第2章

离散信息的度量

上节回顾

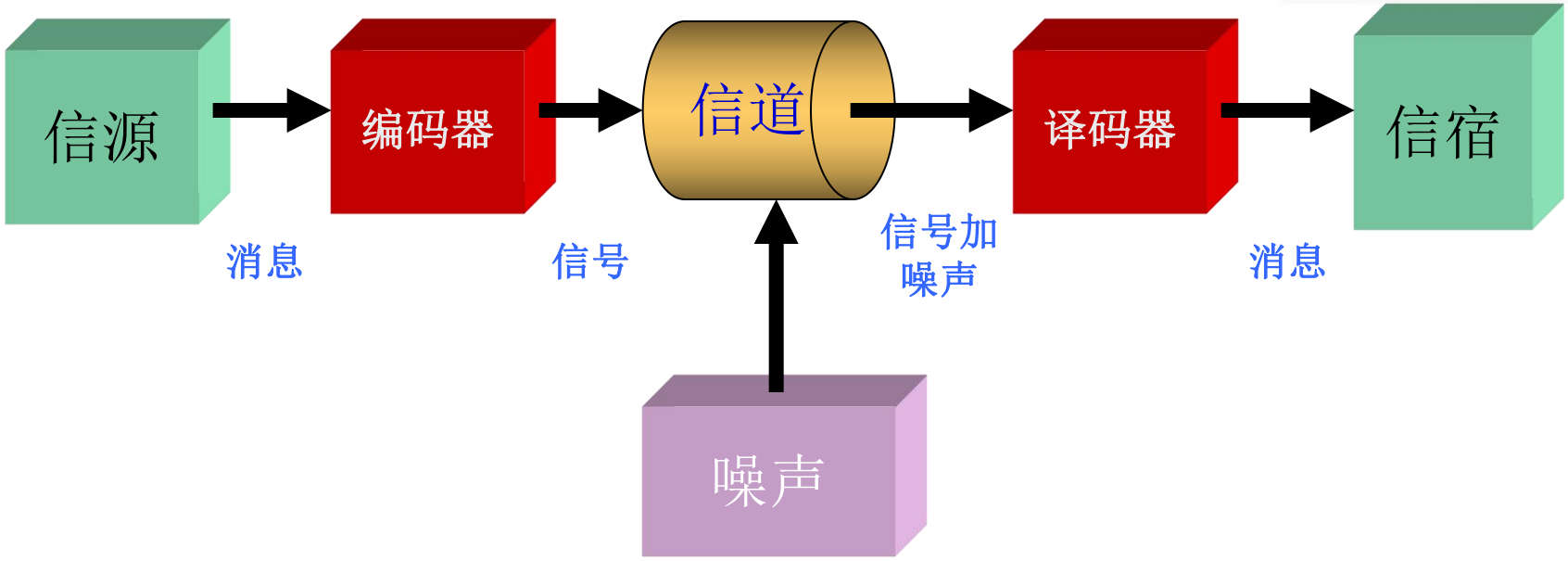


图1.2.1　通信系统模型

信源

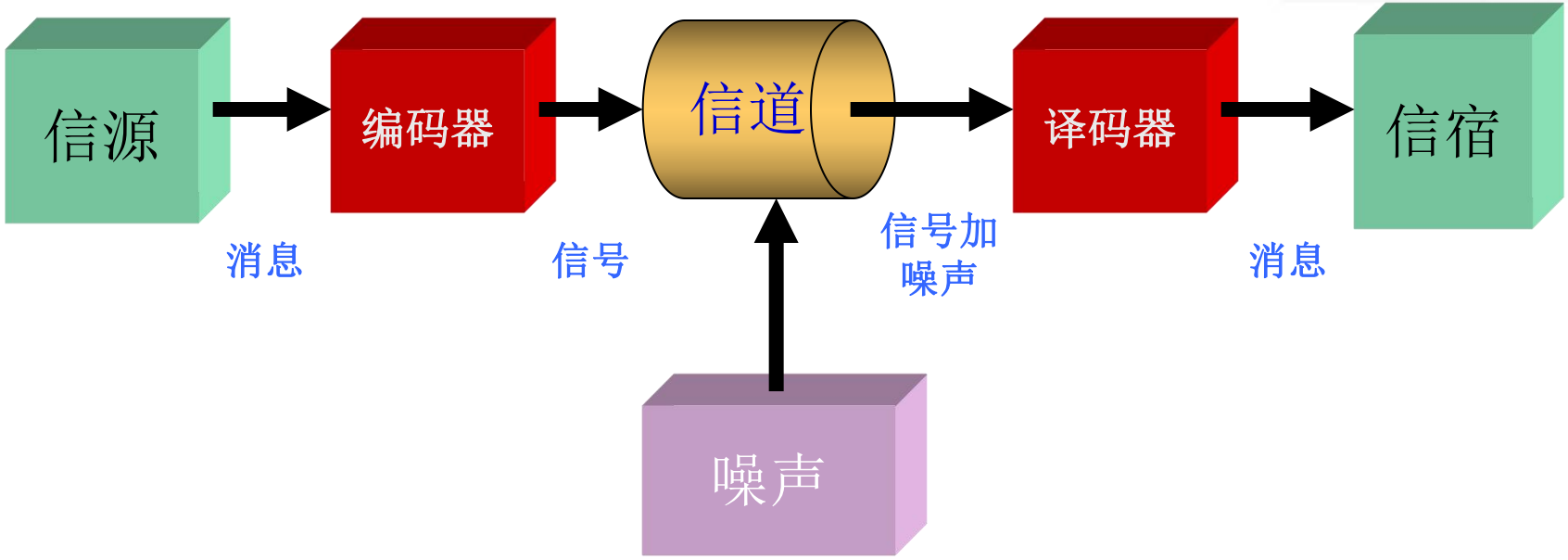


图1.2.1　通信系统模型

信源

信源是信息的来源，其功能是直接产生可能包含信息的消息

核心问题：　信源的消息中所包含的信息量

以及信息如何量度

信源

- 按输出符号的取值分类：离散信源和连续信源

- 连续信源又分为：离散时间连续信源；波形信源或模拟信源。

- 按输出符号之间的依赖关系分类： 无记忆信源，有记忆信源

本章主要内容

2.1 自信息和互信息

2.1.1 自信息

2.1.2 互信息

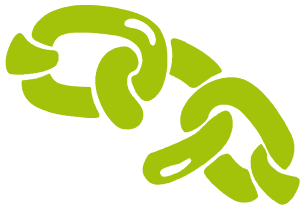
2.2 信息熵的基本概念

2.2.1 信息熵

2.2.2 联合熵与条件熵

2.2.3 相对熵

2.2.4 各类熵之间的关系



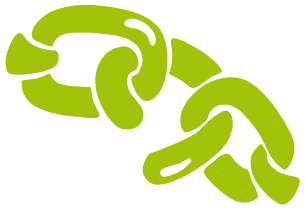
本章主要内容

2.3 信息熵的基本性质

2.3.1 凸函数及其性质
2.3.2 熵的基本性质
2.3.3 熵函数的唯一性
2.3.4 有根概率树与熵的计算

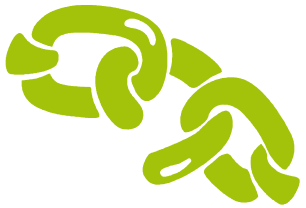
2.4 平均互信息

2.4.1 平均互信息的定义
2.4.2 平均互信息的性质
2.4.3 平均条件互信息



本书符号的约定

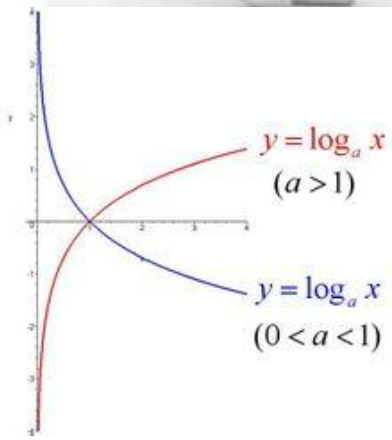
- 大写字母表示随机变量或随机事件集合
- 小写字母表示随机变量的取值或随机事件
- x取自一个有限符号集合（也称字母表）$A=\{a_1,a_2,...,a_n\}$
- 符号集也可以是无限可数的，但如无特别说明，都默认为有限符号集的情况



§2.1.1 自信息

事件集合X中的事件$x=a_i$的自信息定义为：

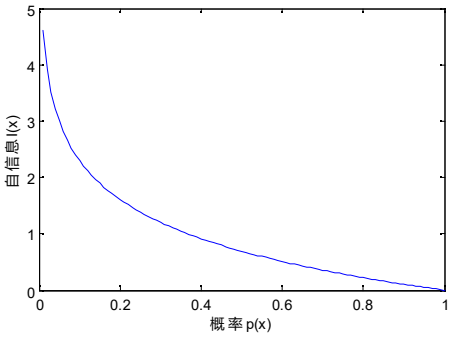
$$I_X(a_i)=-\log P_X(a_i) \quad (2.1a)$$

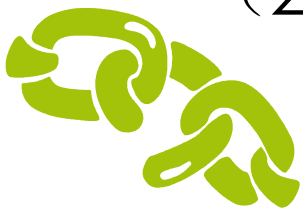


简记为 $$I(x)=-\log p(x) \quad (2.1b)$$

注：（1）$a_i \in A$，且 $\sum_{i=1}^{n} P_X(a_i)=1$，$0 \le P_X(a_i) \le 1$

（2）底数>1，自信息随概率单调递减





自信息

(2) 对数底的选取有如下几种情况：

以2为底：单位为比特（bit），工程上常用；

以3为底：单位为Tit；

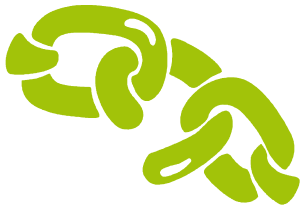
以e为底：单位为奈特（Nat），理论推导时常用；

以10为底：单位为Dit或哈特；

各单位之间的换算关系为：

$$1\text{奈特} = \log_e e = \log_2 e\text{比特} = 1.443\text{比特}$$

$$1Dit = \log_{10}10 = \log_2 10\text{比特} = 1/\log_{10} 2\text{比特} = 3.322\text{比特}$$



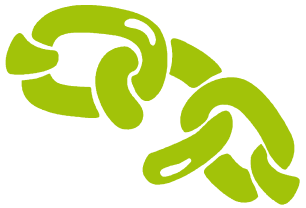
自信息

(3) 自信息为随机变量，且I(x)是p(x)的单调递减函数，即概率大的事件自信息小，而概率小的事件自信息大；

(4) 自信息含义体现在如下两个方面：

①表示事件发生前该事件发生的不确定性。

②表示事件发生后该事件所包含的信息量，也是提供给信宿的信息量，也是解除这种不确定性所需要的信息量。



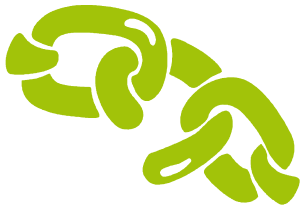
§2.1.1 自信息

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”与事件“取出一个白球”相比较，哪个事件的发生更难猜测？



§2.1.1 自信息

解 （1）设表示“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：

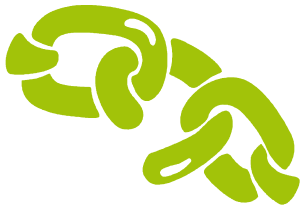
$$I(a_1)=-\log 0.9=0.152\text{比特}$$

（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。

结论：欲求事件的自信息，首先要求事件发生的概率。



联合自信息

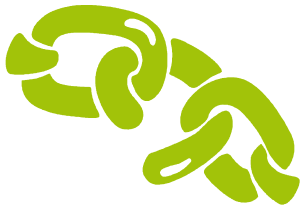
联合事件集合XY中的事件x=a_i，y=b_j包含的**联合自信息**定义为：

$$I_{XY}(a_i,b_j)=-\log P_{XY}(a_i,b_j) \qquad (2.2a)$$

简记为

$$I(xy)=-\log P(xy) \qquad (2.2b)$$

其中，$P(xy)$ 要满足非负和归一化条件。



联合自信息

联合自信息可以推广到多维随机矢量。N维矢量 $\mathrm{x}=(x_1,x_2,...,x_n)$ 的自信息定义为

$$I(\mathbf{x})=-\log p(\mathbf{x}) \qquad (2.3)$$

实际上，如果把联合事件看成一个单一事件，那么联合自信息的含义与自信息的含义相同。

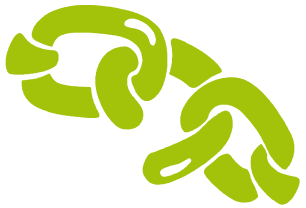
联合自信息

例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”相比较，哪个事件的发生更难猜测？



联合自信息

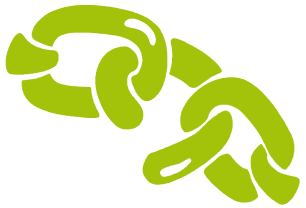
解:三种情况都是求联合自信息。设x为红球数，y为白球数。

(1) $P_{XY}(1,1)=\frac{C_{90}^{1}C_{10}^{1}}{C_{100}^{2}}=\frac{90\times10}{100\times99/2}=2/11$ $I(1,1)=-\log 2/11=2.460$比特

