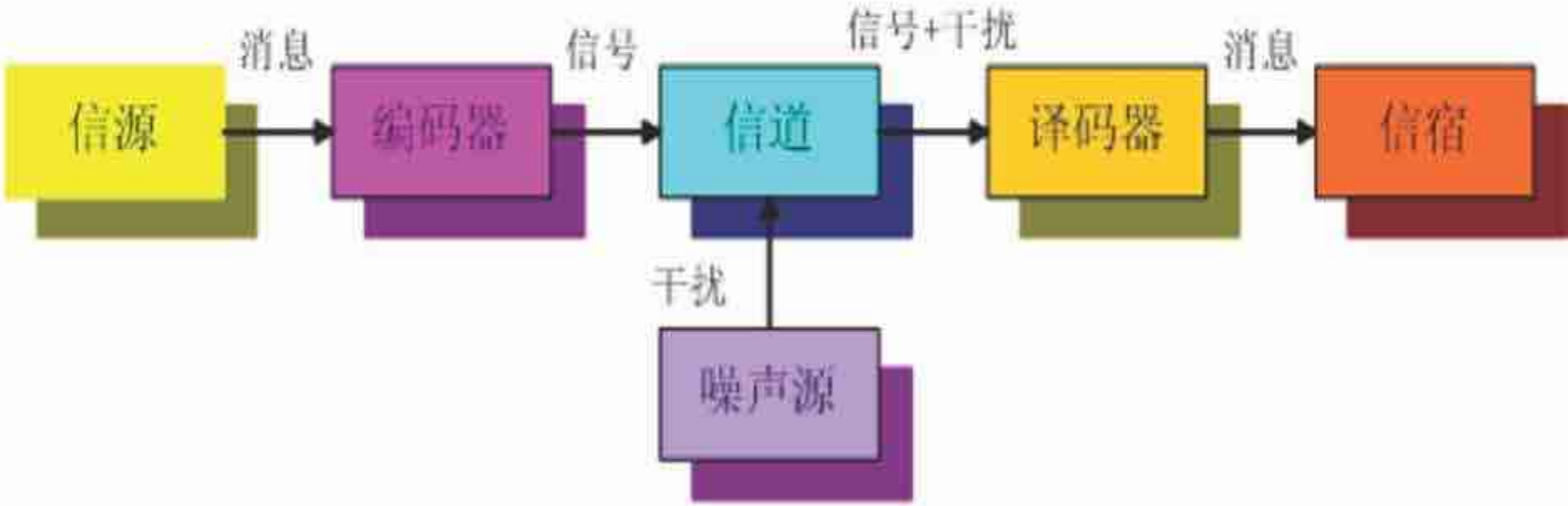

信源
消息
编码器
信号
信道
信号+干扰
译码器
消息
信宿
干扰
噪声源

英语字母并非等概率出现，字母之间有严格的依赖关系。表是对27个符号出现的概率统计结果。

符号	概率	符号	概率	符号	概率
空格	0.2	S	0.052	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

1、下列汉字取自国标(GB 2312-80)中的分级与排列内容：包含所有的第一级汉字和第二级汉字中的常用部分。

2、第一级汉字（16—55区的汉字）以拼音字母为序进行排列，同音字以笔形顺序横、竖、撇、捺、折为序，起笔相同的按第二笔，依次类推；第二级汉字（56-87区的汉字）按部首为序进行排列。

3、对于多音字，仅在表中出现一次。如：柏，音(bai，bo)，表中仅出现在“bai”中。

4、汉字区位码用阿拉伯数字表示，每个汉字对应4个数字。

5、本汉字代码表摘自《字符集和信息编码 国家标准汇编》，(中国标准出版社，1998年编)。

a

啊 1601 阿 1602 吖 6325 嗄 6436 腌 7571 锕 7925

a i

埃 1603 挨 1604 哎 1605 唉 1606 哀 1607 皑 1608 癌 1609 蔼 1610 矮 1611 艾 1612 碍 1613 爱 1614 隘 1615 捱 6263 嗳 6440 嗌 6441 嫒 7040 瑷 7208 暧 7451 砹 7733 锿 7945 霭 8616

a n

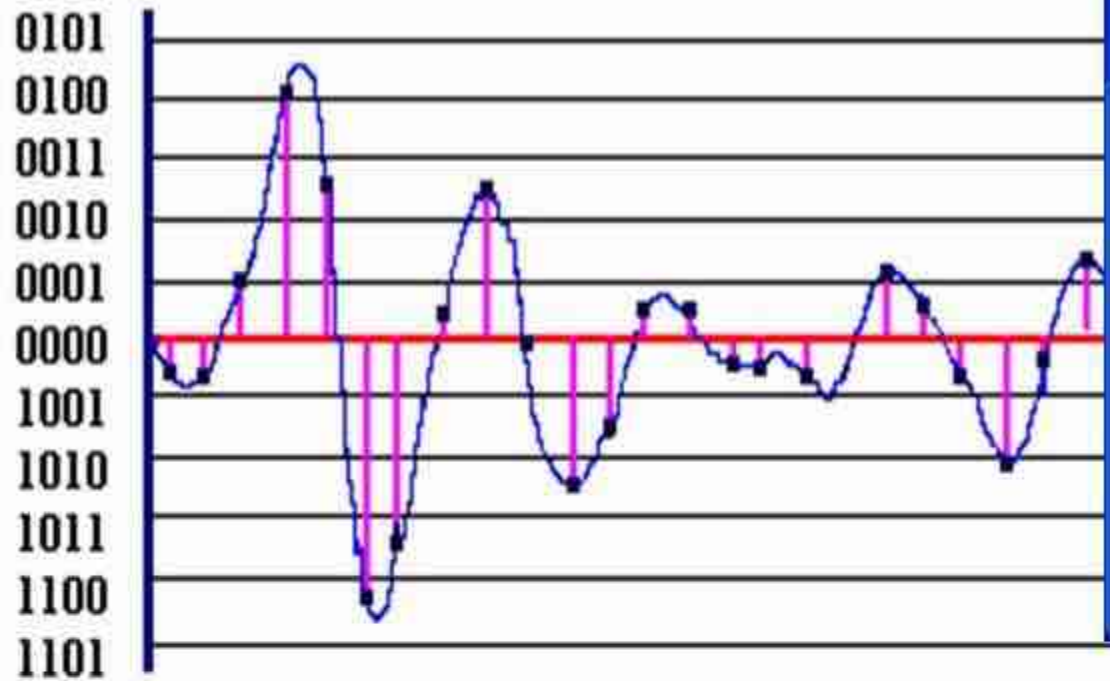
鞍 1616 氨 1617 安 1618 俺 1619 按 1620 暗 1621 岸 1622 胺 1623 案 1624 谙 5847 埯 5991 揞 6278 犴 6577 庵 6654 桉 7281 铵 7907 鹌 8038 黯 8786

