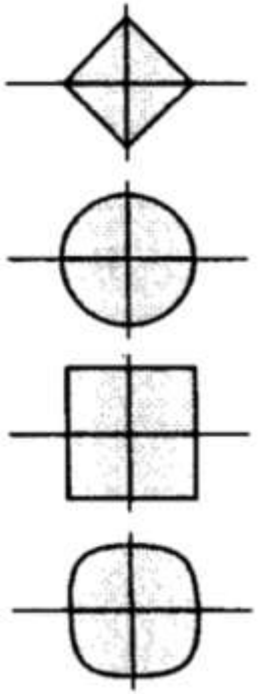
Vector Norms and Matrix Norms

- 向量范数(Vector Norms),Matlab调用:norm(x,1,2,inf,-inf,p)
- $\mathbf{1}-$范数: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N}|x_i|$
- $\mathbf{2}-$范数: $\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{N}|x_i|^2\right)^{\frac{1}{2}}$, Euclidean Norm,欧几里德范数
- $\infty-$范数: $\|\boldsymbol{x}\|_\infty = max_i|x_i|$
- $-\infty-$范数: $\|\boldsymbol{x}\|_{-\infty} = min_i|x_i|$
- $\boldsymbol{p}-$范数:$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{N}|x_i|^p\right)^{1/p}$,也称为Holder范数
- 注意：显然$\boldsymbol{p}-$范数包括前面的几种特殊情况，$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{N}|x_i|^p\right)^{\frac{1}{p}}$(不妨令$x_i$降序排列)= $|x_1|\left(k+\left|\frac{x_{k+1}}{x_1}\right|^p+\cdots+\left|\frac{x_N}{x_1}\right|^p\right)^{\frac{1}{p}} \to |x_1|$

若N=2,则右边图像分别对应哪个范数?



2.0 Vector Norms and Matrix Norms

- General Matrix Norms
- $\|\boldsymbol{A}\| \geq \boldsymbol{0}, \|\boldsymbol{A}\| = \boldsymbol{0} \Leftrightarrow \boldsymbol{A} = \boldsymbol{0}$
- $\|\boldsymbol{\alpha A}\| = |\boldsymbol{\alpha}|\|\boldsymbol{A}\|$
- $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$
- $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$（相容性）
- 诱导矩阵范数(Induced Matrix Norms)
- $\|\boldsymbol{A}\| = \boldsymbol{max}_{\|\boldsymbol{x}\|=\boldsymbol{1}}\|\boldsymbol{Ax}\|$,$\boldsymbol{A} \in \boldsymbol{C}^{\boldsymbol{m}\times\boldsymbol{n}}, \boldsymbol{x} \in \boldsymbol{C}^{\boldsymbol{n}\times\boldsymbol{1}}$则称向量范数诱导出矩阵范数
- 显然有：$\|\boldsymbol{Ax}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|$,若$\boldsymbol{A}$非奇异，则$\boldsymbol{min}_{\|\boldsymbol{x}\|=\boldsymbol{1}}\|\boldsymbol{Ax}\| = \frac{\boldsymbol{1}}{\|\boldsymbol{A}^{-\boldsymbol{1}}\|}$

2.0 Vector Norms and Matrix Norms

- 向量的2范数求导
- $f(x) = \frac{1}{2}||Ax - b||_2^2$,则其导数$f'(x) = A^*(Ax - b)$
- 推导：$f(x) = \frac{1}{2}(Ax - b)^T(Ax - b) \Leftrightarrow f(x) = \frac{1}{2}(x^T A^T Ax - 2b^T Ax + b^T b)$
- 对x求导数得：$f'(x) = A^T Ax - A^T b = A^T(Ax - b)$
- 注意：$\frac{\partial y^T x}{\partial x} = y, \frac{\partial (x^T Ax)}{\partial x} = (A + A^T)x$

正则化

- Tikhonov正则化：选择$\hat{x}$，对给定的$\lambda > 0$，最小化
$$minimize_x \ ||Ax - y||^2 + \lambda||x||^2$$
- 以数学家Andrey Tikhonov的名字命名，目的是求与测量值相容的猜测值$\hat{x}$（即$||A\hat{x} - y|||^2$较小），同时该值也不太大。
- 此时，堆叠矩阵$\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda}I \end{bmatrix}$的列总是线性无关，这时对矩阵$A$没有任何限制条件或约束，并注意到：$\tilde{A}x = \left(Ax, \sqrt{\lambda}x\right) = 0$意味着$\sqrt{\lambda}x = 0$,从而$x = 0$。$\tilde{A}$的Gram矩阵为：$\tilde{A}^T\tilde{A} = A^TA + \lambda I$,因此当$\lambda > 0$时，其总是可逆的。于是**Tikhonov正则近似解为**：
$$\hat{x} = (A^TA + \lambda I)^{-1}A^Ty$$

思考：为什么近似解是这个？

16. $x, y \in R^n$,则点 y 到集合$\{x|Ax = b\}$(A为$m \times n$的矩阵，且$rank(A) = m < n, b \in R^m$)的投

影为 $(A^TA+\lambda I)^{-1}A^Ty$. 。

求解$\boldsymbol{Ax = b}$

$$[A] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

- $Ax = b$
  - 若$rank(A) < rank(A\ b)$，不相容的系统
  - 若$rank(A) = rank(A\ b)$，相容系统
    - 若$rank(A)$ =未知变量的个数，唯一解
    - 若$rank(A)$ <未知变量的个数，无穷多解
- 上述高斯消元法求解方程中前向消元的过程等价于矩阵分解:
  - $A = LU$,其中L是下三角矩阵，U是上三角矩阵
  - 解$Ax = b$,则$LUx = b \Rightarrow Ux = L^{-1}b$
  - 令$L^{-1}b = y$,则得$Ly = b, Ux = y$
  - 注意：$Ly = b$用前向替换解得y；然后$Ux = y$用后向替换解得x
  - 思考：如何获得一个非奇异矩阵A的这种LU分解？

https://math.libretexts.org/Bookshelves/Linear_Algebra/Introduction_to_Matrix_Algebra_(Kaw)/01%3A_Chapters/1.07%3A_LU_Decomposition_Method_for_Solving_Simultaneous_Linear_Equations

LU分解求逆计算时间: $CT|_{invLU} = 1 \times CT|_{DE} + n \times CT|_{FS} + n \times CT|_{BS}$

$$= T\left(\frac{32n^3}{3} + 12n^2 - \frac{20n}{3}\right)$$

高斯消元GE求逆计算时间: $CT|_{invGE} = n \times CT|_{FE} + n \times CT|_{BS}$

$$= n \times T(4n^2 + 12n)$$

$$= T\left(\frac{8n^4}{3} + 12n^3 + \frac{4n^2}{3}\right)$$

$A = LU$

- $n \times n$的方阵A分解为LU形式的计算时间复杂性为
- $CT|_{DE} = T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right)$, T为时钟周期时间clock cycle time:+,-,*: 4, 除: 16, AMD K7处理器
- 前向替换求解$Ly = b$的计算时间$CT|_{FS} = T(4n^2 - 4n)$
- 后向替换求解$Ux = y$的计算时间$CT|_{BS} = T(4n^2 + 12n)$
- 因此总的计算LU分解的计算时间
- $\boldsymbol{CT|_{LU}} = CT|_{DE} + CT|_{FS} + CT|_{BS} = T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right) + T(4n^2 - 4n) + T(4n^2 + 12n) = \boldsymbol{T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right)}$
- 高斯消元法的前向消除计算时间$CT|_{FE} = T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right)$
- 后向替换的计算时间为$CT|_{BS} = T(4n^2 + 12n)$
- 因此，总的高斯消元法的计算时间$\boldsymbol{CT|_{GE}} = CT|_{FS} + CT|_{BS} = T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right) + T(4n^2 + 12n) = \boldsymbol{T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right)}$
- **问题：既然复杂性一样，为何还需要做LU分解?**

回顾求逆: $AA^{-1} = I \Rightarrow A\bar{a_i} = e_i$

30. 高斯消元法和 LU 分解的复杂性一样，为何还需要做 LU 分解，请举例说明。

求逆更快

1. 通过对矩阵进行高斯消元来对矩阵进行 A=LU 分解，假设原矩阵为$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$，则

L=$\cancel{\begin{bmatrix} -2 & 0 & 0 \\ -1 & 1 & 0 \\ -\frac{1}{2} & \frac{3}{10} & \frac{1}{2} \end{bmatrix}}$，U=$\begin{bmatrix} 1 & \frac{1}{2} & 0 \\ 0 & \frac{5}{2} & -1 \\ 0 & 0 & \frac{8}{5} \end{bmatrix}$。

$$\begin{bmatrix} -2 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -\frac{2}{5} & 1 \end{bmatrix}$$

矩阵的条件数

- $||e|| = ||x - \hat{x}|| = ||A^{-1}b - A^{-1}\hat{b}|| = ||A^{-1}(b - \hat{b})|| \le ||A^{-1}|| \cdot ||b - \hat{b}|| = ||A^{-1}|| \cdot ||r||$
- 因此，绝对误差$||e|| \le ||A^{-1}|| \cdot |r||$
- 但经常采用相对误差来衡量
  - $||e|| \le ||A^{-1}|| \cdot ||r|| \cdot. \frac{||Ax||}{||b||} \le ||A^{-1}|| \cdot ||A|| \cdot ||x|| \cdot \frac{||r||}{||b||}$
  - 从而有$\frac{||e||}{||x||} \le ||A^{-1}|| \cdot ||A|| \cdot \frac{||r||}{||b||} = \kappa(A) \cdot \frac{||r||}{||b||}$
  - 而$\kappa(A) = ||A^{-1}|| \cdot ||A||$称为矩阵$A$的条件数