(2) $P_{XY}(0,2)=\frac{C_{10}^{2}}{C_{100}^{2}}=\frac{10\times9/2}{100\times99/2}=1/110$ $I(0,2)=-\log 1/110=6.782$比特

(3) $P_{XY}(2,0)=\frac{C_{90}^{2}}{C_{100}^{2}}=\frac{90\times89/2}{100\times99/2}=89/110$ $I(2,0)=-\log 89/110=0.306$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。

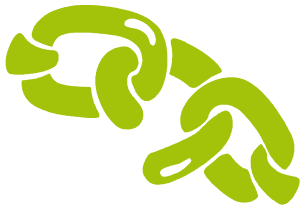


联合自信息

例2．2　设二元随机矢量 $X^N=(X_1X_2...X_N)$ ，其中 $\{X_i\}$ 为独立同分布随机变量，且1符号的概率为 $\theta(0\le\theta\le1)$，求序列 $\mathbf{x}=010011$ 的自信息。

解　所求序列的自信息为

$$I(\mathbf{x})=-\log p(\mathbf{x})=-\log[\theta^3(1-\theta)^3]=-3\log[\theta(1-\theta)]$$



条件自信息

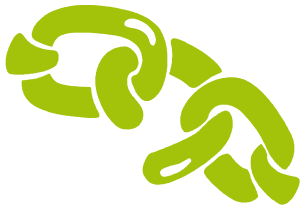
给定联合事件集XY，事件$x=a_i$在事件$y=b_j$给定条件下的条件自信息定义为：

$$I_{X/Y}(a_i|b_j)=-\log P_{X/Y}(a_i|b_j) \qquad (2.4a)$$

简记为

$$I(x|y)=-\log p(x|y) \qquad (2.4b)$$

其中：条件概率p(x|y)也要满足非负和归一化条件。



条件自信息

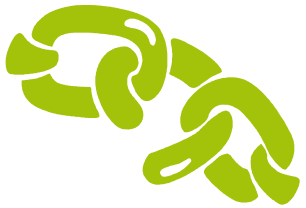
条件自信息含义与自信息类似，只不过是概率空间有变化。条件自信息的含义包含两个方面：

(1)在事件y=b_j给定条件下，在x=a_i发生前的不确定性；

(2)在事件y=b_j给定条件下，在事件x=a_i发生后所得到的信息量。

同样，条件自信息也是随机变量。容易证明，自信息、条件自信息和联合自信息之间有如下关系：

$$I(xy)=I(x)+I(y|x)=I(y)+I(x|y) \qquad (2.5)$$

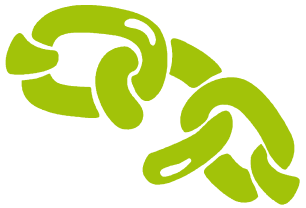


条件自信息

例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。



条件自信息

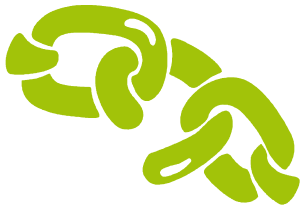
解:这两种情况都是求条件自信息，设r表示红球，w表示白球。

(1) $p(y=w|x=r)=10/99$

$I(y=w|x=r)=-\log 10/99=3.307$ 比特

(2) $p(y=r|x=r)=89/99$

$I(y=r|x=r)=-\log 89/99=0.154$ 比特

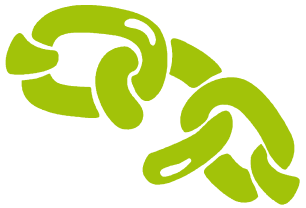


条件自信息

例2．3　有8×8=64个方格，甲将一棋子放入方格中，让乙猜；

（1）将方格按顺序编号后叫乙猜顺序号，其困难程度为何？

（2）将方格按行和列编号并告诉乙方格行号后，让乙猜列顺序号，其困难程度为何？



条件自信息

解:设行列编号分别为x和y，因为没有任何附加信息，故假定甲选择的编号是等可能的，即 $p(xy)=1/64, x=1...8, y=1...8$ ；计算得

$$p(x)=\sum_y p(xy)=1/8, x=1,...,8,\ p(y|x)=p(xy)/p(x)=1/8$$

以上两个问题归结到计算联合自信息和条件自信息的问题；

（1）$I(xy)=-\log_2 p(xy)=\log_2 64=6$ 比特

（2）$I(x|y)=-\log_2 p(y|x)=\log_2 8=3$ 比特

条件互信息

离散随机事件x= a_i 和y=b_j 之间的互信息（x∈X ，y∈Y）定义为：

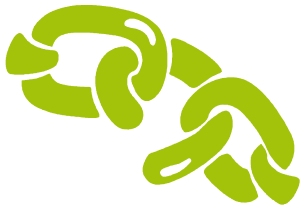
$$I_{X;Y}(a_i;b_j)=\log\frac{P_{X/Y}(a_i|b_j)}{P_X(a_i)} \tag{2.6a}$$

简记为

$$I(x;y)=\log\frac{p(x\mid y)}{p(x)} \tag{2.6b}$$

通过计算可得

$$I(x;y)=I(x)-I(x\mid y) \tag{2.7}$$



§2.1.2 互信息

注：

(1)互信息的单位与自信息单位相同；

(2)x与y的互信息等于x的自信息减去在y条件下x的自信息。I(x)表示 x的不确定性，I(x|y)表示在 y发生条件下x 的不确定性；因此I(x;y)表示当 y发生后x不确定性的变化。两个不确定度之差，是不确定度消除的部分，也就是由y发生所得到的关于x的信息量。

(3)互信息反映了两个随机事件x与y之间的统计关联程度。在通信系统中，互信息的物理意义是，信道输出端接收到某消息（或消息序列）y后，获得的关于输入端某消息（或消息序列）x的信息量。

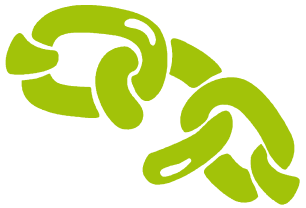
互信息的性质

(1)互易性：I(x；y)=I(y；x)；

(2)当事件x，y统计独立时，互信息为零，即I(x；y)=0；

(3)互信息可正可负；

(4)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。

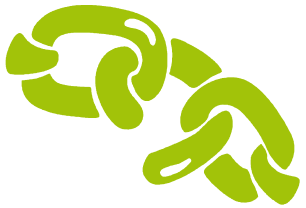


互信息的性质

- 由定义明显看出性质(1）成立，而且

$$I(x;y)=\log\frac{p(x|y)}{p(x)}=\log\frac{p(y|x)}{p(y)}=\log\frac{p(xy)}{p(x)p(y)} \qquad (2.8)$$

- 当事件x ，y 统计独立时，有p(x|y)= p(x)，所以性质（2）成立；
- 因为，当p(x|y)> p(x)时，I(xy)>0； 当p(x|y)<p(x)时，I(xy)<0，所以性质（3）成立；
- 根据(2.7)，并考虑自信息和条件自信息的非负性，可得性质（4）。也可以说，一个事件的自信息是任何其他事件所能提供的关于该事件的最大信息量。



互信息的性质

例2．4 设e表示事件“降雨”，f表示事件“空中有乌云”，且P(e)=0.125, P(e/f)=0.8,

求：（1）事件“降雨”的自信息；

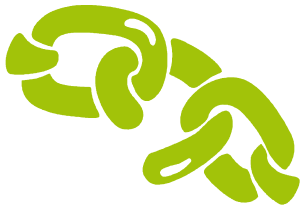
（2）在“空中有乌云”条件下“降雨”的自信息；

（3）事件“无雨”的自信息；

（4）在“空中有乌云”条件下“无雨”的自信息；

（5）“降雨”与“空中有乌云”的互信息；

（6）“无雨”与“空中有乌云”的互信息；



互信息的性质

解　设表示事件“无雨”，则$P(\bar{e})=1-P(e)$；

(1) $I(e)=-\log 0.125 = 3$ bit ；

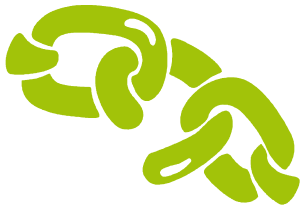
(2) $I(e|f)=-\log 0.8 = 0.322$ bit

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

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

(5) $I(e;f)=3-0.322 = 2.678$ bit

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

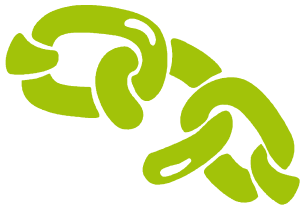


条件互信息

设联合事件集XYZ，在给定z∈Z 条件下，x(∈X) 与y(∈Y) 之间的条件互信息定义为：

$$I(x;y\mid z)=\log\frac{p(x\mid yz)}{p(x\mid z)} \qquad (2.9)$$

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。



条件互信息

例2.5

设三维随机矢量(XYZ),且$p_{XYZ}(000)=1/2, p_{XYZ}(011)=1/4,$

$p_{XYZ}(101)=p_{XYZ}(110)=1/8,$求$I(x=0,y=0|z=0)$和$I(x=1;y=0|z=1)$。

解：由 $p_{X|Z}(0|0)=\frac{p_{XZ}(00)}{p_Z(0)}=\frac{1/2}{1/2+1/8}=4/5$ ，$p_{X|YZ}(0|00)=1$ 得

$$I(x=0;y=0|z=0)=\log 5/4$$

由 $p_{X|Z}(1|1)=\frac{p_{XZ}(11)}{p_Z(1)}=\frac{1/8}{1/4+1/8}=1/3$ ， $p_{X|YZ}(1|01)=1$ 得

$$I(x=0;y=0|z=0)=\log 3$$

§2.2.1 信息熵

离散随机变量X的熵定义为自信息的平均值

$$H(X)=E_{p(x)}[I(x)]=-\sum_x p(x)\log p(x) \qquad (2.10)$$

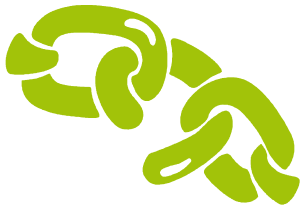
X的概率分布可写成矢量形式，称为概率矢量，记为 $p=(p_1,p_2,\ldots,p_n)$，X的熵可简记为

$$H(X)=H(p)=H(p_1,p_2,\ldots,p_n) \qquad (2.11)$$

因此，$H(p_1,p_2,\ldots,p_n)$ 也称为概率矢量 $p=(p_1,p_2,\ldots,p_n)$的熵。当n=2时，简记为

$$H(p,1-p)=H(p) \qquad (2.12)$$

其中，$p\le 1/2$，为二元信源中一个符号的概率。

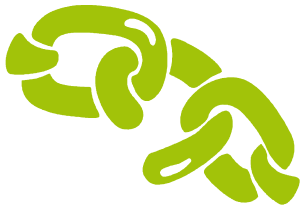


§2.2.1 信息熵

注：(1) I(x)为事件X=x的自信息, $E_{p(x)}$表示对随机变量用p(x)取平均运算；熵的单位为：比特（奈特）/ 符号。

(2) $\sum_{i=1}^{n} p_i = 1$ ，$0 \le p_i \le 1$ ，所以H(X)为n-1元函数。

式(2.10)与统计力学中热熵的表示形式相同（仅差一个常数因子），为与热熵区别，将H(X)称为信息熵，简称熵。



§2.2.1 信息熵

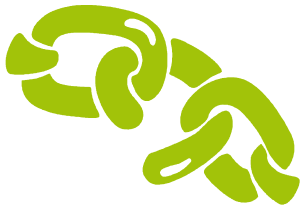
信息熵是从平均意义上表征随机变量总体特性的一个量，其含义体现在如下几方面：

(1) 在事件发生后，表示平均每个事件（或符号）所提供的信息量；

(2) 在事件发生前，表示随机变量取值的平均不确定性；

(3) 表示随机变量随机性大小，熵大的，随机性大；

(4) 当事件发生后，其不确定性就被解除，熵是解除随机变量不确定性平均所需信息量。

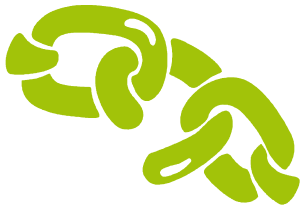


§2.2.1 信息熵

例 2.6　一电视屏幕的格点数为$500\times600=3\times10^5$，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:电视屏幕可能出现的画面数为10^{300000}，所以每个画面出现的概率为$p=10^{-300000}$，每幅画面平均所包含的信息量为:

$H(X)=\log_2(1/p)=\log_2(10^{300000})=10^6$ 比特/画面。

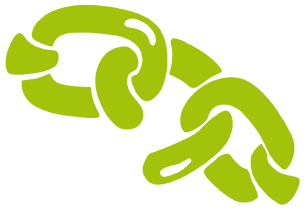


§2.2.1 信息熵

例2.7　A、B两城市天气情况概率分布如表2.1所示，问哪个城市的天气具有更大的不确定性？

表2.1　两种天气的概率分布

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



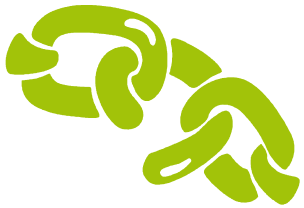
§2.2.1 信息熵

解：

$$H(A)=H(0.8,0.15,0.05)=-0.8\times\log 0.8-0.15\times\log 0.15-0.05\times\log 0.05=0.884$$ 比特/符号

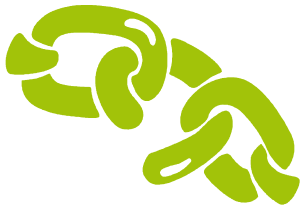
$$H(B)=H(0.4,0.3,0.3)=-.4\times\log .4-.3\times\log .3-.3\times\log .3=1.571$$ 比特/符号

所以，B城市的天气具有更大的不确定性。



§2.2.1 信息熵

例 2.8　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。现做从两箱中分别做随机取一球的实验，问从哪箱中取球的结果随机性更大？



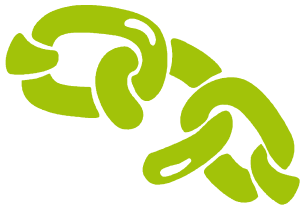
§2.2.1 信息熵

解:设A、B分别代表甲、乙两箱，则

$$H(A)=H(0.5,0.2,03)=-0.5\times\log 0.5-0.2\times\log 0.2-0.3\times\log 0.3=1.486$$ 比特/符号

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

所以，从甲箱中取球的结果随机性更大。



§2.2.2 联合熵与条件熵

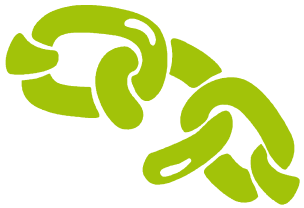
联合熵用于多维随机矢量的信息度量。设N维随机矢量，$\mathrm{X}^N=(X_1X_2,...,X_N)$ 取值为 $x=(x_1,x_2,...,x_n)$，联合熵定义为联合自信息的平均值

$$H(\mathbf{X}^N)=H(X_1X_2\cdots X_n)=E_{p(\mathbf{x})}[-\log p(\mathbf{x})]=-\sum_{\mathbf{x}}p(\mathbf{x})\log p(\mathbf{x}) \qquad (2.13)$$

其中，为矢量的联合概率，式中是N重求和。联合熵是信息熵的扩展，单位是比特/N个符号。

对于二维随机矢量XY，联合熵表示为：

$$H(XY)=E_{p(xy)}[I(xy)]=-\sum_{x}\sum_{y}p(x\,y)\log p(x\,y) \qquad (2.14)$$



§2.2.2 联合熵与条件熵

例 2.9　设随机变量X和Y符号集均为{0，1}，且 $p(x=0)=2/3$，$p(y=0|x=0)=1/2$，$p(y=1|x=1)=1/3$, 求联合熵H(XY)。

解：由 $p(xy)=p(x)p(y|x)$，可得XY的联合概率分布 $p(xy)$ 如表2.2所示：

表2.2　两种天气的联合概率分布

p(xy)　y / x	0	1
0	1/3	1/3
1	2/9	1/9

联合熵可化为一维熵计算，有 $H(XY)=H(1/3,\ 1/3,\ 2/9,\ 1/9)=1.8911$ 比特/2个符号

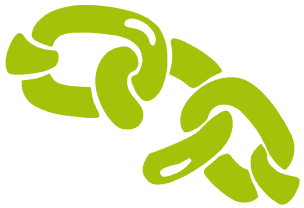
条件熵

对于二维随机矢量XY，条件熵 $H(Y|X)$定义为条件自信息$I(y|x)$的平均值

$$H(Y|X)=\underset{p(xy)}{E}[I(y|x)]=-\sum_{x}\sum_{y}p(x\ y)\log p(y|x) \qquad (2.15a)$$

$$=\sum_{x}p(x)[-\sum_{y}p(y|x)\log p(y|x)]=\sum_{x}p(x)H(Y|x) \qquad (2.15b)$$

其中，$H(Y|x)=-\sum_{y}p(y|x)\log p(y|x)$为在x取某一特定值时Y的熵。



条件熵

例 2.9（续）求条件熵 $H(Y|X)$ 。

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

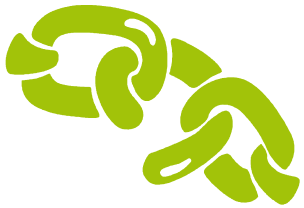
$=(2/3)H(1/2)+(1/3)H(1/3)=0.9728$ 比特/符号

条件熵

例 2．10　设随机变量X与Y之间的条件概率矩阵为：

$$\mathrm{p}=\begin{pmatrix} \mathrm{p}_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{pmatrix} \qquad (2.16)$$

其中 $\mathrm{p}_{ij}=p(y=j\,|\,x=i), i=1,2,\ldots,n, j=1,2,\ldots,m$ 求 $H(Y\,|\,X)$。



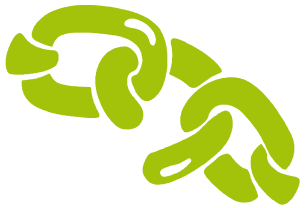
条件熵

解：$H(Y|X)=-\sum_{x,y}p(xy)\log p(y|x)=-\sum_{i,j}p_i\,p_{ij}\,\log p_{ij}=\sum_i p_i H(Y|x=i)$

其中，$H(Y|x=i)=-\sum_j p_{ij}\log p_{ij}$。

如果式(2.16)所表示的条件概率矩阵的各行所包含的元素都相同，则$H(Y|x=i)$与i无关，此时

$$H(Y|X)=H(Y|x=i)=H(p_{11},p_{12},\ldots,p_{1m})$$

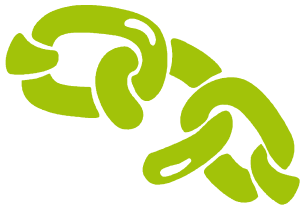


条件熵

条件熵也可扩展到多维矢量的情况。设N维随机矢量$X^N=(X_1...X_N)$和M维随机矢量$Y^M=(Y_1...Y_M)$，其中$x=(x_1...x_N)$，$y=(y_1...y_M)$，联合集X^NY^M上，条件熵定义为

$$H(\mathbf{Y}^M\,|\,\mathbf{X}^N)=-\underset{p(\mathbf{xy})}{E}[p(\mathbf{y}\,|\,\mathbf{x})]=-\sum_{\mathbf{xy}}p(\mathbf{x}\,\mathbf{y})\log p(\mathbf{y}\,|\,\mathbf{x})$$

当M=N=1时，式(2.17)归结于(2.15)。



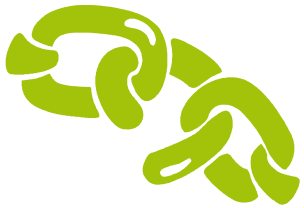
§2.2.3 相对熵

若P和Q为定义在同一概率空间的两个概率测度，定义P相对于Q的相对熵为：

$$D(P//Q)=\sum_{x}P(x)\log\frac{P(x)}{Q(x)} \qquad (2.18)$$

●相对熵又称散度、鉴别信息、方向散度、交叉熵、Kullback_Leibler距离等。

●在（2.18)中，概率分布的维数不限，可以是一维，也可以是多维，也可以是条件概率。



§2.2.3 相对熵

首先介绍一个在信息论中有用的不等式。

对于任意正实数x，下面不等式成立

$$1-1/x \le \ln x \le x-1 \qquad (2.19)$$

证明：①设 $f(x)=\ln x - x + 1$，可求得函数的稳定点为x=1，并可求得在该点的二阶导数小于0，从而可得x=1为f(x)取极大值的点，即 $f(x)=\ln x - x + 1 \le 0$，仅当x=1时式(2.19)右边等号成立。

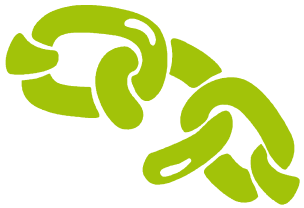
②令y=1/x，可得 $1-1/y \le \ln y$，再将y换成x，就得到左边的不等式。

§2.2.3 相对熵

定理2.1 如果在一个共同有限字母表概率空间上给定两个概率测度P(x)和Q(x)，那么

$$D(P//Q) \geq 0 \qquad (2.20)$$

仅当对所有x, P(x)=Q(x)时, 等式成立。



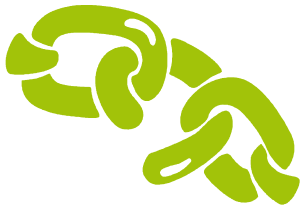
§2.2.3 相对熵

证：因 $P(x),Q(x)\geq 0$，$\sum_x P(x)=\sum_x Q(x)=1$，所以根据式(2.19)，有

$$-D(P//Q)=\sum_x P(x)\log\frac{Q(x)}{P(x)}\leq\sum_x P(x)(\log e)[\frac{Q(x)}{P(x)}-1]$$

$$=(\log e)[\sum_x Q(x)-\sum_x P(x)]=0$$

仅当对所有x，P(x)=Q(x)时，等式成立。



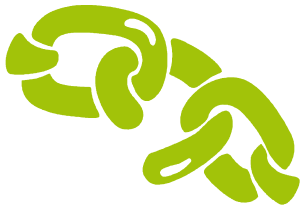
§2.2.3 相对熵

式(2.20)称为散度不等式（divergence inequality）。

●该式说明，一个概率测度相对于另一个概率测度的散度是非负的，仅当两测度相同时，散度为零。

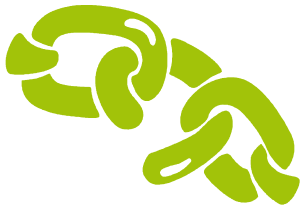
●散度可以解释为两个概率测度之间的“距离”，即两概率测度不同程度的度量。

●散度并不是通常意义下的距离，因为它不满足对称性，也不满足三角不等式。



§2.2.3 相对熵

例2.11　设一个二元信源的符号集为{0，1}，有两个概率分布p和q，并且 $p(0)=1-r, p(1)=r$, $q(0)=1-s$, $q(1)=s$, 求散度 $D(p//q)$ 和 $D(q//p)$ ，并分别求当 $r=s$ 和 $r=2s=1/2$ 时散度的值。



§2.2.3 相对熵

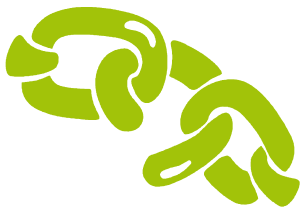
解:根据 (2.18)式，得

$$D(p//q)=(1-r)\log\frac{1-r}{1-s}+r\log\frac{r}{s},D(q//p)=(1-s)\log\frac{1-s}{1-r}+s\log\frac{s}{r}$$

