人工智能数学基础

- 课程引入
  - 人工智能
  - 机器学习
    - 分类/回归
    - 有监督/半监督/无监督
      - 线性回归（LR）
      - 神经网络（NN）(或感知机网络)
    - 深度学习
      - DNN
      - CNN（LeNet/AlexNet/ResNet/VGG/InceptionNet/...）
    - 相关概念
      - 训练/测试（/验证）
      - 损失函数
      - 目标函数
      - 过拟合
      - 最优化
- 矩阵分析
  - 引入
  - 机器学习常用量
    - 向量
      - 定义
      - 向量的模
      - 向量运算：加法/减法/数乘/方向角（方向余弦）/向量投影/点乘/向量的长度/向量外积
      - 应用
    - 矩阵
      - 线性方程组
      - 定义
      - 特殊矩阵
      - 同型矩阵
      - 相等矩阵

- 张量
  - 定义与表示
- 范数
  - 定义
  - 向量范数
    - 向量范数

切比雪夫范数 —— 1. 向量的∞ 范数(最大范数): $||x||_\infty = \max_{1\le i\le n}|x_i|$,

稀疏规则算子，也称曼哈顿范数 —— 2. 向量的1-范数: $||x||_1 = \sum_{i=1}^{n}|x_i|$,

欧几里得范数 —— 3. 向量的2-范数（Euclid范数）: $||x||_2 = \left(\sum_{i=1}^{n}|x_i|^2\right)^{\frac{1}{2}}$,

广义范数 —— 4. 向量的p-范数: $||x||_p = \left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}$。

  - 矩阵范数

几种常用的矩阵范数（$A \in R^{n\times n}$）

1. A 的Frobenius范数： $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\mathrm{tr}(A^TA)}$
2. A 的行范数: $||A||_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$,
3. A 的列范数: $||A||_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|$,
4. A 的2-范数（谱范数）: $||A||_2 = \sqrt{\lambda_{max}(A^TA)}$

$\lambda_{max}(A^TA)$表示最大特征值

  - 范数与目标
- 矩阵运算
  - 矩阵加减法
  - 矩阵数乘
  - 矩阵与向量/矩阵的乘法/应用
  - 矩阵转置
  - 矩阵求逆
  - 伴随矩阵

设矩阵 $A=\begin{pmatrix} a_{11} & a_{12} & \mathbf{L} & a_{1n} \\ a_{21} & a_{22} & \mathbf{L} & a_{2n} \\ \mathbf{M} & \mathbf{M} & & \mathbf{M} \\ a_{n1} & a_{n2} & \mathbf{L} & a_{nn} \end{pmatrix}$ 中元素 a_{ij} 的代数余子式 A_{ij}，

$$A_{ij}=(-1)^{i+j}\det(\mathbf{A}_{ij}).$$

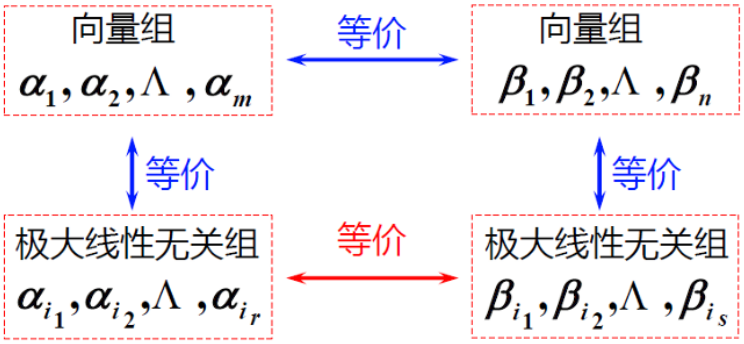
$$A^{*}=\begin{pmatrix} A_{11} & A_{21} & \mathbf{L} & A_{n1} \\ A_{12} & A_{22} & \mathbf{L} & A_{n2} \\ \mathbf{M} & \mathbf{M} & & \mathbf{M} \\ A_{1n} & A_{2n} & \mathbf{L} & A_{nn} \end{pmatrix}$$ 称为A的伴随矩阵。

- 利用伴随矩阵求矩阵的逆矩阵

定理：矩阵A 可逆充分必要条件是$|A|\neq 0$,

且当$|A|\neq 0$ 时 $A^{-1}=\frac{1}{|A|}A^{*}$.