23. 求解方程中，系数矩阵的条件数越大越好（X）

3.2 矩阵的特征值分解-特征值和特征向量 (Eigenvalues and Eigenvectors)

- 根据前面的介绍,高斯消元法中等价于$S = LU$分解，将U中的主元引入到对角矩阵中，则$S = LDL^T$,然后将D分解到两端$A = A^T A$
  - $S = LU$
  - $LU = LDL^T$
  - $S = A^T A$

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & & \\ -\frac{1}{2} & 1 & \\ 0 & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ & \frac{3}{2} & -1 \\ & & \frac{4}{3} \end{bmatrix}$$

$$\begin{bmatrix} 1 & & \\ -\frac{1}{2} & 1 & \\ 0 & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{2} & & \\ & \mathbf{\frac{3}{2}} & \\ & & \mathbf{\frac{4}{3}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ & 1 & -\frac{2}{3} \\ & & 1 \end{bmatrix}$$

$$\begin{bmatrix} \sqrt{2} & & \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{3}{2}} & \\ 0 & -\sqrt{\frac{2}{3}} & \sqrt{\frac{4}{3}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & -\sqrt{\frac{1}{2}} & 0 \\ & \sqrt{\frac{3}{2}} & -\sqrt{\frac{2}{3}} \\ & & \sqrt{\frac{4}{3}} \end{bmatrix}$$

19. 矩阵 A 的唯一特征向量是$(1,4)^T$的倍数，则必定不可逆（ X ）；有重复的特征值（ √ ）；不能对角化为$X\Lambda X^{-1}$（ √ ）

$\prod \lambda_i = |A|$

四个空间的基本关系

· 行空间 $C(A^T)$和零空间$N(A)$正交	$C(A) \cap N(A^T) = \vec{0}$
· 列空间 $C(A)$ 和左零空间$N(A^T)$正交	$C(A^T) \cap N(A) = \vec{0}$

矩阵间关系 ⟹ 特征值间关系.

1. 幂次关系

1. $A^2x = A(Ax) = A(\lambda x) = \lambda(Ax) = \lambda^2 x$，矩阵幂的关系也会使得特征值有幂的关系：当矩阵平方，特征值也平方；当矩阵n次方，特征值也n次方。特征向量是保持不变的，这个对 A 来说“好”的方向对 A^n 也是“好”方向。

2. 线性数乘.

2.$(2A)x = (2\lambda)x$，当矩阵变为c倍，特征值也变为原来的c倍，特征向量不变。

3. 矩阵移动 nI

3.$(A - I)x = Ax - x = (\lambda - 1)x$，矩阵移动 I，特征值改变1；矩阵移动 cI，特征值移动 c。特征向量不变 。

4. 逆

$Ax = \lambda x \Rightarrow A^{-1}Ax = A^{-1}\lambda x \Rightarrow x = \lambda A^{-1}x$，这也就是 $A^{-1}x = \frac{1}{\lambda}x$，可逆矩阵它逆的特征值是它特征值的倒数，特征向量不变。

SVD的实现（以下仅针对实数域）

矩阵	别称	维度	计算方式	含义
U 矩阵	A 的左奇异矩阵	m 行 m 列	列由 AA^T 的特征向量组成，且特征向量为单位向量	包含了有关行的所有信息（代表自己的观点）
$\sum$ 矩阵	A 的奇异值矩阵	m 行 n 列	对角元素来源于 AA^T 或 A^TA 的特征值的平方根，并且按降序排列，值越大可以理解为越重要	记录 SVD 过程（是一种日志）
V 矩阵	A 的右奇异矩阵	n 行 n 列	列由 A^TA 的特征向量组成，且特征向量为单位向量	包含了有关列的所有信息（代表自己的特征）

$$A = U\Sigma V^T \Rightarrow AV = U\Sigma V^T V \Rightarrow AV = U\Sigma \Rightarrow Av_i = \sigma_i u_i \Rightarrow \sigma_i = Av_i / u_i$$

上面还有一个问题没有讲，就是我们说A^TA 的特征向量组成的就是我们SVD中的V矩阵，而

AA^T 的特征向量组成的就是我们SVD中的U矩阵，这有什么根据吗？这个其实很容易证明，我们以V矩阵的证明为例。

$$A = U\Sigma V^T \Rightarrow A^T = V\Sigma U^T \Rightarrow A^TA = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$$

上式证明使用了$U^U = I, \Sigma^T = \Sigma$ 。可以看出 A^TA 的特征向量组成的的确就是我们SVD中的V矩阵。类似的方法可以得到 AA^T 的特征向量组成的就是我们SVD中的U矩阵。

问题4：秩为k的矩阵中，最接近 A 的是 $A_k = \sigma_1 u_1 v_1^t + \sigma_2 u_2 v_2^t + .. + \sigma_k u_k v_k^t$。这个接近是用矩阵的范数衡量的，即 $||A - A_k|| \leq ||A - B||$，其中 B 的秩为k。这个定理叫Eckart-Young-Mirsky Theorem。SVD不仅仅是矩阵分解，它还是最优近似。

3.4稀疏表示-正交匹配追击*Orthogonal-Matching-Pursuit*（OMP）

- 正交匹配追击（OMP）

任务：获取(P_0): $\min_x ||x||_0, s.t. Ax = b$的近似解

输入参数：给定矩阵A, b,以及误差阈值ϵ_0

输出　　：第k次近似解x^k

初始化：$k = 0$,并设定

- 初始解：$x^0 = 0$
- 初始残差：$r^0 = b - Ax^0 = b$
- 初始解支撑$S^0 = Support\{x^0\} = \emptyset$

迭代过程：$k \leftarrow k + 1$

- **Sweep**:计算误差$\epsilon(j) = \min_{z_j} ||a_j z_j - r^{k-1}||_2^2$,其中$z_j$用最优值$z_j^* = \frac{a_j^T r^{k-1}}{||a_j||_2^2}$代入
- **更新支撑**：求$\epsilon(j)$的最小指标j_0: $\forall j \notin S^{k-1}, \epsilon(j_0) \le \epsilon(j)$，更新$S^k = S^{k-1} \cup \{j_0\}$
- **更新临时解**：计算$x^k = minize_x ||Ax - b||_2^2, s.t. Support\{x\} = S^k$;
- **更新残差**:计算$r^k = b - Ax^k$;
- **停止准则**：如果$||r^k||_2 < \epsilon_0$则停止，否则继续循环

输出：输出x^k

$$\epsilon(j) = \min_{z_j} ||a_j z_j - r^{k-1}||_2^2 = ||\frac{a_j^T r^{k-1}}{||a_j||_2^2} a_j - r^{k-1}||_2^2$$

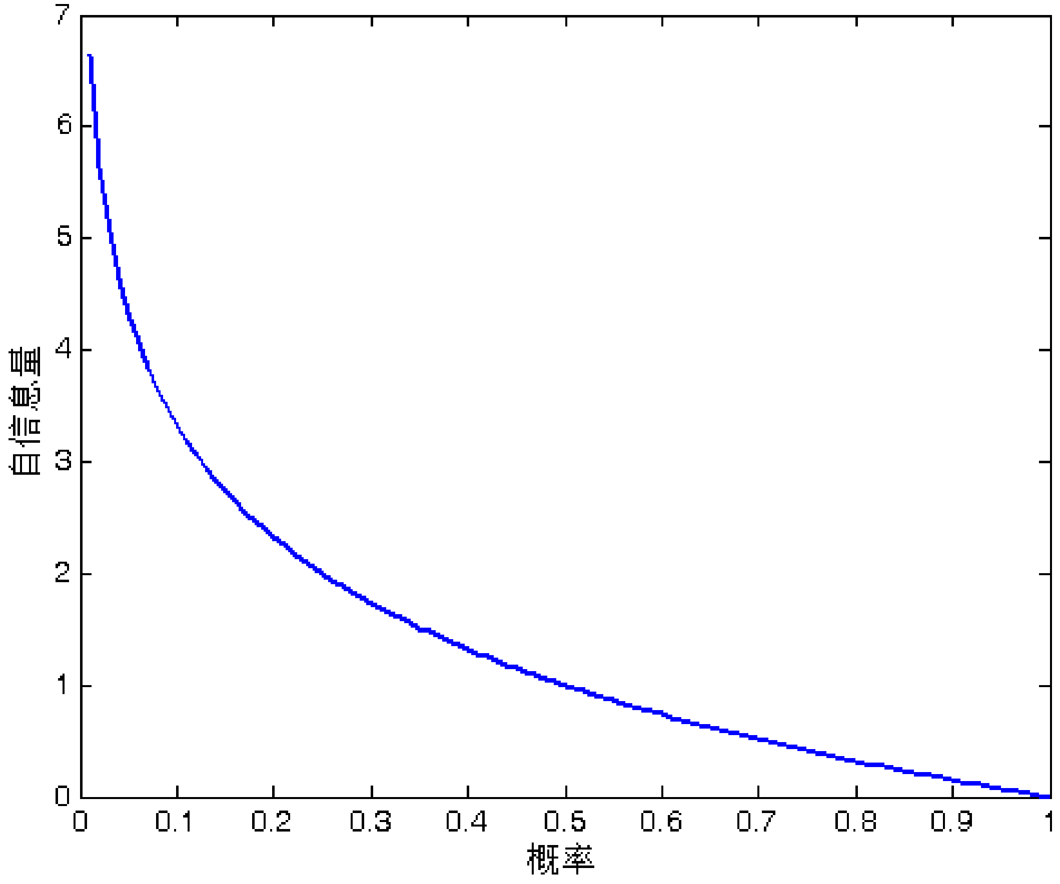
$$= ||r^{k-1}||_2^2 - 2\frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2} + \frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2}$$

$$= = ||r^{k-1}||_2^2 - \frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2}$$

