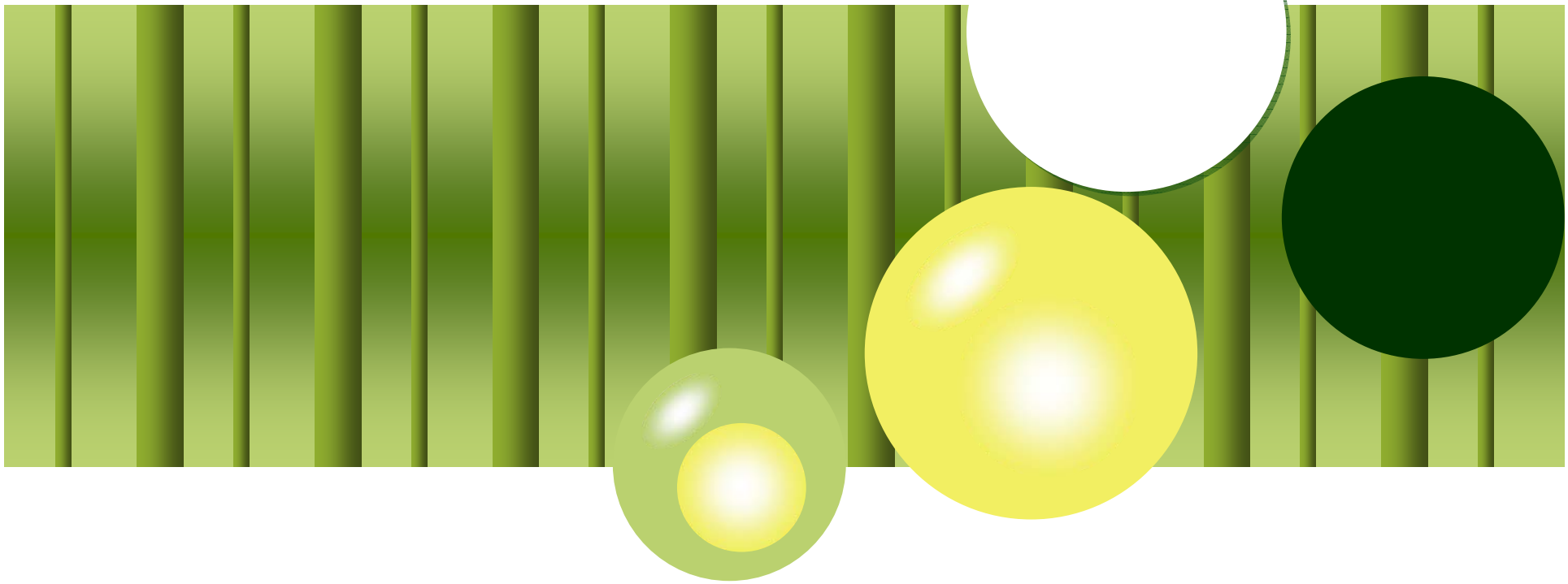
第五章 无失真信源编码

中山大学 信息科学与技术学院



主要内容

本章主要介绍无失真信源编码定理与一些重要的无失真信源编码方法

- 一、概述
- 二、定长码
- 三、变长码
- 四、哈夫曼编码
- 五、几种实用的信源编码方法

- 信源编码：将信源符号序列按一定的数学规律映射成由码符号组成的码序列的过程。
- 信源译码：根据码序列恢复信源序列的过程。
- 无失真信源编码：即信源符号可以通过编码序列无差错地恢复。
  （适用于离散信源的编码）
- 限失真信源编码：信源符号不能通过编码序列无差错地恢复。
  （可以把差错限制在某一个限度内）

- 信源编码的目的：提高传输有效性，即用尽可能短的码符号序列来代表信源符号。

- 无失真信源编码定理证明，如果对信源序列进行编码，当序列长度足够长时 ，存在无失真编码使得传送每信源符号所需的比特数接近信源的熵。因此，采用有效的信源编码会使信息传输效率得到提高。