当r=s时，有 $D(p//q)=D(q//p)=0$，当 $r=2s=1/2$ 时，有

$$D(p//q)=(1-1/2)\log\frac{1-1/2}{1-1/4}+(1/2)\log\frac{1/2}{1/4}=1-\log(3)/2=0.2075\text{比特}$$

$$D(q//p)=(1-1/4)\log\frac{1-1/4}{1-1/2}+(1/4)\log\frac{1/4}{1/2}=\frac{3}{4}\log(3)-1=0.1887\text{比特}$$



注：一般地，$D(p//q)$和$D(q//p)$并不相等，即不满足对称性。

§2.2.4 各类熵之间的关系

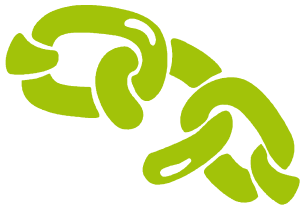
由式（2.18)可得到熵与相对熵的关系，即由

$$D(P//Q) = -E_{p(x)} \log Q(x) - H(X) \geq 0$$

得

$$E_{p(x)} \log[1/Q(x)] \geq H(X) \qquad (2.21)$$

上式表明，同一概率空间的两随机变量集合，如果一种分布的自信息用另一种分布做平均，其值不小于另一种分布的熵。



§2.2.4 各类熵之间的关系

定理2.2 (熵的不增原理)

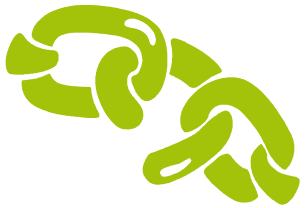
$$H(Y|X) \le H(X) \qquad (2.22)$$

证：设 $p(y)=\sum_x p(x)p(y|x)$ ，那么

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

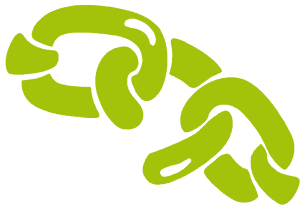
$$=-\sum_y\sum_x p(x)p(y|x)\log p(y)+\sum_x\sum_y p(x)p(y|x)\log p(y|x)$$

$$=\sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{p(y)}=\sum_x p(x)D(p(y|x)\,//\,p(y))\ge 0$$



§2.2.4 各类熵之间的关系

- 上面利用了散度不等式，仅当X、Y 相互独立时，等式成立。
- (2.22)表明，条件熵总是不大于无条件熵，这就是熵的不增原理：在信息处理过程中，已知条件越多，结果的不确定性越小，也就是熵越小。



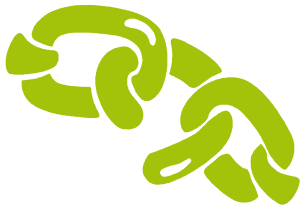
§2.3.1 凸函数及其性质

1．凸函数的定义

多元实值函数 $f(\mathbf{x})=f(x_1,x_2\cdots,x_n)$ 称为定义域上的上凸（cap）函数，若对于任何 $\alpha(0\le\alpha\le1)$，及任意两矢量 $\mathbf{x}_1,\mathbf{x}_2$，有

$$f(\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2)\ge\alpha f(\mathbf{x}_1)+(1-\alpha)f(\mathbf{x}_2) \qquad (2.23)$$

成立；若对于任何 α 及任意 $\mathbf{x}_1,\mathbf{x}_2$，上面不等式反向，则称 $f(\mathbf{x})$ 为下凸（cup）函数；若仅当 $\mathbf{x}_1=\mathbf{x}_2$ 或 $\alpha=0$（或=1）时不等式中等号成立，则称 $f(\mathbf{x})$ 为严格上凸（或下凸）函数。

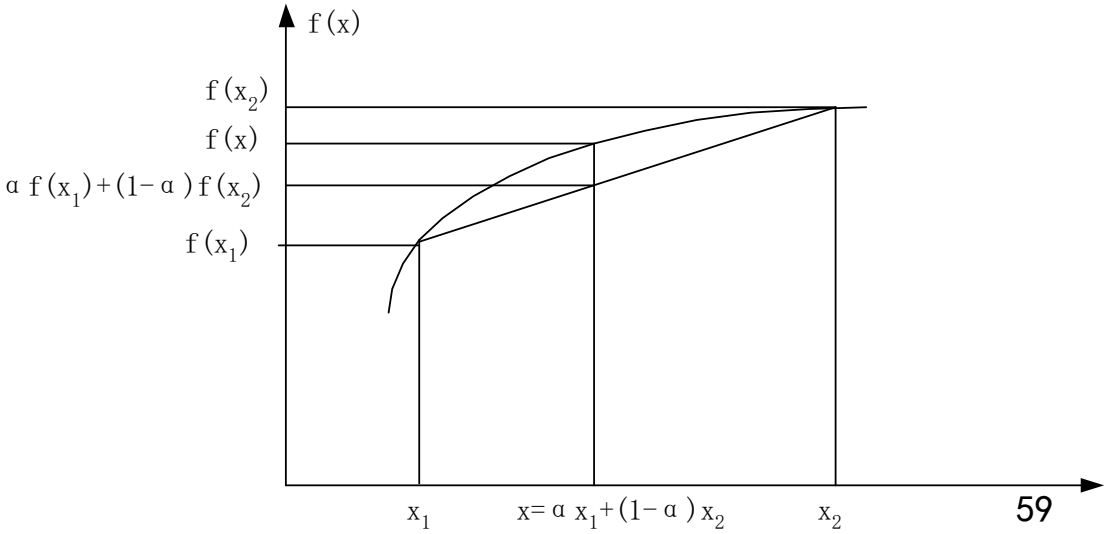


§2.3.1 凸函数及其性质

●一元上凸函数曲线如图所示。称$\mathbf{x}=\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2$为自变量$\mathbf{x}_1,\mathbf{x}_2$的内插值，称$g(\mathbf{x})=\alpha f(\mathbf{x}_1)+(1-\alpha)f(\mathbf{x}_2)$为函数值$f(\mathbf{x}_1)$和$f(\mathbf{x}_2)$的内插值。

●从图中可以看出，当α从0变化到1时，点$(\mathbf{x},g(\mathbf{x}))$从点$(\mathbf{x}_2,g(\mathbf{x}_2))$沿直线段移动到点$(\mathbf{x}_1,f(\mathbf{x}_1))$此线段实际上是连接两点的弦。

●上凸的含义就是：在自变量$\mathbf{x}_1$和$\mathbf{x}_2$之间的区域，函数的图线在连接曲线$f(\mathbf{x})$上对应两点弦的上方。



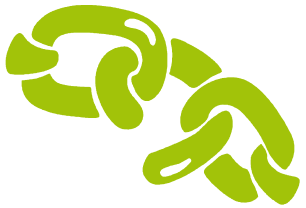
§2.3.1 凸函数及其性质

2．凸函数的性质

（1）若$f_1(x),f_2(x),...,f_k(x)$均为上凸函数，$c_1,c_2,...,c_k$均为正数，那么$\sum_i c_i f_i(\mathbf{x})$ 为上凸函数（或 $f_i(\mathbf{x})$ 严格上凸函数，若中任意一个为严格上凸）。

证：利用式(2.23)，有

$$\sum_i c_i f_i(\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2)\geq\sum_i c_i[\alpha f_i(\mathbf{x}_1)+(1-\alpha)f_i(\mathbf{x}_2)]$$
$$=\alpha\sum_i c_i f_i(\mathbf{x}_1)+(1-\alpha)\sum_i c_i f_i(\mathbf{x}_2)$$



§2.3.1 凸函数及其性质

（2）对于一维随机变量 x，若 $f(x)$在某区间的二阶导数小于等于0，即 $\partial^2 f(x)/\partial x^2 \le 0$，则在此区间内为上凸函数（或严格上凸函数，若 $\partial^2 f(x)/\partial x^2 < 0$ ）。（证明略）

（3）Jensen不等式，由下面的定理来描述。

定理2.3 若 $f(\mathbf{x})$ 是定义在某区间上的上凸函数，则对于任意一组矢量 $(x_1, x_2, ..., x_k)$和任意一组非数 $\lambda_1, \lambda_2, ..., \lambda_k$ ，$\sum_{i=1}^{k} \lambda_i = 1$，有

$$f[\sum_{i=1}^{k} \lambda_i \mathbf{x}_i] \ge \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i) \qquad (2.24)$$

对于严格上凸函数，仅当 $x_1 = ... = x_k$ 或$\lambda_i = 1 (1 \le i \le k)$且 $\lambda_j = 0 (j \ne i)$时，等式成立。

§2.3.1 凸函数及其性质

Jensen不等式的证明

证：利用数学归纳法证明。

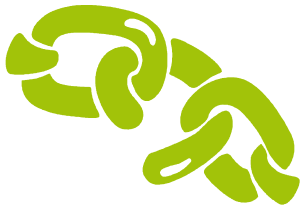
根据上凸函数的定义（2.23），说明当$k=2$时，不等式(2.24)成立。假定$k=n$时(2.23)成立，那么当$k=n+1$时设$\lambda_i \geq 0, \sum_{i=1}^{n+1}\lambda_i = 1$

令$\alpha = \sum_{i=1}^{n}\lambda_i$，有$\lambda_{n+1} = 1-\alpha$，所以

$$\sum_{i=1}^{n+1}\lambda_k f(\mathbf{x}_k) = \sum_{i=1}^{n}\lambda_i f(\mathbf{x}_i) + \lambda_{n+1} f(\mathbf{x}_{n+1}) = \alpha\sum_{i=1}^{n}(\lambda_i/\alpha) f(\mathbf{x}_i) + (1-\alpha) f(\mathbf{x}_{n+1})$$

$$\leq \alpha f[\sum_{i=1}^{n}(\lambda_i/\alpha)\mathbf{x}_i] + (1-\alpha) f(\mathbf{x}_{n+1}) \leq f[\alpha\sum_{i=1}^{n}(\lambda_i/\alpha)\mathbf{x}_i] + \lambda_{n+1}\mathbf{x}_{n+1}]$$

$$= f(\sum_{i=1}^{n+1}\lambda_i\mathbf{x}_i)$$



§2.3.1 凸函数及其性质

式（2.24）称为Jensen不等式。因为可作为随机矢量的概率分布，所以有如下推论。

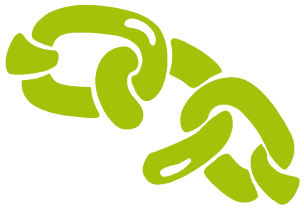
推论2.1　若$f(\mathbf{x})$为上凸函数，那么

$$E[f(\mathbf{x})] \le f[E(\mathbf{x})] \qquad (2.25)$$

在信息论中，对数函数是最常用的上凸函数，根据(2.25)，有

推论2.2　对于一元对数函数$\log(x)$，有

$$E[\log(x)] \le \log[E(x)] \qquad (2.26)$$



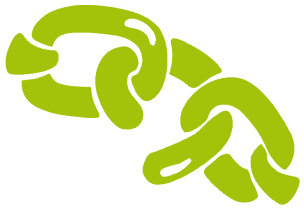
§2.3.2 熵的基本性质

1．对称性

概率矢量 $p=(p_1,p_2,\ldots,p_n)$ 中，各分量的次序任意改变，熵不变，即

$$H(p_1,p_2,\ldots,p_n)=H(p_{j_1},p_{j_2},\ldots,p_{j_n}) \qquad (2.27)$$

其中，$j_1,j_2,\ldots j_n$ 是1, 2, . . . , n的任何一种n级排列。该性质说明熵仅与随机变量总体概率特性（即概率分布）有关，而与随机变量的取值以及符号排列顺序无关。



