机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

北京邮电大学

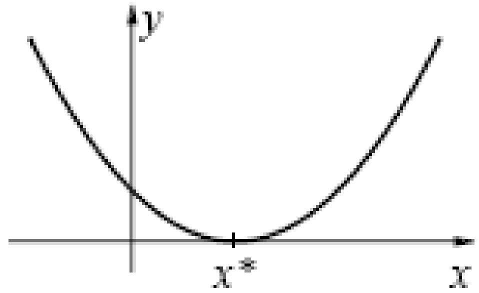
专题 二：线性模型
数学基础知识补充-III

- **内容提要**
  - 基本概念
    - 无约束优化问题 / 迭代下降法 / 最优性条件
    - 凸(convex)优化
    - 下降方向
  - 基本算法
    - 梯度下降法 / 牛顿法 / 共轭梯度法 / 最小二乘问题
  - 最优化问题建模与求解举例
    - 以寻找最快下降方向问题为例 (如何把有约束问题转换为无约束问题)

无约束最优化问题

- 问题定义：

$$\min_{\mathbf{x}} f(\mathbf{x})$$



其中 $\mathbf{x} = (x_1, x_2, ..., x_n)^T \in R^n, \quad f : R^n \to R$

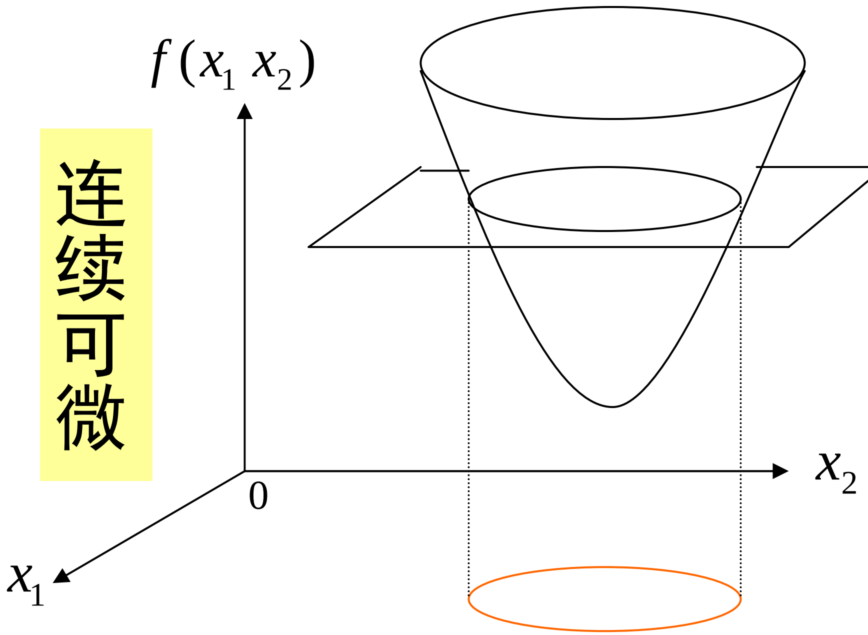
– 如果是最大化问题，则把目标函数加负号、转化为最小化问题

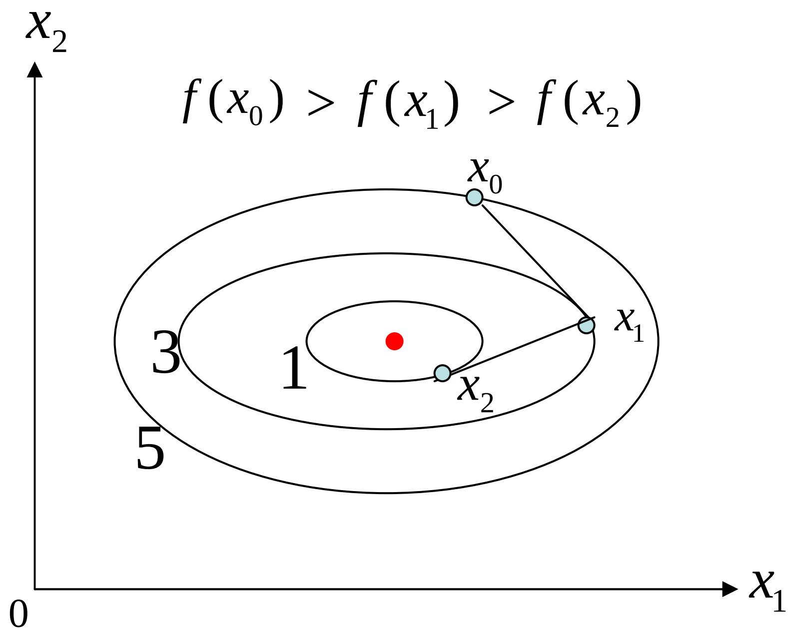
$$\max_{\mathbf{x}} f(\mathbf{x}) \Leftrightarrow \min_{\mathbf{x}} -f(\mathbf{x})$$

迭代下降算法基本思想

- 迭代下降算法：
  - 寻找一个搜索方向 $\mathbf{d}^{(k)}$ ，使得每次迭代时函数值减小，即 $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \mathbf{d}^{(k)}$ ，有 $f\left(\mathbf{x}^{(k+1)}\right) \le f\left(\mathbf{x}^{(k)}\right)$

连续可微





无约束优化问题的最优性条件

- 函数f(x)在x*处为局部最优解的一阶必要条件:

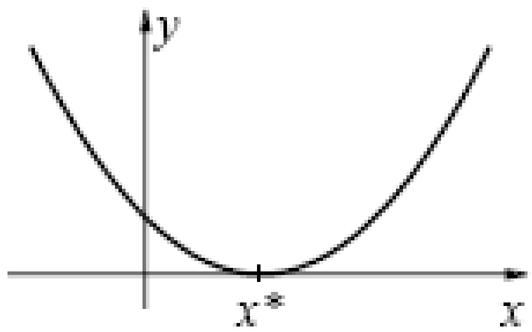
$$\nabla f\left(\mathbf{x}^*\right)=0$$

  – 如果函数**f(x)**为凸函数，则上述条件也为充分条件
  – 对于凸优化问题，则每个局部最优解也是全局最优解

- 最优化问题的本质区别不在于线性与非线性，而在于凸或非凸
  – **convex**与**non-convex** **(**最小化问题**)**
  – **concave**与**non-concave** **(**最大化问题**)**

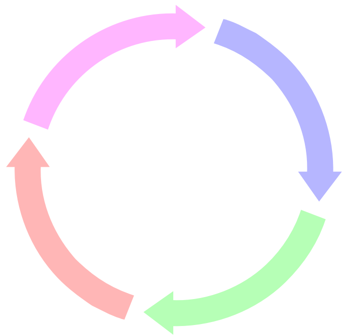
凸优化问题

- 凸函数
  - 为方便起见，最优化理论中的“最优化”一般是指最小化(minimizing)，“凸函数”是指convex函数(下凸函数)
  - 判断方法：
    - 定义法
    - 一阶条件：
    - 二阶条件：Hessian矩阵为半正定, i.e. $H = \nabla^2 f(x) \geq 0$
    - 复合函数规则
      - S. Boyd: Convex Optimization, 2004.



