2022-2023学年秋季学期

信息论与编码

第二讲

2.1. 单符号离散信源

- 这是最简单最基本的信源：
  - 其每次只发出一个符号代表一个消息，
  - 输出的消息是有限或可数的。
  - 它是组成实际信源的基本单元，
  - 仅涉及一个随机事件。

信源每次输出其中一个消息；信源发出的消息是随机的。因此用随机变量X表示，从而概率统计性质确定。

- 用一维离散型随机变量表示：

信源$\mathbb{S}$的数学模型：用离散概率空间表示为

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

满足$0 \le p(x_i) \le 1$且$\sum_{i}^{n} p(x_i)=1$，

其中

- X 表示信源输出消息的整体，
- x_i表示某个消息，
- $p(x_i)$表示消息x_i出现的概率。
- n 为信源可能输出消息数，可以有限也可以是可数的。

在没有收到消息之前，信宿不能确定信源发出的是什么消息，即，有不确定性；

只有当收信者收到通过信道传输过来的消息后才能消除不确定性并获得消息。

说明：收到某消息m获得的信息量为

I(m) = 不确定性的减少量

= (收到 m 之前关于事件 m 发生的不确定性)-(收到 m 后关于事件 m 发生的不确定性（无噪声时为 0）).

显然，这与事件发生的可能性密切相关。

例1.2.

甲乙丙3个袋子，甲中有n个不同阻值的电阻，乙中有m个不同功率的电阻，丙中有n**种**不同的电阻，且每种电阻有m**种**不同的功率，即丙中有mn个电阻不同阻值不同功率的电阻.

说明：

由此例可见定义消息 x 的不确定性为$I(x) = -\log p(x)$是合理的。

对甲中随机选择一个电阻，并对其阻值进行猜测，a_i为电阻阻值，不妨设其概率为$p(a_i)=\frac{1}{n}$，则有离散概率空间：

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

其中猜测的困难度=其不确定性且信息量应为概率的函数，所以接收到“选取阻值为a_i”的电阻所获得的信息量为$I(a_i)=f(p(a_i))=f(\frac{1}{n})$.

对乙类似的有：

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} b_1 & b_2 & \cdots & b_n \\ \frac{1}{m} & \frac{1}{m} & \cdots & \frac{1}{m} \end{pmatrix}$$

接收到“选取功率为b_i”的电阻所获得的信息量为$I(b_i)=f(p(b_i))=f(\frac{1}{m})$.

对丙：

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} c_1 & c_2 & \cdots & c_n \\ \frac{1}{mn} & \frac{1}{mn} & \cdots & \frac{1}{mn} \end{pmatrix}$$

信息量为$I(c_i) = f(p(c_i)) = f(\frac{1}{mn})$.

由于事件丙=事件甲∩事件乙，所以$I(c_i) = I(a_i) + I(b_i) = f(\frac{1}{n}) + f(\frac{1}{m}) = f(\frac{1}{mn})$.满足这个等式的函数是对数函数，因此信息量可以定义为对数函数形式，即$I(m) = -\log p$.

自信息量，简称自信息

一个随机事件发生某一结果后所带来的信息量。

若随机事件x发生的概率为$p(x)$，则

定义其自信息量为$I(x) = -\log p(x)$。

- $I(x)$代表两种含义：

1) 在事件x发生前，表示事件x发生的不确定性大小；
2) 在事件x发生后，表示事件x所含有或所能提供的信息量（即，消除的不确定性的大小）。

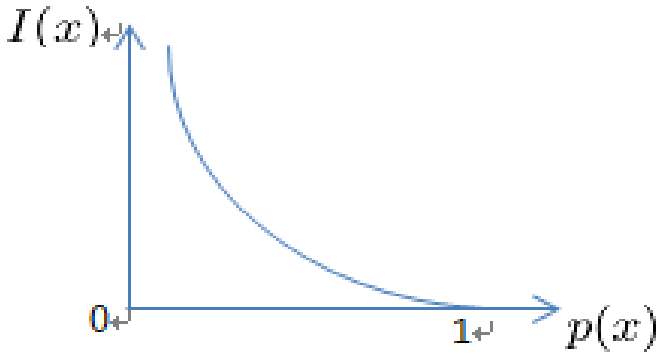
自信息量也是一个随机变量，即，事件所带来的信息量（不确定性）是随机的。

这里

- 以 2 为底的对数，单位称为比特（bit）；
- 以 e 为底的对数，单位称为奈特（nat）；
- 以 10 为底的对数，单位称为哈特（hat）；
- 以 r 为底的对数，为 r-进制。

性质：

1) 非负性：$I(x) \geq 0$。
2) 当$p(x) = 1$时，$I(x) = 0$，即，概率为 1 的确定性事件的自信息量为 0，即，不含任何信息量。
3) 当$p(x) = 0$时，$I(x) = \infty$，即，0 概率事件不可能发生，一旦发生，信息量巨大。
4) $I(x)$是关于$p(x)$单调递减函数。如图所示。