a n g

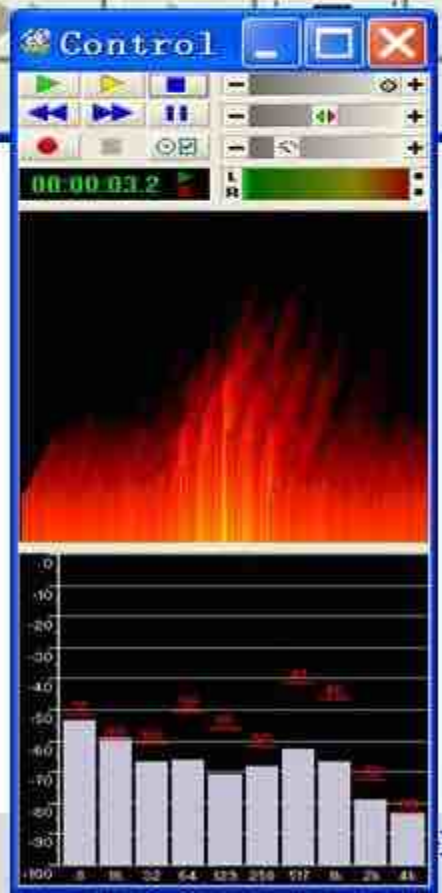
肮 1625 昂 1626 盎 1627

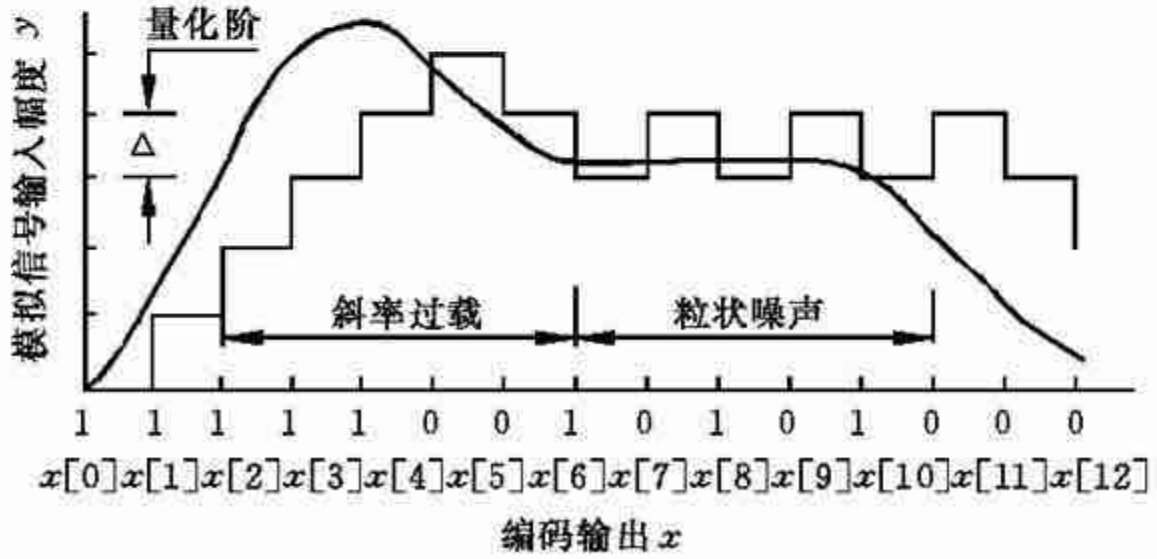
a o

凹 1628 敖 1629 熬 1630 翱 1631 袄 1632 傲 1633 奥 1634 懊 1635 澳 1636 坳 5974 拗 6254 嗷 6427 岙 6514 廒 6658 遨 6959 媪 7033 骜 7081 獒 7365 聱 8190 螯 8292 鏊 8643 鳌 8701 鏖 8773


0101
0100
0011
0010
0001
0000
1001
1010
1011
1100
1101


声音 - 录音机
文件(F) 编辑(E) 效果(S)
帮助(H)
位置:
3.00 秒
长度:
60.00 秒


Control
00:00:03.2


量化阶
Δ
模拟信号输入幅度 y
斜率过载
粒状噪声
1 1 1 1 1 0 0 1 0 1 0 1 0 0 0
x[0]x[1]x[2]x[3]x[4]x[5]x[6]x[7]x[8]x[9]x[10]x[11]x[12]
编码输出 x

IES

信息工程基础

解迎刚
yinggangxie@163.com
Tel:13691117939

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、消息序列以及时间连续消息的信源。

信源的主要问题：

1. 如何描述信源（信源的数学建模问题）
2. 怎样计算信源所含的信息量
3. 怎样有效的表示信源输出的消息，也就是信源编码问题

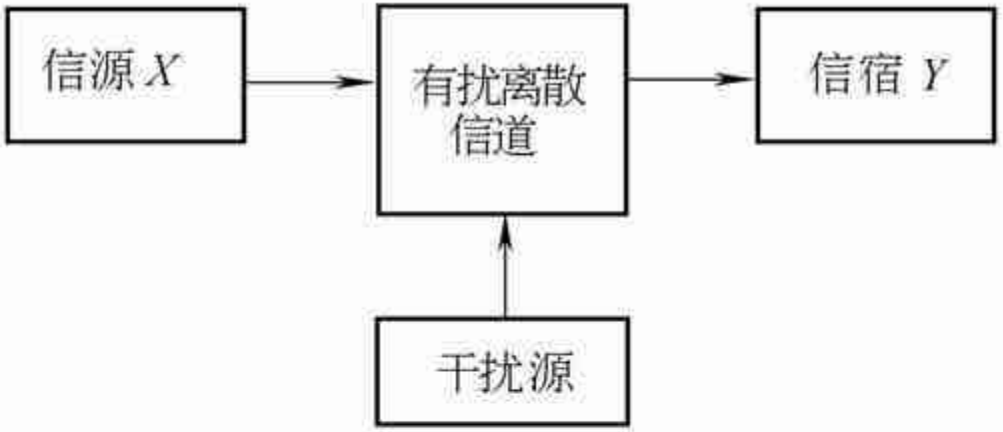
第3章　信源及信源熵

1. 信源的分类及数学模型
2. 离散单符号信源
3. 离散多符号信源
4. 马尔卡夫信源
5. 冗余度

一、信源的分类及数学模型

1. 信源的含义
2. 不同分类模式获得的信息源分类
3. 基于可研究的信源分类方式
3. 信源的数学描述方式

1 信源



- 信源是发出消息的源，信源输出以符号形式出现的具体消息。

  - 发出单个符号的信源：指信源每次只发出一个符号　代表一个消息.
  - 发出符号序列的信源：指信源每次发出一组含二个以上符号的符号序列代表一个消息.

