信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

一个引例

引例：设想你独立重复地掷一枚硬币，并且要把结果告诉远方的朋友，而你朋友只认识0,1。设定如下概率模型

1. 样本空间：{正，反}；
2. 事件集：{∅，{正}，{反}，{正，反}}；
3. 概率：$P(\{正\})=p, P(\{反\})=1-p$。

如果硬币是“理想均匀”的，即$p=1/2$，最优编码映射

$$X(正)=1, X(反)=0.$$

我们下面约定$p=1/4$，即$P(\{正\})=1/4, P(\{反\})=3/4$，考虑几种情形。

(1) 你的朋友不允许延迟:一次试验必须用一个二进制数位{0,1} 表示。

一个引例

引例：设想你独立重复地掷一枚硬币，并且要把结果告诉远方的朋友，而你朋友只认识0,1。设定如下概率模型

1. 样本空间：{正，反}；
2. 事件集：{∅，{正}，{反}，{正，反}}；
3. 概率：$P(\{正\})=p, P(\{反\})=1-p$。

如果硬币是“理想均匀”的，即$p=1/2$，最优编码映射

$$X(正)=1, X(反)=0.$$

我们下面约定$p=1/4$，即$P(\{正\})=1/4, P(\{反\})=3/4$，考虑几种情形。

(1) 你的朋友不允许延迟:一次试验必须用一个二进制数位{0,1} 表示。

(2) 如果允许延迟呢？

一个引例

引例：设想你独立重复地掷一枚硬币，并且要把结果告诉远方的朋友，而你朋友只认识0,1。设定如下概率模型

1. 样本空间：{正，反}；
2. 事件集：{∅，{正}，{反}，{正，反}}；
3. 概率：$P(\{正\})=p, P(\{反\})=1-p$。

如果硬币是“理想均匀”的，即$p=1/2$，最优编码映射

$$X(正)=1, X(反)=0.$$

我们下面约定$p=1/4$，即$P(\{正\})=1/4, P(\{反\})=3/4$，考虑几种情形。

(1) 你的朋友不允许延迟:一次试验必须用一个二进制数位{0,1} 表示。

(2) 如果允许延迟呢？考虑一个极端情形。若正面出现的概率为万分之一，则可以多次（比如一百万次）试验之后，只告诉你朋友正面出现的次数和位置（均用二进制表示）即可。

一个引例

每两次试验结束后再编码发送：

试验结果	概率	码字	码长
正 正	1/16	1 0 1	3
正 反	3/16	1 0 0	3
反 正	3/16	1 1	2
反 反	9/16	0	1

平均$(\frac{1}{16}\times 3+\frac{3}{16}\times 3+\frac{3}{16}\times 2+\frac{9}{16}\times 1)/2\approx 0.8438<1$ 位。
改进的原因在于概率大的所对应的码长短。（Morse 电码）

例子：

- 编码：正反（100）反反（0）正正（101）反正（11）正反（100）
- 译码：1 0 0 0 1 0 1 1 1 1 0 0 （prefix-free code）$\rightarrow$ 正反反反$\cdots$ $\cdots$

一个引例

每三次试验结束后再编码发送：

试验结果	概率	码字	码长
正 正 正	1/64	0 0 0 1 1	5
正 正 反	3/64	0 0 0 1 0	5
正 反 正	3/64	0 0 0 0 1	5
正 反 反	9/64	0 1 1	3
反 正 正	3/64	0 0 0 0 0	5
反 正 反	9/64	0 1 0	3
反 反 正	9/64	0 0 1	3
反 反 反	27/64	1	1

平均$(\frac{27}{64}\times 1+\frac{3\times 9}{64}\times 3+\frac{3\times 3}{64}\times 5+\frac{1}{64}\times 5)/3\approx 0.8229<0.8438$ 位。

一个引例

每三次试验结束后再编码发送：

试验结果	概率	码字	码长
正 正 正	1/64	0 0 0 1 1	5
正 正 反	3/64	0 0 0 1 0	5
正 反 正	3/64	0 0 0 0 1	5
正 反 反	9/64	0 1 1	3
反 正 正	3/64	0 0 0 0 0	5
反 正 反	9/64	0 1 0	3
反 反 正	9/64	0 0 1	3
反 反 反	27/64	1	1