2.1.2 联合自信息

- 两个随机事件集

$X=\{x_1,\cdots,x_n\}$和$Y=\{y_1,\cdots,y_m\}$

- X 和 Y 的联合离散信源模型定义为

$$\begin{pmatrix} XY \\ p(XY) \end{pmatrix}=\begin{pmatrix} x_1y_1 & x_2y_2 & \cdots & x_ny_m \\ p(x_1y_1) & p(x_2y_2) & \cdots & p(x_ny_m) \end{pmatrix},$$

其中$0\le p(x_iy_j)\le 1,\ i=1,\cdots,n,\ j=1,\cdots,m,$

满足$\sum_i\sum_j p(x_iy_j)=1$。

联合自信息量

对于 2 维联合分布 XY 中，若事件x_i与y_j同时发生，其同时发生的可能性为$p(x_iy_j)$，则其联合自信息量定义为$I(x_iy_j) = -\log p(x_iy_j)$。

当x_i与y_j相互独立时，$p(x_iy_j) = p(x_i)p(y_j)$，则$I(x_iy_j) = I(x_i) + I(y_j)$。

两个随机事件相互独立时，“其同时发生所得自信息量”等于“这两个随机事件各自独立发生得到的自信息量之和”。

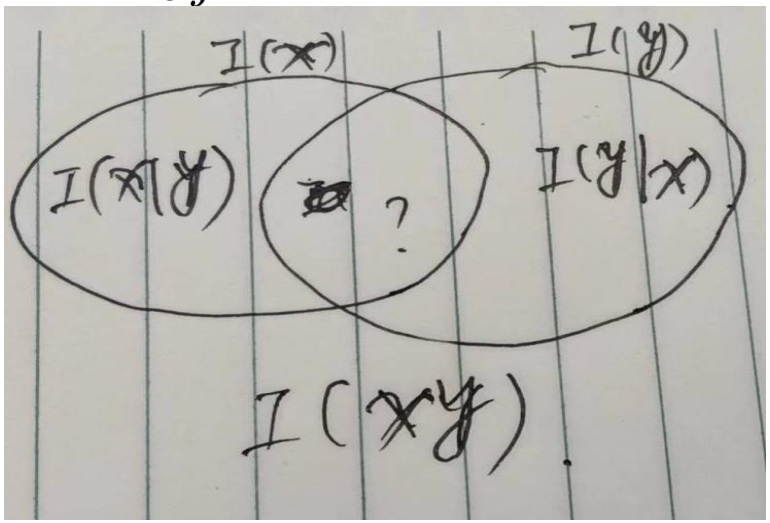
当x_i与y_j不相互独立时，$p(x_iy_j) = p(x_i)p(y_j|x_i)$这就需要条件自信息量的概念。

2.1.3 条件自信息量

对于 2 维联合分布 XY 中，对事件x_i与y_j，在给定事件y_j的条件下，事件x_i的条件自信息量定义为$I(x_i|y_j) = -\log p(x_i|y_j)$。

即，特定条件下（给定事件y_j），随机事件x_i发生所带来的信息量。

同样，事件x_i已知时，发生y_j的条件自信息量为$I(y_j|x_i) = -\log p(y_j|x_i)$。



- 自信息量、条件自信息量和联合自信息量之间有如下关系：

$$I(xy) = -\log p(xy) = -\log p(x)p(y|x)$$
$$= -\log p(x) - \log p(y|x) = I(x) + I(y|x)$$

$$= -\log p(y)P(x|y) = -\log p(y) - \log p(x|y)$$
$$= I(y) + I(x|y)$$

即，

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

例2.3 某地某月的天气概率分布统计如下：

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} x_1(\text{晴}) & x_2(\text{阴}) & x_3(\text{雨}) & x_4(\text{雪}) \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

则4种天气的自信息量：

$$I(x_1) = -\log \tfrac{1}{2} = 1, I(x_2) = 2,$$
$$I(x_3) = I(x_4) = 3$$

最多只需要3bit即可描述天气。

- 例2.4 英文字母中”a”出现的概率是0.064，”c”出现的概率是0.022。

求：a,c的自信息量，ac和ca(相互独立或不相互独立)的联合自信息量。

1) a,c的自信息量：

$$I(a) = -\log p(a) = -\log 0.064 \approx 3.966 bit/symbol$$

$$I(c) = -\log 0.022 \approx 5.506 bit/symbol$$

1) 若前后两个字母出现是相互独立的，则

$$I(ac) = I(a) + I(c) = I(ca) = 9.472$$

1) **若前后两个字母出现不是相互独立的，且a出现后c出现的概率为0.04，则**

$$I(ac) = I(a) + I(c|a) \approx 3.96 - \log 0.04 \approx 3.96 + 4.64 = 8.6bit/symbol$$

这里$I(c|a)$表示 a 出现后 c 出现的条件自信息量

a出现后，c出现带来的信息量。

例2.5 一个正方形棋盘有64个正方形格子，如果甲将一粒棋子随意放在棋盘中的某方格c上，令乙猜测棋子所在位置。

情形 1、将方格按顺序编号$\{1,2,\cdots,64\}$，则乙猜测棋子所在位置的难易程度即为棋子位置的信息量大小：

$p(c)=\frac{1}{64}$，$I(c)=-\log p(c)=6$比特。

情形 2、将方格按行和列分别编号，令$X=\{1,2,\cdots,8\}$为行，$Y=\{1,2,\cdots,8\}$为列，行与列相互独立，每个位置对应(x,y)，其中$(x,y)\in X\times Y$。