2、不同分类模式获得的信息源分类

- 从信源发出的消息在时间上和幅度上的分布
  - 分为离散信源和连续信源；
- 从信源消息是模拟的还是数字的
  - 分为模拟信源和数字信源；
  - 对数字信源还可分为二进制信源和多进制信源。
- 对于离散信源，根据符号的特点以及符号间的关联性，可分无记忆离散信源和有记忆离散信源
  - 对于前者，又可分为发出单个符号的无记忆离散信源和发出符号序列的无记忆离散信源；
  - 对于后者，又可分为发出符号序列的有记忆离散信源和发出符号序列的马尔可夫（**Markov**）信源。

- 从描述信源消息的随机过程的平稳性角度
  - 分为平稳信源和非平稳信源
- 按随机过程的类别
  - 分为高斯信源、马尔可夫信源等
- 根据人们对信源消息的感知
  - 分为数据信源、文本信源、语音信源、图像信源等，其中文本信源和语音信源都是针对人类语言、文字、声乐等感知的，又通称为自然语信源。

离散信源和连续信源

按照信源发出的消息在时间上和幅度上的分布情况：

- 离散信源是指发出在时间和幅度上都是离散分布的离散消息的信源，如文字、数字、数据等符号都是离散消息。

- 连续信源是指发出在时间和幅度上都是连续分布的连续消息（模拟消息）的信源，如语言、图像、图形等都是连续消息。

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、离散化图象	离散随机变量序列 $P(\mathbf{X}) = P(X_1 X_2 \cdots X_N)$
离散	连续	连续信号	跳远比赛的结果、语音信号抽样以后	连续随机变量序列 $P(\mathbf{X}) = P(X_1 X_2 \cdots X_N)$
连续	连续	波形信源（模拟信源）	语音、音乐、热噪声、图形、图象	随机过程 $\{X(e,t)\}$
连续	离散		不常见	

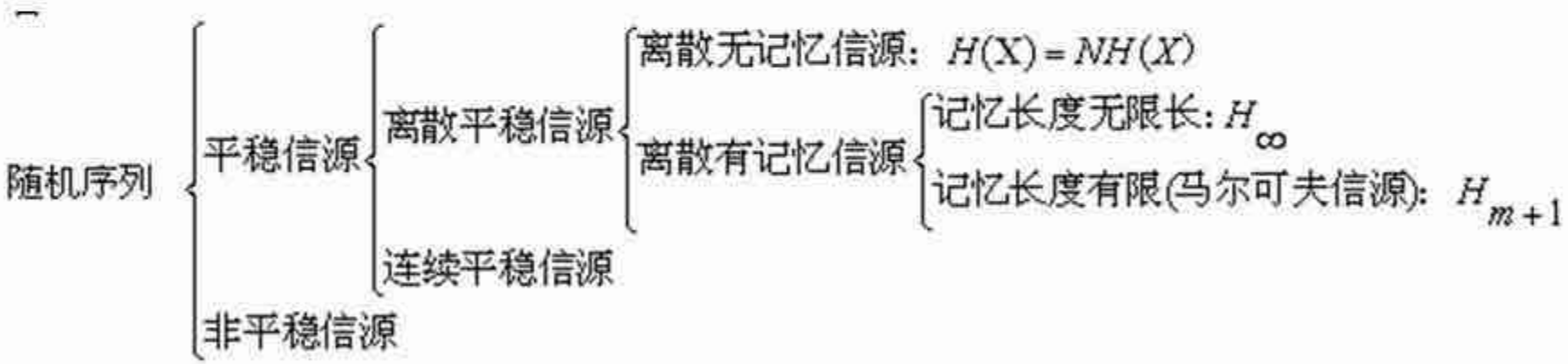
表3.1 信源的分类

3 基于可研究的信源分类方式

- 信源的分类方法可以有多种，但本质上主要基于两方面的考虑：
  - 一是信源消息取值的集合以及消息取值时刻的集合，由此可分为离散信源、连续信源或数字信源、模拟信源等；
  - 二是信源消息的统计特性，由此可分为无记忆（Memoryless）信源、有记忆（Memory）源、平稳信源、非平稳信源、高斯信源、马尔可夫信源等。

可用数学模型近似的实际信源分类

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。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。



4 信源的数学描述方式

例如：一个离散信源发出的各个符号消息的集合

$$X = \{x_1, x_2, \cdots\cdots, x_n\}$$

它们的概率分别为

$$P = \{p(x_1), p(x_2), \cdots\cdots, p(x_n)\}$$

$p(x_i)$ 称为符号 x_i 的先验概率。

把他们写到一起就是概率空间：

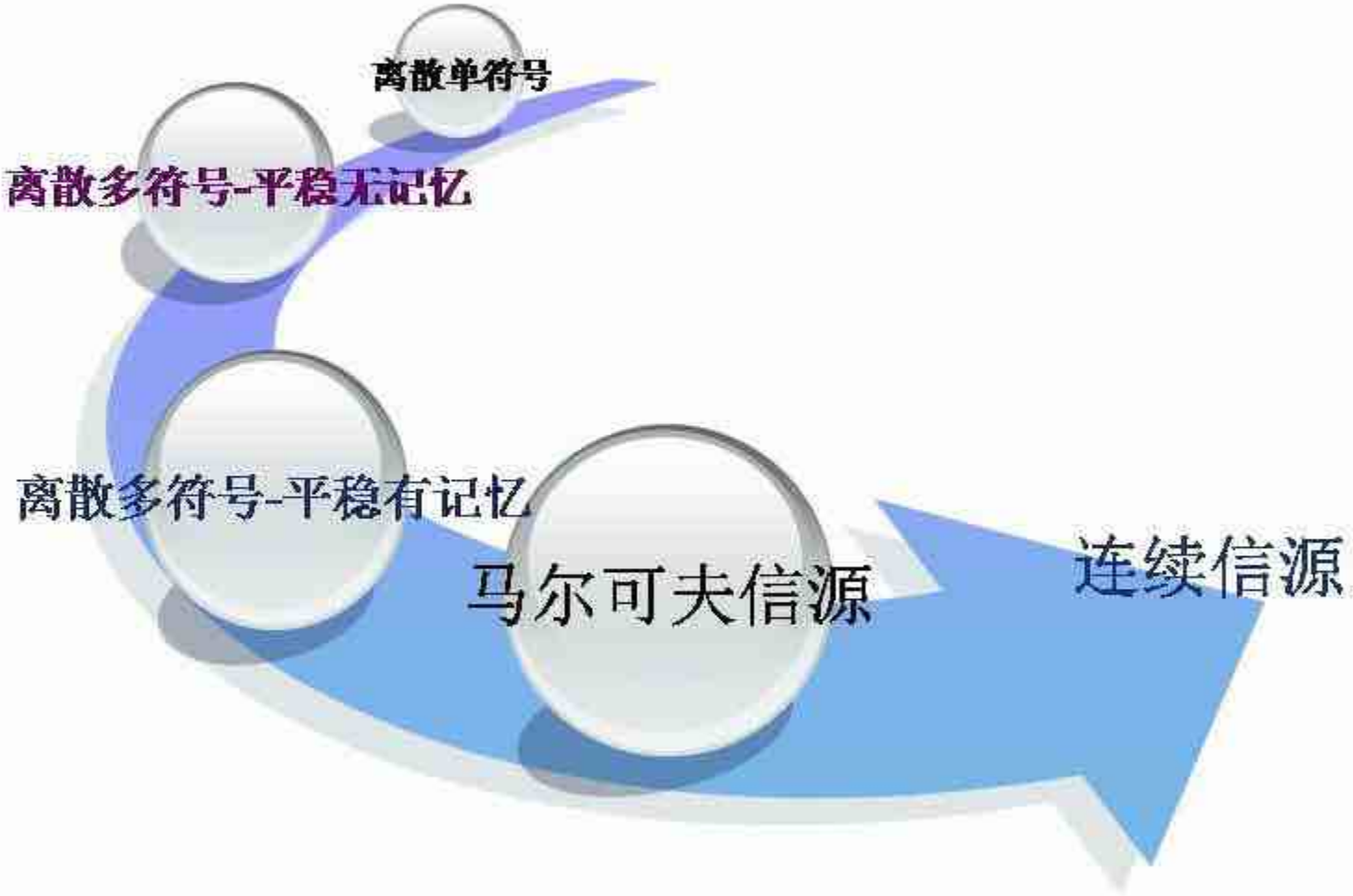
$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots x_n \\ p(x_1) & p(x_2) & \cdots p(x_n) \end{bmatrix}$$

