

# 第三章 无约束优化算法

修贤超

<https://xianchaoxiu.github.io>

- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 非线性最小二乘问题算法

# 引言: 无约束可微优化算法

## ■ 考虑无约束优化问题

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

## ■ 线搜索 $x^{k+1} = x^k + \alpha_k d^k$

- 先确定下降方向: 负梯度、牛顿方向、拟牛顿方向等
- 按某种准则搜索步长

## ■ 信赖域 $z^k = x^k + d^k$

$$d^k = \arg \min_d (g^k)^\top d + d^\top B d \quad \text{s.t.} \quad \|d\|_2 \leq \Delta_k$$

- 给定信赖域半径 (步长)  $\Delta_k$ , 构造信赖域子问题求解方向  $d^k$
- 如果  $z^k$  满足下降性条件, 则  $x^{k+1} = z^k$ , 否则  $x^{k+1} = x^k$  更新  $\Delta_k$

# 线搜索算法

- 求解  $f(x)$  的最小值点如同盲人下山, 无法一眼望知谷底
  - 首先确定下一步该向哪一方向行走
  - 再确定沿着该方向行走多远后停下以便选取下一个下山方向
- 线搜索类算法的数学表述

$$x^{k+1} = x^k + \alpha_k d^k$$

- $\alpha_k$  为步长
- $d^k$  为下降方向, 即  $(d^k)^\top \nabla f(x^k) < 0$

## $\alpha_k$ 的选取

- 首先构造一元辅助函数

$$\phi(\alpha) = f(x^k + \alpha d^k)$$

- 线搜索的目标是选取合适的  $\alpha_k$  使得  $\phi(\alpha_k)$  尽可能减小

- $\alpha_k$  应该使得  $f$  充分下降
- 不应在寻找  $\alpha_k$  上花费过多的计算量

- 一个自然的想法是寻找  $\alpha_k$  使得

$$\alpha_k = \arg \min_{\alpha > 0} \phi(\alpha)$$

- 称为**精确线搜索算法**, 在实际应用中较少使用

- 考虑一维无约束优化问题

$$\min_x f(x) = x^2$$

- 迭代初始点  $x^0 = 1$ , 下降方向  $d^k = -\text{sign}(x^k)$

- 选取如下两种步长

$$\alpha_{k,1} = \frac{1}{3^{k+1}}, \quad \alpha_{k,2} = 1 + \frac{2}{3^{k+1}}$$

- 简单计算可以得到

$$x_1^k = \frac{1}{2}\left(1 + \frac{1}{3^k}\right), \quad x_2^k = \frac{(-1)^k}{2}\left(1 + \frac{1}{3^k}\right)$$

- 序列  $\{f(x_1^k)\}$  和序列  $\{f(x_2^k)\}$  均单调下降, 但序列  $\{x_1^k\}$  收敛的点不是极小值点, 序列  $\{x_2^k\}$  则在原点左右振荡, 不存在极限

- **定义** 设  $d^k$  是点  $x^k$  处的下降方向, 若

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^\top d^k$$

则称步长  $\alpha$  满足 **Armijo 准则**

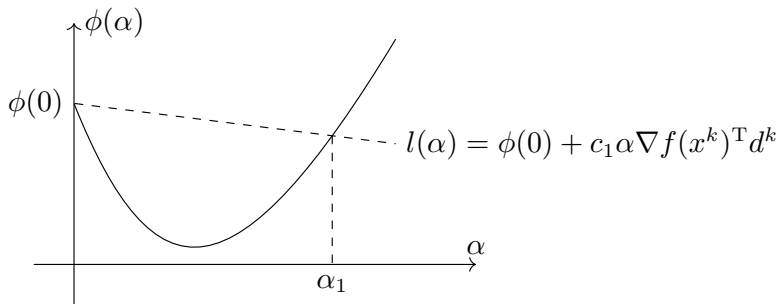
- 参数  $c_1 \in (0, 1)$  是一个常数, 通常取  $c_1 = 10^{-3}$
- 引入 Armijo 准则保证每一步迭代充分下降
- 需要配合其他准则以保证迭代的收敛性, 反例  $\alpha = 0$

# 几何含义

- 点  $(\alpha, \phi(\alpha))$  必须在直线

$$l(\alpha) = \phi(0) + c_1 \alpha \nabla f(x^k)^\top d^k$$

的下方, 图中区间  $[0, \alpha_1]$  中的点均满足 Armijo 准则





- 给定初值  $\hat{\alpha}$ , 以指数方式缩小试探步长, 找到第一个满足 Armijo 准则的点

$$\alpha_k = \gamma^{j_0} \hat{\alpha}$$

其中  $j_0 = \min\{j \mid f(x^k + \gamma^j \hat{\alpha} d^k) \leq f(x^k) + c_1 \gamma^j \hat{\alpha} \nabla f(x^k)^\top d^k\}, \gamma \in (0, 1)$

=====

## 算法 线搜索回退法

- 1 选择初始步长  $\hat{\alpha}$ , 参数  $\gamma, c \in (0, 1)$ . 初始化  $\alpha \leftarrow \hat{\alpha}$
- 2 **while**  $f(x^k + \alpha d^k) > f(x^k) + c\alpha \nabla f(x^k)^\top d^k$  **do**
- 3 令  $\alpha \leftarrow \gamma\alpha$
- 4 **end while**
- 5 输出  $\alpha_k = \alpha$

