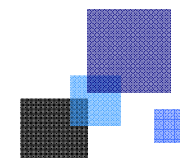


第2章

离散信息的度量



符号约定-离散信源

随机变量和集合：大写字母，例 — X, Y

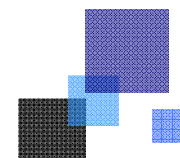
取值空间：以集合表示

随机变量 X 的取值空间 — $A = \{a_1, a_2, \dots, a_n\}$

随机变量实现：小写字母，例 — $x \in A$

概率分布： $P_X(x)$ — 简记为 $p(x)$

$P_{XY}(x, y)$ — 简记为 $p(x, y)$



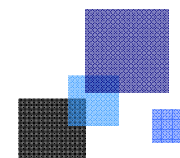
§ 2.1 自信息和互信息

★ 自信息

- 自信息
- 联合自信息
- 条件自信息

★ 互信息

- 互信息
- 互信息的性质
- 条件互信息



§ 2.1.1 自信息

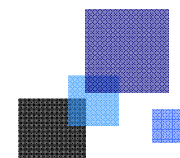
★ 事件集合 X 中的事件 $x = a_i$ 的自信息:

$$I_X(a_i) = -\log P_X(a_i)$$

简记 $I(X) = -\log p(x)$ 或 $I(a_i) = -\log p_i$

其中: 1) $\sum_i p_i = 1$, $0 \leq p_i \leq 1$

2) $I(X)$ 非负 \Rightarrow ? 对数的底数大于1



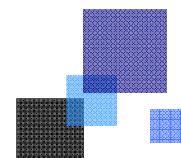
本书符号的约定

集合 $\rightarrow X$

事件 $\rightarrow x$

$x = a_i$ 的概率 $\rightarrow P_x(a_i)$

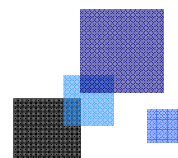
联合概率 $\rightarrow P_{XY}(a_i, b_j)$



关于对数底的选取

$$\left\{ \begin{array}{l} \log_2 x \rightarrow \text{比特} \\ \ln x \rightarrow \text{奈特} \\ \log_{10} x \rightarrow \text{哈特} \end{array} \right.$$

$$\left\{ \begin{array}{l} 1 \text{奈特} = 1.443 \text{比特} \\ 1 \text{哈特} = 3.32 \text{比特} \end{array} \right.$$



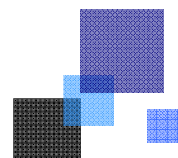
§ 2.1.1 自信息

★自信息为随机变量

★自信息的含义包含两方面：

事件发生前 → 事件发生的不确定性

事件发生后 → 事件包含的信息量



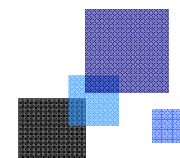


例

2. 1

箱中有90个红球，10个白球。现从箱中随机地取出一个球。求：

- (1) 事件“取出一个红球”的不确定性；
- (2) 事件“取出一个白球”所提供的信息量；
- (3) 事件“取出一个红球”与“取出一个白球”的发生，哪个更难猜测？



解：

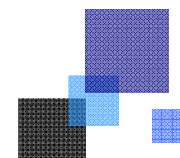
(1) 设 a_1 表示“取出一个红球”的事件，则 $p(a_1) = 0.9$
故事件 a_1 的不确定性为：

$$I(a_1) = -\log 0.9 = 0.152 \text{ 比特}$$

(2) 设 a_2 表示“取出一个白球”的事件，则 $p(a_2) = 0.1$
故事件 a_2 所提供的信息量为：

$$I(a_2) = -\log 0.1 = 3.323 \text{ 比特}$$

(3) 因为 $I(a_2) > I(a_1)$ ，所以事件“取出一个白球”的发生更难猜测。



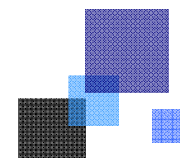
联合自信息

★ 事件集合 XY 中的事件 $x = a_i, y = b_j$ 的自信息:

$$I_{XY}(a_i, b_j) = -\log P_{XY}(a_i, b_j)$$

简记 $I(XY) = -\log p(xy)$

其中: 1) $p(xy)$ 要满足非负和归一化条件

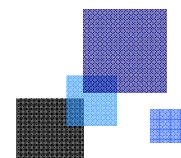




2.1 (续)

箱中球不变，现从箱中随机取出两个球。求：

- (1) 事件“两个球中有红、白球各一个”的不确定性；
- (2) 事件“两个球都是白球”所提供的信息量；
- (3) 事件“两个球都是白球”和“两个球都是红球”的发生，哪个事件更难猜测？



解：

三种情况都是求联合自信息。设x为红球数，y为白球数。

$$(1) \quad P_{XY}(1,1) = \frac{C_{90}^1 C_{10}^1}{C_{100}^2} = \frac{90 \times 10}{100 \times 99 / 2} = 2/11 \quad I(1,1) = -\log 2/11 = 2.460 \text{ 比特}$$

X=1和y=1

$$(2) \quad P_{XY}(0,2) = \frac{C_{10}^2}{C_{100}^2} = \frac{10 \times 9 / 2}{100 \times 99 / 2} = 1/110 \quad I(0,2) = -\log 1/110 = 6.782 \text{ 比特}$$

$$(3) \quad P_{XY}(2,0) = \frac{C_{90}^2}{C_{100}^2} = \frac{90 \times 89 / 2}{100 \times 99 / 2} = 89/110 \quad I(2,0) = -\log 89/110 = 0.306 \text{ 比特}$$

因为 $I(0,2) > I(2,0)$ ，所以事件“两个球都是白球”的发生更难猜测。

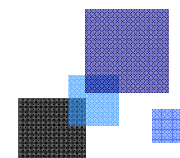
条件自信息

★ 事件 $y = b_j$ 给定, 事件 $x = a_i$ 的自信息:

$$I_{X|Y}(a_i | b_j) = -\log P_{X|Y}(a_i | b_j)$$

简记 $I(X | Y) = -\log p(x | y)$

$p(x | y)$ 要满足非负和归一化条件



★条件自信息的含义包含两方面:

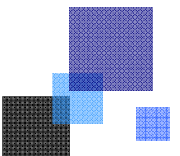
$y = b_j$ 给定, $x = a_i$ 发生前 \rightarrow 事件发生的不确定性

$y = b_j$ 给定, $x = a_i$ 发生后 \rightarrow 事件包含的信息量

★自信息、条件自信息和联合自信息之间的关系

$$I(xy) = I(x) + I(y|x) = I(y) + I(x|y)$$

链式法则



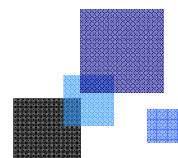


2.1 (续)