显然有：$p(x_n) \geq 0, \sum_{i=1}^{n} p(x_1) = 1.$

第3章　信源及信源熵

1. 信源的分类及数学模型
2. 离散单符号信源
3. 离散多符号信源
4. 马尔卡夫信源
5. 冗余度

信源的研究的顺序（轨迹）



二、离散单符号信源

1 离散单符号信源

2 离散信源的特例—二元信源

1 离散单符号信源

输出单个离散取值的符号的信源称为**离散单符号信源**。它是最简单也是最基本的信源，是组成实际信源的基本单元。它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 $[X, P(X)]$ 称为信源的**概率空间**：

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

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：

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

2 二元信源

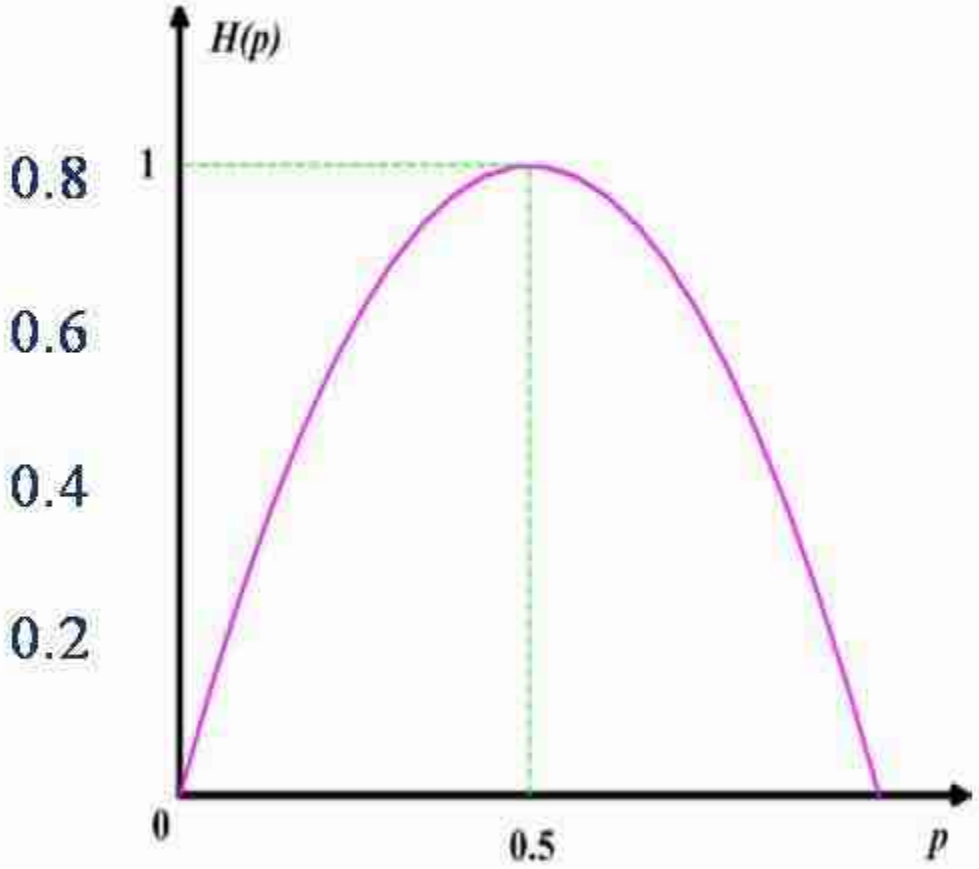
- 二元信源是离散信源的一个特列，该信源X输出符号只有两个，设为0和1。输出符号发生的概率分别为p和q，p＋q=l。即信源的概率空间为

$$\begin{pmatrix} X \\ P \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ p & q \end{pmatrix}$$

则二元信源熵为：

$$H(X) = \sum_{i=1}^{n} p_i Log_2 p_i = -pLog_2 p - qLog_2 q$$

$$= -pLogp - (1-p)Log(1-p) = H(p)$$


H(p)
0.8
1
0.6
0.4
0.2
0
0.5
p

说明：

- 信源信息熵H（X）是概率p的函数，通常用H（p）表示。p取值于[0，1]区间。H（p）函数曲线如图所示。从图中看出，如果二元信源的输出符号是确定的，即p=1或q=1，则该信源不提供任何信息。反之，当二元信源符号0和1以等概率发生时，信源熵达到极大值，等于1比特信息量。

离散多符号信源研究内容

1. 离散多符号信源.—无记忆信源
2. 离散多符号信源.—有记忆信源
3. 离散多符号信源.—马尔可夫信源

三、离散平稳多符号信源

1. 离散多符号信源
2. 离散多符号信源的平均不确定度—熵率
3. 离散多符号无记忆信源定义及熵率计算
4. 离散平稳有记忆信源
5. 离散平稳有记忆信源的熵率计算问题

1 离散多符号信源

定义3.1：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，即对于任意的 $N = 0, 1, 2, \cdots$，$X_i X_{i+1} \cdots X_{i+N} \cdots$ 和 $X_j X_{j+1} \cdots X_{j+N} \cdots$ 具有相同的概率分布。也就是

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

$$P(X_i X_{i+1}) = P(X_j X_{j+1})$$

$$\vdots$$

$$P(X_i X_{i+1} \cdots X_{i+N}) = P(X_j X_{j+1} \cdots X_{j+N})$$

即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。

- 利用联合熵和条件熵的关系，可推出

$$H(X_1)=H(X_2)=\cdots=H(X_N)$$
$$H(X_2|X_1)=H(X_3|X_2)=\cdots=H(X_N|X_{N-1})$$
$$H(X_3|X_1X_2)=H(X_4|X_2X_3)=\cdots=H(X_N|X_{N-2}X_{N-1})$$
$$\vdots$$

对于离散单符号信源，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？

2 熵率（极限熵）

对于离散多符号信源，我们引入熵率的概念，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3.2 随机变量序列中，对前N个随机变量的联合熵求平均：

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

称为平均符号熵。如果当 $N\to\infty$ 时上式极限存在，则 $\lim_{N\to\infty}H_N(X)$ 称为熵率，或称为极限熵，记为

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

3 离散平稳无记忆信源

- 离散无记忆信源：为了方便，假定随机变量序列的长度是有限的，如果信源输出的消息序列中符号之间是无相互依赖关系/统计独立，则称这类信源为离散平稳无记忆信源/离散平稳无记忆信源的扩展。

离散平稳无记忆信源熵率的计算

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，即是统计独立的。假定信源每次输出的是N长符号序列，则它的数学模型是N维离散随机变量序列：$X = X_1 X_2 \cdots X_N$，并且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，我们还可以把一个输出N长符号序列的信源记为：

$$X = X_1 X_2 \cdots X_N = X^N$$

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：

$$H(X) = H(X^N) = N H(X)$$

离散平稳无记忆信源的熵率：

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

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.

