第 2 章

离散信源和离散熵

2.1 信源的数学模型及分类

信息是对事物运动状态或存在方式的不确定性的描述，信息的外在表现形式是消息，由消息承载并以消息的形式传输。那么，作为消息的产生者和发送者，信源的最本质特征是具有不确定性，所以可以用随机变量（Random Variable）来描述信源。在概率论中，随机变量被定义成在基本事件空间上的单值实函数[1]。但是该定义有些晦涩难懂。透彻地理解该定义需要有测度论基础，再加上这个定义对于理解信源和信息并没有太多帮助，因此对于初学者来说，只需要把随机变量理解成按照某种概率分布取某些值的变量即可。基于此，本书将不加区分地交替使用随机变量和信源这两个术语。

随机变量可以划分为离散型随机变量和连续型随机变量两种类型，其概率分布如下。

1．离散型随机变量

有限取值

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_q \\ P(a_1) & P(a_2) & P(a_3) & \cdots & P(a_q) \end{bmatrix} \tag{2-1}$$

无限取值

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots \\ P(a_1) & P(a_2) & P(a_3) & \cdots \end{bmatrix} \tag{2-2}$$

2．连续型随机变量

$$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} (a,b) \\ p_X(x) \end{bmatrix} \tag{2-3}$$

可见，离散型随机变量可以取有限个或无限可列个值，或者说其值域是有限集或无限可列集；连续型随机变量取无限不可列个值，其值域是不可列集。常见的可列集包括自然数、整数和有理数，不可列集包括无理数和实数。

对信源可以从不同的角度加以分类。

（1）根据随机变量的类型可以把信源划分为离散信源和连续信源，分别对应于离散型随机变量和连续型随机变量。

（2）从产生消息的维数可以把信源划分为一维信源和多维信源。一维信源又称单符号信源，N 维信源又称 N 维符号序列信源，其输出的消息要用 N 维随机矢量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_N)$ 来描述。

（3）对于 N 维信源而言，如果输出符号 X_1，X_2，X_3，…，X_N 之间不相互依赖，是统

计独立的，即 $P(X_1, X_2, \cdots, X_N)= P(X_1)P(X_2)\cdots P(X_N)$成立，则称该信源为无记忆信源；反之，则称为有记忆信源。

例 2-1：一维离散信源 X 的 N 次无记忆扩展信源。

设在 N 维随机矢量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_N)$中，每个变量 X_i 都独立同分布（Independent Identical Distribution，IID）于一维离散信源 X，假设 X 的概率分布如式（2-1）所示，另记 $\boldsymbol{X}$ 为 X^N，则 X^N 的概率分布为

$$\begin{bmatrix} X^N \\ P(\alpha_i) \end{bmatrix}=\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{q^N} \\ P(\alpha_1) & P(\alpha_2) & \cdots & P(\alpha_{q^N}) \end{bmatrix}$$

其中，$\begin{cases} \alpha_1 = \overbrace{a_1\ a_1\ \cdots a_1}^{N} \\ \alpha_2 = a_1\ a_1 \cdots a_2 \\ \quad\vdots \\ \alpha_{q^N} = a_q\ a_q \cdots a_q \end{cases}$，则称 X^N 为一维离散信源 X 的 N 次无记忆扩展信源。 ■

（4）如果 N 维信源输出的随机矢量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_N)$的各维概率分布不随时间的推移而改变，则称该信源为平稳信源；反之，则称为非平稳信源。

（5）更为一般地，在工程中常见的时间和取值都连续的信号，如语音、热噪声等，称这类信源为随机波形信源，简称波形信源。波形信源要由随机过程加以描述，在数字通信中，要通过采样（时间离散化）和量化（取值离散化）把波形信源转化成离散信源进行处理。

本书第 2～5 章讨论离散信源，第 6 章讨论连续信源和波形信源。

2.2　离散熵

按照由简单到复杂的顺序，我们首先讨论一维离散信源的信息度量问题。为简洁起见，以下假设一维离散信源 X 取有限值，其概率分布为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_n \\ P(a_1) & P(a_2) & P(a_3) & \cdots & P(a_n) \end{bmatrix}$$

即对于任意一个基本事件 $X=a_i$，其概率 $P(X=a_i)=P(a_i)$。下面来考察事件 $X=a_i$ 包含的信息量，我们首先要思考一个问题：事件 $X=a_i$ 的信息量和其概率 $P(a_i)$之间存在什么关系？或者更简明地说：某个事件发生的概率越大，其所包含的信息量越大；还是某个事件发生的概率越小，其所蕴含的信息量越大？搞明白这个问题才能理解如何度量信息。

来看一个具体例子：设袋子里有 100 个大小相同的球，包括 1 个红色球和 99 个黑色球，随机地抓取 1 个球，以 $X=a_1$ 和 $X=a_2$ 分别表示抓到黑色球和红色球，则 X 的概率分布为

X	a_1	a_2
$P(X)$	0.99	0.01

对于 $X=a_1$ 和 $X=a_2$ 这两个事件，哪个事件包含的信息量更大呢？相信绝大多数人都会选择 $X=a_2$，也就是说小概率事件包含着更丰富的信息。这是符合直观感受的，在一次随机实验中，某个事件发生的概率越小，那么该事件一旦发生，就越让人感觉惊讶和不可思议；反之，对于发生概率很大的事件，比如上面实验中抽到黑色球，该事件的发生并不会让人感觉意外，只会感觉这是很平常的事情，这种事情的发生不会让人诧异。还有一个非常能够说明问题的例子就是，新闻界的一句名言“狗咬人不是新闻，人咬狗才是新闻”。因此，粗略地说，一个事件包含的信息量应该与该事件发生的概率成反比，或者说信息对概率具有递减性。

除递减性之外，信息对概率还应该满足可加性，即两个独立事件 E_1 和 E_2 同时发生所包含的信息量 $I(E_1\cap E_2)$应该等于这两个事件各自信息量的和 $I(E_1)+I(E_2)$。此外，在学习概率时我们知道，在所有事件中，有两个特殊事件，一个是对应着空集∅的不可能事件，其概率为 0；另一个是对应着全集的必然事件，其概率为 1。基于人们的习惯，我们认为不可能事件的信息量为无穷大，必然事件的信息量为 0。综上所述，基于信息对概率的递减性和可加性的要求，以及不可能事件信息量为无穷大，必然事件信息量为 0 等特征，人们提出信息量的对数形式的定义如下。

定义 2-1：称离散信源 X 的某个事件 $X=a_i$ 包含的信息量为该事件的自信息量，记为 $I(X=a_i)$，其计算公式为

$$I(X=a_i)=\log_b\left(\frac{1}{P(a_i)}\right) \tag{2-4}$$

在式（2-4）中，对数可以取不同的底，如果 b=2，则信息的单位为 bit（比特）；如果 b=e，则信息的单位为 nat（奈特）；如果 b=10，则信息的单位为 Hatley（哈特莱）。本书中绝大多数时候都以 2 为底，因此为了行文简洁，在对数符号中略去底 b，默认底为 2。

$I(X=a_i)$度量了事件 $X=a_i$ 包含的信息量，那么 X 的信息量又该如何度量呢？一个很自然的想法就是采用概率中经常使用的统计平均的处理方法，即把所有基本事件 $X=a_i$ 的自信息量的期望作为 X 的信息量的度量。

定义 2-2：称离散信源 X 的基本事件 $X=a_i$ 包含的自信息量的期望为 X 的**信息熵或离散熵**，简称**熵**，记为 $H(X)$，其计算公式为

$$H(X)=E[I(X=a_i)]=E\left[\log\frac{1}{P(a_i)}\right]=-\sum_{i=1}^{n}P(a_i)\log P(a_i)\ \text{（bit / 信源符号）} \tag{2-5}$$

