信息论基础

李　莹
liying2009@ecust.edu.cn

第五章：无失真信源编码

一、信源编码的相关概念

二、定长码及定长信源编码定理

三、变长码及变长信源编码定理

四、变长码的编码方法

1. 信源编码概述

- **信源编码的作用：**
  - **使信源适合于信道的传输，用信道能传输的符号来代表信源发出的消息；**
  - **在不失真或允许一定失真的条件下，用尽可能少的符号来传递信源消息，提高信息传输率。**

- **以提高通信有效性为目的。通常通过压缩信源的冗余度来实现。采用的一般方法是压缩每个信源符号的平均码长。**

- 信源编码理论是信息论的一个重要分支，其理论基础是信源编码的两个定理：
  - 无失真信源编码定理
  - 限失真信源编码定理

- 本章主要介绍无失真信源编码，它实质上是一种统计匹配编码，根据信源的不同概率分布而选用与之相匹配的码。

- 信源的统计剩余度主要决定于以下两个因素：

  1）无记忆信源中，符号概率分布的非均匀性；

  2）有记忆信源中，符号间的相关性及符号概率分布的非均匀性。

- 怎样压缩信源的冗余度？

  1) 去除码符号间的相关性。

  2) 使码符号等概分布。

2. 信源编码器模型

● 信源编码：将信源符号序列按一定的数学规律映射成码符号序列的过程。

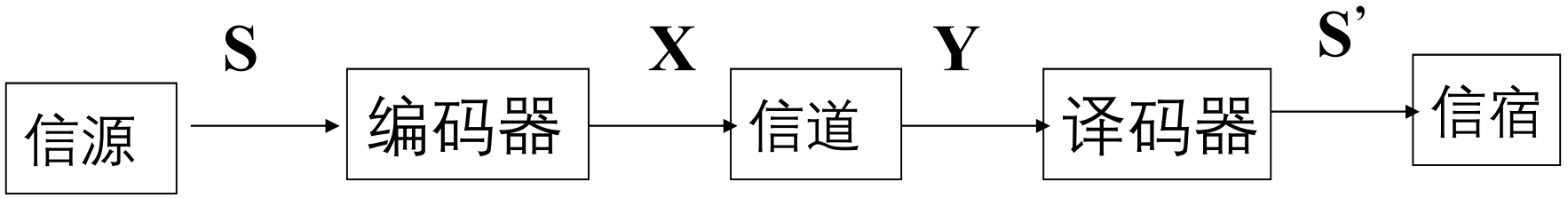
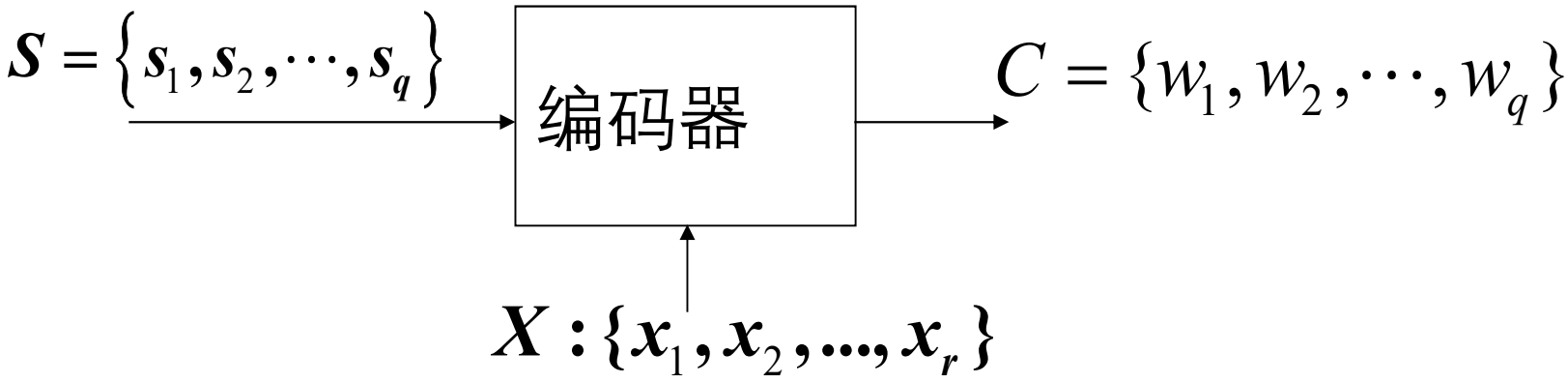


