《信息论基础》

第3章 无失真离散信源编码

吉小鹏

E-mail: 003163@nuist.edu.cn

南京信息工程大学 电子与信息工程学院 尚贤楼209




提纲

3.1 基本概念

3.2 离散无失真信源编码定理

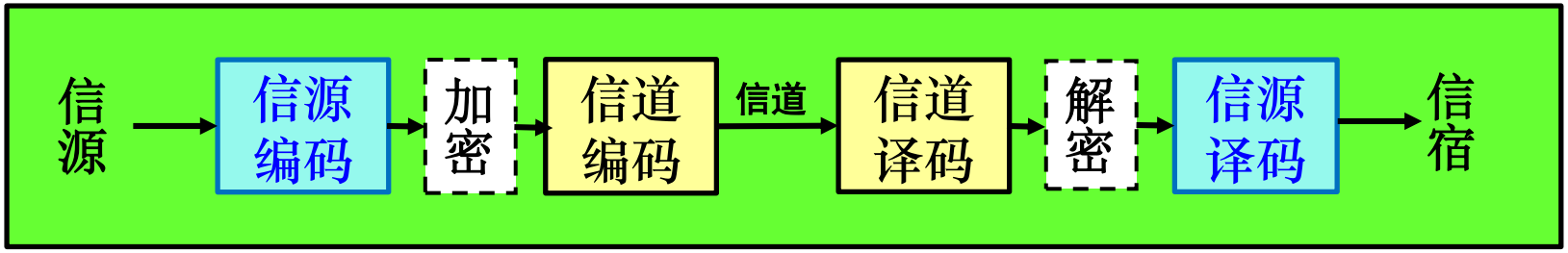
3.3 香农编码

3.4 费诺编码

3.5 霍夫曼编码

3.1 基本概念

- 编码



通信系统的性能指标

- 有效性 ⟹ 信源编码
- 可靠性 ⟹ 信道编码
- 安全性 ⟹ 安全编码（密码）
- 经济性

3.1 基本概念

编码

- 信源编码

在不失真或允许一定失真条件下，如何用尽可能少的符号来传送信源信息，以便提高信息传输率。

- 信道编码

在信道受干扰的情况下如何增加信号的抗干扰能力，同时又使得信息传输率最大。

- **信源编码理论：**

➢ 无失真信源编码定理　（离散信源或数字信号编码）

➢ 限失真信源编码定理　（连续信源或模拟信号编码）

- **信道编码**

3.1 基本概念

- **编码**

问题：为什么可以用较少符号来传输信息？数据为什么可以压缩？

- 由于信源存在冗余度,即存在一些不必要传送的信息,因此信源也就存在进一步压缩其信息率的可能性。
- 信源冗余度越大,其进一步压缩的潜力越大。这是信源编码与数据压缩的前提与理论基础。
- 例：英文字母：

等概率　　$H_0 = \log 27 = 4.76$比特/符号

不等概率　　$H_1 = 4.03$比特/符号

考虑相关性　　$H_2 = 3.32$比特/符号

极限熵　　$H_\infty = 1.4$比特/符号

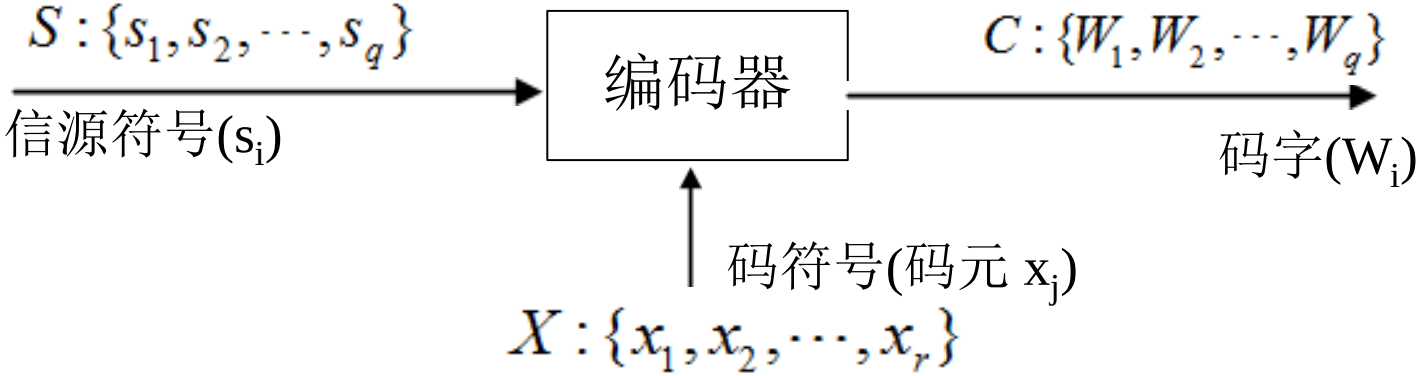
- 冗余度　　$\gamma = (4.76 - 1.4)/4.76 = 0.71$

英语文章有71%是由语言结构定好的，只有29%是自由选择

- **编码**

信源编码:

将信源输出符号，经信源编码器后变换成另外的压缩符号，然后将压缩后信息经信道传送给信宿。



编码器是将信源符号集 S 中的符号 s_i (或者长度为 N 的信源符号序列 α_i) 变换成由 $x_j(j=1,2,\cdots,r)$ 组成的长度为 l_i 的一一对应的序列，即

$$s_i(i=1,\cdots,q)\leftrightarrow W_i=(x_{i_1}x_{i_2}\cdots x_{i_{l_i}}) \qquad \alpha_i=(s_{i_1}s_{i_2}\cdots s_{i_N})\leftrightarrow W_i=(x_{i_1}x_{i_2}\cdots x_{i_{l_i}})$$

3.1 基本概念

- **码的定义**

二元码:

若码符号集为 $X = \{0, 1\}$，所得码字都是一些二元序列，则称二元码。

等长码:

若一组码中所有码字的码长都相同，则称为等长码。

变长码:

若一组码中所有码字的码长不全相同，则称为变长码。

非奇异码:

若一组码中所有码字都不相同，即所有信源符号映射到不同的码符号序列，则称为非奇异码。 $s_i \neq s_j \Rightarrow W_i \neq W_j \quad s_i, s_j \in S \quad W_i, W_j \in C$

奇异码:

若一组码中有相同的码字，则称为奇异码。$s_i \neq s_j \Rightarrow W_i = W_j \quad \exists s_i, s_j \in S$

3.1 基本概念

- **码的定义**

例如，信源符号$S=\{a_1,a_2,a_3,a_4\}$,对应不同码字如表

信源符号	信源符号 出现概率	码 表				
		码0	码1	码2	码3	码4
a_1	$p(a_1)=1/2$	00	0	0	1	1
a_2	$p(a_2)=1/4$	01	11	10	10	01
a_3	$p(a_3)=1/8$	10	00	00	100	001
a_4	$p(a_4)=1/8$	11	11	01	1000	0001

- 等长码：码0
- 变长码：码1，2，3， 4
- 非奇异码：码0, 2, 3, 4
- 奇异码：码1

3.1 基本概念

- **码的定义**

同价码:

若码符号集X中每个码符号x_i所占的传输时间都相同，则所得码C为同价码。

码的 *N* 次扩展码:

假定某码C，将信源S中的符号 s_i 一一变换为码C中的码字 W_i，则码C的 *N* 次扩展码是所有 *N* 个码字组成的码字序列的结合。

唯一可译码:

若码的任意一串有限长的码符号序列只能被唯一地译成所对应的信源符号序列，则此码称为唯一可译码。

要求：① 不同信源符号变换成不同的码字；② 码的任意有限长 *N* 次扩展码都是非奇异码。

3.1 基本概念

- 码的定义

唯一可译码

非即时码：

如果接收端收到一个完整的码字后不能立即译码，还需等下一个码字开始接收后才能判断是否可以译码。

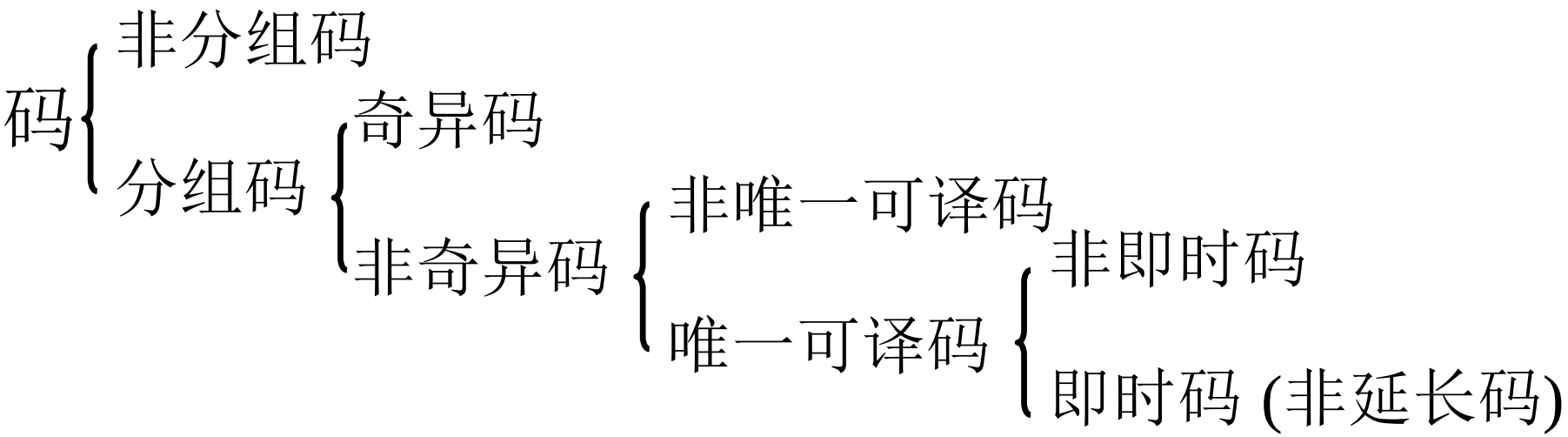
即时码：

在译码时无需参考后续的码符号就能立即作出判断，译成对应的信源符号。

任意一个码字都不是其它码字的前缀部分。

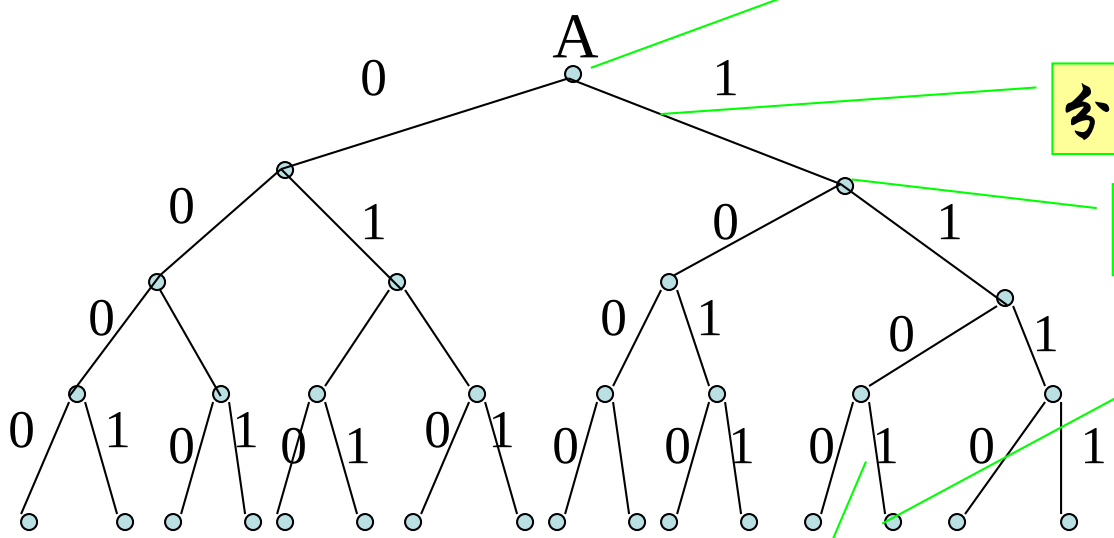
3.1 基本概念

- 码的定义

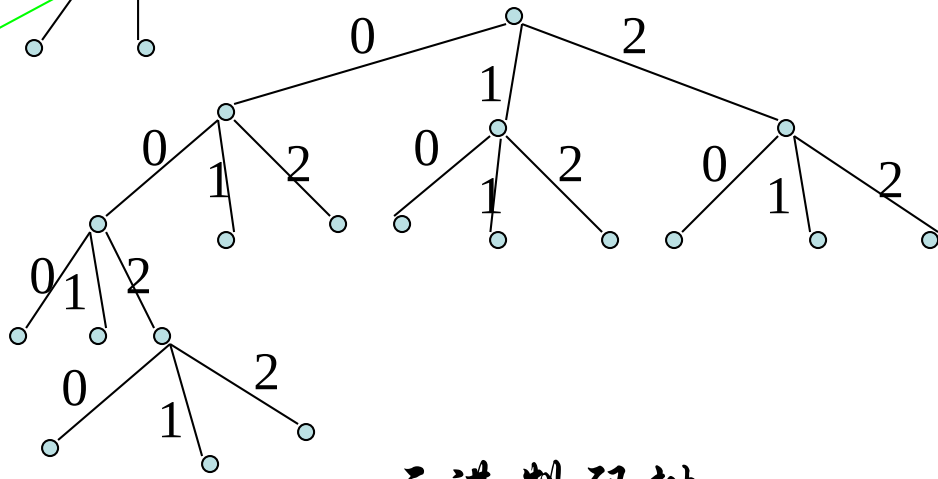


3.1 基本概念

- 码树



二进制码树



三进制码树

树根——码字的起点

分成r个树枝——码的进制数

中间节点——码字的一部分

终端节点——码字1101

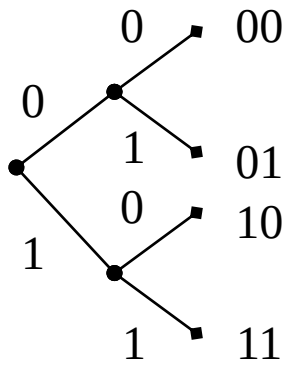
节数——码长

码树

如果有q个信源符号,那么在码树上就要选择q个终端节点,用相应的 *r* 元基本符号表示这些码字。

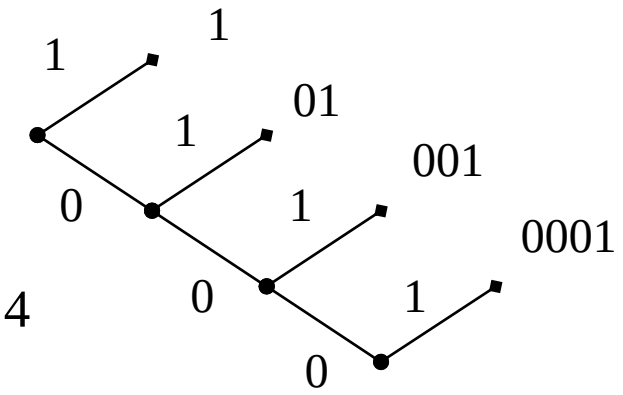
码0
00
01
10
11

码0

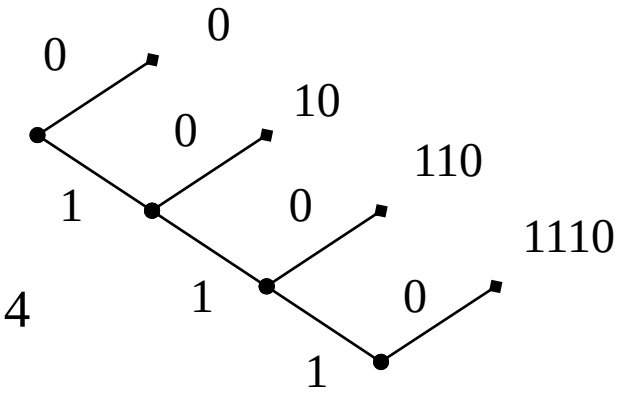


码4
1
01
001
0001

码4



码4

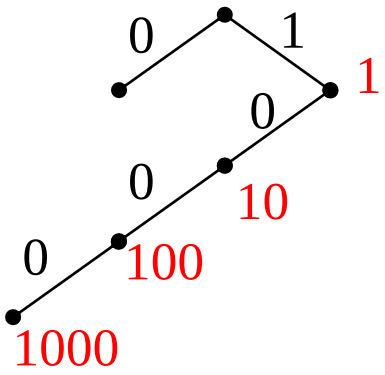


- 任一即时码都可用树图法来表示。
- 当码字长度给定,即时码不是唯一的。

码树

- 码3对应的树如下图:

码3
1
10
100
1000



- 该码树从根到终端节点所经路径上每一个中间节点皆为码字,因此不满足前缀条件。
- 虽然码3不是即时码，但它是唯一可译码。

3.1 基本概念

- **码树**
  - 满树：
    - 每个节点上都有 r 个分枝的树——等长码
  - 非满树：
    - 变长码
  - 用树的概念可导出唯一可译码存在的充分必要条件

定理(Kraft不等式)：对于 r 元字母表上的即时码，码字长度K_i 必定满足不等式 $\sum_{i=1}^{q} r^{-K_i} \le 1$，

r 是进制数，q 是信源符号数

反之，若给定满足以上不等式的一组码字长度，则存在一个相应的即时码，其码字长度就是给定长度。

码字是无限可数的情形，扩展的Kraft不等式 $\sum_{i=1}^{\infty} r^{-K_i} \le 1$ 成立。

码树

- 例：设二进制码树中X=(a_1 , a_2 , a_3 , a_4), K_1=1, K_2=2, K_3=2, K_4=3,

应用Kraft不等式，得：

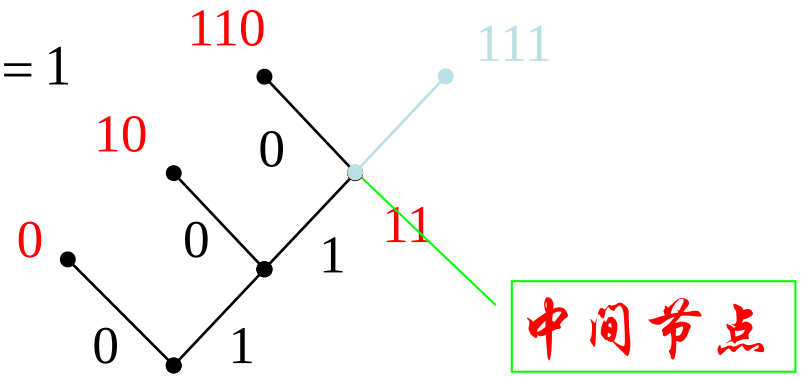
$$\sum_{i=1}^{4} 2^{-K_i} = 2^{-1} + 2^{-2} + 2^{-2} + 2^{-3} = \frac{9}{8} > 1$$

不存在满足这种K_i的唯一可译码

- 如果将各码字长度改成K_1=1, K_2=2, K_3=3, K_4=3, 则

$$\sum_{i=1}^{4} 2^{-K_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$$

这样的码字就存在唯一可译码



码树

- 必须注意：
  - Kraft不等式只是用来说明唯一可译码是否存在，并不能作为唯一可译码的判据。
  - 如码字{0，10，010，111}虽然满足Kraft不等式,但它不是唯一可译码。

提纲

3.1 基本概念

3.2 离散无失真信源编码定理

3.3 香农编码

3.4 费诺编码

3.5 霍夫曼编码

3.2 离散无失真信源编码定理

- **离散无失真信源编码**
  - 信源编码器输入的消息序列:

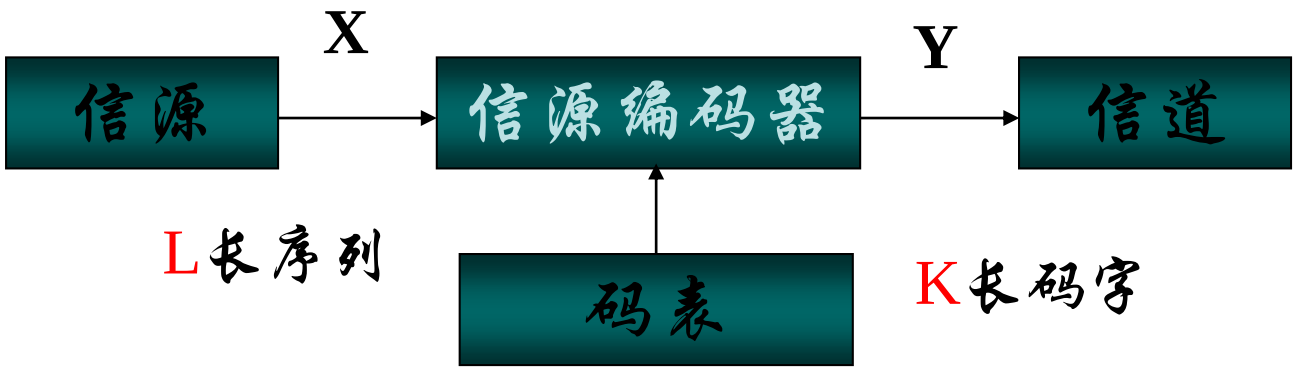
$$\mathbf{X}=(X_1\ X_2 \ldots X_l \ldots X_L),\quad \text{其中}\quad X_l \in \{a_1, \ldots, a_q\},$$

输入的消息总共有 q^L 种可能的组合

  - 输出的码字为:

$$\mathbf{Y}=(Y_1\ Y_2 \ldots Y_k \ldots Y_K),\ \text{其中}\quad Y_k \in \{b_1, \ldots, b_m\}$$

输出的码字总共有 m^K 种可能的组合。



3.2 离散无失真信源编码定理

离散无失真信源编码

- 实现无失真的信源编码，要求：
  - 信源符号 $X_1\ X_2 \ldots X_l \ldots X_L$
  - 码字 $Y_1\ Y_2 \ldots Y_k \ldots\ Y_K$

符号和码字能 **一一对应**

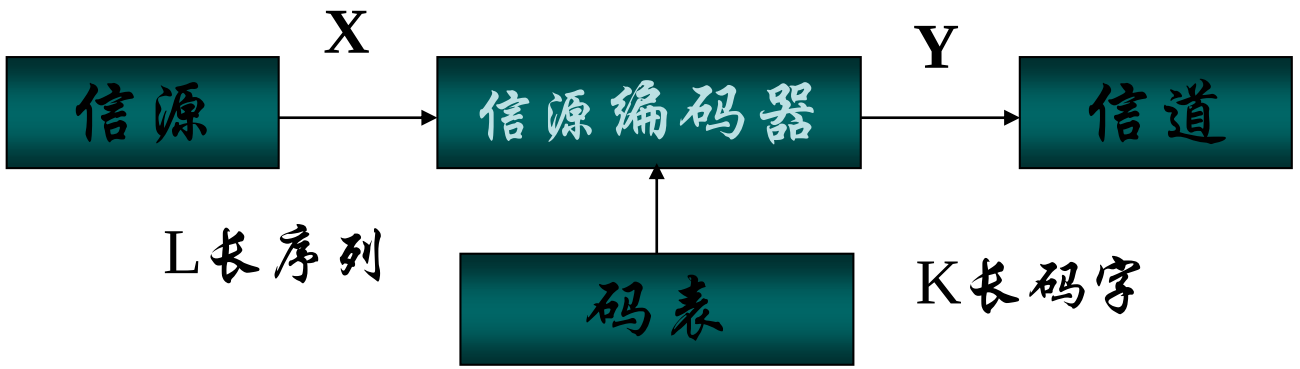
- 能够无失真或无差错地从 Y 恢复 X，也就是能正确地进行反变换或译码；
- 传送 Y 时所需要的信息率最小

信息率最小就是找到一种编码方式使

$$\bar{K} = \frac{K}{L}\log m = \frac{1}{L}\log M \quad \text{最小} \quad M = m^K$$

3.2 离散无失真信源编码定理

- 定长编码定理
  - 在定长编码中，码长 K 是定值。
  - 我们的目的是寻找最小 K 值。
  - 编码器输入$\mathbf{X}=(X_1\ X_2 \ldots X_l \ldots X_L)$，$X_l \in \{a_1, \ldots a_q\}$，
    输入的消息总共有 q^L 种可能的组合
  - 输出的码字$\mathbf{Y}=(Y_1\ Y_2 \ldots Y_k \ldots Y_K)$，$Y_k \in \{b_1, \ldots b_m\}$
    输出的码字总共有 m^K 种可能的组合。