单符号离散信源熵

$$H(X) = -\sum_{i=1}^{q} p_i \log_2 p_i = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = 3/2 \text{ 比特 / 符号}$$

X^2信源的符号	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
对应的消息序列	x_1x_1	x_1x_2	x_1x_3	x_2x_1	x_2x_2	x_2x_3	x_3x_1	x_3x_2	x_3x_3
概率 $p(a_i)$	1/4	1/8	1/8	1/8	1/16	1/16	1/8	1/16	1/16

二次扩展信源的熵 $H(X) = -\sum_{i=1}^{q} p_i \log_2 p_i$

二次扩展信源的熵 $H(\boldsymbol{X}) = 2H(X) = 3$ 比特/二个符号

熵率 $H_\infty = \lim_{N\to\infty} H_N(\boldsymbol{X}) = \lim_{N\to\infty} \frac{1}{N} \times NH(X) = 3/2$ 比特/符号

对上述结论的解释：因为扩展信源X^N的每一个输出符号a_i是由N个x_i所组成的序列，并且序列中前后符号是统计独立的。现已知每个信源符号x_i含有的平均信息量为$H(X)$，那么，N个x_i组成的无记忆序列平均含有的信息量就为$NH(X)$（根据熵的可加性）。因此信源X^N每个输出符号含有的平均信息量为$NH(X)$。

注意，$H(X)$的单位在这里是“比特/二个符号”，

其中每个符号提供的信息量仍然是1.5 bit.

4 离散平稳有记忆信源

实际信源往往是有记忆信源。 对于相互间有依赖关系的N维随机变量的联合熵存在以下关系（熵函数的链规则） ：

$$H(\boldsymbol{X}) = H(X_1X_2\cdots X_N)$$
$$= H(X_1)+H(X_2|X_1)+H(X_3|X_1X_2)+\cdots+H(X_N|X_1X_2\cdots X_{N-1})$$

定理3.1 对于离散平稳信源，有以下几个结论：

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

（2）N给定时平均符号熵大于等于条件熵，即

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

（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$H_\infty = \lim_{N\to\infty} H_N(\boldsymbol{X})$存在，并且

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

例：设二维离散信源$\boldsymbol{X}=X_1X_2$的原始信源X的信源模型为

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

$\boldsymbol{X}=X_1X_2$中前后两个符号的条件概率：

$P(X_2/X_1)$ X_2 / X_1	x_1	x_2	x_3
x_1	7/9	2/9	0
x_2	1/8	3/4	1/8
x_3	0	2/11	9/11

原始信号X的熵

$$H(X)=\sum_{i=1}^{3}p(x_i)\log_2\frac{1}{p(x_i)}=-\frac{1}{4}\log_2(\frac{1}{4})-\frac{4}{9}\log_2(\frac{4}{9})-\frac{11}{36}\log_2(\frac{11}{36})=1.542(比特/符号)$$

由上表的条件概率确定条件熵

$$H(X_2/X_1)=\sum_{i_1=1}^{3}\sum_{i_2=1}^{3}p(x_{i_1})p(x_{i_2}/x_{i_1})\log_2\frac{1}{p(x_{i_2}/x_{i_1})}=0.870(比特/符号)$$

- 条件熵$H(X_2/X_1)$比信源熵/无条件熵$H(X)$减少了0.672比特/符号，这是由于符号之间的依赖性造成的；
- 信源平均每发一个消息提供的信息量，即联合熵

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

- 每一个信源符号提供的平均信息量

$$H_2(X)=(1/2)H(\boldsymbol{X})=(1/2)H(X_1X_2)=1.206(\text{比特/符号})$$

- $H_2(X)$小于信源提供的平均信息量$H(X)$，这同样是由于符号之间的统计相关性所引起。

归纳

- 极限熵的存在性：当离散有记忆信源是平稳信源时，从数学上可以证明，极限熵是存在的，且等于关联长度

$N\to\infty$时，条件熵$H(X_N/X_1X_2\ldots X_{N-1})$的极限值，即

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})=\lim_{N\to\infty}\frac{1}{N}H(X_1X_2\ldots X_{N-1}X_N)=\lim_{N\to\infty}H(X_N/X_1X_2\ldots X_{N-1})$$

- 极限熵的含义：代表了一般离散平稳有记忆信源平均每发一个符号提供的信息量。
- 极限熵的计算：必须测定信源的无穷阶联合概率和条件概率分布。这是相当困难的。有时为了简化分析，往往用条件熵或平均符号熵作为极限熵的近似值。在有些情况下，即使N值并不大，这些熵也很接近极限熵。例如马尔科夫信源

四、马尔卡夫信源

1. 马尔科夫特性
2. 马尔科夫信源
3. 马尔科夫信源的计算条件
4. 马尔可夫信源序列熵的计算

1 马尔可夫(Markov)过程

- 一个随机过程 {Xn，n=0，1，2，…} 就是一族随机变量，而Xn能取的各个不同的值，则称为状态。如果一个随机过程 {Xn，n=0，1，2，…} ，由一种状态转移到另一种状态的转移概率只与现在处于什么状态有关，而与在这时刻之前所处的状态完全无关，即如果过程 {Xn，n=0，1，2，…} 中，Xn+1的条件概率分布只依赖于Xn的值，而与所有更前面的值相互独立，则该过程就是所谓马尔可夫(Markov)过程.

齐次马氏链

- 马尔可夫链是指时间离散，状态也离散的马尔可夫过程。一个马尔可夫链，若从u时刻处于状态i，转移到t＋u时刻处于状态j的转移概率与转移的起始时间u无关，则称之为齐次马尔可夫链，简称齐次马氏链。

2 马尔可夫信源

有一类信源，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，这称为马尔可夫性，这类信源称为马尔可夫信源。马尔可夫信源可以在N不很大时得到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：

$$\begin{aligned} H_\infty &= \lim_{N\to\infty} H(X_N \mid X_1 X_2 \cdots X_{N-1}) \\ &= \lim_{N\to\infty} H(X_N \mid X_{N-m} X_{N-m+1} \cdots X_{N-1}) \quad \text{（马尔可夫性）} \\ &= H(X_{m+1} \mid X_1 X_2 \cdots X_m) \quad \text{（平稳性）} \end{aligned}$$

马尔可夫信源—马尔科夫链

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。因此我们把前面若干个符号看作一个状态，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氏链。