- **定义** 设  $d^k$  是点  $x^k$  处的下降方向, 若

$$\begin{aligned}f(x^k + \alpha d^k) &\leq f(x^k) + c\alpha \nabla f(x^k)^\top d^k, \\f(x^k + \alpha d^k) &\geq f(x^k) + (1 - c)\alpha \nabla f(x^k)^\top d^k\end{aligned}$$

则称步长  $\alpha$  满足 **Goldstein 准则**, 其中  $c \in (0, \frac{1}{2})$

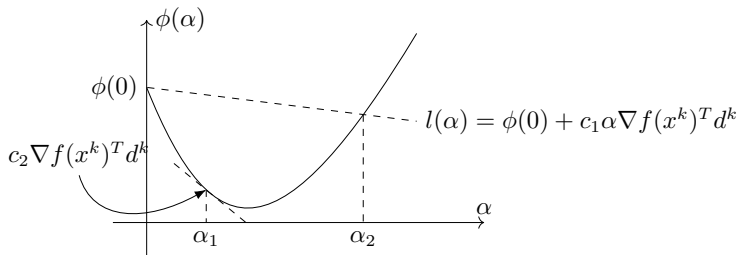
- **定义** 设  $d^k$  是点  $x^k$  处的下降方向, 若

$$\begin{aligned}f(x^k + \alpha d^k) &\leq f(x^k) + c_1 \alpha \nabla f(x^k)^\top d^k, \\ \nabla f(x^k + \alpha d^k)^\top d^k &\geq c_2 \nabla f(x^k)^\top d^k\end{aligned}$$

则称步长  $\alpha$  满足 **Wolfe 准则**, 其中  $c_1, c_2 \in (0, 1)$  为给定的常数且  $c_1 < c_2$

# Wolfe 准则

- Wolfe 准则实际要求  $\phi(\alpha)$  在点  $\alpha$  处切线的斜率不能小于  $\phi'(0)$  的  $c_2$  倍
- $\phi(\alpha)$  的极小值点  $\alpha^*$  处有  $\phi'(\alpha^*) = \nabla f(x^k + \alpha^* d^k)^T d^k = 0$ , 因此  $\alpha^*$  永远满足条件二. 而选择较小的  $c_1$  可使得  $\alpha^*$  同时满足条件一, 即 Wolfe 准则在绝大多数情况下会包含线搜索子问题的精确解



- **定理** 考虑一般的迭代格式  $x^{k+1} = x^k + \alpha_k d^k$ , 其中  $d^k$  是搜索方向,  $\alpha_k$  是步长, 且在迭代过程中 Wolfe 准则满足. 假设目标函数  $f$  下有界、连续可微且梯度  $L$ -利普希茨连续, 即

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

那么

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|^2 < +\infty$$

其中  $\cos \theta_k$  为负梯度  $-\nabla f(x^k)$  和下降方向  $d^k$  夹角的余弦, 即

$$\cos \theta_k = \frac{-\nabla f(x^k)^\top d^k}{\|\nabla f(x^k)\| \|d^k\|}$$

这个不等式也被称为 Zoutendijk 条件

## 线搜索算法的收敛性

- **推论** 对于迭代法  $x^{k+1} = x^k + \alpha_k d^k$ , 设  $\theta_k$  为每一步负梯度  $-\nabla f(x^k)$  与下降方向  $d^k$  的夹角, 并假设对任意的  $k$ , 存在常数  $\gamma > 0$ , 使得

$$\theta_k < \frac{\pi}{2} - \gamma$$

则在 Zoutendijk 定理成立的条件下, 有

$$\lim_{k \rightarrow \infty} \nabla f(x^k) = 0$$

## 线搜索算法收敛性的证明

**证明** 假设结论不成立, 即存在子列  $\{k_l\}$  和正常数  $\delta > 0$ , 使得

$$\|\nabla f(x^{k_l})\| \geq \delta, \quad l = 1, 2, \dots$$

根据  $\theta_k$  的假设, 对任意的  $k$  有

$$\cos \theta_k > \sin \gamma > 0$$

仅考虑 Zoutendijk 条件中第  $k_l$  项的和满足

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|^2 \geq \sum_{l=1}^{\infty} \cos^2 \theta_{k_l} \|\nabla f(x^{k_l})\|^2$$

这显然和 Zoutendijk 定理矛盾

- 35.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 非线性最小二乘问题算法

# 梯度下降法

- 注意到  $\phi(\alpha) = f(x^k + \alpha d^k)$  有泰勒展开

$$\phi(\alpha) = f(x^k) + \alpha \nabla f(x^k)^\top d^k + \mathcal{O}(\alpha^2 \|d^k\|^2)$$

- 由柯西不等式, 当  $\alpha$  足够小时取  $d^k = -\nabla f(x^k)$  会使函数下降最快

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

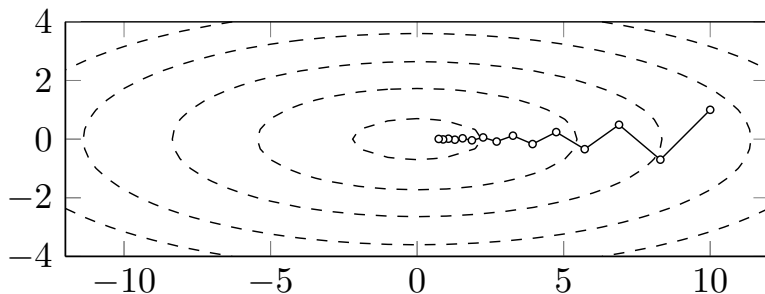
- 另一种理解方式

$$\begin{aligned} x^{k+1} &= \arg \min_x f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{\alpha_k} \|x - x^k\|_2^2 \\ &= \arg \min_x \|x - (x^k - \alpha_k \nabla f(x^k))\|_2^2 \\ &= x^k - \alpha_k \nabla f(x^k) \end{aligned}$$