- 凸优化问题
  - 目标函数为凸函数
  - 可行域为凸集
    - 所有等式约束必须是线性约束
- 对于凸优化问题，其局部极小也是全局极小

迭代下降算法基本步骤



第 1 步　　选取初始点 $x^{(0)}$，$k:=0$;

第 2 步　　构造搜索方向 $d^{(k)}$；

第 3 步　　根据 $d^{(k)}$，确定步长 λ_k；

第 4 步　　令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，

若 $x^{(k+1)}$ 已满足某种终止条件，停止迭代，输出近似解 $x^{(k+1)}$；否则令 $k:=k+1$，转回第 2 步。

– 初始点、搜索方向和步长参数

梯度与下降方向

- 梯度
  - 梯度：增长最快的方向
  - 方向导数：梯度与方向的内积

- 下降方向

设 $f: R^n \mapsto R$ 在点 $\bar{x} \in R^n$ 处可微。若存在 $p \in R^n$，使 $\nabla f(\bar{x})^T p < 0$，则向量 p 是 $f(x)$ 在点 $\bar{x}$ 处的下降方向

- 最速下降方向
  - 最快下降方向：梯度的反方向

最快下降方向

- 寻找最快下降方向等价于如下非线性规划问题:

$$\min_d \nabla f(x)^T d \quad \text{s.t.} \quad d^T d = 1$$

  – 借助柯西-施瓦茨不等式:

$$-\|\nabla f(x)\|_2 \le \nabla f(x)^T d \le \|\nabla f(x)^T\|_2$$

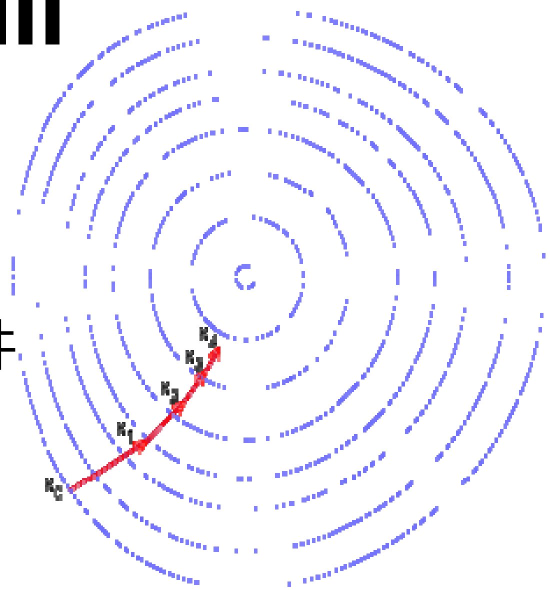
  – 第一个不等号中等式成立条件为:

$$d = -\nabla f(x) / \|\nabla f(x)\|_2$$

    - 即梯度的反方向为下降最快方向

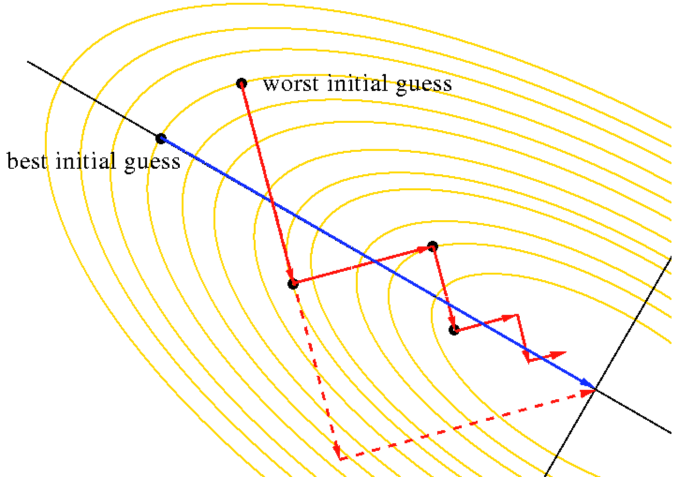
专题 二：线性模型
数学基础知识补充-III

- **内容提要**
  - 基本概念
    - 无约束优化问题 / 迭代下降法 / 最优性条件
    - 凸(convex)优化
    - 下降方向
  - 基本算法
    - 梯度下降法 / 牛顿法 / 共轭梯度法 / 最小二乘问题
  - 最优化问题建模与求解举例
    - 以寻找最快下降方向问题为例 (如何把有约束问题转换为无约束问题)



典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法(梯度下降法)
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法



- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

最速下降法(Steepest Descent)

要求：目标函数 $f(x)$ 一阶连续可微

- 由柯西(Cauchy)在1847年提出的，是求无约束极值的最早的数值算法

步骤：

第 1 步　选取初始点 $x^{(0)}$，给定终止误差 $\boldsymbol{\varepsilon > 0}$，令 $\boldsymbol{k := 0}$；

第 2 步　计算 $\nabla f\left(x^{(k)}\right)$，若 $\left\|\nabla f\left(x^{(k)}\right)\right\| < \varepsilon$，停止迭代，输出 $x^{(k)}$。

否则进行第 3 步；

第 3 步　取 $d^{(k)} = -\nabla f\left(x^{(k)}\right)$

第 4 步　进行一维搜索，求 λ_k，使得

$$f\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min_{\lambda \geq 0} f\left(x^{(k)} + \lambda d^{(k)}\right)$$

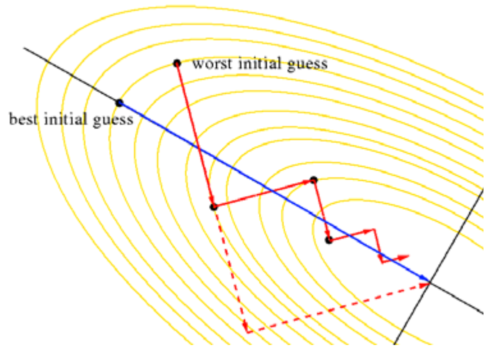
令 $x^{(x+1)} = x^{(k)} + \lambda_k d^{(k)}$，$\boldsymbol{k := k+1}$，转第 2 步。

最速下降法

- 最速下降法
  - 也称为梯度下降法(Gradient Descent)，是一种最基本的迭代下降算法
- 优点：
  - 工作量小，存储变量较少，初始点要求不高；
- 缺点：
  - 收敛慢
  - 最速下降法适用于寻优过程的前期迭代或作为间插步骤，当接近极值点时，宜选用收敛快的算法.

最速下降法的锯齿现象

- 利用最速下降法极小化目标函数时，相邻两个搜索方向是正交的：令
  $\varphi(\lambda)=f\left(x^{(k)}+\lambda d^{(k)}\right)$，$d^{(k)}=-\nabla f\left(x^{(k)}\right)$，为求出从 $x^{(k)}$ 出发沿 $d^{(k)}$ 方向的极小点，令
  $\varphi'(\lambda)=\nabla f\left(x^{(k)}+\lambda d^{(k)}\right)^T d^{(k)}=0$，由此可得：
  $-\nabla f\left(x^{(k+1)}\right)^T \cdot \nabla f\left(x^{(k)}\right)=0$，
  即方向 $d^{(k+1)}=-\nabla f\left(x^{(k+1)}\right)$ 与方向 $d^{(k)}=-\nabla f\left(x^{(k)}\right)$ 正交。
  在这说明迭代所产生的路径是“之”字形的