因此计算最小误差等价于计算残差与矩阵A的剩余列向量之间的内积

4.3 自信息量

- **定义4.3.1** 任意随机事件的自信息量定义为该事件发生概率的对数的负值。
- 假设事件x_i发生的概率为$p(x_i)$，则其自信息定义为$I(x_i) = -logp(x_i)$
  - 自信息量的单位与log函数所选用的对数底数有关，如底数分别取 2、 *e*、10，则自信息量单位分别为：比特、奈特、哈特
  - 事件x_i发生以前，表示事件发生的先验不确定性
  - 当事件x_i发生以后，表示事件x_i所能提供的最大信息量（在无噪情况下）



4.3自信息量-互信息量

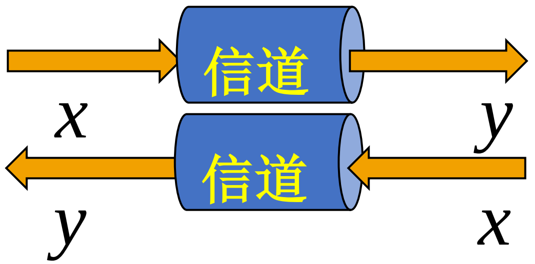
- **定义2** 随机事件y的出现给出关于事件x的信息量，定义为互信息量。定义式：$I(x;y)=\log\frac{p(x|y)}{p(x)}$
- 单位：同自信息量.
- 含义：本身的不确定性，减去知道事件y之后仍然保留的不确定性，就是**由y所提供的关于x的信息量**，或者说**由y所消除的关于x的不确定性**
- 互信息量=**原有的不确定性** - **尚存在的不确定性**
- **定义3** 则平均意义上来说，信源X的不确定程度可以表示为$\boldsymbol{H(X)=-\sum_{i=1}^{n}p_i log p_i}$,称为信源$X$的**熵**！
  - 等价于每个事件的自信息量的平均值（或期望）
  - $H(X)=E(I(x_i))=E[-\log p(x_i)]=-\sum_{i=1}^{n}p(x_i)\log p(x_i)$

4.3互信息量、熵

- 互信息量的性质
  - 对称性$I(x;y) = I(y;x)$：由y所提供的关于x的信息量 = 由x 所提供的关于y的信息量
  - 当事件x、y统计独立时，互信息量为0：p(x|y)=p(x)，x和y之间没有什么关系，无论是否知道y，都对x出现的概率没有影响
  - 可正可负：y的出现有利/不利于确定x的发生
    - 正：$I(x;y) = \log\frac{p(x|y)}{p(x)} > 0 \Rightarrow \frac{p(x|y)}{p(x)} > 1 \Rightarrow p(x|y) > p(x)$，$x$：张三病了。$y$：张三没来上课。
    - 负：$I(x;y) = \log\frac{p(x|y)}{p(x)} < 0 \Rightarrow \frac{p(x|y)}{p(x)} < 1 \Rightarrow p(x|y) < p(x)$，$x$：李四考了全班第一名。$y$：李四没有复习功课。
  - 互信息量不大于任一事件的自信息量

$$I(x;y) = I(x) - I(x|y)$$

$$I(y;x) = I(y) - I(y|x).$$



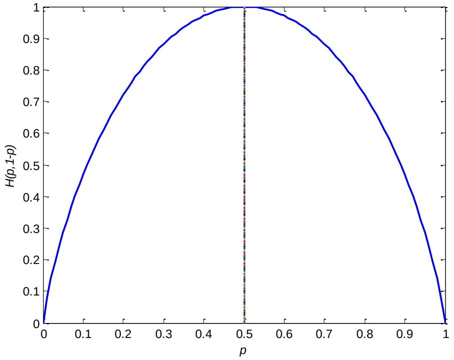
4.4熵-离散信源最大熵定理

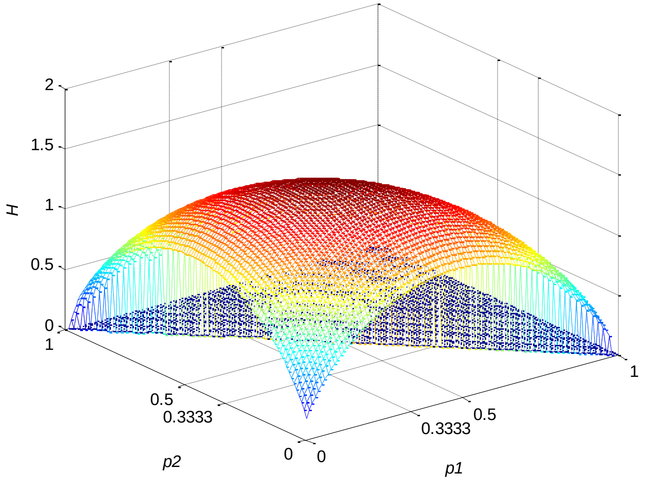
- 熵（entropy）
  - 条件熵：$H(X|Y)=\sum_{XY}p(x,y)I(x|y)=-\sum_{XY}p(x,y)\log p\,(x|y)$
    - 若X表示输入，Y表示输出，$H(X|Y)$表示信道损失
  - 联合熵（共熵）：$H(X,Y)=\sum_{XY}p(x,y)I(x,y)=-\sum_{XY}p(xy)\log p\,(x,y)$
- 熵的性质
- 1. 对称性：与整体有关，个体无关
- 2. 非负性：$H(X)\geq 0$；$H(X)=E[I(x_i)]=\sum_{i=1}^{n}p(x_i)\log\frac{1}{p(x_i)}$
- 3. 扩展性:$\lim_{\varepsilon\to 0}H_{q+1}(p_1,p_2,\cdots,p_q-\varepsilon,\varepsilon)=H_q(p_1,p_2,\cdots,p_q)$
  - **集合X有q个事件，集合Y比X仅仅是多了一个概率接近0的事件，则两个集合的熵值一样**
  - 证明：$x\log x$ 在[0，∞)的连续性，$\lim_{x\to 0}x\log x=0$
  - 含义：集合中，一个事件发生的概率比其它事件发生的概率小得多时，这个事件对于集合的熵值的贡献可以忽略

7. 对于强噪信道的输入输出分别为 X，Y，则 I(X;Y)=___0___，对于一般的信道，则 I(X;Y)=___H(X)−H(X|Y)___。（用 X, Y 的熵和联合熵的表达式表示）

4.4熵-离散信源最大熵定理

- 4. 可加性:设X和Y为两个互相关联的随机变量,X 的概率分布为$\{p_1, p_2, ..., p_m\}$,Y 的概率分布为$\{q_1, q_2, ..., q_n\}$，则 $\mathrm{H(XY) = H(X) + H(Y|X)}$.
  - 当X、Y相互独立时，$\boldsymbol{H(X,Y) = H(X)+H(Y)}$
- **5. 极值性**:$H(p_1, p_2, ..., p_n) \le H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) = \log n$, $H(X) \le logn$
  - **离散信源最大熵定理：**各事件等概率发生时，熵最大
  - $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ p & 1-p \end{bmatrix}$
  - $H(X) = -p\log p - (1-p)\log(1-p)$
  - $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ p_1 & p_2 & 1-p_1-p_2 \end{bmatrix}$
  - $H(X) = -p_1 \log p_1 - p_2 \log p_2 \ - (1-p_1-p_2)\log(1-p_1-p_2)$