图1 信源编码器模型



码字 $w_i = x_{i_1} x_{i_2} \cdots x_{i_{l_i}}$

● 将信源符号集中的符号s_i（或者长为N的信源符号序列）映射成由码符号x_i组成的长度为l_i的一一对应的码符号序列w_i。

例：5.1

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ p(s_1) & p(s_2) & p(s_3) & p(s_4) \end{bmatrix}$$

信源符号s_i	$p(s_i)$	码1	码2
s_1	$p(s_1)=1/2$	00	0
s_2	$p(s_2)=1/4$	01	01
s_3	$p(s_3)=1/8$	10	001
s_4	$p(s_4)=1/8$	11	111

3. N次扩展码

$$S=\{s_1,s_2,\cdots,s_q\} \longleftrightarrow C=\{w_1,w_2,\cdots,w_q\}$$

$$s_i \longleftrightarrow w_i$$

$$S^N=\{\mathbf{s}_1,\mathbf{s}_2,\cdots,\mathbf{s}_{q^N}\} \longleftrightarrow C^N=\{w_1,w_2,\cdots,w_{q^N}\}$$

$$\mathbf{s}_j=s_{j_1}s_{j_2}\cdots s_{j_N} \longleftrightarrow \mathbf{w}_j=w_{j_1}w_{j_2}\cdots w_{j_N}$$

$$j=1,2,\cdots,q^N$$

$$j_1,j_2,\cdots,j_N=1,2,\cdots,q$$

二次扩展信源符号 $\mathbf{s}_j(j=1,2,...,16)$	二次扩展码码字 $\mathbf{w}_j(j=1,2,...,16)$
$\mathbf{s}_1 = s_1 s_1$	$\mathbf{w}_1 = w_1 w_1 = 00$
$\mathbf{s}_2 = s_1 s_2$	$\mathbf{w}_2 = w_1 w_2 = 001$
$\mathbf{s}_3 = s_1 s_3$	$\mathbf{w}_3 = w_1 w_3 = 0001$
$\vdots$	$\vdots$
$\mathbf{s}_{16} = s_4 s_4$	$\mathbf{w}_{16} = w_4 w_4 = 111111$

4. 关于编码的一些术语

- 编码器输出的码符号序列 w_i 称为**码字**；长度 l_i 称为码字长度，简称**码长**；全体码字的集合 C 称为**码**。
- 若码符号集合为 $X=\{0,1\}$，则所得的码字都是二元序列，称为**二元码**。
- 将信源符号集中的每个信源符号 s_i 固定的映射成某一个码字 w_i，这样的码称为分组码。
- 若一个码中所有码字的码长都相等，则称为定长码；否则为变长码。

5. 奇异性

若一个码中所有码字互不相同，则称为**非奇异码**；否则为**奇异码**。

信源符号s_i	码1	码2
s_1	0	0
s_2	11	10
s_3	00	00
s_4	11	01

6. 唯一可译性

● 若任意一串有限长的码符号序列只能被唯一地译为对应的信源符号序列，则称此码为**唯一可译码**。

信源符号s_i	码1	码2	码3
s_1	0	0	0
s_2	11	10	10
s_3	00	00	110
s_4	11	01	111

- 唯一可译码应当满足的条件

1) $w_i(i=1,2,...,q) \leftrightarrow s_i(i=1,2,...,q)$

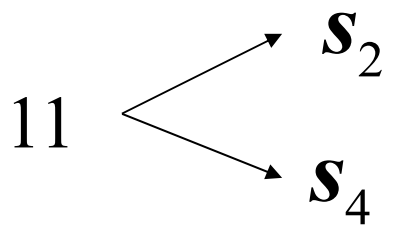
码字与信源符号一一对应

2) 不同的信源符号序列对应不同的码字序列

例1：　1)　奇异码

s_1	0
s_2	11
s_3	00
s_4	11

译码



奇异码一定不是唯一可译码

2) 非奇异码

s_1	0
s_2	10
s_3	00
s_4	01

0 1 0 0 0 0 1 0 译码 $\longrightarrow$ $s_1 s_2 s_3 s_4 s_1$

0 1 0 0 0 0 1 0 译码 $\longrightarrow$ $s_4 s_3 s_3 s_2$

3) 等长码 非奇异码 唯一可译码

s_1	00
s_2	01
s_3	10
s_4	11

0 0 0 1 1 0 1 1 译码 $\longrightarrow$ $s_1 s_2 s_3 s_4$

4) 唯一可译码

s_1 1

s_2 10

s_3 100

s_4 1000

1 1 0 1 0 0 1 0 0 0

1 0 { 0 / 1 } s_2 / s_3 ?