锯齿现象产生的几何解释

- 当Hessian矩阵 $\nabla^2 f\left(x^{(k)}\right)$ 的条件数 $r=\dfrac{\lambda_{\max}}{\lambda_{\min}}$ 很大时, 其收敛速度很慢
  - 收敛速率 $$\rho=\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2$$
  - 在极小点附近，目标函数可用二次函数近似，其等值面接近椭球面
    - 长轴和短轴对应于最小特征值与最大特征值的方向，其长短与特征值的平方根成反比
    - 最小特征值与最大特征值相差越大，椭球面越扁，一维搜索沿着“狭长谷”进行
    - 当条件数很大时，要使迭代点充分接近极小点，需要走很大弯路，因此计算效率很低

梯度下降法的另一种解释

- 对函数f(x)进行二阶近似：

$$g(x) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T \left(x - x^{(k)}\right) + \frac{\eta_k}{2}\left(x - x^{(k)}\right)^T \left(x - x^{(k)}\right)$$

  – 其中 $\eta_k = \sigma_{\max}\left(\nabla^2 f\left(x^{(k)}\right)\right)$ 为**Hessian**矩阵的最大奇异值

  – 在 $x^{(k)}$ 附近，我们有： $f(x) \le g(x)$

- 利用g(*x*)的最优解 x^* 作为 $x^{(k+1)}$ ，即得出梯度下降法的基本更新规则

$$x^{(k+1)} = x^{(k)} - \frac{1}{\eta_k}\nabla f\left(x^{(k)}\right)$$

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

牛顿法

- 设 $f(x)$ 是二次可微的实函数，$x \in R^n$，又设 $x^{(k)}$ 是 $f(x)$ 的极小点的一个估计，我们把 $f(x)$ 在 $x^{(k)}$ 展开 Taylor 级数，并取二阶近似：

$$f(x) \approx \phi(x) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T \left(x - x^{(k)}\right)$$
$$+ \frac{1}{2}\left(x - x^{(k)}\right)^T \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right)$$

其中 $\nabla^2 f\left(x^{(k)}\right)$ 是 $f(x)$ 在 $x^{(k)}$ 处的 Hessian 矩阵. 为求 $\phi(x)$ 的平稳点，令 $\nabla\phi(x) = 0$，即 $\phi(x) = \nabla f\left(x^{(k)}\right) + \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right) = 0$

设 $\nabla^2 f\left(x^{(k)}\right)$ 可逆，那么可以得到牛顿法的迭代公式：

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

其中 $\nabla^2 f\left(x^{(k)}\right)^{-1}$ 是 Hessian 矩阵的逆矩阵

算法的二次终止性

- 二次终止性：算法用于二次凸函数时，经有限次迭代必达到极小点
  - 牛顿法具有二次终止性

设二次凸函数 $f(x)=\frac{1}{2}x^T Ax+b^T x+c$，其中 A 对称正定，

利用极值条件求解：$\nabla f(x)=Ax+b=0$，得出最优解：

$$x=-A^{-1}b$$

利用牛顿法：任取初始点 $x^{(1)}\in R^n$，则

$x^{(2)}=x^{(1)}-A^{-1}\nabla f\left(x^{(1)}\right)=x^{(1)}-A^{-1}\left(Ax^{(1)}+b\right)=-A^{-1}b$，

显然，一次迭代即达到极小点

牛顿法

- 牛顿法
  - 目标函数要求二次可微
  - **Taylor**级数展开，取二阶近似
    - 确定函数的近似平稳点
  - 步骤

第 1 步 选定初始点 $x^{(0)} \in R^n$，给定允许误差 $\varepsilon > 0$，
令 k=0；

第 2 步 求 $\nabla f\left(x^{(k)}\right)$，$\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1}$，检验：若 $\nabla f\left(x^{(k)}\right) < \varepsilon$，
则停止迭代，$x^{(*)} = x^{(k)}$. 否则，转向(3)；

第 3 步 令 $d^{(k)} = -\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1} \nabla f\left(x^{(k)}\right)$（牛顿方向）；

第 4 步 $x^{(k+1)} = x^{(k)} + d^{(k)}$，$k = k+1$，转回(2)

牛顿法

- 优点：
  - 二次终止性:
    - 如果f是对称正定矩阵A的二次函数，则用牛顿法经过一次迭代就可达到最优点
  - 牛顿法的收敛速度快
    - 由于函数在极值点附近和二次函数很近似；如果目标函数不是二次函数，则牛顿法不能一步达到极值点
- 疑问：
  - 沿着牛顿方向函数值一定下降么?
    - 没有确定最优步长的步骤...

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

阻尼牛顿法

- 与原始牛顿法的区别
  - 增加沿牛顿方向的一维搜索
    - 确定最优步长
  - 因为含有一维搜索，故每次迭代目标函数一般有所下降
    - 可以证明，适当条件下，阻尼牛顿法具有全局收敛性且二级收敛

阻尼牛顿法

- 步骤：增加了沿牛顿方向的一维搜索

第 1 步 选定初始点$x^{(0)} \in R^n$，给定允许误差$\varepsilon > 0$，

令 k=0；

第 2 步 求$\nabla f\left(x^{(k)}\right)$，$\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1}$，检验：若$\nabla f\left(x^{(k)}\right) < \varepsilon$，

则停止迭代，$x^{(*)} = x^{(k)}$. 否则，转向(3)；

第 3 步 令 $d^{(k)} = -\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1} \nabla f\left(x^{(k)}\right)$（牛顿方向）；

第 4 步 进行一维搜索，求λ_k，使得

$$f\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min_{\lambda \geq 0} f\left(x^{(k)} + \lambda d^{(k)}\right),$$

令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$， $k = k+1$, 转回(2)

牛顿法

- 优点：
  - 收敛速度快
- 缺点：
  - 需要计算**Hessian**矩阵及其逆矩阵，
    - 加大了计算机计算量和存储量
  - 要求**Hessian**矩阵可逆
    - 未必可逆呢？
    - 未必正定呢？
      - 导致牛顿方向不一定为下降方向

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

修正牛顿法

- 动机：
  - 克服**Hessian**矩阵奇异性和不定性

- 方法：
  - 引入矩阵$\boldsymbol{G_k}$:

$$\boldsymbol{G_k} = \mathbf{Hessian} + \varepsilon_k \mathbf{I}$$

  - 只要 ε_k选择合适，则$\boldsymbol{G_k}$对称正定

修正牛顿法

- 记搜索方向 $d^{(k)}=x-x^{(k)}$，得到 $\nabla^2 f\left(x^{(k)}\right)d^{(k)}=-\nabla f\left(x^{(k)}\right)$

  阻尼牛顿法用的搜索方向是上述方程的解，即

  $d^{(k)}=-\nabla^2 f\left(x^{(k)}\right)^{-1}\nabla f\left(x^{(k)}\right)$，这里假设逆矩阵 $\nabla^2 f\left(x^{(k)}\right)^{-1}$ 存在。

  解决 Hessian 矩阵 $\nabla^2 f\left(x^{(k)}\right)$ 非正定问题的基本思想是：修正 $\nabla^2 f\left(x^{(k)}\right)$，构造一个对称正定矩阵 G_k，用 G_k 取代 $\nabla^2 f\left(x^{(k)}\right)$，从而得到 $G_k d^{(k)}=-\nabla f\left(x^{(k)}\right)$，解此方程，可以得到下降方向为：$d^{(k)}=-G_k^{-1}\nabla f\left(x^{(k)}\right)$，再沿此方向进行一维搜索。

  构造矩阵 G_k 的方法之一是令 $G_k=\nabla^2 f\left(x^{(k)}\right)+\varepsilon_k I$，$I$ 是 n 阶单位矩阵，ε_k 为一个适当的正数。