则乙猜测位置的信息量大小：

$$I(c)=I(xy)=-\log p(xy)=-\log p(x)-\log p(y)=6$$

若已知乙所在的列，即，y，则

$$I(c|y)=I(x|y)=-\log p(x|y)=-\log\frac{p(xy)}{p(y)}=3$$

不确定性减少3比特。

2.1.4 信息熵

- 自信息是信源发出的某一具体消息所含有的信息量，不同消息所含的自信息量不同，即，自信息量本身为随机变量，不能衡量信源的总体信息量。

需要了解信源的总体情况

- 在多数情况下，更关心离散信源符号集的平均信息量问题，即，信源输出中平均每个符号所能提供的信息量。

对于信源

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

信源平均信息量为信源中各个离散消息的自信息量的数学期望，称之为信息熵，简称为熵，记为$H(X)$。

即，

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

- **注意：**
  - **这里熵函数的自变量用大写X，表示信源整体。**
  - **信源熵的单位在以2为底时称为比特/符号(bit/symbol)。**

- **H(X)表示的意义**
  - 信源输出消息后每个离散消息所提供的平均信息量。
  - 信源输出消息前信源的平均不确定度。

- **例2.6 有一袋中装100个球，其中红色80个，白色20个。讨论“随意摸取一个球猜是什么颜色“的熵。**

$$\text{信源为}\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} R & W \\ 0.8 & 0.2 \end{pmatrix}$$

于是

- 摸取红球 R 所得信息量：$I(R) = -\log p(R) = -\log 0.8$,
- 摸取白球 W 所得信息量：$I(W) = -\log p(W) = -\log 0.4$。

可计算信息熵

$$H(X) = -p(R) \cdot \log p(R) - p(W) \cdot \log p(W) \text{bit/symbol.}$$

- 若每次摸出一个球后又放回去，再进行第二次摸取，则摸 n 次后红球出现的次数为$n \cdot p(R)$；白球出现的次数为$n \cdot p(W)$。

  每次摸取事件相互独立，n 次后总共所获得的总信息量为

  $$n \cdot p(R) \cdot I(R) + n \cdot p(W) \cdot I(W) = nH(X)。$$

- **如下例子说明信源与分布的关系**

例2.7 有三个信源X，Y和Z，概率空间分别如下：

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad \begin{pmatrix} Y \\ p(Y) \end{pmatrix} = \begin{pmatrix} y_1 & y_2 \\ 0.99 & 0.01 \end{pmatrix} \quad \begin{pmatrix} Z \\ p(Z) \end{pmatrix} = \begin{pmatrix} z_1 & z_2 \\ 0 & 1 \end{pmatrix}$$

X分布最均匀，随机性最强；Y不均匀，随机性弱；Z无随机性。

信息熵分别为：

$$H(X) = -p(x_1) \cdot \log p(x_1) - p(x_2) \cdot \log p(x_2)$$

- $= (-\frac{1}{2}) \times 2 = 1$ (bit/symbol)；

$$H(Y) = -p(y_1) \cdot \log p(y_1) - p(y_2) \cdot \log p(y_2)$$

- $= -0.99 \log 0.99 - 0.01 \log 0.01 \approx 0.08$ (bit/symbol)；
- $H(Z) = 0$(bit/symbol)。

显然，$H(X) > H(Y) > H(Z)$

信源符号的概率分布越均匀，则平均信息量越大；确定事件不含信息量。

- 总之，信息熵是信源的平均不确定性的描述，熵是分布均匀性的一种度量。
- 例2.8 信源有6种输出符号，概率空间为

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} A & B & C & D & E & F \\ 0.5 & 0.25 & 0.125 & 0.05 & 0.05 & 0.025 \end{pmatrix}$$

1) 计算$H(X)$：

$$H(X) = -\sum_{i=1}^{6} p(x_i) \cdot \log p(x_i) \approx 1.94\text{(bit/symbol)}。$$

2) 求符号序列**ABABBA**和**FDDFDF**的信息量，并将它们与6位序列符号的平均信息量比较。这里所有符号均相互独立。

计算所含的信息量：

$$I_1 = I(ABABBA) = 3I(A) + 3I(B)$$
$$= -3\log p(A) - 3\log p(B) = 9 \quad \text{(bit/symbol)。}$$