箱中球不变，现从箱中先拿出一球，再拿出一球，求：

(1) 事件“在第一个球是红球条件下，第二个球是白球”的不确定性；

(2) 事件“在第一个球是红球条件下，第二个球是红球”所提供的信息量。



解：

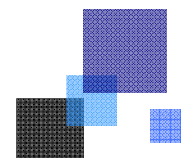
这两种情况都是求条件自信息，设 r 表示红球， w 表示白球。

$$(1) \quad p(y = w \mid x = r) = 10/99$$

$$I(y = w \mid x = r) = -\log 10/99 = 3.307 \text{ 比特}$$

$$(2) \quad p(y = r \mid x = r) = 89/99$$

$$I(y = r \mid x = r) = -\log 89/99 = 0.154 \text{ 比特}$$



例

2.2

有 $8 \times 8 = 64$ 个方格，甲将一棋子放入方格中，让乙猜：

1) 将方格按顺序编号，让乙猜顺序号的困难程度为何？

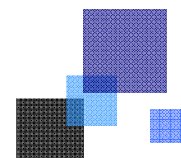
2) 将方格按行和列编号，当甲告诉乙方格的行号后，让乙猜列顺序号的困难程度为何？

解： 两种情况下的不确定性

2个坐标—》2个事件 x, y

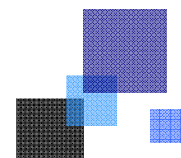
$$1) I(xy) = \log_2 64 = 6 \text{ bit}$$

$$2) I(x|y) = -\log_2 p(x|y) = -\log_2(1/8) = 3 \text{ bit}$$



§ 2.1.2 互信息

- ★ 互信息
- ★ 互信息的性质
- ★ 条件互信息



互信息

★ 离散随机事件 $x = a_i, y = b_j$ 之间的互信息:

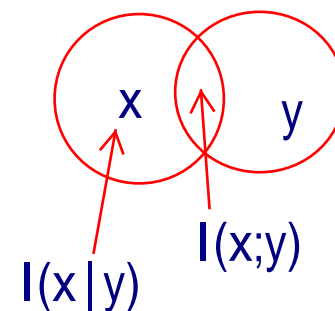
注意：这里没有负号

$$I_{X;Y}(a_i; b_j) = \log \frac{P_{X/Y}(a_i | b_j)}{P_X(a_i)}$$

通过计算

简记 $I(x; y) = \log \frac{p(x/y)}{p(x)}$ 或 $I(a_i; b_j) = \log \frac{p_{ji}}{p_i}$

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



$I(x; y)$ 与 $I(x | y)$ 的区别?

互信息的性质

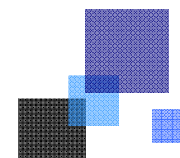
★ 互易性

★ 当事件 x , y 统计独立时, 互信息为0, 即
 $I(x; y) = 0$

★ 互信息可正可负

$\text{Log} p(y|x)/p(y)$,
 $p(y|x) < p(y)$ 时为负

★ 任何两事件之间的互信息不可能大于其中任一事件的自信息



例

2.3

设 e 表示“降雨”， f 表示“空中有乌云”，且
 $P(e)=0.125$ ， $P(e|f)=0.8$

- 求：
- 1) “降雨”的自信息 $I(e)$
 - 2) “空中有乌云”条件下“降雨”的自信息 $I(e|f)$
 - 3) “无雨”的自信息 $I(\bar{e})$
 - 4) “空中有乌云”条件下“无雨”的自信息 $I(\bar{e}|f)$
 - 5) “降雨”与“空中有乌云”的互信息 $I(e;f)$
 - 6) “无雨”与“空中有乌云”的互信息 $I(\bar{e};f)$

解：

1) $I(e) = -\log 0.125 = 3 \text{ bit}$

2) $I(e | f) = -\log 0.8 = 0.322 \text{ bit}$

3) $I(\bar{e}) = -\log 0.875 = 0.193 \text{ bit}$

4) $I(\bar{e} | f) = -\log 0.2 = 2.322 \text{ bit}$

$p(e|f) = 1 - p(\bar{e}|f) = 0.8$

5) $I(e; f) = 3 - 0.322 = 2.678 \text{ bit}$

$I(e) - I(e|f)$

6) $I(\bar{e}; f) = 0.193 - 2.322 = -2.129 \text{ bit}$

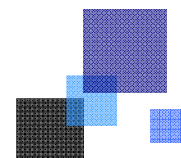
互信息为负——空中有乌云
不利于无雨

条件互信息

★ 设联合集XYZ，在给定 $z \in Z$ 条件下 $x (\in X)$ 与 $y (\in Y)$ 之间的互信息定义为：

$$I(x; y / z) = \log \frac{p(x / yz)}{p(x / z)}$$

除条件外，条件互信息的含义与互信息的含义与性质都相同

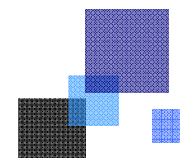


§ 2.2 信息熵

★信息熵的定义与计算

★条件熵与联合熵

★熵的基本性质



信息熵的定义与计算

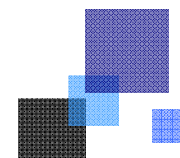
★ 离散信源X的熵定义为自信息的平均值, 记为 $H(X)$

$$H(X) = E_{p(x)}[I(x)] = -\sum_x p(x) \log p(x)$$

→ $I(x)$ 为事件 x 的自信息

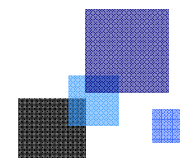
→ $E_{p(x)}$ 表示对随机变量 x 用 $p(x)$ 来进行取平均运算

→ 熵的单位为比特（奈特） / 信源符号



信息熵 $H(X)$ 的含义

- ★ 信源输出前→ 信源的平均不确定性
- ★ 信源输出后→ 一个信源符号所提供的平均信息量
或编码一个信源符号的平均长度
- ★ 表示信源随机性大小： $H(X)$ 大的，随机性大
- ★ 信源输出后，不确定性就解除→ 解除信源不确定性所需信息量



例

2.4

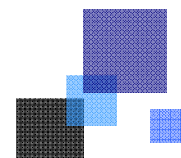
一电视屏幕的格点数为 $500 \times 600 = 300000$ ，每点有10个灰度等级，若每幅画面等概率出现，求每幅画面平均所包含的信息量

解：可能的画面数是多少？ $10^{300000} \Rightarrow p = \frac{1}{10^{300000}}$

代入公式：

出现每幅画面的概率

$$H(X) = \log_2(1/p) = \log_2(10^{300000}) = 10^6 \text{ bit}$$

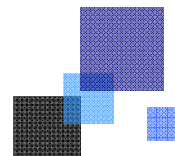


例 2.5

A、B两城市天气情况概率分布如下表：

	晴	阴	雨
A城市	0. 8	0. 15	0. 05
B城市	0. 4	0. 3	0. 3

问哪个城市的天气具有更大的不确定性？

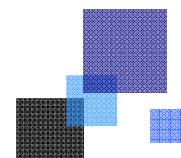


解：

$$\begin{aligned} H(A) &= H(0.8, 0.15, 0.05) = -0.8 \times \log 0.8 - 0.15 \times \log 0.15 - 0.05 \times \log 0.05 \\ &= 0.884 \text{ 比特/符号} \end{aligned}$$

$$\begin{aligned} H(B) &= H(0.4, 0.3, 0.3) = -0.4 \times \log 0.4 - 0.3 \times \log 0.3 - 0.3 \times \log 0.3 \\ &= 1.571 \text{ 比特/符号} \end{aligned}$$

所以，B城市的天气具有更大的不确定性。



例

2.6

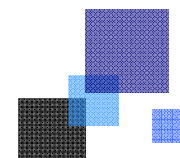
有甲、乙两箱球，甲箱中有红球50、白球20、黑球30；乙箱中有红球90、白球10。现做从两箱中分别随机取一球的实验，问从哪箱中取球的结果随机性更大？