修正牛顿法

- 步骤
  - 增加沿牛顿方向的一维搜索，并引入矩阵$\boldsymbol{G_k}$

第 1 步 选定初始点$x^{(0)} \in R^n$，给定允许误差$\varepsilon > 0$，
令 k=0；

第 2 步 求$\nabla f\left(x^{(k)}\right)$，$\left(\nabla^2 f\left(x^{(k)}\right) + \varepsilon_k I\right)^{-1}$，检验：若$\nabla f\left(x^{(k)}\right) < \varepsilon$，则停止迭代，$x^{(*)} = x^{(k)}$. 否则，转向(3)；

第 3 步 令 $d^{(k)} = -\left(\nabla^2 f\left(x^{(k)}\right) + \varepsilon_k I\right)^{-1} \nabla f\left(x^{(k)}\right)$(修正后的牛顿方向)；

第 4 步 进行一维搜索，求λ_k，使得

$$f\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min_{\lambda \geq 0} f\left(x^{(k)} + \lambda d^{(k)}\right),$$

令$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，$k = k + 1$，转回(2)

牛顿法

- 优点：
  - 收敛快
- 缺点：
  - 要求**Hessian**矩阵要可逆
  - 需要计算二阶导数和逆矩阵，计算量和存储量开销较大

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

拟牛顿法

- 为克服牛顿法的缺点，同时保持较快收敛速度的优点，利用第 k 步和第 k+1 步得到的 $x^{(k)}$， $x^{(k+1)}$， $\nabla f\left(x^{(k)}\right)$， $\nabla f\left(x^{(k+1)}\right)$，构造一个正定矩阵 G_{k+1} 近似代替 $\nabla^2 f\left(x^{(k)}\right)$，或用 H_{k+1} 近似代替 $\nabla^2 f\left(x^{(k)}\right)^{-1}$，将牛顿方向 $d^{(k+1)}$ 改为： $G_{k+1}d^{(k+1)} = -\nabla f\left(x^{(k+1)}\right)$，或者 $d^{(k+1)} = -H_{k+1}\nabla f\left(x^{(k+1)}\right)$ 从而得到下降方向.

- 基本思想：
  - 用不包含二阶导数的矩阵近似牛顿法中的**Hessian**矩阵的逆矩阵

拟牛顿法

- 首先介绍拟牛顿条件，为了构造$\nabla^2 f\left(x^{(k)}\right)^{-1}$的近似矩阵$H_k$，先分析$\nabla^2 f\left(x^{(k)}\right)^{-1}$与一阶导数的关系。设 k 次迭代后，得到点$x^{(k+1)}$，把目标函数在点$x^{(k+1)}$展开 Taylor 级数，并取二阶近似，得到：

$$f(x) \approx f\left(x^{(k+1)}\right) + \nabla f\left(x^{(k+1)}\right)^T \left(x - x^{(k+1)}\right) + \frac{1}{2}\left(x - x^{(k+1)}\right)^T \nabla^2 f\left(x^{(k+1)}\right)\left(x - x^{(k+1)}\right)$$

可知，在点$x^{(k+1)}$附近有：

$$\nabla f(x) \approx \nabla f\left(x^{(k+1)}\right) + \nabla^2 f\left(x^{(k+1)}\right)\left(x - x^{(k+1)}\right)$$

令$x = x^{(k)}$，则

$$\nabla f\left(x^{(k)}\right) \approx \nabla f\left(x^{(k+1)}\right) + \nabla^2 f\left(x^{(k+1)}\right)\left(x^{(k)} - x^{(k+1)}\right)$$

拟牛顿条件

- 记 $p^{(k)} = x^{(k+1)} - x^{(k)}$，$q^{(k)} = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$，则有：

  $$q^{(k)} \approx \nabla^2 f\left(X^{k+1}\right) \cdot p^{(k)},$$

  设 $\nabla^2 f\left(x^{(k)}\right)$ 可逆，则有：$p^{(k)} = \nabla^2 f\left(x^{(k)}\right)^{-1} \cdot q^{(k)}$，

  为了用不包含二阶导数的矩阵 H_{k+1} 代替牛顿法中的 Hessian 矩阵的逆矩阵，有理由令 H_{k+1} 满足：$p^{(k)} = H_{k+1} \cdot q^{(k)}$，称为拟牛顿条件

秩1校正

- 怎样构造满足“拟牛顿条件”的矩阵 H_{k+1} 呢？

  当 $\nabla^2 f\left(x^{(k)}\right)^{-1}$ 为 n 阶对称正定时，满足拟牛顿条件的矩阵 H_{k+1} 也应该是 n 阶对称正定矩阵。可以利用秩 1 校正法来近似。一般策略是：H_1 取为任意 n 阶对称正定矩阵(通常选 n 阶单位矩阵 I)，通过不断修正 H_k 给出 H_{k+1}：令 $H_{k+1} = H_k + \Delta H_k$，其中 ΔH_k 为校正矩阵。

  确定 ΔH_k 的方法之一：$\Delta H_k = \alpha_k z^{(k)}\left(z^{(k)}\right)^T$，$\alpha_k$ 是常数，$z^{(k)}$ 是 n 维列向量

秩1校正

- $\Delta H_k = \alpha_k z^{(k)} \left(z^{(k)}\right)^T$ 的特点：（1）秩为 1；（2）对称；

  $z^{(k)}$ 的选择应该满足拟牛顿条件： $p^{(k)} = H_k q^{(k)} + \alpha_k z^{(k)} \left(z^{(k)}\right)^T q^{(k)}$，

  可以得出：$z^{(k)} = \dfrac{p^{(k)} - H_k q^{(k)}}{\alpha_k \left(z^{(k)}\right)^T q^{(k)}}$， $q^{(k)T}\left(p^{(k)} - H_k q^{(k)}\right) = \alpha_k \left(\left(z^{(k)}\right)^T q^{(k)}\right)^2$

  于是 $H_{k+1} = H_k + \dfrac{\left(p^{(k)} - H_k q^{(k)}\right)\left(p^{(k)} - H_k q^{(k)}\right)^T}{q^{(k)T}\left(p^{(k)} - H_k q^{(k)}\right)}$

  称为**秩 1 校正公式**

基于秩1校正的拟牛顿法

- 利用秩 1 校正极小化函数 $f(x)$，在第 k 次迭代中，令搜索方向 $d^{(k)}=-H_k\nabla f\left(x^{(k)}\right)$，然后沿 $d^{(k)}$ 方向搜索，求步长 λ_k，满足：

$$f\left(x^{(k)}+\lambda_k d^{(k)}\right)=\min_{\lambda\geq 0} f\left(x^{(k)}+\lambda d^{(k)}\right)$$

从而确定后继点为

$$x^{(k+1)}=x^{(k)}+\lambda_k d^{(k)}$$

求出点 $x^{(k+1)}$ 处梯度 $\nabla f\left(x^{(k+1)}\right)$ 以及 $p^{(k)}$ 和 $q^{(k)}$，再利用秩 1 校正公式计算 H_{k+1}，进而求出在点 $x^{(k+1)}$ 处的搜索方向 $d^{(k+1)}$，以此类推。