$I_2 = I(FDDFDF) = 3I(F) + 3I(D) = 28.932$(bit/symbol)。

6 位序列符号的平均信息量：$I_0 = 6H(X) = 11.64$(bit/symbol)。

故$I_1 < I_0 < I_2$。

说明：对于二元信源

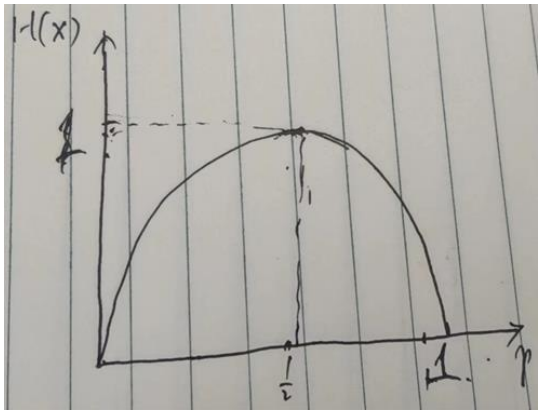
$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ p & 1-p \end{pmatrix}$$

有

$$H(X) = -p \cdot \log p - (1-p) \cdot \log(1-p) := H(p),$$

这里$p \in [0,1]$。

- 当$p = 1$或 0 时，信源输出是确定的，不提供任何信息；
- 当$p = \frac{1}{2}$时，$H(X) = 1$，此时信源熵最大。（如图）



- 均匀分布的二元信源输出的二元数字序列中，每个二元数字提供1bit平均信息量。
- 若非均匀分布，则每一个二元数字所提供的平均信息量总是小于1比特。这也说明二元数字与信息量单位bit/symbol的关系。

2.1.5 条件熵

考虑两个或两个以上的随机变量的相互关系。

对于联合符号集合，**条件熵是条件自信息量的数学期望**。

即，

在已知随机变量$Y=\{y_1,\cdots,y_m\}$的条件下，

随机变量$X=\{x_1,\cdots,x_n\}$的条件熵$H(X|Y)$定义为

$$H(X|Y)=E(I(X|Y)):=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)I(x_i|y_j)$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(x_i|y_j)$$

特别地，给定$Y=y_j$的条件下，X 的条件熵为

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

于是，有

$$
\begin{aligned}
H(X|Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(x_i|y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(y_j)p(x_i|y_j)\log p(x_i|y_j) \\
&= -\sum_{j=1}^{m} p(y_j)\sum_{i=1}^{n} p(x_i|y_j)\log p(x_i|y_j) \\
&= -\sum_{j=1}^{m} p(y_j)H(X|Y=y_j)
\end{aligned}
$$

故$H(X|Y)=\sum_{j=1}^{m} p(y_j)H(X|Y=y_j)$。

显然，当 X 与 Y 互相独立的时候，$H(X|Y)=H(X)$

相应地，在给定条件 X 下，Y 的条件熵$H(Y|X)$为

$$H(Y|X)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(y_j|x_i)=\sum_{i=1}^{n}p(x_i)H(Y|X=x_i)。$$

当 X 与 Y 互相独立的时候，$H(Y|X)=H(Y)$。

思考：如果$p(y|x)=p(x|y)$，会发生什么？

例2.9 已知XY构成的联合概率空间为

$$\begin{pmatrix} XY \\ p(XY) \end{pmatrix}=\begin{pmatrix} 00 & 01 & 10 & 11 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

其中$X,Y\in\{0,1\}$。计算条件熵$H(X|Y)$。

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(x_i|y_j)$$

解：由全概率公式，可得

$$p(y=0)=p(0)=p(y=0)\sum_x p(x|0)$$

$$=(p(0|0)+p(1|0))p(y=0)=p(00)+p(10)=\frac{1}{2},$$

同理可得$p(y=1)=p(1)=p(01)+p(11)=\frac{1}{2}$。

于是，由$p(x|y)=\frac{p(xy)}{p(y)}$，知

$$p(x=0|y=0)=\frac{p(00)}{p(0)}=\frac{1}{4}=p(1|1)$$

$$p(x=0|y=1)=\frac{p(01)}{p(1)}=\frac{3}{4}=\frac{p(10)}{p(0)}=p(x=1|y=0)$$

故

这样的模型我们称为对称模型，后面会重点介绍

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(x_i|y_j)$$

$$=-p(00)\log p(0|0)-p(01)\log p(0|1)-p(10)\log p(1|0)-p(11)\log p(1|1)$$

$$=0.812\text{(bit/symbol)}$$

可以计算$H(X|Y)=H(Y|X)$。（原因：对称模型）

2.1.6 联合熵

定义联合离散符号集合 XY 上的每个元素对$x_i y_j$的联合自信息量的数学期望为联合熵，记为$H(XY)$。

即，

$$H(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)I(x_iy_j)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(x_iy_j)$$

特别地，当 X 与 Y 相互独立时，$H(XY)=H(X)+H(Y)$。

进一步，

$$H(XY) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(x_i y_j)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(y_j|x_i)p(x_i)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(y_j|x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(x_i)$$