熵 $H(X)$的单位是 bit/信源符号，含义是信源 X 所有符号的平均信息量。式（2-5）是香农 1948 年发表的论文中给出的，后来 Feinstein 等人证明，在要求信息满足对概率的递减性和可加性的条件下，信息熵的表达式是唯一的，只能如式（2-5）所示，这个结论被称为**熵函数形式的唯一性**，也被称为**熵的公理化结构**。本书略去对熵函数唯一性的证明，感兴趣的读者请参阅参考文献［2，4］。熵本身是统计热力学中的一个概念，用来描述分子热运动的混乱程度，因此被称为热熵。香农把这种混乱性的含义借用过来描述信源的平均不确定性。

式（2-5）隐含着熵的一个性质，即熵只与信源的概率分布 $P(a_i)$有关，与信源发出的符号 a_i 无关，也就是说，无论信源发出的是 $a_1, a_2, \ldots$，还是 $b_1, b_2, \ldots$，只要二者具有相同

的概率分布，这两个信源的熵就是相同的。

可以从以下几个方面理解熵的物理意义。

（1）熵 $H(X)$是单个随机变量 X 的信息测度。

$H(X)$度量了单个随机变量 X 的信息量，后面还会讨论联合熵、条件熵、平均互信息，这些信息量分别描述了不同情况的信息测度。

（2）熵描述了信源的先验不确定性。

就通信而言，先验和后验分别表示通信发生前和发生后。熵是信源本身的性质，只与信源本身的概率分布有关，与通信与否没有关系，因此熵描述信源的先验不确定性。如果一个信源向信宿传输了一系列符号，则这个符号序列的信息量（后验信息量）应该是各个符号自信息量的和，而不是信源的熵，请参考本章习题第 2 题。

（3）熵描述了信源输出符号的平均信息量。

由定义 2-2 可见，熵是自信息量的统计平均。

（4）熵 $H(X)$是描述信源 X 所需的最少的比特数。

这是熵很重要的一个意义。信源编码是经典信息论的主要研究内容之一，其主要作用就是进行数据压缩，也就是用尽可能少的比特去描述信源。此物理意义告诉我们熵是数据压缩的下限。如果用少于熵的比特去描述信源，则不可避免地会发生失真；如果比特数大于熵，则通过合理的编码设计还可以进一步压缩。让我们通过一个例子来说明。

例 2-2：信源 X 的概率空间为

X	1	2	3	4	5	6	7	8
$P(X)$	1/2	1/4	1/8	1/16	1/64	1/64	1/64	1/64

该信源的熵等于

$$H(X)=-\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{8}\log\frac{1}{8}+\frac{1}{16}\log\frac{1}{16}+\frac{1}{64}\log\frac{1}{64}+\frac{1}{64}\log\frac{1}{64}+\frac{1}{64}\log\frac{1}{64}+\frac{1}{64}\log\frac{1}{64}\right]$$

= 2bit/信源符号

为该信源给出两种编码方案如下。

编码方案 1：1→0 0 0；2→0 0 1；3→0 1 0；4→0 1 1；5→1 0 0；6→1 0 1；7→1 1 0；8→1 1 1；

编码方案 2：1→0；2→1 0；3→1 1 0；4→1 1 1 0；5→1 1 1 1 0 0；6→1 1 1 1 0 1；7→1 1 1 1 1 0；8→1 1 1 1 1 1；

定义**平均码长**为各符号码长的统计平均，则对于编码方案 1，平均码长等于 $\overline{L_1}=3\text{bit}$；对于方案 2，平均码长等于 $\overline{L_2}=\frac{1}{2}\times1+\frac{1}{4}\times2+\frac{1}{8}\times3+\frac{1}{16}\times4+\frac{1}{64}\times6+\frac{1}{64}\times6+\frac{1}{64}\times6=2\text{bit}$。方案 1 是我们在学习数字电路时非常常见的一种编码方式，用 3bit 描述 8 种状态，属于**等长编码**。方案 2 给出的编码比较奇怪，为各符号分配了不同长度的比特串，属于**变长编码**。通过对平均码长的计算我们发现，方案 2 实现了在统计平均意义下用 2bit 描述 8 种状态，且平均码长等于信源的熵，因此编码方案 2 对信源进行了最大限度压缩。在第 3 章还将重点讨论等长码和变长码。 ■

为了进一步体会熵的含义，我们再看一个例子。

例 2-3：4 个信源 X、Y、Z、W 的概率分布为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} a_1 & a_2 \\ 0.01 & 0.99 \end{bmatrix},\ \begin{bmatrix} Y \\ P(Y) \end{bmatrix}=\begin{bmatrix} b_1 & b_2 \\ 0.4 & 0.6 \end{bmatrix},\ \begin{bmatrix} Z \\ P(Z) \end{bmatrix}=\begin{bmatrix} c_1 & c_2 \\ 0.5 & 0.5 \end{bmatrix},$$

$$\begin{bmatrix} W \\ P(W) \end{bmatrix}=\begin{bmatrix} d_1 & d_2 & d_3 & d_4 & d_5 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix}$$

这 4 个信源的熵分别为：$H(X)=0.08$，$H(Y)=0.971$，$H(Z)=1$，$H(W)=2.32$。通过对这 4 个信源的比较发现，在符号个数相同的条件下，等概分布比不等概分布的不确定性大；在等概分布的前提下，可能的取值越多，不确定性越大，这也符合我们的直观感受。比较 X 和 Z，显然 Z 的结果更难猜中；比较 Z 和 W，显然 W 的结果更难猜中，这也验证了熵是对信源不确定性的度量。 ■

2.3 离散熵的性质

由离散熵的定义式（2-5）可知，熵仅与信源的消息数和概率分布有关，如果消息数 n 是固定的，则熵仅是概率分布的函数。为简洁起见，以下记信源的概率分布为

$$\boldsymbol{P}=(P(a_1),P(a_2),\cdots,P(a_n))=(p_1,p_2,\cdots,p_n) \qquad p_i \geqslant 0\ (i=1,2,\cdots,n),\ \sum_{i=1}^{n} p_i = 1$$

$\boldsymbol{P}=(p_1,p_2,\cdots,p_n)$称为概率分布矢量。由于熵 $H(X)$只和 $\boldsymbol{P}$ 有关，因此记

$$H(\boldsymbol{P})=H(p_1,p_2,\cdots,p_n)=-\sum_{i=1}^{n} p_i \log p_i \tag{2-6}$$

由于 $H(\boldsymbol{P})$是 $\boldsymbol{P}$ 的函数，所以称 $H(\boldsymbol{P})$为**熵函数**。需要说明的是，当某个 $p_i=0$ 时，会出现 0log0 型未定式，但因为 $\lim\limits_{\varepsilon\to 0}\varepsilon\log\varepsilon=0$，所以该未定式取值为 0，因此式（2-6）是合理的。

基于熵函数的定义式（2-6），可得离散熵的若干数学性质如下。

（1）对称性：当 $p_1, p_2, p_3, \cdots, p_n$ 的顺序任意互换时，熵 $H(\boldsymbol{P})$的取值不变。

（2）非负性：$H(\boldsymbol{P})=H(p_1,p_2,\cdots,p_n)\geqslant 0$。

（3）确定性：当且仅当某个 $p_i=1$ 时，$H(\boldsymbol{P})=0$，此时随机变量以概率为 1 取值 a_i，因此成为确定事件，没有不确定性。

（4）递增性：若某个 p_i 分割成 m 个概率 $p_{i1}, p_{i2},\cdots,p_{im}$ 之和，则新信源的熵递增，这是因为信源可能的取值增加，带来了不确定性的增加。

（5）极值性：当信源输出消息数 n 固定时，信源熵 $H(\boldsymbol{P})=H(p_1,p_2,\cdots,p_n)$在等概分布时取最大值，即

$$H(p_1,p_2,\cdots,p_n)\leqslant H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)=\log n$$