§2.3.2 熵的基本性质

2．非负性

$$H(\mathbf{p}) = H(p_1, p_2, \cdots, p_n) \geq 0 \qquad (2.28)$$

仅当对某个 $p_i = 1$ 时，等式成立。

●自信息是非负的，熵为自信息的平均，所以也是非负的。

●非负性仅对离散熵有效，而对连续熵来说这一性质并不成立。

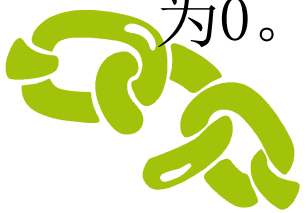
§2.3.2 熵的基本性质

3．确定性

$$H(1,0)=H(1,0,0)=H(1,0,\ldots,0)=0 \qquad (2.29)$$

●这就是说，当随机变量集合中任一事件概率为1时，熵就为0。

●这个性质意味着，从总体来看，事件集合中虽含有许多事件，但如果只有一个事件几乎必然出现，而其他事件几乎都不出现，那么，这就是一个确知的变量，其不确定性为0。

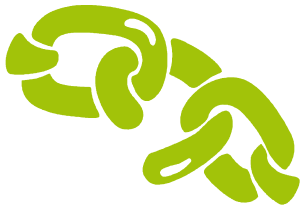


§2.3.2 熵的基本性质

4. 扩展性

$$\lim_{\varepsilon \to 0} H_{n+1}(p_1, p_2, \ldots, p_n - \varepsilon, \varepsilon) = H(p_1, p_2, \ldots, p_n) \qquad (2.30)$$

利用 $\lim_{\varepsilon \to 0} \varepsilon \log \varepsilon = 0$ 可得到上面的结果，其含义是，虽然小概率事件自信息大，但在计算熵时所占比重很小，可以忽略。

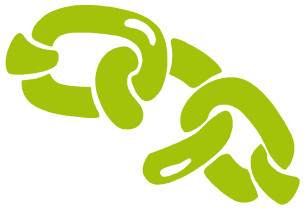


§2.3.2 熵的基本性质

5．可加性

熵的可加性首先由香农提出，含义如下：如果一个事件可以分成两步连续选择来实现（多步产生的事件也称复合事件），那么原来的熵H应为H的单独值的加权和。

注：“H的单独值”是指每次选择的熵值，“权值”就是每次选择的概率。



§2.3.2 熵的基本性质

先看一个简单例子，某随机事件集合有3个事件，概率分别为：$p_1=1/2$，$p_2=1/3$，$p_3=1/6$；

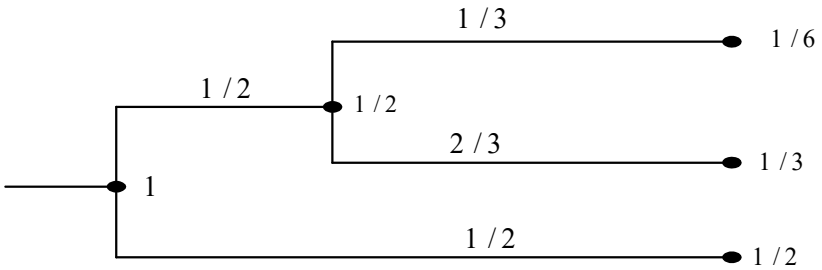
这3个事件可以直接产生，也可分两步产生，即先以1/2的概率产生两个事件，选择其中之一作为输出，或者在另一事件发生条件下再以2/3和1/3的概率产生两事件，选择其中的一个作为输出。

§2.3.2 熵的基本性质

3个事件概率的产生可用下图的树来描述，从根节点开始，通过两步选择生成的3片树叶代表3个事件，节点旁边的数值表示该节点的概率，分支旁的数字表示分支的条件概率，原节点的概率乘分支条件概率就得到产生的下一节点（也称子节点）的概率。熵的可加性意味概率矢量的熵

$$H(1/2,1/3,1/6)=H(1/2,1/2)+(1/2)H(1/3,2/3) \qquad (2.31)$$

上式等号右边的第1项是第1步选择的熵；由于第2步选择只有1/2的概率发生，所以第2项是第2步选择的熵与权值1/2的乘积。



复合事件产生例

§2.3.2 熵的基本性质

上例可以推广到一般情况。设某事件集合$n \times m$包含个事件，概率分别为 $p_1p_{11},\ldots,p_1p_{1m},p_2p_{21},\ldots,p_2p_{2m},\ldots,p_np_{n1},\ldots p_np_{nm}$这 $n \times m$ 个事件可以分两步产生，第一步产生n个事件，其中每个事件的概率分别为$p_1p_2,\ldots,p_n$；第二步再以这n个事件为条件，以$p_{i1},\ldots p_{im}(i=1,\ldots,n)$为条件概率，分别产生m个事件。熵的可加性的一般形式可以表示成：

$$H(p_1p_{11},\ldots,p_1p_{1m},p_2p_{21},\ldots,p_2p_{2m},\ldots,p_np_{n1},\ldots p_np_{nm})$$
$$=H(p_1,\ldots,p_n)+\sum_{i=1}^{n}p_iH(p_{i1},\ldots p_{im}) \qquad (2.32)$$

其中，$p_i,p_{ij}\geq 0$，对所有i，j；$\sum_{i=1}^{m}p_i=1$；$\sum_{j=1}^{m}p_{ij}=p_{i1}+\ldots+p_{im}=1$，$i=1,\ldots,n$。

§2.3.2 熵的基本性质

可把这个事件的概率写成阶矩阵，形式如下：

$$Q=\begin{pmatrix} p_1p_{11} & p_1p_{12} & \cdots & p_1p_{1m} \\ p_2p_{21} & p_2p_{22} & \cdots & p_2p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_np_{n1} & p_np_{n2} & \cdots & p_np_{nm} \end{pmatrix} \qquad (2.33)$$

●矩阵(2.33)从第1行到第n行，每行元素的和分别为$p_1, p_2, \ldots, p_n$，因为$\{p_i, i=1,\ldots,n\}$满足归一性，故可以视为一个事件集合的概率分布，设此集合为X。

●同理也可把每列元素的和$\{q_j, j=1,\ldots,m\}$视为另一个事件集合的概率分布，设此集合为Y，

●矩阵(2.33)就是XY的联合概率矩阵，其中，$p(x=a_i, y=b_j)=p_ip_{ij}$；

所以$H(X)=H(p_1,\ldots,p_n)$，$H(Y|X)=\sum_{i=1}^{n} p_iH(p_{i1},\ldots,p_{im})$，

$H(XY)=H(p_1p_{11},\ldots,p_1p_{1m},p_2p_{21},\ldots,p_2p_{2m},\ldots,p_np_{n1},\ldots p_np_{nm})$。

§2.3.2 熵的基本性质

定理2.4 （熵的可加性）

$$H(XY)=H(X)+H(Y|X)=H(Y)+H(X|Y) \qquad (2.34)$$

证：$$H(XY)=-\sum_x\sum_y p(x\,y)\log p(x\,y)=-\sum_i p_i\sum_j p_{ij}[\log p_i+\log p_{ij}]$$

$$=-\sum_i p_i\log p_i\sum_j p_{ij}-\sum_i p_i\sum_j p_{ij}\log p_{ij}=H(X)+H(Y|X)$$

又 $$H(XY)=-\sum_x\sum_y p(x\,y)\log p(x\,y)=-\sum_j q_j\sum_i (p_i p_{ij}/q_j)[\log q_j+\log(p_i p_{ij}/q_j)]$$

$$=-\sum_j q_j\sum_i (p_i p_{ij}/q_j)[\log q_j+\log(p_i p_{ij}/q_j)]$$

其中，$p(x|y)=p(x)p(y|x)/p(y)=p_i p_{ij}/q_j$

§2.3.2 熵的基本性质

●如果把随机事件产生的过程倒过来看，由$n \times m$个事件的集合变成n个事件的集合的过程相当于原事件集合中符号合并的过程，其中每个m事件合并成一个事件。

●符号合并前的熵由(2.32)等号的左边表示，合并后的熵由(2.32)等号右边的第一项表示，而等号右边第2项大于零。

●结论：随机变量符号集中的符号经合并后，随机变量的熵减小。

§2.3.2 熵的基本性质

例2.12　设某地气象为随机变量X，符号集A={晴，多云，阴，雨，雪，雾，霾}，概率分别为0.3, 0.2, 0.2, 0.05, 0.05, 0.05, 0.15；现将多云和阴用阴代替，雨和雪用降水代替，雾和霾用雾霾代替，得到简化气象Y，符号集B ={晴，阴，降水，雾霾}，求两气象熵的差。

§2.3.2 熵的基本性质

解 ：两气象熵的差可利用熵的可加性公式(2.32)，有

$$H(X)-H(Y)=0.4\times H(1/2)+0.1\times H(1/2)+0.2\times H(1/4)=0.6623\text{比特/符号}$$

实际上，式 (2.32)是熵的可加性常用的一种描述方式，是指通过两步产生随机事件的情况，对于多步产生的随机事件的情况，可用多维随机矢量熵的可加性来描述。

§2.3.2 熵的基本性质

定理2.3.3（熵的链原则）设N维随机矢量 $(X_1X_2...X_n)$，则有

$$H(X_1X_2...X_N)=H(X_1)+H(X_{2|}\,|\,X_1)+...+H(X_n\,|\,X_1...X_{N-1})$$
$$=\sum_{i=1}^{N}H(X_i\,|\,X_1...X_{i-1})\le\sum_{i=1}^{N}H(X_i) \qquad (2.35)$$

仅当 $X_1X_2...X_n$ 统计独立（即 X_i 独立于 $X_1X_2...X_{i-1}$，对 $i=1,...,N$ ）时，等式成立，即

$$H(X_1X_2...X_N)=H(X_1)+H(X_{2|})+...+H(X_n) \qquad (2.36)$$

上式称为熵的强可加性。(2.35)称做熵的链原则，式中规定 $\phi=X_1X_2...X_N$，当 $i<1$ 。

§2.3.2 熵的基本性质

证　通过将联合概率展开，再求平均，得

$$H(X_1X_2...X_N)=-E_{p(x)}\log p(x)=-E_{p(x)}\log p(x_1x_2...x_N)$$
$$=-E_{p(x)}\log[p(x_1)p(x_2\mid x_1)...p(x_N\mid x_1...x_{N-1})]$$
$$=-E_{p(x)}\sum_{i=1}^{N}\log p(x_i\mid x_1...x_{i-1})]=-\sum_{i=1}^{N}E_{p(x)}[\log p(x_i\mid x_1...x_{i-1})]$$
$$=\sum_{i=1}^{N}H(X_i\mid X_1...X_{i-1})\le\sum_{i=1}^{N}H(X_i)$$

上面的不等式用到熵的不增原理，仅当 $X_1,X_2,...,X_N$ 统计独立时，等号成立。

§2.3.2 熵的基本性质

熵的可加性可以从多种角度来理解：

（1）复合事件集合的不确定性为组成该复合事件的各简单事件集合不确定性的和；

（2）对事件输出直接测量所得信息量等于分成若干步测量所得信息量的和；

（3）事件集合的平均不确定性可以分步解除，各步解除不确定性的和等于信息熵。

§2.3.2 熵的基本性质

例 2.13　现有12个外形相同的硬币。知道其中有一个重量不同的假币，但不知它是比真币轻，还是比真币重。现用一无砝码天平对这些硬币进行称重来鉴别假币，无砝码天平的称重有3种结果：平衡，左倾、右倾。问至少称几次才能鉴别出假币并判断出其是轻还是重？

§2.3.2 熵的基本性质

解：●根据熵的可加性，一个复合事件的不确定性可以通过多次实验分步解除，各次试验所得信息量的总和应该不小于随机变量集合的熵。如果使每次实验所获得的信息量最大，那么所需要的总实验次数就最少。

●用无砝码天平一次称重实验所得到的最大信息量为log3，k次称重所得的最大信息量为klog3。设每一个硬币是假币的概率都相同，那么从12个硬币中鉴别其中一个重量不同（不知是否轻或重）的假币所需信息量为log24。而2log3=log9<log24<log27=3log3。所以理论上至少3次称重才能鉴别出假币并判断其轻或重。

§2.3.2 熵的基本性质

6．极值性

定理2.5（离散最大熵定理） 对于有限离散随机变量，当符号集中的符号等概率发生时，熵达到最大值。

证：设随机变量有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

$$D(P//Q)=\sum_x P(x)\log\frac{P(x)}{Q(x)}=\sum_x P(x)\log P(x)-\sum_x P(x)\log(1/n)$$
$$=-H(X)+\log n\geq 0 \qquad (2.37)$$