注意：信源编码的效率主要看平均到每个信源符号上的编码符号长度

§5.1 概述

本节主要内容

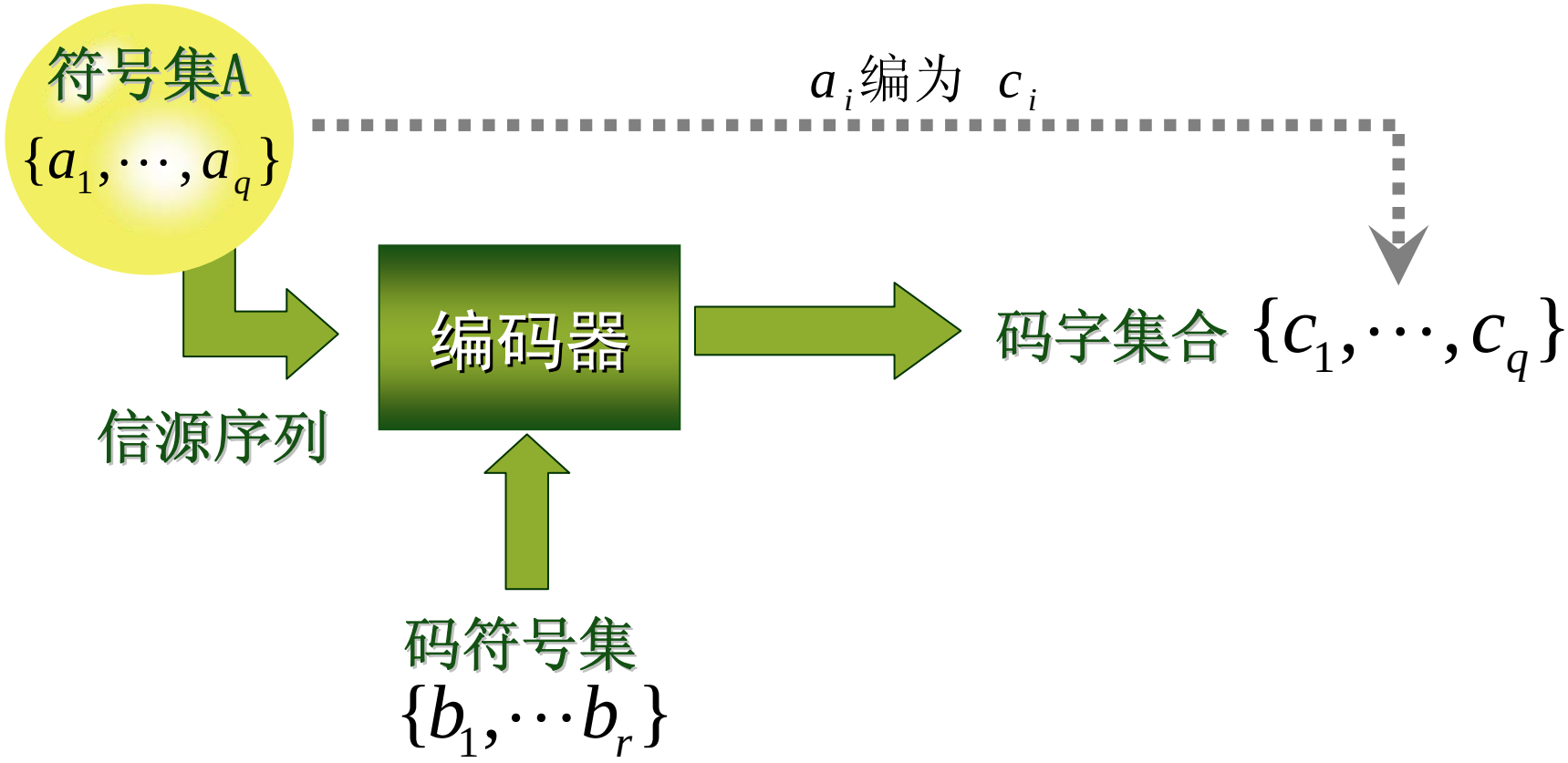
一、信源编码器

二、信源编码的分类

三、分组码

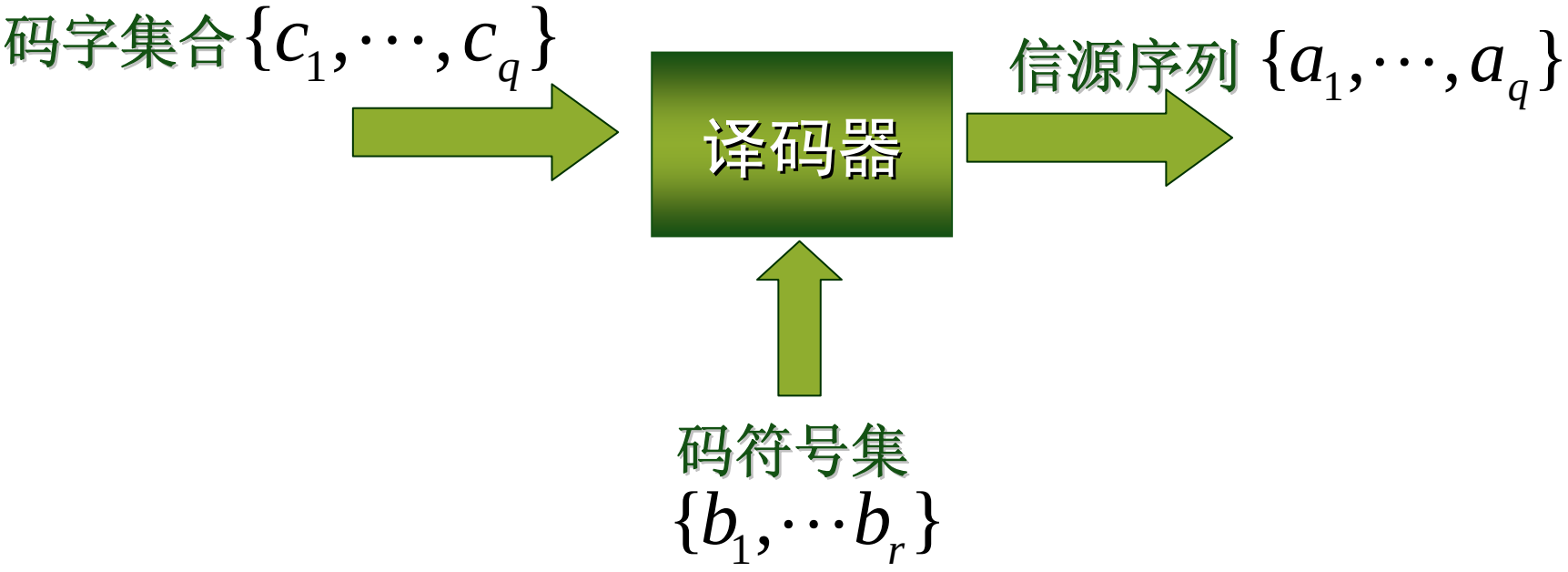
§5.1.1 信源编码器

分组码单符号信源编码器



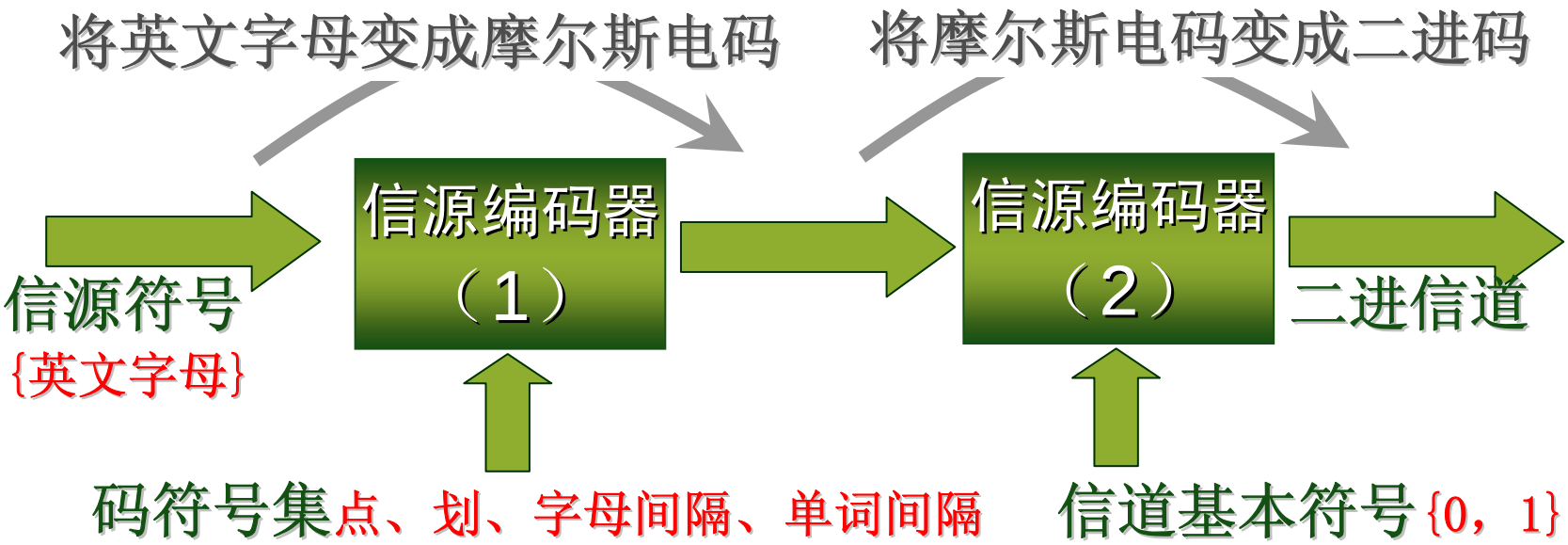
信源译码器

分组码单符号译码器



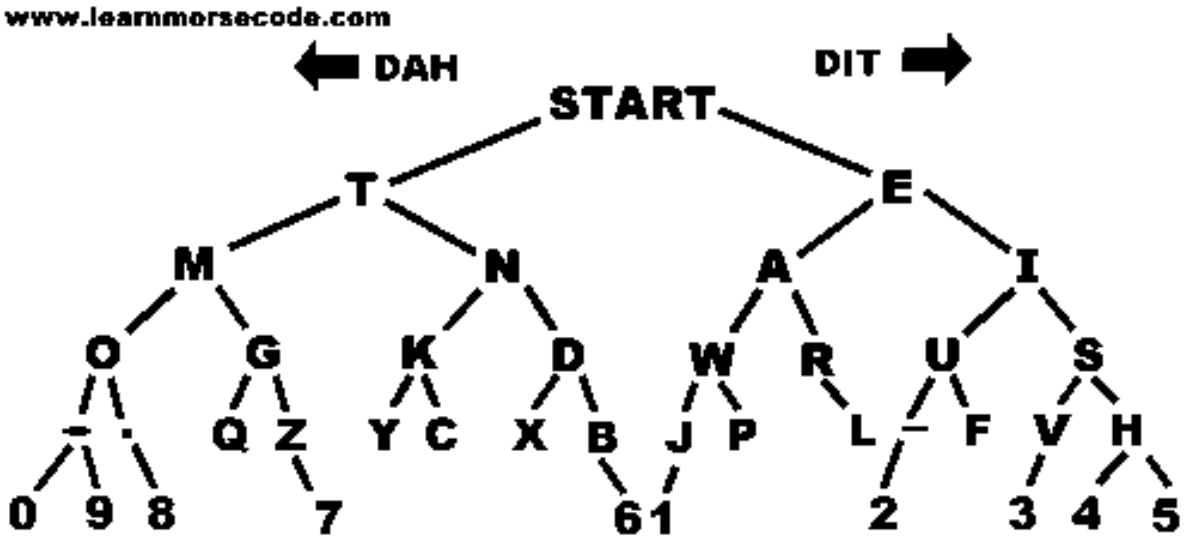
简单信源编码器

摩尔斯信源编码器



符号	点	划	字母间隔	单词间隔
电平	+ -	+ + + -	- - -	- - - - -
二进代码	1 0	1110	000	00000

摩尔斯信源编码器



www.learnmorsecode.com

A ·−	I ··	Q −−·−	Y −·−−
B −···	J ·−−−	R ·−·	Z −−··
C −·−·	K −·−	S ···	Period ·−·−·−
D −··	L ·−··	T −	Comma −−··−−
E ·	M −−	U ··−	? ··−−··
F ··−·	N −·	V ···−	/ −··−·
G −−·	O −−−	W ·−−	@ ·−−·−·
H ····	P ·−−·	X −··−	

