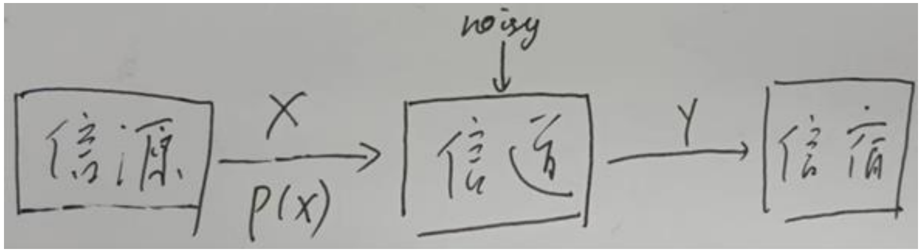
2022-2023学年秋季学期

信息论与编码
第三讲

❍2.1.8 互信息



H(X)为输入变量X的先验不确定量。

$$H(X) = -\sum_{x \in X} p(x) \cdot \log p(x) = -\sum_{i=1}^{n} p(x_i) \cdot \log p(x_i)$$

- 若没有噪声，则信道输出符号与输入符号一一对应，接收到符号后就消除了发送符号的不确定性。

- 但是一般信道中有噪声存在，这导致在接收到Y后对X仍有不确定性。

问题：如何度量“接收到Y后对X仍有不确定性”？

确定了多少，即，熵损失多少？

收到 y 后输入 x 的后验概率为$p(x|y)$。

此时，X 的平均不确定性为

$$H(X|y_i) = -\sum_{i=1}^{n} p(x_i|y_j)\log p(x_i|y_j),$$

这是收到y_j后 X 的后验熵，

即，收到y_j后，关于输入 x 的信息度量。

由极值性，$H(X|Y) \le H(X)$，$H(Y|X) \le H(Y)$，

知

$$\begin{aligned} &H(XY) = H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \le H(X) + H(Y)。\end{aligned}$$

于是，

$X \to Y$：Y 所能接收到的 X 的确定信息为

$H(X) - H(X|Y) = H(Y) - H(Y|X)$

即为互信息

1. 互信息量

称一个事件y所给出的关于另一事件x的信息为“y对x的”互信息量，表示为$I(x;y)$。

定义其为事件x的后验概率与先验概率比值的对数，

即，

$$I(x;y) = \log \frac{p(x|y)}{p(x)}。$$

显然，$I(x;y) = I(x) - I(x|y)$。

类似，可定义“x对y的”互信息量为

$$I(y;x) = \log \frac{p(y|x)}{p(y)} = I(y) - I(y|x)。$$

由于$p(x|y)=\frac{p(xy)}{p(y)}$，故

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

注：由于无法判断$p(x|y)$与$p(x)$的大小关系，因此$I(x;y)$不一定是否不小于 0。

2. 条件互信息量：

对于联合分布 XYZ 中，在给定$z\in Z$的条件下，x 与 y 之间的互信息量定义如下条件互信息量，记为$I(x;y|z)$：

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

3. 平均互信息量：

- 互信息量表示某一事件发生后给出的关于另一事件的信息量，随着x与y的变化而变化。
- 为了从整体上表示一个随机变量Y所给出的关于另一个随机变量X的信息量，我们引入平均互信息量。

称互信息量$I(x;y)$在 XY 的联合概率空间中的统计平均值为平均互信息量，表示为$I(X;Y)$，

即，$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{p(x|y)}{p(x)}$。

称之为Y对X的平均互信息量。

同理，定义 X 对 Y 的平均互信息量为

$$I(Y;X)=\sum_{y\in Y}\sum_{x\in X}p(xy)\log\frac{p(y|x)}{p(y)}\text{。}$$

由于$p(x|y)=\frac{p(xy)}{p(y)}$，知

$$I(X;Y)=\sum_i\sum_j p(x_iy_j)\log\frac{p(x_iy_j)}{p(x_i)p(y_j)}$$
$$=H(X)+H(Y)-H(XY)=I(Y;X)$$

- **说明：**

1）Y 对 X 的平均互信息量表示接收到输出 Y 的前后 X 的平均不确定度减少的量，即，从 Y 获得 X 的平均信息量。

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{p(x|y)}{p(x)}$$
$$=\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{1}{p(x)}-\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{1}{p(x|y)}$$
$$=H(X)-H(X|Y)$$

2）$I(Y;X)=\sum_{y\in Y}\sum_{x\in X}p(xy)\log\frac{p(y|x)}{p(y)}=H(Y)-H(Y|X)$。

即，发出信源 X 前后，Y 的平均不确定度的减少量。

3）

$$I(X;Y)=\sum_{i}\sum_{j}p(x_iy_j)\log\frac{p(x_iy_j)}{p(x_i)p(y_j)}=H(X)+H(Y)-H(XY)$$

表示通信前后整个系统不确定性的减少量。

4）$I(X;Y)=I(Y;X)$表示“接收到输出 Y 的前后 X 的平均不确定度减少的量”=“发出信源 X 前后，Y 的平均不确定度的减少量”。

5）$I(X;Y)=I(Y;X)=0$当且仅当 X 与 Y 互相独立。

6）$I(X;Y)=I(Y;X)\le min\{H(X),H(Y)\}$。

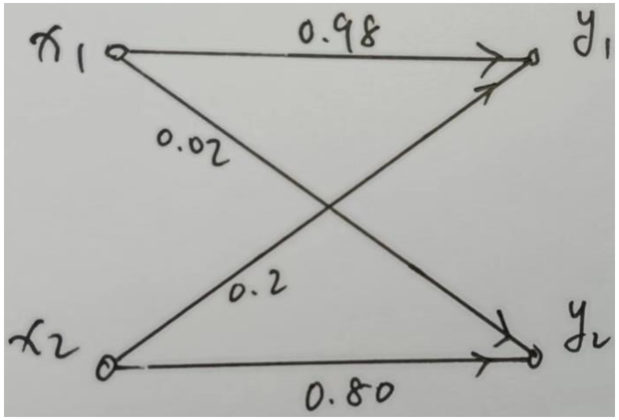
❍噪声存在，这导致在接收到Y或X后对X或Y仍有不确定性。

- 信道疑义度：称条件熵
  $H(X|Y)=\sum_{y\in Y}p(y)H(X|Y=y)=-\sum_{x\in X}\sum_{y\in Y}p(xy)\log p(x|y)$ 为信道的疑义度，表示收到 Y 后 X 的不确定性。
- 噪声熵：称$H(Y|X)=H(Y)-I(X;Y)$为噪声熵，表示为已知 X 的条件下 Y 尚存的不确定性，这由信道中噪声引起的。

❍例2.12 已知信源

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

在如图所示的信道上，求该信道上传输的平均互信息量$I(X;Y)$、信道疑义度$H(X|Y)$、噪声熵$H(Y|X)$和联合熵$H(XY)$。



解： 已知

$p(x_1) = \frac{1}{2} = p(x_2)$，$p(y_1|x_1) = 0.98$，$p(y_2|x_1) = 0.02$，

$p(y_1|x_2) = 0.20$，$p(y_2|x_2) = 0.80$，

于是

$p(x_1y_1) = p(x_1)p(y_1|x_1) = \frac{1}{2} * 0.98 = 0.49$，

$p(x_1y_2) = 0.01$，$p(x_2y_1) = 0.10$，$p(x_2y_2) = 0.40$.