- 线性相关与线性无关
  - 定义
    - **向量组的线性相关性：**对于n维向量组$a_1, a_2, \ldots, a_n$，如果存在不全为零的实数$k_1, k_2, \ldots, k_n$使得
      $$k_1a_1 + k_2a_2 + \ldots + k_na_n = 0,$$
      则称n维向量组$a_1, a_2, \ldots, a_n$**线性相关**，否则称该向量组**线性无关**。
  - 极大线性无关组
    - 秩
    - 极大线性无关组（不唯一）
    - 向量组等价



    - 向量组之间的线性表示：A=B*C
  - 向量组的秩
    - 定义
    - 向量组的秩与矩阵的秩之间的关系

    - 向量组的秩及其极大线性无关组求解方法
        - 1. 将向量组按列向量排列，构成矩阵 A
        - 2. 对矩阵 A 进行初等行变换，形成行标准型矩阵 B
        - 3. 矩阵 B 的秩及其列向量之间的关系与矩阵 A 的秩及其列向量之间的关系一致
    - 应用
        - 主成分分析
        - 特征脸
        - 线性回归
- 矩阵的秩
    - 定义
        - K 阶子行列式（K 阶子式）
        - 矩阵的秩
        - 满秩矩阵
        - 不可逆矩阵（降秩矩阵）
    - 初等变换求矩阵的秩
    - 矩阵秩的性质
        - 0 ≤ R(A) ≤ min{m,n}
        - R(A)=R(A^T)
        - 若 A~B，则 R(A)=R(B)
        - 若 P, Q 可逆，则 R(PAQ)=R(PA)=R(AQ)=R(A)
        - max{R(A),R(B)} ≤ R(A,B) ≤ R(A) + R(B)
        - R(A+B) ≤ R(A) + R(B)
        - R(AB) ≤ min{R(A),R(B)}
    - 应用
        - 低秩矩阵分解
        - 交叉验证
        - 矩阵填充
- 矩阵的特征值和特征向量
    - 向量的内积和正交化
        - 向量内积定义及性质
        - 向量正交及正交向量组定义

- 定理：正交向量组线性无关

- 标准正交基

- 施密特正交化准则

  施密特正交化过程：设$\alpha_1, \alpha_2, \ldots, \alpha_r$是线性无关向量组，通过递归方法逐个构造正交向量，每次构造新向量时，减去其在已构造正交向量上的投影，以消除相关性。

  ➢ **正交化：**取 $\beta_1 = \alpha_1$

  $$\beta_2 = \alpha_2 - \frac{(\beta_1, \alpha_2)}{(\beta_1, \beta_1)}\beta_1,$$

  $$\beta_3 = \alpha_3 - \frac{(\beta_1, \alpha_3)}{(\beta_1, \beta_1)}\beta_1 - \frac{(\beta_2, \alpha_3)}{(\beta_2, \beta_2)}\beta_2$$

  $$\beta_r = \alpha_r - \frac{(\beta_1, \alpha_r)}{(\beta_1, \beta_1)}\beta_1 - \frac{(\beta_2, \alpha_r)}{(\beta_2, \beta_2)}\beta_2 - \Lambda - \frac{(\beta_{r-1}, \alpha_r)}{(\beta_{r-1}, \beta_{r-1})}\beta_{r-1}$$

  ➢ **单位化：**取 $e_1 = \frac{\beta_1}{|\beta_1|}$，$e_2 = \frac{\beta_2}{|\beta_2|}$，$\Lambda\Lambda$ ，$e_r = \frac{\beta_r}{|\beta_r|}$，

- 正交矩阵（A A^T = E，即 A^-1=A^T=A*）

- 矩阵的特征值与特征向量

  - 定义及性质：$Ax = \lambda x$，λ为特征值，x为其对应的特征向量

  - 求特征值特征向量

    - 特征方程|λE-A|=0 的根为特征值λ

    - Ax =λx 即(λE-A)x=0, 求该齐次线性方程组的一个基础解系为特征向量

- 相似矩阵

  - 定义

    定义9：设A与B是n 阶方阵，若存在一个可逆矩阵P，使得：

    $$B = P^{-1}AP$$

    则称B与A相似。

  - 性质

  - 相似矩阵的特征值相同

    - 推论 1：n 阶方阵 A 与对角矩阵相似，则 A 的特征值为对角矩阵对角线值

    - 推论 2：对 n 阶方阵 A，若存在可逆矩阵 P，使得P^(−1)AP =∧ 为对角矩阵，则方阵 A 可对角化.

    - n阶方阵A可对角化的充分必要条件是A有n个线性无关的特征向量.