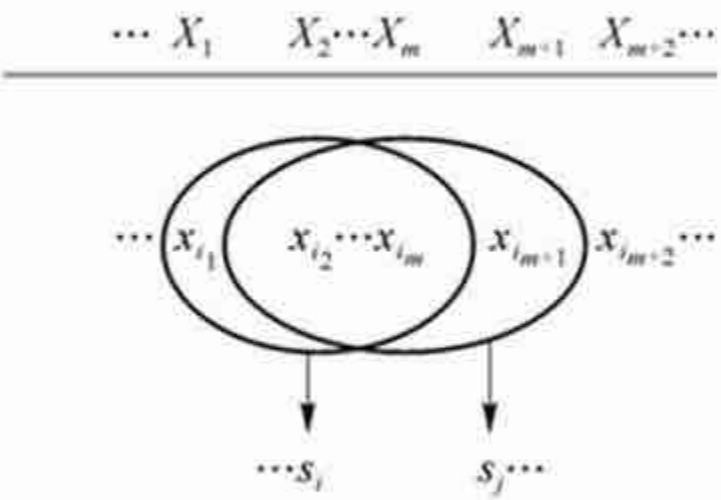


图3.1 马尔可夫信源

M 阶马尔可夫信源

对于一般的m 阶马尔可夫信源，它的概率空间可以表示成：

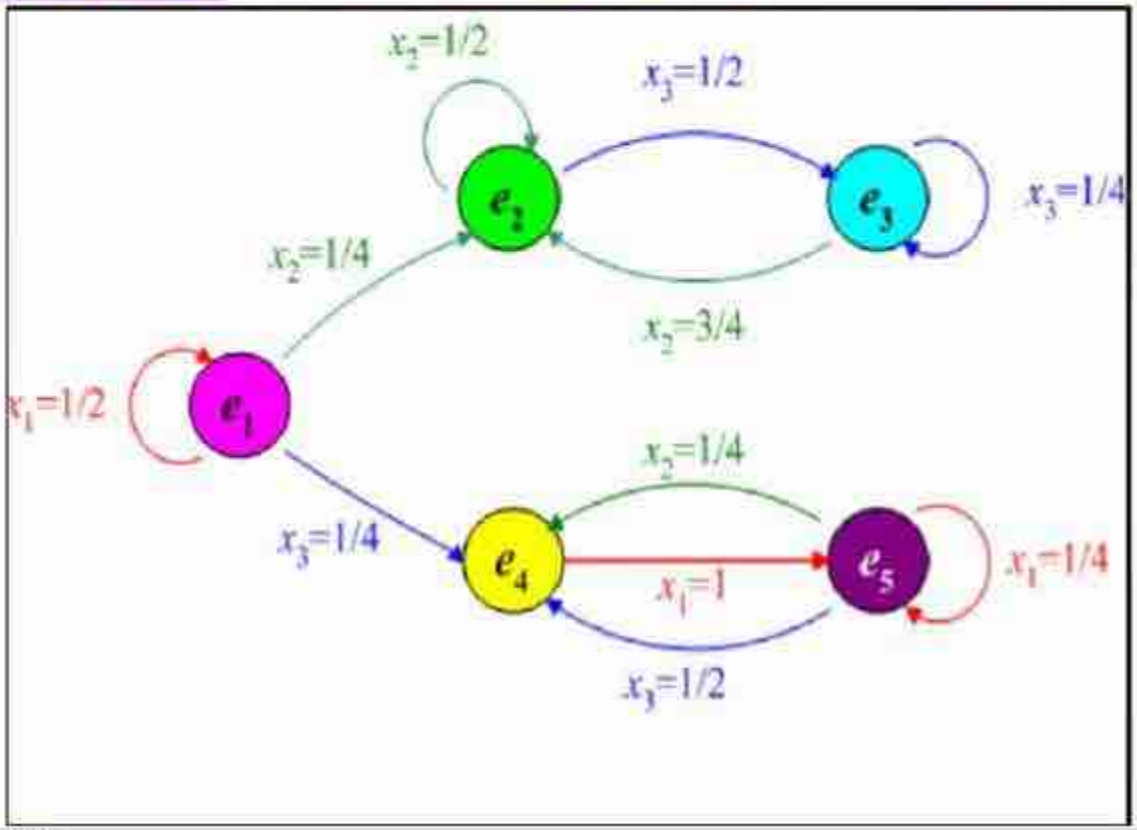
$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_q \\ p(x_{i_{m+1}} \mid x_{i_1} x_{i_2} \cdots x_{i_m}) \end{bmatrix}$$

令 $s_i = x_{i_1} x_{i_2} \cdots x_{i_m}$，$i_1, i_2, \cdots, i_m \in \{1, 2, \cdots, q\}$，从而得到马尔可夫信源的状态空间：

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

- 设信源符号 $X \in \{x_1, x_2, x_3\}$ 信源所在的状态。

$S \in \{e_1, e_2, e_3, e_4, e_5\}$ 各状态之间的转移情况由图给出。



- 若信源处于某一状态 e_i，当它发出一个符号后，所处的状态就变化了；
- 任何时刻信源处在什么状态完全由前一刻的状态和发出的符号决定；
- 因为条件概率 $P(x_k/e_i)$ 已给定，所以状态的转移满足一定的概率分布，并可求出状态的一步转移概率 $P(e_j/e_i)$。
- 马尔科夫链的状态转移图：每个圆圈代表一种状态，状态之间的有向线表示某一状态向另一状态的转移。有向线一侧的符号和数字分别代表发出的符号和条件概率