## 二次函数的梯度法

- 设二次函数  $f(x, y) = x^2 + 10y^2$ , 初始点  $(x^0, y^0)$  取为  $(10, 1)$ , 取固定步长  $\alpha_k = 0.085$ , 使用梯度法  $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$  进行 15 次迭代



# 二次函数的收敛定理

■ **定理** 考虑正定二次函数

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x$$

设最优值点为  $x^*$ . 若使用梯度法  $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$  并选取  $\alpha_k$  为精确线搜索步长, 即

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^\top A \nabla f(x^k)}$$

则梯度法关于迭代点列  $\{x^k\}$  是 **Q-线性收敛**, 即

$$\|x^{k+1} - x^*\|_A^2 \leq \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 \|x^k - x^*\|_A^2$$

# 梯度法在凸函数上的收敛性

- 对于可微函数  $f$ , 若存在  $L > 0$ , 对任意的  $x, y \in \text{dom } f$  有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

则称  $f$  是**梯度利普希茨连续的**, 相应利普希茨常数为  $L$

- **定理** 设  $f(x)$  为 **凸的梯度  $L$ -利普希茨连续函数**,  $f^* = f(x^*) = \inf_x f(x)$  存在且可达, 如果步长  $\alpha_k$  取为常数  $\alpha$  且满足  $0 < \alpha < \frac{1}{L}$ , 那么点列  $\{x^k\}$  的函数值收敛到最优值, 且在函数值的意义下收敛速度为  $\mathcal{O}(\frac{1}{k})$

# 梯度法在强凸函数上的收敛性

■ **引理** 设函数  $f(x)$  是  $\mathbb{R}^n$  上的凸可微函数, 则以下结论等价

- $f$  的梯度为  $L$ -利普希茨连续的
- 函数  $g(x) = \frac{L}{2}x^\top x - f(x)$  是凸函数
- $\nabla f(x)$  有**余强制性**, 即对任意的  $x, y \in \mathbb{R}^n$ , 有

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2$$

■ **定理** 设  $f(x)$  为  $m$ -强凸的梯度  $L$ -利普希茨连续函数,  $f(x^*) = \inf_x f(x)$  存在且可达. 如果步长  $\alpha$  满足  $0 < \alpha < \frac{2}{m+L}$ , 那么由梯度下降法迭代得到的点列  $\{x^k\}$  收敛到  $x^*$ , 且为**Q-线性收敛**

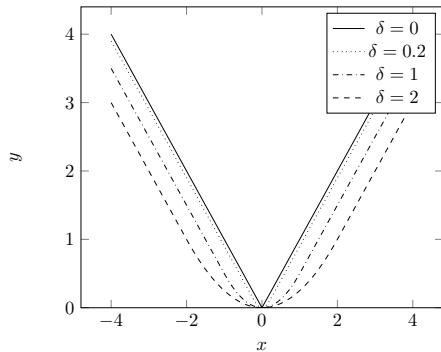
## 应用举例: LASSO 问题求解

### ■ 考虑

$$\min f(x) = \frac{1}{2} \|Ax - b\|^2 + \mu \|x\|_1$$

### ■ 由于 $\|x\|_1$ 不光滑, 考虑 Huber 光滑函数

$$l_{\delta}(x) = \begin{cases} \frac{1}{2\delta}x^2, & |x| < \delta \\ |x| - \frac{\delta}{2}, & \text{其他} \end{cases}$$



# 应用举例: LASSO 问题求解

## ■ 光滑化 LASSO 问题为

$$\min f_{\delta}(x) = \frac{1}{2} \|Ax - b\|^2 + \mu L_{\delta}(x), \quad \text{其中} \quad L_{\delta}(x) = \sum_{i=1}^n l_{\delta}(x_i)$$

## ■ $f_{\delta}(x)$ 的梯度为

$$\nabla f_{\delta}(x) = A^{\top}(Ax - b) + \mu \nabla L_{\delta}(x)$$

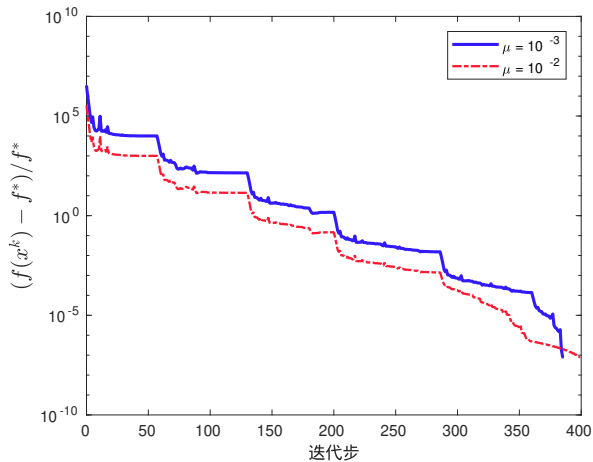
其中

$$(\nabla L_{\delta}(x))_i = \begin{cases} \text{sign}(x_i), & |x_i| > \delta \\ \frac{x_i}{\delta}, & |x_i| \leq \delta \end{cases}$$

