第 2 章　信源及信源熵

信源（Information Source）是信息的来源，是产生消息（符号）、时间离散的消息序列（符号序列）以及时间连续的消息的来源。

信源输出的消息都是随机的，因此可用概率来描述其统计特性。在信息论中，用随机变量 X、随机矢量 $\boldsymbol{X}$、随机过程 $\{X(e,t)\}$ 分别表示产生消息、消息序列和时间连续消息的信源。

信源的主要问题包括：① 如何描述信源（信源的数学建模问题）；② 怎样定量描述信源输出信息的能力；③ 怎样有效地表示信源输出的消息，也就是信源编码问题。本章介绍前两个问题，重点是第二个问题，即计算信源输出信息的能力——熵率，在第 4、6 章将介绍第三个问题。下面分类介绍信源的数学模型及其熵率的计算。

2.1　信源的分类及其数学模型

第 1 章中已经介绍了离散随机变量及信息熵，离散随机变量表示信源输出的是一个符号的消息，如掷一颗骰子的试验，而通常实际信源输出的消息往往是时间（或空间）的函数，如掷多颗骰子的试验，消息的取值还可能是连续的，如多人跳远比赛的结果。

信源的分类有多种方法，我们常根据信源输出的消息在时间和取值上是离散的或连续的进行分类，如表 2.1 所示。

表 2.1　信源的分类

时间（空间）	取值	信源种类	举　例	数学描述
离散	离散	离散信源 （数字信源）	文字、数据、离散化图像	离散随机变量序列 $P(\boldsymbol{X})=P(X_1X_2\cdots X_N)$
离散	连续	连续信源	跳远比赛的结果、语音信号抽样以后	连续随机变量序列 $P(\boldsymbol{X})=P(X_1X_2\cdots X_N)$
连续	连续	波形信源 （模拟信源）	语音、音乐、热噪声、图形、图像	随机过程 $\{X(e,t)\}$
连续	离散		不常见	

实际信源输出的消息，如平时说话的语声和图像，在时间（或空间）和取值上都是连续的。这样的信源称为波形信源，用随机过程 $\{X(e,t)\}$ 描述。对于频率或时间受限的随机过程，根据抽样定理，人们通常把它转化成时间（或频率）离散的随机序列来处理，这样的信源称为连续信源。取样后的值通常还是连续的，因此还可以进一步经过分层量化，将连续随机变量转化成离散随机变量，连续信源变成离散信源来处理。

此外，根据各维随机变量的概率分布是否随时间的推移而变化，信源可以分为平稳信源和非平稳信源；根据随机变量间是否统计独立，信源可以分为有记忆信源和无记忆信源。

一个实际信源的统计特性往往是相当复杂的，要想找到精确的数学模型很困难。实际应

用时常常用一些可以处理的数学模型来近似。比如，语音信号就是非平稳随机过程，但人们常常用平稳随机过程来近似。平稳随机过程抽样以后的结果就是平稳随机序列。在数学上，随机序列是随机过程的一种，是时间参数离散的随机过程，这里把它单列出来。随机序列特别是离散平稳随机序列是我们研究的主要内容。实际信源的分类如下：

随机过程：波形信源
- 平稳信源
  - 离散平稳信源
    - 离散无记忆信源
    - 离散有记忆信源
      - 记忆长度无限长
      - 记忆长度有限（马尔可夫信源）
  - 连续平稳信源
- 非平稳信源

2.2 离散单符号信源

输出离散取值的单个符号的信源称为离散单符号信源，它是最简单也最基本的信源，是组成实际信源的基本单元，用一个离散随机变量表示。

信源所有可能输出的消息和消息对应的概率共同组成的二元序对$[X,P(X)]$被称为信源的概率空间，即

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} X=x_1 & \cdots & X=x_i & \cdots & X=x_q \\ p(x_1) & \cdots & p(x_i) & \cdots & p(x_q) \end{bmatrix}$$

其中，X 表示信源输出的消息的整体，x_i 表示某个消息，$p(x_i)$表示消息 x_i 出现的概率。q 是信源可能输出的消息数（信源可能输出的消息数可以是有限个，也可以是无限个，通常是有限个），这些消息两两互不相容，信源每次输出其中的一个消息。

注意，$p(x_i)$满足概率空间的非负性和完备性：

$$0\leqslant p(x_i)\leqslant 1$$

$$\sum_{i=1}^{q} p(x_i) = 1$$

信源输出的所有消息的自信息的统计平均值，定义为信源的平均自信息量（信息熵），它表示离散单符号信源的平均不确定性，即

$$H(X) = E[-\log p(x_i)] = -\sum_{i=1}^{q} p(x_i)\log p(x_i) \tag{2.1}$$

【例 2.1】二元信源$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ p & q \end{bmatrix}$，$p+q=1$，求 $H(X)$。

【解】$$\begin{aligned} H(X) &= -\sum_{i=1}^{q} p_i\log p_i \\ &= -p\log p-(1-p)\log(1-p) \\ &= H(p) \end{aligned}$$

$H(X)$是概率 p 的函数，通常用 $H(p)$表示，p 取值于$[0,1]$区间，如图 2.1 所示。

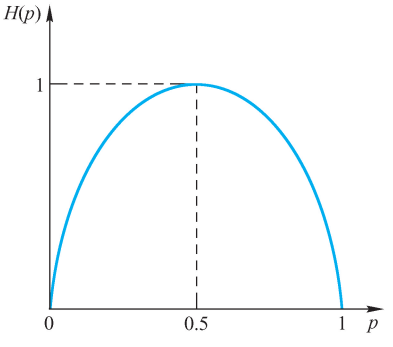

图 2.1 二元熵函数

若输出符号是确定的，即 $p=1$ 或 $p=0$，则 $H(p)=0$，信源不提供任何信息。若 $p=0.5$，即输出符号 0、1 以等概率发生时，信源熵达到极大值，平均每符号等于

1 比特信息量。

2.3 离散多符号信源

前面介绍的单符号信源是最简单的信源模型，用一个离散随机变量表示。实际信源输出的往往是符号序列，称为离散多符号信源，通常用离散随机变量序列（随机矢量）来表示：$\boldsymbol{X}=X_1X_2\cdots$。例如，电报系统发出的是一串有无脉冲的信号（用有脉冲表示 1，无脉冲表示 0），因此电报系统是输出一串 0、1 序列的二元信源。