3.2 离散无失真信源编码定理

- 定长编码定理
  - 若对信源进行 **定长** 编码，必须满足：

$$q^L \le m^K \quad 或 \quad \frac{K}{L} \ge \frac{\log q}{\log m}$$

  - 只有当 K 长的码符号序列数 m^K 大于或等于信源的符号数 q^L 时，才可能存在定长非奇异码。
  - 例如英文电报有27个符号，q = 27， L = 1，m = 2 (二元编码)

$$K \ge L \frac{\log_2 q}{\log_2 m} = \log_2 27 \approx 5$$

每个英文电报符号至少要用**5**位二元符号编码；

实际英文电报符号信源，平均每个英文电报符号所提供的信息量约等于**1.4** bit。

3.2 离散无失真信源编码定理

- 渐近等分割性和 ε 典型序列

定理1（渐近等分割性定理）：

若随机序列 $X_1 \ldots X_i \ldots X_N$ 中 X_i 相互统计独立并且服从同一概率分布 P(x)，又 $\alpha_i = (x_{i_1} x_{i_2} \cdots x_{i_N}) \in X_1 X_2 \cdots X_N$，则 $-\frac{1}{N}\log P(\alpha_i) = -\frac{1}{N}\log P(x_{i_1} x_{i_2} \cdots x_{i_N})$ 依概率收敛于$H(X)$。 $\lim_{N\to\infty} P\{\left|\frac{I(\alpha_i)}{N} - H(X)\right| < \varepsilon\} = 1$

定理2：

对于任意小的正数$\varepsilon > 0, \delta > 0$，当 N 足够大时，则

$$G_{\varepsilon N} = \{\alpha_i : \left|\frac{I(\alpha_i)}{N} - H(X)\right| < \varepsilon\}$$

$$\overline{G_{\varepsilon N}} = \{\alpha_i : \left|\frac{I(\alpha_i)}{N} - H(X)\right| \geq \varepsilon\}$$

(1) $P(G_{\varepsilon N}) > 1-\delta, \quad P(\overline{G_{\varepsilon N}}) \leq \delta$

(2) 若 $\alpha_i \in G_{\varepsilon N}$，则 $2^{-N[H(X)+\varepsilon]} < P(\alpha_i) < 2^{-N[H(X)-\varepsilon]}$

(3) 设 $\|G_{\varepsilon N}\|$ 表示集合 $G_{\varepsilon N}$ 中包含的典型序列的个数，则有

$$(1-\delta)2^{N[H(X)-\varepsilon]} \leq \|G_{\varepsilon N}\| \leq 2^{N[H(X)+\varepsilon]}$$

3.2 离散无失真信源编码定理

- 定长编码定理

定长信源编码定理:

由 L 个符号组成的、每个符号的熵为 $H(X)$ 的**无记忆平稳信源**符号序列 $X_1 \ldots X_i \ldots X_L$, 可用 K 个符号 $Y_1 \ldots Y_k \ldots Y_K$ (每个符号有 m 种可能值) 进行定长编码。对任意 $\varepsilon > 0,\ \delta > 0$ ，只要

$$\frac{K}{L}\log m \geq H(X)+\varepsilon ,$$

则当 L 足够大时，必可使译码差错小于 δ，即可实现几乎无失真编码;

反之，当 $\frac{K}{L}\log m \leq H(X)-2\varepsilon$ 时，译码差错一定是有限值，即不可能实现无失真编码；而当 L 足够大时，译码几乎必定出错。

3.2 离散无失真信源编码定理

- **定长编码定理**

$$\frac{K}{L}\log m \geq H(X)+\varepsilon$$

(1) 定理中实现几乎无失真编码的条件可改写为

$$K\log m > LH(X)$$

其中：左边：K 长码字所能携带的最大信息，

右边：L 长信源序列平均携带的信息量。

定理表明：

只要码字所能携带的信息量大于信源序列输出的信息量，则可以使传输几乎无失真，当然条件是 L 足够大。

3.2 离散无失真信源编码定理

- **定长编码定理**

$$\frac{K}{L}\log m \geq H(X)+\varepsilon$$

(2) 定义编码后平均每个信源符号能载荷的最大信息量

$$R' = \frac{K}{L}\log m = \frac{1}{L}\log M$$

称为编码后**信源**的**信息传输率**。

定理表明：

只要 $R' > H(X)$，这种编码器一定可以做到几乎无失真，也就是收端的译码差错概率接近于零，条件是所取的符号数 L 足够大。

3.2 离散无失真信源编码定理

- **定长编码定理**

$$\frac{K}{L}\log m \geq H(X)+\varepsilon$$

对定长编码，若要实现几乎无失真编码，当允许错误概率小于 δ 时，则信源序列长度必须满足:

$$L \geq \frac{\sigma^2(X)}{\varepsilon^2\delta}$$

$$\sigma^2(X) = E\{[I(x_i) - H(X)]^2\}$$

– 信源序列的自信息方差

3.2 离散无失真信源编码定理

- **定长编码定理**

$$\frac{K}{L}\log m \geq H(X)+\varepsilon$$

为了衡量编码效果，定义编码效率

$$\eta = \frac{H(X)}{R'} = \frac{H(X)}{\frac{K}{L}\log m}.$$

由定长信源编码定理可知，最佳等长编码的效率为

$$\eta = \frac{H(X)}{H(X)+\varepsilon} \quad (\varepsilon > 0)$$

最佳等长编码的效率可接近于1。

3.2 离散无失真信源编码定理

- **定长编码定理** $L \geq \dfrac{\sigma^2(X)}{\varepsilon^2\delta}$ $\sigma^2(X) = E\{[I(x_i) - H(X)]^2\}$ $\eta = \dfrac{H(X)}{H(X)+\varepsilon}$

例：设离散无记忆信源概率空间 $= E[I^2(x_i)] - [H(X)]^2$

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

若取差错率 $\delta \leq 10^{-6}$，编码效率为90%，则 L 应满足?

- 信源熵： $H(X) = -\sum_i p(x_i)\log p(x_i) = 2.55bit/符号$
- 方差： $\sigma^2(X) = \sum_{i=1}^{8} p_i[-\log p_i - H(X)]^2 = 7.82bit^2$

$$\eta = \frac{H(X)}{H(X)+\varepsilon} = 0.9 \Rightarrow \varepsilon = 0.28 \qquad L \geq \frac{\sigma^2(X)}{\varepsilon^2\delta} = \frac{7.82}{0.28^2 \times 10^{-6}} = 9.8 \times 10^7$$

在差错率和编码效率要求并不十分苛刻的条件下，就需要 $L = 10^8$ 个信源符号进行联合编码，这显然是很难实现的。

3.2 离散无失真信源编码定理

- 变长编码定理
  - 在变长编码中，码长 K 是变化的。
  - 我们的目的是寻找最小 K 平均值。
  - 我们可根据信源各个符号的统计特性,如概率大的符号用短码，概率小的用较长的码，这样在大量信源符号编成码后平均每个信源符号所需的输出符号数就可以降低，从而提高编码效率。

3.2 离散无失真信源编码定理

- 单符号变长编码定理

Morse电报字符

A	· –	K	– · –	U	· · –	1	· – – – –
B	– · · ·	L	· – · ·	V	· · · –	2	· · – – –
C	– · – ·	M	– –	W	· – –	3	· · · – –
D	– · ·	N	– ·	X	– · · –	4	· · · · –
E	·	O	– – –	Y	– · – –	5	· · · · ·
F	· · – ·	P	· – – ·	Z	– – · ·	6	– · · · ·
G	– – ·	Q	– – · –	,	– – · · – –	7	– – · · ·
H	· · · ·	R	· – ·	.	· – · – ·	8	– – – · ·
I	· ·	S	· · ·		–	9	– – – – ·
J	· – – –	T	–			0	– – – – –

3.2 离散无失真信源编码定理

- **变长编码定理**

设信源为：$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_q \\ P(a_1) & P(a_2) & \cdots & P(a_q) \end{bmatrix}$