解： 设A、B分别代表甲、乙两箱，则

$$\begin{aligned} H(A) &= H(0.5, 0.2, 0.3) = -0.5 \times \log 0.5 - 0.2 \times \log 0.2 - 0.3 \times \log 0.3 \\ &= 1.486 \text{ 比特/符号} \end{aligned}$$

$$H(B) = H(0.9, 0.1) = -0.9 \times \log 0.9 - 0.1 \times \log 0.1 = 0.469 \text{ 比特/符号}$$

所以，从甲箱中取球的结果随机性更大。

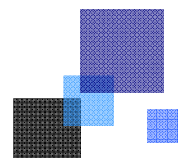


信息熵的计算

★定理2.1 离散信源的熵等于所对应的有根概率树上所有节点(包括根节点, 不包括叶)的分支熵用该节点概率加权的和, 即

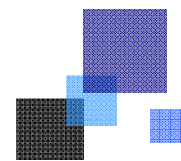
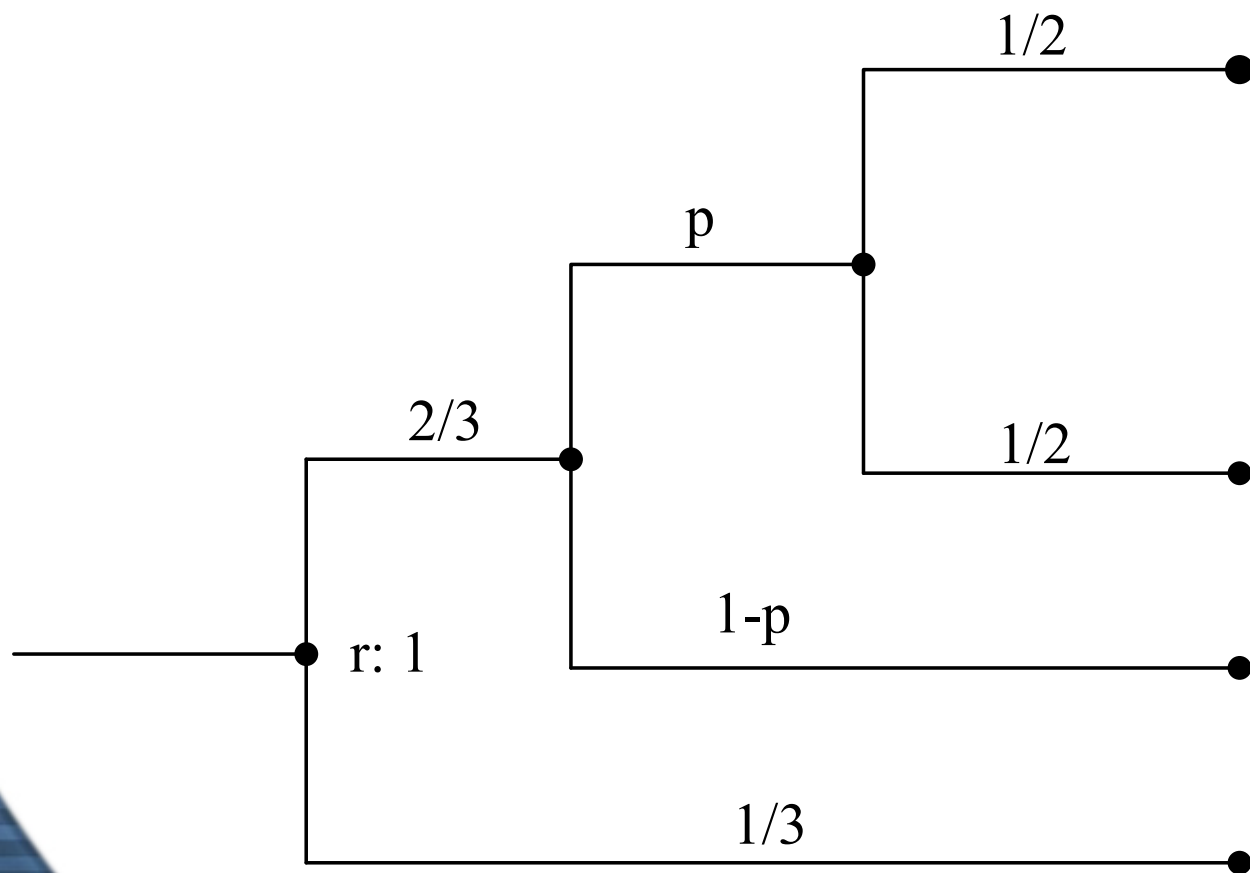
$$H(X) = \sum_i q(u_i) H(u_i)$$

其中, $q(u_i)$ 为节点 u_i 的概率, $H(u_i)$ 为节点 u_i 的分支熵。





2. 6



条件熵

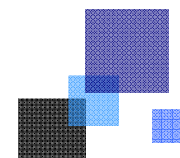
★ 条件熵：联合集XY上，条件自信息 $I(y|x)$ 的平均值

$$\begin{aligned} H(Y/X) &= E_{p(xy)} [I(y/x)] \\ &= -\sum_x \sum_y p(x, y) \log p(y/x) \\ &= \sum_x p(x) \left[-\sum_y p(y/x) \log p(y/x) \right] \\ &= \sum_x p(x) H(Y/x) \end{aligned}$$

遍历x取平均

$$H(Y/x) = -\sum_y p(y/x) \log p(y/x)$$

为在x取某一特定值时 Y的熵



例

2.7

随机变量X和Y，符号集均为{0, 1} $p_x(0) = \frac{2}{3}$ $p_x(1) = \frac{1}{3}$

$$p(y=0|x=0) = p(y=1|x=0) = \frac{1}{2} \quad p(y=1|x=1) = 1 \quad \text{求} H(Y|X)$$

解： $H(Y|X) = \sum_x p(x)H(Y|x) = p(x=0)H(Y|x=0) + p(x=1)H(Y|x=1)$

X=1时，Y确定，
信息量为0

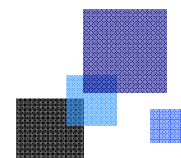
$$= \frac{2}{3}H\left(\frac{1}{2}\right) + \frac{1}{3}H(1) = \frac{2}{3} \text{ 比特/符号}$$

联合熵

★ 联合熵：联合集XY上，对联合自信息 $I(xy)$ 的平均值

$$H(XY) = E_{p(xy)} [I(x, y)]$$

$$= - \sum_x \sum_y p(x, y) \log p(x, y)$$



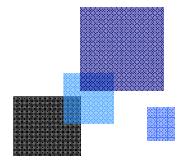
例 2.7 (续)

求 $H(XY)$

解：由已知条件可得 XY 的联合概率分布，如下表所示

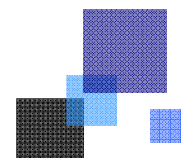
		Y	
		0	1
X	0	$\frac{1}{3}$	$\frac{1}{3}$
	1	0	$\frac{1}{3}$

$$H(XY) = H\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = \log 3 = 1.585 \text{ 比特/符号}$$