此极值性也被称为**最大离散熵定理**。

（6）上凸性：$H(\boldsymbol{P})$是 $\boldsymbol{P}$ 的上凸函数。

该性质的证明比较复杂，感兴趣的读者请参阅文献［2，4］，此处仅以二元信源为例给予直观性的解释说明。

二元信源 X 的概率分布为

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

该信源的熵 $H(X)=H(p,1-p)=-[p\log p+(1-p)\log(1-p)]$，$p\in[0,1]$，绘制 $H(X)$随 p 的函数曲线如图 2-1 所示，可见熵 $H(X)$是概率分布的上凸函数。

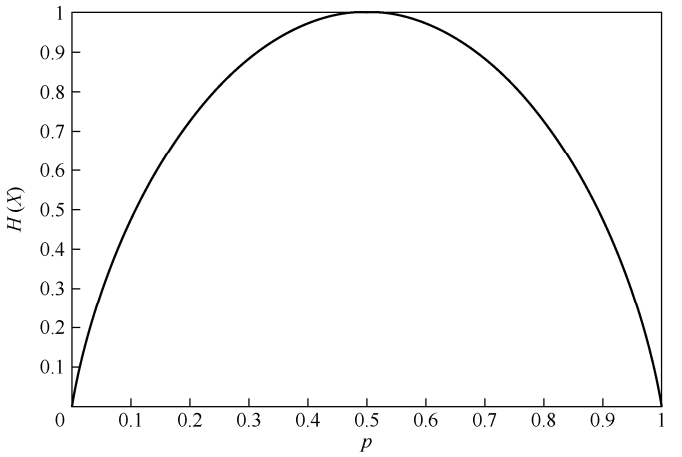

图 2-1　熵函数 $H(X)$的上凸性

2.4　二维信源的联合熵和条件熵

$H(X)$度量了单个随机变量 X 的信息量。进一步地，我们还需要度量由 N 维随机矢量 $\boldsymbol{X}=(X_1, X_2,\cdots, X_N)$构成的 N 维信源的信息量，这就要用到联合熵的概念。为了行文方便，以下仅以二维信源为例加以说明，相关的概念和计算公式可以自然推广到三维以上的信源。

定义 2-3：设二维信源(X_1,X_2)的联合概率分布为

X_1 \ X_2	b_1	b_2	$\cdots$	b_n
a_1	$p(a_1,b_1)$	$p(a_1,b_2)$	$\cdots$	$p(a_1,b_n)$
a_2	$p(a_2,b_1)$	$p(a_2,b_2)$	$\cdots$	$p(a_2,b_n)$
$\vdots$		$\vdots$		
a_m	$p(a_m,b_1)$	$p(a_m,b_2)$	$\cdots$	$p(a_m,b_n)$

定义(X_1,X_2)的**联合熵**为

$$H(X_1,X_2)=-\sum_i\sum_j p(a_i,b_j)\log p(a_i,b_j)\ \text{（bit/2 信源符号）}\qquad (2\text{-}7)$$

可见，联合熵 $H(X_1,X_2)$表征了二维随机变量(X_1,X_2)的总不确定性，因此其单位为 bit/2 信源符号。把联合熵定义 2-3 由二维信源推广到 N 维信源，其联合熵定义为

$$H(X_1,X_2,\cdots,X_N)=-\sum_{i_1}\cdots\sum_{i_N} p(x_{i_1},\ldots,x_{i_N})\log p(x_{i_1},\ldots,x_{i_N})\ \text{（bit/}N\text{ 信源符号）}\qquad (2\text{-}8)$$

作为 N 维信源的一个典型例子，我们考察 N 次无记忆扩展信源的联合熵。

定理 2-1：一维离散信源 X 的 N 次无记忆扩展信源 X^N 的熵为 $H(X^N)=NH(X)$。

该定理的证明请参阅文献［2］，此处仅以一例示之。

例 2-4：一维信源 X 的概率分布为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} a_1 & a_2 & a_3 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

则 X 的熵 $H(X)=-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$ bit/信源符号

对 X 做 2 次无记忆扩展可得 X^2 的联合概率分布为

	a_1	a_2	a_3
a_1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
a_2	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$
a_3	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

计算 X^2 的熵，有 $H(X^2)=\frac{1}{4}\log4+\frac{4}{8}\log8+\frac{4}{16}\log16=\frac{1}{2}+\frac{3}{2}+1=3$ bit/信源符号，可见 $H(X^2)=2H(X)$。 ■

对于定义 2-3 中的二维随机变量(X_1,X_2)，当已知 $X_1=a_i$ 时，X_2 的不确定性为

$$H(X_2\mid X_1=a_i)=-\sum_j p(b_j\mid a_i)\log p(b_j\mid a_i)\qquad (2\text{-}9)$$

式（2-9）描述了在 $X_1=a_i$ 的条件下，X_2 依然存在的不确定性。进一步，把式（2-9）对 a_i 取统计平均可得，当 X_1 已知时，X_2 存在的不确定性为

$$\begin{aligned} H(X_2\mid X_1)&=\sum_i p(a_i)H(X_2\mid X_1=a_i)\\ &=\sum_i p(a_i)\left[-\sum_j p(b_j\mid a_i)\log p(b_j\mid a_i)\right]\\ &=-\sum_i\sum_j p(a_i)p(b_j\mid a_i)\log p(b_j\mid a_i)\\ &=-\sum_i\sum_j p(b_j,a_i)\log p(b_j\mid a_i) \end{aligned}$$

定义 2-4：称 $H(X_2|X_1)$为在 X_1 已知的条件下 X_2 的**条件熵**，其计算公式为

$$H(X_2|X_1)=-\sum_i\sum_j p(b_j,a_i)\log p(b_j\mid a_i) \tag{2-10}$$

可见，条件熵 $H(X_2|X_1)$表示在 X_1 已知的条件下，X_2 仍然存在的不确定性。条件熵 $H(X_2|X_1)$与无条件熵 $H(X_2)$之间满足

$$H(X_2|X_1)\leqslant H(X_2) \tag{2-11}$$

其物理意义是，由于 X_1 的存在对 X_2 提供了一定信息量，使 X_2 的不确定性有一定程度的减小，式（2-11）中等号当且仅当 X_1 和 X_2 统计独立时成立。

定理 2-2（**链式法则**）：联合熵 $H(X_1,X_2)$、条件熵 $H(X_2|X_1)$和无条件熵 $H(X_1)$之间满足

$$H(X_1,X_2)=H(X_2|X_1)+H(X_1) \tag{2-12}$$

证明：
$$\begin{aligned}H(X_1,X_2)&=-\sum_i\sum_j p(a_i,b_j)\log p(a_i,b_j)\\&=-\sum_i\sum_j p(a_i,b_j)\log[p(b_j\mid a_i)p(a_i)]\\&=-\sum_i\sum_j p(a_i,b_j)\log p(b_j\mid a_i)-\sum_i\sum_j p(a_i,b_j)\log p(a_i)\\&=-\sum_i\sum_j p(a_i,b_j)\log p(b_j\mid a_i)-\sum_i p(a_i)\log p(a_i)\\&=H(X_2\mid X_1)+H(X_1)\end{aligned}$$ ■

推广到多个变量的情况，链式法则形如

$$\begin{aligned}&H(X_1,X_2,\cdots,X_N)\\&=H(X_N\mid X_{N-1},\cdots,X_2,X_1)+H(X_{N-1}\mid X_{N-2},\cdots,X_2,X_1)+\cdots+H(X_2\mid X_1)+H(X_1)\end{aligned}$$

链式法则又被称为**熵的强可加性**。特别地，当两个随机变量 X_1 和 X_2 统计独立时，由于条件熵 $H(X_2|X_1)$等于无条件熵 $H(X_2)$，所以有

$$H(X_1,X_2)=H(X_2)+H(X_1)\qquad \text{当且仅当 } X_1 \text{ 和 } X_2 \text{ 统计独立时成立} \tag{2-13}$$