- 实对称矩阵的对角化

  - 定理

    - 定理：实对称矩阵的特征值为实数

- 定理：设λ是n阶实对称矩阵A的k重特征值，则 A 的属于特征值λ的线性无关的特征向量个数恰为k。即$R(\lambda E-A)=n-k$
- 设"λ" _1, "λ" _2 是对称矩阵A的两个特征值，p_1, p_2 是对应的特征向量，若"λ" _1≠"λ" _2，则p_1 与 p_2 正交。
- 设A为n阶对称矩阵,则必有正交矩阵P,使P^(−1)AP =∧

- （矩阵的特征分解）求正交矩阵P使$\boldsymbol{P}$^(−**1**)AP 为对角矩阵
  - 求出A的所有不同的特征值$\lambda_1, \lambda_2, \cdots \lambda_s$.
  - 求出A对应每个特征值λ_i的线性无关的特征向量.
  - 利用施密特标准化将对应的特征向量进行单位正交化处理
  - 以求出的n个两两正交的单位特征向量作为列向量，所得的n 阶方阵即为所求的正交矩阵P

- 矩阵分解
  - 矩阵分解的含义
  - 满秩分解
    - 定义：将一个非零矩阵（长方形）分解成一个列满秩矩阵与一个行满秩矩阵的乘积问题.
    - 基于 Hermite 标准形 H 的满秩分解
      - Hermite 标准形 H 的定义
      - 若 A 的 Hermite 标准形 H，则取 A 的k_1,k_2 ", "...，k_r 列构成矩阵 B，取 H 的前 r 行构成矩阵 C，则 A=BC 即为矩阵 A 的满秩分解。
        - **满秩分解定理：**设A∈ $C_r^{m\times n}$ $(\boldsymbol{r>0})$，且A的Hermite 标准形H为

$$H=\begin{pmatrix} 0 & \Lambda & 0 & \overset{k_1}{1} & * & \Lambda & * & \overset{k_2}{0} & * & \Lambda & * & \overset{k_r}{0} & * & \Lambda & * \\ 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 1 & * & \Lambda & * & 0 & * & \Lambda & * \\ M & M & M & M & M & M & M & M & M & M & M & M & M & \Lambda & M \\ 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 0 & * & \Lambda & 0 & 1 & * & \Lambda & * \\ 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 \\ \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda & \Lambda \\ 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 & 0 & 0 & \Lambda & 0 \end{pmatrix} \text{第}r\text{行}$$

        则取**A**的$k_1, k_2, \ldots, k_r$列构成矩阵B，取**H**的前r 行构成矩阵C，则**A=BC即为矩阵A的满秩分解**。
      - 矩阵 A 的满秩分解是不唯一的。
  - QR 分解（正三角分解/酉三角分解）
    - 定义：设 A∈C^(m×r)，如果存在 n 阶酉矩阵 Q 和 n 阶上三角矩阵 R，使 A=QR，则称之为 A 的 QR 分解或酉三角分解。当 A∈R^(n×n)时，则称为 A 的正三角分解。
    - QR 分解定理：任意一个满秩实(复）矩阵 A，都可唯一地分解 A = QR，其中 Q 为正交（酉）矩阵，R 是具有正对角元的上三角矩阵。

- QR 分解方法
  - 1. 将 A 矩阵各列进行单位正交化表示，得到矩阵 Q
  - 2. 将矩阵 A 的各列向量表示成 Q 列向量的线性组合，得到稀疏上三角矩阵 R。(A=QR, Q^(-1)=Q^T,-->R=Q^T A)
  - 3. A=QR

- SVD 分解
  - 设 A∈C_r^(m×n), λ_i 是（格拉姆）矩阵 A^H A 的特征值， μ_i 是（格拉姆）矩阵 AA^H 的特征值，它们都是实数。那么 λ_i=μ_i>0 (i=1,2,...,r)，即 A^H A 与 AA^H 的非零特征值相等.
  - 奇异值定义：α_i=√(λ_i)= √(μ_i)>0 (i=1,2,...,r)为矩阵 A 的正奇异值，简称奇异值。
  - 奇异值分解定理

    定理3 (奇异值分解定理)： 设A∈ $C_r^{m\times n}$, $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r$是A的r个奇异值，那么存在m阶酉矩阵U和n阶酉矩阵V使得

    $$A = U\begin{pmatrix} \Delta & O \\ O & O \end{pmatrix} V^H$$

    其中 $\Delta = \begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \mathrm{O} & \\ & & & \alpha_r \end{bmatrix}$，且满足$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r$.