熵的基本性质

- ★ 凸函数
- ★ 信息散度
- ★ 熵的基本性质
- ★ 各类熵的关系
- ★ 熵函数的唯一性



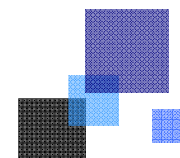
凸函数

$$\left. \begin{array}{l} H(X) = H(p) = -\sum p_i \log p_i \\ \sum p_i = 1 \end{array} \right\} \Rightarrow H(X) \text{ 为 } n-1 \text{ 元函数}$$

$$\text{当 } n=2 \Rightarrow H(p) = H(p_1, p_2) = H(p_1, 1-p_1) = H(p_1)$$

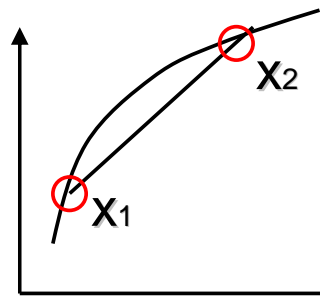
下面我们来定义凸函数

LOOK一下



★ 对于 α ($0 \leq \alpha \leq 1$) 及任意两矢量 x_1, x_2 , 有

$$f[\alpha x_1 + (1-\alpha)x_2] \geq \alpha f(x_1) + (1-\alpha)f(x_2) \quad \text{上凸函数 (cap)}$$

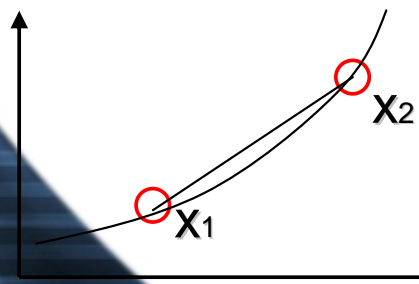


若当且仅当 $x_1 = x_2$ 或 $\alpha = 0, 1$ 时等式成立 \Rightarrow

严格上凸函数

★ 对于 α ($0 \leq \alpha \leq 1$) 及任意两矢量 x_1, x_2 , 有

$$f[\alpha x_1 + (1-\alpha)x_2] \leq \alpha f(x_1) + (1-\alpha)f(x_2) \quad \text{下凸函数 (cup)}$$

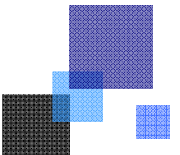


若当且仅当 $x_1 = x_2$ 或 $\alpha = 0, 1$ 时等式成立 \Rightarrow

严格下凸函数

说明:

- ✓ 在很多数学或信息论专著中，需要注意凸和凹的使用含义；
- ✓ 函数 $f(x)$ 是凸的 (convex) 《——》 本书中的下凸 (cup)
- ✓ 函数 $f(x)$ 是凹的 (concave) 《——》 本书中的上凸 (cap)



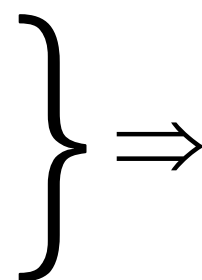
★ 定理：

注意：lambda为小于等于1的正数

$f(x)$ 是区间上的实值连续严格上凸函数

任意一组 x_1, x_2, \dots, x_q

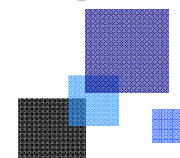
$\lambda_1, \lambda_2, \dots, \lambda_q, \sum \lambda_k = 1$



Jenson不等式

$$f\left[\sum_{k=1}^q \lambda_k x_k\right] \geq \sum_{k=1}^q \lambda_k f(x_k)$$

当且仅当 $x_1 = x_2 = \dots = x_q$ 或 $\lambda_k = 1$ ($1 \leq k \leq q$) 且 $\lambda_j = 0$ ($j \neq k$) 时，等式成立



证 利用数学归纳法。根据上凸函数的定义有

$$f[\alpha x_1 + (1-\alpha)x_2] \geq \alpha f(x_1) + (1-\alpha)f(x_2)$$

其中 $0 < \alpha < 1$, 即 $q=2$ 时成立。

今假定 $q=n$ 成立。现考虑 $q=n+1$ 的情况

设 $\lambda_k \geq 0, \sum_{k=1}^{n+1} \lambda_k = 1$, 令 $\alpha = \sum_{k=1}^n \lambda_k$, 则 $\lambda_{n+1} = 1 - \alpha$,

$$\sum_{k=1}^{n+1} \lambda_k f(x_k) = \sum_{k=1}^n \lambda_k f(x_k) + \lambda_{n+1} f(x_{n+1})$$

$$= \alpha \sum_{k=1}^n (\lambda_k / \alpha) f(x_k) + \lambda_{n+1} f(x_{n+1})$$

利用n时的归纳假定

$$\leq \alpha f\left[\sum_{k=1}^n (\lambda_k / \alpha) x_k\right] + \lambda_{n+1} f(x_{n+1})$$

二次利用上凸函数的性质

$$\leq f\left[\alpha \sum_{k=1}^n (\lambda_k / \alpha) x_k + \lambda_{n+1} x_{n+1}\right]$$

$$= f\left[\sum_{k=1}^{n+1} \lambda_k x_k\right]$$

当且仅当
 $x_1 = x_2 = \dots = x_q$ 或
 $\lambda_k = 1$ ($1 \leq k \leq q$) 且 $\lambda_j = 0$ ($j \neq k$) 时, 等式成立。

- ★ 特别地，当 x_k 为离散信源符号的取值， λ_k 为相应的概率， $f(x)$ 为对数函数时，有

$$E[\log(x)] \leq \log[E(x)]$$

对数是上凸函数

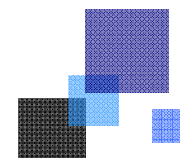
- ★ 对于一般的凸函数有

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

★ 上凸函数的判定方法

1) 在某区间的二阶导数小于0, 则在此区间内为严格上凸函数。

2) 利用Jenson不等式 $f[\sum_{k=1}^q \lambda_k x_k] \geq \sum_{k=1}^q \lambda_k f(x_k)$



一个在以后的公式推导中反复使用的不等式：

对于任意 x ，有：

$$1 - \frac{1}{x} \leq \ln x \leq x - 1$$

这是怎么得来的？

设 $f(x) = \ln x - x + 1 \Rightarrow \begin{cases} \text{① } x=1 \text{ 为稳定点} \\ \text{② } x=1 \text{ 时, 2阶导数小于0} \end{cases} \Rightarrow x=1 \text{ 处有极大值}$

$y = \frac{1}{x}$ 代入等式 $\Rightarrow 1 - \frac{1}{y} \leq \ln y$ y换成x

不等式应用：信息论中，很多场合带有对数符号，使用该不等式去掉对数符号

$$f'(x) = (1/x) - 1 = 0 \\ \rightarrow x=1$$

$f''(x) < 0 \rightarrow$ 上凸
最大值：1

信息散度

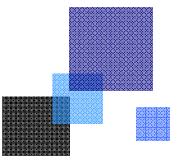
★ P和Q为定义在同一概率空间的两个概率测度，则P相对于Q的散度：

$$D(P // Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

注意本章仅讨论离散信息

注意：散度不对称，
即 $D(P || Q) \neq D(Q || P)$

上式中，概率分布的维数不限，可以是一维，也可以是多维。



★ 定理： 如果在一个共同的有限字母表(指标集)的概率空间上给定的两个概率测度 $P(x)$ 和 $Q(x)$

$$D(P // Q) \geq 0$$

当且仅当对所有 x , $P(x) = Q(x)$ 时,等式成立

$\log x$ 是上凸函数

$$\begin{aligned} -D(P // Q) &= \sum_x P(x) \log \frac{Q(x)}{P(x)} \leq \log \left[\sum_x P(x) \frac{Q(x)}{P(x)} \right] \\ &\leq \log \left[\sum_x Q(x) \right] = 0 \end{aligned}$$