为简单起见，这里我们只研究离散平稳信源，也就是统计特性不随时间改变的信源。下面先给出离散平稳信源的严格数学定义。

【定义 2-1】对于随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$，在任意两个不同时刻 i 和 j（i 和 j 为大于 1 的任意整数），信源发出消息的概率分布完全相同，也就是对于任意 $N=0,1,2,\cdots$，$X_iX_{i+1}\cdots X_{i+N}\cdots$和 $X_jX_{j+1}\cdots X_{j+N}\cdots$具有相同的概率分布，即

$$P(X_i)=P(X_j) \tag{2.2}$$

$$P(X_iX_{i+1})=P(X_jX_{j+1}) \tag{2.3}$$

$$\vdots$$

$$P(X_iX_{i+1}\cdots X_{i+N})=P(X_jX_{j+1}\cdots X_{j+N}) \tag{2.4}$$

各维联合概率分布均与时间起点无关的信源称为离散平稳信源。

根据式(2.2)至式(2.4)以及联合概率与条件概率的关系，可得

$$P(X_{i+1}\mid X_i)=P(X_{j+1}\mid X_j) \tag{2.5}$$

$$\vdots$$

$$P(X_{i+N}\mid X_iX_{i+1}\cdots X_{i+N-1})=P(X_{j+N}\mid X_jX_{j+1}\cdots X_{j+N-1}) \tag{2.6}$$

即离散平稳信源的条件概率分布均与时间起点无关，而只与关联长度 N 有关。这样我们很容易推出

$$H(X_1)=H(X_2)=\cdots=H(X_N) \tag{2.7}$$

$$H(X_2\mid X_1)=H(X_3\mid X_2)=\cdots=H(X_N\mid X_{N-1}) \tag{2.8}$$

$$H(X_3\mid X_1X_2)=H(X_4\mid X_2X_3)=\cdots=H(X_N\mid X_{N-2}X_{N-1}) \tag{2.9}$$

$$\vdots$$

对于离散单符号信源，我们用信息熵来表示信源的平均不确定性。对于离散多符号信源，怎样表示信源的平均不确定性呢？下面引入“熵率”的概念，它表示信源输出的符号序列中，平均每个符号所携带的信息量。

【定义 2-2】随机变量序列中，对前 N 个随机变量的联合熵求平均称为平均符号熵，即

$$H_N(\boldsymbol{X})=\frac{1}{N}H(X_1X_2\cdots X_N) \tag{2.10}$$

如果 $N\to\infty$ 时上式的极限存在，则 $\lim\limits_{N\to\infty}H_N(\boldsymbol{X})$ 称为熵率，或称为极限熵，记为

$$H_\infty\overset{\text{def}}{=}\lim_{N\to\infty}H_N(\boldsymbol{X}) \tag{2.11}$$

2.3.1 离散平稳无记忆信源

一般情况下，信源输出序列中的每一位出现什么符号是随机的，但是前后符号的出现有

一定的统计关系。为简单起见，我们先假定消息符号序列中前后符号的出现是无关的，即我们首先讨论无记忆信源。

离散平稳无记忆信源输出的符号序列是平稳随机序列，并且符号之间是无关的，即统计独立的。为了研究离散平稳无记忆信源的熵率，我们假定信源每次输出的是 N 长的符号序列，这可视为一个新信源，称为离散平稳无记忆信源的 N 次扩展信源，它的数学模型是 N 维离散随机变量序列（随机矢量）$\boldsymbol{X}=X_1X_2\cdots X_N$，其中每个随机变量之间统计独立。同时，由于是平稳信源，每个随机变量的统计特性都相同，我们还可以把 N 次扩展信源的输出记为 $\boldsymbol{X}=X_1X_2\cdots X_N=X^N$。

根据统计独立的多维随机变量的联合熵与信息熵之间的关系，可以推出

$$H(\boldsymbol{X})=H(X^N)=NH(X) \tag{2.12}$$

即 N 次扩展信源的熵等于单符号离散信源熵的 N 倍，信源输出的 N 长符号序列平均提供的信息量，是单符号离散信源平均每个符号所提供信息量的 N 倍。这似乎很好理解，如抛掷一枚均匀硬币的试验每次可以得到 1 比特的信息量，抛掷 N 枚均匀硬币的试验则可以得到 N 比特的信息量。

离散平稳无记忆信源的熵率

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})=\lim_{N\to\infty}\frac{1}{N}\cdot NH(X)=H(X) \tag{2.13}$$

【例 2.2】设有一离散无记忆信源 X，其概率空间为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & x_3 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

求该信源的熵率及其二次扩展信源（信源每次输出两个符号）的熵。

【解】

单符号离散信源熵为

$$\begin{aligned} H(X) &= -\sum_{i=1}^{q}p_i\log p_i \\ &=\frac{1}{2}\log 2+\frac{1}{4}\log 4+\frac{1}{4}\log 4 \\ &=1.5\ \text{比特/符号} \end{aligned}$$

熵率为

$$\begin{aligned} H_\infty &= \lim_{N\to\infty}H_N(\boldsymbol{X}) \\ &=\lim_{N\to\infty}\frac{1}{N}\cdot NH(X) \\ &=H(X) \\ &=1.5\ \text{比特/符号} \end{aligned}$$

二次扩展信源的熵为

$$H(\boldsymbol{X})=2H(X)=3\ \text{比特/二个符号}$$

注意，$H(\boldsymbol{X})$ 的单位是“比特/二个符号”，其中每个符号提供的信息量仍然是 1.5 比特。

2.3.2 离散平稳有记忆信源

前面介绍了离散平稳信源中最简单的离散平稳无记忆信源，而实际信源往往是有记忆信源。假定信源输出 N 长的符号序列，则它的数学模型是 N 维离散随机变量序列（随机矢量）$\boldsymbol{X}=X_1X_2\cdots X_N$，其中每个随机变量之间存在统计依赖关系。

相互间有依赖关系的 N 维随机变量的联合熵可以用式(2.14)表示，这称为熵函数的链规则：

$$\begin{aligned}H(\boldsymbol{X})&=H(X_1X_2\cdots X_N)\\&=H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\end{aligned}\tag{2.14}$$

N 维随机变量的联合熵等于起始时刻随机变量 X_1 的熵与各阶条件熵之和。

【定理 2-1】对于离散平稳信源，有以下几个结论：

（1）条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 随 N 的增加是递减的。

（2）N 给定时，平均符号熵≥条件熵，即

$$H_N(\boldsymbol{X})\geqslant H(X_N\mid X_1X_2\cdots X_{N-1})\tag{2.15}$$

（3）平均符号熵 $H_N(\boldsymbol{X})$ 随 N 的增加是递减的。

（4）如果 $H(X_1)<\infty$，则 $H_\infty=\lim\limits_{N\to\infty}H_N(\boldsymbol{X})$ 存在，并且

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})=\lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\tag{2.16}$$

【证明】

（1）$\begin{aligned}H(X_N\mid X_1X_2\cdots X_{N-1})&\leqslant H(X_N\mid X_2\cdots X_{N-1})\quad\text{（条件熵小于等于无条件熵）}\\&=H(X_{N-1}\mid X_1X_2\cdots X_{N-2})\quad\text{（序列的平稳性）}\end{aligned}$

所以，条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 随着 N 的增加是递减的。这表明记忆长度越长，条件熵越小，也就是序列的统计约束关系增加时，不确定性减少。