1 ·−−−−	
2 ··−−−	
3 ···−−	
4 ····−	
5 ·····	
6 −····	
7 −−···	
8 −−−··	
9 −−−−·	
0 −−−−−	

原信源的N次扩展码

将N个信源符号编成一个码字。相当于对原信源的N次扩展源的信源符号进行编码。

例 信源X={0, 1}的二次扩展源X_2的符号集为：{00, 01, 10, 11}。对X_2编码，即为原信源X的二次扩展码。

§5.1.2 信源编码的分类

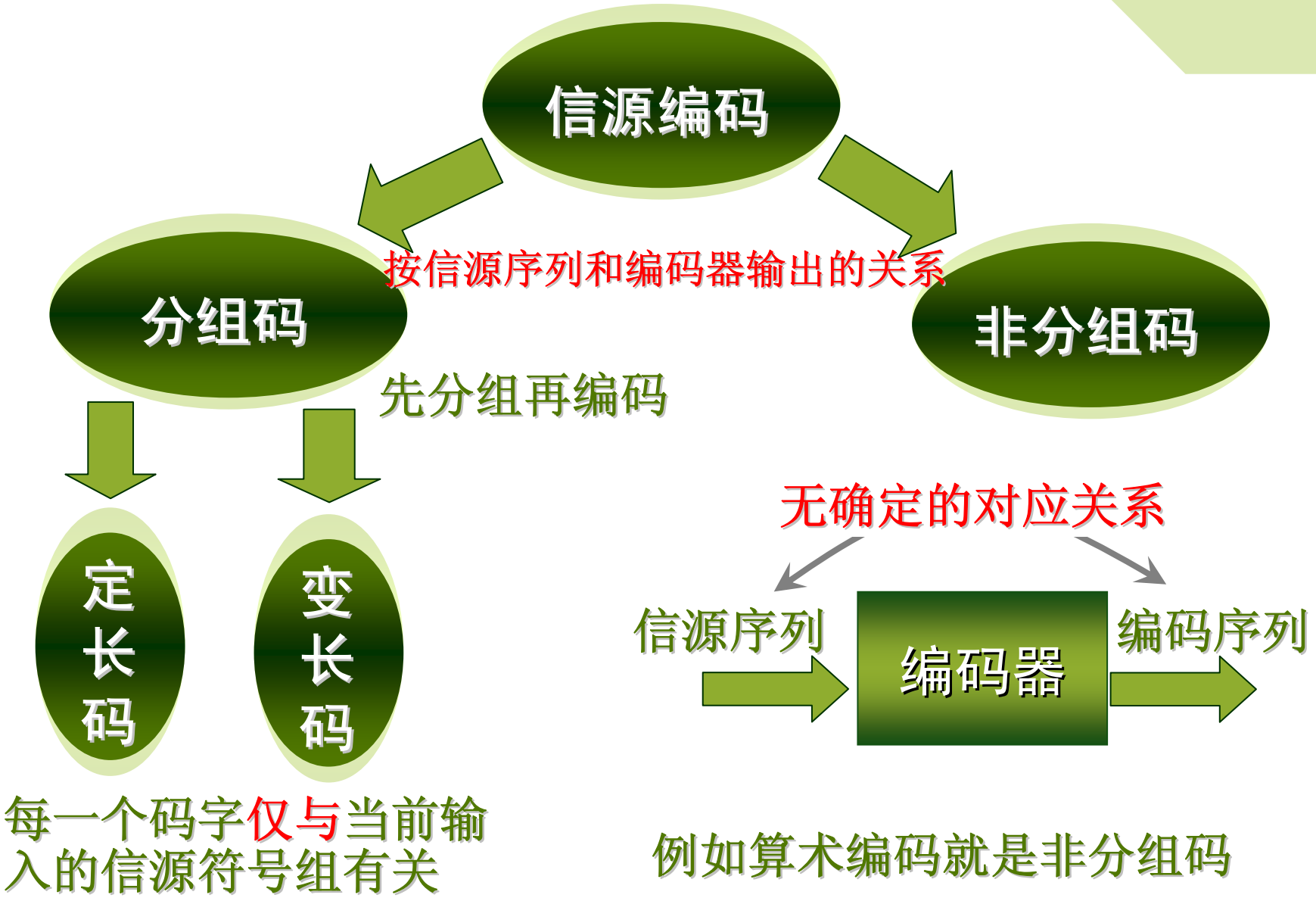
✓ **概率匹配编码**：信源符号的概率已知。

- 分组码：先分组再编码。在分组码中，每一个码字仅与当前输入的信源符号组有关，与其他信源符号无关。

  包括：定长码、变长码（Huffman编码、费诺编码）

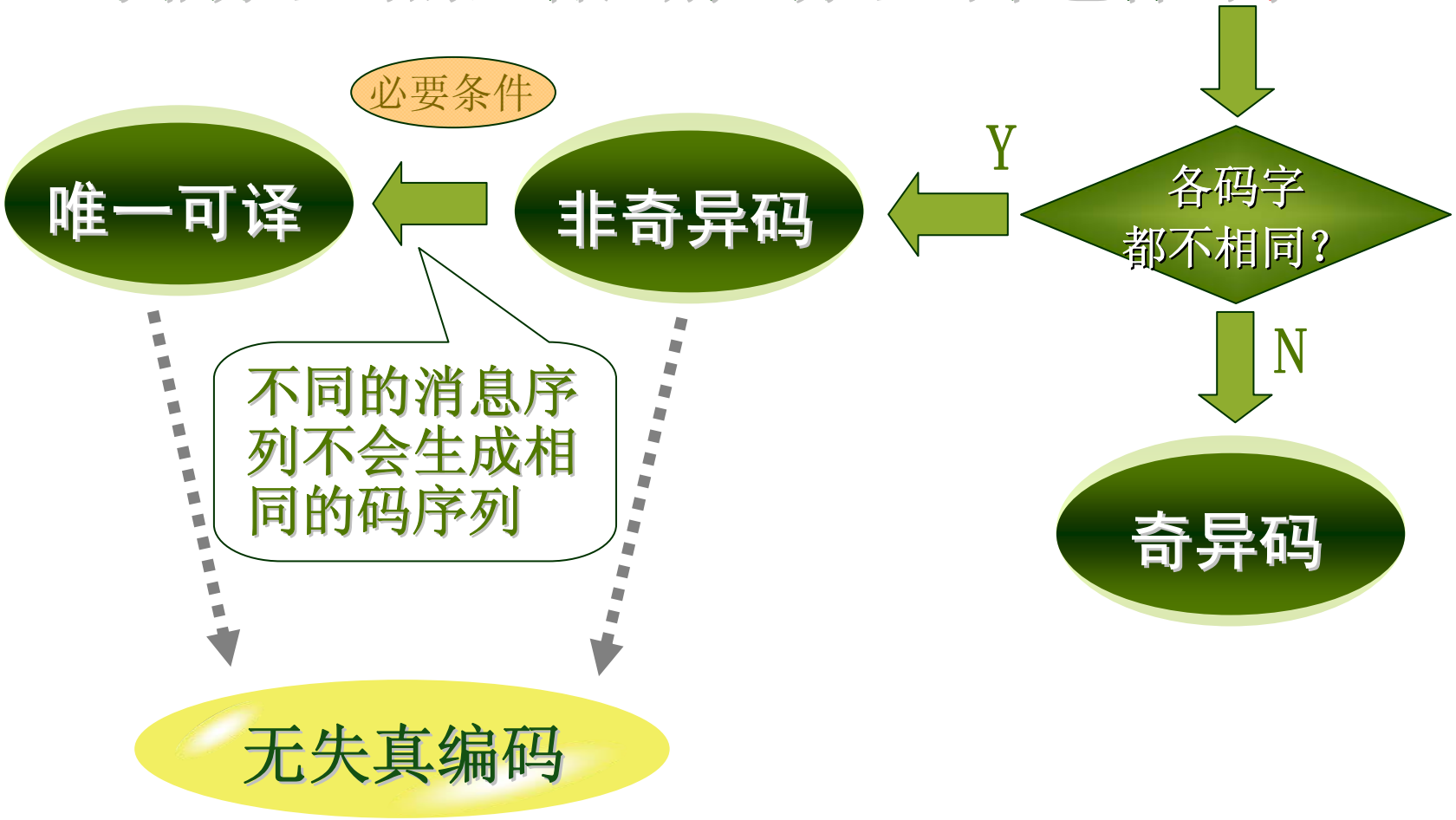
- 非分组码：码序列中的符号与信源序列中的符号无确定的对应关系。例如算术编码。 2021年数学阿贝尔奖