平均$(\frac{27}{64}\times 1+\frac{3\times 9}{64}\times 3+\frac{3\times 3}{64}\times 5+\frac{1}{64}\times 5)/3\approx 0.8229<0.8438$ 位。

如果我们把更多的是按结果分组然后再编码传输，是不是可以更有效呢？能不能无限制地减少数位呢？

一个引例

(3) 二进制数位受限，但允许“误报”。

- 三次试验后必须立即告诉结果，但只允许用2 位二进制数位。

试验结果	概率	码字	码长
反 反 反	27/64	0 0	2
反 正 反	9/64	0 1	2
反 反 正	9/64	1 0	2
正 反 反	9/64	1 1	2
正 正 正	1/64	1 1	2
正 正 反	3/64	1 1	2
正 反 正	3/64	1 1	2
反 正 正	3/64	1 0	2

由于数位受限，为降低错误概率，我们可以把试验结果按发生的概率从大到小进行排序；然后只保证概率大的结果的正确性。

一个引例

例子：掷一不均匀硬币，结果为正面的概率是1/4，反面的概率是3/4。掷这枚硬币10次，为其中出现了大于等于7次反面的序列提供等长的不同码字。求：

(a) 最短码长；

(b) 错误概率。

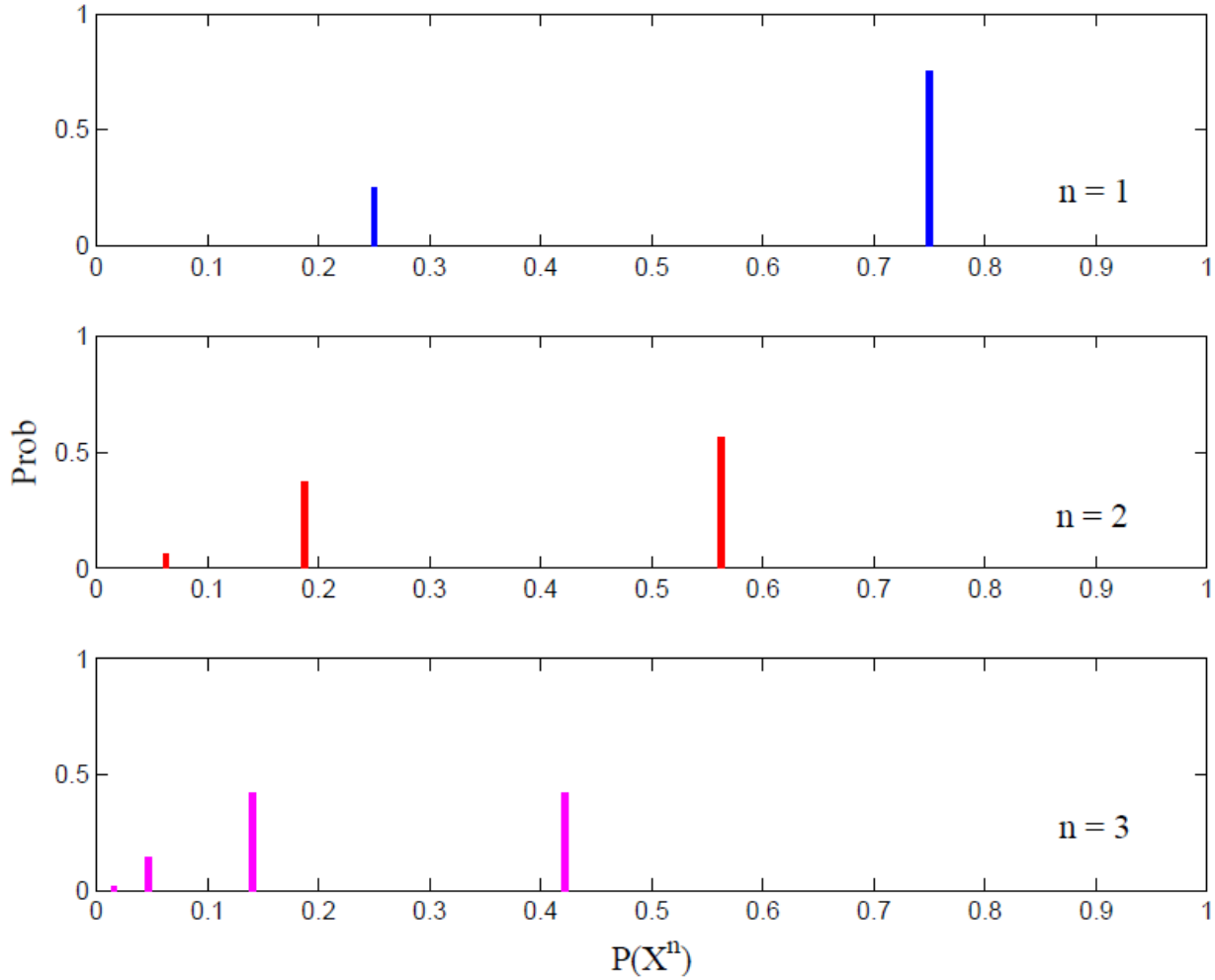
一个引例

直观：

1. 延迟有可能缩短平均码长；
2. 发生概率大的结果应分配较少的二进制数位；
3. 若二进制数位受限，则应先按照概率大的结果。

由此，自然想到把试验结果按照发生概率大小排序。下面是掷硬币试验中 $P_{X^n}(x^n)$ 的排序。

一个引例



一个引例

问题：随着n 的增大，

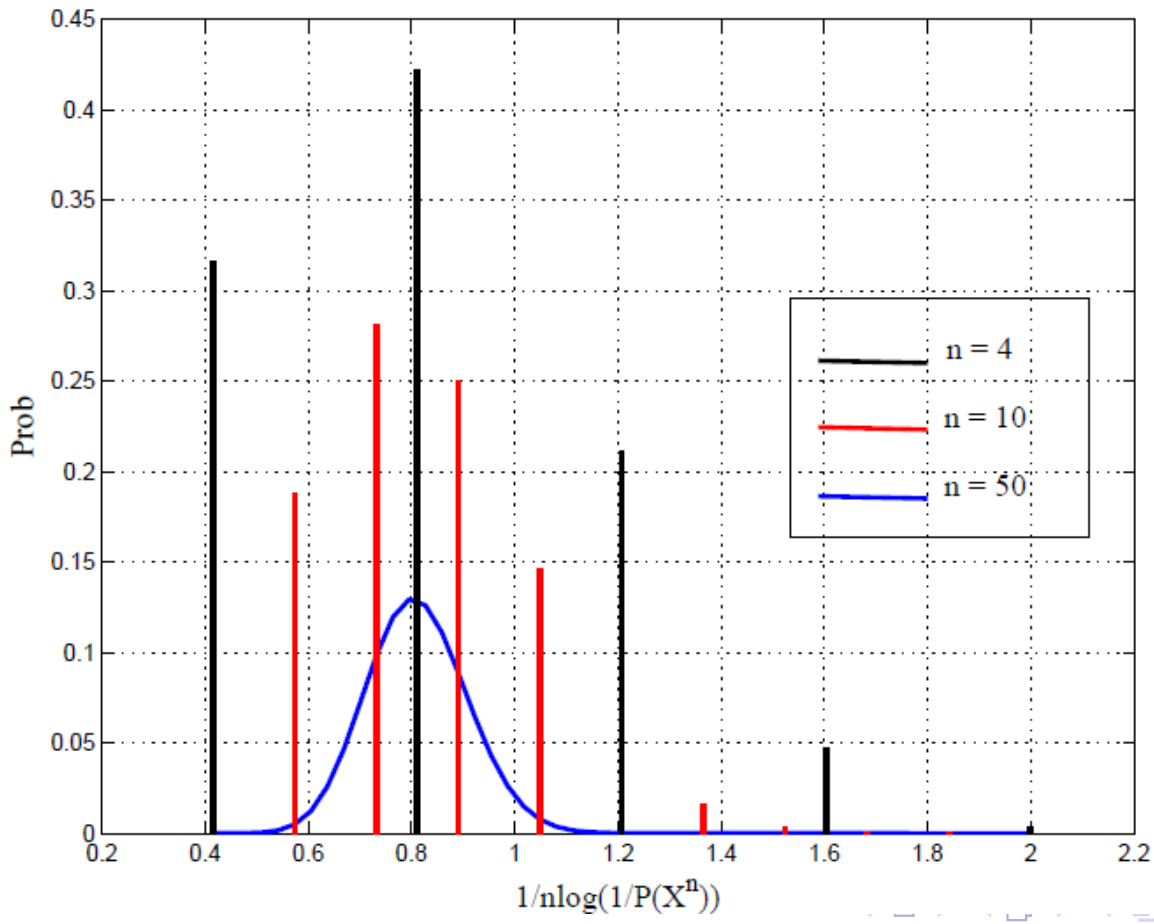
1. 序列的个数指数增长；
2. 单个序列$P_{X^n}(x^n)$ 的概率通常是指数衰减；
3. 若不归一化，没有规律可循。（回想大数定律的情形）