- 缺陷：

只有当 $q^{(k)T}\left(p^{(k)}-H_k q^{(k)}\right)>0$，才能保证 H_{k+1} 的正定性，而这一点是没有保证的

基于DFP公式的秩2校正

- DFP（Davidon-Fletcher-Powell）公式，也称作变尺度法：

  定义校正矩阵为：$\Delta H_k = \frac{p^{(k)}p^{(k)T}}{q^{(k)T}p^{(k)}} - \frac{H_k q^{(k)} q^{(k)T} H_k}{q^{(k)T} H_k q^{(k)}}$，于是

$$H_{k+1} = H_k + \frac{p^{(k)}p^{(k)T}}{q^{(k)T}p^{(k)}} - \frac{H_k q^{(k)} q^{(k)T} H_k}{q^{(k)T} H_k q^{(k)}}$$

$$G_{k+1} = G_k + \left(1 + \frac{p^{(k)T} G_k p^{(k)}}{p^{(k)T} q^{(k)}}\right) \frac{q^{(k)} q^{(k)T}}{q^{(k)T} p^{(k)}} - \frac{q^{(k)} p^{(k)T} G_k - G_k p^{(k)} q^{(k)T}}{p^{(k)T} q^{(k)}}$$

- 可以证明：DFP 方法构造的矩阵 H_{k+1} 均对称正定矩阵，因此搜索向均为下降方向，每次迭代后均使函数值有所下降

基于BFGS公式的秩2校正

- 也可以用不含二阶导数的矩阵 G_{k+1} 近似 Hessian 矩阵 $\nabla^2 f\left(x^{(k+1)}\right)$，
  我们利用另一种形式的拟牛顿条件： $q^{(k)} \approx \nabla^2 f\left(X^{k+1}\right)\cdot p^{(k)}$，则有
  $$q^{(k)} = G_{k+1}\cdot p^{(k)}$$
  关于 B_{k+1} 的修正公式为： $G_{k+1} = G_k + \dfrac{q^{(k)}q^{(k)T}}{q^{(k)T}p^{(k)}} - \dfrac{G_k p^{(k)} p^{(k)T} G_k}{p^{(k)T} G_k p^{(k)}}$

  称为 BFGS（Boryden-Fletcher-Goldfarb-Shanno）修正公式，是 DFP 修正公式的对偶形式

- 可以得到关于 H_{k+1} 的 BFGS 公式，
  $$H_{k+1}^{BFGS} = H_k + \left(1 + \frac{q^{(k)T} H_k q^{(k)}}{q^{(k)T} p^{(k)}}\right)\frac{p^{(k)}p^{(k)T}}{q^{(k)T}p^{(k)}} - \frac{p^{(k)}q^{(k)T}H_k + H_k q^{(k)} p^{(k)T}}{q^{(k)T}p^{(k)}}$$

基于DFP公式的拟牛顿法

第 1 步 选定初始点 $x^{(1)} \in R^n$，给定允许误差 $\varepsilon > 0$，令 k=0；

第 2 步 设 $H_1 = I$，计算出在 $x^{(1)}$ 处的梯度 $\nabla f\left(x^{(1)}\right)$，令 k=1；

第 3 步 令 $d^{(k)} = -H_k \nabla f\left(x^{(k)}\right)$；

第 4 步 从 $x^{(k)}$ 出发，沿着方向 $d^{(k)}$ 进行一维搜索，求 λ_k，使得

$$f\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min_{\lambda \geq 0} f\left(x^{(k)} + \lambda d^{(k)}\right),$$

令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$；

第 5 步 检查是否满足收敛准则，若 $\left\|\nabla f\left(x^{(k+1)}\right)\right\| < \varepsilon$，则停止迭代，

得到 $\bar{x} = x^{(k+1)}$；否则进行步骤(6)；

第 6 步 若 $k = n$，则令 $x^{(1)} = x^{(k+1)}$，返回步骤(2)；否则，进行步骤(7)；

第 7 步 令 $p^{(k)} = x^{(k+1)} - x^{(k)}$，$q^{(k)} = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$，计算 H_{k+1}：

$$H_{k+1} = H_k + \frac{p^{(k)} p^{(k)T}}{q^{(k)T} p^{(k)}} - \frac{H_k q^{(k)} q^{(k)T} H_k}{q^{(k)T} H_k q^{(k)}},$$

$k = k+1$ 返回步骤(3)

拟牛顿法

- 拟牛顿法是迭代下降方法中最为有效的一类算法
  - 迭代中仅需一阶导数，无需计算**Hessian**矩阵
  - 当$\mathbf{H}_k$正定时，算法产生的方向一定为下降方向
- 缺点：
  - 所需存储量较大

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

共轭方向(Conjugate Directions)

- 设 A 是 n 阶实对称正定矩阵， $\boldsymbol{p}^i \in \boldsymbol{R}^n(\boldsymbol{i}=\mathbf{0},\mathbf{1},\ldots,\boldsymbol{n}-\mathbf{1})$ 是非零向量。若 $\boldsymbol{p}^0,\boldsymbol{p}^1,\ldots,\boldsymbol{p}^{n-1}$ 是一组 A 共轭方向，则它们一定是线性无关的。

- 设 A 为 n 阶实对称正定矩阵，对于非零向量 $\boldsymbol{p},\boldsymbol{q} \in \boldsymbol{R}^n$，若有 $\boldsymbol{p}^T \boldsymbol{A}\boldsymbol{q}=\mathbf{0}$，则称 p 和 q 是相互 A 共轭的。

- 对于非零向量组 $\boldsymbol{p}^i \in \boldsymbol{R}^n, \boldsymbol{i}=\mathbf{0},\mathbf{1},\ldots,\boldsymbol{n}-\mathbf{1}$，若有 $\left(p^i\right)^T Ap^j=0$， $i,j=0,1,\ldots,n-1, i\neq j$，则称 $\boldsymbol{p}^0,\boldsymbol{p}^1,\ldots,\boldsymbol{p}^n$ 是 A 共轭方向组，也称它们为一组 A 共轭方向。

共轭梯度法

- 为什么选择共轭方向？
  - 对于二次凸函数，若沿着一组共轭方向搜索，经过有限步迭代必到达极小点
    - x^*-x_0 = 共轭方向的线性组合，可以证明线性组合系数恰好为最优步长参数
    - 根据这种性质构造具有二次终止性的算法
- 为何选择共轭梯度方向？
  - 易于计算：仅仅利用前一步的下降方向**p**和当前位置的梯度向量

$$p^{(k)} = -g^{(k)} + \beta_k p^{(k-1)}, \qquad \beta_k = \frac{g^{(k)T} A p^{(k-1)}}{p^{(k-1)T} A p^{(k-1)}}$$

- 共轭梯度法的基本思想
  - 把共轭性与最速下降方法相结合，利用已知点处的梯度构造一组共轭方向，并沿这组方向进行搜索，求出目标函数的极小点
    - 线性共轭梯度法源于求解大规模线性方程组(1950s)
    - 非线性共轭梯度法被Fletcher & Reeves(1960s)提出，为最早提出的求解大规模非线性优化问题的算法