    称表达式 $A = U\begin{pmatrix} \Delta & O \\ O & O \end{pmatrix} V^H$ 为**矩阵A 的奇异值分解式**。

  - 奇异值分解步骤（方法 1）

    - **SVD分解步骤：**

      ① 求出AA^H或A^HA的全部特征值$\boldsymbol{\lambda}_i$，则$\alpha_i = \sqrt{\boldsymbol{\lambda}_i}$为A 的正奇异值，求得$\boldsymbol{\Sigma}$

      ② 求酉矩阵U ∈ $U^{m\times m}$,(U的列向量为AA^H的单位化正交特征向量)，使得:

      $$U^H AA^H U = diag[\alpha_1^2, \alpha_2^2, \dots, \alpha_r^2, 0, \dots, 0] = diag[\lambda_1, \lambda_2, \dots, \lambda_r, 0, \dots, 0]$$

      ③ 设$\boldsymbol{U} = [\boldsymbol{U_1}, \boldsymbol{U_2}]$，$V_1 = A^H U_1 \Delta^{-H}$，则$\boldsymbol{V_1}$为次酉阵，于是求$V_2 = \mathrm{U}_{n-r}^{n\times(n-r)}$，使得$\boldsymbol{V} = [\boldsymbol{V_1}, \boldsymbol{V_2}]$，$\boldsymbol{V} \in \boldsymbol{U^{n\times n}}$

      ④ $A = U\begin{pmatrix} \Delta & O \\ O & O \end{pmatrix} V^H$

    - 上述步骤是通过Gram矩阵AA^H的特征值确定奇异值矩阵Σ，并通过其特征向量单位正交化后确定酉（正交）矩阵$U = [U_1, U_2]$，然后通过$V_1 = A^H U_1 \Delta^{-H}$求次酉阵$V_1$，进而通过$V_1$求与其互补的次酉阵$V_2$确定酉（正交）矩阵$V = [V_1, V_2]$，则$A = U\Sigma V^H$。
    - 也可以先通过A^HA求Σ 和$V = [V_1, V_2]$，再通过$U_1 = AV_1\Delta^{-1}$确定次酉阵U_1,进而通过U_1求与其互补的次酉阵U_2确定$U = [U_1, U_2]$，则$A = U\Sigma V^H$。

  - 奇异值分解步骤（方法 2）

- **SVD分解步骤（方法2）：**

  ① 求出AA^H或A^HA的全部特征值λ_i,则$\alpha_i=\sqrt{\lambda_i}$为A 的正奇异值，求得Σ

  ② 求酉矩阵$U\in U^{m\times m}$，U的列向量为AA^H的单位正交特征向量，使得:
  $UAA^HU^H=diag[\alpha_1^2,\alpha_2^2,\dots,\alpha_r^2,0,\dots,0]=diag[\lambda_1,\lambda_2,\dots,\lambda_r,0,\dots,0]_{m\times m}$
  其中，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r>0$

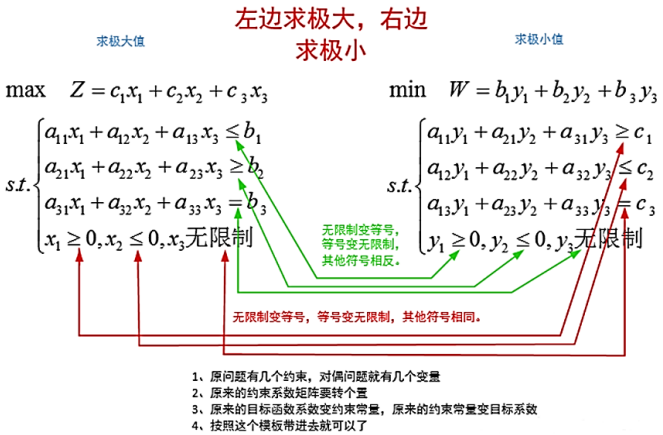
  ③ 同理，求酉矩阵$V\in U^{n\times n}$，V的列向量为A^HA的单位正交特征向量，使得:
  $VA^HAV^H=diag[\alpha_1^2,\alpha_2^2,\dots,\alpha_r^2,0,\dots,0]=diag[\lambda_1,\lambda_2,\dots,\lambda_r,0,\dots,0]_{n\times n}$
  其中，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r>0$

  ④ $A=U\Sigma V^H$

- 最优化
  - 最优化问题概述
    - 优化的定义
    - 实例
      - 食谱问题
      - 资金使用问题
    - 最优化数学定义
      - 定义：目标函数+约束条件
      - 最优化问题的求解步骤
      - 最优化问题分类
      - 最优化应用场景
  - 最优化数学基础
    - 方向导数（梯度向量与方向余弦的点乘）
    - 梯度（偏导向量）
    - 黑塞矩阵（Hessian Matrix）
      - 定义：二阶偏导矩阵
      - 物理含义：空间局部曲率
    - 凸集
      - 线性组合：$ax\_1+(1-a)x\_2$
      - 定义：设集合 S⊂$R\^n$，如果$x_1,x_2\in S$，有$ax_1+(1-a)x_2\in S$，$\forall a\in[0,1]$，则称 S 为凸集。
      - 常见凸集及其证明
    - 凸函数
      - 定义
      - 几何意义：一元凸函数表示连接函数图形上任意两点之间的连线总是位于曲线弧的上方。