## ■ $f_{\delta}(x)$ 的梯度是利普希茨连续的, 且相应常数为 $L = \|A^{\top}A\|_2 + \frac{\mu}{\delta}$

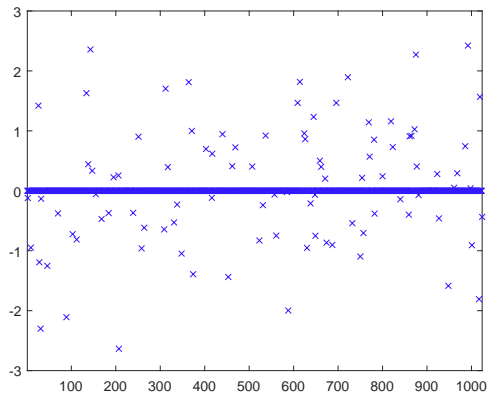
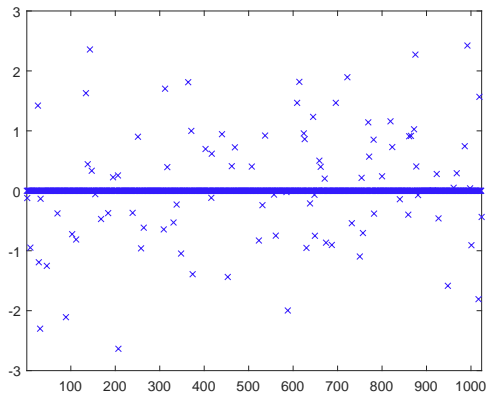
# 应用举例: LASSO 问题求解

## ■ 光滑化 LASSO 问题求解迭代过程



# 应用举例: LASSO 问题求解

## ■ 精确解 (左) v.s. 梯度法解 (右)





- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 非线性最小二乘问题算法

# 次梯度算法结构

## ■ 回顾一阶充要条件

$$x^* \text{ 是一个全局极小点 } \Leftrightarrow 0 \in \partial f(x^*)$$

## ■ 类似梯度法构造如下次梯度算法的迭代格式

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k)$$

- 固定步长  $\alpha_k = \alpha$
- 固定  $\|x^{k+1} - x^k\|$ , 即  $\alpha_k \|g^k\|$  为常数
- 消失步长  $\alpha_k \rightarrow 0$  且  $\sum_{k=0}^{\infty} \alpha_k = +\infty$
- 选取  $\alpha_k$  使其满足某种线搜索准则

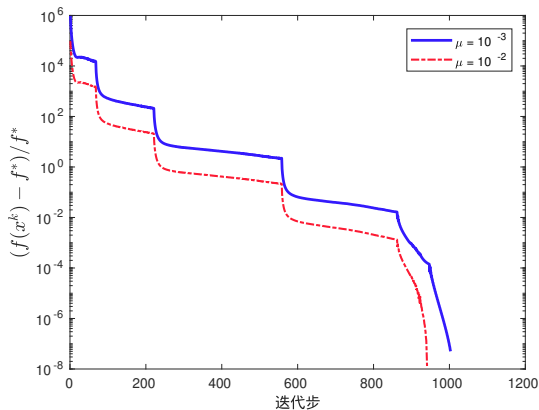
# 应用举例: LASSO 问题求解

## ■ 考虑 LASSO 问题

$$\min f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1$$

## ■ 次梯度算法

$$x^{k+1} = x^k - \alpha_k(A^\top(Ax^k - b) + \mu\text{sign}(x^k))$$



- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 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, 得到**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** 假设目标函数  $f$  是二阶连续可微函数, 且海瑟矩阵在最优值点  $x^*$  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  $x^*$
- $\{x^k\}$  是 Q-二次收敛到  $x^*$
- $\{\|\nabla f(x^k)\|\}$  是 Q-二次收敛到 0

# 收敛速度分析

## ■ 经典牛顿法说明

- 初始点  $x^0$  需要距离最优解充分近, 即只有**局部收敛性**
- $\nabla^2 f(x^*)$  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

## ■ 解决方案

- 先以梯度类算法求得较低精度的解, 然后用牛顿法加速
- 修正牛顿法
- 非精确牛顿法
- 拟牛顿类算法



# 修正牛顿法

## 算法 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

# 非精确牛顿法

- 当变量维数很大时, 海瑟矩阵  $\nabla^2 f(x)$  计算存在困难, 且求逆代价很高
- 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

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

## ■ 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

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

## 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

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

- 重写梯度和海瑟矩阵为

$$\begin{aligned}\nabla \ell(x) &= -\frac{1}{m} A^\top (b - b \odot p(x)) + 2\lambda x \\ \nabla^2 \ell(x) &= \frac{1}{m} A^\top W(x) A + 2\lambda I\end{aligned}$$