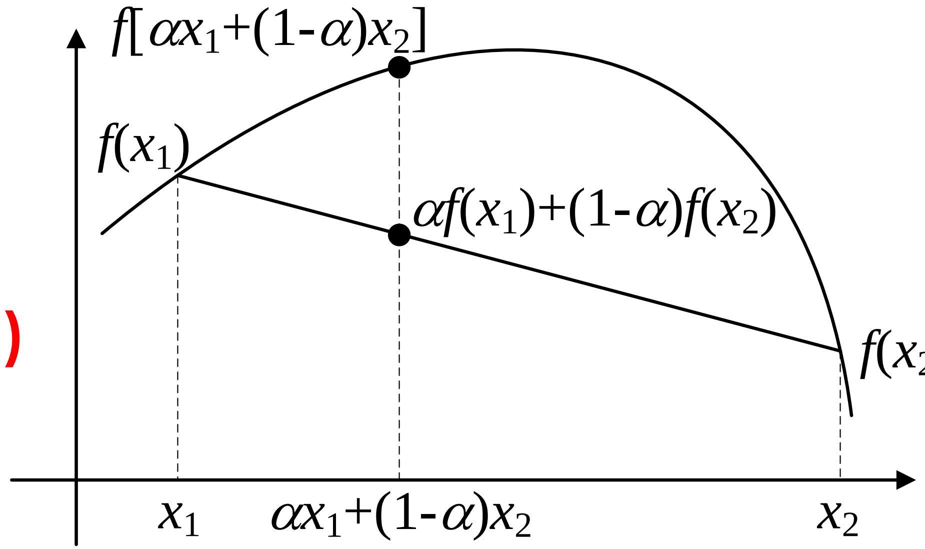


- 6. 确定性
- 7. 上凸性

4.4熵-离散信源最大熵定理

- 问题：求熵的最大值$H(X) = -\sum_{i=1}^{n} p_i log p_i$，subject to $\sum_{i=1}^{n} p_i = 1$
- **定义4.3.2** 如果$f[\alpha x_1 + (1-\alpha)x_2] \geq \alpha f(x_1) + (1-\alpha)f(x_2)$，其中$0 < \alpha < 1$，称$f(x)$为**凹函数（上凸函数）**。如果$f[\alpha x_1 + (1-\alpha)x_2] > \alpha f(x_1) + (1-\alpha)f(x_2)$，则称$f(X)$为**严格凹函数（上凸函数）**
- **类似的，有凸函数（下凸函数）以及严格凸函数的定义**
- 对于凹函数（上凸函数），有**詹森（Jenson）不等式**

$$f(E[x]) \geq E[f(x)]$$



4.4熵-离散信源最大熵定理

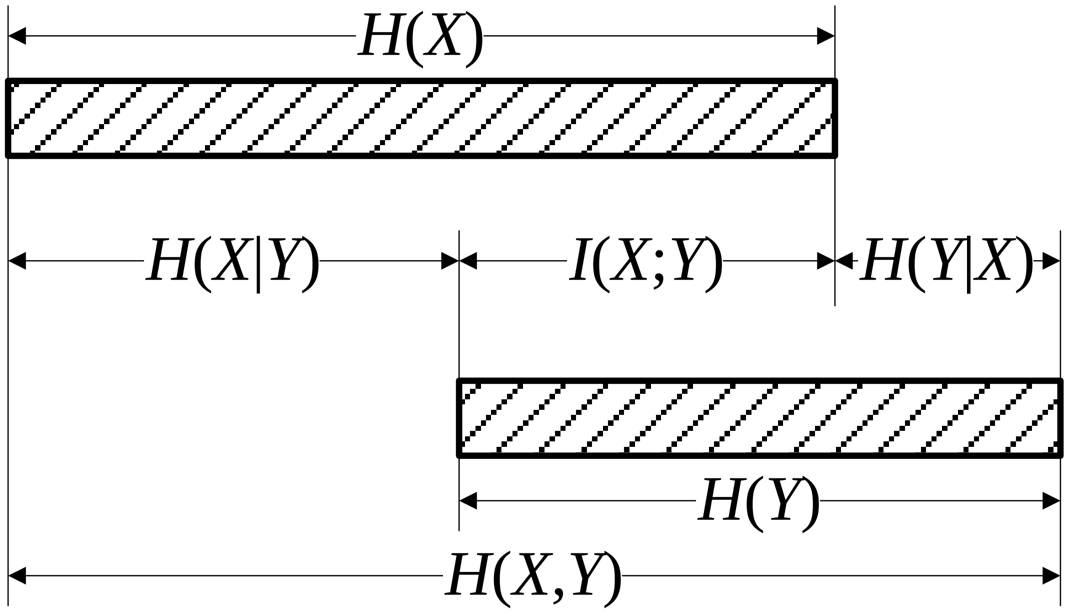
- **问题**：求熵的最大值$H(X) = -\sum_{i=1}^{n} p_i logp_i$, subject to $\sum_{i=1}^{n} p_i = 1$
- **引理**：如果$\sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} q_i = 1$,则$\sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) \le \sum_{i=1}^{n} p_i \log\left(\frac{1}{q_i}\right)$，当前仅当对$\forall i, p_i = q_i$时等号成立!
- **证明**：$\because \forall x > 0, logx \le x - 1,$等号仅当$x = 1$成立$, \therefore \sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) - \sum_{i=1}^{n} p_i \log\left(\frac{1}{q_i}\right) = \sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) \le \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n}(q_i - p_i) = 0$■
- 按照之前的思路，需要证明$H(X)$是概率的上凸函数，即设P_1, P_2是两个概率分布，$0 < \alpha < 1$，证明：$H[\alpha P_1 + (1-\alpha)P_2] > \alpha H(P_1) + (1-\alpha)H(P_2)$
- 证明：根据引理，令$q_i = \alpha p_{1i} + (1-\alpha)p_{2i}$,因为$0 < \alpha < 1$,因此$p_{1i} \ne q_i(\forall i)$，不满足引理等号成立的条件。$\therefore \sum_{i=1}^{n} p_{1i} \log\left(\frac{1}{p_{1i}}\right) \le \sum_{i=1}^{n} p_{1i} \log\left(\frac{1}{q_i}\right)$,因此$\sum_{i=1}^{n} p_{1i} \log\left(\frac{q_i}{p_{1i}}\right) < 0$,代入$q_i$,
- $H[\alpha P_1 + (1-\alpha)P_2] - \alpha H(P_1) + (1-\alpha)H(P_2) = -\sum_{i=1}^{n}(\alpha p_{1i} + (1-\alpha)p_{2i}) \log(\alpha p_{1i} + (1-\alpha)p_{2i}) + \alpha \sum_{i=1}^{n} p_{1i} logp_{1i} + (1-\alpha)\sum_{i=1}^{n} p_{2i} \log p_{2i} = \alpha \sum_{i=1}^{n} p_{1i} \log\left(\frac{p_{1i}}{\alpha p_{1i}+(1-\alpha)p_{2i}}\right) + (1-\alpha)\sum_{i=1}^{n} p_{2i} \log\left(\frac{p_{2i}}{\alpha p_{1i}+(1-\alpha)p_{2i}}\right) = \alpha \sum_{i=1}^{n} p_{1i} \log\left(\frac{p_{1i}}{q_i}\right) + (1-\alpha)\sum_{i=1}^{n} p_{2i} \log\left(\frac{p_{2i}}{q_i}\right) > 0$
- 因此，得证$H(X)$是凹函数或上凸函数!

4.4熵-离散信源最大熵定理

- 求熵的最大值$H(X) = -\sum_{i=1}^{n} p_i log p_i$, subject to $\sum_{i=1}^{n} p_i = 1$
- **解**：根据拉格朗日乘子法，目标函数$f(p_i, \lambda) = H(X) + \lambda(\sum p_i - 1)$,对参数求偏导数，并令其为0，则有
- $\begin{cases} \frac{\partial f}{\partial p_i} = -log p_i - p_i \cdot \frac{1}{p_i} + \lambda = 0 \\ \sum p_i - 1 = 0 \end{cases} \Rightarrow \begin{cases} log p_i = \lambda - 1 \\ \sum p_i = 1 \end{cases} \Rightarrow \begin{cases} p_i = e^{\lambda - 1} \\ n e^{\lambda - 1} = 1 \end{cases}$ 从而，可得$p_i = \frac{1}{n}$
- 熵函数的形式除了一个常数倍外是唯一确定的！

4.5平均互信息量、相对熵

- 各种信息量之间的关系：平均互信息量与信源熵、条件熵的关系
  - $I(X;Y) = H(X) - H(X \mid Y)$
  - $I(X;Y) = H(Y) - H(Y \mid X)$
  - $I(X;Y) = H(X) + H(Y) - H(XY)$



4.5平均互信息量、相对熵

- **相对熵**：$D(p||q) = \sum_{x\in X} p(x)\log\left(\frac{p(x)}{q(x)}\right)$:相对熵,交叉熵，Kullback熵，K-L散度，K-L距离，方向散度)
  - 相对熵非负，互信息非负
  - **证明**：1)对任意$x > 0$,都有$1 - \frac{1}{x} \le lnx \le x - 1$成立，等式成立的充要条件是$x = 1$,令$f(x) = x - 1 - lnx, x > 0$，则$f''(x) = \frac{1}{x^2} > 0$,从而$f(x)$为下凸函数，且最小值$x = 1$取得$f(1) = 0$，从而$x - 1 \ge lnx$,然后令$x := \frac{1}{x}$，代入得证;
  - 2)$\log\left(\frac{q_i}{p_i}\right) = loge \cdot \ln\left(\frac{q_i}{p_i}\right) \le loge \cdot \left(\frac{q_i}{p_i} - 1\right)$,从而：

$$\sum_i p_i \log\left(\frac{q_i}{p_i}\right) \le \sum_i p_i loge \cdot \left(\frac{q_i}{p_i} - 1\right) = loge \cdot \sum_i (q_i - p_i) = 0$$

  - 从而$\sum_i p_i \log\left(\frac{p_i}{q_i}\right) \ge 0$,等号成立当且仅当$\frac{p_i}{q_i} = 1$

25. 相对熵$D(p||q) = \sum p(x)\log\left(\frac{p(x)}{q(x)}\right)$非负（✓）

4.6 决策树-信息熵的直接应用

决策树分类器是一种基于树状结构的机器学习模型，用于对物体或数据进行分类。在构建决策树时，选择合适的特征进行分类是至关重要的。信息增益是决策树中一种用于选择特征的方法，它结合了熵的概念。

在信息理论中，熵是用于衡量系统的不确定性或混乱程度的指标。当系统的状态越不确定或混乱时，熵越高；当系统的状态越确定或有序时，熵越低。在决策树中，熵被用来衡量数据集的不确定性。