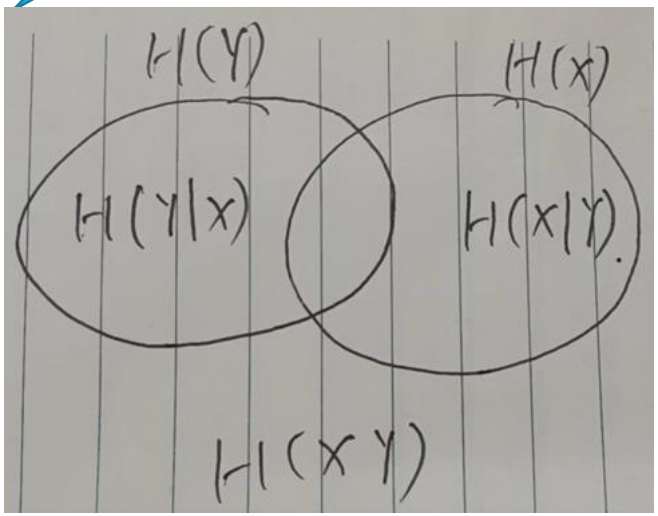
$$= H(Y|X) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)p(y_j|x_i)\log p(x_i)$$

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

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

同理，有$H(XY) = H(X|Y) + H(Y)$。

这两条性质称为可加性（对于n个分布情形如何？）



例2.10二元通信系统

设符号集为$\{0,1\}$。有如下事件发生：

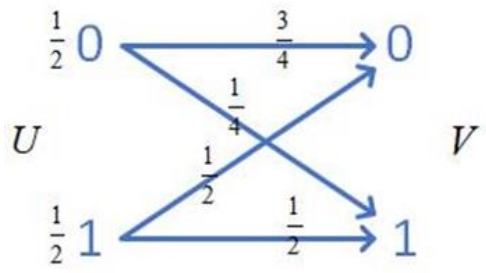
U 事件：u_0是”发出 0”。 u_1是 “发出 1”。

V 事件：v_0是“收到 0”。v_1是“收到 1”。

由于存在失真，因此传输时会产生误码。

给定概率$p(u_0)=\frac{1}{2}$，$p(v_0|u_0)=\frac{3}{4}$ ，$p(v_0|u_1)=\frac{1}{2}$。

相关的示意图和概率如下



计算:

1）已知发出 0，收到符号后得到的信息量。即，u_0发生时，v_0和v_1发生的可能性。

$$H(V|u_0) = -p(v_0|u_0)\log p(v_0|u_0) - p(v_1|u_0)\log p(v_1|u_0)$$

$$p(u_0) = \frac{1}{2},\ p(v_0|u_0) = \frac{3}{4}\ ,\ p(v_0|u_1) = \frac{1}{2},$$

因此

$$H(V|u_0) = -p(v_0|u_0)\log p(v_0|u_0) - p(v_1|u_0)\log p(v_1|u_0)$$
$$= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.82(\text{bits/symbol})$$

2）已知发出符号，收到符号后得到的信息量：即，U 发生后 V 发生。

$$H(V|U) = -\sum_{i=0}^{1}\sum_{j=0}^{1} p(u_i v_j)\log p(v_j|u_i)$$

$$p(u_0v_0) = p(u_0)p(v_0|u_0) = \frac{1}{2}\cdot\frac{3}{4} = \frac{3}{8}$$

$$p(u_0v_1) = p(u_0)p(v_1|u_0) = \frac{1}{2}\cdot(1-\frac{3}{4}) = \frac{1}{8},$$

$$p(u_1v_0) = p(u_1)p(v_0|u_1) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

$$p(u_1v_1) = p(u_1)p(v_1|u_1) = \frac{1}{2}\cdot(1-\frac{1}{2}) = \frac{1}{4}。$$

于是，

$$H(V|U) = -\sum_{i=0}^{1}\sum_{j=0}^{1} p(u_i v_j)\log p(v_j|u_i)$$

$$= -\frac{3}{8}\log\frac{3}{4} - \frac{1}{8}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{2} \approx 0.91 \text{ bit/symbol}$$

3）已知发出和收到符号，则得到的信息量

$$H(UV) = -\sum_{i=0}^{1}\sum_{j=0}^{1} p(u_i v_j)\log p(u_i v_j)$$

$$= -\frac{3}{8}\log\frac{3}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} \approx 1.91\text{bit/symbol}$$

4）已知收到了符号(V occurs)，被告知发出的符号，能到的信息量: $H(U|V) = -\sum_{i=0}^{1}\sum_{j=1}^{1} p(u_i v_j)\log p(u_i|v_j)$

由$p(v_0|u_0) = \frac{p(u_0)p(v_0|u_0)}{p(v_0)} = \frac{3}{5}$,

同理得$p(u_1|v_0) = \frac{2}{5}$，$p(u_0|v_1) = \frac{1}{3}$，$p(u_1|v_1) = \frac{2}{3}$。

于是

$H(U|V) = -\sum_{i=0}^{1}\sum_{j=1}^{1} p(u_i v_j)\log p(u_i|v_j) = \approx 0.951$ bit/symbol。

2.1.7 信息熵的性质

1、非负性：

只要有一个事件确定，其余事件不会发生

- $H(X) \geq 0$，且等号成立 iff 存在 i 使得x_i满足$p(x_i) = 1$而$p(x_j) = 0$，$j \neq i$。
- $H(XY) \geq 0$
- $H(X|Y) \geq 0$