由于我们只关心大小顺序，因而可按照$-\frac{1}{n}\log P_{X^n}(x^n)$ 的排序。下面是掷硬币试验的排序结果。

一个引例

由大数定律，（引出熵的定义）

$$-\frac{1}{n}\log P_{X^n}(x^n) = \frac{1}{n}\sum_t \log\frac{1}{P_X(X_t)} \to \mathrm{E}(\log\frac{1}{P_X(X_t)}) \triangleq H(X) \approx 0.8113$$



离散无记忆信源

Definition 1

一个离散信源（source）可以用随机序列表示，即

$$\mathbf{X} = (X_1, X_2, \cdots, X_n, \cdots), \tag{1}$$

$X_t \in \mathcal{X}$。为方便起见，我们记$X^n \triangleq (X_1, X_2, \cdots, X_n)$。假定对于任意给定$n$，概率质量函数$P_{X^n}(x^n), x^n \in \mathcal{X}^n$ 是已知的。

信源编码的功能是用二进制序列表示信源产生的消息，目标是在允许的“错误”范围内，用尽可能少的二进制数位。

信源编码的压缩速率（平均每个信源符号所需要的二进制数位数）的极限完全由

$$\frac{1}{n} \log \frac{1}{P_{X^n}(X^n)}$$

的“谱线”（可以称为熵谱）的极限行为来决定。

离散无记忆信源

Definition 2 (离散无记忆信源)

设信源$\mathbf{X} = (X_1, X_2, ..., X_n, ...)$, 满足:

1. **无记忆的:** $P_{X^n}(x^n) = \prod_{1 \le t \le n} P_{X_t}(x_t)$ 对于任意 $n > 1$
2. **平稳的:** $P_{X_t}(x) \equiv P_{X_1}(x) \triangleq P_X(x)$ 对于任意 $t > 1$

则称该信源为离散平稳无记忆信源，也称作独立同分布信源。

熵

Definition 3

离散随机变量X，概率质量函数为 $P_X(x)$， $x \in \mathcal{X}$，则的X 熵，记作 $H(X)$，

$$H(X) = \mathrm{E}(\log \frac{1}{P_X(X)}) = -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x). \tag{2}$$

熵

说明：

(1) 根据熵的定义，若 log 以2 为底，则熵的单位是“比特/符号”；若 log 以 e 为底，则熵的单位是“奈特/符号”。如果我们已知信源每秒发出的符号数，则熵的单位可以是“比特/秒”或“奈特/秒”。

(2) 我们约定 $0\log 0 \triangleq 0$，这样约定是考虑到 $\lim\limits_{x\to 0^+} x\log x = 0$ 。

(3) $I(x) \triangleq \log(1/P_X(x))$ 也被称为 x 的**自信息量**。所以，我们也可以说熵是自信息量的数学期望。概率小的样本点具有大的自信息量，但是在熵中权重较小；而概率大的样本点的自信息量小，但在熵中权重较大。