（2）$\begin{aligned}NH_N(\boldsymbol{X})&=H(X_1X_2\cdots X_N)\\&=H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\\&=H(X_N)+H(X_N\mid X_{N-1})+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\quad\text{（序列的平稳性）}\\&\geqslant NH(X_N\mid X_1X_2\cdots X_{N-1})\quad\text{（条件熵小于无条件熵）}\end{aligned}$

所以，$H_N(\boldsymbol{X})\geqslant H(X_N\mid X_1X_2\cdots X_{N-1})$。即 N 给定时，平均符号熵≥条件熵。

（3）$\begin{aligned}NH_N(\boldsymbol{X})&=H(X_1X_2\cdots X_N)=H(X_N\mid X_1X_2\cdots X_{N-1})+H(X_1X_2\cdots X_{N-1})\\&=H(X_N\mid X_1X_2\cdots X_{N-1})+(N-1)H_{N-1}(\boldsymbol{X})\\&\leqslant H_N(\boldsymbol{X})+(N-1)H_{N-1}(\boldsymbol{X})\quad[\text{利用式(2.15)}]\end{aligned}$

所以，$H_N(\boldsymbol{X})\leqslant H_{N-1}(\boldsymbol{X})$。即序列的统计约束关系增加时，由于符号间的相关性，平均每个符号所携带的信息量减少。

（4）只要 X_1 的样本空间是有限的，则必然有 $H(X_1)<\infty$，因此

$$0\leqslant H(X_N\mid X_1X_2\cdots X_{N-1})\leqslant H(X_{N-1}\mid X_1X_2\cdots X_{N-2})\leqslant\cdots\leqslant H(X_1)<\infty$$

从而 $H(X_N\mid X_1X_2\cdots X_{N-1})(N=1,2,\cdots)$ 是一个单调有界数列，极限 $\lim\limits_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})$

必然存在，且极限为 0 和 $H(X_1)$ 之间的某一值。

对于收敛的实数列有以下等式成立：如果 $a_1, a_2, a_3, \cdots$ 是一个收敛的实数列，那么

$$\lim_{N\to\infty}\frac{1}{N}(a_1+a_2+\cdots+a_N)=\lim_{N\to\infty}a_N \tag{2.17}$$

利用式(2.17)可以推出

$$\begin{aligned}\lim_{N\to\infty}H_N(\boldsymbol{X}) &= \lim_{N\to\infty}\frac{1}{N}[H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})]\\ &= \lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\end{aligned}$$

证毕。

定理 2-1 表明，由于信源输出序列前后符号之间的统计依存关系，随着 N 的增加，即统计约束条件不断增加，平均符号熵 $H_N(\boldsymbol{X})$ 及条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 均随之减小。当 $N\to\infty$ 时，$H_N(\boldsymbol{X})=H(X_N\mid X_1X_2\cdots X_{N-1})$，即为熵率，它表示信源输出的符号序列中，平均每个符号所携带的信息量。所以，求离散平稳有记忆信源的熵率时可以有两种途径，可以求它的极限平均符号熵，也可以求它的极限条件熵：

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

一般情况下，平稳信源输出的符号序列中，符号之间的相关性可以追溯到最初的一个符号，如一篇文章的最后一句话可以一直追溯到与开篇第一句话相关。要准确地计算出这个熵率，必须测定信源的无穷维联合概率和条件概率分布，这相当困难。为简化分析，往往用 N 不太大时的平均符号熵或条件熵作为熵率的近似值。比如，英文信源的熵率通常用 $N=5$ 时的条件熵近似。

有一类信源在某时刻发出的符号仅与在此之前发出的有限个符号有关，而与更早些时候发出的符号无关，这类信源称为马尔可夫信源。马尔可夫信源可以在 N 不是很大时得到 H_∞。如果信源在某时刻发出的符号仅与在此之前发出的 m 个符号有关，则称为 m 阶马尔可夫信源，它的熵率为

$$\begin{aligned}H_\infty &= \lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\\ &= \lim_{N\to\infty}H(X_N\mid X_{N-m}X_{N-m+1}\cdots X_{N-1}) \quad (\text{马尔可夫性})\\ &= H(X_{m+1}\mid X_1X_2\cdots X_m) \quad (\text{序列的平稳性})\end{aligned} \tag{2.18}$$

式中，$H(X_{m+1}\mid X_1X_2\cdots X_m)$ 通常记为 H_{m+1}。

【例 2.3】信源 X 的信源模型为

$$\begin{bmatrix}X\\P(X)\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & x_3\\ \frac{1}{4} & \frac{4}{9} & \frac{11}{36}\end{bmatrix}$$

输出符号序列中只有前后两个符号有记忆。条件概率 $P(X_2\mid X_1)$ 列于表 2.2 中。求熵率，并比较 $H(X_2\mid X_1)$、$\frac{1}{2}H(X_1X_2)$ 和 $H(X)$ 的大小。

表 2.2　条件概率 $P(X_2 \mid X_1)$

X_1 \ X_2	x_1	x_2	x_3
x_1	$\frac{7}{9}$	$\frac{2}{9}$	0
x_2	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$
x_3	0	$\frac{2}{11}$	$\frac{9}{11}$

【解】

熵率为

$$H_\infty = H_{m+1} = H(X_2 \mid X_1)$$
$$= 0.870 \text{ 比特/符号}$$

如果不考虑符号间的相关性，则信源熵为

$$H(X) = \frac{1}{4}\log 4 + \frac{4}{9}\log\frac{9}{4} + \frac{11}{36}\log\frac{36}{11}$$
$$= 1.542 \text{ 比特/符号}$$

可见，$H(X_2 \mid X_1) < H(X) = H(X_2)$。这是由于 X_1 和 X_2 之间存在统计依赖关系，在 X_1 已知的情况下，X_2 的不确定度减少，即条件熵 $H(X_2 \mid X_1)$ 小于无条件熵 $H(X_2)$。因此，在考虑序列符号之间的相关性之后，序列的熵减小。

如果把信源输出的符号序列看成是分组发出的，每两个符号作为一组，这样可以把符号序列视为由一个新信源发出的，新信源每次发出的是由两个符号构成的消息。新信源的数学模型是一个二维随机变量，新信源的熵为

$$H(X_1X_2) = H(X_1) + H(X_2 \mid X_1)$$
$$= 1.542 + 0.870$$
$$= 2.412 \text{ 比特/二个符号}$$