从而可计算

$$p(y_1) = \sum_{i=1}^{2} p(x_iy_1) = 0.59，\quad p(y_2) = 1 - p(y_1) = 0.41$$

再由$p(x|y) = \frac{p(xy)}{p(y)}$，知

$p(x_1|y_1) = 0.831$，$p(x_2|y_1) = 1 - p(x_1|y_1) = 0.169$；

$p(x_1|y_2) = 0.024$，$p(x_2|y_2) = 0.976$.

直接计算即可。

于是，

$$H(X) = -\sum_{i=1}^{2} p(x_i)\cdot\log p(x_i) = 1(\text{bits/symbol})$$

$$H(Y) = -\sum_{i=1}^{2} p(y_i)\cdot\log p(y_i) \approx 0.98(\text{bits/symbol})$$

$$H(XY) = -\sum_{i=1}^{2}\sum_{j=1}^{2} p(x_i y_j)\cdot\log p(x_i y_j) \approx 1.43(\text{bits/symbol})$$

平均互信息

$$I(X;Y) = H(X) + H(Y) - H(XY) = o.55(\text{bits/symbol})$$

信道疑义度$H(X|Y) = H(X) - I(X;Y) = 0.45(\text{bits/symbol})$

噪声熵$H(Y|X) = H(Y) - I(X;Y) = 0.43(\text{bits/symbol})$

❍4. 平均互信息的性质：

1）非负性：$I(X;Y)\geq 0$；当且仅当 X 和 Y 相互独立时，即，$p(xy)=p(x)p(y)$对所有 x 和 y 成立，式中等号成立。

证： $I(X;Y)=H(X)-H(X|Y)$， 又 $H(X|Y)\leq H(X)$， 故 $I(X;Y)\geq 0$。

等号成立当且仅当$H(X|Y)=H(X)$，即，X 与 Y 相互独立。

❍**说明**：给定随机变量Y后，一般地总能消除一部分关于X的不确定性。

即，

从一个事件提取关于另一个事件的信息，最坏情况是0；不会由于知道了一个事件反而使另一个事件的不确定性增加。

2）对称性：

$$I(X;Y)=I(Y;X)$$

证明：由于$p(x|y)=\frac{p(xy)}{p(y)}$，知

$$I(X;Y)==\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{p(x|y)}{p(x)}$$

$$=\sum_{i}\sum_{j}p(x_iy_j)\log\frac{p(x_iy_j)}{p(x_i)p(y_j)}$$

$$=\sum_{y\in Y}\sum_{x\in X}p(xy)\log\frac{p(y|x)}{p(y)}=I(Y;X)。$$

说明：

“从Y获得的关于X的信息量”等于”从X获得的关于Y的信息量”。

3）极值性：

$$I(X;Y)=I(Y;X)\le min\{H(X),H(Y)\}。$$

证明：由$H(X|Y)\le H(X)$和$H(Y|X)\le H(Y)$，以及 $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$，即可。

- **说明：**

“从一个事件获得的关于另一个事件的信息量”至多只能是另一个事件的平均自信息量，不会超过另一事件本身所含信息量。

4）凸性：

称条件概率$p(y|x)$为信道的转移概率。由平均互信息量的定义知

$$I(X;Y)=I(Y;X)=\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{p(y|x)}{p(y)}$$

$$=\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log\frac{p(y|x)}{p(y)}$$

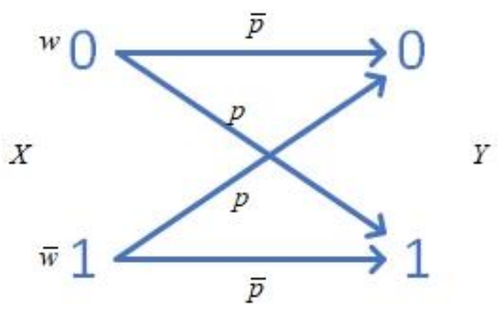
$$=\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log\frac{p(y|x)}{\sum\limits_{x\in X}p(x)p(y|x)}$$

这是关于$p(x)$和$p(y|x)$的函数，即，$I(X;Y)=f(p(x),p(y|x))$。

若信道固定，则转移概率固定，即，常值，则$I(X;Y)=f(p(x))$。

若信源固定，则$I(X;Y)=f(p(y|x))$。

例2.1 二元对称信道(如图)



信源为$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ w & \overline{w} \end{pmatrix}$，计算其平均互信息量。

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_i \sum_j p(x_i y_j) \log \frac{1}{p(y_j|x_i)}$$

$$= H(Y) - \sum_{i=1}^{2} p(x_i) \sum_{j=1}^{2} p(y_j|x_i) \log \frac{1}{p(y_j|x_i)}$$

由于

$$\sum_{j=1}^{2} p(y_j|x_i) \log \frac{1}{p(y_j|x_i)} = \begin{cases} -p\log p - \overline{p}\log \overline{p} = H(p), & \text{when } i = 1 \\ -p\log p - \overline{p}\log \overline{p} = H(p), & \text{when } i = 2 \end{cases}$$

于是

$$I(X;Y) = H(Y) - \sum_{i=1}^{2} p(x_i) \sum_{j=1}^{2} p(y_j|x_i) \log \frac{1}{p(y_j|x_i)}$$

$$= H(Y) - \sum_{i=1}^{2} p(x_i) H(p)$$

$$= H(Y) - H(p)$$

下面计算H(Y):

$$p(Y=0)=p(X=0)p(Y=0|X=0)+P(X=1)p(Y=0|X=1)$$

$$=w\overline{p}+\overline{w}p$$

$$p(Y=1)=p(X=0)p(Y=1|X=0)+P(X=1)p(Y=1|X=1)$$

$$=wp+\overline{wp}$$

所以

$$I(X;Y)=H(p(Y=0),p(Y=1))-H(p)$$

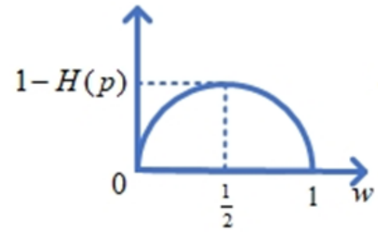
$$=H(w\overline{p}+\overline{w}p)-H(p)$$

。

说明：

当信道固定时，即，p 固定，则 $I(X;Y)$ 仅是 w 的函数 $I(X;Y)=f(w)$，其函数图像如图，此时为上凸函数，且满足：

$$\begin{cases} w=0,1 \Rightarrow f(w)=0 \\ w=\dfrac{1}{2} \Rightarrow f(w)=1-H(p) \end{cases}$$



即，在信道固定时，选择不同的信源（符号概率分布不同）与信道连接，在信道输出端接收到每个符号后，所获得的信息量是不同的，而且存在一种信源使得$I(X;Y)$最大。

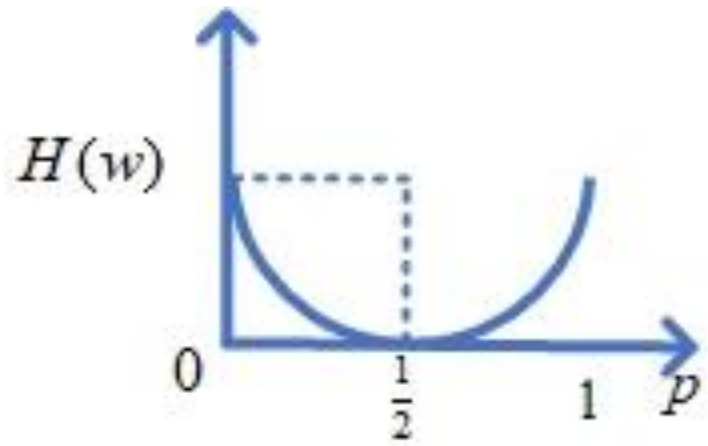
当信源固定时，即，w 固定，

记$I(X;Y)=H(w\overline{p}+\overline{w}p)-H(p)=f(p)$，

此时为下凸函数，且满足：

$$\begin{cases} p=0,1 \Rightarrow f(p)=H(w) \\ p=\dfrac{1}{2} \Rightarrow f(p)=0 \end{cases}$$