熵

例子1. 抛一枚均匀硬币，观察到它的结果X 为正面或反面，获得$\log 1/P_X(x)$ 的信息量，它的熵即为：

$$\begin{aligned} H(X) = \mathrm{E}(\log \frac{1}{P_X(X)}) &= -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x). \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \\ &= 1 \text{ 比特/符号} \end{aligned}$$

再抛一枚不均匀硬币，观察到它的结果X' 为正面的概率是1/4，反面的概率是3/4，则抛一次硬币的熵为：

$$\begin{aligned} H(X') &= -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} \\ &\approx 0.8113 \text{ 比特/符号} \end{aligned}$$

熵

例子2. 掷一个骰子，观察到它的结果$X \in \{1,2,3,4,5,6\}$，则获得$\log 1/P_X(x)$ 的信息量，而它的熵可以理解为平均意义上获得的信息量，即为：

$$\begin{aligned} H(X) = \mathrm{E}(\log \frac{1}{P_X(X)}) &= -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x). \\ &= -\sum_{x\in\mathcal{X}} \frac{1}{6} \log \frac{1}{6} \\ &\approx 2.58 \text{ 比特/符号} \end{aligned}$$

熵

Definition 4

一对随机变量(X, Y) 服从联合分布$p(x, y)$，则它们的**联合熵**定义为：

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y, x). \tag{3}$$

一对随机变量(X, Y) 服从联合分布$p(x, y)$，则它们的**条件熵**定义为：

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X = x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x). \end{aligned} \tag{4}$$

熵

对于服从联合分布为$p(x,y,z)$ 的三个随机变量(X,Y,Z)，有**联合条件熵**

$$\begin{aligned} H(Y,Z|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y,Z|X=x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z)\log p(y,z|x). \end{aligned} \tag{5}$$

熵

例子 设$X \in \{\text{少云}, \text{多云}\}$，$Y \in \{\text{有雨}, \text{无雨}\}$。根据一天晴天或阴天的情况，有不同的下雨概率

X\Y	有雨	无雨
少云	1/8	3/8
多云	3/8	1/8

求在不知道某天阴晴的情况下，关于下雨情况的熵$H(Y)$。以及已知某天阴晴的情况下，下雨情况的条件熵$H(Y|X)$?

$$
\begin{aligned}
H(Y) = \mathsf{E}(\log \frac{1}{P_Y(Y)}) &= -\sum_{Y \in \mathcal{Y}} P_Y(y) \log P_Y(y). \\
&= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \\
&= 1 \text{ 比特/符号}
\end{aligned}
$$

熵

X\Y	有雨	无雨
少云	1/8	3/8
多云	3/8	1/8

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x). \\
&= -\frac{1}{8}\log\frac{1}{4} - \frac{3}{8}\log\frac{3}{4} - \frac{3}{8}\log\frac{3}{4} - \frac{1}{8}\log\frac{1}{4} \\
&\approx 0.8113 \text{ 比特/符号}
\end{aligned}
$$

熵的性质

1. 对称性

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

$$H(p_1,p_2,\cdots,p_n)=H(p_2,p_1,\cdots,p_n)$$

说明熵仅与信源的总体概率特性有关，而与随机变量的取值无关，例如下列信源的熵都是相等的。

$$\begin{bmatrix} X \\ P \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & x_3 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

$$\begin{bmatrix} Y \\ P \end{bmatrix}=\begin{bmatrix} y_1 & y_2 & y_3 \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

$$\begin{bmatrix} Z \\ P \end{bmatrix}=\begin{bmatrix} z_1 & z_2 & z_3 \\ 1/2 & 1/3 & 1/6 \end{bmatrix}$$

熵的性质

2. 非负性

$$H(X) = H(p_1, p_2, \cdots, p_n) \geq 0$$

3. 确定性

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

4. 拓展性

$$\lim_{\varepsilon \to 0} H_{K+1}(P_1, P_2, \cdots, P_K - \varepsilon, \varepsilon) = H_K(P_1, P_2, \cdots, P_K)$$