由图中可得状态的一步转移概率：

该信源满足马尔可夫信源定义。

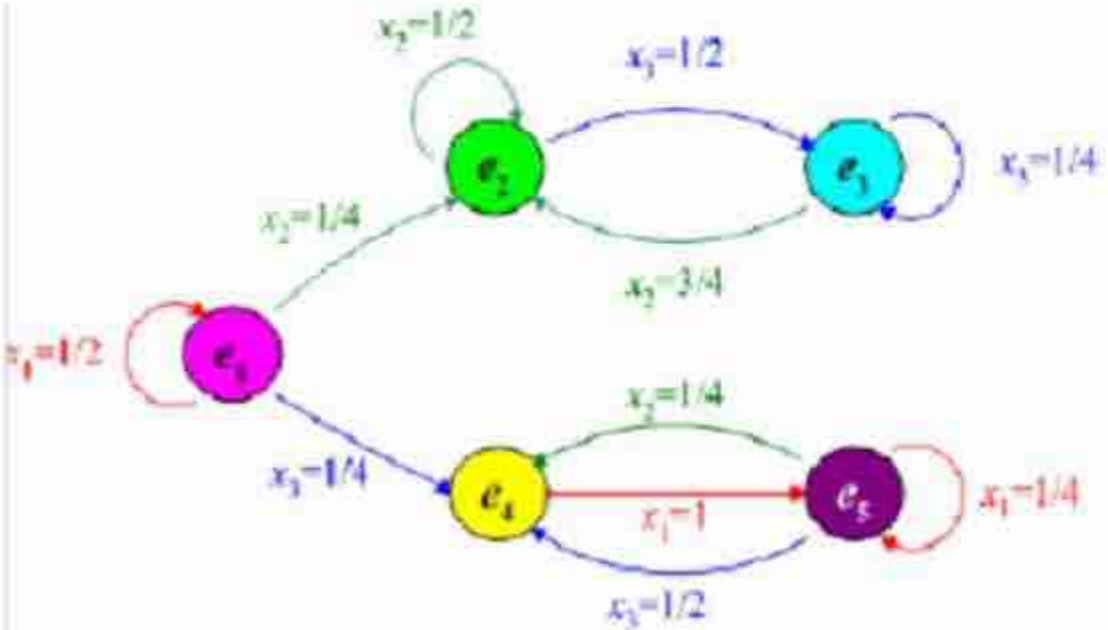


图2.2.1 状态转移图

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

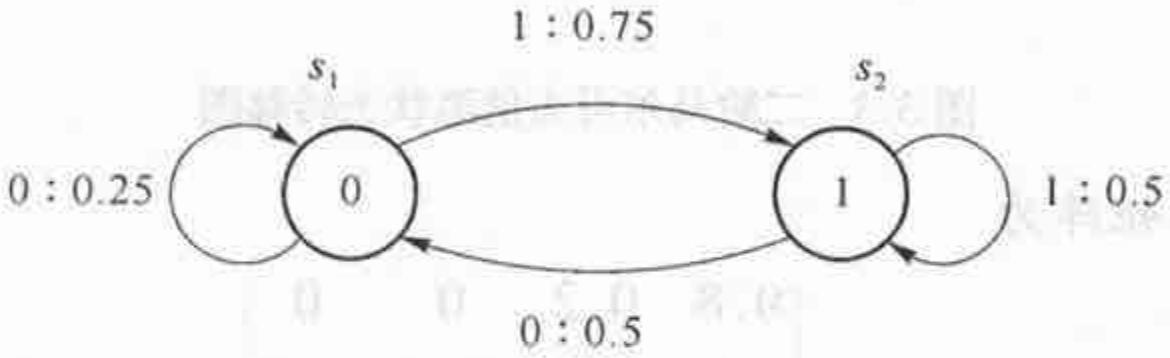
结论

- 一般有记忆信源发出的是有关联的各符号构成的整体消息，即发出的是符号序列，并用符号间的联合概率描述这种关联性；
- 马尔科夫信源的不同之处在于它用符号之间的转移概率/条件概率来描述这种管理关系。即马尔科夫信源是以转移概率发出每个信源符号。
- 转移概率的大小取决于它与前面符号之间的关联性。

设一个二元一阶马尔可夫信源，信源符号集为 $X=\{0,1\}$，信源输出符号的条件概率为

$$p(0|0)=0.25, p(0|1)=0.5, p(1|0)=0.75, p(1|1)=0.5$$

求状态转移概率.



得马尔可夫链的状态转移概率：

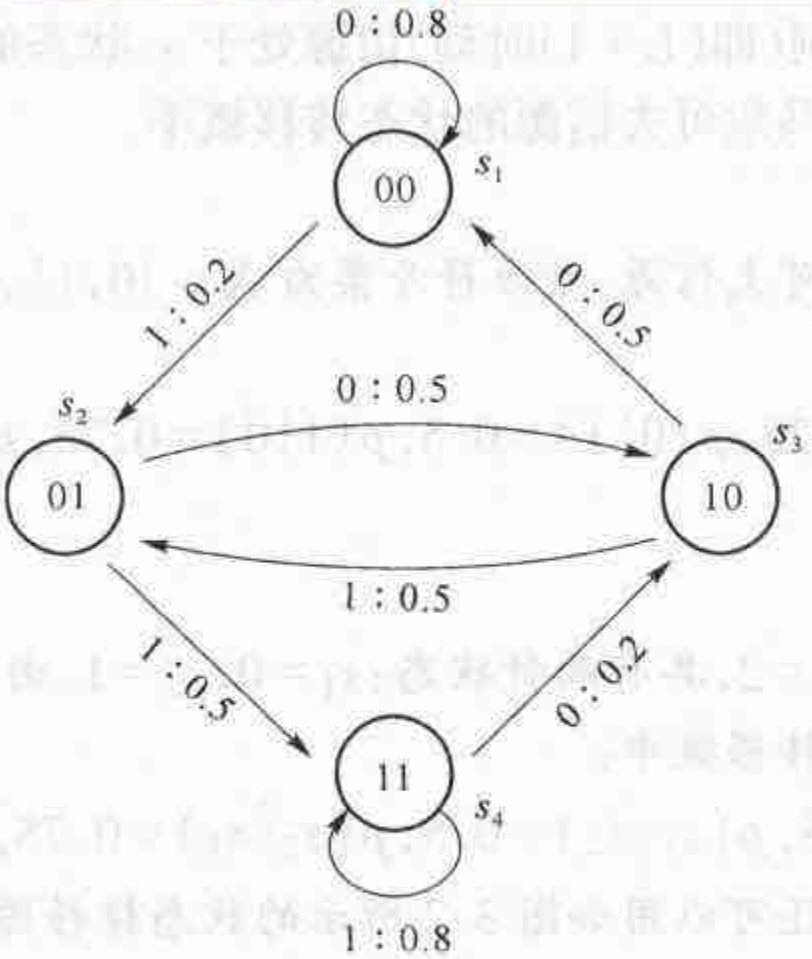
$$p(s_1|s_1)=0.25, p(s_1|s_2)=0.5, p(s_2|s_1)=0.75, p(s_2|s_2)=0.5$$

$$P=\begin{Bmatrix} & s_1=0 & s_2=1 \\ s_1=0 & \begin{bmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{bmatrix} \\ s_2=1 & \end{Bmatrix}$$

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 $p(0|00)=p(1|11)=0.8$，$p(0|01)=p(0|10)=p(1|01)=p(1|10)=0.5$，$p(1|00)=p(0|11)=0.2$.

求状态转移概率矩阵.

这里 $q=2, m=2$，故共有 $q^m=4$ 个可能的状态：$s_1=00, s_2=01, s_3=10, s_4=11$. 但由于信源只可能发出 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.



得状态转移概率：

$$p(s_1|s_1)=p(s_4|s_4)=0.8$$

$$p(s_2|s_1)=p(s_3|s_4)=0.2$$

$$p(s_3|s_2)=p(s_1|s_3)=p(s_4|s_2)=p(s_2|s_3)=0.5$$

信源的状态转移概率矩阵为

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

3 马尔科夫信源的计算条件—齐次遍历性

平稳性

当转移概率$P_{ij}(m, m+n)$只与i, j及时间间距n

有关时, 称转移概率具有平稳性.

同时也称此链是齐次的或时齐的.

马尔可夫信源的编历性

定义 设齐次马氏链的状态空间为I,若对于所有的$a_i, a_j \in I$,转移概率$P_{ij}(n)$存在极限

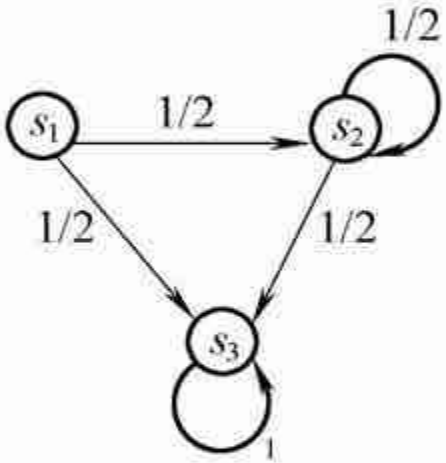
$$\lim_{n\to\infty} P_{ij}(n) = W_j \qquad (\text{不依赖于} i)$$

$$\text{并且} W_j = \sum_i W_i P_{ij}, \quad \sum_j W_j = 1, \quad Wj \geq 0$$

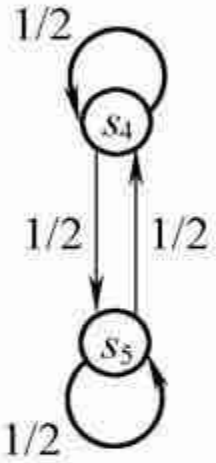
则称此链具有遍历性.

$W = (w_1, w_2, \cdots)$为链的极限分布.