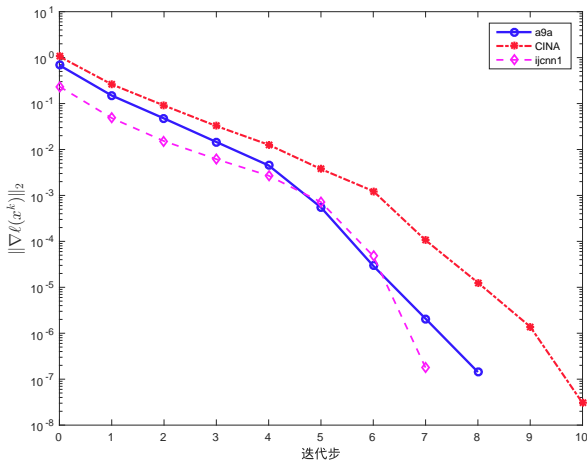
- 最终牛顿法迭代格式为

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

# 应用举例：逻辑回归模型

## ■ 设置精度条件为

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



- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 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

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

- 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

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

忽略高阶项  $\|s^k\|^2$ , 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

- 近似矩阵  $B^k$  正定, 即有必要条件

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

- 如果线搜索使用 Wolfe 准则

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

$$\Downarrow$$

$$\nabla f(x^{k+1})^\top s^k \geq c_2 \nabla f(x^k)^\top s^k$$

两边同时减去  $\nabla f(x^k)^\top s^k$ , 由于  $c_2 - 1 < 0$  且  $s^k$  是下降方向得到

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

# 拟牛顿算法的基本框架

## 算法 拟牛顿算法框架

- 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**

=====

- 基于  $H^k$  的拟牛顿算法更实用
- 基于  $B^k$  的拟牛顿算法有较好的理论性质

## 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  $B^{k+1}s^k = y^k$ , 代入秩一更新得到

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

$$\Downarrow$$

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

- 令  $u = y^k - B^k s^k$ , 代入上式有

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

# 秩一更新公式

■ 假设  $(a \cdot (y^k - B^k s^k)^\top s^k) \neq 0$ , 则  $a = \frac{1}{(y^k - B^k s^k)^\top s^k}$

■ 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

- 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  $B^{k+1} s^k = y^k$ , 将秩二更新的待定参量式代入得到

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

$$\Downarrow$$

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

- 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

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

- 在拟牛顿类算法中, 基于  $B^k$  的 BFGS 公式为

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

利用 Sherman-Morrison-Woodbury (SMW) 公式, 基于  $H^k$  的 BFGS 公式为

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

# 拟牛顿法的全局收敛性

- **定理** 假设初始矩阵  $B^0$  是对称正定矩阵, 目标函数  $f(x)$  是二阶连续可微函数, 下水平集

$$\mathcal{L} = \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$$

是凸的, 且存在  $m, M \in \mathbb{R}^+$  使得对  $\forall z \in \mathbb{R}^n, x \in \mathcal{L}$  满足

$$m \|z\|^2 \leq z^\top \nabla^2 f(x) z \leq M \|z\|^2$$

那么 BFGS 结合 Wolfe 线搜索的拟牛顿算法全局收敛到  $f(x)$  的极小值点  $x^*$

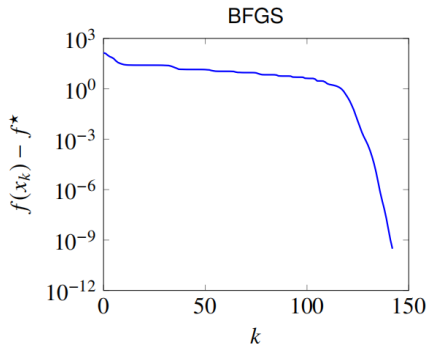
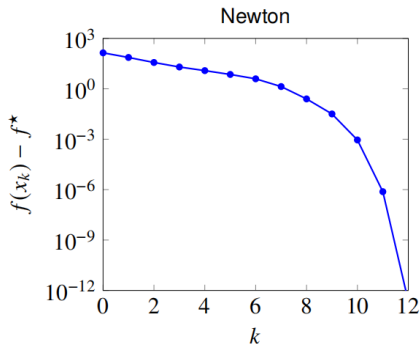
- 如果海瑟矩阵在  $x^*$  处 Lip-连续, 则迭代点列  $\{x^k\}$  为 **Q-超线性收敛**到  $x^*$

# 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, 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





- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 非线性最小二乘问题算法

- 在当前迭代点  $x^k$  建立局部模型, 求出最优解

$$d^k = \arg \min_d \quad (g^k)^\top d + d^\top B d \quad \text{s.t.} \quad \|d\|_2 \leq \Delta_k$$

- 更新模型信赖域的半径

- 模型足够好  $\Rightarrow$  增大半径
- 模型比较差  $\Rightarrow$  缩小半径
- 否则半径不变

- 对模型进行评价

- 好  $\Rightarrow$  子问题的解即下一个迭代点
- 差  $\Rightarrow$  迭代点不改变

- 根据带拉格朗日余项的泰勒展开

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

- 利用  $f(x)$  的二阶近似来刻画  $f(x)$  在点  $x^k$  处的性质

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

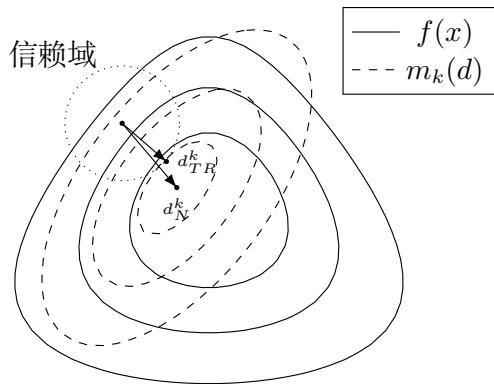
- 由于泰勒展开的局部性, 需对上述模型添加信赖域约束

$$\Omega_k = \{x^k + d \mid \|d\| \leq \Delta_k\}$$

# 信赖域子问题

- 信赖域算法每一步都需要求解如下子问题

$$\min_{d \in \mathbb{R}^n} m_k(d) \quad \text{s.t.} \quad \|d\| \leq \Delta_k \quad (3)$$