熵的性质

5. 可加性

$$
\begin{aligned}
&H_M\left(P_1Q_{11}, P_1Q_{21}, \cdots, P_1Q_{m_1 1}, P_2Q_{12}, P_2Q_{22}, \cdots,\right.\\
&\left.P_2Q_{m_2 2}, \cdots, P_KQ_{1K}, P_KQ_{2K}, \cdots, P_KQ_{m_K K}\right)\\
&= H_K\left(P_1, P_2, \cdots, P_K\right) + \sum_{k=1}^{K} P_k H_{m_k}\left(Q_{1k}, Q_{2k}, \cdots, Q_{m_k k}\right)
\end{aligned}
$$

其中

$$
\begin{aligned}
&\sum_{k=1}^{K} P_k = 1, P_k \geqslant 0\\
&\sum_{j=1}^{m_k} Q_{jk} = 1, Q_{jk} \geqslant 0\\
&M = \sum_{k=1}^{K} m_k
\end{aligned}
$$

可加性可以从概率树的角度描述。

链式法则： 设N 维随机变量集$(X_1X_2\cdots X_n)$，则有

$$
H\left(X_1X_2\cdots X_n\right) = H\left(X_1\right) + H\left(X_2 \mid X_1\right) + \cdots + H\left(X_n \mid X_1\cdots X_{n-1}\right)
$$

熵的性质

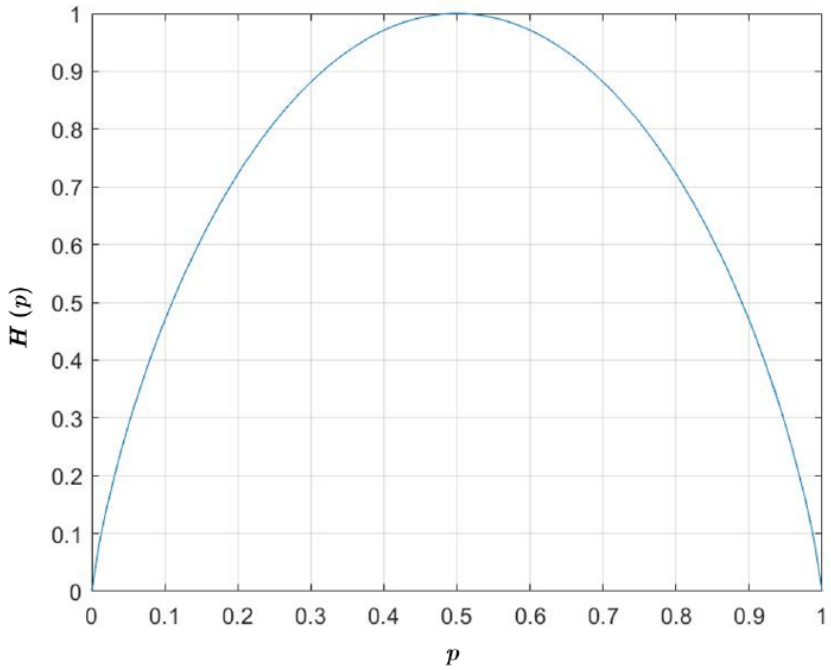
6. 极值性

当随机变量X 的各个取值概率相等时，熵最大。因为出现任何取值的可能性相等，不确定性最大，即

$$H(X) \leq H(\frac{1}{M}, \frac{1}{M}, \cdots, \frac{1}{M}) = \log M$$

熵的性质

例子1： 对于二元随机变量，$P_X(1) = p$，$P_X(0) = 1 - p$，我们定义熵函数 $H(p) = -p\log p - (1-p)\log(1-p)$。该函数如下图所示。可以看出，$H(0) = H(1) = 0$，而对于 $p > 0$，$H(p) > 0$，且在 $p = \frac{1}{2}$ 时达到最大熵 $H(\frac{1}{2}) = 1$ 比特/符号。



熵的性质

例子2： 计算$H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right)$：

$$\begin{aligned} H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)+\frac{1}{4}H\left(\frac{1}{2},\frac{1}{2}\right) \\ &= H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{4}H\left(\frac{1}{2},\frac{1}{2}\right) \\ &= \frac{7}{4}\ \text{bits/ 符号} \end{aligned}$$