- 证明函数为凸函数
- 判定条件
  - 一阶条件：设D是开凸集， $f(x)$在D上具有一阶连续导数，则$f(x)$是D上的凸函数的充要条件是：对D上的任意两个不同点x,y，f(y)≥f(x)+f'(x)(y-x)
  - 二阶条件：设D是开凸集， $f(x)$在D上具有二次可微，则$f(x)$是D上的凸函数的充要条件是： $f(x)$在D上的 Hessian 矩阵 D^2 $f(\boldsymbol{x})$是半正定的。
- 凸优化
  - 定义：设约束集 D∈$\boldsymbol{R}$^$\boldsymbol{n}$是凸集，目标函数$\boldsymbol{f}$是定义在 D 上的凸函数，则称此类规划 min$f(x)$ ($x\in D$)为凸优化问题或凸规划。
  - 在凸优化问题中，局部极小点就是全局极小点
  - 凸优化问题求解的意义

- 标准形式和对偶形式（05.26）
  - 对偶问题引入：线性规划的对偶问题
  - 原问题与对偶问题
    - 不等式约束的对偶问题
    - 等式/无条件约束的对偶问题
    - 一般线性规划问题的对偶问题



- 符号变换（变量符号与约束条件符号）：“大同小异”

  - 对偶问题的性质
    - 弱对偶定理
    - 强对偶定理
    - 支持向量机与拉格朗日对偶问题

- 连续优化
  - 最小二乘问题（LSP）

- LSP 定义及基本原理
  - n 个点与一条直线的接近程度，可以通过这组点的 y 值与直线上的函数值之间的二范数进行衡量
  - 线性拟合
- LSP 求解线性回归方程
  - 线性回归方程 y=a+bx，利用 LSP 确定 a、b 系数

    则可以求得 $b=$

    $$\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}$$

    $$=\frac{x_1y_1+x_2y_2+\cdots+x_ny_n-n\overline{x}\,\overline{y}}{x_1^2+x_2^2+\cdots+x_n^2-n\overline{x}^2}. \qquad a=\overline{y}-b\overline{x}$$

  -
    - **线性回归方程的求解步骤：**
      - **Step 1:** 列表x_i, y_i, x_iy_i;
      - **Step 2:** 计算$\bar{x}, \bar{y}, \sum_{i=1}^n x_i^2, \sum_{i=1}^n x_iy_i$;
      - **Step 3:** 代入公式，计算b和a;

        $$b=\frac{x_1y_1+x_2y_2+\cdots+x_ny_n-n\overline{x}\,\overline{y}}{x_1^2+x_2^2+\cdots+x_n^2-n\overline{x}^2} \qquad a=\bar{y}-b\bar{x}$$

      - **Step 4:** 写出线性回归方程$y=a+bx$;
    - **利用回归直线对总体进行估计**
      - 利用回归直线进行预测：若回归方程为$y=a+bx$，则在$x=x_0$处的估计值为$y=a+bx_0$

- 一阶优化
  - 梯度下降法
    - 定义：**梯度下降法**又称**最速下降法**，函数 J(a)在某点 x_k 的梯度∇J(a_k)是一个向量，其方向为 J(a) 在该点增长最快的方向。显然负梯度方向为 J(a) 减少最快的方向
    - 搜索方向：求函数 J(a)极小值的问题，可以选择任意初始点 a_0，从 a_0 出发沿着负梯度方向走，可使得 J(a)下降最快。
      - 对于任意点a_k，可以定义a_k点的负梯度搜索方向的单位向量为：

        $$\hat{s}^{(k)}=-\frac{\nabla J(a_k)}{\|\nabla J(a_k)\|}$$

      - 从a_k点出发，沿着$\hat{s}^{(k)}$方向走一步，步长为ρ_k，得到的新点a_{k+1}表示为：

        $$a_{k+1}=a_k+\rho_k\hat{s}^{(k)}$$

        因此，在新点a_{k+1}，函数$J(a)$的函数值为：

        $$J(a_{k+1})=J(a_k+\rho_k\hat{s}^{(k)})$$

    - 梯度算法步骤（关键问题：如何设计步长？）

梯度算法步骤：

① 给定初始点 $x^1 \in R^n$，允许误差$\varepsilon > 0$，令$k = 1$；

② 计算搜索方向d^k =-$\nabla f(x^k)$;

③ 若$\|d^k\| \le \varepsilon$，则停止计算，x^k为所求极值点，否则，求最优步长λ_k，使得$f(x^k + \lambda_k d^k) = \min_{\lambda} f(x^k + \lambda d^k)$；

④ 令$x^{k+1} = x^k + \lambda_k d^k$，令$k = k + 1$，转向步骤2.