✓ **通用编码**：信源符号的概率未知。

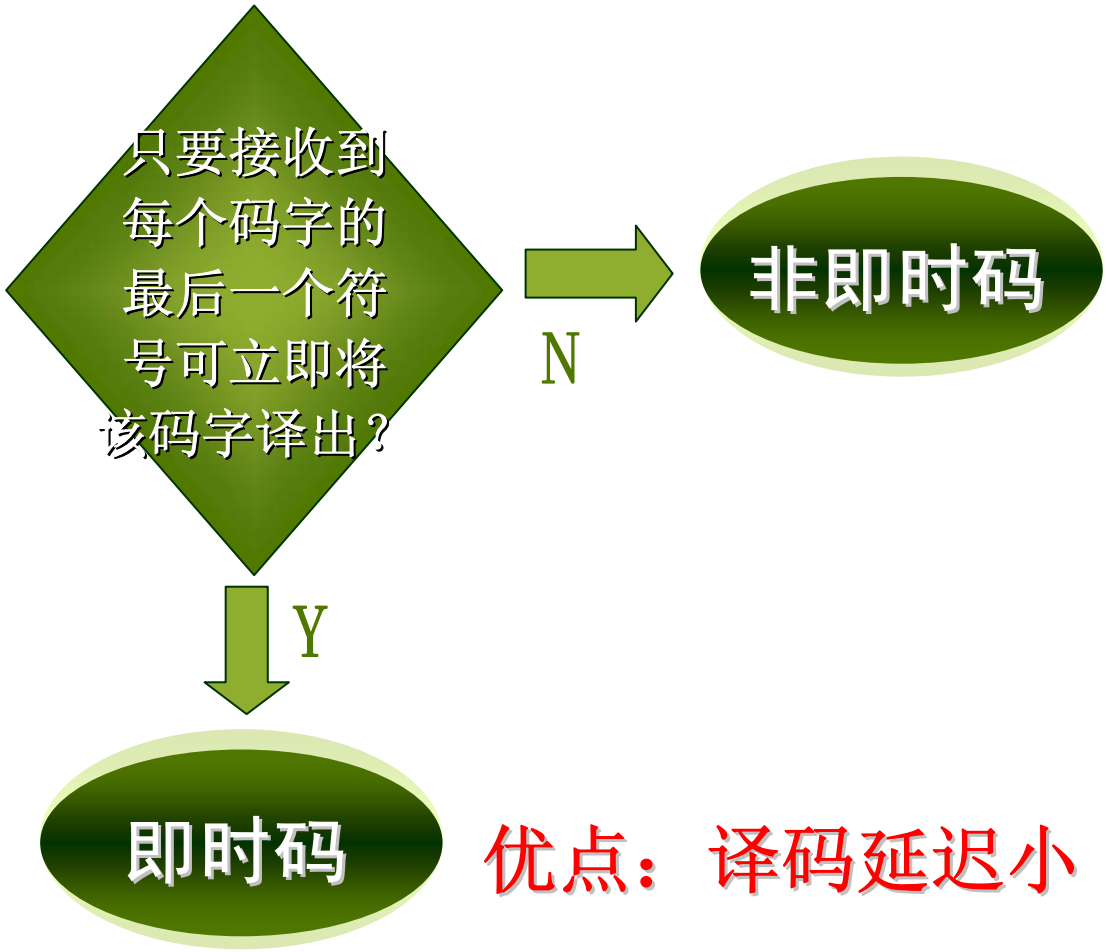

信源编码
按信源序列和编码器输出的关系
分组码
非分组码
先分组再编码
定长码
变长码
每一个码字仅与当前输入的信源符号组有关
无确定的对应关系
信源序列
编码器
编码序列
例如算术编码就是非分组码

§5.1.3 分组码

✓ 与非分组码的显著区别：分组码中包含码字



即时码与非即时码



优点：译码延迟小

✓ 异前置码

❖设$\vec{x}_k$ 为长度为 k 的码字，即 $\vec{x}_k = x_1, \cdots, x_k$ ，

称 $x_1 x_2 \cdots x_j (1 \le j \le k)$ 为 $\vec{x}_k$ 的前置。

- ❖ 一个码中无任何码字是其他码字的前置
- ❖ 异前置码是唯一可译码
- ❖ 异前置码与即时码是等价的

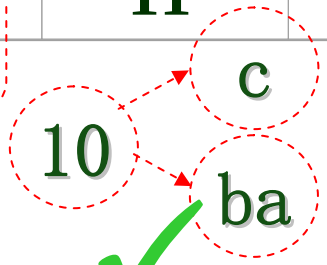
✓ 逗号码

- ❖ 用一个特定的码符号表示所有码字的结尾
- ❖ 逗号码是唯一可译码

例 5.1 设信源符号集为{a, b, c, d}，采用6种分组编码如下表，分析每一个码的唯一可译性

符号	码A	码B	码C	码D	码E	码F
a	0	0	00	0	1	0
b	0	1	01	10	01	01
c	1	10	10	110	001	011
d	10	11	11	111	0001	0111