信息增益是指在使用某个特征对数据集进行划分后，熵的减少量。换句话说，信息增益衡量了使用特定特征后，数据集的不确定性减少了多少。***决策树分类器会选择能够使信息增益最大化的特征来进行分类，因为这样可以最大程度地减少数据集的不确定性，使得分类结果更加可靠。***

具体而言，决策树分类器在选择特征进行分类时，会计算每个特征的信息增益，然后***选择信息增益最大的特征作为当前节点的划分依据。***这个过程会不断迭代进行，直到达到某个停止条件（例如达到最大深度、节点中的样本数小于某个阈值等）为止，从而构建出整个决策树模型。

总的来说，信息增益的概念结合了熵的概念，帮助决策树分类器选择最优的特征进行分类，从而构建出高效且准确的分类模型。

5.1 基本概率论

- 全概率公式

$$P(B)=\sum_{i=1}^{n}P\left(A_j\right)P\left(B|A_j\right)$$

- 贝叶斯公式

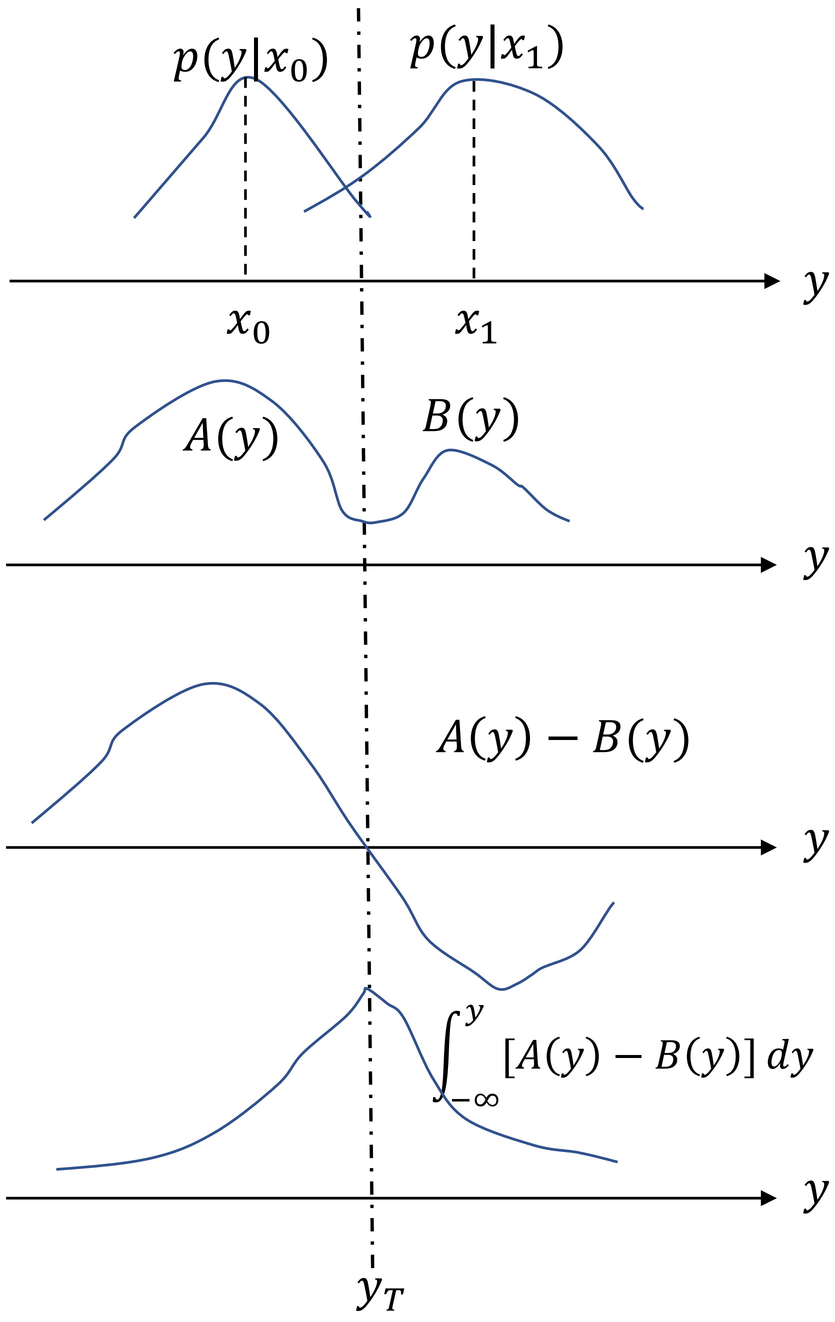
$$P\left(A_j|B\right)=\frac{P\left(A_j\right)P\left(B|A_j\right)}{\sum_{i=1}^{\infty}P(A_i)P(B|A_i)},$$

$$P(A|B)=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})}$$

$$\min\left\{\int_{S_{y_0}}[B(y)-A(y)]\,dy\right\}$$

令$A(y)=(C_{10}-C_{00})p(x_0)p(y|x_0)$
$B(y)=(C_{01}-C_{11})p(x_1)p(y|x_1)$

$$\max\left\{\int_{-\infty}^{y}[A(y)-B(y)]\,dy\right\}$$



5.3检测准则-贝叶斯准则

- $(C_{01}-C_{11})\,p(x_1)p(y|x_1)-(C_{10}-C_{00})p(x_0)p(y|x_0)>0$等价于
- $y=\begin{cases}1, \dfrac{p(y|x_1)}{p(y|x_0)}>\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}\\ 0, \dfrac{p(y|x_1)}{p(y|x_0)}\le\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}\end{cases}$
- 令$L(y)=\dfrac{p(y|x_1)}{p(y|x_0)}$为似然比，$\lambda=\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}$为检测阈值，则条件转化为
- $L(y)\gtreqless\lambda$来判断接收的类别!
- 这就是贝叶斯准则!

原式 $= p(x_0)C_{00} + p(x_1)C_{11} + p(x_0)\beta(C_{10}-C_{00}) + p(x_1)\alpha(C_{01}-C_{11})$

$$= p(x_0)C_{00} + p(x_1)C_{11} + \int_{S_{y_1}} (C_{10}-C_{00})p(y|x_0)p(x_0)dy + \int_{S_{y_0}} (C_{01}-C_{11})p(y|x_1)p(x_1)dy$$

$$= p(x_0)C_{00} + p(x_1)C_{11} + (C_{10}-C_{00})p(x_0)\left(1-\int_{S_{y_0}} p(y|x_0)dy\right) + \int_{S_{y_0}} (C_{01}-C_{11})p(y|x_1)p(x_1)dy$$

$$= p(x_0)C_{00} + p(x_1)C_{11} + (C_{10}-C_{00})p(x_0) - \int_{S_{y_0}} (C_{01}-C_{11})p(y|x_1)p(x_1) - (C_{10}-C_{00})p(x_0)p(y|x_0)dy$$

5.3检测准则-最大后验概率准则

- 贝叶斯准则要求损失函数、转移概率分布和事件的先验概率都为已知，但有些情况无法或者不必要知道损失函数时，可以简单假设错误判决的损失为1而正确判决的损失为0，即$C_{00}=C_{11}=0$，$C_{01}=C_{10}=1$，这时贝叶斯准则中右边的常数
- $\lambda=\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}=\dfrac{p(x_0)}{p(x_1)}$
- 根据贝叶斯公式：
- $\dfrac{p(y|x_1)}{p(y|x_0)}\gtrless\dfrac{p(x_0)}{p(x_1)}\Leftrightarrow\dfrac{p(y|x_1)}{p(y|x_0)}=\dfrac{p(x_1|y)p(y)}{p(x_1)}\cdot\dfrac{p(x_0)}{p(x_0|y)p(y)}\gtrless\dfrac{p(x_0)}{p(x_1)}\Leftrightarrow$
$\dfrac{p(x_1|y)}{p(x_0|y)}\gtrless 1$

5.3检测准则-最大似然准则

- 如果不仅损失函数不知道，先验概率也不知道，那么这时候，就只能假定两种判决出错的概率和代价基本相同
- 这种最合理的假设就是假定常数$\lambda=\frac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}$，这时判决准则变为
- $\frac{p(y|x_1)}{p(y|x_0)} \gtreqless 1$

5.3检测准则-极小极大检测准则

  - $C'_M=[(C_{11}-C_{00})+\alpha(p')(C_{01}-C_{11})-\beta(p')(C_{10}-C_{00})]p(x_1)+C_{00}+\beta(p')(C_{10}-C_{00})$
  - 思考：一旦假想先验概率$p'(x_1)$后，平均贝叶斯风险C'_M与$p(x_1)$的关系是否为一条直线？
- 极小极大准则
  - 选使贝叶斯风险具有最大值时的先验概率作为假想的先验概率，这时真正风险C_{MT}（极小极大的平均损失）应不小于贝叶斯风险C_{MB}。
  - 当$p(x_1)$与$p'_0(x_1)$一致时，两者相同。保守，但保险！
  - 如果实际$p(x_1)$接近于1（实际贝叶斯风险很小），此时选择$p'_0(x_1)$作为先验概率所引起的平均风险C'_M很大！但按照极小极大准则选择，平均风险C_{MT}为与C_{MB}最大点相切的一条水平直线，无论实际先验概率为多少，其平均风险保持不变，若接近1，则$C_{MT}\ll C'_M$
  - $\frac{dC_{MB}(p)}{dp(x_1)}=0$或者$\frac{dC'_M(p)}{dp(x_1)}=0$求得假想先验概率，从而确定阈值和判决界，最后确定平均分险
  - 将其代入，可得：$(C_{11}-C_{00})+\alpha(p')(C_{01}-C_{11})-\beta(p')(C_{10}-C_{00})=0$