即，存在信道使得互信息为 0。



由平均互信息量的定义知

$$I(X;Y)=I(Y;X)=\sum_{x\in X}\sum_{y\in Y}p(xy)\log\frac{p(y|x)}{p(y)}$$

$$=\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log\frac{p(y|x)}{p(y)}$$

$$=\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log\frac{p(y|x)}{\sum\limits_{x\in X}p(x)p(y|x)}$$

这是关于$p(x)$和$p(y|x)$的函数，即，$I(X;Y)=f(p(x),p(y|x))$。

若信道固定，则转移概率固定，即，常值，则$I(X;Y)=f(p(x))$。

若信源固定，则$I(X;Y)=f(p(y|x))$。

定理2.1 当条件概率分布$P(Y|X)$固定时，$I(X;Y)$是输入信源概率分布$p(x)$的严格上凸函数。

说明：当信道固定时，一定存在一个最佳信源输入分布$P(X)$，使得$I(X;Y)$的值最大。

所谓上凸性函数是指，同一信源集合$\{x_1, \cdots, x_n\}$，对于任两个不同的概率分布 $P_1(X) = \{p_1(x_1), \cdots, p_1(x_n)\}$ 和 $P_2(X) = \{p_2(x_1), \cdots, p_2(x_n)\}$，对于小于 1 的正数$0 < \alpha, \beta < 1$，满足$\alpha + \beta = 1$，有

$$f(\alpha p_1(x_i) + \beta p_2(x_i)) \geq \alpha f(p_1(x_i)) + \beta f(p_2(x_i))。$$

称f为$P(X) = \{p(x_1), \cdots, p(x_n)\}$的上凸函数。（若是大于号，则是严格上凸。此外，对于下凸函数，类似定义。）

定理2.1 当输入分布$P(X)$固定时，$I(X;Y)$是条件概率分布$p(y|x)$的严格下凸函数。

说明：如果把条件概率分布$P(Y|X)$视为信道的转移概率分布，则对于固定的输入分布，一定存在一种最差信道，此信道的干扰最大，接收者获得的信息量最小。

前一定理是研究信道容量的基础；后一定理是研究信源的信息率失真函数的理论基础。

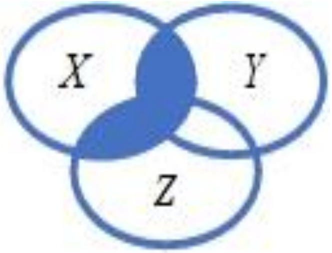
❍2.1.9 数据处理

即高维的数据互信息，我们以三维为例。为了了解数据处理定理，我们需要引入三元随机变量X，Y，Z的平均条件互信息和平均联合互信息。

定义1.1. 平均条件互信息量为

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

这表示给定 Z，从 Y 得到 X 的信息量平均值。



定义1.1. 平均联合互信息量为

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

这表明从二维随机变量 YZ 所得到的关于随机变量 X 的信息量。

于是，有下列等式成立：

$$I(X;YZ) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$$

$$I(XY;Z) = I(Y;Z) + I(X;Z|Y) = I(X;Z) + I(Y;Z|X)$$

马尔可夫(Markov)链

对随机变量序列$\cdots, \mathcal{S}_{i-2}, \mathcal{S}_{i-1}, \mathcal{S}_i, \mathcal{S}_{i+1}, \mathcal{S}_{i+2}, \cdots$，每个$\mathcal{S}_i$事件发生的可能性只与其前后相邻的事件$\mathcal{S}_{i-1}$和$\mathcal{S}_{i+1}$有关。通过这种关联传递，使得距离很远的两个状态变量之间也能关联起来形成一条链$\cdots \to \mathcal{S}_{i-1} \to \mathcal{S}_i \to \mathcal{S}_{i+1} \to \cdots$。

定理2.1 [数据处理定理、信息不增定理] 若 X，Y，Z 构成马尔可夫链，即，$X \to Y \to Z$.则下列关系成立：$I(X;Z) \le I(X;Y)$和$I(X;Z) \le I(Y;Z)$。等号成立的条件是 对任意 x,y,z，有$p(x|yz) = p(x|z)$和$p(z|xy) = p(z|x)$。

此时，若 Y 给定，X 与 Z 的互信息为 0。

说明：

1） $I(X;Z) \leq I(X;Y)$表明从 Z 所得到的关于 X 的信息量不超过从 Y 所得到的 X 的信息量。

2) 若将$Y \to Z$视为数据处理系统，则通过数据处理后，虽然可以满足某种要求，但是，从信息量来看，处理后会损失一部分信息，最多保持原有的信息，即，对收到的 Y 处理后，绝不会减少关于 X 的不确定性。此即为数据处理定理。

3) 整个定理说明后面得到前面的信息量逐渐减少，每次处理总会丢失一些信息在任何传输信息系统中，最后获得的信息至多是信源所提供的信息。一旦在某一过程中丢失一些信息，以后的系统不管如何处理，如果不触及丢失信息的输入端，就不能再恢复已丢失的信息。此为信息不增定理。□

❍熵关系小结：

$$I(X;Y) = H(X) - H(X|Y)$$

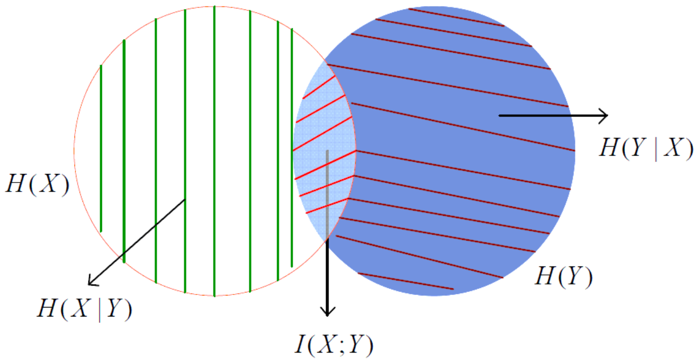
1） $= H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$;

2）若信道无噪声，则$p(x|y) = p(y|x)$=1，即，输入与输出完全一一对应。此时，H(X|Y)=0=H(Y|X)且$I(X;Y) = H(X) = H(Y)$。