- 确定最优步长

一维搜索

- 采用数学规划法求函数极值点的迭代计算：

$$x^{k+1} = x^k + a_k d^k$$

K+1次迭代的搜索方向

搜索的最佳步长因子

- 当搜索方向d^k给定，求最佳步长a_k就是求一元函数的极值

$$f(x^{k+1}) = f(x^k + a_k d^k) = \varphi(a_k)$$

- 此方法称为**一维搜索**，是优化搜索方法的基础。

- 求解一元方程$\varphi(a)$的极小点a^*可用解析法。

一维搜索

$$f(x+ad) \approx f(x) + ad^T \nabla f(x) + \frac{1}{2}(ad)^T G(ad)$$ 多维泰勒展开

$$= f(x) + \alpha d^T \nabla f(x) + \frac{1}{2}\alpha^2 d^T G d$$

- 上式求a的极值，即求a导数为零。

$$d^T \nabla f(x) + \alpha^* d^T G d = 0$$

则 $$\alpha^* = -\frac{d^T \nabla f(x)}{d^T G d}$$

- 问题求解：给定目标函数及其初始点，计算迭代 k 步后的点是否为最优解（二阶判定法）
- 梯度下降法的收敛性：

设 $f(x)$ 有一阶连续偏导数，若步长 λ_k 满足

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

则有 $\nabla f(x^k + \lambda_k d^k)^T d^k = 0$。

- 第 k+1 步的搜索方向与第 k 步的搜索方向垂直

注：因为梯度法的搜索方向 $d^{k+1} = -\nabla f(x^k + \lambda_k d^k)$，所以

$(d^{k+1})^T d^k = 0 \Rightarrow d^{k+1} \perp d^k$。

- 梯度下降的优缺点
  - 梯度下降算法**并不能保证被优化函数达到全局最优解**
  - **计算时间太长**
  - 随机梯度下降+批梯度下降

- 共轭梯度法
  - **共轭方向**定义

共轭方向

- **定义**

设 A 是 $n\times n$ 的对称正定矩阵，对于 R^n 中的两个非零向量 d^1 和 d^2，

若有 $d^{1T}Ad^2=0$，则称 d^1 和 d^2 关于 A 共轭。

设 d^1, d^2, Λ, d^k 是 R^n 中一组非零向量，如果它们两两关于 A 共轭，即 $d^{iT}Ad^j=0, i\neq j, i,j=1,2,\Lambda,k$。

则称这组方向是关于 A 共轭的，也称它们是一组 A 共轭方向。

注：如果 A 是单位矩阵，则

$$d^{1T}\cdot I\cdot d^2=0\Rightarrow d^{1T}\cdot d^2=0$$
$$\Rightarrow d^1\perp d^2$$

  - 共轭方向的几何解释
  - 定理

设 A 是 n 阶对称正定矩阵，d^1, d^2, Λ, d^k 是 k 个 A 共轭的非零向量，则这个向量组线性无关。

  - 共轭梯度算法

对于极小化问题 $\min \quad f(x)=\frac{1}{2}x^tAx+b^tx+c$，

其中 A 是正定矩阵，称下述算法为共轭方向法：

(1) 取定一组 A 共轭方向 $d^{(1)}, d^{(2)}, \Lambda, d^{(n)}$；

(2) 任取初始点 $x^{(1)}$，依次按照下式由 $x^{(k)}$ 确定点 $x^{(k+1)}$，

$$\begin{cases} x^{(k+1)}=x^{(k)}+\lambda_k d^{(k)} \\ f(x^{(k)}+\lambda_k d^{(k)})=\min\limits_{\lambda} f(x^{(k)}+\lambda d^{(k)}) \end{cases}$$

直到某个 $x^{(k)}$ 满足 $\nabla f(x^{(k)})=0$。

注 由定理2可知，利用共轭方向法求解上述极小化问题，至多经过 n 次迭代必可得到最优解。

  - **如何选取一组共轭方向?**
    - **二次函数情形**
      - 正定二次函数基本性质

定理3 对于正定二次函数 $f(x)=\frac{1}{2}x^TAx+b^Tx+c$，FR算法在$m\le n$次一维搜索后即终止，并且对所有的 $i(1\le i\le m)$，下列关系成立

(1) $d^{(i)^T}Ad^{(j)}=0,\ j=1,2,\Lambda\ ,i-1;$

(2) $g_i^{\ T}g_j=0,\ j=1,2,\Lambda\ ,i-1;$

(3) $g_i^Td^{(i)}=-g_i^Tg_i$。 ⟵ $d^{(k+1)}=-g_{k+1}+\beta_k d^{(k)}$

(1) 由定理3可知搜索方向$d^{(1)},d^{(2)},\Lambda\ ,d^{(m)}$是$A$共轭的。

(2) 算法中第一个搜索方向必须取负梯度方向，否则构造的搜索方向不能保证共轭性。

(3) 由定理3的（3）可知，$g_i^Td^{(i)}=-g_i^Tg_i=-\|g_i\|^2<0$，所以$d^{(i)}$是迭代点$x^{(i)}$处的下降方向。