5.3检测准则-极小极大检测准则

- $(\boldsymbol{C}_{11}-\boldsymbol{C}_{00})+\boldsymbol{\alpha}(\boldsymbol{p}')(\boldsymbol{C}_{01}-\boldsymbol{C}_{11})-\boldsymbol{\beta}(\boldsymbol{p}')(\mathbf{C}_{10}-\mathbf{C}_{00})=0$
- 令$\boldsymbol{C}_{11}=\boldsymbol{C}_{00}$，则上式成为：

$$\boldsymbol{\alpha}(\boldsymbol{p}')(\boldsymbol{C}_{01}-\boldsymbol{C}_{11})=\boldsymbol{\beta}(\boldsymbol{p}')(\mathbf{C}_{10}-\mathbf{C}_{00})$$ ---（等风险条件）

- 从中可求出$\boldsymbol{p}_0'(\boldsymbol{x}_1)$，于是可求出极小极大准则下的最佳阈值$\boldsymbol{\lambda}$:
- $\boldsymbol{\lambda}=\frac{[1-\boldsymbol{p}_0'(\boldsymbol{x}_1)](\boldsymbol{C}_{10}-\boldsymbol{C}_{00})}{\boldsymbol{p}_0'(\boldsymbol{x}_1)(\boldsymbol{C}_{01}-\boldsymbol{C}_{11})}$

$$y=\begin{cases}1, \dfrac{p(y|x_1)}{p(y|x_0)}>\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}\\ 0, \dfrac{p(y|x_1)}{p(y|x_0)}\le\dfrac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}\end{cases}$$

6.0简单回顾一下线性代数里面的概念凸性(Convexity)

- 矩阵$\boldsymbol{A}$的秩表示为$\boldsymbol{rank(A)}$,非空集合$\boldsymbol{C} \subseteq \boldsymbol{R^n}$的维数$\boldsymbol{dimC}$定义为：
  - $\boldsymbol{n - max\{rank(A): A \in R^{n\times n}, Ax = Ay}$,对所有$\boldsymbol{x, y \in C\}}$
- 凸组合(convex combination):点$\boldsymbol{x_1, x_2, \cdots, x_n}$的凸组合是指点$\boldsymbol{x = \Sigma_{i=1}^{n} \alpha_i x_i, \alpha_i \geq 0}$, 且$\boldsymbol{\Sigma\alpha_i = 1}$
- 仿射组合：点$\boldsymbol{x_1, x_2, \cdots, x_n}$的仿射组合是指点$\boldsymbol{x = \Sigma_{i=1}^{n} \alpha_i x_i}$, 且$\boldsymbol{\Sigma\alpha_i = 1}$
- 锥组合：点$\boldsymbol{x_1, x_2}$的锥组合是指形如$\boldsymbol{x = \alpha_1 x_1 + \alpha_2 x_2, \alpha_1 > 0, \alpha_2 > 0}$
- 凸集(convex set):令$\boldsymbol{A \subseteq R^n}$是凸的，如果$\boldsymbol{\forall x, y \in A}$,则$\boldsymbol{\alpha x + (1-\alpha) y \in A}$,$\boldsymbol{\alpha \in (0,1)}$
- 凸包（convex Hull)
  - 集合$\mathbf{A}$的凸包$\boldsymbol{conv(A)}$定义为$\boldsymbol{A}$中点的所有凸组合.
  - 凸包是包含集合的最小凸集
  - 离散点的凸包示例，扇形凸包示例
- 仿射包（affine Hull)
  - 集合$\boldsymbol{A \subseteq R^n}$的仿射包为$\boldsymbol{A}$中点的组合：affine$\boldsymbol{A} \coloneqq \boldsymbol{\{x | x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k, x_1, x_2, \cdots, x_k \in A, \sum_i \alpha_i = 1\}}$
  - 一般情况下，一个集合的仿射包实际上是包含该集合 的最小的仿射集

6.1 线性规划的基本概念-模型的标准形式

- 目标函数:$\boldsymbol{Max(Min)\ z = c_1x_1 + c_2x_2 + \cdots + c_nx_n}$
- 约束条件

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n = b_1 \\ \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots a_{in}x_n = b_i \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots a_{mn}x_n = b_m \\ x_1, x_2, x_3, \cdots, x_n \geq 0 \end{cases}$$

注意：右端项要求非负

松弛变量(Slack Variable)：化不等式为等式约束

把一般的LP化成标准型的过程称为线性规划问题的标准化

方法：

1 目标标准化

max Z 等价于 min (- Z)

min Z=- $\sum c_jx_j$

2 化约束为等式,加松弛变量、减剩余变量

3 变量非负化

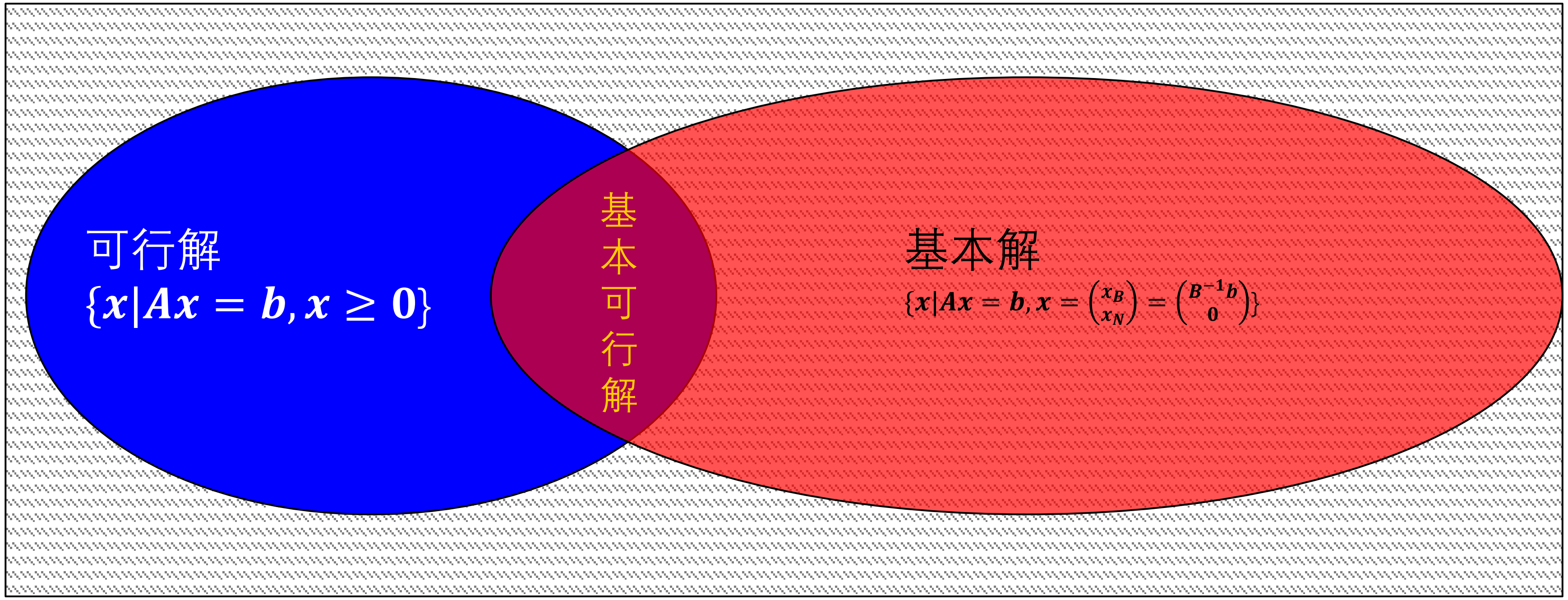
4 右端非负

6.1 线性规划的基本概念-解的基本概念

- **可行解**(Feasible solution)**有时也称为可行域**
  - $\{x \mid Ax = b, x \geq 0\}$
- **最优解:使目标函数取最优的可行解**
- **基**(basis), **基变量**(basic Var.), **非基变量**(non-basic var.)
  - 若$B = [P_1\ P_2\ \ldots\ P_m]$为$A$的一个$m$阶可逆矩阵, $A = [B\ N]$, $\boldsymbol{x} = [\boldsymbol{x}_{\boldsymbol{B}}\ \boldsymbol{x}_{\boldsymbol{N}}]^{\boldsymbol{T}}$, 则称$B$为一个基或基矩阵, 对应$\boldsymbol{x}_{\boldsymbol{B}}$: $x_1, \cdots, x_m$称为基变量, 剩余的$\boldsymbol{x}_{\boldsymbol{N}}$: $n - m$个变量称为非基变量。
- **基本解:非基变量为0时, 满足约束$Ax = b$的解**
  - 基本解至少有$n - m$个分量为0, 至多有m个非零分量
  - 非零分量的个数少于m时, 称为退化的基本解
  - 基本解的个数最多有$C(n, m) = n!/(m!\,(n - m)!)$

