

# 第五章 无约束优化算法

修贤超

<https://xianchaoxiu.github.io>

- 5.1 线搜索方法
- 5.2 梯度类算法
- 5.3 次梯度算法
- 5.4 牛顿类算法
- 5.5 拟牛顿类算法
- 5.6 信赖域算法
- 5.7 非线性最小二乘问题算法

# 梯度法的困难

- 考虑无约束优化问题

$$\min_{x \in \mathbb{R}^n} f(x)$$

- 梯度下降法

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

- 当  $\nabla^2 f(x)$  的条件数较大时, 收敛速度比较缓慢
- 如果  $f(x)$  足够光滑, 可以利用  $f(x)$  的二阶信息改进下降方向以加速迭代

- 对于可微二次函数  $f(x)$ , 考虑目标函数  $f$  在点  $x^k$  的二阶泰勒近似

$$f(x^k + d^k) = f(x^k) + \nabla f(x^k)^\top d^k + \frac{1}{2}(d^k)^\top \nabla^2 f(x^k) d^k + o(\|d^k\|^2)$$

- 将等式右边视作  $d^k$  的函数并极小化, 得牛顿方程

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k)$$

- 若  $\nabla^2 f(x^k)$  非奇异, 可构造迭代格式

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

当步长  $\alpha_k = 1$  时, 称为**经典牛顿法**

# 经典牛顿法的收敛性

- **定理 5.6** 假设目标函数  $f$  是二阶连续可微函数, 且海森矩阵在最优值点  $x^*$  的一个邻域  $N_\delta(x^*)$  内是利普希茨连续的, 即存在常数  $L > 0$  使得

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in N_\delta(x^*)$$

如果  $f(x)$  在点  $x^*$  处满足  $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$ , 则对于经典牛顿法有

- 如果初始点离  $x^*$  足够近, 则迭代点列  $\{x^k\}$  收敛到  $x^*$
- $\{x^k\}$  Q-二次收敛到  $x^*$
- $\{\|\nabla f(x^k)\|\}$  Q-二次收敛到 0

# 收敛速度分析

- 初始点 $x^0$  需要距离最优解充分近
- 常以梯度类算法先求得较低精度的解, 后用牛顿法加速
- $\nabla^2 f(x^*)$  需正定, 半正定条件下可能退化到 Q-线性收敛
- $\nabla^2 f$  的条件数较高时, 将对初值的选择作出较严苛的要求

- 经典牛顿法的基本格式如下

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

- 海瑟矩阵可能非正定, 导致牛顿方向其实并非下降方向;
- 初始点离最优值点较远时候迭代不稳定, 算法可能不收敛

- 为提高算法的稳定性, 应该

- 对  $\nabla^2 f(x)$  进行修正, 使其正定
- 用线搜索确定步长来增加算法的稳定性 (Wolfe, Goldstein, Armijo)

# 带线搜索的修正牛顿法

- 1 给定初始点  $x^0$
- 2 **for**  $k = 0, 1, 2, \dots$  **do**
- 3 确定矩阵  $E^k$  使得矩阵  $B^k = \nabla^2 f(x^k) + E^k$  正定且条件数较小
- 4 求解修正的牛顿方程  $B^k d^k = -\nabla f(x^k)$  得方向  $d^k$
- 5 使用任意一种线搜索准则确定步长  $\alpha_k$
- 6 更新  $x^{k+1} = x^k + \alpha_k d^k$
- 7 **end for**