- FR 共轭梯度法

Fletcher – Reeves 共轭梯度法：

$$\min\quad f(x)=\frac{1}{2}x^TAx+b^Tx+c$$

其中$x\in R^n$，A是对称正定矩阵，$b\in R^n$，c是常数。

基本思想：将共轭性和最速下降方向相结合，利用已知迭代点处的梯度方向构造一组共轭方向，并沿此方向进行搜索，求出函数的极小点。

- 利用 FR 法求共轭梯度具体步骤：
  - **以下为分析算法的具体步骤：**

(1) 任取初始点 $x^{(1)}$，第一个搜索方向取为 $d^{(1)}=-\nabla f(x^{(1)})$；

(2) 设已求得点 $x^{(k+1)}$，若$\nabla f(x^{(k+1)})\ne 0$，令 $g_{k+1}=\nabla f(x^{(k+1)})$，则下一个搜索方向$d^{(k+1)}$按如下方式确定：

令 $d^{(k+1)}=-g_{k+1}+\beta_k d^{(k)}$ (1)

如何确定 β_k?

要求$d^{(k+1)}$和$d^{(k)}$关于A共轭。

则在（1）式两边同时左乘 $d^{(k)^T}A$，得

$$0=d^{(k)^T}Ad^{(k+1)}=-d^{(k)^T}Ag_{k+1}+\beta_k d^{(k)^T}Ad^{(k)}$$

解得 $\beta_k=\frac{d^{(k)^T}Ag_{k+1}}{d^{(k)^T}Ad^{(k)}}$ (2)

- 非二次函数情形
  - 利用泰勒展开的一维搜索确定最优步长

$$\min \quad f(x)$$
$$s.t. \quad x \in R^n$$

对用于正定二次函数的共轭梯度法进行修改：

(1) 第一个搜索方向仍取最速下降方向，即$d^{(1)} = -\nabla f(x^{(1)})$。

其它搜索方向按下式计算：

$$d^{(i+1)} = -\nabla f(x^{(i+1)}) + \beta_i d^{(i)},$$

其中$\beta_i = \frac{\| \nabla f(x^{(i+1)}) \|^2}{\| \nabla f(x^{(i)}) \|^2}$。

(2) 搜索步长λ_i不能利用公式（3）计算，需由一维搜索确定。

- 二阶优化
  - 牛顿法
    - 基本思想
      - 在求目标函数 $f(x)$ 的极小值时，先将它在 $x^{(k)}$ 点附近展开成泰勒级数的二次函数式，然后求出函数的极小值点，并以此点作为欲求目标函数的极小值点 x^* 的一次近似值。
      - 设目标函数是连续二阶可微的，将函数在点 $x^{(k)}$ 按泰勒级数展开，并取到二次项：

$$f(x) \approx \Phi(x^{(k)}) = f(x^{(k)}) + [\nabla f(x^{(k)})]^T (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T H(x^{(k)})(x - x^{(k)})$$

      - 对x求导，其极值点必满足一阶导数为零，所以，

$$\nabla \Phi(x) = \frac{\partial f(x)}{\partial x} = \nabla f(x^{(k)}) + \left(x - x^{(k)}\right) H(x^{(k)}) = 0$$

得到 $$x_{\min} = x^{(k)} - \left[H(x^{(k)})\right]^{-1} \nabla f(x^{(k)}) \qquad [1]$$

式中，$\left[H(x^{(k)})\right]^{-1}$essian矩阵的逆矩阵。

在一般情况下，$f(x)$ 不一定是二次函数，因而 $x_{\min}$ 也不可能是 的极值点。但是在 $x^{(k)}$ 点附近， 函数 $\Phi(x)$ 和 $f(x)$ 是近似的，所以可以用 $x^{(k+1)}$ 点作为下一次迭代，即得

$$x^{k+1} = x^{(k)} - \left[H(x^{(k)})\right]^{-1} \nabla f(x^{(k)}) \qquad [2]$$

如果目标函数 $f(x)$ 是正定二次函数，那么 $H(x)$ 是个正矩阵，逼近式 [1]是准确的。因此由 $x^{(k)}$ 点出发只要迭代一次既可以求 $f(x)$ 的极小点。

- 迭代步骤
  -

    ① 给定初始点$x^{(0)}$，计算精度ε，令$k = 0$;