2.5　平均互信息

2.5.1　两个随机变量的平均互信息

对于一个通信系统，如果以 X 表示信源发出的消息符号，以 Y 表示信宿收到的消息符号，那么 X 的熵 $H(X)$描述了信源的先验不确定性。当信宿收到符号 Y 之后关于 X 的后验不确定性是条件熵 $H(X|Y)$，且总有 $H(X|Y)\leqslant H(X)$成立，这就说明信宿在接收到信道传送来的符号 Y 之后得到了一些关于 X 的信息量，使得对信源 X 的不确定性有所减少，把这部分由信源传递给信宿的信息量称为平均互信息。

定义 2-5：称 $I(X;Y)=H(X)-H(X|Y)$为 X 和 Y 的**平均互信息**，简称**互信息**。

以下简记 X 和 Y 的边缘概率分布和联合概率分布分别为 $p(x)$、$p(y)$、$p(x,y)$，则根据定

义 2-5 可得平均互信息的计算公式为

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= \sum_x p(x) \log \frac{1}{p(x)} - \sum_x \sum_y p(x,y) \log \frac{1}{p(x \mid y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{1}{p(x)} - \sum_x \sum_y p(x,y) \log \frac{1}{p(x \mid y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x \mid y)}{p(x)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{2-14}$$

也有文献称 $I(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$ 为符号 x 和 y 的互信息，所以平均互信息 $I(X;Y)$也是所有符号互信息 $I(x;y)$的统计平均。平均互信息的物理意义是，在平均意义下，每个信道输出符号 Y 携带的关于 X 的信息量，其单位是 bit/符号。

提醒读者注意的是，联合熵 $H(X,Y)$中各变量是以“,”分隔，但平均互信息 $I(X;Y)$中两个变量是以“;”分隔，这是信息论著作的标准表示法。这是因为平均互信息也可以定义在两个变量组之间，每个变量组由若干个变量构成，如 $I(X_1,X_2;Y_1,Y_2)$，其含义是 X_1,X_2 作为一个整体和 Y_1,Y_2 这个整体之间的平均互信息，很明显“;”和“,”的作用是不同的。

总结熵 $H(X)$、条件熵 $H(X|Y)$、联合熵 $H(X,Y)$和平均互信息 $I(X;Y)$这 4 种信息测度的含义列于表 2-1，它们之间的关系如图 2-2 所示。

表 2-1　4 种信息测度

熵	$H(X)$	一个随机变量的信息量
联合熵	$H(X, Y)$	多个随机变量的信息量
条件熵	$H(X\|Y)$	在 Y 已知的条件下，X 的信息量
平均互信息	$I(X;Y)$	Y 所包含的关于 X 的信息量，反之亦然

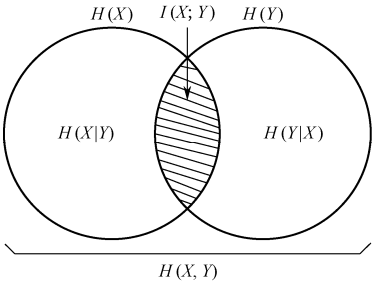

图 2-2　4 种信息测度的关系

根据图 2-2 可知，平均互信息满足如下等式：

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{2-15}$$

平均互信息具有如下性质。

（1）非负性：$I(X;Y) \geqslant 0$。两个随机变量的平均互信息是非负的。注：两个符号 x 和 y

的平均互信息 $I(x;y)$可以为负值；另外，对于 3 个以上随机变量的平均互信息 $I(X;Y;Z)$可能取负值，请见附录 D 中例 D-3。

（2）极值性：$I(X;Y)\leqslant H(X)$。非负性和极值性可以根据平均互信息的定义 $I(X;Y)=H(X)-H(X|Y)$及 $H(X|Y)\leqslant H(X)$直接得到。

（3）对称性：$I(X;Y)=I(Y;X)$。这个性质可以通过观察图 2-2 得到。

（4）凸性。2.3 节离散熵的性质中有一条性质是，“离散熵是信源概率分布的上凸函数，当信源等概分布时，熵取最大值。”平均互信息也具有凸性，但它的凸性既和信源概率分布有关，也和信道转移概率有关，我们把平均互信息的凸性放在第 4 章具体分析。

2.5.2　多个随机变量的平均互信息

类似于熵和条件熵，平均互信息也可以推广到 3 个及 3 个以上变量。

定义 2-6：称

$$I(X;Y\mid Z)=\sum_{x}\sum_{y}\sum_{z}p(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)} \tag{2-16}$$

为在 Z 已知的条件下，X 和 Y 的平均互信息。

$I(X;Y|Z)$满足

$$I(X;Y|Z)=H(X|Z)-H(X|Y,Z) \tag{2-17a}$$

$$=H(Y|Z)-H(Y|X,Z) \tag{2-17b}$$

$$=H(X|Z)+H(Y|Z)-H(X,Y|Z) \tag{2-17c}$$

证明：以式（2-17a）为例，有

$$H(X\mid Z)=-\sum_{x}\sum_{z}p(x,z)\log p(x\mid z)=-\sum_{x}\sum_{y}\sum_{z}p(x,y,z)\log p(x\mid z)$$

$$H(X\mid Y,Z)=-\sum_{x}\sum_{y}\sum_{z}p(x,y,z)\log p(x\mid y,z)$$

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

代入式（2-17a）可证。 ■

比较式（2-15）可以发现，条件平均互信息 $I(X;Y|Z)$只是在平均互信息 $I(X;Y)$的基础上增加了以 Z 为条件，其他形式都是相同的。条件互信息也满足非负性，即

$$I(X;Y|Z)\geqslant 0 \tag{2-18}$$

定义 2-7：称

$$I(X;Y,Z)=H(X)-H(X|Y,Z)=H(Y,Z)-H(Y,Z|X)=H(X)+H(Y,Z)-H(X,Y,Z) \tag{2-19}$$

为随机变量 X 和随机矢量(Y,Z)的联合平均互信息。

$I(X;Y,Z)$表示随机变量 X 和随机矢量(Y,Z)之间互相包含的信息量或统计依存度。$I(X;Y,Z)$也满足对称性和非负性，即

$$I(X;Y,Z)=I(Y,Z;X) \tag{2-20}$$

$$I(X;Y,Z)\geqslant 0 \tag{2-21}$$

联合平均互信息 $I(X;Y,Z)$和条件平均互信息 $I(X;Y|Z)$满足如下关系，即

$$I(X;Y,Z) = I(X;Y|Z) + I(X;Z) = I(X;Z|Y) + I(X;Y) \tag{2-22}$$

证明：$I(X;Y,Z) = H(X) - H(X|Y,Z)$

$$= H(X) - H(X|Z) + H(X|Z) - H(X|Y,Z)$$

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

同理可证，$I(X;Y,Z) = I(X;Z|Y) + I(X;Y)$ ■

式（2-22）也被称为**平均互信息的链式法则**，一般地，有

$$I(X_1, X_2, \cdots, X_n; Y)$$

$$= \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \cdots, X_1)$$

$$= I(Y; X_n|X_{n-1}, \cdots, X_1) + I(Y; X_{n-1}|X_{n-2}, \cdots, X_1) + \cdots + I(Y; X_3|X_2, X_1) + I(Y; X_2|X_1) + I(Y; X_1)$$

2.6 离散平稳信源的熵率

若信源发出的符号序列的各维概率分布与时间起点无关，即对于任意 $n \geqslant 1$，$l \geqslant 1$，$X_1, X_2, \cdots, X_n$ 和 $X_{1+l}, X_{2+l}, \cdots, X_{n+l}$ 具有相同的概率分布，也就是说任意维概率分布不随时间推移而改变，则称这种信源为**离散平稳信源**。

