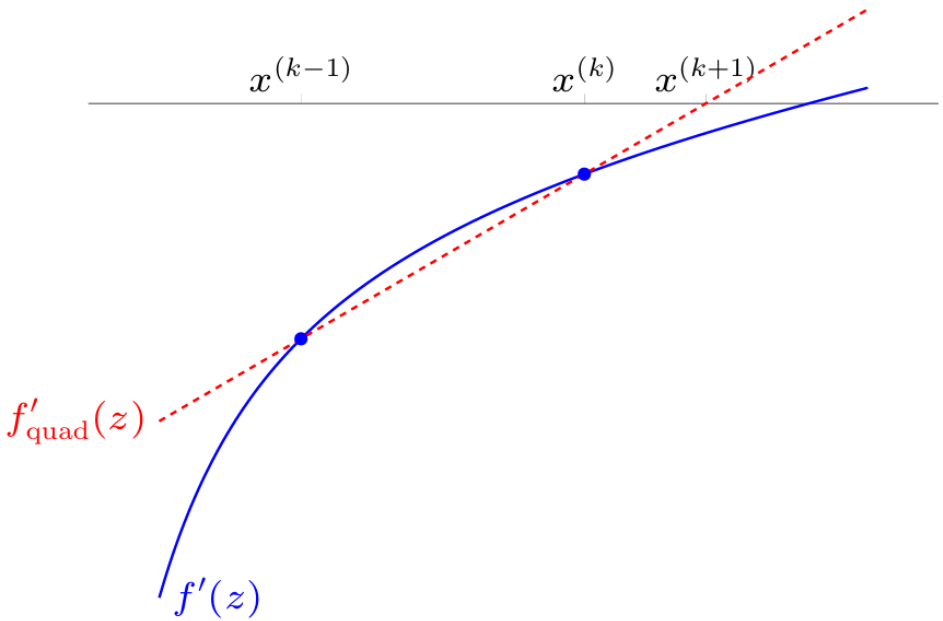
割线条件意味着 f_{quad} 的梯度在 $x^{(k-1)}$ 处与 f 一致：

$$\nabla f_{\text{quad}}(x^{(k-1)}) = \nabla f(x^{(k)}) + H_k(x^{(k-1)} - x^{(k)}) = \nabla f(x^{(k-1)})$$

割线法

对于 $f:\mathbb{R}\to\mathbb{R}$，单位步长的 BFGS 给出割线法

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{H_k}, \quad H_k = \frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$



目录

讲　　员

郭加熠，江波，刘慧康

Broyden-Fletcher-Goldfarb-Shanno (BFGS) 更新

BFGS 更新

$$H_k = H_{k-1} + \frac{yy^T}{y^T s} - \frac{H_{k-1} ss^T H_{k-1}}{s^T H_{k-1} s}$$

其中

$$s = x^{(k)} - x^{(k-1)}, \quad y = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$$

逆更新

$$H_k^{-1} = \left(I - \frac{sy^T}{y^T s}\right) H_{k-1}^{-1} \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s}$$

- 注意 $y^T s > 0$ 对于严格凸的 f；也适用于线搜索
- 更新或逆更新的成本是 $O(n^2)$ 操作

正定性

如果 $y^T s > 0$，则 BFGS 更新保持 H_k 的正定性。(可适用于非凸函数)

证明： 从逆更新公式，

$$v^T H_k^{-1} v = \left(v - \frac{s^T v}{s^T y} y\right)^T H_{k-1}^{-1} \left(v - \frac{s^T v}{s^T y} y\right) + \frac{(s^T v)^2}{y^T s}$$

- 如果 $H_{k-1} \succ 0$，则对于所有 v，两项都是非负的
- 第二项仅在 $s^T v = 0$ 时为零；那么第一项仅在 $v = 0$ 时为零

这确保了 $\Delta x = -H_k^{-1} \nabla f(x^k)$ 是一个下降方向

BFGS 更新的最优性

$X = H_k$ 是下面凸优化问题的最优解

$$\begin{aligned} \min \quad & \mathrm{tr}(H_{k-1}^{-1}X) - \log\det(H_{k-1}^{-1}X) - n \\ \text{s.t.} \quad & Xs = y \end{aligned}$$

- 目标函数函数是非负的，仅当 $X = H_{k-1}$ 时为零
- 目标函数也代表 $N(0, X)$ 和 $N(0, H_{k-1})$ 之间的相对熵

最优性结果是由于 KKT 条件：$X = H_k$ 满足

$$X^{-1} = H_{k-1}^{-1} - \frac{1}{2}(sv^T + vs^T), \quad Xs = y, \quad X \succ 0$$

其中

$$v = \frac{1}{s^T y}\left(2H_{k-1}^{-1}y - \left(1 + \frac{y^T H_{k-1}^{-1} y}{y^T s}\right)s\right)$$

Broyden, Fletcher, Goldfarb, Shanno

Davidon-Fletcher-Powell (DFP) 更新

在之前的目标中交换 H_{k-1} 和 X

$$\begin{aligned} \min \quad & \operatorname{tr}(H_{k-1}X^{-1}) - \log\det(H_{k-1}X^{-1}) - n \\ \text{s.t.} \quad & Xs = y \end{aligned}$$

- 最小化 $N(0, H_{k-1})$ 和 $N(0, X)$ 之间的相对熵
- 问题关于 X^{-1} 中是凸的（约束条件写为 $s = X^{-1}y$）
- 相当于 BFGS 公式的 “对偶”

$$H_k = \left(I - \frac{ys^T}{s^Ty}\right) H_{k-1} \left(I - \frac{sy^T}{s^Ty}\right) + \frac{yy^T}{s^Ty}$$