信息散度的物理含义：

对于服从 P 的信源 X 采用基于分布 Q 的编码平均所需要的额外比特数

1) 对于信源 X 服从 $P(X)$ ，符号 x 的最优编码长度为：

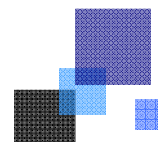
$$\log \frac{1}{P(x)}, \text{ 平均编码长度为: } H(x) = \sum P(x) \log \frac{1}{P(x)}$$

2) 符号 x 采用分布 $Q(x)$ 的编码长度为： $\log \frac{1}{Q(x)}$ ，平均编

$$\text{码长度为: } \sum P(x) \log \frac{1}{Q(x)}$$

3) 编码的额外长度为： $\sum P(x) \log \frac{1}{Q(x)} - \sum P(x) \log \frac{1}{P(x)}$

$$= \sum P(x) \log \frac{P(x)}{Q(x)} = D_{KL}(P \parallel Q)$$



熵的基本性质 (1)

★ 对称性: $\mathbf{p}=(p_1,p_2,\dots,p_n)$ 中, 各分量的次序可以任意改变

★ 非负性: 自信息非负, 熵为自信息的平均 \Rightarrow 熵非负

★ 扩展性: $\lim_{\varepsilon \rightarrow 0} \varepsilon \log \varepsilon = 0 \Rightarrow \lim_{\varepsilon \rightarrow 0} H_{n+1}(p_1, p_2, \dots, p_n - \varepsilon, \varepsilon) = H_n(p_1, p_2, \dots, p_n)$
即: 小概率事件对熵的影响很小, 可以忽略

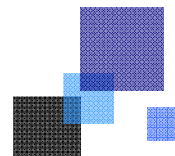
★ 可加性: $H(XY) = H(X) + H(Y|X)$

$$H(X_1X_2\dots X_N) = H(X_1) + H(X_2|X_1) + \dots + H(X_N|X_1\dots X_{N-1})$$

复合事件集合的不确定性为各个分事件集合的不确定性的和

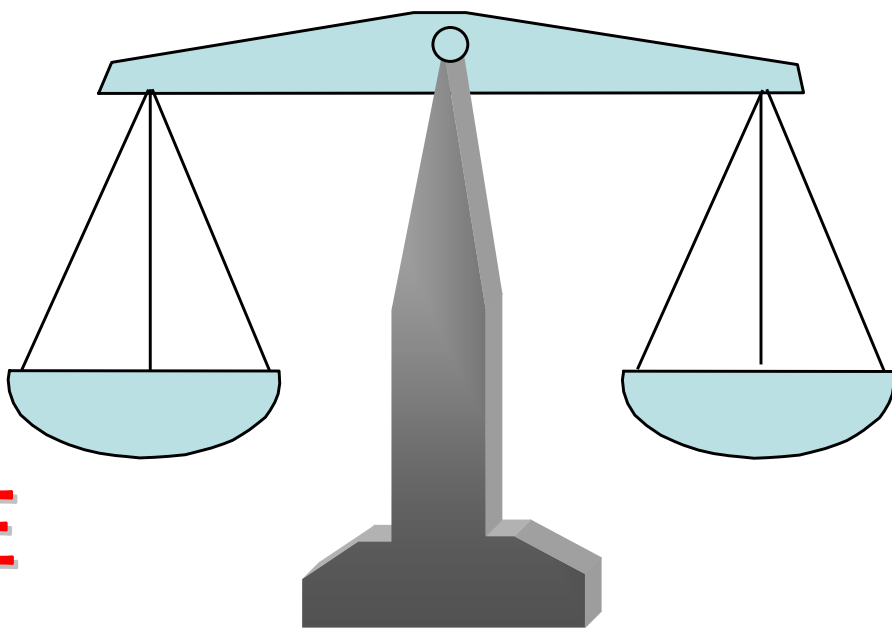
说明:

- ✓ 熵的对称性 — 改变概率矢量中各分量的顺序，熵不改变
- ✓ 熵的非负性 — 只适用于离散信源，对连续信源不成立
- ✓ 熵的可加性 — 熵的链式法则 (Chain Rule)

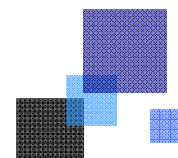
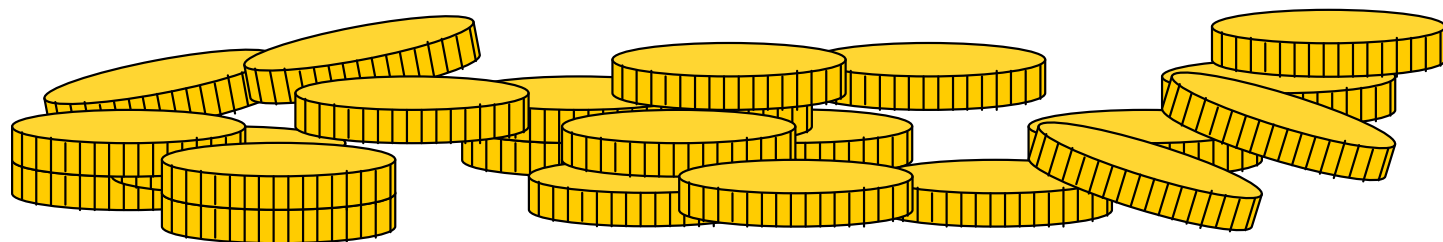


熵的链原则举例

找找假币在哪里？



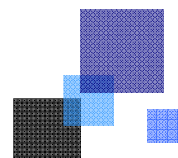
1. 12个外形相同的硬
币，有一个重量不同
的假币
2. 天平称重：平衡、左
倾、右倾
3. 最少称几次可鉴别假
币？



解题要点:

- ✓ 据熵的链式法则，复合事件的不确定性可通过多次实验去除
- ✓ 每次实验所获得信息量最大，则总实验次数最少
- ✓ 天平一次称重信息量 $\log 3$ ， k 次为 $k\log 3$
- ✓ 12个硬币，每个有真/假2状态，共24种可能状态，信息量为 $\log 24$
- ✓ 注意到 $3\log 3 = \log 27 > \log 24$ ，故理论上最少3次

$$2\log 3 = \log 9 < \log 24$$



熵的基本性质 (2)

★ 极值性:

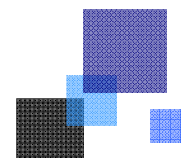
定理2.4.3 — 离散最大熵定理

对于离散随机信源，当其中的事件等概率发生时，熵达到最大值

$$H(X) \leq \log n$$

证 设随机变量集合有 n 个符号，概率分布为 $P(x)$ ； $Q(x)$ 为等概率分布，即 $Q(x)=1/n$ 。根据散度不等式有

$$\begin{aligned} D(P // Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_x P(x) \log P(x) - \sum_x P(x) \log(1/n) \\ &= -H(X) + \log n \geq 0 \\ \Rightarrow H(X) &\leq \log n \end{aligned}$$