平均符号熵为

$$\frac{1}{2}H(X_1X_2) = 1.206 \text{ 比特/符号}$$

可见

$$H(X_2 \mid X_1) < \frac{1}{2}H(X_1X_2) < H(X)$$

这是因为 $H(X_1X_2)$ 考虑了同一组的两个符号之间的相关性，所以 $H(X_1X_2)$ 小于不考虑符号间的相关性时的信源熵 $H(X)$，但 $H(X_1X_2)$ 未考虑前一组的后一符号与后一组的前一符号之间的关联，所以

$$H(X_2 \mid X_1) < \frac{1}{2}H(X_1X_2)$$

2.3.3　马尔可夫信源

前面讨论了离散平稳信源的熵率，由于符号间的相关性可以追溯到很远，使得熵率的计算比较复杂。马尔可夫信源是一类相对简单的有记忆信源，信源在某一时刻发出某一符号的

概率除与该符号有关外，只与此前发出的有限个符号有关。例如，m 阶马尔可夫信源只与前面发出的 m 个符号有关，而 1 阶马尔可夫信源只与前面 1 个符号有关。因此，如果把前面若干个符号视为一个状态（若信源有 q 个可能的输出符号，则一共有 q^m 个可能的状态），那么可以认为，信源在某一时刻发出某一符号的概率除了与该符号有关外，只与该时刻信源所处的状态有关，而与过去的状态无关。信源发出一个符号后，信源所处的状态即发生改变，这些状态的变化组成了马尔可夫链。因此我们可以把对马尔可夫信源的研究转化对马尔可夫链的研究。

如图 2.2 所示，信源在某时刻处于某状态 s_i，当它发出一个符号 $x_{i_{m+1}}$ 后，所处状态就变了，转移到状态 s_j，因此信源输出的符号序列 $X_1X_2\cdots X_mX_{m+1}\cdots$ 变换成信源状态序列 $S_1S_2\cdots S_LS_{L+1}\cdots$，于是讨论信源输出符号的不确定性问题变成了讨论信源状态转移的问题。

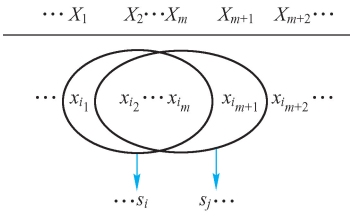

图 2.2　马尔可夫信源

状态之间的一步转移概率 $p_{ij}=P(S_{L+1}=s_j\mid S_L=s_i)$ 表示前一时刻即 L 时刻信源处于 s_i 状态下，在下一时刻即 $L+1$ 时刻信源处于 s_j 状态的概率。用马尔可夫链的状态转移图可以方便地描述离散马尔可夫信源的状态转移概率。

【例 2.4】设有一个二元一阶马尔可夫信源，信源符号集为 $X=\{0,1\}$，输出符号的条件概率为 $p(0\mid 0)=0.25$，$p(0\mid 1)=0.5$，$p(1\mid 0)=0.75$，$p(1\mid 1)=0.5$，求状态转移概率。

【解】

这里 $q=2$，$m=1$，$q^m=2$，信源有两种状态：$s_1=0$，$s_2=1$。

由输出符号的条件概率，可求得信源的状态转移概率：

$$p(s_1\mid s_1)=0.25$$
$$p(s_1\mid s_2)=0.5$$
$$p(s_2\mid s_1)=0.75$$
$$p(s_2\mid s_2)=0.5$$

信源的状态转移概率还可以用如图 2.3 所示的状态转移图表示。

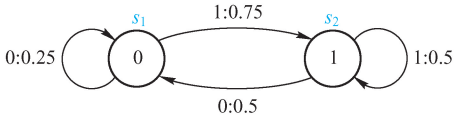

图 2.3　一阶马尔可夫信源状态转移图

对于一阶马尔可夫信源，它的状态转移概率和信源输出符号的条件概率即符号转移概率相同。

【例 2.5】设有一个二元二阶马尔可夫信源，其信源符号集为 $X=\{0,1\}$，输出符号的条件概率为 $p(0\mid 00)=p(1\mid 11)=0.8$，$p(0\mid 01)=p(0\mid 10)=p(1\mid 01)=p(1\mid 10)=$

0.5，$p(1\mid 00)=p(0\mid 11)=0.2$，求状态转移概率矩阵。

【解】

这里 $q=2$，$m=2$，故信源共有 $q^m=4$ 个可能的状态：$s_1=00$，$s_2=01$，$s_3=10$，$s_4=11$。由于信源每次只可能发出 0 或 1，所以信源下一时刻只可能转移到其中的两种状态之一。比如，如果信源原来所处状态为 $s_1=00$，则下一时刻信源只可能转移到 00 或 01 状态，而不会转移到 10 或 11 状态。

由输出符号的条件概率，容易求得状态转移概率：

$$p(s_1\mid s_1)=p(s_4\mid s_4)=0.8$$

$$p(s_2\mid s_1)=p(s_3\mid s_4)=0.2$$

$$p(s_3\mid s_2)=p(s_1\mid s_3)=p(s_4\mid s_2)=p(s_2\mid s_3)=0.5$$

其余状态转移概率为 0，该信源的状态转移图如图 2.4 所示。信源的状态转移概率还可以用状态转移矩阵表示：

$$\boldsymbol{P}=\begin{bmatrix}0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.2 & 0.8\end{bmatrix}$$

解毕。

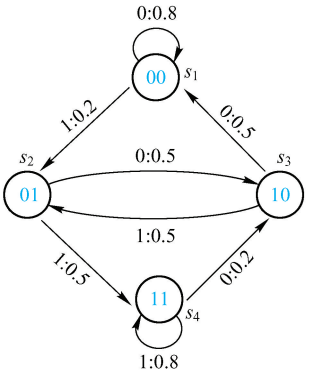

图 2.4　二阶马尔可夫信源状态转移图

对于一般的 m 阶马尔可夫信源，它的所有可能的输出符号及输出符号的条件概率可以组成马尔可夫信源的概率空间：

$$\begin{bmatrix}X\\ P(X)\end{bmatrix}=\begin{bmatrix}x_1 & \cdots & x_i & \cdots & x_q\\ p(x_{i_{m+1}}\mid x_{i_1}x_{i_2}\cdots x_{i_m})\end{bmatrix}$$

令 $s_i=x_{i_1}x_{i_2}\cdots x_{i_m}$（$i_1,i_2,\cdots,i_m\in\{1,2,\cdots,q\}$），则由信源输出符号的条件概率 $p(x_{i_{m+1}}\mid x_{i_1}x_{i_2}\cdots x_{i_m})$ 可以确定状态转移概率 $p(s_j\mid s_i)$（$i,j\in\{1,2,\cdots,q^m\}$），从而得到马尔可夫信源的状态空间：

$$\begin{bmatrix}s_1 & \cdots & s_i & \cdots & s_{q^m}\\ & & p(s_j\mid s_i) & & \end{bmatrix}$$