共轭梯度方向与二次终止性

扩展子空间定理： 设函数 $f(x)=\frac{1}{2}x^T Ax+b^T x+c$，其中 A 是 n 阶对称正定阵，$d^{(1)},d^{(2)},\ldots,d^{(k)}$ 是 A 共轭的非零向量. 以任意 $x^{(1)}\in R^n$ 为初始点，依次沿 $d^{(1)},d^{(2)},\ldots,d^{(k)}$ 一维搜索，得到点 $x^{(2)},x^{(3)},\ldots,x^{(k+1)}$，则点 $x^{(k+1)}$ 是函数 $f(x)$ 在线性流形 $x^{(1)}+\beta_k$ 上的唯一极小点。特别地，当 $k=n$ 时，点是 $x^{(k+1)}$ 是函数 $f(x)$ 在 R^n 上的唯一极小点. 其中

$$\beta_k=\left\{x \mid x=\sum_{i=1}^{k}\lambda_i d^{(i)},\lambda_i\in R^1\right\}$$ 是 $d^{(1)},d^{(2)},\ldots,d^{(k)}$ 生成的子空间。

- 共轭梯度法的二次终止性
  - 牛顿法一步到达极小点
  - 共轭梯度法最多经过有限步(at most n steps)达到极小点

线性共轭梯度法

- **对于二次凸函数：FR (Fletcher-Reeves) CG法**

第 1 步　选取初始点 $x^{(1)} \in R^n$，令 $k=1$

第 2 步　计算 $\nabla f\left(x^{(k)}\right)$，若 $\left\|\nabla f\left(x^{(k)}\right)\right\| < \varepsilon$，则停止迭代，得点 $\bar{x} = x^{(k)}$；否则，进行下一步

第 3 步　构造搜索方向，令 $d^{(k)} = -\nabla f\left(x^{(k)}\right) + \beta_{k-1} d^{(k-1)}$，其中 $k=1$时，$\beta_{k-1}=0$，$k>1$时，按照 $\beta_k = \dfrac{\left\|\nabla f\left(x^{(k+1)}\right)\right\|^2}{\left\|\nabla f\left(x^{(k)}\right)\right\|^2}$，计算 β_{k-1}

第 4 步　令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，其中，步长 $\lambda_k = -\dfrac{\nabla f\left(x^{(k)}\right)^T d^{(k)}}{d^{(k)T} A d^{(k)}}$；

第 5 步　若 $k=n$，则停止计算，得点 $\bar{x} = x^{(k+1)}$；否则 $k=k+1$，返回步骤（2）

非线性共轭梯度法

- 用于一般函数的非线性FRCG法，区别在于
  - 步长需要利用一维搜索确定
  - “重置”策略
    - 把n步作为一轮，每搜索一轮结束后，取一次最速下降方向，开始下一轮

共轭梯度法

- **第 1 步**　选取初始点 $x^{(1)} \in R^n$，给定终止误差 $\boldsymbol{\varepsilon} > \mathbf{0}$，设 $y^{(1)} = x^{(1)}$，$d^{(1)} = -\nabla f\left(y^{(1)}\right), k = j = 1$；

  第 2 步　若 $\left\|\nabla f\left(y^{(1)}\right)\right\| < \varepsilon$，停止迭代；否则，进行一维搜索，求 λ_j，

  满足：$f\left(y^{(j)} + \lambda_j d^{(j)}\right) = \min\limits_{\lambda \geq 0} f\left(y^{(j)} + \lambda d^{(j)}\right)$，令

  $y^{(j+1)} = y^{(j)} + \lambda_j d^{(j)}$；

  第 3 步　如果 $j < n$，则进行步骤(4)；否则，进行步骤(5)；

  第 4 步　令 $d^{(j+1)} = -\nabla f\left(y^{(j+1)}\right) + \beta_j d^{(j)}$，其中 $\beta_j = \dfrac{\left\|\nabla f\left(y^{(j+1)}\right)\right\|^2}{\left\|\nabla f\left(y^{(j)}\right)\right\|^2}$，

  令 $j = j + 1$，转步骤(2)；

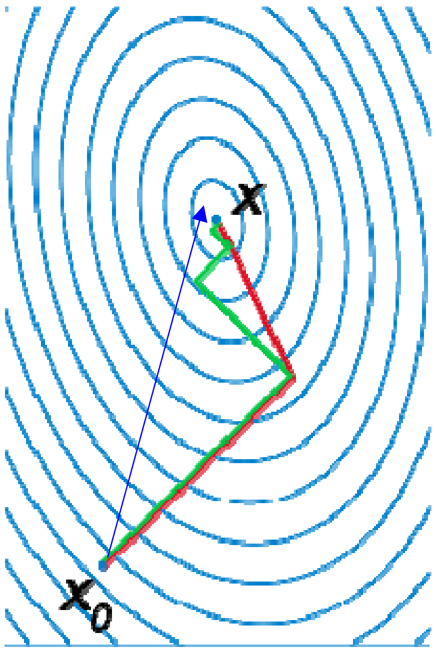
  第 5 步　令 $x^{(k+1)} = y^{(n+1)}$，$y^{(1)} = x^{(k+1)}$，$d^{(1)} = -\nabla f\left(y^{(1)}\right)$，令

  $j = 1, k = k + 1$，转步骤（2）

共轭梯度法

- 优点
  - 存储量小
    - **FRCG**法只需存储3个n维向量
      - 求解变量多的大规模问题时，可以利用共轭梯度法
  - 收敛速度快于最速下降法

    - 右图中:
      - 红色路径为共轭梯度法
      - 绿色路径为梯度下降法
      - 蓝色路径为牛顿法



典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

最小二乘问题

- 最小二乘(Least Square)问题

目标函数由若干个函数的平方和构成，$F(x)=\sum_{i=1}^{m} f_i^2(x)$，其中 $x \in R^n$。一般 $m \ge n$，把极小化这类函数的问题 $\min F(x)=\sum_{i=1}^{m} f_i^2(x)$，其中 $x \in R^n$，称为最小二乘问题。当 $f_i(x)$ 为线性函数时，为线性最小二乘问题；当 $f_i(x)$ 为非线性函数时，为非线性最小二乘问题。

线性最小二乘(Linear Least Square)

假设 $f_i(x)=\mathbf{p_i^T}x-b_i, i=1,\dots,m$，其中 $\mathbf{p_i}$ 是 n 维列向量，b_i 是实数。把问题写成矩阵形式，令 $A=\begin{pmatrix}\mathbf{p_1^T}\\ \vdots \\ \mathbf{p_m^T}\end{pmatrix}$，$\mathbf{b}=\begin{pmatrix}b_1\\ \vdots \\ b_m\end{pmatrix}$，$A$ 是 $m\times n$ 矩阵，$\mathbf{b}$ 是 m 维列向量，则

$$F(x)=\sum_{i=1}^{m}f_i^2(x)=\left(f_1(x),f_2(x),\dots,f_m(x)\right)^T\begin{pmatrix}f_1(x)\\ f_2(x)\\ \cdots \\ f_m(x)\end{pmatrix}=(Ax-\mathbf{b})^T(Ax-\mathbf{b})$$

$$=x^TA^TAx-2\mathbf{b}^TAx+\mathbf{b}^T\mathbf{b}$$

求平稳点，令 $\nabla F(x)=2A^TAx-2A^T\mathbf{b}=0$，则平稳点满足：$A^TAx=A^T\mathbf{b}$；设 A 列满秩，则 A^TA 为 n 阶对称正定矩阵，由此得到目标函数的平稳点：$x=\left(A^TA\right)^{-1}A^T\mathbf{b}$

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

非线性最小二乘

- 基本思想：

通过解一系列线性最小二乘问题来求非线性最小二乘问题的解.
设 $x^{(k)}$ 是解的第 k 次近似，在 $x^{(k)}$ 处将函数 $f_i(x)$ 线性化，把问题转化为求线性最小二乘问题，找到极小点 $x^{(k+1)}$ 后，把它作为非线性最小二乘问题的解的第 $k+1$ 次近似。再从 $x^{(k+1)}$ 出发，重复上述过程。