# 模型近似程度好坏的的衡量

## ■ 引入

$$\rho_k = \frac{f(x^k) - f(x^k + d^k)}{m_k(0) - m_k(d^k)} \quad (4)$$

- 函数值实际下降量与预估下降量（即二阶近似模型下降量）的比值
- 如果  $\rho_k$  接近 1, 说明  $m_k(d)$  来近似  $f(x)$  是比较成功的, 则扩大  $\Delta_k$
- 如果  $\rho_k$  非常小甚至为负, 说明过分地相信了二阶模型  $m_k(d)$ , 则缩小  $\Delta_k$

# 算法: 信赖域算法

- 1 给定最大半径  $\Delta_{\max}$ , 初始半径  $\Delta_0$ , 初始点  $x^0, k \leftarrow 0$
- 2 给定参数  $0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1, \gamma_1 < 1 < \gamma_2$
- 3 **while** 未达到停机准则 **do**
- 4 计算子问题 (3) 得到迭代方向  $d^k$
- 5 根据 (4) 计算下降率  $\rho_k$
- 6 更新信赖域半径

$$\Delta_{k+1} = \begin{cases} \gamma_1 \Delta_k, & \rho_k < \bar{\rho}_1 \\ \min\{\gamma_2 \Delta_k, \Delta_{\max}\}, & \rho_k > \bar{\rho}_2 \text{ 以及 } \|d^k\| = \Delta_k \\ \Delta_k, & \text{其他} \end{cases}$$

- 7 更新自变量

$$x^{k+1} = \begin{cases} x^k + d^k, & \rho_k > \eta \\ x^k, & \text{其他} \end{cases} \quad /* \text{ 只有下降比例足够大才更新 } */$$

- 8  $k \leftarrow k + 1$
- 9 **end while**

- 3.1 线搜索方法
- 3.2 梯度类算法
- 3.3 次梯度算法
- 3.4 牛顿类算法
- 3.5 拟牛顿类算法
- 3.6 信赖域算法
- 3.7 非线性最小二乘问题算法

# 非线性最小二乘问题

## ■ 考虑最小二乘问题

$$\min_x f(x) = \frac{1}{2} \sum_{j=1}^m r_j^2(x)$$

## ■ 记 $r(x) = (r_1(x), r_2(x), \dots, r_m(x))^\top$ , 问题可以表述为

$$\min f(x) = \frac{1}{2} \|r(x)\|_2^2$$

## ■ 记 $J(x) \in \mathbb{R}^{m \times n}$ 是向量值函数 $r(x)$ 在点 $x$ 处的雅可比矩阵

$$J(x) = \begin{bmatrix} \nabla r_1(x)^\top \\ \nabla r_2(x)^\top \\ \vdots \\ \nabla r_m(x)^\top \end{bmatrix}$$



## ■ $f(x)$ 的梯度和海瑟矩阵

$$\nabla f(x) = \sum_{j=1}^m r_j(x) \nabla r_j(x) = J(x)^\top r(x)$$

$$\begin{aligned}\nabla^2 f(x) &= \sum_{j=1}^m \nabla r_j(x) \nabla r_j(x) + \sum_{i=1}^m r_i(x) \nabla^2 r_i(x) \\ &= J(x)^\top J(x) + \sum_{i=1}^m r_i(x) \nabla^2 r_i(x)\end{aligned}$$

■ 小残差  $\Rightarrow$  高斯-牛顿方法和 Levenberg-Marquardt 方法

■ 大残差  $\Rightarrow$  引入带结构的拟牛顿方法

# 高斯-牛顿方法

- 使用近似  $\nabla^2 f_k \approx J_k^\top J_k$ , 省略  $\nabla^2 r_j$  的计算, 减少了计算量
- 高斯-牛顿法的迭代方向  $d_k^{GN}$  满足

$$J_k^\top J_k d_k^{GN} = -J_k^\top r_k$$

- 另一种理解: 在点  $x_k$  处, 考虑近似  $r(x_k + d) \approx r_k + J_k d$  得到

$$\min_d f(x_k + d) = \frac{1}{2} \|r(x_k + d)\|^2 \approx \frac{1}{2} \|r_k + J_k d\|^2$$

- 然后更新  $x_{k+1} = x_k + \alpha_k d_k$

# 算法: 高斯-牛顿方法

- 1 给定始值  $x_0, k \leftarrow 0$
- 2 **while** 未达到停机准则 **do**
- 3 计算残差向量  $r_k$ , 雅可比矩阵  $J_k$
- 4 求解线性最小二乘问题  $\min_d \frac{1}{2} \|r_k + J_k d\|^2$  确定下降方向  $d_k$
- 5 使用线搜索准则计算步长  $\alpha_k$
- 6 更新  $x_{k+1} = x_k + \alpha_k d_k$
- 7  $k \leftarrow k + 1$
- 8 **end while**