∴ 为非即时码

5) 唯一可译码

s_1 1

s_2 01

s_3 001

s_4 0001

1 0 1 0 0 1 0 0 0 1

0 1 → 即时

s_2

∴ 为即时码

任何一个码字不是其它码字的前缀

7. 即时码

- 若某个唯一可译码在接收到一个完整的码字时无需参考后续的码符号就能立即译码，则称此码为**即时码**。

- 问题：

1) 判断下面的码是否即时码？

0

10

110

111

2) 等长码是否即时码？

- 唯一可译码成为即时码的充要条件：

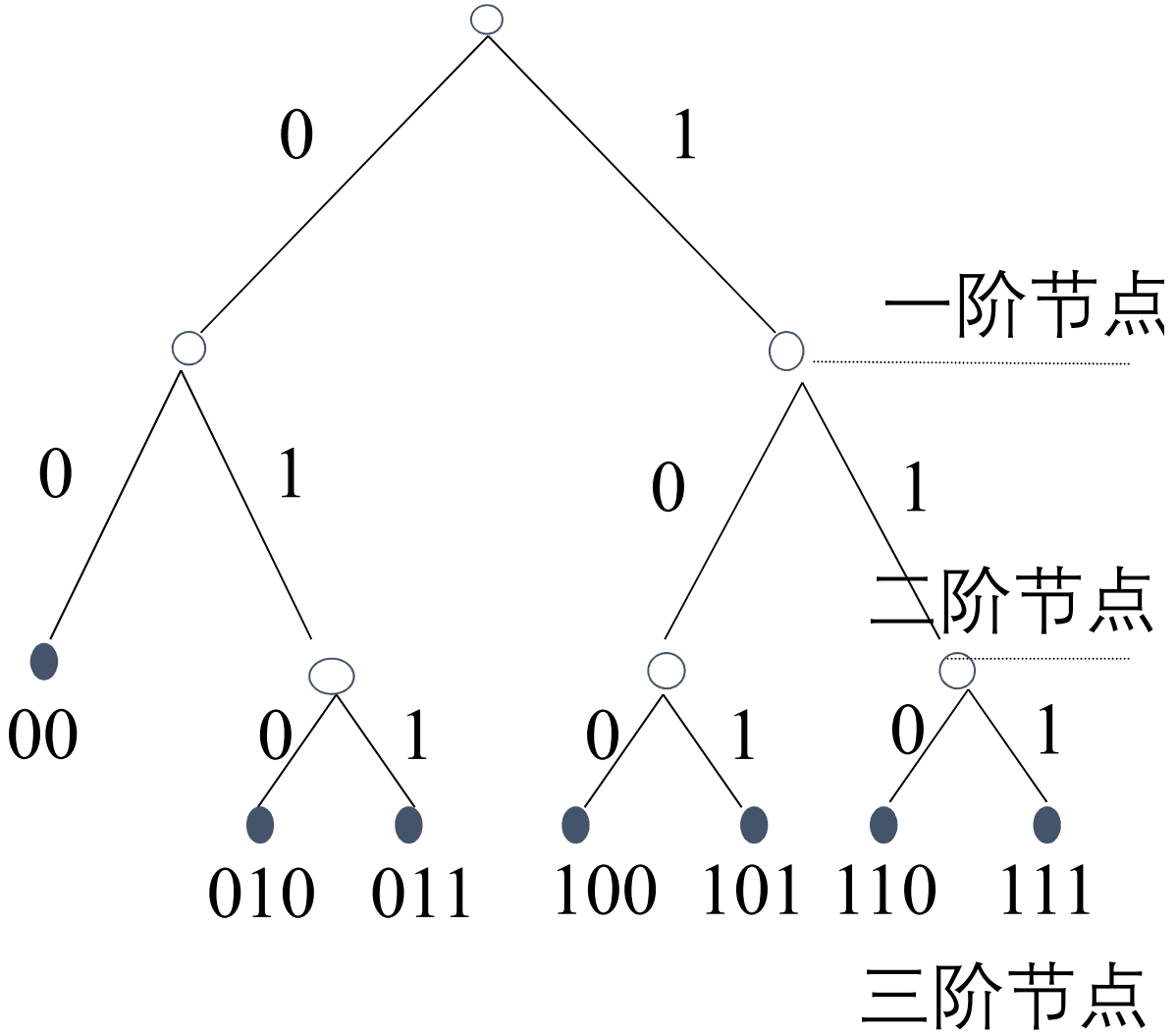
定理5.1 一个唯一可译码成为即时码的充要条件是其中任何一个码字都不是其他码字的前缀。