=====

- $B^k$  应具有较低的条件数
- 对  $\nabla^2 f(x)$  的改动较小, 以保存二阶信息
- $B^k$  本身的计算代价不应太高



# 非精确牛顿法

- 当变量维数很大时，牛顿法可能有如下困难

- 海瑟矩阵  $\nabla^2 f(x)$  的计算、存储存在困难
- 对  $\nabla^2 f(x)$  求逆或者做 Cholesky 分解的代价很高

- 非精确牛顿法

- 使用迭代法求解牛顿方程，在一定的精度下**提前停机**，以提高求解效率
- 引入向量  $r^k$  来表示残差，将上述方程记为

$$\nabla^2 f(x^k)d^k = -\nabla f(x^k) + r^k$$

因此终止条件可设置为

$$\|r^k\| \leq \eta_k \|\nabla f(x^k)\|$$

- 不同的  $\{\eta_k\}$  将导致不同的精度要求，使算法有不同的收敛速度

# 应用举例：逻辑回归模型

## ■ 考虑二分类的逻辑回归模型

$$\min_x \ell(x) = \frac{1}{m} \sum_{i=1}^m \ln(1 + \exp(-b_i a_i^\top x)) + \lambda \|x\|_2^2$$

## ■ 为使用牛顿法, 需要计算目标函数的梯度与海瑟矩阵

$$\begin{aligned} \nabla \ell(x) &= \frac{1}{m} \sum_{i=1}^m \frac{1}{1 + \exp(-b_i a_i^\top x)} \cdot \exp(-b_i a_i^\top x) \cdot (-b_i a_i) + 2\lambda x \\ &= -\frac{1}{m} \sum_{i=1}^m (1 - p_i(x)) b_i a_i + 2\lambda x \end{aligned}$$

其中  $p_i(x) = \frac{1}{1 + \exp(-b_i a_i^\top x)}$

- 进一步对  $\nabla \ell(x)$  求导, 成立

$$\begin{aligned}\nabla^2 \ell(x) &= \frac{1}{m} \sum_{i=1}^m b_i \cdot \nabla p_i(x) a_i^\top + 2\lambda I \\ &= \frac{1}{m} \sum_{i=1}^m b_i \frac{-1}{(1 + \exp(-b_i a_i^\top x))^2} \cdot \exp(-b_i a_i^\top x) \cdot (-b_i a_i a_i^\top) + 2\lambda I \\ &= \frac{1}{m} \sum_{i=1}^m (1 - p_i(x)) p_i(x) a_i a_i^\top + 2\lambda I \quad (b_i^2 = 1)\end{aligned}$$

## 应用举例：逻辑回归模型

- 引入矩阵  $A = [a_1, a_2, \dots, a_m]^\top \in \mathbb{R}^{m \times n}$ , 向量  $b = (b_1, b_2, \dots, b_m)^\top$ , 以及

$$p(x) = (p_1(x), p_2(x), \dots, p_m(x))^\top,$$

- 则可重写梯度和海瑟矩阵为

$$\nabla \ell(x) = -\frac{1}{m} A^\top (b - b \odot p(x)) + 2\lambda x$$

$$\nabla^2 \ell(x) = \frac{1}{m} A^\top W(x) A + 2\lambda I$$

其中  $W(x)$  为由  $\{p_i(x)(1 - p_i(x))\}_{i=1}^m$  生成的对角矩阵

- 最终牛顿法迭代格式为

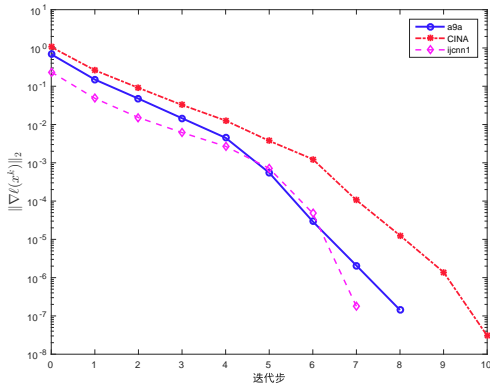
$$x^{k+1} = x^k + \left( \frac{1}{m} A^\top W(x^k) A + 2\lambda I \right)^{-1} \left( \frac{1}{m} A^\top (b - b \odot p(x^k)) - 2\lambda x^k \right)$$

# 应用举例：逻辑回归模型

## ■ 设置精度条件为

$$\|\nabla^2 \ell(x^k) d^k + \nabla \ell(x^k)\|_2 \leq \min\{\|\nabla \ell(x^k)\|_2^2, 0.1 \|\nabla \ell(x^k)\|_2\}$$