非线性最小二乘

- 把 $f_i(x)$ 在点 $x^{(k)}$ 展开一阶 Taylor 级数，

$$\varphi_i(x) = f_i\left(x^{(k)}\right) + \nabla f_i\left(x^{(k)}\right)^T\left(x - x^{(k)}\right)$$
$$= \nabla f_i\left(x^{(k)}\right)^T x - \left[\nabla f_i\left(x^{(k)}\right)^T x^{(k)} - f_i\left(x^{(k)}\right)\right]$$

令 $\phi(x) = \sum_{i=1}^{m} \varphi_i^2(x)$，用 $\phi(x)$ 近似 $F(x)$，从而用 $\phi(x)$ 的极小点作为目标函数 $F(x)$ 的极小点的估计。

非线性最小二乘

现在求解线性最小二乘问题：$\min \phi(x)$，记

$$A_k = \begin{pmatrix} \nabla f_1\left(x^{(k)}\right)^T \\ \vdots \\ \nabla f_m\left(x^{(k)}\right)^T \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f_1\left(x^{(k)}\right)}{\partial x_1} & \dfrac{\partial f_1\left(x^{(k)}\right)}{\partial x_2} & \cdots & \dfrac{\partial f_1\left(x^{(k)}\right)}{\partial x_n} \\ \vdots & & & \\ \dfrac{\partial f_m\left(x^{(k)}\right)}{\partial x_1} & \dfrac{\partial f_m\left(x^{(k)}\right)}{\partial x_2} & \cdots & \dfrac{\partial f_m\left(x^{(k)}\right)}{\partial x_n} \end{pmatrix}$$

$$\mathbf{b}_k = \begin{pmatrix} \nabla f_1\left(x^{(k)}\right)^T x^{(k)} - f_1\left(x^{(k)}\right) \\ \vdots \\ \nabla f_m\left(x^{(k)}\right)^T x^{(k)} - f_m\left(x^{(k)}\right) \end{pmatrix} = A_k x^{(k)} - f^{(k)}，\text{其中 } f^{(k)} = \begin{pmatrix} f_1\left(x^{(k)}\right) \\ \vdots \\ f_m\left(x^{(k)}\right) \end{pmatrix}$$

非线性最小二乘

- $\phi(x)=\sum_{i=1}^{m}\varphi_i^2(x)=(A_k x-\mathbf{b}_k)^T(A_k x-\mathbf{b}_k)$，不难得出：

  $A_k{}^T A_k x = A_k{}^T \mathbf{b}_k$， $A_k{}^T A_k x = A_k{}^T\left(A_k x^{(k)}-f^{(k)}\right)$，

  $A_k{}^T A_k\left(x-x^{(k)}\right)=-A_k{}^T f^{(k)}$

  如果 A_k 列满秩，则 $A_k{}^T A_k$ 为对称正定，逆矩阵存在，

  于是得出 $\phi(x)$ 的极小点 $x^{(k+1)}=x^{(k)}-\left(A_k{}^T A_k\right)^{-1}A_k{}^T f^{(k)}$，

  把 $x^{(k+1)}$ 作为 $F(x)$ 的极小点的第 $k+1$ 次近似。

非线性最小二乘

- 把 $x^{(k+1)} = x^{(k)} - H_k^{-1}\nabla F\left(x^{(k)}\right)$，或

  $x^{(k+1)} = x^{(k)} - \left(A_k^T A_k\right)^{-1} A_k^T f^{(k)}$ 称作 Gauss-Newton 公式，

  向量 $d^{(k)} = -\left(A_k^T A_k\right)^{-1} A_k^T f^{(k)}$ 成为在点 $x^{(k)}$ 处的

  Gauss-Newton 方向。

  为保证每次迭代都保证目标函数值下降，应该沿着 $d^{(k)}$ 方向进行一维搜索：$\min\limits_{\lambda} F\left(x^{(k)} + \lambda d^{(k)}\right)$，求得步长后，令 $x^{(k+1)} = x^{(k)} + \lambda d^{(k)}$，把 $x^{(k+1)}$ 作为第 $k+1$ 次近似。

非线性最小二乘

- 不难看出，在 $2A_k^T A_k\left(x-x^{(k)}\right)=-2A_k^T f^{(k)}$ 中，$2A_k^T f^{(k)}$ 和 $2A_k^T A_k$ 分别是 $\phi(x)$ 的梯度和 Hessian 矩阵，于是记作：$H_k\left(x-x^{(k)}\right)=-\nabla F\left(x^{(k)}\right)$，得出：$x^{(k+1)}=x^{(k)}-H_k^{-1}\nabla F\left(x^{(k)}\right)$。与牛顿迭代类似，差别在于 H_k 是逼近函数 $\phi(x)$ 的 Hessian 矩阵，而不是目标函数 $F(x)$ 的。

非线性最小二乘

- **第 1 步**　选取初始点 $x^{(1)} \in R^n$，给定终止误差 $\varepsilon > 0$，令 $k=1$；

  第 2 步　计算函数值 $f_i\left(x^{(k)}\right)$，得向量 $f^{(k)} = \begin{pmatrix} f_1\left(x^{(k)}\right) \\ \vdots \\ f_m\left(x^{(k)}\right) \end{pmatrix}$，再计算一阶偏导数 $a_{ij} = \dfrac{\partial f_i\left(x^{(k)}\right)}{\partial x_j}$ 得到 $m \times n$ 矩阵 $A = \left(a_{ij}\right)_{m\times n}$；

  第 3 步　解方程组，令 $A_k{}^T A_k d^{(k)} = -A_k{}^T f^{(k)}$，求得 Gauss-Newton 方向 $d^{(k)}$；

  第 4 步　从 $x^{(k)}$ 出发，沿 $d^{(k)}$ = 作一维搜索，求步长，使得 $F\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min\limits_{\lambda \geq 0} F\left(x^{(k)} + \lambda d^{(k)}\right)$，令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$；

  第 5 步　若 $\left\| x^{(k+1)} - x^{(k)} \right\| < \varepsilon$，则停止计算，得解 $\bar{x} = x^{(k+1)}$；否则，令 $k = k+1$，返回步骤（2）

典型的无约束优化算法

- 根据搜索方向的不同，分为:
  - 最速下降法
  - 牛顿法
    - 阻尼牛顿法
    - 修正牛顿法
    - 伪(Pseudo)牛顿法
  - 共轭梯度法
- 最小二乘问题
  - 线性最小二乘
  - 非线性最小二乘
  - 修正最小二乘

非线性最小二乘

- 算法的修正
  – Levenberg-Marquardt方法

有时 $A_k^T A_k$ 会出现奇异或接近奇异的情形，此时可以作修正。基本技巧是把一个正定对角矩阵加到 $A_k^T A_k$ 上去，改变原矩阵的特征值结构，使其变成条件数较好的对称正定矩阵，得到行之有效的修正最小二乘法。

修正方法之一，Marquardt 方法：$d^{(k)} = -\left(A_k^T A_k + \alpha_k I\right)^{-1} A_k^T f^{(k)}$，