信源	概率p_i	编码I	编码II	编码III	编码IV	编码V
s_1	1/2	00	0	0	0	0
s_2	1/4	01	0	1	10	01
s_3	1/8	10	1	00	110	011
s_4	1/8	11	10	11	111	0111

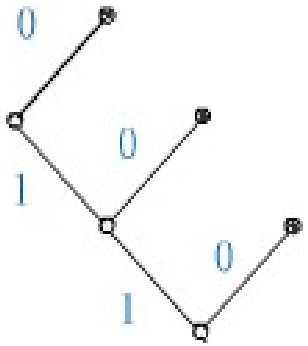
消息	A	B	C	D	E	F
s_1	000	010	0	1	00	00
s_2	010	011	10	00	01	01
s_3	011	000	1110	0110	10	100
s_4	100	011	1111	0111	110	101
s_5	101	101	1100	0100	1110	110
s_6	111	100	1101	0101	1111	111

8. 即时码的构造方法

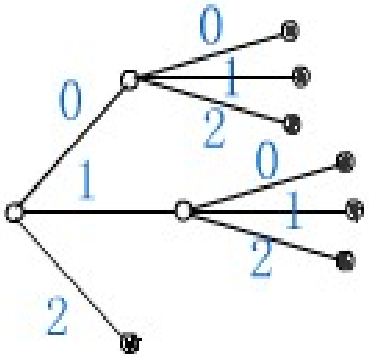
- 用树图法可以方便地构造即时码。树中每个中间节点都伸出1至r个树枝，将所有的码字都安排在终端节点上就可以得到即时码。



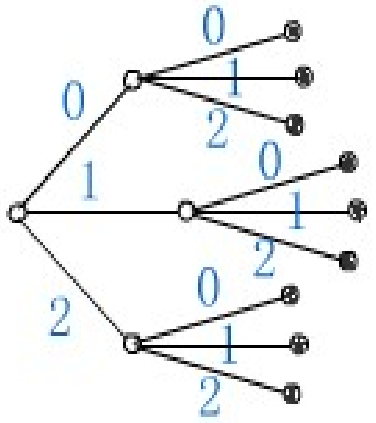
- 用树图法可以方便地构造即时码。树中每个中间节点都伸出1至*r*个树枝，将所有的码字都安排在终端节点上就可以得到即时码。
- 每个中间节点都正好有*r*个分枝的树称为**整树（满树）**。
- 所有终端节点的阶数都相等的树为**完全树**