思考其物理意义

2、对称性：

熵的对称性是指$H(X)$中的$p(x_1), \cdots, p(x_n)$的顺序任意互换时熵的值不变。

由熵的定义显然，因为这仅涉及到求和的顺序不同。

即，

$$H(X) = H(p(x_1), \cdots, p(x_n)) = H(p(x_{s_1}), p(x_{s_2}), \cdots, p(x_{s_n}))$$

其中$s_1, \cdots, s_n$为$1, 2, \cdots, n$的一个置换。

这说明，熵只与信源的总体结构有关，而与个别消息的概率无关。也就是说，事件的发生顺序对熵无影响。

例2.11 两个信源

$$\begin{pmatrix} Y \\ p(Y) \end{pmatrix} = \begin{pmatrix} y_1(\text{晴}) & y_2(\text{雾}) & y_3(\text{雨}) \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \end{pmatrix} \quad \begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} x_1(\text{红}) & x_2(\text{黄}) & x_3(\text{蓝}) \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}$$

直接计算，两个信源的熵相同，原因在于信源的总体统计特性相同。

注：

信息熵只抽取了信源输出的统计特征，而不考虑信息的具体含义和效用，即，不关心具体的信息。

这是香农(shannon)信息论的缺陷。

3、最大离散熵定理

定理2.1(最大离散熵定理) 若信源X中包含n个不同的离散消息，

则信源熵$H(X) \le \log n$，且当且仅当 X 中各消息出现的概率相等(即，均匀分布) 时等号成立。

即证明$H(X) - \log n \le 0$即可。

证明：注意，当 x>0 时，有$\ln x \le x - 1$，并且当且仅当 x=1 时等号成立。

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

$$= \sum_{i=1}^{n} p(x_i) \log \frac{1}{np(x_i)}$$

$$= \sum_{i=1}^{n} p(x_i) \ln \frac{1}{np(x_i)} \cdot \log_2 e$$

于是，由上述不等式得

$$H(x) - \log n \leq \sum_{i=1}^{n} p(x_i)\left(\frac{1}{np(x_i)} - 1\right) \cdot \log_2 e$$
$$= (1-1) \cdot \log_2 e$$
$$= 0$$

- **说明：**
  - **等概率分布信源的熵最大；只要信源中某一信源符号出现的概率大，就会引起整个信源熵下降。**
  - **信源符号数目越多，其熵值就越大。信源信息多，不确定性大。**

- **4、可加性：**

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

- **推广：**

设有 N 个概率空间$X_1, X_2, \cdots, X_N$，则

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

若这 N 个随机变量相互独立，则

$$H(X_1X_2\cdots X_N) = \sum_{i=1}^{N} H(X_i)。$$

若这 N 个随机变量相互独立且同分布，则

$$H(X_1X_2\cdots X_N) = NH(X_i)。$$

5、极值性：

$H(X|Y) \le H(X)$，$H(Y|X) \le H(Y)$；

等号成立当且仅当 X 与 Y 互相独立。

我们需要证明下述Shannon辅助定理

Shannon 辅助定理：

任意两消息数相同的信源 X,Y，不妨设$X = \{x_1, \cdots, x_n\}$，$Y = \{y_1, \cdots, y_n\}$，这里$\sum_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} p(y_i) = 1$，则

$$-\sum_{k=1}^{n} p(x_k)\log p(x_k) \le -\sum_{k=1}^{n} p(x_k)\log p(y_k)。$$

说明？？？

任意一概率分布对其他分布的自信息量取数学期望必大于其自身熵。

证明：利用$\log n$的凸性即可。

$$\sum_k p(x_k)\log\frac{p(y_k)}{p(x_k)} \overset{\text{凸性}}{\leq} \log\sum_k p(x_k)\frac{p(y_k)}{p(x_k)}$$
$$= \log\sum_k p(y_k) = \log 1 = 0 \qquad 。$$

- **由此可以证明：条件熵不超过无条件熵。**
  - $H(X|Y) \leq H(X)$。

事实上，

$$
\begin{aligned}
H(X|Y) &= -\sum_i \sum_j p(x_i y_j) \log p(x_i|y_j) \\
&= \sum_i \sum_j p(x_i y_j) \log \frac{p(y_j)}{p(x_i y_j)} \\
&= \sum_i \sum_j p(x_i y_j) \log \frac{p(y_j)p(x_i)}{p(x_i y_j)} - \sum_i \sum_j p(x_i y_j) \log p(x_i) \\
&\leq 0 - \sum_i \sum_j p(x_i y_j) \log p(x_i) \\
&= -\sum_i p(x_i) \log p(x_i) \\
&= H(X)
\end{aligned}
$$

- 类似地，$H(Y|X) \leq H(Y)$。

注：

在上式证明中，$\sum_i \sum_j p(x_i y_j) \log \frac{p(y_j)p(x_i)}{p(x_i y_j)}$即可导出可加性的证明。

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

于是，

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

- 注：当X和Y独立时，有
  - $H(XY) = H(X) + H(Y)$
  - $H(X|Y) = H(X)$
  - $H(Y|X) = H(Y)$

- **6、确定性：**

$$H(1,0)=H(1,0,0)=\cdots=H(1,0,\cdots,0)=0$$

只要信源符号中有一个符号出现概率为1，信源熵为0。

即，总体来看，信源虽然有不同的输出符号，但它只有一个符号是必然出现的，而其它符号则是不可能出现的，这个信源是确定信源。

7、扩展性：

1）$\lim_{\epsilon \to 0} H_{s+1}(p_1, p_2, \cdots, p_s - \epsilon, \epsilon) = H_s(p_1, p_2, \cdots, p_s)$。