其中 I 为单位矩阵，α_k 是一个正实数。

非线性最小二乘

- 第 1 步　选取初始点 $x^{(1)} \in R^n$，设定初始参数 $\alpha_1 > 0$，增长因子 $\beta > 1$，允许误差 $\varepsilon > 0$，计算 $F\left(x^{(1)}\right)$，令 $\alpha = \alpha_1$，$k = 1$；

  第 2 步　令 $\alpha = \alpha / \beta$，计算函数值 $f_i\left(x^{(k)}\right)$，得向量

  $f^{(k)} = \left(f_1\left(x^{(k)}\right) \cdots f_m\left(x^{(k)}\right)\right)^T$，再计算一阶偏导数

  $a_{ij} = \dfrac{\partial f_i\left(x^{(k)}\right)}{\partial x_j}$ 得到 $m \times n$ 矩阵 $A = \left(a_{ij}\right)_{m \times n}$；

  第 3 步　解方程组，令 $\left(A_k^T A_k + \alpha I\right) d^{(k)} = -A_k^T f^{(k)}$，求得方向 $d^{(k)}$，令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$；

  第 4 步　计算 $F\left(x^{(k+1)}\right)$，若 $F\left(x^{(k+1)}\right) < F\left(x^{(k)}\right)$ 则转步骤（6）；否则进行步骤(5)；

  第 5 步　若 $\left\| A_k^T f^{(k)} \right\| < \varepsilon$，则停止计算，得解 $\overline{x} = x^{(k+1)}$；否则，令 $\alpha = \beta\alpha$，转步骤（3）；

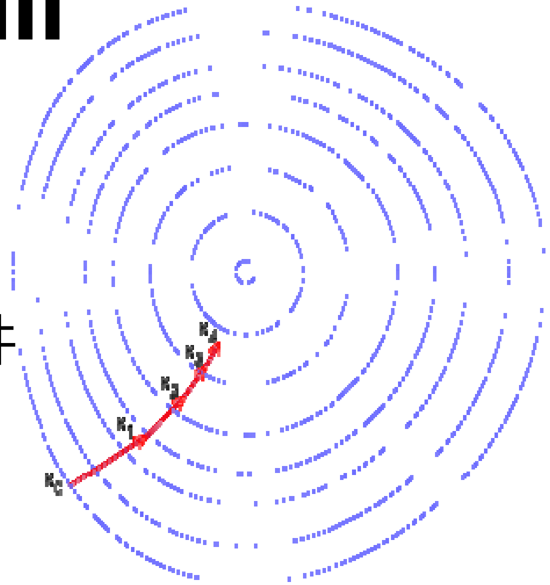
  第 6 步　若 $\left\| A_k^T f^{(k)} \right\| < \varepsilon$，则停止计算，得解 $\overline{x} = x^{(k+1)}$；否则，令 $k = k+1$，转步骤（2）；

Q / A

- Any Question? …

专题 二：线性模型
数学基础知识补充-III

- **内容提要**
  - 基本概念
    - 无约束优化问题 / 迭代下降法 / 最优性条件
    - 凸(convex)优化
    - 下降方向
  - 基本算法
    - 梯度下降法 / 牛顿法 / 共轭梯度法 / 最小二乘问题
  - 最优化问题建模与求解举例
    - 以寻找最快下降方向问题为例 (如何把有约束问题转换为无约束问题)



问题建模：寻找最快下降方向

- 从下降方向的定义出发，选定目标函数：

$$\min_{\mathbf{d}} \nabla f(\mathbf{x})^T \mathbf{d}$$

  – 寻找约束条件：

$$\mathbf{d}^T \mathbf{d} = 1$$

- 构造有约束最优化问题：

$$\min_{\mathbf{d}} \nabla f(\mathbf{x})^T \mathbf{d} \quad \text{s.t.} \quad \mathbf{d}^T \mathbf{d} = 1$$

问题求解

- 寻找最快下降方向即如下非线性规划问题:

$$\min_{d} \nabla f(x)^T d \quad \text{s.t.} \quad d^T d = 1$$

- 如何求解？
  - 方法1：借助不等式
  - 方法2：拉格朗日乘子法a
  - 方法3：拉格朗日乘子法b

方法-1

- 寻找最快下降方向等价于如下非线性规划问题:

$$\min_{d} \nabla f(x)^T d \quad \text{s.t.} \quad d^T d = 1$$

  - 借助柯西-施瓦茨不等式:

$$-\|\nabla f(x)\|_2 \le \nabla f(x)^T d \le \|\nabla f(x)^T\|_2$$

  - 第一个不等号中等式成立条件为:

$$d = -\nabla f(x) / \|\nabla f(x)\|_2$$

    - 即梯度的反方向为下降最快方向

方法-2

- 寻找最快下降方向等价于如下非线性规划问题:

$$\min_{d} \ \nabla f(x)^T d \quad \text{s.t.} \ d^T d = 1$$

  – 非凸优化问题，但该问题最优解唯一

- 应用拉格朗日乘子法：

$$L(d,\lambda) = \nabla f(x)^T d + \lambda\left(d^T d - 1\right),$$

  – 原问题**(Primal Problem)**最优性条件为：

$$\nabla_d L(d,\lambda) = \nabla f(x) + 2\lambda d = 0 \ \Rightarrow \ d^* = -\nabla f(x)/2\lambda$$

  – 带入原问题的可行性条件： $\Rightarrow$ $\lambda^* = \left\|\nabla f(x)\right\|_2 / 2$

- 结论：梯度反方向为下降最快方向

方法-3

- 把等式约束放松为不等式约束: $\min_{d} \nabla f(x)^T d \quad \text{s.t.} \quad d^T d \le 1$
  - 得到一个不等式约束的凸优化问题
    - 目标函数为线性，可行域为凸集
- 拉格朗日乘子法：$L(d,\lambda) = \nabla f(x)^T d + \lambda\left(d^T d - 1\right), \quad \lambda \ge 0$
  - 原问题(**Primal Problem**)最优性条件为：

$$\nabla_d L(d,\lambda) = \nabla f(x) + 2\lambda d = 0 \Rightarrow d^* = -\nabla f(x)/2\lambda$$

  - 带入到辅助函数中，得到对偶问题：

$$\max_{\lambda} D(\lambda) = -\frac{1}{4\lambda}\left\|\nabla f(x)\right\|_2^2 - \lambda \Rightarrow \lambda^* = \left\|\nabla f(x)\right\|_2 / 2$$

    - 带入目标函数可得：梯度反方向为下降最快方向
  - 目标函数为线性函数，可以知最优解在可行域边界上，因此放松后最优化问题的最优解与原问题最优解一致

讨论：3种方法的比较

- 方法1：
  - 考虑了目标函数的特殊形式，借助了特定不等式
- 方法2：
  - 拉格朗日乘子法解决等式约束的最优化问题
  - 具有通用性，同时又考虑了目标函数和约束条件的特殊形式，即问题最优解的唯一性
- 方法3：
  - 拉格朗日乘子法解决不等式约束的最优化问题
  - 把非凸优化问题放松为凸优化问题，具有通用性，是处理非凸优化问题的一种常用策略

Q / A

- Any Question? …

参考资料

- 陈宝林，最优化理论与算法(第二版)，清华大学出版社，2005年10月.
  - **Stephen Boyd and Liewen Vandenberghe，Convex Optimization, Cambridge Univ. Press, 2004.**
  - **Jorge Nocedal and Stephen J. Weight, Numerical Optimization, Springer-Verlag, 1999.[影印版，科学出版社2006]**