状态空间由所有状态及状态间的状态转移概率组成。因此通过引入状态转移概率，可以把对马尔可夫信源的研究转化为对马尔可夫链的研究。

我们主要研究遍历的 m 阶马尔可夫信源的熵率。

当时间足够长后，遍历的马尔可夫信源可视为平稳信源来处理，又因为 m 阶马尔可夫信源发出的符号只与最近的 m 个符号有关，所以

$$\begin{aligned}H_\infty&=\lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\\&=\lim_{N\to\infty}H(X_N\mid X_{N-m}X_{N-m+1}\cdots X_{N-1})\quad\text{（马尔可夫性）}\\&=H(X_{m+1}\mid X_1X_2\cdots X_m)\quad\text{（序列的平稳性）}\\&=H_{m+1}\end{aligned}\tag{2.19}$$

即 m 阶马尔可夫信源的极限熵 H_∞ 等于条件熵 H_{m+1}。H_{m+1} 表示已知前面 m 个符号的条件下，

输出下一个符号的平均不确定性。

对于齐次遍历的马尔可夫链，其状态 s_i 由 $x_{i_1}x_{i_2}\cdots x_{i_m}$ 唯一确定，因此

$$p(x_{i_{m+1}} \mid x_{i_1}x_{i_2}\cdots x_{i_m}) = p(x_{i_{m+1}} \mid s_i) = p(s_j \mid s_i) \tag{2.20}$$

所以

$$\begin{aligned} H_{m+1} &= H(X_{m+1} \mid X_1X_2\cdots X_m) \\ &= E[p(x_{i_{m+1}} \mid x_{i_1}x_{i_2}\cdots x_{i_m})] \\ &= E[p(x_{i_{m+1}} \mid s_i)] \\ &= -\sum_{i=1}^{q^m}\sum_{i_{m+1}=1}^{q} p(s_i)p(x_{i_{m+1}} \mid s_i)\log p(x_{i_{m+1}} \mid s_i) \\ &= \sum_i p(s_i)H(X \mid s_i) \qquad (2.21) \\ &= -\sum_i\sum_j p(s_i)p(s_j \mid s_i)\log p(s_j \mid s_i) \qquad (2.22) \end{aligned}$$

式中，$p(s_i)$是马尔可夫链的平稳分布或称状态极限概率，$H(X \mid s_i)$表示信源处于某一状态 s_i 时发出下一个符号的平均不确定性，$p(s_j \mid s_i)$是状态的一步转移概率。

【例 2.6】求图 2.4 中的二阶马尔可夫信源的极限熵。

【解】

根据马尔可夫链状态的分类方法可以判断，图 2.4 中的 4 个状态是不可约的非周期常返态，因此是遍历的。设状态的平稳分布为 $\boldsymbol{W} = [W_1 \quad W_2 \quad W_3 \quad W_4]$，其中 $W_1 = p(s_1)$，$W_2 = p(s_2)$，$W_3 = p(s_3)$，$W_4 = p(s_4)$，遍历的马尔可夫链满足方程 $\boldsymbol{WP} = \boldsymbol{W}$，即

$$\begin{cases} 0.8W_1 + 0.5W_3 = W_1 \\ 0.2W_1 + 0.5W_3 = W_2 \\ 0.5W_2 + 0.2W_4 = W_3 \\ 0.5W_2 + 0.8W_4 = W_4 \end{cases}$$

并且满足

$$W_1 + W_2 + W_3 + W_4 = 1$$

因此，可解得 $W_1 = p(s_1) = \frac{5}{14}$，$W_2 = p(s_2) = \frac{1}{7}$，$W_3 = p(s_3) = \frac{1}{7}$，$W_4 = p(s_4) = \frac{5}{14}$。

所以

$$\begin{aligned} H_\infty &= H_3 = \sum_i p(s_i)H(X \mid s_i) \\ &= \frac{5}{14}H(0.8,0.2) + \frac{1}{7}H(0.5,0.5) + \frac{1}{7}H(0.5,0.5) + \frac{5}{14}H(0.8,0.2) \\ &= 0.80 \text{ 比特/符号} \end{aligned}$$

注意，这时符号的平稳概率分布为

$$\begin{aligned} p(0) &= 0.8p(s_1) + 0.5p(s_2) + 0.5p(s_3) + 0.2p(s_4) = 0.5 \\ p(1) &= 0.2p(s_1) + 0.5p(s_2) + 0.5p(s_3) + 0.8p(s_4) = 0.5 \end{aligned}$$

它与状态的稳定分布是有区别的。

如果不考虑符号间的相关性，则由符号的平稳概率分布可得信源熵 $H(X) = 1$ 比特/符

号，而考虑符号间的相关性后，该信源的熵率为

$$H_\infty = H_{m+1} = H_3 = 0.80 \text{ 比特/符号}$$

2.3.4 信源的相关性和剩余度

前面讨论了离散平稳信源及其熵率。实际的离散信源可能是非平稳的，对于非平稳信源来说，其 H_∞ 不一定存在，但为了方便，通常假定它是平稳的，用平稳信源的 H_∞ 来近似。对于一般的离散平稳信源，求 H_∞ 值也是很困难的，那么进一步假定它是 m 阶马尔可夫信源，用 m 阶马尔可夫信源的条件熵 H_{m+1} 来近似（大多数平稳信源可用马尔可夫信源来近似，即认为输出符号只与前面的有限个符号有关）。$m=1$ 是最简单的离散平稳有记忆信源，这时 $H_{m+1} = H_2 = H(X_2 \mid X_1)$。若再进一步简化信源模型，则可以假设信源为离散平稳无记忆信源，这时可用单符号离散信源的平均自信息量来近似，$H_1 = H(X)$。最后，可以假定信源输出的符号是等概分布的，因此可以用最大离散熵来近似，$H_0 = \log q$。所以，对于一般的离散信源，根据我们研究的目的不同，可以用不同的信源模型来近似。

由定理 2-1 可知

$$\log q = H_0 \geqslant H_1 \geqslant H_2 \geqslant \cdots \geqslant H_{m+1} \geqslant \cdots \geqslant H_\infty$$

对于一个信源，其输出的每个符号实际所携带的平均信息量用熵率 H_∞ 表示。由于信源输出符号间的依赖关系也就是信源的相关性，使信源的 H_∞ 减小，信源输出符号间统计约束关系越长，信源的 H_∞ 越小。当信源输出符号间彼此不存在依赖关系且为等概分布时，信源的 H_∞ 等于最大熵 H_0。例如，信源符号集有 4 个符号，最大熵为 2 比特/符号，输出一个由 10 个符号构成的符号序列，最多包含 $10 \times 2 = 20$ 比特的信息量。假如，由于符号间的相关性或不等概分布，使信源的 H_∞ 减小到 1.2 比特/符号，则输出的符号序列平均所含有的信息量为 $10 \times 1.2 = 12$ 比特，如果信源输出符号间没有相关性并且符号等概分布，则输出 12 比特的信息量只需输出 6 个符号就可以了，说明信源存在剩余。因此，我们引入信源剩余度（冗余度）的概念。

【定义 2-3】 一个信源的熵率（极限熵）与具有相同符号集的最大熵的比值称为熵的相对率，即

$$\eta = \frac{H_\infty}{H_0} \tag{2.23}$$

信源剩余度为

$$\gamma = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log q} \tag{2.24}$$