# Levenberg-Marquardt (LM) 方法

- LM 方法本质为信赖域方法, 更新方向为如下问题的解

$$\min_d \quad \frac{1}{2} \|J^k d + r^k\|^2 \quad \text{s.t.} \quad \|d\| \leq \Delta_k \quad (5)$$

- 将如下近似当作信赖域方法中的  $m_k$

$$m_k(d) = \frac{1}{2} \|r^k\|^2 + d^\top (J^k)^\top r^k + \frac{1}{2} d^\top (J^k)^\top J^k d$$

# Levenberg-Marquardt 方法

- 1 给定最大半径  $\Delta_{\max}$ , 初始半径  $\Delta_0$ , 初始点  $x^0, k \leftarrow 0$
- 2 给定参数  $0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1, \gamma_1 < 1 < \gamma_2$
- 3 **while** 未达到停机准则 **do**
- 4 计算子问题 (5) 得到迭代方向  $d^k$
- 5 根据 (4) 计算下降率  $\rho_k$
- 6 更新信赖域半径

$$\Delta_{k+1} = \begin{cases} \gamma_1 \Delta_k, & \rho_k < \bar{\rho}_1 \\ \min\{\gamma_2 \Delta_k, \Delta_{\max}\}, & \rho_k > \bar{\rho}_2 \text{ 以及 } \|d^k\| = \Delta_k \\ \Delta_k, & \text{其他} \end{cases}$$

- 7 更新自变量

$$x^{k+1} = \begin{cases} x^k + d^k, & \rho_k > \eta \\ x^k, & \text{其他} \end{cases} \quad /* \text{ 只有下降比例足够大才更新 } */$$

- 8  $k \leftarrow k + 1$
- 9 **end while**

## 子问题 (3) 求解

- **推论** 向量  $d^*$  是信赖域子问题

$$\min_d \quad \frac{1}{2} \|Jd + r\|^2 \quad \text{s.t.} \quad \|d\| \leq \Delta$$

的解当且仅当  $d^*$  是可行解并且存在数  $\lambda \geq 0$  使得

$$\begin{aligned}(J^\top J + \lambda I)d^* &= -J^\top r \\ \lambda(\Delta - \|d^*\|) &= 0\end{aligned}$$

- 实际上,  $(J^\top J + \lambda I)d^* = -J^\top r$  是最小二乘问题的最优性条件

$$\min_d \quad \frac{1}{2} \left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2$$

- 信赖域型 LM 方法本质上是固定信赖域半径  $\Delta$ , 通过迭代寻找满足条件的乘子  $\lambda$ , 每一步迭代需要求解线性方程组

$$(J^\top J + \lambda I)d = -J^\top r$$

- LM 的更新基于  $\Delta$ , LMF 的更新直接基于  $\lambda$ , 每一步求解子问题

$$\min_d \quad \frac{1}{2} \|J^k d + r^k\|^2 \quad \text{s.t.} \quad \|d\| \leq \Delta$$

$$\Downarrow$$

$$\min_d \quad \|Jd + r\|_2^2 + \lambda \|d\|_2^2$$

- 调整  $\lambda$  的原则可以参考信赖域半径的调整原则

# 算法 LMF 方法

- 1 给定初始点  $x_0$ , 初始乘子  $\lambda_0, k \leftarrow 0$
- 2 给定参数  $0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1, \gamma_1 < 1 < \gamma_2$
- 3 **while** 未达到停机准则 **do**
- 4 求解 LM 方程  $((J_k)^\top J_k + \lambda I)d = -(J_k)^\top r_k$  得到迭代方向  $d_k$
- 5 根据 (4) 式计算下降率  $\rho_k$
- 6 更新信赖域半径

$$\lambda_{k+1} = \begin{cases} \gamma_2 \lambda_k, & \rho_k < \bar{\rho}_1 & /* \text{ 扩大乘子 (缩小信赖域半径) } */ \\ \gamma_1 \lambda_k, & \rho_k > \bar{\rho}_2 & /* \text{ 缩小乘子 (扩大信赖域半径) } */ \\ \lambda_k, & \text{其他} & /* \text{ 乘子不变 } */ \end{cases}$$

- 7 更新自变量

$$x_{k+1} = \begin{cases} x_k + d_k, & \rho_k > \eta & /* \text{ 只有下降比例足够大才更新 } */ \\ x_k, & \text{其他} \end{cases}$$

- 8  $k \leftarrow k + 1$
- 9 **end while**



# 大残量问题的拟牛顿算法

- 大残量问题中, 海瑟矩阵的第二部分不可忽视, 此时高斯 - 牛顿法和 LM 方法可能只有线性的收敛速度

$$\nabla^2 f(x) = J(x)^\top J(x) + \sum_{i=1}^m r_i(x) \nabla^2 r_i(x)$$

- 记  $s_k = x_{k+1} - x_k, T_{k+1}$  应保留原海瑟矩阵的性质

$$\begin{aligned} T_{k+1} s_k &\approx \sum_{j=1}^m r_j(x_{k+1}) (\nabla^2 r_j(x_{k+1})) s_k \\ &\approx \sum_{j=1}^m r_j(x_{k+1}) (\nabla r_j(x_{k+1}) - \nabla r_j(x_k)) \\ &= (J_{k+1})^\top r_{k+1} - (J_k)^\top r_{k+1} \end{aligned}$$