熵的基本性质 (3)

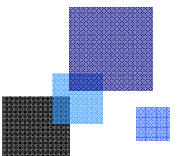
★ 确定性:

$$H(1,0) = H(1,0,0) = \dots = H(1,0,\dots,0) = 0。$$

当随机变量集合中任一事件概率为1时，熵为0

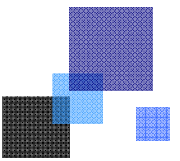
★ 上凸性:

$H(p)=H(p_1,p_2,\dots,p_n)$ 是 (p_1,p_2,\dots,p_n) 的严格的上凸函数



有关熵严格上凸性的说明:

- ✓ 熵函数 $H(P)$ 是概率矢量 $P=(p_1, p_2, \dots, p_n)$ 的严格上凸函数
- ✓ 对任意概率分布 P_1 和 P_2 ，及任意 $0 < a < 1$ ，有
$$H[aP_1 + (1-a)P_2] > aH(P_1) + (1-a)H(P_2)$$
- ✓ 说明存在一个概率分布 P ，使信源熵最大。对离散信源，该分布是均匀分布；对于连续信源，该分布是高斯
- ✓ 证明留给大家



各类熵的关系

★ 条件熵不大于信息熵:

定理 (熵的不增原理)

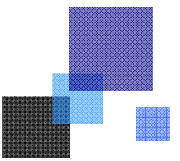
$$H(Y | X) \leq H(Y)$$

$$H(Y) - H(Y | X) = - \sum_Y q(y) \log q(y) + \sum_x \sum_y p(x) p(y/x) \log p(y/x)$$

$$q(y) = p(x) p(y/x)$$

$$= \sum_x p(x) \sum_y p(y/x) \log \frac{p(y/x)}{q(y)}$$

$$= \sum_x p(x) D(p(y/x) // q(y)) \geq 0$$



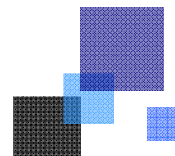
熵不增加原理 — 证明要点和讨论

1、 $q(y) = \sum_x p(x)p(y|x)$

2、 $\sum_x p(y|x) \log \frac{p(y|x)}{q(y)} = D[p(y|x) \| q(y)] \geq 0$

3、 $I(X;Y) = H(Y) - H(Y|X) \geq 0$ 这里是平均互信息

4、 信息处理中，条件越多，熵越小



各类熵的关系

★ 联合熵不大于个信息熵的和:

$$H(X_1 X_2 \cdots X_N) \leq \sum_{i=1}^N H(X_i)$$

联合信源编码可以获得高的编码效率

证明思路: 熵的可加性和熵不增加原理 $\} \Rightarrow$
 $H(Y | X) \leq H(Y)$

★ 联合熵与信息熵、条件熵的关系

$$H(XY) = H(X) + H(Y/X)$$

熵函数的唯一性

如果要求熵函数满足以下条件：

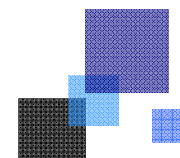
★ 是概率的连续函数

$$H(X) = \log n$$

★ 信源符号等概率时是 n （信源符号数）的增函数

★ 可加性或链式法则

那么，以概率矢量为变量的熵函数的表示是唯一的。



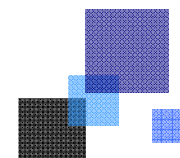
§ 2.3 平均互信息

★ 平均互信息的定义

涉及二个随机变量之间

★ 平均互信息的性质

★ 平均条件互信息



集合与事件之间的互信息

★ 集合X与事件 $y=b_j$ 之间的互信息:

$$I(X; y) = \sum_x P(x|y) \log \frac{P(x|y)}{p(x)}$$

2个消息x和y间的互信息：
 $\log(p(x|y)/p(x))$

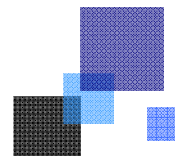
由事件 $y=b_j$ 提供的关于集合X的**平均条件互信息**（用条件概率平均）

定理 $I(X; y) \geq 0$,

仅当y与所有x独立时，等式成立。

证 根据散度的定义，有 $I(X; y) = D(P_{X/y} // P_X) \geq 0$

仅当对所有x， $p(x) = p(x/y)$ 时，等式成立，证毕。



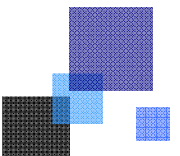
有关集合X和事件y互信息的说明:

1、事件x和y互信息的定义

$$I(x, y) = \log \frac{p(x/y)}{p(x)} \text{ 或 } I(a_i; b_j) = \log \frac{p_{ji}}{p_i}$$

2、对条件概率 $P(x|y)$ 中x取平均, 有:

$$I(X; y) = \sum_x P(x|y) \log \frac{P(x|y)}{p(x)}$$



平均互信息

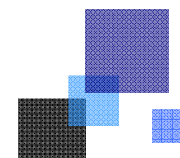
★ 集合X、Y之间的平均互信息：

$$I(X;Y) = \sum_x p(x) I(Y;x)$$

$$= \sum_{x,y} p(x) p(y/x) \log \frac{p(y/x)}{p(y)}$$

$$= \sum_{x,y} p(x) p(y/x) \log \frac{p(y/x)}{\sum_x p(x) p(y/x)}$$

$$= \sum_{i,j} p_i p_{ij} \log \frac{p_{ij}}{\sum_i p_i p_{ij}}$$

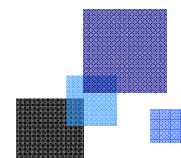
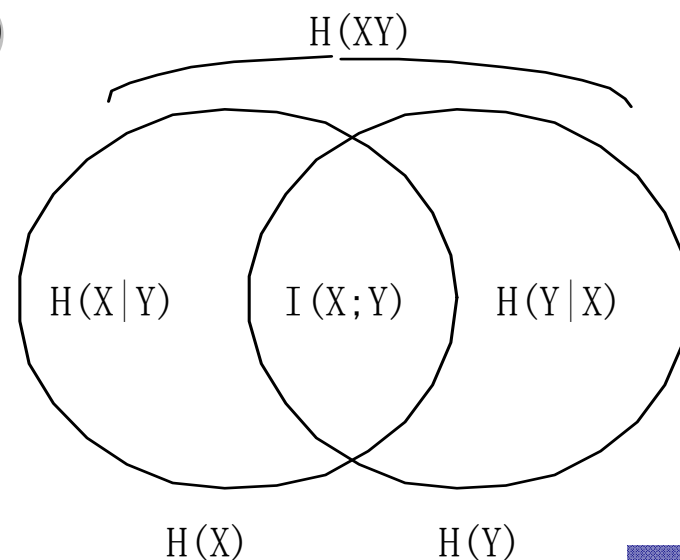


平均互信息与熵的关系

★ $I(X; Y) = H(X) - H(X|Y)$

★ $I(X; Y) = H(Y) - H(Y|X)$

★ $I(X; Y) = H(X) + H(Y) - H(XY)$



平均互信息的性质

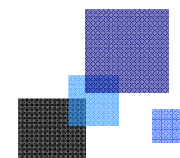
★ 非负性:

定理： $I(X;Y) \geq 0$ 仅当 X, Y 独立时，等式成立

证明思路： $I(X; y) \geq 0 \Rightarrow$ 其均值 ≥ 0

★ 互易性(对称性): $I(X;Y) = I(Y;X)$

★ 凸函数性



★ 定理:

$I(X;Y)$ 为概率分布 $p(x)$ 的上凸函数。

证明思路:

- 1、设 $p(y|x)$ 固定, 用于描述特定信道
- 2、任取二种输入分布 $p_1(x)$ 和 $p_2(x)$, $0 < \theta < 1$, 构造 $p(x) = \theta p_1(x) + (1-\theta)p_2(x)$, 验证 $p(x)$ 满足概率分布条件 $\sum p(x) = 1$, 记 $I(X;Y) = I[p(x)]$
- 3、利用 Jensen 不等式证明:
$$I[\theta p_1(x) + (1-\theta)p_2(x)] \geq \theta I[p_1(x)] + (1-\theta)I[p_2(x)]$$

例

二元信源 X 输出符号为 $\{0, 1\}$ ， $P_X(0)=\omega$ ，条件概率分别为 $P_{Y|X}(0|0) = P_{Y|X}(1|1)=1-p$ ， $P_{Y|X}(1|0)=P_{Y|X}(0|1)=p$ ，求 $I(X;Y)$ 。

二元对称信道

解：

$$\left. \begin{array}{l} P_Y(0) \rightarrow q(0) \\ P_Y(1) \rightarrow q(1) \end{array} \right\} \Rightarrow [q(0) \quad q(1)] = [\omega \quad 1-\omega] \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

$$q(0) = \omega(1-p) + (1-\omega)p = p + \omega - 2\omega p$$

$$q(1) = \omega p + (1-\omega)(1-p) = \omega p + (1-\omega) - (1-\omega)p = 1 - \omega - p + 2\omega p$$

$$\Rightarrow H(Y) = H(p + \omega - 2\omega p)$$

$x=0 \rightarrow y=0:1-p; y=1:p$

$x=1 \rightarrow y=0:p; y=1:1-p$

$$\begin{aligned} H(Y|X) &= \omega[-(1-p)\log(1-p) - p\log p] + (1-\omega)[-p\log p - (1-p)\log(1-p)] \\ &= -p\log p - (1-p)\log(1-p) = H(p) \end{aligned}$$

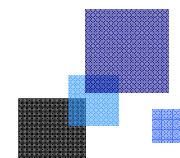
$$\begin{aligned} \Rightarrow I(X;Y) &= H(Y) - H(Y|X) \\ &= H(p + \omega - 2\omega p) - H(p) \end{aligned}$$

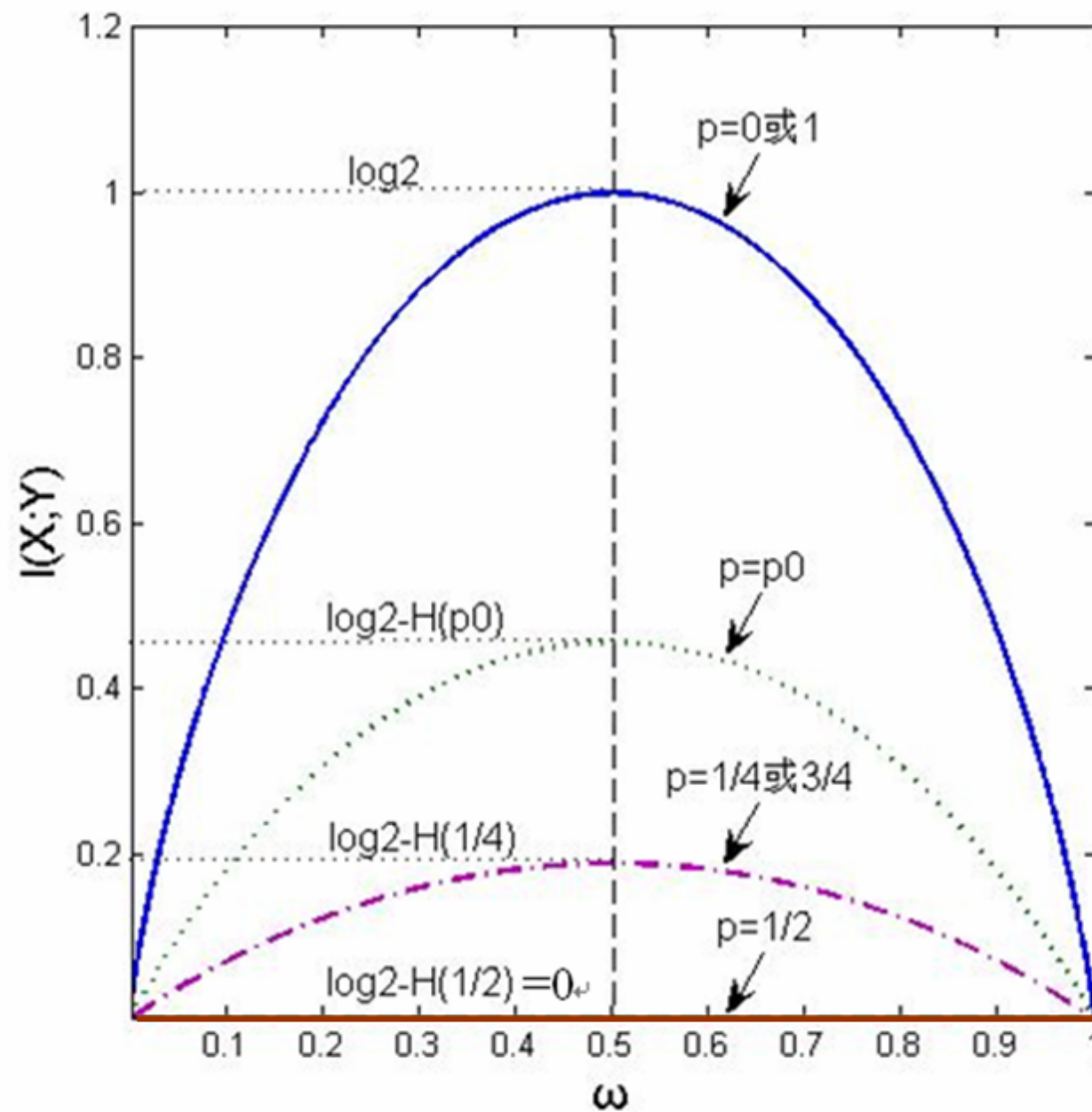
- 1. 2个上凸的和 \rightarrow 上凸，有最大值
- 2. 最大值取决于第一项
- 3. 输入为等概率分布时，互信息最大

可见：

- 1) 这是一个上凸函数，注意区别输入分布和信道误码率；
- 2) $p + \omega - 2\omega p = \frac{1}{2}$ 移项 $\Rightarrow (1-2\omega)(p-1/2) = 0$

$$\Rightarrow \begin{cases} \text{当 } p \neq \frac{1}{2}, \omega = \frac{1}{2} \text{ 时, 有极大值} \\ \text{当 } p = \frac{1}{2}, I(X;Y) = H(\frac{1}{2}) - H(\frac{1}{2}) = 0 \end{cases}$$





★ 定理:

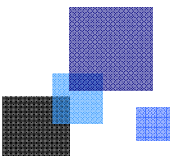
对于固定的概率分布 $p(x)$, $I(X;Y)$ 为条件概率 $p(y|x)$ 的下凸函数。