即，当信源消息集中添加一些小概率消息时，信源熵不变。(有限个小概率事件不改变熵)

2）若信源 X 的 n 个符号的概率分别为$p_1, \cdots, p_n$，其中有一个元素被划分成 m 个元素（或符号），而这 m 个概率之和等于原来元素的概率，如，$x_i \to (y_{i1}, \cdots, y_{im})$，$p_i \to (q_{i1}, \cdots, q_{im})$，满足$\sum_j q_{ij} = p_i$，即，由原本发送 n 个符号变为发送 n+m-1 个符号，此时所得到一个新的信源$\tilde{X}$，$H(\tilde{X}) \geq H(X)$。

新信源的熵增大了

事实上，令

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

则不妨设

$$\begin{pmatrix} \tilde{X} \\ p(\tilde{X}) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_{n-1} & y_{n1} & \cdots & y_{nm} \\ p_1 & p_2 & \cdots & p_{n-1} & q_{n1} & \cdots & q_{nm} \end{pmatrix}$$

于是

$$\begin{aligned} H(\tilde{X}) &= H_{n+m-1}(p_1, \cdots, p_{n-1}, q_{n1}, \cdots, q_{nm}) \\ &= H_n(p_1, \cdots, p_n) + p_n H_m(\tfrac{q_{n1}}{p_n}, \cdots, \tfrac{q_{nm}}{p_n}) \\ &\geq H_n(p_1, \cdots, p_n) \\ &= H(X) \end{aligned}$$

其中，

第一个等式按定义展开后加一项减一项$-p_n \log p_n + p_n \log p_n$，整理得第二个等式。此时也就证明了(2).

由于(1)是(2)的特例，即，在 m=2 时，$q_{n1} = p_s - \epsilon$，$q_{n2} = \epsilon$。

当$\epsilon \to 0$时，后面的$H_2(\frac{q_{n1}}{p_n} \cdot \frac{q_{n2}}{p_n}) \to 0$，因此有

$$\lim_{\epsilon \to 0} H_{s+1}(p_1, p_2, \cdots, p_s - \epsilon, \epsilon) = H_s(p_1, p_2, \cdots, p_s)。$$

- **把计算n维熵转换为2维熵的和：**

  思路是反向利用(2)逐步把两个信源消息合并成一个，

  即，$(x_{n-1}, x_n) \to y_{n-1}$满足

  $(p_{n-1}, p_n) \to p_{n-1} + p_n = q_{n-1} = p(y_{n-1})$。

  令$q_i = \sum_{j=i}^{n} p_j$，$2 \le i \le n-1$，有

$$
\begin{aligned}
&H_n(p_1,\cdots,p_n)\\
&=H_{n-1}(p_1,\cdots,p_{n-2},p_{n-1}+p_n)+(p_{n-1}+p_n)H_2(\tfrac{p_{n-1}}{p_{n-1}+p_n},\tfrac{p_n}{p_{n-1}+p_n})\\
&=H_{n-1}(p_1,\cdots,p_{n-2},q_{n-1})+q_{n-1}H_2(\tfrac{p_{n-1}}{q_{n-1}},\tfrac{p_n}{q_{n-1}})
\end{aligned}
$$

$$
\begin{aligned}
&H_{n-1}(p_1,\cdots,p_{n-2},q_{n-1})\\
&=H_{n-2}(p_1,\cdots,p_{n-3},p_{n-2}+q_{n-1})+(p_{n-2}+q_{n-1})H_2(\tfrac{p_{n-2}}{p_{n-2}+q_{n-1}},\tfrac{q_{n-1}}{p_{n-2}+q_{n-1}})\\
&=H_{n-2}(p_1,\cdots,p_{n-3},q_{n-2})+q_{n-2}H_2(\tfrac{p_{n-2}}{q_{n-2}},\tfrac{q_{n-1}}{q_{n-2}})
\end{aligned}
$$

如此下去，

$$H_3(p_1,p_2,q_3)=H_2(p_1,q_2)+q_2H_2(\tfrac{p_2}{q_2},\tfrac{q_3}{q_2})$$

于是，

$$H_n(p_1,\cdots,p_n)=H_2(p_1,q_2)+q_2H_2(\tfrac{p_2}{q_2},\tfrac{q_3}{q_2})+\cdots+q_{n-1}H_2(\tfrac{p_{n-1}}{q_{n-1}},\tfrac{p_n}{q_{n-1}})$$