3）$H(XY) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

$H(Y|X) \leq H(Y)$，$H(X|Y) \leq H(X)$

$H(XY) \leq H(X) + H(Y)$

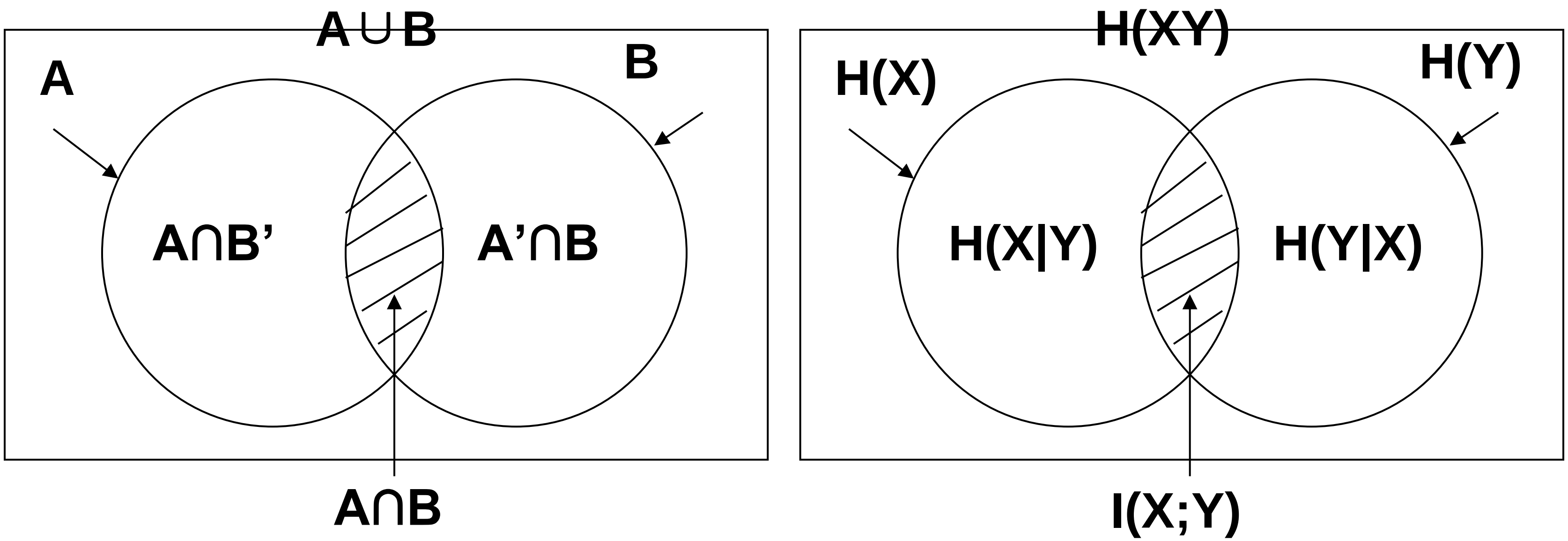


3）对三元 X, Y, Z，有如下关系：

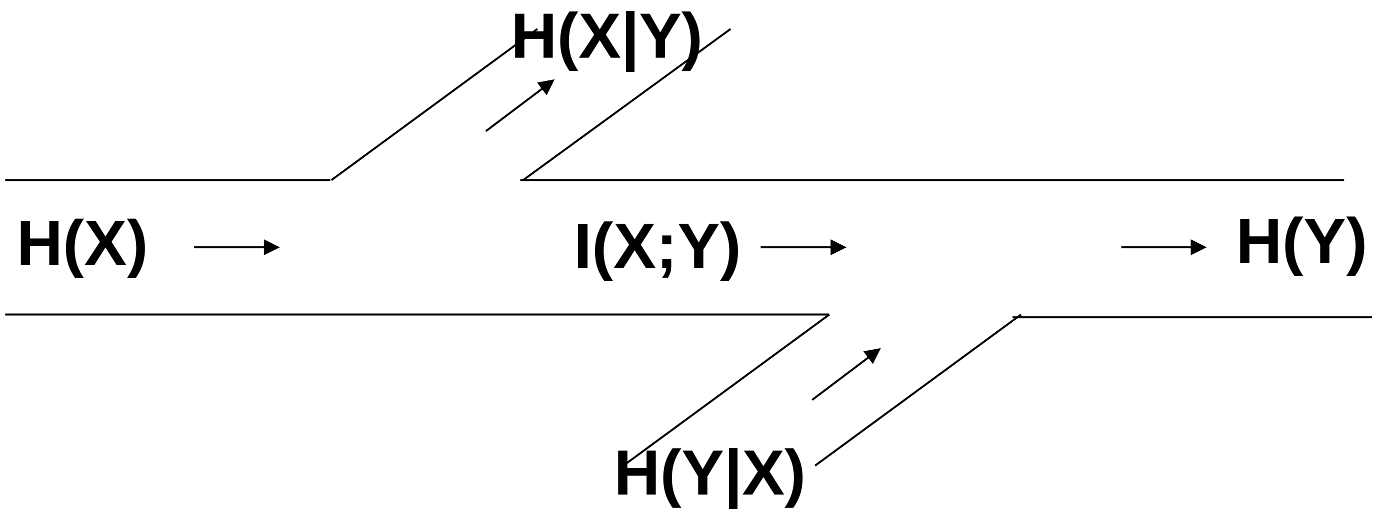
$$I(X;YZ) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$$

$$I(XY;Z) = I(Y;Z) + I(X;Z|Y) = I(X;Z) + I(Y;Z|X)$$

各类熵与集合图的类比

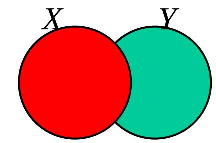
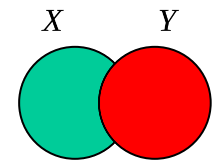
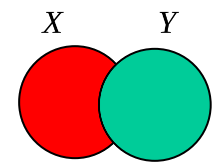
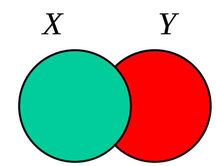
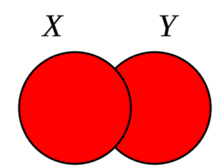
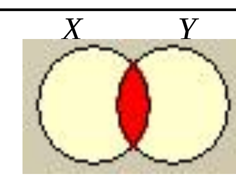


信道中熵的信息流图



- H(Y|X) ：噪声熵；
- H(X|Y) ：信道疑义度；
- 它们都是由于噪声干扰的存在而存在的。信道中存在噪声干扰，是减低信道传信能力的基本原因。

各种熵之间的关系

名称	符号	关　系	图　示
无条件熵	$H(X)$	$H(X)\ge H(X/Y)$ $=H(X/Y)+I(X;Y)$ $H(X)=H(XY)-H(Y/X)$	
	$H(Y)$	$H(Y)\ge H(Y/X)$ $=H(Y/X)+I(X;Y)$ $H(Y)=H(XY)-H(X/Y)$	
条件熵	$H(X/Y)$	$H(X/Y)=H(XY)-H(Y)$ $=H(X)-I(X;Y)$	
条件熵	$H(Y/X)$	$H(Y/X)=H(XY)-H(X)$ $=H(Y)-I(X;Y)$	
联合熵	$H(XY)=H(YX)$	$H(XY)=H(X)+H(Y/X)$ $=H(Y)+H(X/Y)$ $=H(X)+H(Y)-I(X;Y)$ $=H(X/Y)+H(Y/X)+I(X;Y)$	
交互熵	$I(X;Y)=I(Y;X)$	$I(X;Y)=H(X)-H(X/Y)$ $=H(Y)-H(Y/X)$ $=H(XY)-H(Y/X)-H(X/Y)$ $=H(X)+H(Y)-H(XY)$	

❍2.2. 扩展信源

前面讨论单符号离散信源，但现实中信源输出消息是时间或空间的离散序列，是多符号组成序列，即，每次发出一组含两个以上符号的序列表示一个消息。称这样的信源为离散多符号信源。

用$\vec{X}$表示$X_1 X_2 \cdots X_N \cdots$。

$$\text{离散信源序列}\begin{cases}\text{无记忆的，即，序列中符号之间不相关；}\\ \text{有记忆的，即，序列中符号间有相关性。}\\ \text{（这需要用联合概率和条件概率来描述信源发出的符号之间的关系。）}\end{cases}$$

下面讨论离散无记忆扩展信源和两类较简单的离散有记忆信源—离散平稳信源和Markov信源。

2.2.1 无记忆扩展信源

每次发出一组含两个以上符号的序列表示一个消息，而且所发出的各个符号是互相独立的信源，称为无记忆扩展信源（DM）。该信源特点如下：

1) 一个消息是一个符号序列—符号组；

2) 符号之间相互独立；

3) 各个符号的出现概率是其自身的先验概率。

称序列中符号组的长度为扩展次数(假定有限)。

DM二元二次扩展

二次扩展二元信源：输出的消息是符号序列是分组发出，每组有两个信源符号，则得到新的等效信源，其输出符号为 。如二进制时，则为00，01，10，11.