证明思路: 设 $p_1(y|x), p_2(y|x)$

令 $0 < \theta < 1$, 取 $p(y|x) = \theta p_1(y|x) + (1-\theta)p_2(y|x)$

$$\Rightarrow \sum_y p(y|x) = 1 \quad \text{记 } I(X;Y) = I[p(y|x)]$$

那么只要证 $I[p(y|x)] \leq \theta I[p_1(y|x)] + (1-\theta)I[p_2(y|x)]$



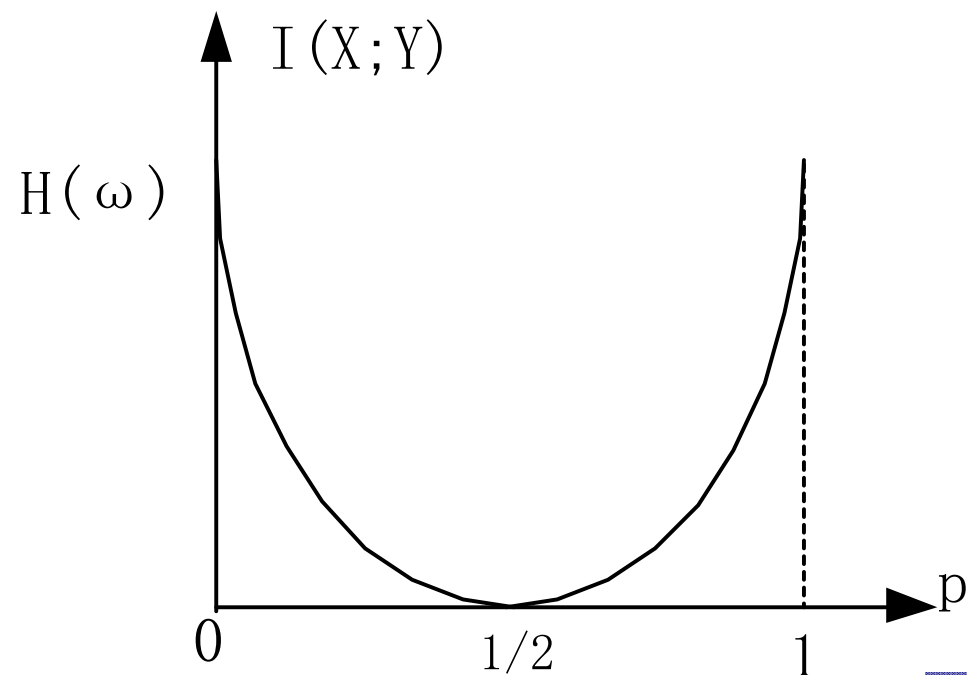
ω

例

续

$$I(X;Y) = H(p + \omega - 2\omega p) - H(p), \quad \omega \text{ 固定}$$

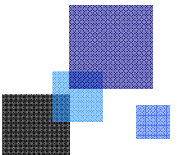
解：这是一个下凸函数



★ 极值性:

$$I(X; Y) \leq H(X)$$

$$I(X; Y) \leq H(Y)$$



平均条件互信息

★ 设联合集XYZ，Z条件下，X与Y之间的平均互信息：

$$I(X;Y|Z) = E_{p(xyz)} \left\{ \log \frac{p(x|yz)}{p(x|z)} \right\} = \sum_{x,y,z} p(xyz) \log \frac{p(x|yz)}{p(x|z)}$$

链式法则

$$\text{由于 } I(x;yz) = \log \frac{p(x|yz)}{p(x)} = \log \frac{p(x|yz)}{p(x|z)} \frac{p(x|z)}{p(x)} = I(x;y|z) + I(x;z)$$

同理得—》

$$I(x;yz) = I(x;z/y) + I(x;y)$$

两边求平均，得： $I(X;YZ) = I(X;Z/Y) + I(X;Y) = I(X;Y/Z) + I(X;Z)$

★ 定理：平均条件互信息是非负的：

$$I(X; Y / Z) \geq 0$$

当且仅当 $p(x|z) = p(x|yz)$ ，等式成立

证明：

$$\begin{aligned} I(X; Y / Z) &= \sum_{x,y,z} p(xyz) \log \frac{p(x|yz)}{p(x|z)} \\ &= \sum_{x,y,z} p(xyz) \log \frac{p(xyz)}{p(x|z)p(yz)} = D(P_{XYZ} // P_{X/Z} \times P_{YZ}) \geq 0 \end{aligned}$$

构造KL散度D，并利用
 $D \geq 0$ 是证明类似问题的
常用方法

当且仅当 $p(x|z) = p(x|yz)$ 时， $I(X; Y / Z) = D(P_{XYZ} // P_{X/Z} \times P_{YZ}) \geq 0$ 等式成立

★ 定理:

$$I(X;YZ) \geq I(X;Z)$$

表明给定Z时, 有关Y的信息对X无贡献

当且仅当 $p(x|z) = p(x|yz)$ 时等式成立

$$I(X;YZ) \geq I(X;Y)$$

当且仅当 $p(x|y) = p(x|yz)$ 时等式成立

定理证明依据: $I(X;YZ) = I(X;Z/Y) + I(X;Y) = I(X;Y/Z) + I(X;Z)$

设Y、Z为独立随机变量集合，其中Y含n个事件，Z含k个事件，则联合集YZ含nk个事件。Z集合可看成YZ集合中某些事件的合并处理， $\{yz\}_y \rightarrow \{z\}$ ，于是nk个事件合并成k个事件。

- 1) 随机事件合并后，获得的信息量减少， $\log(nk) \rightarrow \log k$
- 2) 如果YZ为二维取值空间，则Z的取值空间是对YZ取值空间的合并；而YZ取值空间是对Z或Y取值空间的细化。可见，通过对取值空间的细化，可使获得的信息量增加。

本章小结

★ 自信息的平均值为熵 $H(X) = E_{p(x)} [-\log p(x)]$

★ 条件自信息的平均值为条件熵 $H(X/Y) = E_{p(xy)} [-\log p(x/y)]$

★ 联合自信息的平均值为联合熵 $H(XY) = E_{p(xy)} [-\log p(xy)]$

★ 互信息的平均值为平均互信息 $I(X;Y) = E_{p(xy)} [\log \frac{p(y/x)}{q(y)}]$

★ 条件互信息的平均值为平均条件互信息

$$I(X;Y/Z) = E_{p(xyz)} [\log \frac{p(y/xz)}{q(y/z)}]$$

本章小结

★ 熵的可加性 $H(X_1 X_2 \cdots X_n) = H(X_1) + H(X_2 / X_1) + \cdots + H(X_n / X_1 \cdots X_{n-1})$

★ 平均互信息与熵的关系

$$\begin{aligned} I(X; Y) &= H(X) - H(X / Y) \\ &= H(Y) - H(Y / X) \\ &= H(X) + H(Y) - H(XY) \end{aligned}$$

★ 离散熵与平均互信息都具有非负性

★ 离散最大熵定理 $H(X) = \log n$ (n 为信源符号数)

★ 平均互信息的凸函数性质

对输入

对信道

平均互信息为输入概率的上凸函数，为条件概率的下凸函数

课后习题

Page 35

2.6、2.7、2.10