设一维离散信源 X 的概率分布为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_n \\ P(a_1) & P(a_2) & P(a_3) & \cdots & P(a_n) \end{bmatrix}$$

由 X 进行 N 次扩展得到 N 维离散平稳信源$(X_1, X_2, \cdots, X_N)$。

如果是无记忆扩展，则根据定理 2-1，生成的 N 维无记忆扩展信源 X^N 的熵满足 $H(X^N)=NH(X)$，那么平均到每个消息符号上的信息量为

$$\frac{H(X^N)}{N} = H(X) \text{（bit/信源符号）}$$

如果是有记忆扩展，则 N 维序列$(X_1, X_2, \cdots, X_N)$的熵 $H(X_1, X_2, \cdots, X_N)$由联合概率分布 $P(X_1, X_2, \cdots, X_N)$决定，即

$$H(X_1, X_2, \cdots, X_N) = -\sum_{i_1} \cdots \sum_{i_N} p(x_{i_1}, \cdots, x_{i_N}) \log p(x_{i_1}, \cdots, x_{i_N}) \text{（bit/}N\text{ 信源符号）} \tag{2-23}$$

定义 2-8：称联合熵 $H(X_1, X_2, \cdots, X_N)$的算术平均值

$$H_N(X) = \frac{H(X_1, X_2, \cdots, X_N)}{N} \text{（bit/信源符号）} \tag{2-24}$$

为 N 维符号序列$(X_1, X_2, \cdots, X_N)$的**平均符号熵**。

可见，平均符号熵近似度量了 N 维符号序列$(X_1, X_2, \cdots, X_N)$中每个符号的信息。随着 N 取不同的值，可以得到一系列的平均符号熵 $H_N(X)$（N=1,2,3,$\cdots$）。此外，还可以得到一系列的条件熵 $H(X_N|X_{N-1}, \cdots, X_2, X_1)$（$N$=2,3,$\cdots$）。对于离散平稳信源来说，这些熵有一些非常有趣且有用的性质。

性质 1：平均符号熵 $H_N(X)$随着 N 的增加而非递增。

性质 2：条件熵 $H(X_N|X_{N-1},\cdots,X_2,X_1)$随着 N 的增加而非递增。

性质 3：$H_N(X)\geqslant H(X_N|X_{N-1},\cdots,X_2,X_1)$。

性质 4：对于离散平稳信源来说，有

$$\lim_{N\to\infty}H_N(X)=\lim_{N\to\infty}H(X_N\,|\,X_{N-1},X_{N-2},\cdots,X_1)=H_\infty$$

该极限一定存在，称 H_∞ 为**熵率或极限熵**。

证明：根据信源的平稳性可得

$$H(X_{N-1}|X_{N-2},\cdots,X_2,X_1)=H(X_N\,|\,X_{N-1},\cdots,X_2) \tag{2-25}$$

此外，条件熵具有性质

$$H(X_N|X_{N-1},\cdots,X_2)\geqslant H(X_N|X_{N-1},\cdots,X_2,X_1) \tag{2-26}$$

这是因为不等式右侧增加条件 X_1 会提供一定的信息，从而降低 X_N 的不确定性。联立式（2-25）和式（2-26）可得

$$H(X_{N-1}|X_{N-2},\cdots,X_2,X_1)\geqslant H(X_N|X_{N-1},\cdots,X_2,X_1) \tag{2-27}$$

即性质 2 成立。

此外，根据性质 2 和链式法则可得

$$\begin{aligned}NH_N(X)&=H(X_1,X_2,\cdots,X_N)\\&=H(X_N|X_{N-1},\cdots,X_2,X_1)+H(X_{N-1}|X_{N-2},\cdots,X_2,X_1)+\cdots+H(X_2|X_1)+H(X_1)\\&\geqslant NH(X_N|X_{N-1},\cdots,X_2,X_1)\end{aligned}$$

所以有 $H_N(X)\geqslant H(X_N|X_{N-1},\cdots,X_2,X_1)$，即性质 3 成立。又因为

$$\begin{aligned}NH_N(X)&=H(X_1,X_2,\cdots,X_N)\\&=H(X_N|X_{N-1},\cdots,X_2,X_1)+H(X_{N-1},\cdots,X_2,X_1)\\&=H(X_N|X_{N-1},\cdots,X_2,X_1)+(N-1)H_{N-1}(X)\\&\leqslant H_N(X)+(N-1)H_{N-1}(X)\end{aligned}$$

因此有 $H_N(X)\leqslant H_{N-1}(X)$，即性质 1 成立。

综上可知，$H_N(X)$是非负的、单调不增的有界数列，即 $0\leqslant\cdots\leqslant H_N(X)\leqslant H_{N-1}(X)\leqslant H_1(X)\leqslant H_0(X)=\log n$，所以 $H_N(X)$必然存在极限，即

$$\lim_{N\to\infty}H_N(X)=H_\infty$$

最后证明 $H(X_N|X_{N-1},\cdots,X_2)$的极限也等于熵率 H_∞。根据平均符号熵的定义和链式法则，对于任意的正整数 N 和 K，有

$$\begin{aligned}&(N+K)H_{N+K}(X)\\&=H(X_1,X_2,\cdots,X_{N+K})\\&=H(X_{N+K}\,|\,X_{N+K-1},\cdots,X_2,X_1)+H(X_{N+K-1}|\,X_{N+K-2},\cdots,X_2,X_1)+\cdots+\\&\quad H(X_N|X_{N-1},\cdots,X_2,X_1)+H(X_1,X_2,\cdots,X_{N-1})\\&\leqslant(K+1)H(X_N|X_{N-1},\cdots,X_2,X_1)+(N-1)H_{N-1}(X)\end{aligned}$$

由此推出

$$H_{N+K}(X)\leqslant\frac{K+1}{N+K}H(X_N\,|\,X_{N-1},\cdots,X_2,X_1)+\frac{N-1}{N+K}H_{N-1}(X)$$

固定 N，令不等式两端 $K\to\infty$，可得

$$H_\infty \leqslant \lim_{K\to\infty} H(X_N \mid X_{N-1},\cdots,X_2,X_1) \tag{2-28}$$

又根据性质 3 有 $H(X_N|X_{N-1},\cdots,X_2,X_1)\leqslant H_N(X)$，令 $N\to\infty$ 可得

$$\lim_{N\to\infty} H(X_N \mid X_{N-1},\cdots,X_2,X_1) \leqslant H_\infty \tag{2-29}$$

结合式（2-28）和式（2-29）可知，$H(X_N|X_{N-1},\cdots,X_1)$的极限也等于熵率 H_∞，即性质 4 也成立。 ■

对于性质 4 而言，如果不满足平稳性条件，则极限熵 H_∞ 有可能不存在，见例 2-5。

例 2-5：设$\{X_k\}$是一个信源，X_k 相互独立，且 $H(X_k)=k$（$k\geqslant 1$），则

$$\frac{1}{n}H(X_1,X_2,\cdots,X_n)=\frac{1}{n}\sum_{k=1}^{n}k=\frac{n+1}{2}$$

可见该信源的平均符号熵随 $n\to\infty$ 而发散。 ■

对于一般的离散平稳信源，求极限熵 H_∞ 是很困难的。但对于大部分信源，当 N 不是很大时，$H_N(X)$和 $H(X_N|X_{N-1},\cdots,X_2,X_1)$就可以非常接近 H_∞。因此，可以用平均符号熵或条件熵作为 H_∞ 的近似值。

例 2-6：一维信源 X 的概率分布为

$$\begin{bmatrix} X \\ p(x) \end{bmatrix}=\begin{bmatrix} a_1 & a_2 & a_3 \\ \dfrac{11}{36} & \dfrac{4}{9} & \dfrac{1}{4} \end{bmatrix}$$