即时码



	等长	异前置码	逗号码	0表示开头
非奇异	✗	✓	✓	✓	✓
唯一可译	✗	✗	✓	✓	✓	✓

变长和定长码比较

一般指信源符号速率，不是编码速率

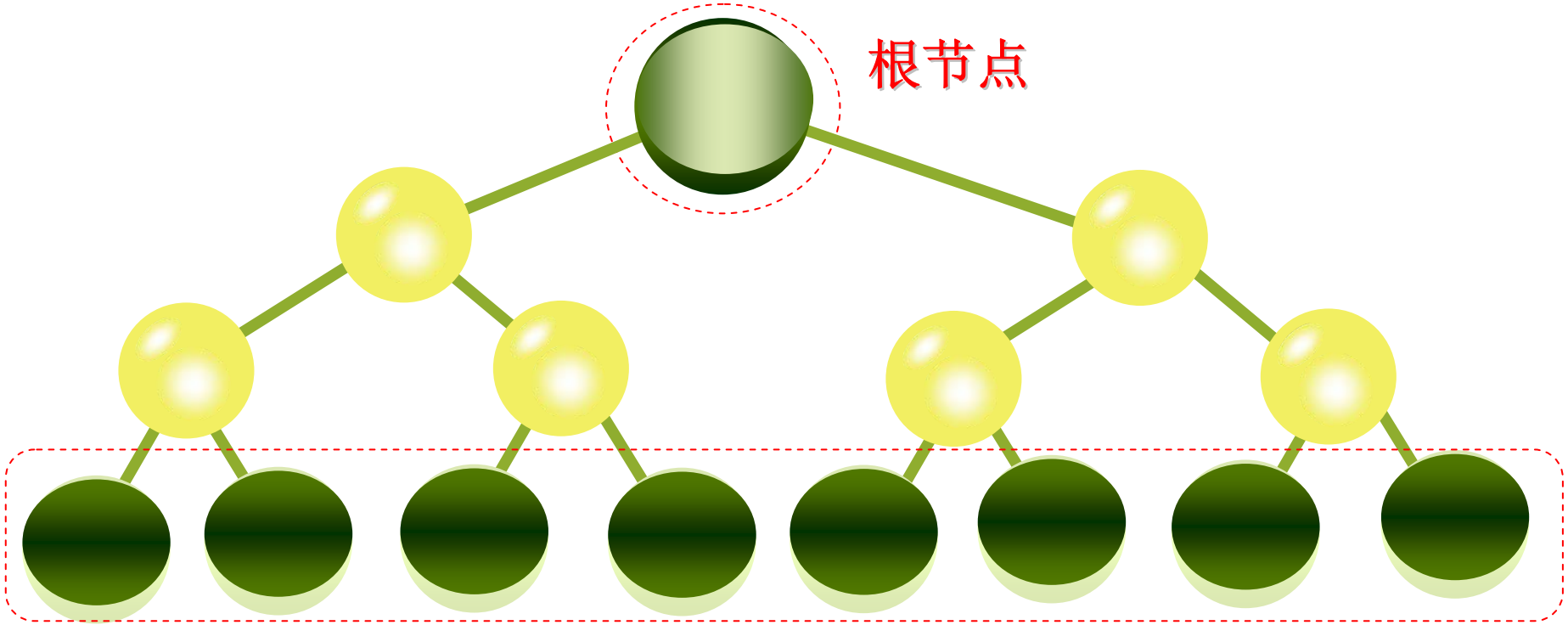
变长码	定长码
非奇异且异前置就唯一可译	只要非奇异，就唯一可译
速率变化→设置缓冲器	速率恒定→不需缓冲器
受误码影响大，逗号码除外	码长已知→容易同步
容易产生差错传播	无差错传播

1、恒定速率指信源符号
2、恒定编码速率不一定是恒定信源符号速率

使用上页码D传“a”和”b”—》0 10 —》第1个“0”翻转—》110—》译码—》“c”—》影响到第2个码—》信源失真

码树

✓ 码树是表示信源编码码字的重要工具之一

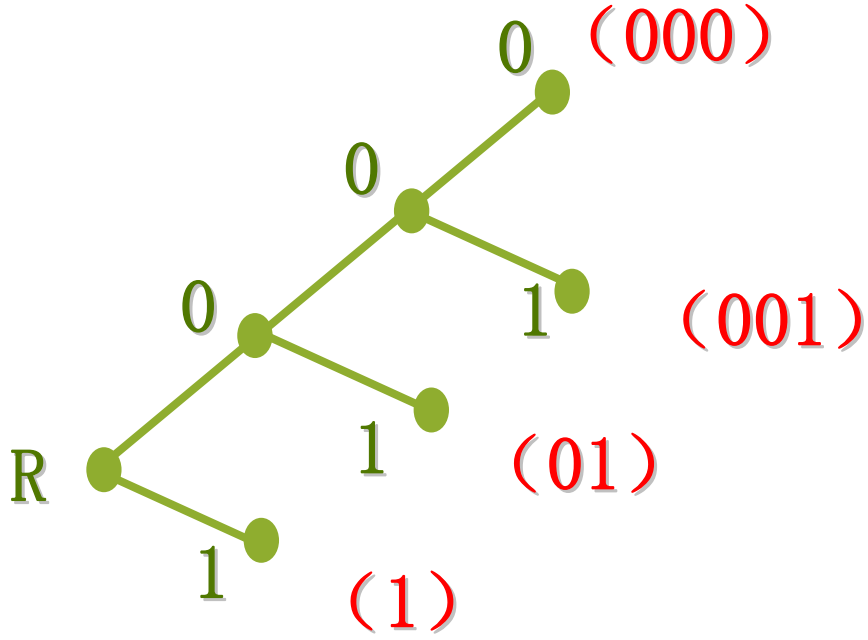


1、根节点代表码集合（code），有X个编码符号，就是X叉树

2、所有叶子是码字（codeword）

例 5.2　一个码C包含4个码字：{1,01,000,001}，试用码树来表示

解：采用二进制码树



一些结论

- ✓ 在码树中，n阶节点的个数最多为 r^n
  - ❖例：2进码树中，r阶节点数目最多为 2^r

- ✓ 非奇异码字总能与码树建立一一对应的关系

§5.2 定长码

本节主要内容

一、无失真编码条件

二、信源序列分组定理

三、定长码信源编码定理

§5.2.1 无失真编码条件

✓ 对于定长码，只要非奇异就唯一可译。这就要求码字的数目不少于被编码的信源序列的个数

✓ 单信源符号编码：

非奇异+唯一可译——》无失真

❖设信源X包含q个符号，码符号集包含的符号数为r

$$q \le r^{l}$$

→ 码长

✓ N长信源符号序列编码（N次扩展码）

N长信源符号数 ← $q^{N} \le r^{l}$ 或 $\frac{l}{N} \ge \frac{\log q}{\log r}$ → 平均每个信源符号所需码符号数