即 $H(X)\leq \log\ n$，仅当 $P(x)$ 等概率分布时等号成立。

注：离散最大熵定理仅适用于有限离散随机变量，对于无限可数符号集，只有附加其他约束才有可能求得最大熵。

§2.3.2 熵的基本性质

7．上凸性

$H(\mathbf{p})=H(p_1,p_2,\cdots,p_n)$ 是概率矢量 $\mathbf{p}$ 的严格的上凸函数。

这就是说，若 $\mathbf{p}=\theta\mathbf{p}_1+(1-\theta)\mathbf{p}_2$，那么 $H(\mathbf{p})>\theta H(\mathbf{p}_1)+(1-\theta)H(\mathbf{p}_2)$ 其中 $\mathbf{p},\mathbf{p}_1,\mathbf{p}_2$ 均为n维概率矢量，$0\le\theta\le1$ 。该性质可用凸函数性质（1）来证明（提示：先证明 $-p_i\log p_i$ 是严格上凸的，见习题2.14)。

§2.3.2 熵的基本性质

8．一一对应变换下的不变性

离散随机变量的变换包含两种含义，一是符号集中符号到符号的映射，二是符号序列到序列的变换。首先研究第一种情况。设两随机变量X、Y，符号集分别为A、B，其中Y是X的映射，可以表示为$A \to B, x \to f(x)$ 。因此有

$$p(y \mid x) = \begin{cases} 1 & y = f(x) \\ 0 & y \neq f(x) \end{cases} \qquad (2.38)$$

§2.3.2 熵的基本性质

所以 $H(Y|X)=0$ ；$H(XY)=H(X)+H(Y|X)=H(X)$ ，而另一方面 $H(XY)=H(Y)+H(X|Y)\geq H(Y)$ ，所以，$H(X|Y)=0$，仅当 f 是一一对应映射时等号成立，此时 $H(X)\geq H(Y)$ 。应用类似的论证也可推广到多维随机矢量情况，因此得到如下定理。

定理2.6　离散随机变量（或矢量）经符号映射后的熵不大于原来的熵，仅当一一对应映射时熵不变。

§2.3.2 熵的基本性质

例 2.14　设二维随机矢量XY，其中X、Y为独立同分布随机变量，符号集为A={0，1，2}，对应的概率为{1/3，1/3，1/3}，做变换 $u=x+y$，$v=x-y$，得到二维随机矢量UV；求 $H(U),H(V),H(UV)$。

§2.3.2 熵的基本性质

解 很明显，u,v 都是x,y的函数，并且在x,y 给定条件下u,v 独立，所以，

$$p(u)=\sum_{x,y}p(xyu)=\sum_{x,y}p(xy)p(u\mid xy)=\sum_{x,y}p(x)p(y)|_{u=x+y}=\sum_{x}p_X(x)p_Y(u-x)$$

上式表明$p(u)$ 是 $p(x)$ 与$p(y)$ 的卷积，可用图解法计算，得U的符号集为{-2，-1，0，1，2}，概率分布为$(1/9,2/9,1/3,2/9,1/9)$，所以

$$H(U)=H(1/9,2/9,1/3,2/9,1/9)=2.1972 \quad \text{比特/符号}$$

同理可得 $$H(V)=H(U)=2.1972 \quad \text{比特/符号}$$

因为变换是一一对应的，所以

$$H(UV)=H(XY)=H(X)+H(Y)=2\log 3=3.1699 \quad \text{比特/2个符号}$$

因为$H(UV)<H(U)+H(V)$ ，因此u,v 不独立（条件独立）。

§2.3.2 熵的基本性质

下面研究第二种情况。设随机变量X构成的长度为N的序列变换到随机变量Y构成的长度为M的序列，称为由X^N 到Y^M 的变换，记为 $X^N \to Y^M$。其中最有意义的是一一对应的变换，此时有$H(X^N \mid Y^M)=H(Y^M \mid X^N)=0$。由此可以推出

$$H(X^N)=H(Y^M) \qquad (2.39)$$

其中，$H(X^N)$ 为变换前N维联合熵，$H(Y^M)$ 为变换后M维联合熵。

§2.3.2 熵的基本性质

定理2.7 离散随机序列经一一对应变换后，序列的熵不变，但单符号熵可能改变。

实际上，经过这种一一对应变换后，有两种极端情况：

（1）X 的一个符号用 Y 的多个符号表示，例如信源编码器；

（2）X 的多个符号表示 Y 的一个符号，例如第3章的扩展源。

§2.3.3 熵函数的唯一性

可以证明，如果要求熵函数满足以下条件：

（1）是概率的连续函数；

（2）当各事件等概率时是n（信源符号数）的增函数；

（3）可加性；

那么，熵函数的表示是唯一的，即与(2.10)的表示形式仅差一个常数因子。

§2.3.4 有根概率树与熵的计算

●在有根树中，从根节点延伸的分支端点构成1阶节点；从1阶节点延伸的分支端点构成2阶节点；……最后到末端节点，称作树叶，树叶不再继续延伸。

●树上的每一个i阶节点是其延伸产生的i+1阶节点的父节点，而这些i+1阶节点又是产生它们的i阶节点的子节点。

●从根节点开始到叶节点终止，每个节点分配相应的概率，此树称有根概率树。

§2.3.4 有根概率树与熵的计算

●其中根节点的概率为1；每个父节点（设为u）的概率p(u)是其所有子节点（设为v_i）概率p(v_i)的和，即$p(u)=\sum_i p(v_i)$，对应分支p($u \to v_i$）的概率为子节点概率除以父节点概率所得的商，即$p(v_i \mid u)=p(v_i)/p(u)$。

●每个节点的所有分支概率的和为1，利用这些分支概率计算得到的熵称为该节点的分支熵。

●除叶节点外，每个内部节点与其向后延伸的所有分支和节点构成一棵子树，这个内部节点作为子树的根。

§2.3.4 有根概率树与熵的计算

●设随机变量包含n个符号，各符号的概率分为$p_1, p_2, \cdots, p_n$。如果用有根概率树描述该随机变量，那么树叶的数目等于n，树叶对应的概率就是符号的概率。

●有根概率树可按如下方法构造：将这些树叶任意分组，各组分别合并形成各自的父节点；形成的所有父节点既可以继续分组、合并，也可与未合并的树叶继续分组、合并，形成阶数更低的父节点；……最后合并成一个节点，就是树根。

●可见同一个随机变量可以有多种结构的有根概率树。

§2.3.4 有根概率树与熵的计算

在有根概率树中，从树叶到树根所经过的分支数目称为该叶到树根的距离，也称叶的深度。树叶的平均深度定义为：

$$\bar{l} = \sum_{i=1}^{M} p_i l_i \qquad (2.40)$$

其中，p_i 为树叶的概率，l_i 为树叶的深度，M为树叶的数目。图2.2中有根概率树叶的平均深度为 $\bar{l} = (1/2)\times 1 + (1/3)\times 2 + (1/6)\times 2 = 1.5$ 。

§2.3.4 有根概率树与熵的计算

设有根树有n片树叶，概率分别为 $p_1, p_2, \ldots, p_n$，定义有根树叶的熵为

$$H_{leaf} = -\sum_{k=1}^{n} p_k \log p_k \qquad (2.34)$$

很明显，如果树叶与随机变量所取符号一一对应，那么树叶的熵就等于随机变量的熵。图2.2中有根概率树的叶熵为 $H = H(1/2, 1/3, 1/6) = 1.4591$ 比特/符号。

§2.3.4 有根概率树与熵的计算

设有根树的某节点m的子节点的概率分别为 $p_{m1}, p_{m2}, \ldots, p_{mr}$

定义该节点的分支熵为

$$H_m = -\sum_{j=1}^{r} (p_{mj} \mid p_m) \log(p_{mj} \mid p_m) \qquad (2.42)$$

其中 $p_m = p_{m1} + p_{m2} + , \ldots, p_{mr}$， r为节点m的子节点数。很明显，叶节点没有分支熵。

§2.3.4 有根概率树与熵的计算

定理2.7（路径长引理） 在一棵有根概率树中，叶的平均深度等于除叶之外所有节点（包括根）概率的和。也就是说，如果有根概率树所有内部节点的概率分别为$q_1, q_2, \cdots, q_M$，那么叶节点平均深度为

$$\overline{l} = \sum_{j=1}^{M} q_j \tag{2.43}$$

§2.3.4 有根概率树与熵的计算

定理2.8（叶熵定理）离散随机变量的熵等于所对应的有根概率树上所有内部节点(包括根节点，不包括叶)的分支熵用该节点概率加权的和，即

$$H(X)=\sum_i q(u_i)H(u_i) \qquad (2.44)$$

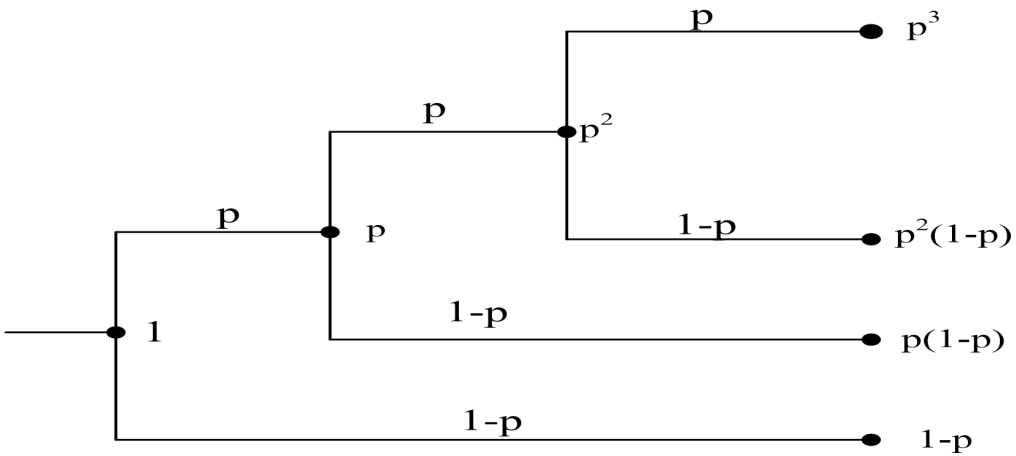
其中，$q(u_i)$ 为节点 u_i 的概率，$\mathrm{H}(u_i)$ 为节点 u_i 的分支熵。

§2.3.4 有根概率树与熵的计算

例 2.15　离散随机变量符号集 $A=\{a_1,a_2,a_3,a_4\}$，概率为 $\{1-p, p(1-p), p^2(1-p), p^3\}$ 所对应的有根概率树如图所示，计算树叶的平均深度和该随机变量的熵。

解　概率树叶的平均深度：$\overline{l}=1+p+p^2$，

随机变量的熵：$H=H(p)+pH(p)+p^2H(p)=(1+p+p^2)H(p)$



§2.3.4 有根概率树与熵的计算

例2.16　利用熵的可加性或有根概率树，计算 $H(1/3,1/3,1/6,1/6)$ 。

解法1：将两个1/6概率合并成一个概率，再将三个概率合并，形成树根，有

$$H(1/3,1/3,1/6,1/6)=H(1/3,1/3,1/3)+1/3=1.918 \text{ 比特/符号}$$

解法2：将两个1/3概率分解为两个1/6概率的和，形成扩展树，有 $\log_2 6=H(1/3,1/3,1/6,1/6)+(2/3)\log_2 2$

得 $H(1/3,1/3,1/6,1/6)=1.918$　比特/符号

§2.3.4 有根概率树与熵的计算

例2.17　一离散随机变量有9个符号，概率分别为$p_i, i=1,...,9$，其中 $p_i(i=1,...4)=(1-\varepsilon)^2/4, p(i=5,...,8)=\varepsilon(1-\varepsilon)/2, p_9=\varepsilon^2$；计算此离散随机变量的熵。

解：因为$\sum_{i=1}^{4} p_i=(1-\varepsilon)^2, \sum_{i=5}^{6} p_i=\sum_{i=7}^{8}=\varepsilon(1-\varepsilon)$，根据熵的可加性，所求熵为

$$H(p_1,p_2,\cdots,p_9)=H((1-\varepsilon)^2,\varepsilon(1-\varepsilon),\varepsilon(1-\varepsilon),\varepsilon^2)+(1-\varepsilon)^2H(1/4,1/4,1/4,1/4)+\varepsilon(1-\varepsilon)H(1/2,1/2)$$
$$+\varepsilon(1-\varepsilon)H(1/2,1/2)=H((1-\varepsilon)^2,\varepsilon(1-\varepsilon),\varepsilon(1-\varepsilon),\varepsilon^2)+(1-\varepsilon)^2\log 4+2\varepsilon(1-\varepsilon)\log 2$$
$$=H(1-\varepsilon,\varepsilon)+(1-\varepsilon)H(1-\varepsilon,\varepsilon)+\varepsilon H(1-\varepsilon,\varepsilon)+2(1-\varepsilon)^2\log 2+2\varepsilon(1-\varepsilon)\log 2$$
$$=2[H(\varepsilon)+(1-\varepsilon)\log 2]$$