可得一维信源 X 的熵为 $H_1(X)=-\sum_{i=1}^{3}P(a_i)\log P(a_i)=1.542$（bit/信源符号）

把 X 进行二维有记忆扩展得到二维平稳信源(X_1,X_2)，假设二维信源的联合概率分布为

X_1 \ X_2	a_1	a_2	a_3
a_1	$\frac{1}{4}$	$\frac{1}{18}$	0
a_2	$\frac{1}{18}$	$\frac{1}{3}$	$\frac{1}{18}$
a_3	0	$\frac{1}{18}$	$\frac{7}{36}$

由 X 的边缘分布和联合分布不难得到条件概率分布为

X_1 \ X_2	a_1	a_2	a_3
a_1	$\frac{9}{11}$	$\frac{1}{8}$	0
a_2	$\frac{2}{11}$	$\frac{3}{4}$	$\frac{2}{9}$
a_3	0	$\frac{1}{8}$	$\frac{7}{9}$

条件熵 $H(X_2 \mid X_1) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p(a_i, a_j)\log p(a_j \mid a_i) = 0.87$（bit/信源符号）

联合熵 $H(X_2, X_1) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p(a_i, a_j)\log p(a_i, a_j) = 2.41$（bit/2 信源符号）

所以，平均符号熵 $H_2(X) = \frac{1}{2}H(X_2, X_1) = 1.205$（bit/信源符号）

可见，$H_2(X) \leqslant H_1(X) \leqslant H_0(X) = \log 3 = 1.585$，且 $H_2(X) \geqslant H(X_2|X_1)$。 ■

离散平稳信源熵的 4 个性质说明：当符号序列长度 N 趋于无穷大时，平均符号熵和条件熵都非递增地趋于熵率，因此下面的不等式成立。

$$0 \leqslant H_\infty \leqslant \cdots \leqslant H_N(X) \leqslant \cdots \leqslant H_1(X) \leqslant H_0(X) = \log n$$

其中，$H_0(X)$表示信源 X 的各符号取等概分布，这也是输出 n 个符号的信源可能达到的最大熵（见最大离散熵定理）。这个结论提示我们，对于离散平稳信源而言，实际的信息熵 H_∞ 与具有同样符号集的最大熵 $H_0(X)$之间存在冗余，这些冗余的产生原因有两方面：

（1）信源的一维概率分布非等概，造成了 $H_1(X)$和 $H_0(X)$之间的差异；

（2）信源输出的各符号之间不独立，存在统计依赖关系。

定义 2-9：定义离散平稳信源的冗余度为

$$\gamma = 1 - \frac{H_\infty}{\log n} \tag{2-30}$$

冗余度衡量了信源符号携带信息的有效性，冗余度越大，信源有效性越低，可压缩空间就越大。人类自然语言属于离散平稳信源，以英语为例，英语由 26 个英文字母和一个空格字符构成，共 27 个字符，则

$H_0(X) = \log 27 = 4.76$（bit/信源符号），$H_1(X) = 4.03$（bit/信源符号）

$H_2(X) = 3.32$（bit/信源符号），$H_3(X) = 3.1$（bit/信源符号），$H_\infty = 1.4$（bit/信源符号）

H_1、H_2、H_3、H_∞ 是基于统计得出的，因此英语的冗余度为 0.71。这说明英语中有 71%是由单词中字母的顺序、语句中单词的搭配关系、语法结构和语言习惯决定的，只有 29%是语言使用者可以自由支配的。因此，在存储或传输英语文字时，只需要保留全部篇幅的 29%即可，剩下的 71%可以根据英文的统计特性来恢复，这就是冗余度压缩的理论基础。

2.7　马尔可夫信源

2.7.1　马尔可夫链

马尔可夫随机过程是具有无后效性的随机过程[41,45]，马尔可夫性又称为无后效性。无后效性是指当过程 t 时刻的状态已知时，大于 t 时刻的随机过程所处状态的概率特性只与过程 t 时刻的状态有关，而与 t 时刻之前的状态无关。根据时间参数集合 T 和状态空间集合 E 类型的不同，马尔可夫过程可以分为 3 种类型：

（1）时间离散、状态离散的马尔可夫过程；

（2）时间连续、状态离散的马尔可夫过程；

（3）时间连续、状态连续的马尔可夫过程。

本书仅讨论第 1 种情况，即时间离散、状态离散的马尔可夫过程，又称马尔可夫链。其物理实验原型是直线上的随机游动，如图 2-3 所示，一点以概率 p 或 $1-p$ 分别向右或向左移动，一次移动一步，该点在某个时刻所处的位置仅与其前一时刻所处的位置有关，与更前面的时刻所处的位置无关。

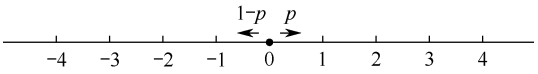

图 2-3　直线上的随机游动

不失一般性，设参数集合 $T=\{0,1,2,3,\cdots\}$；状态集合 $E=\{1,2,3,\cdots\}$或$\{1,2,3,\cdots,n\}$，则马尔可夫链 $X(t)$的无后效性可以表示为

$$P[X(t+k)=j|X(t)=i_t, X(t-1)=i_{t-1},\cdots, X(0)=i_0]= P[X(t+k)=j|X(t)=i_t] \tag{2-31}$$

其中，$t\in T$，$k\in T$，$j, i, i_t, i_{t-1},\cdots, i_0\in E$。

称 $P[X(t+k)=j|X(t)=i]=p_{ij}(t,t+k)$为马尔可夫链在时刻 t 由状态 i 到状态 j 的 k 步转移概率。若该转移概率与时刻 t 无关，即 $p_{ij}(t,t+k)=p_{ij}(t',t'+k)$，则称为**时齐马尔可夫链**，此时简记 $p_{ij}(t,t+k)$为 $p_{ij}(k)$。特别地，时齐马尔可夫链的一步转移概率 $p_{ij}(1)$常记为 p_{ij}。对于有限状态空间 $E=\{1,2,3,\cdots,n\}$，一步转移概率可以写成矩阵，即

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{2-32}$$

式中，$\boldsymbol{P}$ 为时齐马尔可夫链的一步转移概率矩阵。如果状态空间 E 为无限集合，则该矩阵为半无限的。对于状态 i 来说，经过一步转移之后达到的状态一定属于$\{1,2,3,\cdots,n\}$，因此一步转移概率矩阵 $\boldsymbol{P}$ 一定满足 $\sum_j p_{ij}=1$，即矩阵 $\boldsymbol{P}$ 的行取和为 1。同理，k 步转移概率也满足 $\sum_j p_{ij}(k)=1$。k 步转移概率可以写成矩阵的形式，即

$$\boldsymbol{P}(k)=\begin{bmatrix} p_{11}(k) & p_{12}(k) & \cdots & p_{1n}(k) \\ p_{21}(k) & p_{22}(k) & \cdots & p_{2n}(k) \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1}(k) & p_{n2}(k) & \cdots & p_{nn}(k) \end{bmatrix} \tag{2-33}$$