名称	$m$	$n$
a9a	16281	122
CINA	3206	132
ijcnn1	91701	22



- 5.1 线搜索方法
- 5.2 梯度类算法
- 5.3 次梯度算法
- 5.4 牛顿类算法
- 5.5 拟牛顿类算法
- 5.6 信赖域算法
- 5.7 非线性最小二乘问题算法

# 割线方程的推导

- 设  $f(x)$  是二阶连续可微函数. 对  $\nabla f(x)$  在点  $x^{k+1}$  处一阶泰勒近似, 得

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

- 令  $x = x^k$ , 且  $s^k = x^{k+1} - x^k$  为点差,  $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$  为梯度差, 得

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

- 现忽略高阶项  $\|s^k\|^2$ , 只希望近似海瑟矩阵的矩阵  $B^{k+1}$  满足方程

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

或其逆矩阵  $H^{k+1}$  满足

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

- 上述两个方程称为割线方程

## 曲率条件

- 由于近似矩阵必须保证迭代收敛, 正如牛顿法要求海瑟矩阵正定,  $B^k$  正定也是必须的, 即有必要条件

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

- 在迭代过程中满足  $(s^k)^\top y^k > 0, \forall k \in \mathbb{N}^+$
- 如果线搜索使用 Wolfe 准则

$$\nabla f(x^k + \alpha d^k)^\top d^k \geq c_2 \nabla f(x^k)^\top d^k$$

其中  $c_2 \in (0, 1)$ . 上式即  $\nabla f(x^{k+1})^\top s^k \geq c_2 \nabla f(x^k)^\top s^k$ . 在不等式两边同时减去  $\nabla f(x^k)^\top s^k$ , 由于  $c_2 - 1 < 0$  且  $s^k$  是下降方向, 因此最终有

$$(y^k)^\top s^k \geq (c_2 - 1) \nabla f(x^k)^\top s^k > 0$$



# 拟牛顿算法的基本框架

- 1 给定初始坐标  $x^0 \in \mathbb{R}^n$ , 初始矩阵  $B^0 \in \mathbb{R}^{n \times n}$  (或  $H^0$ ),  $k = 0$
- 2 **while** 未达到停机准则 **do**
- 3 计算方向  $d^k = -(B^k)^{-1} \nabla f(x^k)$  或  $d^k = -H^k \nabla f(x^k)$
- 4 通过线搜索 (Wolfe) 产生步长  $\alpha_k > 0$ , 令  $x^{k+1} = x^k + \alpha_k d^k$
- 5 更新海瑟矩阵的近似矩阵  $B^{k+1}$  或其逆矩阵  $H^{k+1}$
- 6  $k \leftarrow k + 1$
- 7 **end while**

## 秩一更新 (SR1)

- 对于拟牛顿矩阵  $B^k \in \mathbb{R}^{n \times n}$ , 设  $0 \neq u \in \mathbb{R}^n$  且  $a \in \mathbb{R}$  待定, 则  $uu^\top$  是秩一矩阵, 且有秩一更新

$$B^{k+1} = B^k + a uu^\top$$

- 根据割线方程  $B^{k+1}s^k = y^k$ , 代入秩一更新的结果, 得到

$$(B^k + a uu^\top)s^k = y^k$$

整理得

$$a uu^\top s^k = (a \cdot u^\top s^k)u = y^k - B^k s^k$$

- 由于  $a \cdot u^\top s^k$  是标量, 因此上式表明  $u$  和  $y^k - B^k s^k$  同向. 令  $u$  和  $y^k - B^k s^k$  相等, 即  $u = y^k - B^k s^k$ . 代入上式得