由此可见，对于一个单符号离散信源

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ p(x_1) & p(x_2) \end{pmatrix},$$

其二次扩展信源的数学模型为

$$\begin{pmatrix} X^2 \\ P(X^2) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ p(a_1) & p(a_2) & p(a_3) & p(a_4) \end{pmatrix} = \begin{pmatrix} x_1x_1 & x_1x_2 & x_2x_1 & x_2x_2 \\ p(x_1x_1) & p(x_1x_2) & p(x_2x_1) & p(x_2x_2) \end{pmatrix}$$

其中$a_i = x_{i_1}x_{i_2}$，$i_1, i_2 \in \{1, 2\}$，且$p(a_i) = p(x_{i_1})p(x_{i_2})$。

❍2、DM的N次扩展

设一个单符号离散信源的数学模型为

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n) \end{pmatrix},$$

满足$\sum_{i=1}^{n} p(x_i) = 1$，则其 N 次扩展信源用$X^N$表示，其数学模型为

$$\begin{pmatrix} X^N \\ P(X^N) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \cdots & a_{n^N} \\ p(a_1) & p(a_2) & \cdots & p(a_{n^N}) \end{pmatrix},$$

满足$\sum_{i=1}^{n^N} p(a_i) = 1$，$a_i = x_{i_1} \cdots x_{i_N}$，$x_{i_j} \in \{x_1, \cdots, x_n\}$。

在 N 次扩展信源X^N中，符号序列组成的向量a_i的各个分量之间是彼此统计独立的，

即，$p(a_i)=\prod_{j=1}^{N} p(x_{i_j})$，$i_1,\cdots,i_N \in \{1,2,\cdots,n\}$。

定义 N 次扩展信源X^N的熵为

$$H(X^N) = -\sum_{i=1}^{n^N} p(a_i)\cdot\log p(a_i) \text{ (bits/symbol)}。$$

由熵的可加性，得

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

再由X_i之间的相互独立性且$H(X_i)=H(X)$，则

$$H(X^N)=\sum_{i=1}^{N} H(X_i)=NH(X)。$$

即，DM N 次扩展的熵是原信源 X 熵的 N 倍。

例 2.2 单符号离散二元信源

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ w & \overline{w} \end{pmatrix}$$，其中$w+\overline{w}=1$。

其 DM 2 次扩展信源为

$$\begin{pmatrix} X^2 \\ P(X^2) \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ p(\alpha_1) & p(\alpha_2) & p(\alpha_3) & p(\alpha_4) \end{pmatrix} = \begin{pmatrix} 00 & 01 & 10 & 11 \\ p(00) & p(01) & p(10) & p(11) \end{pmatrix}$$

且$\sum_{i=1}^{2^2} p(\alpha_i)=1$，

$$p(\alpha_i)=p(x_{i_1})p(x_{i_2})=\begin{cases} w^2, \mathrm{if} x_{i_1}=x_{i_2} \\ w(1-w)\mathrm{if} x_{i_1}\neq x_{i_2} \end{cases}$$。

于是，$H(X^2)=2H(X)$。□

例 2.3 单符号离散信源

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$，其 DM 2 次扩展信源为

$$\begin{pmatrix} X^2 \\ \text{符号列} \\ P(X^2) \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & \alpha_9 \\ x_1x_1 & x_1x_2 & x_1x_3 & x_2x_1 & x_2x_2 & x_2x_3 & x_3x_1 & x_3x_2 & x_3x_3 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} \end{pmatrix}$$

于是可以计算

$$H(X) = -\sum_{i=1}^{3} p(x_i) \log p(x_i) = -\tfrac{1}{2} \log \tfrac{1}{2} - \tfrac{1}{4} \log \tfrac{1}{4} - \tfrac{1}{4} \log \tfrac{1}{4}$$

$$= -\tfrac{3}{2}(\text{bits/symbol})$$

$$H(X^2) = -\sum_{i=1}^{3^2} p(a_i) \log p(a_i) = 3(\text{bits/symbol}) = 2H(X)$$。 □

2.2.2 离散平稳信源

- 前面讨论的DM信源及其扩展信源太理想化，而现实中大多数信源不会如此简单。一般地，离散信源的输出是随空间和时间变化而变化的离散符号序列，通常有如下特性：
- 1) 序列中符号之间有依赖关系，这可用联合概率分布的函数来描述（即，有记忆）；
- 2) 序列的统计特性可能会随时间而改变，即，不同时刻输出的序列的概率分布可能不同（即，非平稳性）。

一般来说，信源的前后消息之间有前后依赖关系，可以用随机矢量描述：

$$X = (\ldots, X_1, X_2, \ldots X_i, \ldots)$$

信源在某一时刻发出什么样的值取决于两方面

1、这一时刻该变量的概率分布

2、这一时刻以前发出的消息：如一个人讲话

- 下面考虑一种特殊信源，即，统计特性无变化或在较短时间内统计特性稳定的信源，称为平稳信源；否则，为非平稳信源。显然，平稳信源中符号序列的概率分布与时间起点无关。

我们现在讨论平稳的随机序列.

所谓平稳是指序列的统计性质与时间的推移无关（两个任意时刻信源发出符号的概率分布完全相同）。信源所发符号序列的概率分布与时间的起点无关，这种信源我们称之为离散序列平稳信源。

- 对于随机变量序列$X=X_1X_2\cdots X_N$，若任意两个不同时刻i j,信源发出消息的概率分布完全相同，则称这种信源为一维平稳信源。

- 除上述条件外，如果联合概率分布也与时间起点无关，则称信源为二维平稳信源。这种信源在任何时刻发出两个符号的概率完全相同。

- 各维联合概率均与时间起点无关的完全平稳信源称为离散平稳信源。

$$P(X_i)=P(X_j)$$

$$P(X_iX_{i+1})=P(X_jX_{j+1})$$

$$\cdots$$

$$P(X_iX_{i+1}X_{i+2}\cdots X_{i+N})=P(X_jX_{j+1}X_{j+2}\cdots X_{j+N})$$

离散平稳信源的熵

最简单的平稳信源——二维平稳信源，信源发出序列中只有前后两个符号间有依赖关系，我们可以对其二维扩展信源进行分析。

单符号信源的概率空间:

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} a_1, & a_2, & \cdots & ,a_n \\ p(a_1), & p(a_2), & \cdots, & p(a_n) \end{bmatrix}, \qquad \sum_{i=1}^{n} p(a_i) = 1$$

假定$\boldsymbol{X}=\boldsymbol{X}_1\boldsymbol{X}_2$，则可得到一个新的信源

$$\begin{bmatrix} \boldsymbol{X} \\ P(\boldsymbol{X}) \end{bmatrix} = \begin{Bmatrix} \alpha_1, & \alpha_2, & \cdots, & \alpha_i, & \cdots, & \alpha_{n^2} \\ p(\alpha_1), & p(\alpha_2), & \cdots, & p(\alpha_i), & \cdots, & p(\alpha_{n^2}) \end{Bmatrix},$$

式中, $p(\alpha_i) = p(a_{i_1} a_{i_2}) = p(a_{i_1}) p(a_{i_2} / a_{i_1}),\quad \sum_{i=1}^{n^2} p(\alpha_i) = 1$

- 二维离散平稳有记忆信源的熵小于等于二维平稳无记忆信源的熵。

根据信息熵的定义，可得：

(1) 联合熵 $$H(X_1X_2) = -\sum_{i=1}^{q}\sum_{j=1}^{q} P(a_ia_j)\log P(a_ia_j)$$

可以表征信源输出长度为2的符号的平均不确定性，或所含有的信息量。因此可以用$H_2(X) = \frac{1}{2}H(X_1X_2)$ 作为二维平稳信源的信息熵的近似值.