定理 2-3：对于一个时齐马尔可夫链，设 $m=k+l$，则有

$$p_{ij}(m)=p_{ij}(k+l)=\sum_r p_{ir}(k)p_{rj}(l) \tag{2-34}$$

式（2-34）被称为 Chapman-Kolmogorov 方程（简称 C-K 方程），写成矩阵的形式就是

$$\boldsymbol{P}(k+l)=\boldsymbol{P}(k)\boldsymbol{P}(l) \tag{2-35}$$

由此可得，$\boldsymbol{P}(2)=\boldsymbol{P}(1)\boldsymbol{P}(1)=[\boldsymbol{P}(1)]^2$，$\boldsymbol{P}(3)=[\boldsymbol{P}(1)]^3$。一般地，有

$$\boldsymbol{P}(m)=[\boldsymbol{P}(1)]^m \tag{2-36}$$

可见，时齐马尔可夫链的 m 步转移概率完全由其一步转移概率决定。

马尔可夫链在 t 时刻处于各个状态 i 的概率分布被称为绝对概率分布，记为

$$p_i^t = P\{X(t)=i\}\ （i\in E） \tag{2-37}$$

特别地，在初始时刻（0 时刻）的概率分布称为初始概率分布，记为

$$p_i^0 = P\{X(0)=i\}\ （i\in E） \tag{2-38}$$

根据概率的非负性和规范性，有 $p_i^0 \geqslant 0$，$\sum_i p_i^0 = 1$；$p_i^t \geqslant 0$，$\sum_i p_i^t = 1$。在已知初始概率分布和 k 步转移概率的条件下，应用全概公式不难得到 k 时刻的绝对概率分布，即

$$p_j^k = \sum_i p_i^0 p_{ij}(k) \tag{2-39}$$

若马尔可夫链转移概率的极限 $\lim_{k\to\infty} p_{ij}(k) = \pi_j$ 存在，且与 i 无关，则称该马尔可夫链具有**遍历性**，称 π_j 为马尔可夫链的极限概率分布。对于一个遍历的马尔可夫链，有

$$\lim_{k\to\infty} p_j^k = \lim_{k\to\infty}\sum_i p_i^0 p_{ij}(k) = \sum_i p_i^0 \lim_{k\to\infty} p_{ij}(k) = \pi_j \sum_i p_i^0 = \pi_j$$

可见其绝对概率的极限也等于极限概率分布 π_j。遍历性表示马尔可夫链经过一段时间的暂态过程后会进入稳态过程，概率分布会达到一种平稳状态，称平稳状态的概率分布为极限概率分布，概率分布不再随时间而改变，也不依赖于初始状态。

那么，对于一个马尔可夫链，如何判断其是否具有遍历性，又如何求得其极限分布呢？下面的定理给出了对这个问题的解答。

定理 2-4：对于有限状态马尔可夫链，如果存在正整数 k，使得由状态 i 到状态 j 的 k 步转移概率 $p_{ij}(k)$ 满足 $p_{ij}(k)>0$（$i,j=1,2,\cdots,n$），那么这个马尔可夫链是遍历的，即 $\lim_{k\to\infty} p_{ij}(k) = \pi_j$，其中 π_j 表示状态 j 的极限概率，且极限概率 π_j（$j=1,2,\cdots,n$）是方程组

$$\pi_j = \sum_{i=1}^{n} \pi_i p_{ij}\ （j=1,2,\cdots,n）$$

满足条件 $\pi_j>0$ 和 $\sum_{j=1}^{n} \pi_j = 1$ 的唯一解。

2.7.2　马尔可夫信源概述

马尔可夫链在状态转移的同时可以伴随符号的输出，从而构成马尔可夫信源。如果某时刻 t 输出的符号 X_t 仅与之前输出的 m 个符号 $X_{t-1},X_{t-2},\cdots,X_{t-m}$ 有关系，而与 X_{t-m-1} 之前的所有符号都没有关系，则称该马尔可夫信源的记忆阶数是 m，或者说该马尔可夫信源是 m 阶的。为了突出马尔可夫信源的记忆阶数，可以用 m 重符号$(X_1,\cdots,X_m)$作为系统的状态表示，而忽略状态本身的物理意义。

例 2-7：如图 2-4（a）所示的移位寄存器电路的输入和输出相同，输入（输出）0 或 1bit 的概率取决于寄存器当前的状态，系统的 4 个状态为 S_0、S_1、S_2、S_3，由于每个时刻的输出符号仅与之前两个输出符号（寄存器内容）有关，与更前面的输出符号无关，所

以该系统构成一个 2 阶马尔可夫链，状态转移示意如图 2-4（b）所示，其中 x/p 表示在状态转移过程中以概率 p 输出符号 x。

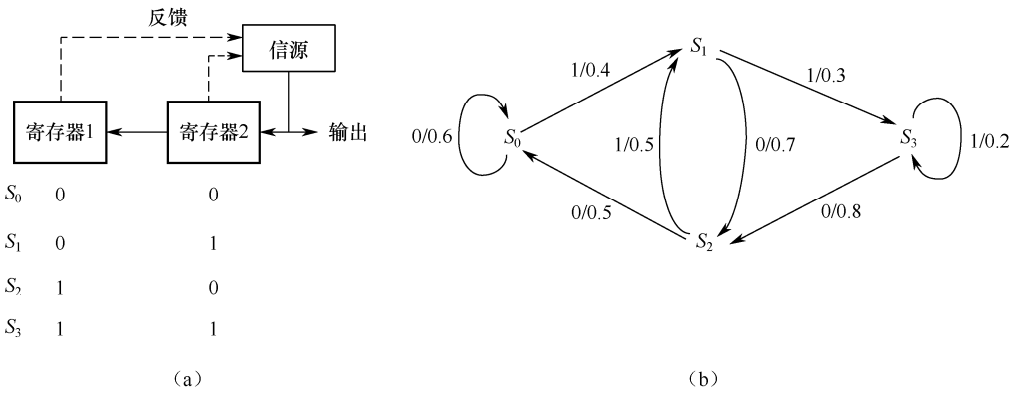

图 2-4　2 阶马尔可夫链 ■

显然，如果信源符号集为$(x_1,\cdots,x_q)$，那么 m 阶马尔可夫信源一共有 q^m 个状态。当马尔可夫链经过暂态过程，进入平稳状态后，马尔可夫信源成为离散平稳信源，其熵率为

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

对于 m 阶马尔可夫信源而言，由于某时刻输出的符号仅与前面的 m 个符号有关，所以上式可改写为

$$H_\infty = H(X_{m+1} \mid X_m,\cdots,X_1) \tag{2-40}$$

可见，m 阶马尔可夫信源的熵率等于 m 阶条件熵。

进一步，根据 m 重符号$(X_1,\cdots,X_m)$和系统状态的一一对应关系，可以用 s_i 表示$(X_1,\cdots,X_m)$所对应的状态，以 $H(X|s_i)$表示系统从 s_i 出发下一个时刻存在的不确定性，则有

$$H_\infty = \sum_i H(X \mid s_i)\pi(s_i) \tag{2-41}$$

可以证明

$$H_\infty = -\sum_i\sum_j p(s_i,s_j)p(s_j \mid s_i) = -\sum_i\sum_j \pi(s_i)p(s_j \mid s_i)\log p(s_j \mid s_i) \tag{2-42}$$

其中，$\pi(s_i)$表示状态 s_i 的极限概率。

例 2-7（续）：应用定理 2-4 计算图 2-4 所示 2 阶马尔可夫链的极限概率，可列方程组如下：

$$\begin{aligned}
\pi_0 &= 0.6\pi_0 + 0.5\pi_2 \\
\pi_1 &= 0.4\pi_0 + 0.5\pi_2 \\
\pi_2 &= 0.7\pi_1 + 0.8\pi_3 \\
\pi_3 &= 0.3\pi_1 + 0.2\pi_3 \\
\pi_0 &+ \pi_1 + \pi_2 + \pi_3 = 1
\end{aligned}$$

解之得，$\pi_0 = 0.345$，$\pi_1 = 0.276$，$\pi_2 = 0.276$，$\pi_3 = 0.103$，代入式（2-41）可得

$H_\infty = 0.345H(0.4,0.6) + 0.276H(0.3,0.7) + 0.276H(0.5,0.5) + 0.103H(0.2,0.8) = 0.9286\text{bit}$ ■