多元信息熵的计算可以转化为计算若干个二元信息熵。

如，

$$H(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{6},\tfrac{1}{6})$$
$$= H(\tfrac{1}{3},\tfrac{1}{3}+\tfrac{1}{6}+\tfrac{1}{6}) + (\tfrac{1}{3}+\tfrac{1}{6}+\tfrac{1}{6})H(\tfrac{\frac{1}{3}}{\frac{2}{3}},\tfrac{\frac{1}{3}}{\frac{2}{3}}) + (\tfrac{1}{6}+\tfrac{1}{6})H(\tfrac{\frac{1}{6}}{\frac{1}{6}+\frac{1}{6}},\tfrac{\frac{1}{6}}{\frac{1}{6}+\frac{1}{6}})$$
$$\approx 1.918(\text{bits/symbol})$$

8、上凸性：

$$\forall 0 \le \theta \le 1,\ H(\theta p_1 + (1-\theta)p_2) \ge \theta H(p_1) + (1-\theta)H(p_2)$$

做作业

小结：

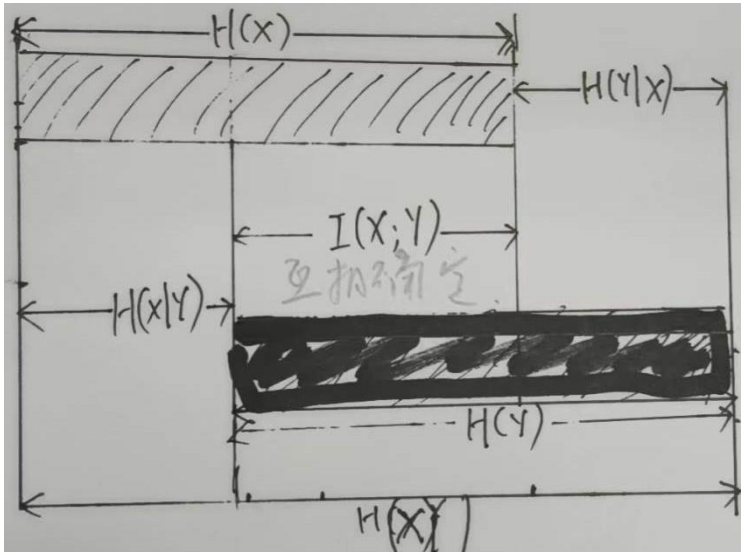
联合熵与条件熵之间的关系：

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

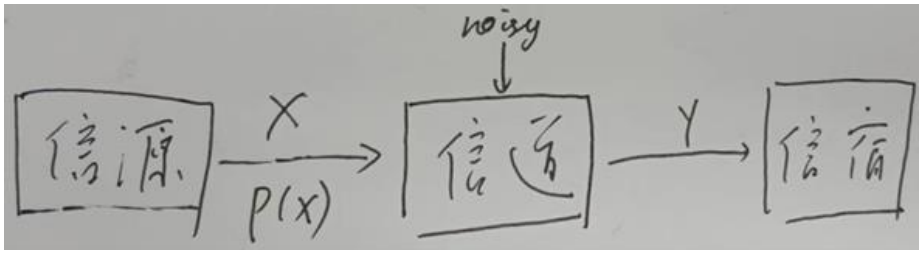
$$H(X|Y) \le H(X),\ \ H(Y|X) \le H(Y);$$

$$H(XY) \le H(X) + H(Y)。$$

如下图所示：



2.1.8 互信息



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

H(X)为输入变量X的先验不确定量。

- 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”？

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。

确定了多少，即，熵损失多少？

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

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

由于$p(x|y)=\frac{p(xy)}{p(y)}$，知

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

性质：

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

- 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。

- 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。

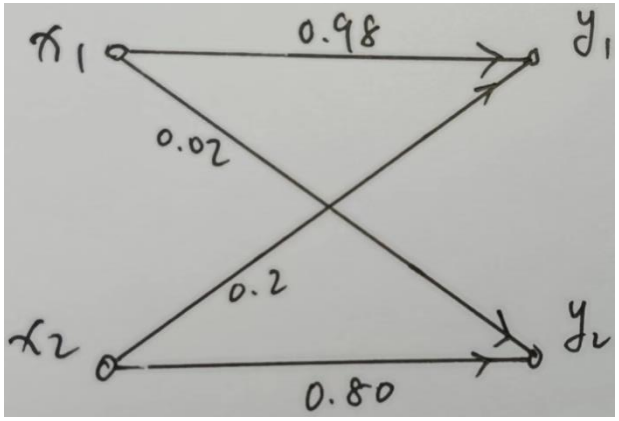
- 噪声熵：称$H(Y|X)=H(Y)-I(X;Y)$为噪声熵，表示为已知 X 的条件下 Y 尚存的不确定性，这由信道中噪声引起的。

思考题：找例子满足$I(X;Y)<0$。这说明，在未收到 y 之前对 x 是否出现的猜测疑义度较小，但是由于噪声存在，使得收到 y 之后反而使得对 x 是否出现的猜测疑义度增加了。

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

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

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$H(X|Y) = H(X) - I(X;Y) = 0.45$(bits/symbol)

噪声熵$H(Y|X) = H(Y) - I(X;Y) = 0.43$(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：

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

证明：由$H(X|Y)\leq H(X)$和$H(Y|X)\leq H(Y)$，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$I(X;Y)=f(p(y|x))$。

定理：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。

定理：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。