编码后的码字为 $W_1, W_2, \cdots, W_q$

其码长分别为 $l_1, l_2, \cdots, l_q$

因为对唯一可译码，信源符号和码字一一对应，所以

$$P(W_i) = P(a_i)$$

则该码的**平均码长**为 $\overline{L} = \sum_{i=1}^{q} P(a_i) \cdot l_i.$，单位：码符号/信源符号，是每个信源符号平均需要的码元数。平均每个码元携带的信息量，即编码后**信道**的信息传输率，又称**码率**，为 $R = H(Y) = \dfrac{H(X)}{\overline{L}}.$

最佳码（紧致码）：对于某一信源和某一码符号集来说，若有一唯一可译码，其平均码长小于所有其他唯一可译码的平均长度。

3.2 离散无失真信源编码定理

- 单符号变长编码定理

单符号变长编码定理：

若一离散无记忆信源的 X 符号熵为 $H(X)$，每个信源符号用 m 进制码元进行变长编码，一定存在一种无失真编码方法，构成唯一可译码，其码字平均长度满足下列不等式:

$$\frac{H(X)}{\log m} \le \bar{L} < \frac{H(X)}{\log m} + 1 \quad . \qquad H_m(X) \le \bar{L} < H_m(X) + 1 \quad .$$

（表示以 m 进制信息量单位测度）

$$\min \sum_i p_i l_i \qquad \Rightarrow l_i^* = -\log_m p_i$$

$$s.t. \ \sum_i m^{-l_i} \le 1$$

$$\Rightarrow \bar{L} = \sum_i p_i l_i^* = -\sum_i p_i \log_m p_i = \frac{H(X)}{\log m}$$

3.2 离散无失真信源编码定理

- 无失真变长信源编码定理

无失真变长信源编码定理（香农第一定理）：

离散无记忆信源 X 的 N 次扩展信源 X^N，其熵为 $H(X^N)$。对信源 X^N 进行 m 进制编码，总可以找到一种编码方式，构成唯一可译码，使信源 X 中的每个信源符号所需的平均码长满足：

$$\frac{H(X)}{\log m} \le \frac{\overline{L_N}}{N} < \frac{H(X)}{\log m} + \frac{1}{N}$$

$$H_m(X) \le \frac{\overline{L_N}}{N} < H_m(X) + \frac{1}{N}$$

$$H(X) \le \frac{\overline{L_N}}{N} \log m < H(X) + \frac{\log m}{N}$$

当 $N \to \infty$ 时，

$$\lim_{N\to\infty} \frac{\overline{L_N}}{N} = \frac{H(X)}{\log m} = H_m(X)$$

（表示以 m 进制信息量单位测度）

3.2 离散无失真信源编码定理

- 无失真变长信源编码定理

推广到平稳有记忆信源：

$$\frac{H(X_1X_2\cdots X_N)}{\log m}\le \overline{L_N} < \frac{H(X_1X_2\cdots X_N)}{\log m}+1$$

$$\frac{H(X_1X_2\cdots X_N)}{N\log m}\le \frac{\overline{L_N}}{N} < \frac{H(X_1X_2\cdots X_N)}{N\log m}+\frac{1}{N}$$

$$\lim_{N\to\infty}\frac{\overline{L_N}}{N}=\frac{1}{\log m}\lim_{N\to\infty}\frac{H(X_1X_2\cdots X_N)}{N}=\frac{H_\infty(X)}{\log m}$$

3.2 离散无失真信源编码定理

- 无失真变长信源编码定理

$$H(X) \le \frac{\overline{L_N}}{N}\log m < H(X) + \frac{\log m}{N}$$

变长信源编码后**信源**的信息传输率 $R' = \frac{\overline{L_N}}{N}\log m < H(X) + \frac{\log m}{N}$

（编码后平均每个信源符号能载荷的最大信息量）

香农第一定理也可陈述为：当 $R' > H(X)$ 就存在唯一可译变长编码；

若 $R' < H(X)$，不存在唯一可译变长码，不能实现无失真的信源编码。

若从**信道**角度来看，变长信源编码后**信道**的信息传输率（码率）

$$R = \frac{H(X)}{\overline{L}} = \frac{H(X)}{\overline{L}} \quad \text{（比特/码符号）}$$

$$\frac{\text{（比特/信源符号）}}{\text{（码符号/信源符号）}}$$

$$\because \overline{L} = \frac{\overline{L_N}}{N} \ge \frac{H(X)}{\log m} \qquad \therefore R = \frac{H(X)}{\overline{L}} \le \log m$$

当平均码长 $\overline{L}$ 达到极限值 $H(X)/\log m$ 时取等号，即 $R = \log m$ 。

3.2 离散无失真信源编码定理

- 无失真变长信源编码定理

为了衡量各种编码距离极限压缩值的情况，定义变长码的编码效率。因为平均码长 $\bar{L}$ 一定是大于或者等于极限值的，所以定义变长码的编码效率为极限值与平均码长 $\bar{L}=\frac{\bar{L}_N}{N}$ 之比。

$$H(X)\le\frac{\bar{L}_N}{N}\log m<H(X)+\frac{\log m}{N}$$

一般对于**平稳有记忆信源**有：

$$\eta=\frac{\frac{H_\infty}{\log m}}{\bar{L}}=\frac{H_\infty}{\bar{L}\log m}$$

对于**无记忆信源**则有：

$$\eta=\frac{H_m(X)}{\bar{L}}=\frac{H(X)}{\bar{L}\log m}>\frac{H(X)}{H(X)+\frac{\log m}{N}}$$

为了衡量各种编码与最佳码的差距，定义码的剩余度为

$$1-\eta=1-\frac{H_m(X)}{\bar{L}}$$

3.2 离散无失真信源编码定理

- 无失真变长信源编码定理 $H(X) \le \frac{\overline{L_N}}{N}\log m < H(X) + \frac{\log m}{N}$

例：设离散无记忆信源概率空间

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

若取差错率 $\delta \le 10^{-6}$，编码效率为90%，则 L 应满足?

- 信源熵：$H(X) = -\sum_i p(x_i)\log p(x_i) = 2.55bit/符号$

$$\eta > \frac{H(X)}{H(X) + \frac{\log m}{N}} \qquad 0.9 = \frac{2.55}{2.55 + \frac{1}{L}} \qquad \Rightarrow L = 4$$

3.2 离散无失真信源编码定理

- **无失真变长信源编码定理**

  例：设离散无记忆信源概率空间 $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$

  - 信源熵：$H(X) = -\sum_i p(x_i)\log p(x_i) = 0.811bit/符号$
  - 若用二元定长编码(0,1)来构造一个即时码:

    $a_1 \to 0, a_2 \to 1$
  - 平均码长为 $\overline{K} = 1$
  - 编码效率为 $\eta = \frac{H_L(X)}{\overline{K}} = 0.811$
  - 信道的信息传输率为 $R = 0.811bit/二元码符号$

$$\eta = \frac{H(X)}{R'} = \frac{H(X)}{\frac{K}{L}\log m}.$$

3.2 离散无失真信源编码定理

- **无失真变长信源编码定理**

  例：设离散无记忆信源概率空间 $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$

  - 信源熵：$H(X) = -\sum_i p(x_i) \log p(x_i) = 0.811 bit/符号$
  - 再对长度为 $N=2$ 的信源序列进行变长编码，其即时码如表
  - 平均码长为 $\bar{L}_2 = \frac{9}{16}\times 1 + \frac{3}{16}\times 2 + \frac{3}{16}\times 3 + \frac{1}{16}\times 3 = \frac{27}{16}$
  - 单个符号的平均码长 $\bar{L} = \frac{\bar{L}_2}{2} = \frac{27}{32}$
  - 编码效率 $\eta_2 = \frac{H(X)}{\bar{L}} = 0.961$
  - 信道的信息传输率为 $R_2 = 0.961 bit/二元码符号$

a_i	$p(a_i)$	即时码
$a_1 a_1$	9/16	0
$a_1 a_2$	3/16	10
$a_2 a_1$	3/16	110
$a_2 a_2$	1/16	111

$$\eta = \frac{H_m(X)}{\bar{L}} = \frac{H(X)}{\bar{L}\log m}$$

3.2 离散无失真信源编码定理

- **无失真变长信源编码定理**

  例：设离散无记忆信源概率空间 $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$

  - 信源熵：$H(X) = -\sum_i p(x_i)\log p(x_i) = 0.811bit/符号$
  - 将信源序列长度增加，对 $N=3$ 和 $N=4$ 的信源序列进行变长编码，
  - 编码效率分别为 $\eta_3 = 0.985, \eta_4 = 0.991$
  - 信道的信息传输率为 $R_3 = 0.985bit/二元码符号$