(2) $$H(X_1X_2) = H(X_1) + H(X_2 \mid X_1)$$

另外， $$H(X_2 \mid X_1) \le H(X_2)$$

$$H(X_1X_2) \le H(X_1) + H(X_2) = 2H(X)$$

只有信源统计独立时等号成立。

[例] 设某二维离散信源的原始信源的信源空间

X={x1, x2, x3}；　P(X)={1/4, 1/4, 1/2}，　一维条件概率为：

p(x1|x1)=1/2；　p(x2|x1)=1/2；　p(x3|x1)=0；

p(x1|x2)=1/8；　p(x2|x2)=3/4；　p(x3|x2)=1/8；

p(x1|x3)=0；　p(x2|x3)=1/4；　p(x3|x3)=3/4；

原始信源的熵为：　H(X)=1.5 bit/符号

条件熵：　H(X2|X1)=1.4 bit/符号

可见：　H(X2|X1)<H(X)

二维信源的熵：H(X1X2)=H(X1)+H(X2|X1)=2.9 bit/消息

每个信源符号提供的平均信息量为：

H_2(X1X2)=H(X1X2)/2=1.45 bit/符号 < H(X)。

相关性导致不确定性减少，编码变短

练习：概率分布换成P(X)={1/4, 4/9, 11/36}?

❍2.2.3 N维离散平稳有记忆信源

对于一般的离散平稳信源，符号的依赖关系往往不仅存在于相邻的两个符号之间，而是存在于更多符号之间，因此需要将二维离散平稳有记忆信源推广到N维情形。

计算很繁琐

❍联合熵

已知N维联合概率分布，离散平稳信源的联合熵表示发送一个消息（由N个符号组成）所提供的平均信息量。

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

注：计算熵的目的之一是估算消息编码时，平均情况下每个消息符号可能编出的最短码长。

$\vec{X}$消息符号长为 N，而单符号信源消息符号长度为 1，这难以比较哪个对编码更有利。

于是，考虑取平均长度，即，将$H(X^N)=H(\vec{X})$除以消息长度 N，

从而得到平均意义上的单符号携带信息量，从而具有可比性。

离散平稳有记忆信源的联合熵 $H(X_1X_2\cdots X_N)$ 表示平均发一个消息(由N个符号组成)所提供的信息量。从数学角度出发，信源平均发一个符号所提供的信息量应为

$$H_N(X)=\frac{1}{N}H(X_1X_2\cdots X_N)$$

这仍然难以计算以及用于编码

称$H_N(\vec{X})$为平均符号熵。

当对$\vec{X}=X_1\cdots X_N$进行编码时，不仅把$X_1X_2\cdots X_N$放在一起，而且还要考虑它们之间的相关性，这种相关性在计算联合概率和条件概率中体现。

直观地，符号间有相关性后，平均码长肯定会更短，如，若 e 与 n 总是一起出现，则在这种关联下对 en 编码与对 e 或 n 编码一样处理即可，从而缩短码长。

特别地，随着 N 的增大，编码符号数将更多，更多符号间有相关性，从而更接近实际，此时平均码长更短。

❍极限熵

我们称 $H_N(X)$ 为平均符号熵，当 $N \to \infty$ 时，平均符号熵取极限值，称为极限熵或极限信息量，即

$$H_\infty = \lim_{N\to\infty} \frac{1}{N} H(X_1X_2\cdots X_N) = \lim_{N\to\infty} H(X_N \mid X_1X_2\cdots X_{N-1})$$

极限是否存在？

对于离散平稳信源，当 $H_1(X)<\infty$时，具有以下性质：

如何证明？

✓ N给定时，平均符号熵≥条件熵；

$$H_N(X)\geq H(X_N/X_1X_2\cdots X_{N-1})$$

✓ 平均符号熵 $H_N(X)$随N增加是非递增的；

$$0\leq H_N(X)\leq H_{N-1}(X)\leq\cdots\leq H_1(X)\leq\infty$$

✓ 条件较多的熵必小于或等于条件较少的熵，而条件熵必小于等于无条件熵。

对于二维离散平稳信源，条件熵等于极限熵，因此条件熵就是二维离散平稳信源的真实熵。对于一般信源，求出极限熵是很困难的。

一般，取N不大时就可以得到与极限熵非常接近的条件熵和平均符号熵，因此可以用条件熵和平均符号熵来近似极限熵。

■ 马尔可夫信源

在很多信源的输出序列中，符号之间的依赖关系是有限的，任何时刻信源符号发生的概率只与前边已经发出的若干个符号有关，而与更前面的符号无关。

为了描述这类信源除了信源符号集外还要引入状态集。这时，信源输出消息符号还与信源所处的状态有关。

若一个信源满足下面两个条件，则称为马尔可夫信源：

（1）某一时刻信源输出的符号的概率只与当前所处的状态有关，而与以前的状态无关；

（2）信源的当前状态由当前输出符号和前一时刻信源状态唯一确定。

所谓“状态”，指与当前输出符号有关的前m个随机变量序列($X_1X_2\ldots X_m$)的某一具体消息，用s_i表示，把这个具体消息看作是某个状态。

m时刻的状态

$$s_i=\{a_{k_1},a_{k_2},\ldots,a_{k_m}\}$$

$$a_{k_1},a_{k_2},\ldots,a_{k_m}\in\{a_1,a_2,\ldots,a_n\}$$

$$k_1,k_1,\ldots,k_m=1,2,\ldots,n$$

$$i=1,2,\ldots,n^m$$

当信源在(m+1)时刻发出符号$a_{k_{m+1}}$时，我们可把s_j看成另一种状态：

m+1时刻的状态

$$s_j=s_{i+1}=\{a_{k_2},\ldots,a_{k_m},a_{k_{m+1}}\}$$

$$a_{k_2},\ldots,a_{k_m},a_{k_{m+1}}\in\{a_1,a_2,\ldots,a_n\}$$

$$k_2,\ldots,k_m,k_{m+1}=1,2,\ldots,n$$

$$i=1,2,\ldots,n^m$$

➢ 某一时刻信源输出的符号的概率只与当前所处的状态有关，而与以前的状态无关。或者说只与此前已输出的若干个符号有关。即 $p(a_{k_{m+1}} \mid s_i)$

$$\ldots, X_1, \overbrace{X_2, \ldots, X_{m-1}, X_m, X_{m+1}}^{s_j}, \ldots$$

$$\underbrace{X_1, X_2, \ldots, X_{m-1}, X_m}_{s_i}$$

把此前已输出的符号视为状态

➢ 信源的当前状态由当前输出符号和前一时刻信源状态唯一确定。

该条件表明，若信源处于某一状态 s_i ，当它发出一个符号后，所处的状态就变了，一定转移到另一状态。状态的转移依赖于发出的信源符号，因此任何时刻信源处在什么状态完全由前一时刻的状态和当前发出的符号决定。将信源输出符号的不确定性问题变换为讨论信源状态转换问题。状态之间的一步转移概率为：

$$p(s_j|s_i)$$

由上可知，m阶马尔可夫信源符号集共有n个符号，则信源共有 n^m 个不同状态。信源在某一时刻时，必然处于某一种状态，等到下一个符号输出时，转移到另外一个状态。

从而得到马尔可夫信源的状态空间。

$$\begin{bmatrix} s_1 & s_2 & \cdots & s_{n^m} \\ p(s_1) & p(s_2) & \cdots & p(s_{n^m}) \end{bmatrix}$$