二进制非满树



三进制满树



三进制完全树

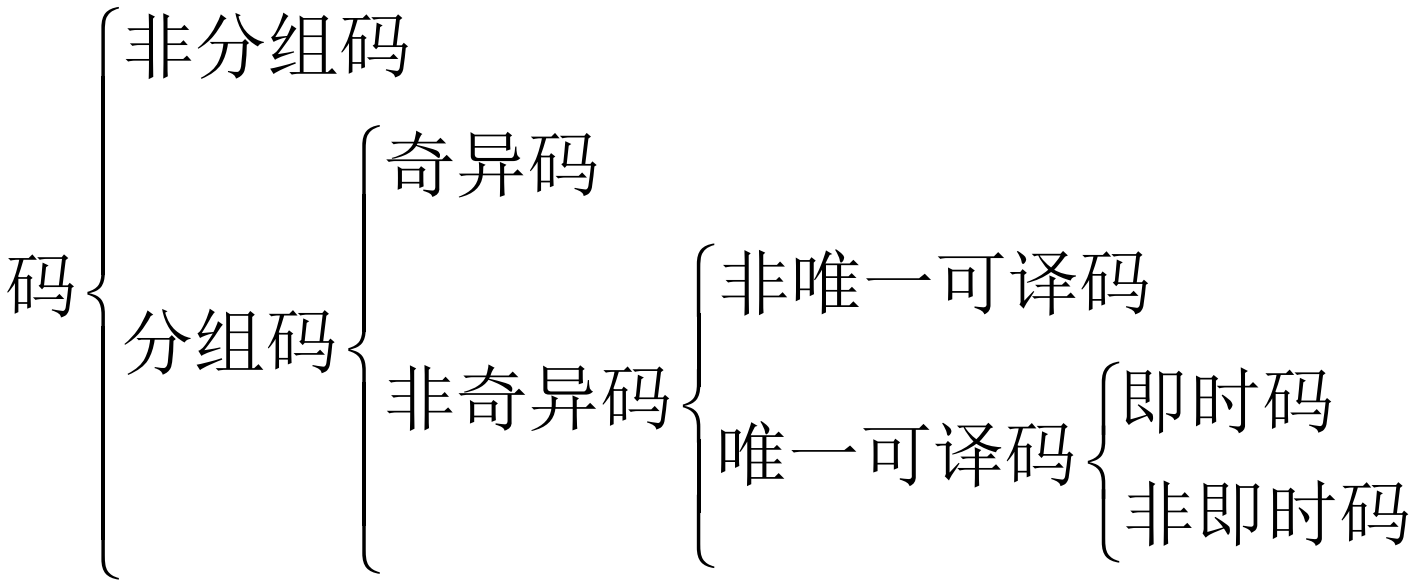


图2 各类码之间的关系

第五章：无失真信源编码

一、信源编码的相关概念

二、定长码及定长信源编码定理

三、变长码及变长信源编码定理

四、变长码的编码方法

1. 唯一可译定长码存在的条件

- 对于定长码，非奇异码一定是唯一可译码。
- 所谓非奇异码，即信源符号集中的每一个信源符号 s_i 与码中的某一个码字 w_i 一一对应。
- 设信源符号集中共有 q 个符号，$S=\{s_1,s_2,\cdots,s_q\}$，码符号集中共有 r 种码元，$X:\{x_1,x_2,...,x_r\}$，定长码码长为 l，则要满足非奇异性必然有

$$q \le r^l$$

该条件是必要条件，而不是充分条件。

- 如果对N次扩展信源 S^N 进行定长编码，要满足非奇异性，须满足条件 $q^N \le r^l$

其中 $S^N = \left\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{q^N}\right\}$，　$X : \{x_1, x_2, ..., x_r\}$

当r=2 时，$\dfrac{l}{N} \ge \log q$

当N=1时，

$$l \ge \log q$$

例2：英文字母表中，每一字母用定长编码转换成二进制表示，码字的最短长度是多少？

解：

信源符号数 $q = 26$

码符号数 $r = 2$

$$r^l \geq q \quad \Rightarrow \quad l \geq \frac{\log q}{\log r} = \frac{\log 26}{\log 2} = 4.7$$

$$\therefore \qquad l_{\min} = 5$$

2. 定长信源编码定理

- 定理5.3.1 设离散平稳无记忆信源的熵为$H(S)$, 若对N次扩展信源 S^N进行定长编码，则对于任意ε >0，只要满足

$$\frac{l}{N} \geq \frac{H(S)+\varepsilon}{\log r}$$

则当N足够大时，可实现几乎无失真编码，即译码错误概率 P_E 为任意小；反之，则不可能实现无失真编码，

如果

$$\frac{l}{N} \leq \frac{H(S)-2\varepsilon}{\log r}$$

当N足够大时，译码错误概率 P_E 为1。

- 定长编码定理同样也适用于离散平稳有记忆信源，差别是要将信源熵 $H(S)$改为极限熵 H_∞ 。

当r=2时 $$\frac{l}{N} \geq H(S) + \varepsilon$$

可以验证，对定长信源编码来说，要想实现无失真信源编码，N常常要取非常大的值。这在实际应用中很难实现。

比特/信源符号

- 定义： $R' \overset{def}{=} \frac{l}{N}\log r$ 为**编码信息率**

- 定义： $\eta = \frac{H(S)}{R'} = \frac{H(S)}{\frac{l}{N}\log r}$ 为**编码效率**。

$$\frac{\text{比特/信源符号}}{\text{比特/信源符号}}$$

- 根据**编码效率的**定义，最佳编码效率为：$\eta = \dfrac{H(S)}{H(S)+\varepsilon}$

- 在已知方差和信源熵的条件下，信源符号序列长度N与最佳编码效率η和允许编码错误概率δ 之间的关系为：

$$N \geq \frac{D\left[I(s_i)\right]}{\varepsilon^2 \delta} = \frac{D\left[I(s_i)\right]}{H^2(S)} \frac{\eta^2}{(1-\eta)^2 \delta}$$

例 5.2 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$

如果对信源符号采用定长二元编码，要求编码效率 η=90%，允许错误概率 $\delta \le 10^{-6}$，求所需信源序列长度 N。

$$N \ge 9.8 \times 10^7$$

例5.3 设离散无记忆信源 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$，要求 $\eta = 0.96, \delta \le 10^{-5}$ 求N。

$$N \ge 4.13 \times 10^7$$