    $R_4 = 0.991bit/二元码符号$
  - 如果对这一信源采用定长二元码编码，要求编码效率达到96%时，允许译码错误概率 $\delta \le 10^{-5}$
  - 自信息的方差 $\sigma^2(X) = \sum_{i=1}^{2} p_i(\log p_i)^2 - [H(X)]^2 = 0.4715$
  - 所需要的信源序列长度 $L \ge \frac{0.4715}{(0.811)^2} \bullet \frac{(0.96)^2}{0.04^2 \times 10^{-5}} = 4.13 \times 10^7$

$$\eta = \frac{H(X)}{H(X)+\varepsilon}$$

$$L \ge \frac{\sigma^2(X)}{\varepsilon^2 \delta}$$

提纲

3.1 基本概念

3.2 离散无失真信源编码定理

3.3 香农编码

3.4 费诺编码

3.5 霍夫曼编码

3.3 香农编码

- **香农编码**

变长编码定理指出了平均码长与信源之间的关系，同时也指出了可以通过编码使平均码长达到极限值，这是一个很重要的极限定理。

变长编码定理指出，选择每个码字的长度 K_i 满足下式：

$$K_i = \left\lceil \log \frac{1}{p(x_i)} \right\rceil$$

或： $-\log_2 p(x_i) \le K_i < 1 - \log_2 p(x_i)$

就可以得到这种码，这种编码方式称为**香农编码**。

3.3 香农编码

- 二元香农编码的步骤

(1) 将信源符号按概率从大到小的顺序排列,

$$p(a_1) \geq p(a_2) \geq \cdots \geq p(a_n)$$

(2) 令 $P_1 = p(a_0) = 0$，用 P_i 表示第 i 个码字的累加概率

$$P_i = \sum_{k=1}^{i-1} p(a_k)$$

(3) 确定满足下列不等式的整数 K_i

$$-\log_2 p(a_i) \leq K_i < 1 - \log_2 p(a_i)$$

(4) 将 P_i 用二进制表示，并取小数点后 K_i 位作为符号 a_i 的编码。

3.3 香农编码

二元香农编码的步骤

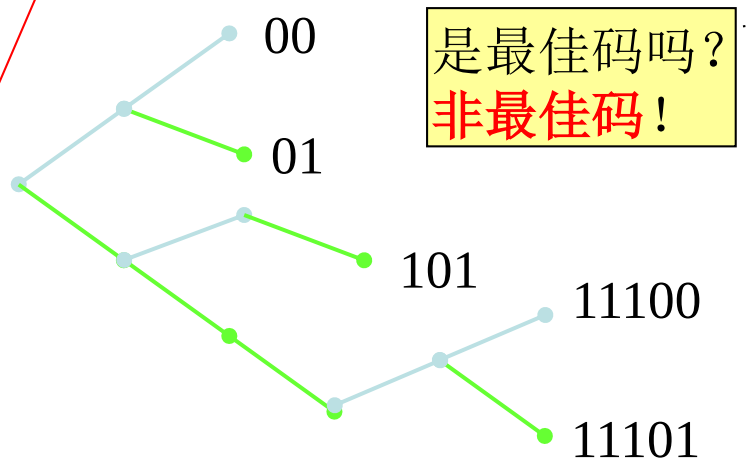
例：有一单符号离散无记忆信源

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 0.4 & 0.3 & 0.2 & 0.05 & 0.05 \end{bmatrix}$$

对该信源编二进制香农码。其编码过程如表所示

信源符号 x_i	符号概率 $p(x_i)$	累加概率 P_i	$-\log p(x_i)$	码长 K_i	码字
x_1	0.4	0	1.32	2	00
x_2	0.3	0.4	1.73	2	01
x_3	0.2	0.7	2.32	3	101
x_4	0.05	0.9	4.3	5	11100
x_5	0.05	0.95	4.3	5	11101

以$i = 3$为例:
码字长度:
$K_3 = [-\log 0.2] = 3$
累加概率
$P_i = 0.70 \rightarrow 0.10110\ldots$



是最佳码吗？
非最佳码！

3.3 香农编码

二元香农编码的步骤

- 香农码的平均码长

$$\overline{K} = \sum_{i=1}^{5} p(x_i)K_i = 0.4 \times 2 + 0.3 \times 2 + 0.2 \times 3 + 0.05 \times 5 \times 2 = 2.5$$

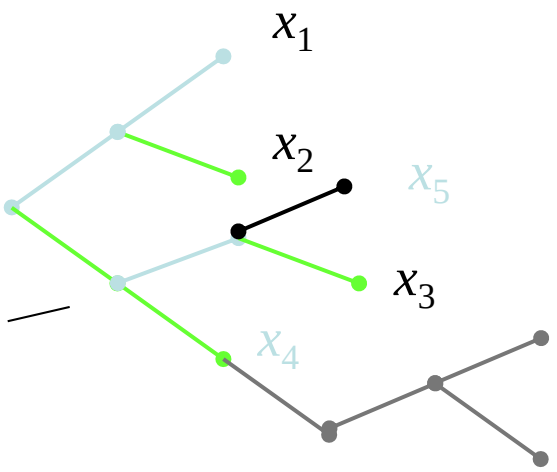
- 熵

$$H(X) = -0.4 \log 0.4 - 0.3 \log 0.3 - 0.2 \log 0.2 - 2 \times 0.05 \log 0.05$$

$$= 1.95$$

- 编码效率

$$\eta = \frac{H(X)}{\overline{K}} = \frac{1.95}{2.5} = 78\%$$



为提高编码效率，首先应达到满树。如把x_4x_5换成前面的节点，可减小平均码长。 不应先规定码长，而是由码树来规定码字，可得更好的结果。

提纲

3.1 基本概念

3.2 离散无失真信源编码定理

3.3 香农编码

3.4 费诺编码

3.5 霍夫曼编码

3.4 费诺编码

- 费诺编码

费诺编码属于概率匹配编码。编码步骤如下：

（1）将信源符号按概率从大到小的顺序排列，

$$p(a_1) \geq p(a_2) \geq \cdots \geq p(a_n)$$

（2）按编码进制数将概率分组，使每组概率尽可能相等或接近。如编二进制码就分成两组，编 m 进制码就分成 m 组。

（3）给每一组分配1位码元；

（4）将每一组再按同样原则划分，重复步骤(2)和(3)，直至概率不再可分为止。

费诺编码

例：设有一单符号离散信源

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 0.4 & 0.3 & 0.2 & 0.05 & 0.05 \end{bmatrix}$$

平均码长：K= 2.1
编码效率：η=93%

平均码长：K= 2.0
编码效率：η=97.5%

- 对该信源编二进制费诺码。

信源符号 x_i	符号概率 $p(x_i)$	第1次分组	第2次分组	第3次分组	码字	码长
x_1	0.4	0	0		00	2
x_4	0.05		1	0	010	3
x_5	0.05			1	011	3
x_2	0.3	1	0		10	2
x_3	0.2		1		11	2

信源符号 x_i	符号概率 $p(x_i)$	第1分组	第2分组	第3分组	第4分组	码字	码长
x_1	0.4	0				0	1
x_2	0.3	1	0			10	2
x_3	0.2		1	0		110	3
x_4	0.05			1	0	1110	4
x_5	0.05				1	1111	4

3.4 费诺编码

费诺编码

- 平均码长

$$\overline{K_1} = \sum_{i=1}^{5} p(x_i)K_i = 0.4\times 2 + 0.3\times 2 + 0.2\times 2 + 2\times(0.05\times 3) = 2.1$$

$$\overline{K_2} = \sum_{i=1}^{5} p(x_i)K_i = 0.4\times 1 + 0.3\times 2 + 0.2\times 3 + 2\times(0.05\times 4) = 2.0$$

- 编码效率

$$\eta_1 = \frac{H(X)}{\overline{K_1}} = \frac{1.95}{2.1} = 93\% \qquad \eta_2 = \frac{H(X)}{\overline{K_2}} = \frac{1.95}{2.0} = 97.5\%$$