其状态转移图如下页。在状态转换图中，把信源的每一种状态用圆圈表示，用有向箭头表示信源发出某一符号后由一种状态到另一状态的转移。

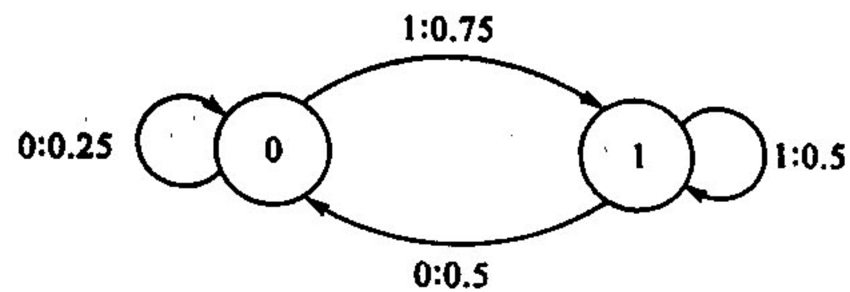
例：设有一二进制一阶马尔可夫信源，信源符号集为{1, 0}，

条件概率定为：$P(0|0)=0.25$

$P(0|1)=0.50$

$P(1|0)=0.75$

$P(1|1)=0.50$



由于信源符号数n=2，因此二进制一阶信源仅有两个状态：

$S_0=0, S_1=1.$

由条件概率求得信源的状态转移概率为：

$P(S_1|S_1)=0.25$；$P(S_1|S_2)=0.50$；

$P(S_2|S_1)=0.75$；$P(S_2|S_2)=0.50.$

［例］设信源符号 $X \in \{x1, x2, x3\}$，信源所处的状态 $S \in \{e1, e2, e3, e4, e5\}$。各状态之间的转移情况由下图给出。

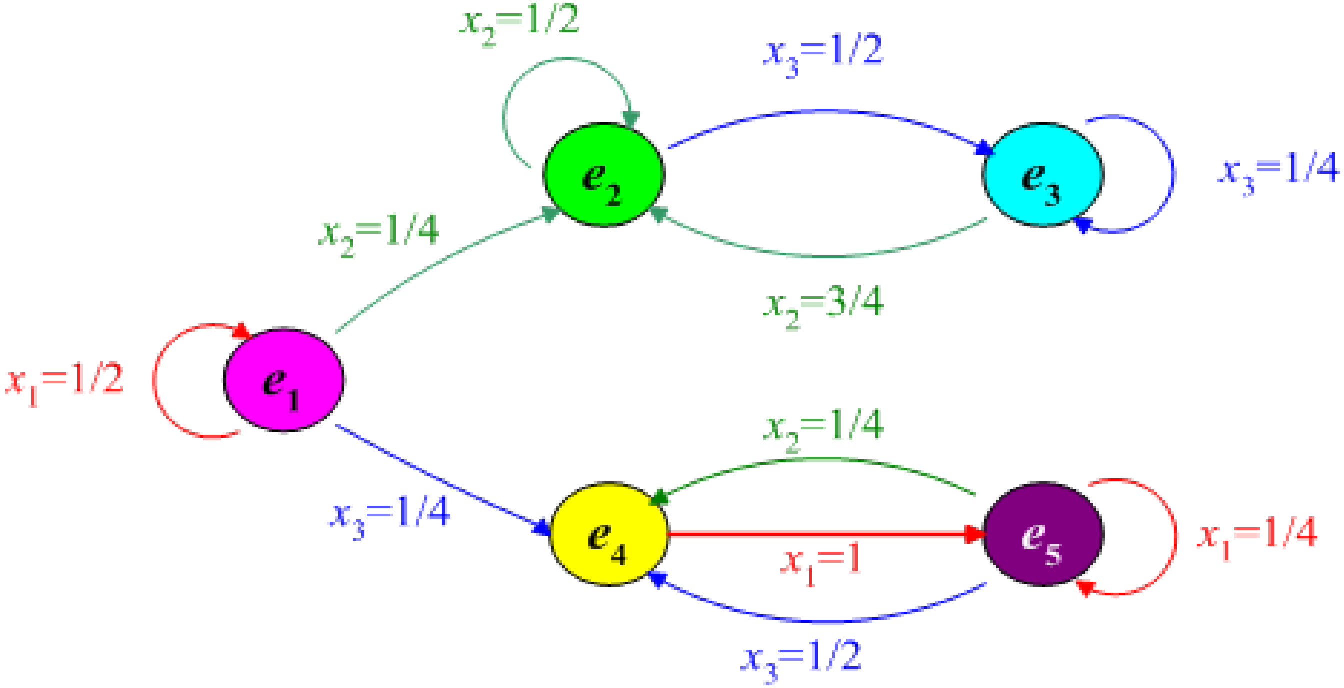


图2.2.1 状态转移图

将图中信源在e_i状态下发符号x_k 的条件概率$p(x_k/e_i)$用矩阵表示

可以看出 $\boxed{\sum_{k=1}^{3} p(x_k \mid e_i) = 1, \qquad i = 1,2,3,4,5}$

$$\begin{cases} P(S_l = e_2 \mid X_l = x_1, S_{l-1} = e_1) = 0 \\ P(S_l = e_1 \mid X_l = x_1, S_{l-1} = e_1) = 1 \\ P(S_l = e_2 \mid X_l = x_2, S_{l-1} = e_1) = 1 \\ P(S_l = e_3 \mid X_l = x_2, S_{l-1} = e_1) = 0 \\ \cdots \end{cases}$$

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{3}{4} & \frac{1}{4} \\ 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \end{array}$$

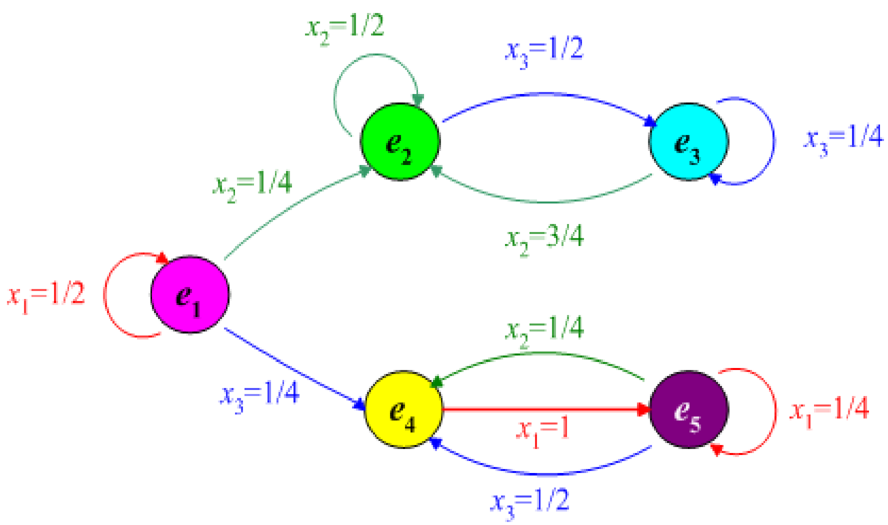


图2.2.1 状态转移图

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_5 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{4} & \frac{1}{4} \end{bmatrix} \end{array}$$