    ② 计算$x^{(k)}$点的梯度$\nabla f(x^{(k)})$、$H(x^{(k)})$及其逆矩阵$\left[H\left(x^{(k)}\right)\right]^{-1}$;

    ③ 构造搜索方向
    $$S^{(k)} = -\left[H\left(x^{(k)}\right)\right]^{-1} \nabla f(x^{(k)})$$

    ④ 沿$S^{(k)}$方向进行一维搜索，得到迭代点$= x^{(k)} + S^{(k)}$
    - 收敛判断：若$\left\|\nabla f(x^{(k+1)})\right\| \le \varepsilon$，则$x^{(k+1)}$为近似最优点，迭代停止，输出最优解$x_{min} = x^{(k+1)}$和$f(x_{min}) = f(x^{(k+1)})$，终止计算。
    - 若不满足$\left\|\nabla f(x^{(k+1)})\right\| \le \varepsilon$，则令$k = k + 1$，转到步骤2继续迭代。

- 拟牛顿法
  - 基本思想
    - 牛顿法收敛很快，但需要计算黑塞矩阵，而此矩阵可能是非正定的，可能导致搜索方向不是下降方向;
    - 牛顿法的一大优势在于：如果初始点离极小点比较近，牛顿法可以表现出相当好的收敛性。
    - 拟牛顿法使用不包含二阶导数的矩阵近似黑塞矩阵。
  - **拟 Newton 条件**
    - 拟 Newton 条件:
      - 考虑无约束非线性优化问题(UNP)，即： $\min_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})$,其中 $f(x)$ 二阶连续可微。
      - 设$\boldsymbol{x^k}$是当前迭代点，$\boldsymbol{H} \in \boldsymbol{R^{n\times n}}$是对称矩阵，$\boldsymbol{B} = \boldsymbol{H^{-1}}$，令
      $$\boldsymbol{d}^k = -H_k \nabla f(\boldsymbol{x}^k) = -B_k^{-1} \nabla f(\boldsymbol{x}^k)$$
      - 将$\boldsymbol{d^k}$作为f在$\boldsymbol{x^k}$处的搜索方向。显然，当H是正定时，$\boldsymbol{d^k}$是f在$\boldsymbol{x^k}$处的下降方向，并且$\boldsymbol{H^{-1}}$一范意义下和$\boldsymbol{B_k}$一范意义下的最速下降方向，所以可以称$\boldsymbol{d^k}$为f在$\boldsymbol{x^k}$处的变尺度方向。
    - 关于 H_k 的拟 Newton 条件

利用梯度函数 $\nabla f(\boldsymbol{x})$ 的 Taylor展开式

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

即

$$\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1}) \approx \nabla^2 f(\boldsymbol{x}^k)(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$$

记 $\boldsymbol{g}^k = \nabla f(\boldsymbol{x}^k)$， $\boldsymbol{p}^{k-1} = \boldsymbol{x}^k - \boldsymbol{x}^{k-1}$， $\boldsymbol{q}^{k-1} = \boldsymbol{g}^k - \boldsymbol{g}^{k-1}$，则由上式，

$$\boldsymbol{q}^{k-1} \approx \nabla^2 f(\boldsymbol{x}^k)\boldsymbol{p}^{k-1}，\text{或} \left[\nabla^2 f(\boldsymbol{x}^k)\right]^{-1} \boldsymbol{q}^{k-1} \approx \boldsymbol{p}^{k-1}$$

因此，为使 H_k 是 $\left[\nabla^2 f(\boldsymbol{x}^k)\right]^{-1}$ 的某种近似，要求满足

$$\boldsymbol{p}^{k-1} = H_k \boldsymbol{q}^{k-1}$$

称为关于$\boldsymbol{H_k}$的拟Newton条件，刻画了H_k近似于$[\nabla^2 f(x^k)]^{-1}$时应具有的一个重要特性。

-

由上式可知，B_k应满足：$q^{k-1} = B_k p^{k-1}$
同样，上式称为关于B_k的拟Newton条件，此时

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

称为f在x^k处的**拟Newton方向**。

其实，拟Newton条件使二次函数

$$\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^T(\boldsymbol{x} - \boldsymbol{x}^k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^k)^T B_k(\boldsymbol{x} - \boldsymbol{x}^k)$$

具有插值性质：

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

- **DFP 法- H_k 的秩 2 修正法**
  - 基本思想
- **求解 UNP 的 DFP 方法**
- **BFGS 方法-DFP 法的对偶方法**

- ~~图论和离散优化~~
  - ~~常见的图论问题~~
    - ~~Hamilton 问题~~
    - ~~旅行商问题~~
  - ~~最短路径问题~~
    - ~~Dijkstra 算法~~