编历性的充分必要条件-平稳分布及不可约



可约马氏链



非不可约马氏链

4 M阶马尔可夫信源熵率的计算

下面计算遍历的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以视作平稳信源来处理，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，所以极限熵H_∞等于条件熵 H_{m+1}。

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

所以：

$$H_{m+1} = H(X_{m+1} \mid X_1 X_2 \cdots X_m) = E\left[p(x_{i_{m+1}} \mid x_{i_1} x_{i_2} \cdots x_{i_m})\right]$$

$$= E\left[p(x_{i_{m+1}} \mid s_i)\right] = -\sum_{i=1}^{q^m} \sum_{i_{m+1}=1}^{q} p(s_i) p(x_{i_{m+1}} \mid s_i) \log p(x_{i_{m+1}} \mid s_i)$$

$$= \sum_i p(s_i) H(X \mid s_i) == -\sum_i \sum_j p(s_i) p(s_j \mid s_i) \log p(s_j \mid s_i)$$

其中，$p(s_i)$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；$p(s_j \mid s_i)$表示下一步状态转移概率.

马尔可夫信源序列熵的计算步骤

1、获得信源的状态转移概率矩阵

(*判断 遍历性--不可约、非周期)

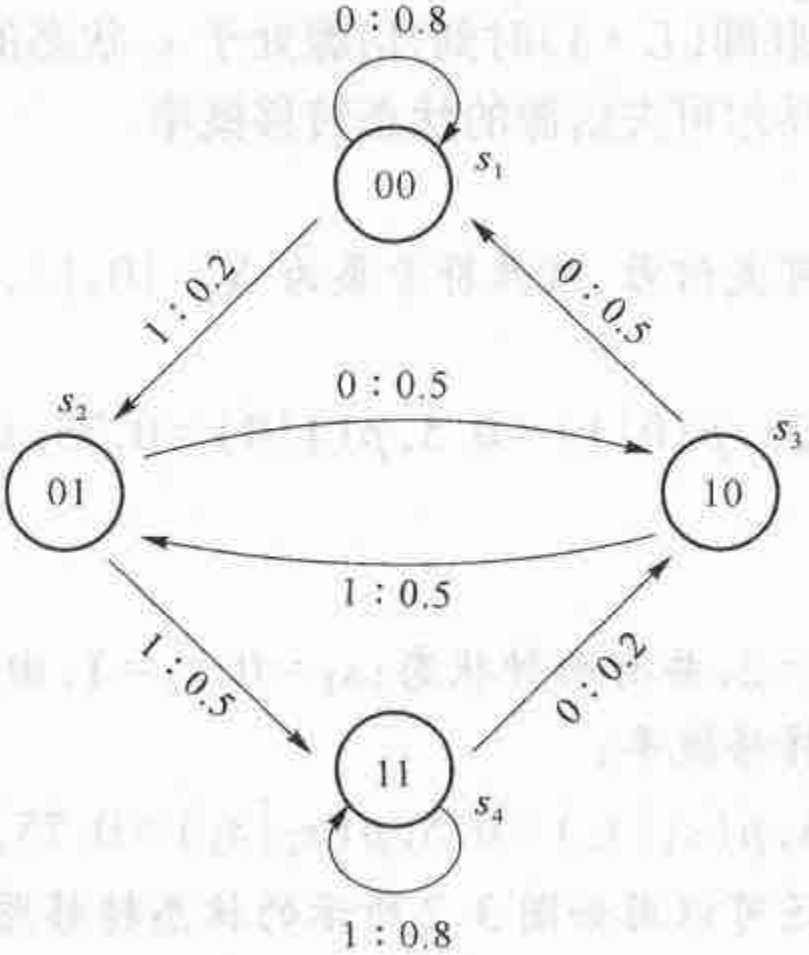
2、利用 $W_j = \sum_i W_i P_{ij}$, $\sum_j W_j = 1$, 求解 W_j

3、利用公式 $H_{m+1} = \sum_i p(s_i) H(X_i / s_i)$ 求 H_{m+1}

4、$H_\infty(X) = \lim_{L\to\infty} H(X_L / X_1 X_2 \cdots X_{L-1})$

$= H(X_{m+1} / X_1 X_2 \cdots X_m) = H_{m+1}(X)$

求图中的二阶马尔可夫信源的极限熵.



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

图中的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 $W_1=p(s_1)$，$W_2=p(s_2)$，$W_3=p(s_3)$，$W_4=p(s_4)$，根据马尔可夫链遍历的充分条件：$\boldsymbol{WP}=\boldsymbol{W}$，得

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

并且满足 $W_1+W_2+W_3+W_4=1$，因此可解得

$$W_1=p(s_1)=5/14, \quad W_2=p(s_2)=1/7,$$
$$W_3=p(s_3)=1/7, \quad W_4=p(s_4)=5/14$$

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

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

五 信源的相关性和剩余度—冗余度

1 信源的相关性和剩余度

2 冗余度

1 信源的相关性和剩余度

根据定理3.1可得 $\log q = H_0 \geq H_1 \geq H_2 \geq \cdots \geq H_{m+1} \geq \cdots \geq H_\infty$

由此看出，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实际熵减小。信源输出符号间统计约束关系越长，信源的实际熵越小。当信源输出符号间彼此不存在依赖关系且为等概率分布时，信源的实际熵等于最大熵 H_0。

定义3.3 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**：

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

信源剩余度为：$\gamma = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log q}$

2 冗余度

- 冗余度

定义:它表示给定信源在实际发出消息时所包含的多余信息,也称余度或剩余度。

$$\gamma = 1 - \eta = 1 - \frac{H_\infty(X)}{H_m(X)}$$

- 冗余度的来源

1. 信源符号间的相关性;
2. 信源符号分布的不均匀性，当等概率分布时信源熵最大

- 信息效率

$$\eta = \frac{H_\infty(X)}{H_m(X)}$$

信源冗余度及信息变差

信源的相关性和剩余度

信源的**剩余度**来自两个方面，一是信源符号间的**相关性**，相关程度越大，符号间的依赖关系越长，信源的实际熵越小，另一方面是信源符号分布的**不均匀性**使信源的实际熵越小。

为了更经济有效的传送信息，需要尽量压缩信源的剩余度，**压缩剩余度的方法就是尽量减小符号间的相关性，并且尽可能的使信源符号等概率分布。**

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。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。

信源编码是减少或消除信源的剩余度以提高信息的传输效率，而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。

3.4 信源的相关性和剩余度

根据定理3.1可得 $\log q = H_0 \geq H_1 \geq H_2 \geq \cdots \geq H_{m+1} \geq \cdots \geq H_\infty$

由此看出，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实际熵减小。信源输出符号间统计约束关系越长，信源的实际熵越小。当信源输出符号间彼此不存在依赖关系且为等概率分布时，信源的实际熵等于最大熵 H_0。

定义3.3 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**：

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

信源剩余度为：$\gamma = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log q}$

- 冗余度的来源
  - 信源符号间的相关性;
  - 信源符号分布的不均匀性，当等概率分布时信源熵最大

第3章　信源及信源熵

1. 信源的分类及数学模型
2. 离散单符号信源
3. 离散多符号信源
4. 马尔卡夫信源
5. 冗余度