- M阶马尔可夫信源的平均信息量，即，信源的极限熵

$$H_\infty = \lim_{N\to\infty} H(X_N \mid X_1, X_2, ..., X_{N-1}) = H(X_{m+1} \mid X_1 X_2 \cdots X_m)$$

$$= H_{m+1} = \sum_{i=1}^{n^m} p(s_i) H(X \mid s_i) --- m\text{阶条件熵}$$

$$= \sum_{i=1}^{n^m} \sum_{j=1}^{n} p(s_i) p(a_j \mid s_i) \log_2 p(a_j \mid s_i)$$

$$= \sum_{i=1}^{n^m} \sum_{j=1}^{n^m} p(s_i) p(s_j \mid s_i) \log_2 p(s_j \mid s_i)$$

➢马尔可夫链的平稳分布

若齐次马尔可夫链对一切i, j存在不依赖于i的极限

$$\lim_{n\to\infty} p_n(s_j \mid s_i) = p(s_j)$$

且满足

$$p(s_j) > 0, \quad \sum_{j=1}^{n^m} p(s_j) = 1$$

则称其具有遍历性，$\boldsymbol{p}(\boldsymbol{s}_\mathrm{i})$称为平稳分布。

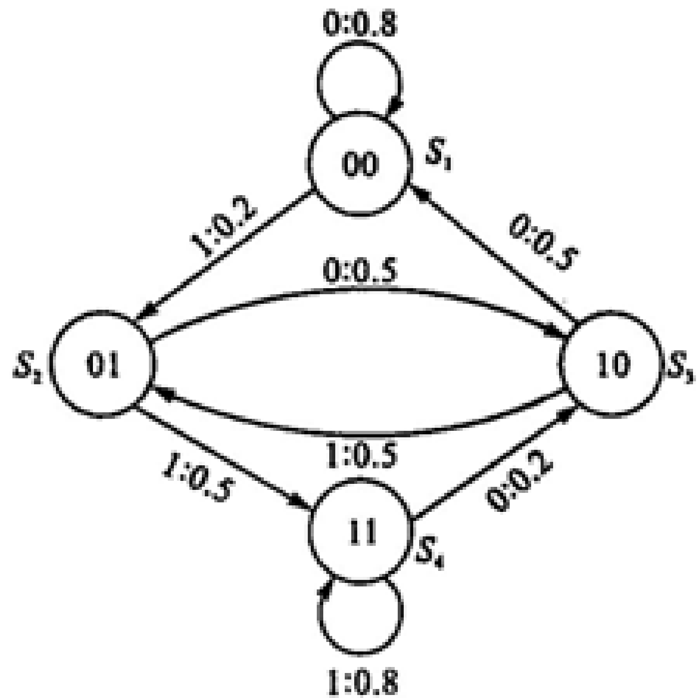
$$p(s_j) = \sum_{i=1}^{n^m} p(s_i) p(s_j / s_i), \quad j = 1, 2, \cdots, n^m$$

[例] 二元2阶马尔可夫信源，原始信号X的符号集为$\{X1=0, X2=1\}$，其状态空间共有$n^m=2^2=4$个不同的状态$s1,s2,s3,s4$，即

E: $\{s1=00,s2=01,s3=10,s4=11\}$, $p(0|00)=p(1|11)=0.8$;

$p(1|00)=p(0|11)=0.2$; $p(0|01)=p(0|10)=p(1|01)=p(1|10)=0.5$

状态转移图见右图所示。



解：$p(s1/s1)= p(x1/s1)=p(0/00)=0.8$

$p(s2/s1)= p(x2/s1)=p(1/00)=0.2$

$p(s3/s2)= p(x1/s2)=p(0/01)=0.5$

$p(s4/s2)= p(x2/s2)=p(1/01)=0.5$

$p(s1/s3)= p(x1/s3)=p(0/10)=0.5$

$p(s2/s3)= p(x2/s3)=p(1/10)=0.5$

$p(s3/s4)= p(x1/s4)=p(0/11)=0.2$

$p(s4/s4)= p(x2/s4)=p(1/11)=0.8$

对于由二元信源$X\in\{0,1\}$得到的状态空间(s_1,s_2,s_3,s_4), 容易验证

$$\sum_{j=1}^{4} p(s_j \mid s_i) = 1, \qquad p(s_i) > 0$$

由 $$p(s_j) = \sum_{i=1}^{4} p(s_i) p(s_j \mid s_i) \qquad j = 1,2,3,4$$

求出稳定状态下的$p(s_j)$，称为<u>状态极限概率</u>。

将一步转移概率代入上式得

$$p(s_1) = 0.8\, p(s_1) + 0.5\, p(s_3)$$

$$p(s_2) = 0.2\, p(s_1) + 0.5\, p(s_3)$$

$$p(s_3) = 0.5\, p(s_2) + 0.2\, p(s_4)$$

$$p(s_4) = 0.5\, p(s_2) + 0.8\, p(s_4)$$

解方程组得

$$p(s_1)=p(s_4)=5/14$$

$$p(s_2)=p(s_3)=1/7$$

计算极限熵

$$H_\infty = H_{2+1} = -\sum_{i=1}^{4}\sum_{j=1}^{4} p(s_i)\,p(s_j \mid s_i)\log_2 p(s_j \mid s_i)$$

$$= 0.801\ \text{(bit/sign)}$$

■ 信源的相关性和剩余度

平均符号熵随N增加是非递增的, 也就是说信源输出符号间的相关程度越长，信源的实际熵越小，趋近于极限熵；若相关程度减小，信源实际熵增大。

$$0 \le H_N(X) \le H_{N-1}(X) \le \cdots \le H_1(X) \le \infty$$

当信源输出符号间彼此不存在依存关系，且为等概率分布时，信源实际熵趋于最大熵$\boldsymbol{H_0}$

$$H_0(X) = \log_2 n$$

为了衡量信源的相关程度，引入信源剩余度(冗余度)的概念。

相对熵率 $\eta = H_\infty / H_0$　　　冗余度　$R = 1 - \eta$

冗余度（多余度、剩余度）表示给定信源在实际发出消息时所包含的多余信息。冗余度来自两个方面：

一是信源符号间的相关性，由于信源输出符号间的依赖关系使得信源熵减小，这就是信源的相关性。相关程度越大，信源的实际熵越小，趋于极限熵$H_\infty(X)$；反之相关程度减小，信源实际熵就增大。

另一个方面是信源符号分布的不均匀性，当等概率分布时信源熵最大。而实际应用中大多是不均匀分布，使得实际熵减小。当信源输出符号间彼此不存在依赖关系且为等概率分布时，信源实际熵趋于最大$H_0(X)$。

➢自然语言的熵

（1）对于英文字母

$$R = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{1.4}{\log 27} = 0.71$$

（2）对于中文

$$R = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{9.733}{\log 10^4} = 0.264$$

我们可以压缩剩余度来压缩信源，提高通信的可靠性。

- 正是因为原始的信息都有冗余，才有可能对信息进行压缩，以尽量减少冗余，提高每个符号携带的信息量；但另一方面，冗余信息可以提高信息的抗干扰能力，如果信息的某部分在传输中被损坏，则通过冗余有可能将其恢复。

  (冗余小, 有效)　　　　　　(冗余大, 可靠)

  中国　　　⇔　　　中华人民共和国

- 从提高信息传输效率的角度出发，总是希望减少剩余度（压缩），这是信源编码的作用；从提高信息抗干扰能力来看，总是希望增加或保留剩余度，这是信道编码的作用。