§2.4.1 平均互信息的定义

1. **离散随机变量与事件之间的互信息**

离散随机变量X与Y的某一取值y之间的互信息定义为：

$$I(X;y)=\sum_{x}p(x\mid y)\log\frac{p(x\mid y)}{p(x)} \qquad (2.45)$$

(2.45)表示由y提供的关于X的信息量（注意：用条件概率平均）。

§2.4.1 平均互信息的定义

定理2.9

$$I(X;y)\geq 0 \qquad (2.46)$$

仅当y与所有x 独立时，等式成立。

证：根据散度的定义与散度不等式，有$I(X;y)=D(p(x\mid y)\,//\,p(x))\geq 0$

仅当对所有x，$p(x)=p(x\mid y)$ 时，等式成立。

类似地，也可证明 $I(x;Y)\geq 0$。

§2.4.1 平均互信息的定义

2. **平均互信息**

离散随机变量X、Y之间的平均互信息定义为：

$$
\begin{aligned}
I(X;Y) &= \sum_{x} p(x) I(Y;x) = \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y)} \\
&= \sum_{x,y} p(x) p(y|x) \log \frac{p(y|x)}{\sum_{x} p(x) p(y|x)} \\
&= \sum_{i,j} p_i p_{ij} \log \frac{p_{ij}}{\sum_{i} p_i p_{ij}}
\end{aligned}
\qquad (2.47)
$$

§2.4.1 平均互信息的定义

平均互信息I(X；Y）其实就是互信息I(x；y)在概率空间XY中求统计平均的结果，是从整体上表示一个随机变量Y所提供的关于另一个随机变量X的信息量。平均互信息的单位为：比特（奈特）/符号或比特（奈特）。

平均互信息的概念可扩展到随机矢量之间。随机矢量 X^N, Y^M 之间的平均互信息定义为

$$I(\mathbf{X}^N;\mathbf{Y}^M)=\sum_{\mathbf{x}} p(\mathbf{x})\sum_{\mathbf{y}} p(\mathbf{y}\mid\mathbf{x})\log\frac{p(\mathbf{y}\mid\mathbf{x})}{p(\mathbf{y})} \qquad (2.48)$$

其中 $x=(x_1,x_2,\ldots,x_N), y=(y_1,y_2,\ldots,y_M)$。很明显，当M=N=1时，式(2.48)归结为(2.47)。

§2.4.1 平均互信息的定义

3. 平均互信息与熵的关系

定理2.10 对于离散随机变量X、Y，下面的关系式成立：

$$I(X;Y)=H(X)-H(X|Y) \qquad (2.49)$$

$$I(X;Y)=H(Y)-H(Y|X) \qquad (2.50)$$

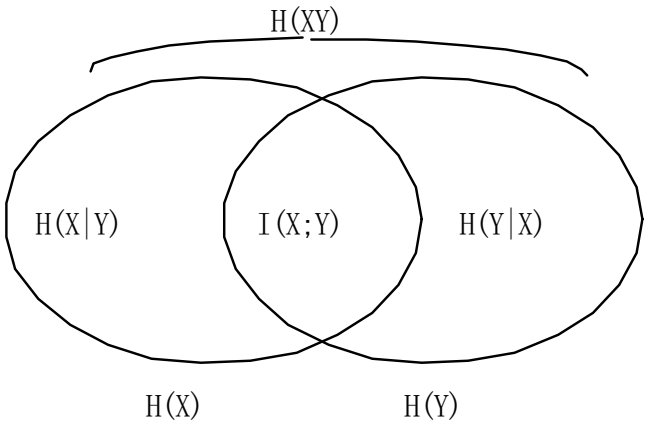
$$I(X;Y)=H(X)+H(Y)-H(XY) \qquad (2.51)$$

由(2.49)式我们可以进一步理解平均互信息的物理意义。在通信系统中，X为信道输入；Y为信道输出；H(X)表示X的不确定性，而H(X|Y)表示接收到Y后关于X的不确定性，平均互信息为二者之差，表示关于X的不确定性的变化，也就是通过Y所获得关于X的信息量。

§2.4.1 平均互信息的定义

4. 香农信息度量与集合运算的关系

上面所介绍的信息熵、条件熵以及平均互信息等统称为香农信息度量，这种度量与集合论中的运算有一一对应的关系。对于两随机变量X、Y，这种对应关系可由图所示的信息图来解释。图中的映射关系为：



$H(X) \to X$，$H(Y) \to Y$，

$H(XY) \to X \cup Y$，$I(X;Y) \to X \cap Y$，

$H(X|Y) \to X \cap Y^C$，$H(Y|X) \to Y \cap X^C$。

§2.4.1 平均互信息的定义

例2.18 对某城市进行交通忙闲的调查，并把天气分成晴雨两种状态，气温分成冷暖两种状态。调查结果得到的各数据联合出现相对频率如表所示，

忙				闲			
晴		雨		晴		雨	
冷	暖	冷	暖	冷	暖	冷	暖
12	8	27	16	8	13	4	12

若把这些频度看作概率测度，求:

（1） 忙闲的无条件熵；

（2） 天气状态和气温状态同时已知时忙闲的条件熵；

（3）从天气状态和气温状态同时获得的关于忙闲的信息量。

§2.4.1 平均互信息的定义

解　令 X={忙，闲}={0, 1}， Y={晴，雨}={0, 1}， Z={冷，暖}={0, 1}，则有下面的联合分布：

XY					
P(xyz)		00	01	10	11
X	0	0.12	0.08	0.27	0.16
	1	0.08	0.13	0.04	0.12
		$p_{yz}(00)=0.2$	$p_{yz}(01)=0.21$	$p_{yz}(10)=0.31$	$p_{yz}(11)=0.28$

(1) $H(\mathrm{X}) = H(0.63,0.37) = 0.951$　比特/符号

(2) $H(X\,|\,YZ) = \mathrm{H}(XYZ) - H(YZ) = H(.12,.08,.27,.16,.08,.13,.04,.12)$

$- H(.2,.21,.31,.28) = 2.819 - 1.976 = 0.843$　比特/符号

(3) $I(X;YZ) = H(X) - H(X/YZ) = 0.951 - 0.843 = 0.108$　比特　/符号

§2.4.2 平均互信息的性质

1. 非负性

定理2.11

$$I(X;Y)\geq 0 \qquad (2.52)$$

仅当X，Y 独立时，等式成立。

证　根据(2.46)式 $I(X;y)\geq 0$，其平均值也大于或等于0。实际上，$I(X;Y)=D(p(xy)\,//\,p(x)p(y))\geq 0$，其中，$p(xy)$ 为XY的联合概率分布，$p(x)p(y)$ 为X和Y概率的乘积。

§2.4.2 平均互信息的性质

2. 对称性

$$I(X;Y)=I(Y;X) \tag{2.53}$$

3. 凸函数性

定理2.12 $I(X;Y)$ 为概率分布 $p(x)$ 的上凸函数。

定理2.13 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。

§2.4.2 平均互信息的性质

例2.19　已知二元随机变量X、Y，输出符号均为$\{0,1\}$，$p_X(0)=\omega$，$(0\le\omega\le1)$，条件概率$p_{Y|X}(0|0)=p_{Y|X}(1|1)=1-p$，$(0\le p\le1)$，求$I(X;Y)$，并讨论其凸函数性。

解：根据题意，有 $p_{Y|X}(1|0)=p_{Y|X}(0|1)=p$

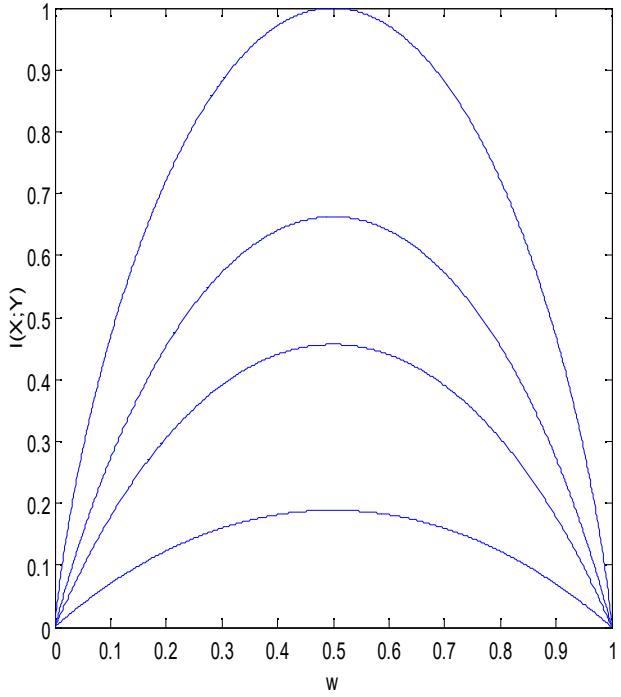
所以
$$\begin{aligned}\begin{pmatrix}p_Y(0) & p_Y(1)\end{pmatrix}&=\begin{pmatrix}\omega & 1-\omega\end{pmatrix}\begin{pmatrix}1-p & p\\ p & 1-p\end{pmatrix}\\&=\begin{pmatrix}p+\omega-2\omega p & 1-p-\omega+2\omega p\end{pmatrix}\end{aligned}$$

故得 $H(Y)=H(p+\omega-2\omega p)$, $H(Y|X)=\omega H(p)+(1-\omega)H(p)=H(p)$

因此 $I(X;Y)=H(p+\omega-2\omega p)-H(p)$

§2.4.2 平均互信息的性质

根据定理2.12 可知，$I(X;Y)$ 是 ω 的上凸函数，选择若干 p 值所做的 $I(X;Y)$ 的曲线如图所示。当 $p=1/2$时，$I(X;Y)=0$；当$p\neq 1/2$ 有如下结论：



（1）因$I(\omega,p)=I(\omega,1-p)$，所以p和1−p对应的是同一条曲线；

（2）因$I(\omega,p)=I(1-\omega,p)$，所以曲线关于$\omega=1/2$对称；

（3）当 $p+\omega-2\omega p=1/2$ 时，有 $(1-2\omega)(p-1/2)=0$，所以当$\omega=1/2$ 时，$I(X;Y)=\log 2-H(p)$，达到极大值，当 $\omega=0$ 或1时，$I(X;Y)$ 取最小值0。

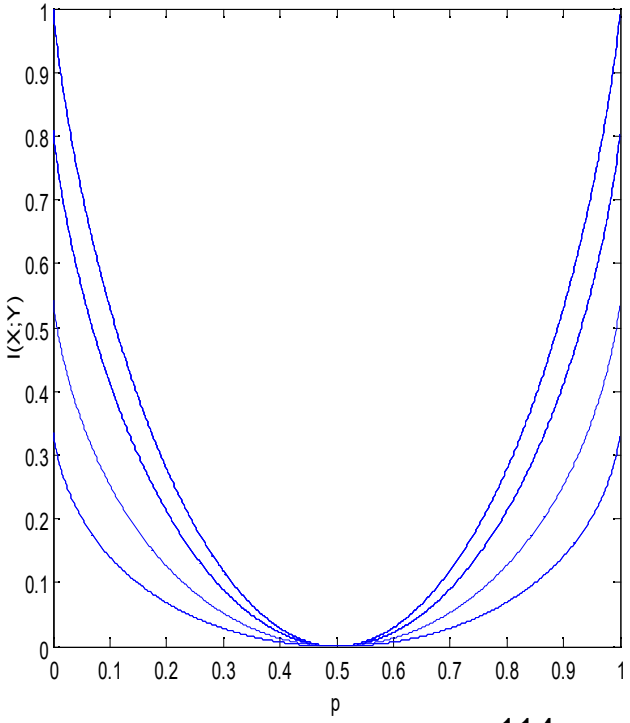
§2.4.2 平均互信息的性质

根据定理2.13 可知，$I(X;Y)$ 是p的下凸函数，选择若干 ω 值所做的 $I(X;Y)$ 的曲线如图所示。当 $\omega=0$或1时$I(X;Y)=0$；当 $\omega\neq 0$或1时有如下结论：