$$(a \cdot (y^k - B^k s^k)^\top s^k)(y^k - B^k s^k) = y^k - B^k s^k$$

# 秩一更新公式

- 再令  $(a \cdot (y^k - B^k s^k)^\top s^k) \neq 0$ , 则可以确定  $a$  为

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

- 拟牛顿矩阵  $B^k$  的秩一更新公式为

$$B^{k+1} = B^k + \frac{uu^\top}{u^\top s^k}, \quad u = y^k - B^k s^k$$

拟牛顿矩阵  $H^k$  的秩一更新公式为

$$H^{k+1} = H^k + \frac{vv^\top}{v^\top y^k}, \quad v = s^k - H^k y^k$$

- $B^k$  和  $H^k$  的公式在形式上互为对偶

- 对于拟牛顿矩阵  $B^k \in \mathbb{R}^{n \times n}$ , 设  $0 \neq u, v \in \mathbb{R}^n$  且  $a, b \in \mathbb{R}$  待定, 则有秩二更新形式

$$B^{k+1} = B^k + a u u^\top + b v v^\top$$

- 根据割线方程, 将秩二更新的待定参量式代入, 得

$$B^{k+1} s^k = (B^k + a u u^\top + b v v^\top) s^k = y^k,$$

整理可得

$$(a \cdot u^\top s^k) u + (b \cdot v^\top s^k) v = y^k - B^k s^k.$$

- 令  $(a \cdot u^\top s^k) u$  对应  $y^k$  相等,  $(b \cdot v^\top s^k) v$  对应  $-B^k s^k$  相等, 即有

$$a \cdot u^\top s^k = 1, \quad u = y^k, \quad b \cdot v^\top s^k = -1, \quad v = B^k s^k$$

- 将上述参量代入割线方程, 即得 BFGS 更新公式

$$B^{k+1} = B^k + \frac{uu^\top}{(s^k)^\top u} - \frac{vv^\top}{(s^k)^\top v}$$

利用 SMW 公式以及  $H^k = (B^k)^{-1}$ , 可以推出关于  $H^k$  的 BFGS 公式

- 在拟牛顿类算法中, 基于  $B^k$  的 BFGS 公式为

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

基于  $H^k$  的 BFGS 公式为

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

- DFP 公式利用与 BFGS 公式类似的推导方法, 不同的是其以割线方程  $H^{k+1}y^k = s^k$  为基础进行对  $H^k$  的秩二更新
- 基于  $H^k$  满足的 DFP 公式, 利用 SMW 公式以及  $B^k = (H^k)^{-1}$ , 可以推出关于  $B^k$  的 DFP 公式
- 基于  $H^k$  的 DFP 更新公式为

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

基于  $B^k$  的 DFP 更新公式为

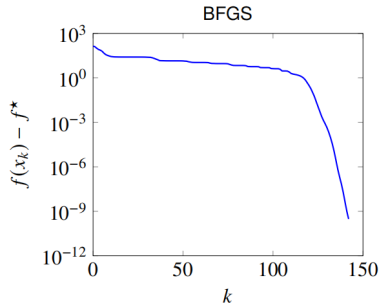
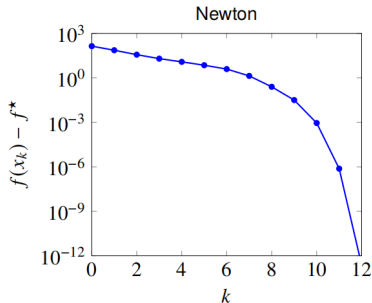
$$B^{k+1} = (I - \frac{y^k (s^k)^\top}{(s^k)^\top y^k})^\top B^k (I - \frac{y^k (s^k)^\top}{(s^k)^\top y^k}) + \frac{y^k (y^k)^\top}{(s^k)^\top y^k}$$

# 例子

## ■ 考虑极小化问题

$$\min_{x \in \mathbb{R}^{100}} c^\top x - \sum_{i=1}^{500} \ln(b_i - a_i^\top x)$$

- 牛顿法每次迭代的计算代价为  $\mathcal{O}(n^3)$  加上计算海瑟矩阵的代价, 而 BFGS 方法的每步计算代价仅为  $\mathcal{O}(n^2)$ , 因此 BFGS 算法可能更快取得最优解



*Q&A*

*Thank you!*

感谢您的聆听和反馈