$H_0 - H_\infty$ 越大，信源的剩余度越大。

信源的剩余度来自两方面：一方面是信源符号间的相关性，相关程度越大，符号间的依赖关系越长，信源的 H_∞ 越小；另一方面是信源符号分布的不均匀性使信源的 H_∞ 减小。当信源输出符号间不存在相关性并且符号为等概分布时，信源的 H_∞ 最大，等于 H_0。一般，平稳信源的极限熵 H_∞ 远小于 H_0。传输一个信源的信息实际只需传输的信息量为 H_∞，如果用二元符号来表示，只需用 H_∞ 个二元符号。

为了最有效地传输信源的信息，就需要掌握信源全部的概率统计特性，即任意 N 维的

概率分布，这显然是不现实的。实际上，往往只能掌握有限 N 维的概率分布，这时传输的信息量为 H_N，因此与理论值 H_∞ 相比，就多传输了 $H_N - H_\infty$。

为了更经济有效地传输信息，需要尽量压缩信源的剩余度，压缩剩余度的方法是尽量减小符号间的相关性，并且尽可能地使信源符号等概分布。第 4 章无失真信源编码中将研究具体的信源剩余度压缩方法。

下面以英文字母为例来说明，信源模型的近似程度不同，计算的信源熵不同。

① 英文字母共 26 个，加上空格 27 个符号。最大熵 $H_0 = \log q = 4.76$ 比特/符号。

② 对在英文书中各字母（包括空格）出现的概率加以统计，不考虑字母之间的依赖关系，可以得到每个字母的概率分布，如表 2.3 所示。

表 2.3　英文字母概率表

字　母	P_i	字　母	P_i	字　母	P_i
空格	0.2	S	0.0502	Y/W	0.012
E	0.105	H	0.047	G	0.011
T	0.072	D	0.035	B	0.0105
O	0.0654	L	0.029	V	0.008
A	0.063	C	0.023	K	0.003
N	0.059	F/U	0.0225	X	0.002
I	0.055	M	0.021	J/Q	0.001
R	0.054	P	0.0175	Z	0.001

因此，如果认为英语字母间是无记忆的，则根据表中的概率可求得

$$H_1 = -\sum_{i=1}^{q} p(x_i)\log p(x_i)$$
$$= 4.03 \text{ 比特 / 符号}$$

③ 若考虑前后两个、三个、若干个字母之间存在相关性，则可根据字母出现的条件概率求得

$$H_2 = 3.32 \text{ 比特/符号}$$
$$H_3 = 3.1 \text{ 比特/符号}$$
$$\vdots$$
$$H_5 = 1.65 \text{ 比特/符号}$$
$$H_\infty = 1.4 \text{ 比特/符号（利用统计推断方法）}$$

如果考虑 5 个字母间的相关性（约等于英文单词的平均长度 4.5），所计算的信源熵已非常接近英文符号的实际信源熵 H_∞。

$$\eta = \frac{H_\infty}{H_0} = \frac{1.4}{4.76} = 0.29$$
$$\gamma = 1 - \eta = 0.71$$

这说明，写英语文章时，71% 是由语言结构定好的，是多余成分，只有 29% 是写文章的人可以自由选择的。可以这样理解，100 页英文书理论上看仅有 29 页是有效的，其余 71 页是多余的。正是由于这一多余量的存在，才有可能对英文信源进行压缩编码。如果对英文信源进行恰当的编码，传输或存储这些符号时可大量压缩篇幅，100 页的英语大约只要 29 页即可。

表 2.4 给出了 5 种文字在不同近似程度下的熵。

表 2.4　5 种文字在不同近似程度下的熵

文字	H_0	H_1	H_2	H_3	…	H_∞	η	γ
英文	4.7	4.03	3.32	3.1		1.4	0.29	0.71
法文	4.7					3	0.63	0.37
德文	4.7					1.08	0.23	0.77
西班牙文	4.7					1.97	0.42	0.58
中文（按 8000 汉字计算）	≈13	9.41	8.1	7.7		4.1	0.315	0.685

【例 2.7】计算汉字的剩余度。假设常用汉字约为 10000 个，其中 140 个汉字出现的概率占 50%，625 个汉字（含 140 个）出现的概率占 85%，2400 个汉字（含 625 个）出现的概率占 99.7%，其余 7600 个汉字出现的概率占 0.3%，不考虑符号间的相关性，只考虑它的概率分布，在这一级近似下计算汉字的剩余度。

【解】为了计算方便，假设每类中汉字出现是等概的，得到表 2.5。

表 2.5　汉字的近似概率表

类　别	汉字个数	所占概率	每个汉字的概率
1	140	0.5	$\frac{0.5}{140}$
2	625 − 140 = 485	0.85 − 0.5 = 0.35	$\frac{0.35}{485}$
3	2400 − 625 = 1775	0.997 − 0.85 = 0.147	$\frac{0.147}{1775}$
4	7600	0.003	$\frac{0.003}{7600}$

不考虑符号间的相关性，只考虑它的概率分布，因此信源的实际熵近似为 $H(X)=9.773$ 比特/汉字，$H_0=13.288$ 比特/汉字，从而 $\gamma=1-\frac{H(X)}{H_0}=0.264$。

从提高信息传输效率的观点出发，人们总是希望尽量去掉剩余度。比如发电报，人们都知道尽可能把电文写得简洁些以去除相关性，如“母病愈”三个字的中文电报就可以表达母亲身体情况好转的消息。但是从提高抗干扰能力角度来看，却希望增加或保留信源的剩余度，因为剩余度大的消息抗干扰能力强。比如，收到电文“母亲病 X，身体健康”，人们很容易把电文纠正为“母亲病愈，身体健康”，而收到电文“母病 X”我们就不知道对方发的是“母病愈”还是“母病危”。

本书从第 4 章开始将讨论信源编码和信道编码，我们可以进一步理解：信源编码减少或消除信源的剩余度，以提高信息的传输效率，信道编码则通过增加冗余度来提高信息传输的抗干扰能力。

2.4　连续信源

连续随机变量的取值是连续的，一般用概率密度函数来描述其统计特征。

单变量连续信源的数学模型为

$$X:\begin{bmatrix}\mathbf{R}\\ p(x)\end{bmatrix}$$

并满足

$$\int_{\boldsymbol{R}} p(x)\,\mathrm{d}x = 1$$

其中，$\mathbf{R}$ 是实数域，表示 X 的取值范围。

对于取值范围有限的连续信源，其数学模型还可表示成

$$X:\begin{bmatrix}(a,b)\\ p(x)\end{bmatrix}$$

并满足

$$\int_a^b p(x)\,\mathrm{d}x = 1$$

其中，(a,b) 是 X 的取值范围。

通过对连续变量的取值进行量化分层，可以将连续随机变量用离散随机变量来逼近。量化间隔越小，离散随机变量与连续随机变量越接近。当量化间隔趋于 0 时，离散随机变量就变成了连续随机变量。通过对离散随机变量的熵取极限，可以推导出连续随机变量熵的计算公式。