（1）因 $I(\omega,p)=I(1-\omega,p)$，所以 ω 和 $1-\omega$ 对应的是同一条曲线；

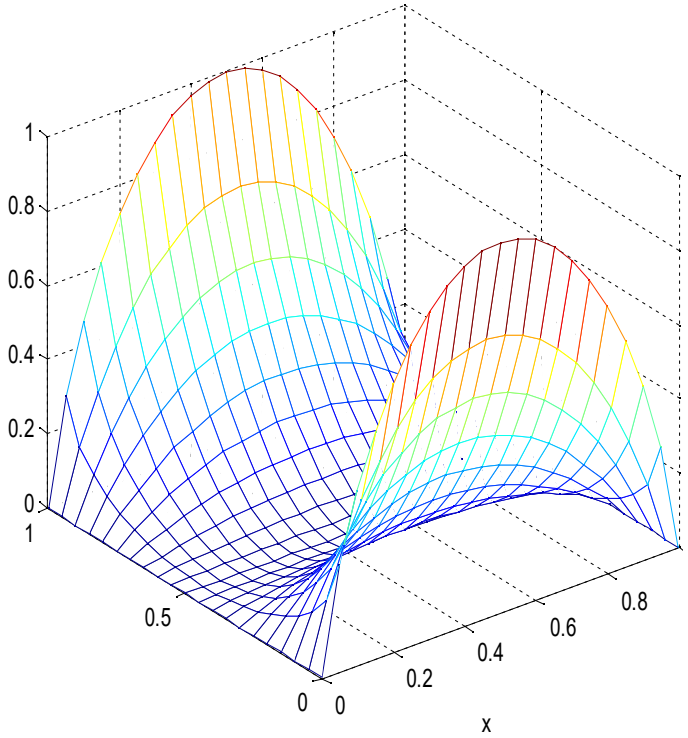
（2）因 $I(\omega,p)=I(\omega,1-p)$，所以曲线关于 $p=1/2$ 对称；

（3）当 $p=0$ 或1时，$I(X;Y)=H(\omega)$或 $H(1-\omega)$ 达到最大值，当 $p=1/2$ 时，$I(X;Y)$取最小值0。



§2.4.2 平均互信息的性质

为更好理解本题 $I(X;Y)$ 的凸函数性，将二元函数 $I(\omega, p)$的图形示于图。将 ω, p 视为函数的两个坐标轴，凸函数性描述函数在不同坐标轴上所表现的性质，这些表现有时可能有很大的差别。



§2.4.2 平均互信息的性质

4．极值性

定理2.14

$$I(X;Y) \le H(X) \qquad (2.56)$$

$$I(X;Y) \le H(Y) \qquad (2.57)$$

平均互信息的极值性说明从一个事件提取关于另一个事件的信息量，至多是另一个事件的熵，不会超过另一个事件自身所含的信息量。

§2.4.3 平均条件互信息

设联合集XYZ，在Z 条件下，X与Y之间的平均互信息定义为条件互信息 $I(x;y|z)$ 的平均值，即

$$I(X;Y|Z)=\underset{p(xyz)}{E}[I(x;y|z)]=\underset{p(xyz)}{E}\{\log\frac{p(x|yz)}{p(x|z)}\}=\sum_{x,y,z}p(xyz)\log\frac{p(x|yz)}{p(x|z)}$$

由于

$$\begin{aligned}I(x;yz)&=\log\frac{p(x|yz)}{p(x)}=\log\frac{p(x|yz)}{p(x|z)}\frac{p(x|z)}{p(x)}\\&=I(x;y|z)+I(x;z)\end{aligned}\qquad(2.59)$$

同理可得

$$I(x;yz)=I(x;z|y)+I(x;y)\qquad(2.60)$$

对（2.59），（2.60）两边求平均，得

$$\begin{aligned}I(X;YZ)&=I(X;Z|Y)+I(X;Y)\\&=I(X;Y|Z)+I(X;Z)\end{aligned}\qquad(2.61)$$

§2.4.3 平均条件互信息

定理2.15 平均条件互信息是非负的，即

$$I(X;Y|Z)\geq 0 \qquad (2.62)$$

仅当$p(x|z)=p(x|yz)$时，等式成立。

证：

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

$$=\sum_{x,y,z}p(xyz)\log\frac{p(xyz)}{p(x|z)p(yz)}$$

$$=D(p(xyz)//p(x|z)p(yz))\geq 0$$

仅当$p(x/z)=p(x/yz)$时，等式成立。

§2.4.3 平均条件互信息

定理2.16

$$I(X;YZ) \geq I(X;Z) \tag{2.63}$$

仅当$p(x|z) = p(x|yz)$时，等式成立

$$I(X;YZ) \geq I(X;Y) \tag{2.64}$$

仅当$p(x|y) = p(x|yz)$时，等式成立。

证：由(2.61)和(2.62)，可得(2.63)式。将Y、Z的位置互换可得(2.64)。

§2.4.3 平均条件互信息

设Y、Z为独立随机变量，其中Y取n个符号, Z取k个符号，则YZ含nk个符号。Z可看成YZ中某些符号的合并处理，由nk个符号合并成k个符号。因此：

（1）YZ中的符号进行合并处理后，使其获得的关于X的信息量减少；

（2）如果YZ在二维空间取值，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。

（3）通过对Z或Y取值空间的细化，可使其获得的关于X的信息量增加。

§2.4.3 平均条件互信息

定理2.15（平均互信息的链原则）

$$I(X_1X_2\ldots X_n;Y)=\sum_{i=1}^{n}I(X_i;Y\mid X_1X_2\ldots X_{i-1}) \qquad (2.65)$$

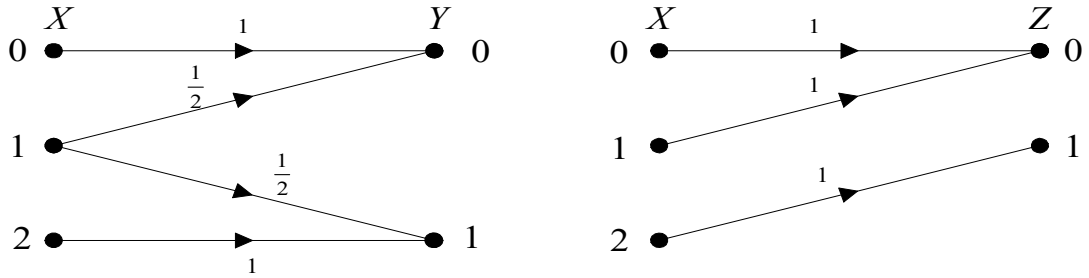
证

$$\begin{aligned}I(X_1X_2\ldots X_n;Y)&=H(X_1X_2\ldots X_n)-H(X_1X_2\ldots X_n\mid Y)\\&=\sum_{i=1}^{n}H(X_i\mid X_1\ldots X_{i-1})-\sum_{i=1}^{n}H(X_i\mid X_1\ldots X_{i-1}Y)\\&=\sum_{i=1}^{n}I(X_i;Y\mid X_1X_2\ldots X_{i-1})\end{aligned}$$

上述证明利用了熵的可加性，(2.65)称做平均互信息的链原则。

§2.4.3 平均条件互信息

例2.20 设信源X的符号集为{0，1，2}，其概率分布为 $P_0 = P_1 = 1/4$ ，$P_2 = 1/2$ ，每信源符号通过两个信道同时传输，输出分别为Y，Z，两信道转移概率如图所示；求



（1）H(Y)，H(Z)；

（2）H(XY)，H(XZ)，H(YZ)，H(XYZ)；

（3）I(X;Y)，I(X;Z)，I(Y;Z)；

（4）I(X;Y|Z)，I(X;YZ)。

§2.4.3 平均条件互信息

解：设y、z的符号集均为{0，1}，yz符号集为{00，01，10，11}，由已知条件得

$X-Y$条件概率矩阵：$\begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$，$X-Z$ 条件概率矩阵：$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$

和 $H(X)=H(1/4,1/4,1/2)$ 比特/符号

因为在x给定条件下y、z独立，即$p(yz|x)=p(y|x)p(z|x)$，故有

$X-YZ$ 条件概率矩阵：$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

§2.4.3 平均条件互信息

（1） 由$(p(y=0) \quad p(y=1)) = \left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{2}\right) \times \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} = \left(\frac{3}{8} \quad \frac{5}{8}\right)$，得

$$H(Y) = H(3/8, 5/8) = 0.955\text{比特/符号}$$

由$(p(z=0) \quad p(z=1)) = \left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{2}\right) \times \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \left(\frac{1}{2} \quad \frac{1}{2}\right)$，得

$$H(Z) = H(1/2) = 1\text{比特/符号}$$

§2.4.3 平均条件互信息

（2）由$H(Y|X)=(1/4)H(1)+(1/4)H(1/2)+(1/2)H(1)]=0.25$，得

$H(XY)=H(X)+H(Y|X)=1.5+0.25=1.75$比特/2个符号

由$H(Z|X)=(1/4)H(1)+(1/4)H(1)+(1/2)H(1)]=0$，得

$H(XZ)=H(X)+H(Z|X)=1.5+0=1.5$比特/2个符号

由$\begin{pmatrix} p(yz=00) & p(yz=01) & p(yz=10) & p(yz=11) \end{pmatrix}$

$$=\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}\times\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}=\begin{pmatrix} \frac{3}{8} & 0 & \frac{1}{8} & \frac{1}{2} \end{pmatrix}$$，得

$H(YZ)=H(3/8,1/8,1/2)=1.406$比特/2个符号

由$H(YZ|X)=(1/4)H(1)+(1/4)H(1/2)+(1/2)H(1)]=0.25$，得

$H(XYZ)=H(X)+H(YZ|X)=1.5+0.25=1.75$比特/3个符号

§2.4.3 平均条件互信息

（3）$I(X;Y)=H(X)+H(Y)-H(XY)=1.5+0.955-1.75=0.705$比特

$I(X;Z)=H(X)+H(Z)-H(XZ)=1.5+1-1.5=1$比特/符号

$I(Y;Z)=H(Y)+H(Z)-H(YZ)=0.955+1-1.406=0.549$ 比特

（4）$I(X;Y|Z)=H(X|Z)-H(X|YZ)=H(XZ)-H(Z)-H(XYZ)+H(YZ)$

$=1.5-1-1.75+1.406=0.156$比特/符号

$I(X;YZ)=I(X;Z)+I(X;Y|Z)=1+0.156=1.156$ 比特

本章小结

1．自信息、联合自信息与条件自信息

自信息 $I(x) = -\log p(x)$

联合自信息 $I(x) = -\log p(x_1 x_2 \ldots x_N)$

条件自信息 $I(y|x) = -\log p(y|x)$

2．信息熵、联合熵、条件熵与相对熵

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

联合熵 $\mathrm{H}(X^N) = H(X_1 X_2 \ldots X_N) = E_{p(x)}[-\log p(x)]$

条件熵 $H(Y|X) = E_{p(\mathbf{xy})}[-\log p(y|x)]$

多维条件熵 $H(\mathbf{Y}^M|\mathbf{X}^N) = E_{p(\mathbf{x},\mathbf{y})}[\log p(\mathbf{y}|\mathbf{x})]$

相对熵（信息散度） $D(P//Q) = E_{p(xy)}\{\log[p(x)/Q(x)]\}$

本章小结

3．互信息与条件互信息

互信息 $I(x;y)=\log[p(y\mid x)/\,p(y)]$

条件互信息 $I(x;y\mid z)=\log[p(y\mid xz)/\,p(y\mid z)]$

平均互信息 $I(X;Y)=E_{p(xy)}[\log(p(y\mid x)/\,p(y))]$

矢量平均互信息 $I(\mathbf{X}^N;\mathbf{Y}^M)=E_{p(\mathbf{x},\mathbf{y})}[\log(p(\mathbf{y}\mid\mathbf{x})/\,p(\mathbf{y}))]$

条件平均互信息 $I(X;Y\mid Z)=E_{p(xyz)}[\log(p(y\mid xz)/\,p(y\mid z))]$

本章小结

4．上凸函数的性质（Jensen不等式）

$$f(E\mathbf{x}) \geq E[f(\mathbf{x})]$$

5．　离散熵的重要性质：①非负性，②不增原理，③可加性，④极值性，⑤可逆变换下熵的不变性

6．平均互信息重要性质：①非负性，②凸函数性，③极值性

7．有根概率树的性质与熵的计算

谢谢!