6.1 线性规划的基本概念-解的基本概念

- 可行解、基本解



6.1 线性规划的基本概念-解的基本情况

- **解的存在性**：若(LP)的可行域(带约束的多面体)非空，则可行域是个凸集，且(LP)一定存在有限最优解或无界最优解
- **解在顶点的可达性**：若(LP)存在有限最优解，则最优解可在某个顶点处达到
- **顶点与基本可行解的关系**：x_0是(LP)的可行域顶点的充分必要条件是x_0是(LP)的基本可行解

→可通过求基本可行解得到有限最优解

6.1 线性规划的基本概念-线性规划问题的基本理论(续4)(注3)

• 单纯形表

c_j			1	4	0	0	
c_B	x_B	$\bar{b}$	x_1	x_2	x_3	x_4	θ
0	x_3	8	1	2	1	0	$\frac{8}{2}$
0	x_4	2	0	1	0	1	$\frac{2}{1}$
$-z$		0	1	4	0	0	

c_j			1	4	0	0	
c_B	x_B	$\bar{b}$	x_1	x_2	x_3	x_4	θ
0	x_3	4	1	0	1	-2	∞
4	x_2	2	0	1	0	1	$\frac{2}{1}$
$-z$		-8	1	0	0	-4	

c_j			1	4	0	0	
c_B	x_B	$\bar{b}$	x_1	x_2	x_3	x_4	θ
0	x_3	4	1	0	1	-2	4
4	x_2	2	0	1	0	1	∞
$-z$		-8	1	0	0	-4	

c_j			1	4	0	0	
c_B	x_B	$\bar{b}$	x_1	x_2	x_3	x_4	θ
1	x_1	4	1	0	1	-2	
4	x_2	2	0	1	0	1	
$-z$		-12	0	0	-1	-2	

6.1线性规划-对偶问题(续2)

- 如果原问题是标准形式，如何定义其对偶问题
- $\boldsymbol{LP}: \boldsymbol{max}\ \boldsymbol{c^T x, s.t.} \begin{cases} \boldsymbol{Ax = b} \\ \boldsymbol{x \geq 0} \end{cases} \Rightarrow \begin{cases} \boldsymbol{Ax \leq b} \\ \boldsymbol{Ax \geq b} \\ \boldsymbol{x \geq 0} \end{cases} \Rightarrow \begin{cases} \boldsymbol{Ax \leq b} \\ \boldsymbol{-Ax \leq -b} \\ \boldsymbol{x \geq 0} \end{cases} \Rightarrow$

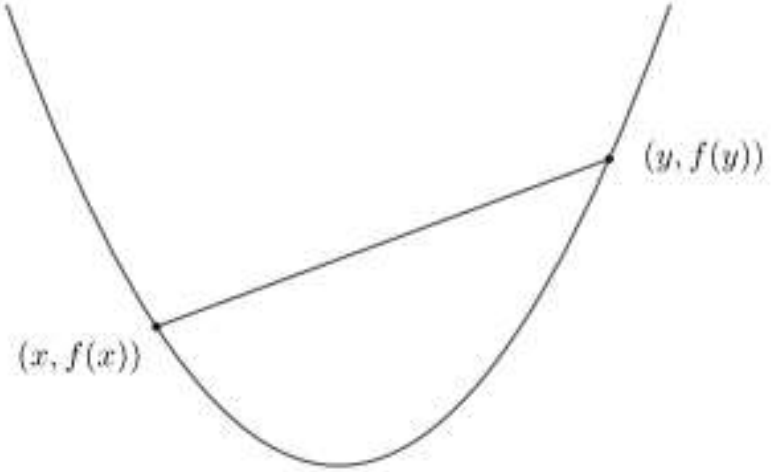
$$\begin{cases} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{-A} \end{pmatrix} \boldsymbol{x} \leq \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{-b} \end{pmatrix} \\ \boldsymbol{x \geq 0} \end{cases}$$

- $\boldsymbol{DP}: \boldsymbol{min}\ \boldsymbol{b^T w; s.t.} \begin{cases} \boldsymbol{A^T w \geq c} \\ \boldsymbol{w}\text{无正负限制} \end{cases}$
- 如何化为非对称的对偶规划?

6.1线性规划-对偶问题(续4)

- 对偶定理LP: $\begin{cases} \max z = c^T x \\ s.t. Ax \le b \\ x \ge 0 \end{cases}$ DP: $\begin{cases} \min f = b^T y \\ A^T y \ge c \\ y \ge 0 \end{cases}$
- 定理(弱对偶定理)：若x, y分别是LP, DP问题的可行解，则$c^T x \le b^T y$
  - 推论(最优性准则定理)：若x^0, y^0分别是LP, DP问题的可行解，当$c^T x^0 = b^T y^0$时,若x^0, y^0分别是LP, DP问题的最优解
  - 推论：若LP有可行解，则LP有最优解的充要条件是DP有可行解
  - 推论：若DP有可行解，则DP有最优解的充要条件是LP有可行解
- 从而利用对偶理论容易判断LP问题是否存在最优解
  - 若LP存在可行解，而其DP问题没有可行解，则LP问题无最优解
  - 若LP存在可行解，而其DP问题也存在可行解，则两个问题都有最优解
- 定理(主对偶定理)：若原规划LP问题有最优解，则对偶规划问题DP也有最优解，反之亦然，并且两者的目标函数值相等

6.3 凸优化基本概念

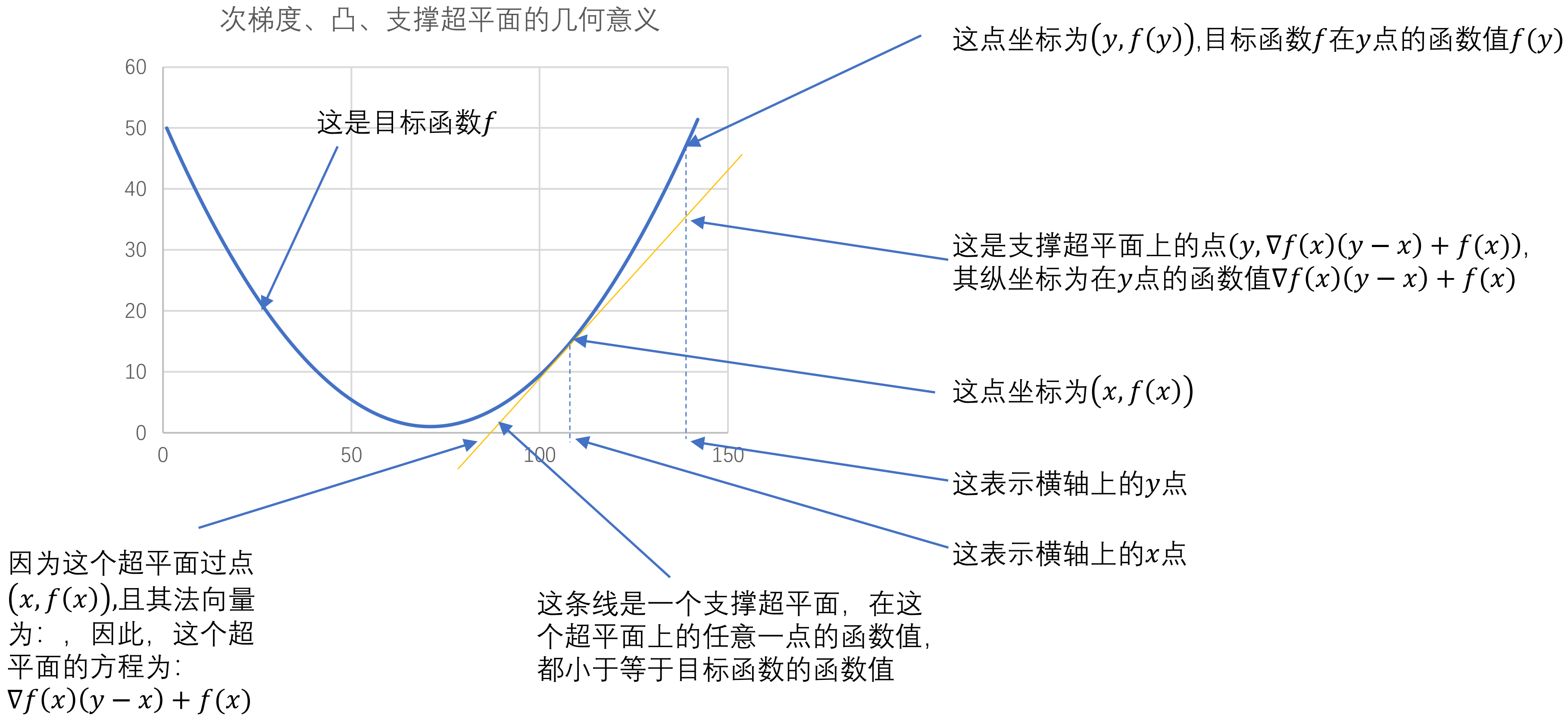


- 凸函数的重要性质
  - 凸函数存在局部最优值，该局部最优值就是全局最优
- 凸集的交集是凸集，凸函数也有一系列的保凸运
  - 凸函数f_i的和$\sum_{i=1}^{m} f_i$总是凸的
  - 凸函数f_i的极大值$f(x) = \max\{f_1(x), f_2(x), \cdots, f_m(x)\}$是凸的
  - 非负凸函数f_1的平方$f(x) = f_1^2(x)$是凸的
  - 凸函数f的仿射变换是凸的：$f(Ax + b)$是凸的
  - 若对每个$y \in domf, f(x, y)$是关于x的凸函数，则$g(x) = \sup_{y \in domf} f(x, y)$是凸函数
  - 给定函数$g: R^n \to R, h: R \to R$，$f(x) = h\big(g(x)\big), g$是凸函数，$h$是凸函数且单调不减，那么$f$是凸函数；若$g$是凹函数，$h$是凸函数且单调不增，那么$f$也是凸函数
  - 给定$g: R^n \to R^k, h: R^k \to R, f(x) = h\big(g(x)\big) = h(g_1(x), g_2(x), \cdots, g_k(x))$，若$g_i$是凹函数，$h$是凸函数且关于每个分量单调不增，那么$f$是凸函数
  - 若$f(x, y)$关于(x, y)整体是凸函数，C是凸集，则$g(x) = \inf_{y \in C} f(x, y)$是凸函数