（称为 DFP 更新）

- 方法提出先于 BFGS 更新，但实际效果不如 BFGS，实际使用较少

收敛性

全局结果（2021 年）

- 如果 f 是强凸的，采用线搜索的 BFGS 算法，对于任意初始设置 $x^{(0)}, H_0 \succ 0$ ，迭代复杂度（全局）为**超线性收敛**

局部收敛（1976 年）

- 如果 f 是强凸的且 $\nabla^2 f(x)$ 是 Lipschitz 连续的，局部收敛是超线性的：对于足够大的 k，

$$\|x^{(k+1)} - x^*\|_2 \leq c_k \|x^{(k)} - x^*\|_2 \to 0$$

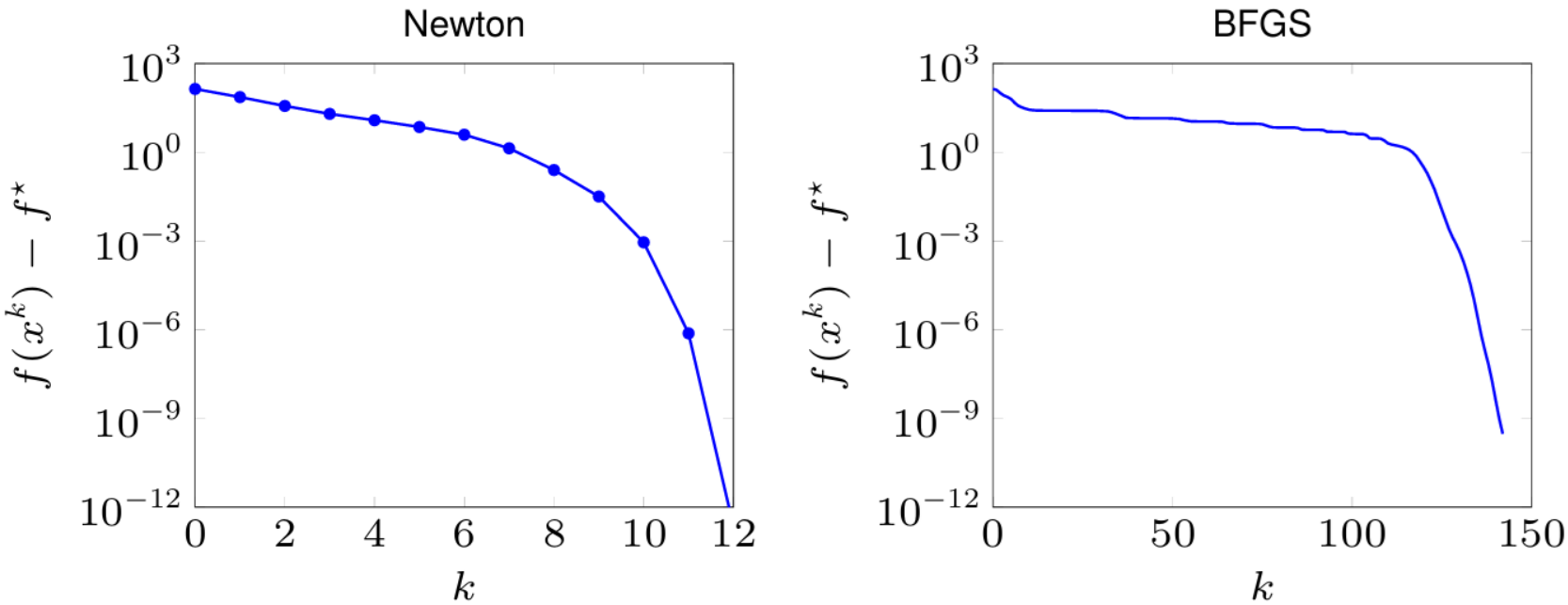
  其中 $c_k \to 0$

（参见牛顿法的二次局部收敛性）

例子

$$\min \quad c^T x - \sum_{i=1}^m \log(b_i - a_i^T x)$$

- $n = 100, m = 500$



- 每次牛顿迭代的成本：$O(n^3)$ 加上计算 $\nabla^2 f(x)$
- 每次 BFGS 迭代的成本：$O(n^2)$

目录

讲　　员

郭加熠，江波，刘慧康

有限内存拟牛顿方法

拟牛顿方法的主要缺点是需要存储 H_k 或 H_k^{-1}

有限内存 BFGS (L-BFGS)：不需要存储 H_k^{-1}

- 存储 $2m$ （例如，$m = 30$）个向量，代表最近的 m 次迭代中的

$$s_j = x^{(j)} - x^{(j-1)}, \quad y_j = \nabla f(x^{(j)}) - \nabla f(x^{(j-1)})$$

- 假设例如，$H_{k-m}^{-1} = I$，利用储存向量对于 $j = k, k-1, \ldots, k-m+1$，递归地计算 $\Delta x = H_k^{-1} \nabla f(x^{(k)})$，使用

$$H_j^{-1} = \left(I - \frac{s_j y_j^T}{y_j^T s_j} \right) H_{j-1}^{-1} \left(I - \frac{y_j s_j^T}{y_j^T s_j} \right) + \frac{s_j s_j^T}{y_j^T s_j}$$

（如果足够聪明，不需要计算逆矩阵）

- 每次迭代的成本是 $O(nm)$；存储是 $O(nm)$

总结

	梯度下降法	牛顿法	拟牛顿法	L-BFGS
内存/迭代	$O(n)$	$O(n^2)$	$O(n^2)$	$O(mn)$
计算量/迭代	$O(n)$	$O(n^3)$	$O(n^2)$	$O(mn)$
局部收敛率	线性	二次	超线性	线性
条件数	影响严重	不受影响	不受影响	温和
脆弱性	稳健	敏感	温和	稳健