- 费诺码比较适合于每次分组概率都很接近的信源。
- 特别是对每次分组概率都相等的信源进行编码时，可达到理想的编码效率。

3.4 费诺编码

费诺编码

例：有一单符号离散无记忆信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6 & x_7 & x_8 \\ 1/4 & 1/4 & 1/8 & 1/8 & 1/16 & 1/16 & 1/16 & 1/16 \end{bmatrix}$$

- 对该信源编二进制费诺码，编码过程如表：

信源符号	概率	编码				码字	码长
x_1	0.25	0	0			00	2
x_2	0.25		1			01	2
x_3	0.125	1	0	0		100	3
x_4	0.125			1		101	3
x_5	0.0625		1	0	0	1100	4
x_6	0.0625				1	1101	4
x_7	0.0625			1	0	1110	4
x_8	0.0625				1	1111	4

费诺编码

- 信源熵为 $H(X) = 2.75$ (比特/符号)
- 平均码长为

$$\overline{K} = (0.25+0.25)\times 2 + 0.12\times 2\times 3 + 0.0625\times 4\times 4 = 2.75(\text{比特/符号})$$

- 编码效率为

$$\eta = 1$$

- 之所以如此,因为每次所分两组的概率恰好相等。

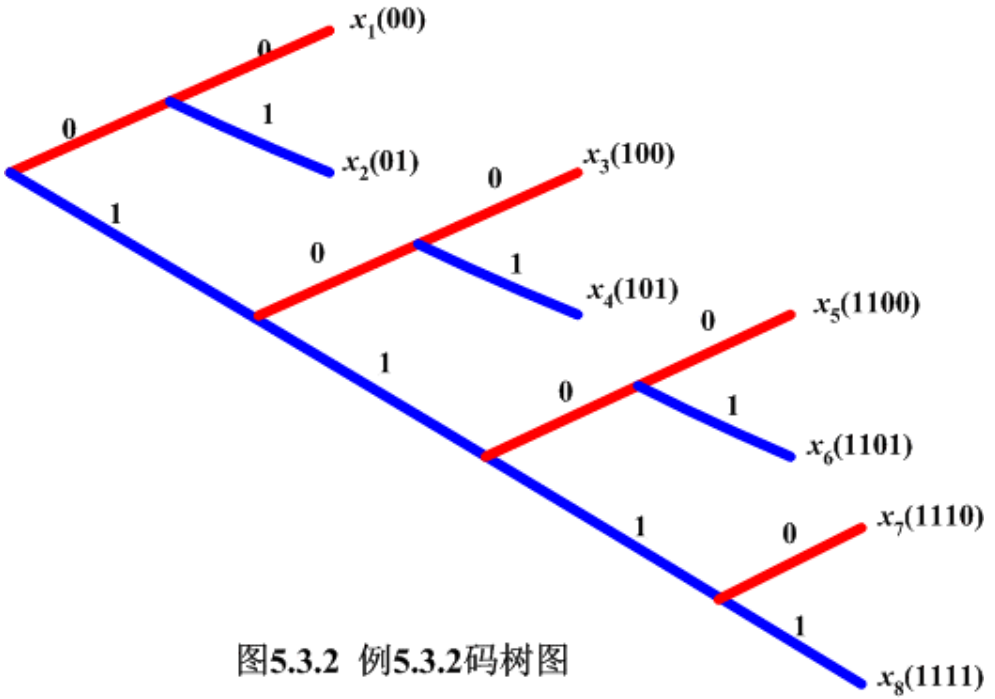


图5.3.2 例5.3.2码树图

提纲

3.1 基本概念

3.2 离散无失真信源编码定理

3.3 香农编码

3.4 费诺编码

3.5 霍夫曼编码

3.5 霍夫曼编码

霍夫曼编码

- 哈夫曼编码也是用码树来分配各符号的码字。
- 费诺码是从树根开始，把各节点分给某子集，若子集已是单点集，它就是一片树叶而作为码字。
- 哈夫曼编码是先给每一符号一片树叶，逐步合并成节点直到树根。
- 哈夫曼(*Huffman*)编码是一种效率比较高的变长无失真信源编码方法。

霍夫曼编码

- 哈夫曼编码的步骤如下：

(1) 将信源消息符号按其出现的概率大小依次排列

$$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_n)$$

(2) 取两个概率最小的字母分别配以0和1两码元，并将这两个概率相加作为一个新字母的概率，与未分配的二进符号的字母重新排队。

(3) 对重排后的两个概率最小符号重复步骤(2)的过程。

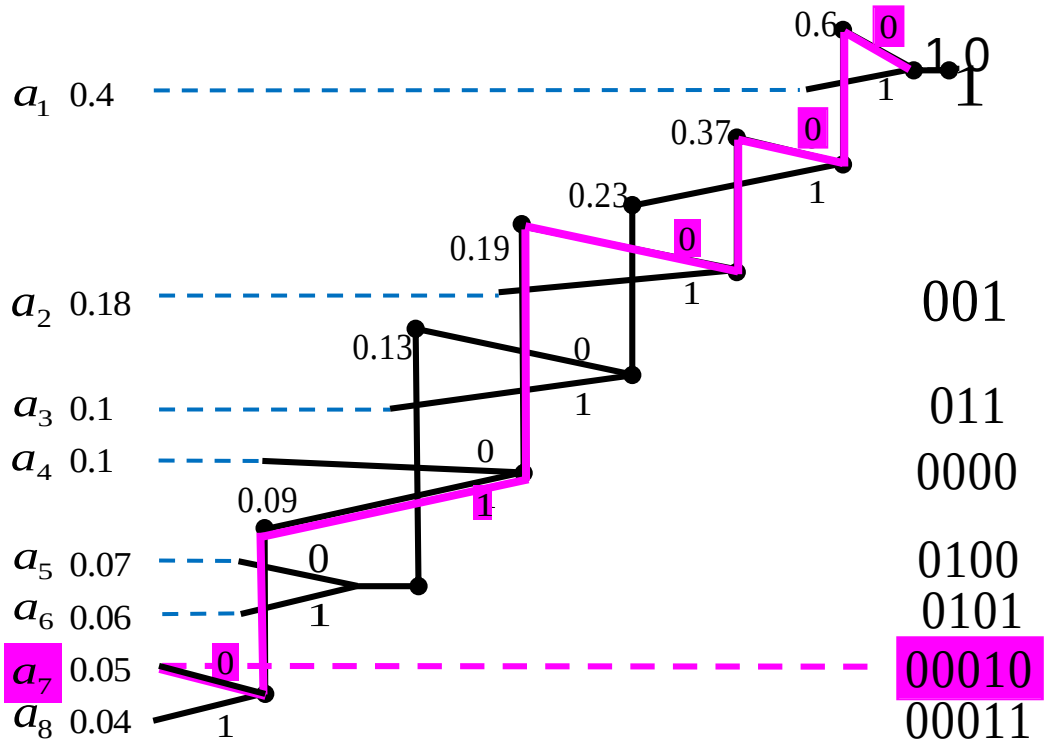
(4) 不断继续上述过程，直到最后两个符号配以0和1为止。

(5) 从最后一级开始，向前返回得到各个信源符号所对应的码元序列，即相应的码字。

- 霍夫曼编码

例：设有一单符号离散无记忆信源，试对该信源编二进制赫夫曼码。

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$



$$H(X) = 2.55(bit/sign)$$

$$\bar{K} = 2.61$$

$$\eta = \frac{H(X)}{R} = 97.7\%$$

若采用定长编码，码长$K=3$,

则编码效率$\eta = \frac{2.55}{3} = 85\%$

霍夫曼编码的效率提高了12.7%

3.5 霍夫曼编码

霍夫曼编码

- 哈夫曼的编法并不惟一。
- 每次对缩减信源两个概率最小的符号分配“0”和“1”码元是任意的，所以可得到不同的码字。只要在各次缩减信源中保持码元分配的一致性，即能得到可分离码字。
- 不同的码元分配，得到的具体码字不同，但码长 K_i 不变，平均码长也不变，所以没有本质区别。
- 缩减信源时，若合并后的新符号概率与其他符号概率相等，从编码方法上来说，这几个符号的次序可任意排列，编出的码都是正确的，但得到的码字不相同。
- 不同的编法得到的码字长度 K_i 也不尽相同。