信源编码基本框架

一般地，一个时间离散的信源可以表示为一个随机变量序列：$X_1, X_2, ..., X_n, ...$，其中X_t 取值在$\mathcal{X}$ 上，其统计规律可以用一族联合分布律$\{P_{\mathbf{X}}(\mathbf{x})\}, n = 1, 2, ...$ 来表征。设$\mathcal{D} = \{0, 1, ..., D-1\}$ 是字符集，我们用$\mathcal{D}^*$ 表示由$\mathcal{D}$ 构成的字符串的全体，包括空字符串，即$\mathcal{D}^* = \bigcup_{\ell \geq 0} \mathcal{D}^\ell$。信源编码的一般框架可以描述为：

$$\text{编码 } \phi_n : \mathcal{X}^n \longmapsto \mathcal{D}^*$$

$$\text{译码 } \psi_n : \mathcal{D}^* \longmapsto \mathcal{X}^n$$

$$\text{码率 } R_n = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} P_{\mathbf{X}}(\mathbf{x}) \ell\left(\phi_n(\mathbf{x})\right)$$

$$\text{译码错误 } \epsilon_n = \Pr\left\{\psi_n\left(\phi_n(\mathbf{X})\right) \neq \mathbf{X}\right\}$$

信源编码基本框架

R_n 中的$\ell\left(\phi_n(\mathbf{x})\right)$ 表示码字$\phi_n(\mathbf{x})$ 的长度。由此，我们知道码率表示在统计意义下每个信源符号所用的码字的平均长度。离散信源编码的问题就是通过证明ϕ_n 与ψ_n 的存在性，寻找满足$\lim_{n\to\infty}\epsilon_n = 0$ 的码率R_n 的下极限。

此外，根据序列或码字的长度是否固定，编译码大致可以分为四种类型：

1. 定长 $\longmapsto$ 定长
2. 定长 $\longmapsto$ 变长
3. 变长 $\longmapsto$ 定长
4. 变长 $\longmapsto$ 变长

信源编码定理

Theorem 5 (信源编码定理)

给定一个离散无记忆信源$X_1, X_2, \cdots, X_i, \cdots$，其中各$X_i$ 独立同分布，熵为$H(X)$。设$R > H(X)$，则一定存在一个编译码方案ϕ_n 和ψ_n，使得其码率R_n 满足$R_n \leq R$，并且$\lim_{n\to\infty} \epsilon_n = 0$。

作业：

Exercise 1. [Cover(2006)]

Minimum entropy. What is the minimum value of $H(p_1, \ldots, p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s that achieve this minimum.

Exercise 2. [Cover(2006)]

World Series. The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

作业：

Exercise 3. [王育民(2013)]

3.1 试证明长为 N 的 D 元不等长码至多有 $D(D^N-1)/(D-1)$ 个码字。

Exercise 4. [王育民(2013)]

3.2 设有一离散无记忆信源 $U=\begin{Bmatrix} a_1 & a_2 \\ 0.004 & 0.996 \end{Bmatrix}$。若对其输出的长为 100 的事件序列中含有两个或少于两个 a_1 的序列提供不同的码字。

(a) 在等长编码下，求二元码的最短码长。

(b) 求错误概率(误组率)。

作业：

Exercise 5.

有一离散无记忆信源X，对于该信源中的序列$(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$，如果它满足

$$2^{-n(H(X)+\epsilon)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\epsilon)},$$

则称它是$\epsilon-$典型 的。投掷一枚不均匀硬币，其正面朝上的概率为$1/4$，反面朝上的概率为$3/4$。对应有随机变量$X \in \{0, 1\}$，其概率质量函数为$P(1) = 1/4$ 和$P(0) = 3/4$。现在独立投掷该硬币n 次，求：

(a) 熵$H(X)$。

(b) 设$n = 5$，则当ϵ 等于0.1 时，哪些序列是$\epsilon-$典型 的？

(c)$\epsilon = 0.01$ 呢？

谢谢！