2.8　本章小结

信息的本质是不确定性，信源是具有或表现出不确定性的事物，因此信源的恰当的数学模型是随机变量。为了度量离散随机变量的不确定性，香农提出了离散熵的概念和计算方法，离散熵是整个香农信息论的核心概念。基于离散熵派生出联合熵、条件熵、平均互信息等概念，这些概念代表了 4 种最典型的香农信息测度，具有不同的含义（见表 2-1），需要理解它们的相互关系（见图 2-2）。多维的离散平稳信源的不确定性用平均符号熵来描述，对于有记忆信源，平均符号熵随着符号序列长度 n 的增大而减少，基于离散熵的非负性，平均符号熵最终将趋于一个极限，称为熵率 H_∞。熵率是离散平稳信源的实际熵，由于熵率小于 $H_0(X)$，所以信源存在冗余，这就为信源压缩提供了可能。本章最后讨论了一类特殊的离散平稳信源——马尔可夫信源。马尔可夫信源的根本属性是无后效性，这种信源用转移概率、极限概率（平稳概率）和状态转移图来描述。一般的离散平稳信源的熵率 H_∞ 很难计算，但马尔可夫信源的熵率的计算比较容易。

习　　题

1．请画图表示离散随机变量的熵、联合熵、条件熵、平均互信息之间的关系。

2．设离散无记忆信源

$$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} a_1=0 & a_2=1 & a_3=2 & a_4=3 \\ 3/8 & 1/4 & 1/4 & 1/8 \end{bmatrix}$$

发出的消息为(2 0 2 1 2 0 1 3 0 2 1 3 0 0 1 2 0 3 2 1 0 1 1 0 3 2 1 0 1 0 0 2 1 0 3 2 0 1 1 2 2 3 2 1 0)，求：

（1）此消息的自信息量；

（2）在此消息中平均每个符号携带的信息量。

3．请问每个 QPSK 和 16QAM 脉冲所含的信息量是多少比特（假设发送消息等概分布）？

4．从 0,1,2,⋯,9 这 10 个数字中，任意选出 3 个不同的数字，试求下列事件所含的自信息量：A_1=“3 个数字中不含 0 和 5”，A_2=“3 个数字中含有 0 和 5”，A_3=“3 个数字中含 0 但不含 5”。

5．有两个二元随机变量 X 和 Y，它们的联合概率如下图所示，并定义另一个随机变量 $Z=XY$。

$p(x,y)$	0	1
0	1/8	3/8
1	3/8	1/8

试计算：

（1）$H(X)$，$H(Y)$，$H(Z)$，$H(X,Y)$，$H(X,Z)$，$H(Y,Z)$，$H(X,Y,Z)$；

（2）$H(X|Y)$，$H(Y|X)$，$H(X|Z)$，$H(Z|X)$，$H(Y|Z)$，$H(Z|Y)$，$H(X|Y,Z)$，$H(Y|X,Z)$，$H(Z|X,Y)$；

（3）$I(X;Y)$，$I(X;Z)$，$I(Y;Z)$，$I(X;Y|Z)$，$I(Y;Z|X)$，$I(X;Z|Y)$。

6．一个汽车牌照编号系统使用 3 个字母后接 3 个数字作为代码，问一个汽车牌照所提供的信息量是多少？如果所有 6 个符号任意选用字母、数字作为代码，问一个汽车牌照所提供的信息量是多少？假定有 26 个字母，10 个数字。

7．袋中有 7 个球，其中，红球 5 个，白球 2 个，从袋中取球两次，每次随机地取 1 个球，且第 1 次取出的球不放回袋中，求下列事件包含的信息量：

（1）第 1 次取得白球，第 2 次取得红球；

（2）两次取得的球中 1 个白球，1 个红球；

（3）取得两个球颜色相同。

8．随机将 15 名新生分配到 3 个班级中去，在这 15 名新生中有 3 名优秀生，求下列事件包含的信息量：

（1）每个班级各有 1 名优秀生；

（2）3 名优秀生分在同一班级里。

9．从 5 双不同的鞋子中任取 4 只，“其中至少有两只鞋子配成一双”这一事件包含的信息量是多少？

10．将 C、C、E、E、I、N、S 共 7 个字母随机地排成一行，“恰好排成单词 SCIENCE”这一事件包含的信息量是多少？

11．有 13 个形状、颜色均相同的小球，其中只有一个与其他的重量不同，使用一个没有砝码的天平，采用两边称重的方法找到该小球，问至少得称几次？

12．在彩色电视传输中，每帧图像约有 5×10^5 个像素，为了能很好地重现图像，每像素分为 64 种不同的色彩，每种色彩分为 16 个亮度电平，并假设色彩和亮度均为等概分布，这帧图像包含的信息量是多少比特？

13．设有一个信源，它在开始时以 $P(a)=0.6$，$P(b)=0.3$，$P(c)=0.1$ 的概率输出 X_1。当 X_1 为 a 时，则 X_2 为 a、b、c 的概率为 $1/3$；当 X_1 为 b 时，则 X_2 为 a、b、c 的概率为 $1/3$；当 X_1 为 c 时，则 X_2 为 a、b 的概率为 $1/2$；而且后面输出 X_i 的概率只与 X_{i-1} 有关，且 $P(X_i\,|\,X_{i-1})=P(X_2|X_1)$ $(i\geqslant 3)$。试利用马尔可夫信源的图示法画出状态转移图，并计算熵率 H_∞。

14．黑白气象传真图的消息只有黑色和白色两种，即信源 $X=\{$黑,白$\}$，设黑色出现的概率为 $P(\text{黑})=0.3$，白色出现的概率 $P(\text{白})=0.7$。

（1）假设图上黑白消息出现前后没有关联，求熵 $H(X)$；

（2）假设消息出现前后有关联，其依赖关系为 $P(\text{白}|\text{白})=0.9$，$P(\text{黑}|\text{白})=0.1$，$P(\text{白}|\text{黑})=0.2$，$P(\text{黑}|\text{黑})=0.8$，求此一阶马尔可夫信源的熵 H_2；

（3）分别求上述两种信源的剩余度，并比较 $H(X)$ 和 H_2 的大小，并说明其物理意义。

15．在打靶实验中，假设单次打靶击中的概率为 p，脱靶的概率为 $q=1-p$，持续射击，直到射中为止，以随机变量 X 表示射击次数，求 $H(X)$。

16．X_1 和 X_2 是两个独立同分布的随机变量，概率分布如下表，计算 $H(2X_1)$和 $H(X_1+X_2)$，思考 $H(2X_1)$是否等于 $2H(X_1)$，$H(X_1+X_2)$是否等于 $H(X_1)+H(X_2)$？并对结果加以解释。

X	0	1
$P(X)$	p	q

17．X 是一个有限取值的随机变量，$Y=f(X)$是 X 的函数，证明 $H(X)\geqslant H(Y)$，$H(X\,|\,Y)\geqslant H(Y\,|\,X)$。另外，上述两个不等式成为等式的充要条件是什么？请分别用函数 $Y=2X$和 $Y=\cos X$检查这些不等式。

18．$H\left(p_1,p_2,\cdots,p_n\right)=H(\boldsymbol{P})$ 表示一个 n 维概率矢量所对应的熵，根据离散最大熵定理，当 $\boldsymbol{P}$ 为等概分布时，离散熵 $H(\boldsymbol{P})$ 取最大值 $\log n$，请思考当 $\boldsymbol{P}$ 取何种分布时，离散熵 $H(\boldsymbol{P})$ 取最小值？该最小值等于多少？

19．有两个二元随机变量 X 和 Y，它们的联合概率分布为

X \ Y	0	1
0	1/3	1/3
1	0	1/3

请计算 $H(X)$、$H(Y)$、$H(X,Y)$、$H(X|Y)$、$H(Y|X)$、$I(X;Y)$。

20．一阶齐次马尔可夫信源状态集为 $\{S_1, S_2, S_3\}$，状态转移概率矩阵为

$$\boldsymbol{P}(S_j|S_i)=\begin{bmatrix}1/4 & 1/4 & 1/2\\ 1/3 & 1/3 & 1/3\\ 2/3 & 1/3 & 0\end{bmatrix}$$

（1）画出状态转移图；

（2）计算极限熵。