3.5 霍夫曼编码

- 霍夫曼编码

例：设有一单符号离散无记忆信源，试对该信源编二进制赫夫曼码。

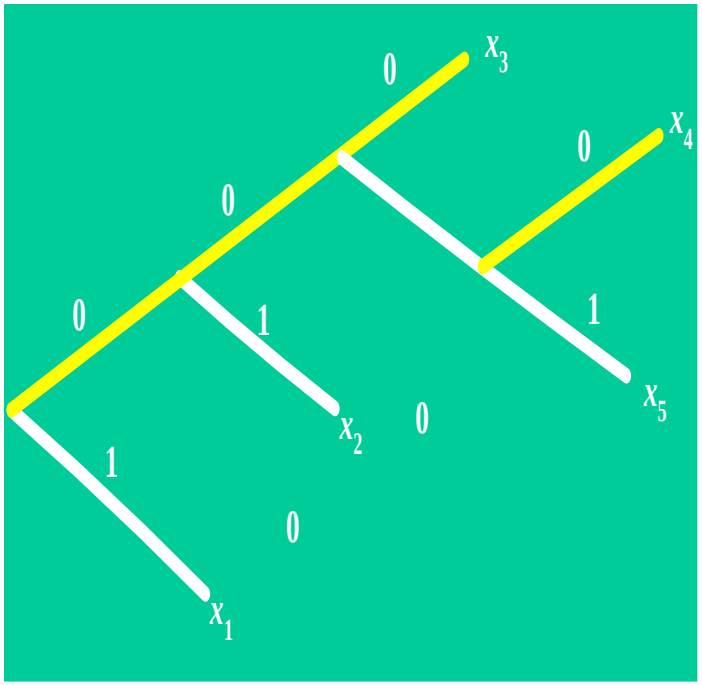
$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$$

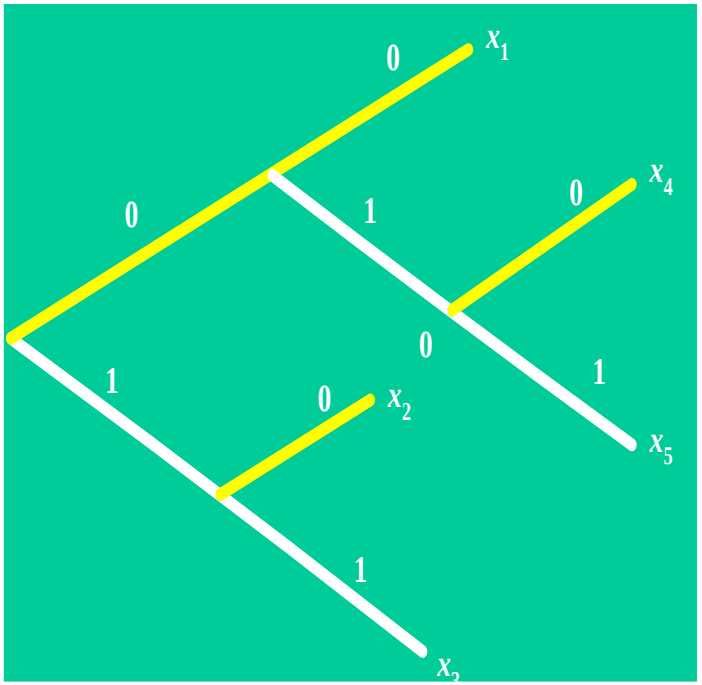
信源符号 x_i	符号概率 $p(x_i)$	编码过程	码字	码字
x_1	0.4	0.4 0.4 0.6 0	1	00
x_2	0.2	0.2 0.4 0 0.4 1	01	10
x_3	0.2	0.2 0 0.2 1	000	11
x_4	0.1	0 0.2 1	0010	010
x_5	0.1	1	0011	011

- 霍夫曼编码

例：设有一单符号离散无记忆信源，试对该信源编二进制赫夫曼码。

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$$





3.5 霍夫曼编码

- 霍夫曼编码

例：设有一单符号离散无记忆信源，试对该信源编二进制赫夫曼码。

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$$

信源符号x_i	概率$p(a_i)$	码字W_{i1}	码长K_{i1}	码字W_{i2}	码长K_{i2}
x_1	0.4	1	1	00	2
x_2	0.2	01	2	10	2
x_3	0.2	000	3	11	2
x_4	0.1	0010	4	010	3
x_5	0.1	0011	4	011	3

3.5 霍夫曼编码

霍夫曼编码

- 单符号信源编二进制哈夫曼码，编码效率主要决定于信源熵和平均码长之比。
- 对相同的信源编码，其熵是一样的，采用不同的编法，得到的平均码长可能不同。
- 平均码长越短，编码效率就越高。
- 编法一的平均码长为

$$\overline{K_1} = \sum_{i=1}^{5} p(x_i)K_i = 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 0.1 \times 4 \times 2 = 2.2$$

- 编法二的平均码长为

$$\overline{K_2} = \sum_{i=1}^{5} p(x_i)K_i = 0.4 \times 2 + 0.2 \times 2 \times 2 + 0.1 \times 3 \times 2 = 2.2$$

- 两种编法的平均码长相同，所以编码效率相同。 $\eta = \frac{H(X)}{\overline{K}} = 0.965$

霍夫曼编码

讨论：哪种方法更好？

定义码字长度的方差σ^2：

$$\sigma^2 = E[(K_i - \overline{K})^2] = \sum_{i=1}^{5} p(x_i)(K_i - \overline{K})^2$$

$$\sigma_1^2 = 0.4(1-2.2)^2 + 0.2(2-2.2)^2 + 0.2(3-2.2)^2 + 0.1(4-2.2)^2 \times 2 = 1.36$$

$$\sigma_2^2 = 0.4(2-2.2)^2 + 0.2(2-2.2)^2 \times 2 + 0.1(3-2.2)^2 \times 2 = 0.16$$

- 第一种方法编出的5个码字有4种不同的码长，第二种方法编出的码字只有2种不同的码长；
- 第二种编码方法的码长方差要小许多，比较接近于平均码长；
- 第二种编码方法更简单、更容易实现，所以更好。

3.5 霍夫曼编码

霍夫曼编码

讨论：哪种方法更好？

结论：

- 在哈夫曼编码过程中，对缩减信源符号按概率由大到小的顺序重新排列时，应使合并后的新符号尽可能排在靠前的位置，这样可使合并后的新符号重复编码次数减少，使短码得到充分利用。

总结

克拉夫特不等式：即时码和唯一可译码存在的充要条件

$$\sum_{i=1}^{q} m^{-l_i} \le 1$$

最佳等长编码的效率：$\eta = \dfrac{H(X)}{H(X)+\varepsilon}$

信源序列长度 N 与最佳编码效率和允许错误概率 δ 关系：$N \ge \dfrac{D(I(\alpha_i))}{\varepsilon^2 \delta}$

$$\ge \frac{D(I(\alpha_i))}{H^2(X)} \frac{\eta^2}{(1-\eta)^2 \delta}$$

无失真信源编码定理，香农第一定理：

离散无记忆信源无失真压缩的极限值：

$$H_m(X) \le \bar{L} = \sum_i p(a_i) l_i < H_m(X) + 1 \quad \text{或} \quad H_m(X) \le \frac{\overline{L_N}}{N} < H_m(X) + \frac{1}{N}$$

当 $N \to \infty$ 则 $\lim\limits_{N\to\infty} \dfrac{\overline{L_N}}{N} = H_m(X)$

离散平稳信源无失真压缩的极限值：$\lim\limits_{N\to\infty} \dfrac{\overline{L_N}}{N} = \dfrac{H_\infty}{\log m}$

总结

编码后信源的信息传输率：

等长码：$R' = \frac{K}{L}\log m$ （比特/信源符号） K 为码长，L 为信源序列长度

变长码：$R' = \frac{\overline{L_N}}{N}\log m$（比特/信源符号）$\overline{L_N}$ 为平均码长，N 为信源长度

编码后信道的信息传输率（码率）： $R = \frac{H_\infty}{\overline{L_N}/N}$ （比特/码符号）

离散无记忆信源 $H_\infty = H(X)$

无失真信源编码效率： $\eta = \frac{H_\infty}{\overline{L}\log m}$ 其中 $\overline{L} = \frac{\overline{L_N}}{N}$

离散无记忆信源 $\eta = \frac{H(X)}{\overline{L}\log m}$

无失真信源编码剩余度： $1-\eta$