假定概率密度函数 $p(x)$ 如图 2.5 所示，把连续随机变量 X 的取值分割成 n 个小区间，各小区间等宽，区间宽度 $\Delta=\dfrac{b-a}{n}$，则变量落在第 i 个小区间的概率为

$$P[a+(i-1)\Delta \leqslant x \leqslant a+i\Delta] = \int_{a+(i-1)\Delta}^{a+i\Delta} p(x)\,\mathrm{d}x = p(x_i)\Delta \tag{2.25}$$

其中，x_i 是 $a+(i-1)\Delta$ 到 $a+i\Delta$ 之间的某个值。当 $p(x)$ 是 X 的连续函数时，由中值定理可知，必存在一个 x_i 值使上式成立，这样，连续变量 X 就可用取值为 $x_i(i=1,2,\cdots,n)$ 的离散变量来近似，连续信源就被量化成离散信源。

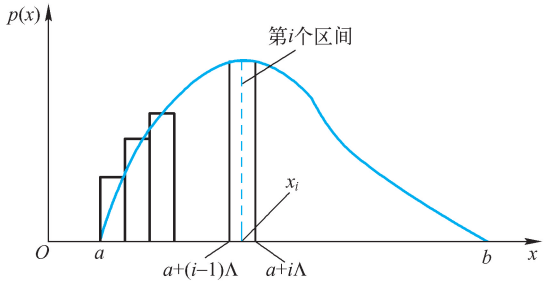

图 2.5　连续随机变量的概率密度函数

这 n 个取值对应的概率分布为 $p_i=p(x_i)\Delta$，这时的离散信源熵是

$$\begin{aligned}H(X) &= -\sum_{i=1}^{n} p(x_i)\Delta\log[p(x_i)\Delta]\\ &= -\sum_{i=1}^{n} p(x_i)\Delta\log p(x_i) - \sum_{i=1}^{n} p(x_i)\Delta\log\Delta\end{aligned} \tag{2.26}$$

当 $n\to\infty$，$\Delta\to 0$ 时，如果上式极限存在，离散信源熵就变成了连续信源的熵：

$$\lim_{\substack{n\to\infty\\ \Delta\to 0}} H(X) = \lim_{\substack{n\to\infty\\ \Delta\to 0}} -\sum_{i=1}^{n} p(x_i)\Delta \log p(x_i) - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \sum_{i=1}^{n} p(x_i)\Delta \log \Delta \tag{2.27}$$

$$= -\int_a^b p(x)\log p(x)\,\mathrm{d}x - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \log\Delta \int_a^b p(x)\,\mathrm{d}x \tag{2.28}$$

$$= -\int_a^b p(x)\log p(x)\,\mathrm{d}x - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \log\Delta \tag{2.29}$$

式(2.29)中的第一项一般是定值，第二项为无穷大量，因此连续信源的熵实际上是无穷大量。这一点是可以理解的，因为连续信源的可能取值是无限多的，它的不确定性是无限大的，当确知输出为某值后，获得的信息量也是无限大的。在丢掉第二项后，我们定义第一项为连续信源的微分熵：

$$h(X) = -\int_{\mathbf{R}} p(x)\log p(x)\,\mathrm{d}x \tag{2.30}$$

微分熵又称为差熵。虽然 $h(X)$ 已不能代表连续信源的平均不确定性，也不能代表连续信源输出的信息量，但是它具有与离散熵相同的形式，也满足离散熵的主要特性，如可加性，但是不具有非负性。另外，我们在实际问题中常常考虑的是熵差，如平均互信息，在讨论熵差时，只要两者离散逼近时所取的间隔 Δ 一致，这两个无限大量将互相抵消，所以熵差具有信息的特性，如非负性。由此可见，连续信源的熵 $h(X)$ 具有相对性。

同样，可以定义两个连续随机变量的联合熵为

$$h(XY) = -\iint_{\mathbf{R}^2} p(xy)\log p(xy)\,\mathrm{d}x\mathrm{d}y \tag{2.31}$$

及条件熵为

$$h(Y\mid X) = -\iint_{\mathbf{R}^2} p(xy)\log p(y\mid x)\,\mathrm{d}x\mathrm{d}y \tag{2.32}$$

$$h(X\mid Y) = -\iint_{\mathbf{R}^2} p(xy)\log p(x\mid y)\,\mathrm{d}x\mathrm{d}y \tag{2.33}$$

并且它们之间也有与离散随机变量一样的相互关系，即

$$h(XY) = h(X) + h(Y\mid X) = H_C(Y) + h(X\mid Y) \tag{2.34}$$

$$h(X\mid Y) \leqslant h(X) \tag{2.35}$$

$$h(Y\mid X) \leqslant H_C(Y) \tag{2.36}$$

【例 2.8】X 是在区间(a,b)内服从均匀分布的连续随机变量，求微分熵。

$$p(x) = \begin{cases} \dfrac{1}{b-a} & x\in(a,b) \\ 0 & x\notin(a,b) \end{cases}$$

【解】

$$h(X) = -\int_a^b p(x)\log p(x)\,\mathrm{d}x$$

$$= -\int_a^b \frac{1}{b-a}\log\frac{1}{b-a}\mathrm{d}x$$

$$= \log(b-a)$$

当$(b-a)>1$ 时，$h(X)>0$；

当$(b-a)=1$时，$h(X)=0$；

当$(b-a)<1$时，$h(X)<0$。

这说明连续熵不具有非负性，失去了信息的部分含义和性质，但是熵差具有信息的特性。

【例 2.9】求均值为 m、方差为 σ^2 的高斯分布的随机变量的微分熵。

【解】高斯随机变量的概率密度为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$$

微分熵

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty}p(x)\log p(x)\,\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty}p(x)\log\left[\frac{1}{\sqrt{2\pi\sigma}}\mathrm{e}^{\frac{(x-m)^2}{2\sigma^2}}\right]\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty}p(x)\log\frac{1}{\sqrt{2\pi}\sigma}-\int_{-\infty}^{+\infty}p(x)\left[-\frac{(x-m)^2}{2\sigma^2}\right]\mathrm{d}x\log\mathrm{e} \\
&= \log\sqrt{2\pi\sigma}+\int_{-\infty}^{+\infty}p(x)\frac{(x-m)^2}{2\sigma^2}p(x)\,\mathrm{d}x\log\mathrm{e} \\
&= \log\sqrt{2\pi}\sigma+\frac{1}{2}\log\mathrm{e} \\
&= \log\sqrt{2\pi\mathrm{e}}\sigma
\end{aligned}$$