# 大残量问题的拟牛顿算法

- 拟牛顿条件为

$$T_{k+1}s_k = (J_{k+1})^\top r_{k+1} - (J_k)^\top r_{k+1}$$

- Dennis, Gay 和 Welsch 给出的一种更新格式

$$T_{k+1} = T_k + \frac{(y^\# - T_k s_k)y^\top + y(y^\# - T_k s_k)^\top}{y^\top s_k} - \frac{(y^\# - T_k s_k)^\top s_k}{(y^\top s)^2} y y^\top$$

其中

$$s_k = x_{k+1} - x_k$$

$$y = J_{k+1}^\top r_{k+1} - J_k^\top r_k$$

$$y^\# = J_{k+1}^\top r_{k+1} - J_k^\top r_{k+1}$$

## 应用实例：相位恢复

- 相位恢复是最小二乘法的重要应用，原始模型为

$$\min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2} \sum_{j=1}^m (|\bar{a}_j^\top z|^2 - b_j)^2$$

其中  $a_j \in \mathbb{C}^n$  是已知的采样向量,  $b_j \in \mathbb{R}$  是观测的模长

- 根据 Wirtinger 导数知

$$\nabla f(\mathbf{z}) = \left[ \frac{\partial f}{\partial z}, \frac{\partial f}{\partial \bar{z}} \right]^*$$

其中

$$\frac{\partial f}{\partial z} = \sum_{j=1}^m (|\bar{a}_j^\top x|^2 - b_j) \bar{z}^\top a_j \bar{a}_j^\top, \quad \frac{\partial f}{\partial \bar{z}} = \sum_{j=1}^m (|\bar{a}_j^\top x|^2 - b_j) z^\top \bar{a}_j a_j^\top$$

## 应用实例：相位恢复

- 雅可比矩阵和高斯 – 牛顿矩阵分别为

$$J(\mathbf{z}) = \overline{\begin{bmatrix} a_1(\bar{a}_1^\top z), & a_2(\bar{a}_2^\top z), & \cdots, & a_m(\bar{a}_m^\top z) \\ \bar{a}_1(a_1^\top \bar{z}), & \bar{a}_2(a_2^\top \bar{z}), & \cdots, & \bar{a}_m(a_m^\top \bar{z}) \end{bmatrix}}^\top$$
$$\Psi(\mathbf{z})\overline{J(\mathbf{z})}^\top J(\mathbf{z}) = \sum_{j=1}^m \begin{bmatrix} |\bar{a}_j^\top z|^2 a_j \bar{a}_j^\top & (\bar{a}_j^\top z)^2 a_j a_j^\top \\ (\bar{a}_j^\top z)^2 \bar{a}_j \bar{a}_j^\top & |\bar{a}_j^\top z|^2 \bar{a}_j a_j^\top \end{bmatrix}$$

- 在第  $k$  步, 高斯 – 牛顿法求解方程

$$\Psi(\mathbf{z}^k)d^k = -\nabla f(\mathbf{z}^k)$$

## 应用实例：相位恢复

### ■ LM 方法求解正则化方程

$$(\Psi(\mathbf{z}^k) + \lambda_k)d^k = -\nabla f(\mathbf{z}^k) \quad (6)$$

### ■ 选取

$$\lambda_k = \begin{cases} 70000n\sqrt{nf(z^k)}, & f(z^k) \geq \frac{1}{900n}\|z^k\|_2^2 \\ \sqrt{f(z^k)}, & \text{其他} \end{cases}$$

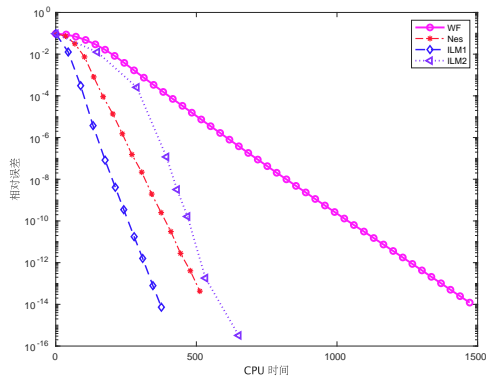
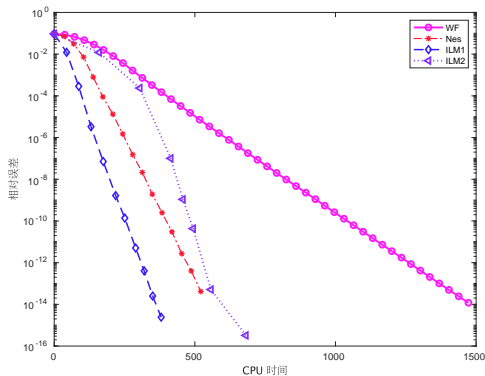
### ■ 利用共轭梯度法求解线性方程 (6), 使得

$$\|(\Psi(\mathbf{z}^k) + \lambda_k)d^k + \nabla f(\mathbf{z}^k)\| \leq \eta_k \|\nabla f(\mathbf{z}^k)\|$$

# 应用实例：相位恢复

■ WF 求解 Wirtinger 梯度下降方法

■ LM ILM1 ( $\eta_k = 0.1$ ), ILM2 ( $\eta_k = \min\{0.1, \|\nabla f(\mathbf{z}^k)\|\}$ ), Nes (Nesterov 加速)



*Q&A*

*Thank you!*

感谢您的聆听和反馈