6.3 凸优化基本概念

这个的几何解释，参看下页的PPT

- (次梯度(subgradient))多元函数$\boldsymbol{f}$定义在非空开凸集$\boldsymbol{S} \subset \boldsymbol{R^n}$上的凸函数，但$\boldsymbol{f}$不一定处处可微。此时采用次梯度的概念来类比梯度的概念。函数$\boldsymbol{f}$在$\boldsymbol{x_0}$处的次梯度是一个向量$\boldsymbol{v}$,如果对于$\boldsymbol{S}$内的任意$\boldsymbol{x}$，都有:$\boldsymbol{f(x) - f(x_0) \geq v^T \cdot (x - x_0)}$成立，所有次梯度的集合称为次微分(subdifferential)，记为$\boldsymbol{\partial f(x_0)}$
  - 从后续凸函数判定的一阶充要条件可以看出，这定义借鉴了这点
- 如果为一维，则可称为次导数(subderivative),例如凸函数$\boldsymbol{f(x) = |x|}$,其在原点的次微分是区间$[\boldsymbol{-1, 1}]$,而在$\boldsymbol{x_0 < 0}$,$\boldsymbol{x_0 > 0}$的次微分分别为单元素集合$\{\boldsymbol{-1}\}, \{\boldsymbol{1}\}$;
  - 又如$\boldsymbol{f(x) = max\{0, \frac{1}{2}(x^2 - 1)\}}$，练习求出其次梯度
- 函数$\boldsymbol{f}$在$\boldsymbol{x_0}$以$\boldsymbol{\nabla f(x_0)}$可微,当且仅当它有$\boldsymbol{\nabla f(x_0)}$作为在$\boldsymbol{x_0}$的唯一次梯度。
  - 注意，次微分是非空的，凸的，和紧的。
- 凸集上的可微函数$\boldsymbol{f}$,其为凸函数等价于：$\boldsymbol{f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y \in dom\ f}$
- 可微函数$\boldsymbol{f}$为凸函数等价于dom$\boldsymbol{f}$为凸，且$\boldsymbol{\nabla f}$单调：$\boldsymbol{\left(\nabla f(x) - \nabla f(y)\right)^T (x - y) \geq 0, \forall x, y \in domf}$
- 复合优化问题$\boldsymbol{min_{x \in R^n} \psi(x) = f(x) + h(x)}$, $\boldsymbol{f}$光滑,$\boldsymbol{h}$为凸，则有一阶必要条件
  - $\boldsymbol{x^*}$为局部极小点，则$\boldsymbol{-\nabla f(x^*) \in \partial h(x^*)}$，其中$\boldsymbol{\partial h(x^*)}$为函数在点$\boldsymbol{x^*}$的次梯度集合



15. 函数$y = |x|, x \in R$在 x=0 点的次微分为 [-1, 1] 。

6.4 约束最优化理论(Theory of Constrained Optimization)

- 约束非线性优化问题P

$$Min\ f(x), s.t. \begin{cases} g_i(x) \le 0, i = 1, 2, \cdots, m \\ h_i(x) = 0, i = 1, 2, \cdots, l \\ x \in X \end{cases} \quad (6\text{-}3)$$

- 如果约束g_i, h_i都是线性的，则称为线性约束优化问题，如果此时目标函数f是二次函数，则称为二次规划问题
- 与以前的线性规划类似，也有可行域等概念
- 如假设x^*是上述问题的一个局部极小点，如果存在$i_0 \in [1, m]$,使得$g_{i_0}(x^*) < 0$,则这时候可以将第i_0个不等式约束去掉，且x^*仍然是去掉这个约束后的问题的局部极小点，称第i_0个约束在x^*处是非积极的.

6.3 凸优化基本概念-共轭函数

- Conjugate function: Fenchel conjugate, 如果$\boldsymbol{f}$可微，又称为 Legendre 变换
  - 令$\boldsymbol{f}: \boldsymbol{A} \to \boldsymbol{R}$是定义在子集$\boldsymbol{A} \subset \boldsymbol{R^n}$上的一个函数。其共轭函数$\boldsymbol{f}^*$:$\boldsymbol{R^n} \to \boldsymbol{R}$定义为：
  - $\boldsymbol{f}^*(\boldsymbol{y}) = \boldsymbol{sup}_{\boldsymbol{x \in A}} \left(\boldsymbol{y^T x} - \boldsymbol{f}(\boldsymbol{x})\right), \boldsymbol{y} \in \boldsymbol{R^n}$
  - 例子：假设$\boldsymbol{f} = \boldsymbol{ax} + \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{R}$,则$\boldsymbol{f}^*(\boldsymbol{y}) = \begin{cases} -\boldsymbol{b}, \boldsymbol{y} = \boldsymbol{a} \\ +\infty, \boldsymbol{otherwise} \end{cases}$

38. 给出函数$f: R^n \to R$的共轭函数$f^*(y)$的定义，并计算$f(x) = e^x, x \in R$的共轭函数。

$$f^*(y) = \sup(yx - f(x))$$

$$f^*(u) = \begin{cases} u\ln u - u, & u > 0 \\ +\infty, & u \le 0 \end{cases}$$

6.5 有约束优化：　对偶

- 线性规划问题的对偶
  - 如果保留$\boldsymbol{x \geq 0}$,直接写等式约束的乘子为$\boldsymbol{y}$,拉格朗日函数应该如何写？
  - $\boldsymbol{L(x,y) = c^T x - y^T(Ax - b) = b^T y + \left(c - A^T y\right)^T x}$
  - 则对偶问题需要将$\boldsymbol{x \geq 0}$添加到约束中
  - $\boldsymbol{max_y \left\{inf_x b^T y + \left(c - A^T y\right)^T x, s.t. x \geq 0\right\}}$,
  - 从而得：
  - $\boldsymbol{max_y b^T y, s.t. A^T y \leq c \quad (*)}$
- 问题：上述（*）式的对偶问题是什么？
  - 等价地：$\boldsymbol{min_y - b^T y, s.t. A^T y \leq c}$
  - 对不等式约束引入拉格朗日乘子$\boldsymbol{x \geq 0}$,则拉格朗日函数为$\boldsymbol{L(y,x) = -b^T y + x^T\left(A^T y - c\right) = -c^T x + (Ax - b)^T y}$
  - 因此对偶函数$\boldsymbol{\theta(x) = inf_y L(y,x) = \begin{cases} -c^T x & Ax = b \\ -\infty & \text{其他} \end{cases}}$
  - 相应的对偶问题是：$\boldsymbol{max_x - c^T x, s.t. Ax = b, x \geq 0}$,这与原问题等价！

18. 非线性规划问题：$\min f(x), s.t.\ g(x) \le 0, h(x)=0, x \in R^n$.请写出该问题的拉格朗日函数，拉格朗日对偶函数，拉格朗日对偶问题。

Lagrangian 函数： $L(x,\lambda,v) = f(x) + \lambda^T g(x) + v^T h(x)$

对偶函数： $G(v,\lambda) = \inf\limits_{x \in D} (f(x) + \lambda^T g(x) + v^T h(x))$

$= \inf\limits_{x \in D} (f(x) + \lambda^T g(x) + v^T h(x))$

对偶问题： maxmize： $G(v,\lambda)$

subject. to： $\lambda \succeq 0$

$$
\begin{cases}
\left.\dfrac{\partial L(x,\lambda^*,\nu^*)}{\partial x}\right|_{x=x^*}=0 & \leftarrow \text{稳定解} \\
\lambda_i^* f_i(x^*)=0 & \leftarrow \text{互补松弛} \\
f_i(x^*)\le 0,\quad h_j(x^*)=0 & \leftarrow \text{LP可行性} \\
\lambda_i^* \geq 0 & \leftarrow \text{DP可行性}
\end{cases}
$$

弱对偶性：$p^* \geq d^*$，即 max-min 不等式：

$$\inf_x \sup_{\lambda, v} L(x, \lambda, v) \geq \sup_{\lambda, v} \inf_x L(x, \lambda, v)$$

强对偶性 $p^* = d^*$

$$\inf_x \sup_{\lambda, v} L(x, \lambda, v) = \sup_{\lambda, v} \inf_x L(x, \lambda, v)$$

$$\parallel \qquad\qquad \parallel$$

$$L(x^*, \lambda^*, v^*)$$

x^*，λ^*，v^* 为 $L(x, \lambda, v)$ 的鞍点

Duality

$$\min_x \quad f_0(x)$$

$$\text{s.t} \quad \underline{f_i(x) \le 0} \qquad 1 \sim m$$

$$h_j(x) = 0 \qquad 1 \sim n$$

Langrangian Function

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} v_j h_j(x)$$

L-D-F

$$g(\lambda, v) = \inf_x L(x, \lambda, v)$$