这里对数以 2 为底，微分熵的单位为比特/样值，如果对数取以 e 为底，则得到

$$h(X)=\ln\sqrt{2\pi\mathrm{e}}\sigma \text{ 奈特/样值}$$

可以看到，正态分布的连续信源的微分熵与数学期望 m 无关，只与方差 σ^2 有关。

【例 2.10】求指数分布的随机变量的微分熵。

$$p(x)=\begin{cases}\dfrac{1}{a}\mathrm{e}^{-\frac{x}{a}} & x>0\\ 0 & x\leqslant 0\end{cases}$$

【解】

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty}p(x)\ln p(x)\,\mathrm{d}x \\
&= -\int_{0}^{+\infty}p(x)\ln\left(\frac{1}{a}\mathrm{e}^{-\frac{x}{a}}\right)\mathrm{d}x \\
&= -\int_{0}^{+\infty}p(x)\ln\frac{1}{a}\mathrm{d}x-\int_{0}^{+\infty}p(x)\ln\mathrm{e}^{-\frac{x}{a}}\mathrm{d}x \\
&= \ln a\int_{0}^{+\infty}p(x)\,\mathrm{d}x+\frac{1}{a}\ln\mathrm{e}\int_{0}^{+\infty}xp(x)\,\mathrm{d}x \quad \left(\int_{0}^{+\infty}xp(x)\,\mathrm{d}x=a,\int_{0}^{+\infty}p(x)\,\mathrm{d}x=1\right) \\
&= \ln a+\ln\mathrm{e}=\ln a\mathrm{e}
\end{aligned}$$

所以，指数分布的相对熵只取决于信源的均值 a。

【例 2.11】求 N 维高斯信源的熵。

【解】

把 N 维高斯信源输出的 N 维连续随机矢量记为列向量，则其转置为行向量

$$\boldsymbol{X}=[X_1,X_2,\cdots,X_N]^{\mathrm{T}}$$

其均值矢量为

$$\boldsymbol{M}=[m_1,m_2,\cdots,m_N]^{\mathrm{T}}$$

协方差矩阵为

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{bmatrix}$$

其中

$$r_{ij}=E[(x_i-m_i)(x_j-m_j)]\,(i,j=1,2,\cdots,N)$$

N 维联合概率密度为

$$p(x_1x_2\cdots x_N)=\frac{1}{(2\pi)^{\frac{N}{2}}|\boldsymbol{R}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right\}$$

N 维联合熵为

$$\begin{aligned} &h(X_1X_2\cdots X_N)\\ &=-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p(x_1x_2\cdots x_N)\ln p(x_1x_2\cdots x_N)\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p(x_1x_2\cdots x_N)\left[-\ln\sqrt{(2\pi)^N|\boldsymbol{R}|}-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right]\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=\frac{1}{2}\ln[(2\pi)^N|\boldsymbol{R}|]+\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})p(x_1x_2\cdots x_N)\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=\frac{1}{2}\ln[(2\pi)^N|\boldsymbol{R}|]+\frac{N}{2} \end{aligned}$$

当 $X_1,X_2,\cdots,X_N$ 统计独立时，$|\boldsymbol{R}|=\prod_{i=1}^{N}\sigma_i^2$，这时有

$$h(X_1X_2\cdots X_N)=\frac{1}{2}\sum_{i=1}^{N}\ln\sigma_i^2+\frac{N}{2}\ln 2\pi+\frac{N}{2}$$

2.4.1 连续信源的最大熵

离散信源当信源符号为等概分布时有最大熵。连续信源微分熵也有极大值，但是与约束条件有关，当约束条件不同时，信源的最大熵不同。我们一般关心的是下面两种约束下的最大熵。

【定理 2-2】在输出幅度受限的情况，服从均匀分布的随机变量 X 具有最大输出熵。

【证明】设输出幅度限制在 $[a,b]$ 内，则约束条件为

$$\int_a^b p(x)\,\mathrm{d}x=1$$

因此，这是在约束条件下求极值的问题，用拉格朗日乘子法。

令

$$F[p(x)]=h(X)+\lambda\int_a^b p(x)\,\mathrm{d}x$$

由$\frac{\partial F}{\partial p(x)}=-\log p(x)-1+\lambda=0$，得$p(x)=e^{\lambda-1}$。

由$\int_a^b p(x)\mathrm{d}x=\int_a^b e^{\lambda-1}\mathrm{d}x=1$，得$e^{\lambda-1}=\frac{1}{b-a}$，即

$$p(x)=\begin{cases}\frac{1}{b-a} & a\leqslant x\leqslant b\\ 0 & \text{其他}\end{cases}$$

从而

$$\begin{aligned}h(X)&=-\int_a^b p(x)\log p(x)\mathrm{d}x\\&=-\int_a^b\frac{1}{b-a}\log\frac{1}{b-a}\mathrm{d}x\\&=\log(b-a)\end{aligned}$$

因此，对于输出信号幅度受限的连续信源，当满足均匀分布时达到最大熵。这个结论与离散信源在等概分布时达到最大熵的结论类似。

【定理 2-3】对于平均功率受限的连续随机变量，当服从高斯分布时具有最大熵。

【证明】对于均值为m、方差为σ^2的连续随机变量，平均功率P=直流功率+交流功率$=m^2+\sigma^2$。因此，平均功率受限相当于约束条件

$$\int_{-\infty}^{+\infty}p(x)\mathrm{d}x=1$$

$$\int_{-\infty}^{+\infty}xp(x)\mathrm{d}x=m$$

$$\int_{-\infty}^{+\infty}(x-m)^2p(x)\mathrm{d}x=\sigma^2$$

这仍然是在约束条件下求极值的问题。令

$$F[p(x)]=h(X)+\lambda_1\int_{-\infty}^{+\infty}p(x)\mathrm{d}x+\lambda_2\int_{-\infty}^{+\infty}xp(x)\mathrm{d}x+\lambda_3\int_{-\infty}^{+\infty}(x-m)^2p(x)\mathrm{d}x$$

由$\frac{\partial F[p(x)]}{\partial p(x)}=-\ln p(x)-1+\lambda_1+\lambda_2x+\lambda_3(x-m)^2=0$，得$p(x)=e^{\lambda_3(x-m)^2+\lambda_2x+\lambda_1-1}$。代入约束条件中，可得

$$e^{\lambda_1-1}=\frac{1}{\sqrt{2\pi}\sigma}$$

$$\lambda_2=0$$

$$\lambda_3=-\frac{1}{2\sigma^2}$$

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

可以求出$h(X)=\log\sqrt{2\pi e}\sigma$（见例 2.9）。

这说明当平均功率受限时，高斯分布的连续信源的熵最大，也就是说，高斯信源输出的每个样值（也称为自由度）提供的平均信息量最大，其大小随交流功率σ^2的增加而增加。

2.4.2 连续信源的熵功率

与离散信源一样，在讨论了连续信源的最大熵问题之后，我们也要考虑没有达到最大熵