例 英文字母26个加1个空格可看成共27个符号的信源。 如对单符号进行编码：

$$27 \le 2^l \Rightarrow l \ge \log 27 = 4.755, \quad 取 l \ge 5$$

但是，如果采用适当的信源编码，理论上每信源符号所需二进码符号数可以远小于上面的值， 在理想情况下可以压缩到接近信源的熵1.4左右。本节就是从理论上证明这种压缩是可以实现的。

§5.2.2 信源序列分组定理

定理5.2.1

总可以找到N_0

离散无记忆信源

任意给定$\varepsilon > 0,\ \delta > 0$

使得

长度为$N \geq N_0$的信源序列都可以分成两组

① 序列 $\vec{x}$ 出现的概率 $p(\vec{x})$ 满足：

$$\left| \frac{1}{N} \log p(\vec{x}) + H(X) \right| < \delta \qquad (5.2.3)$$

② 所有符号序列出现概率之和小于ε

证：　我们先证明（5. 2. 3）式。　设信源符号集为 $A=\{a_1,a_2,\cdots a_q\}$，各符号出现的概率分别为 p_i，$\vec{x}=x_1x_2\cdots x_N$ 为长度为 N 的序列，N_i 为 $\vec{x}$中符号 a_i 出现的次数。　将信源序列按下列原则分成两：G_1和 G_2 其中，

$$G_1：\quad \{\vec{x}:\ \left|\frac{N_i}{N}-p_i\right|<\varsigma,\quad i=1,\cdots,q\} \qquad （5.2.4）$$

$$G_2：\quad \{\vec{x}:\text{其它}\}$$

根据大数定律，当序列足够长时，信源符号 a_i 出现的次数接近 Np_i。因此，G_1 中的序列意味其各符号出现的次数符合大数定律，称典型序列。

从（**5. 2. 4**）中可以看出，G_1 随 ς 的不同而改变。

设 $\vec{x} \in G_1$，则对于 $\vec{x}$ 中的信源符号 a_i，有

$$-\varsigma < \frac{N_i}{N} - p_i < \varsigma, \quad i = 1, \cdots q$$

或 $\frac{N_i}{N} = p_i + \theta_i \varsigma$ ，其中 $|\theta_i| < 1$

考虑到信源是无记忆（独立）的，所以 $\vec{x}$ 的概率 $p(\vec{x})$ $= p_1^{N_1} \cdots p_q^{N_q}$ ， $\vec{x}$ 自信息负值为：

$$\log p(\vec{x}) = \sum_{i=1}^{q} N_i \log p_i = \sum_{i=1}^{q} N(p_i + \theta_i \varsigma) \log p_i$$

$$= -NH(X) + N\varsigma \sum_{i=1}^{q} \theta_i \log p_i$$

所以 $$\frac{\log p(\vec{x})}{N}+H(X)=\varsigma\sum_{i=1}^{q}\theta_i\log p_i$$

$$\left|\frac{\log p(\vec{x})}{N}+H(X)\right|=\varsigma\left|\sum_{i=1}^{q}\theta_i\log p_i\right|\prec\varsigma\sum_{i=1}^{q}\left|\log p_i\right|$$

选择 ς，使得

$$\varsigma=\frac{\delta}{\sum_{i=1}^{q}\left|\log p_i\right|} \qquad (5.2.5)$$

则式（5.2.3）成立。

落入G2中序列的概率随N增加，任意减小

下面证明定理的后半部分。设 $\vec{x} \in G_2$ ， 根据（5. 2. 3）， 有

$$\left|\frac{\log p(\vec{x})}{N} + H(X)\right| \geq \delta \qquad （5. 2.6）$$

因为信源是无记忆的，所以 $p(\vec{x}) = p(x_1)\cdots p(x_N)$ ，得到

$$\log p(\vec{x}) = \sum_{i=1}^{N} \log p(x_i) \qquad （5. 2. 7）$$

将（5. 2. 7）代入（5. 2. 6），得

$$\left|\frac{1}{N}\sum_{i=1}^{N} \log p(x_i) + H(X)\right| \geq \delta \qquad （5. 2. 8）$$

令 $z_i = \log p(x_i)$，可得 $E(z_i) = -H(X)$，所以

$$E\left\{\frac{1}{N}\sum_{i=1}^{N}\log p(x_i)\right\} = \frac{1}{N}\sum_{i=1}^{N}E(z_i) = -H(X)$$

根据**Chebyshev**不等式：$p\left\{\left|\xi - \bar{\xi}\right| > \delta\right\} \le \frac{Var(\xi)}{\delta^2}$，其中 ξ 为随机变量；这样就得到：

Var(x/N)=Var(x)/N^2

$$p_r\left\{\vec{z} : \left|\frac{1}{N}\sum_{i=1}^{N} z_i - \bar{z}\right| \ge \delta\right\} \le \frac{\sigma^2}{N\delta^2} \quad (5.2.9)$$

其中 $\vec{z} = (z_1, z_2, \cdots z_N)$，$\bar{z} = E(\frac{1}{N}\sum_{i=1}^{N} z_i)$，$\sigma^2 = Var(z_i)$

所以，

$$p_r\left\{\vec{x} : \left|\frac{\log p(\vec{x})}{N} + H(X)\right| \ge \delta\right\} \le \frac{\sigma^2}{N\delta^2} \quad (5.2.10)$$

其中，自信息的方差

$$\sigma^2 = Var\left[\log\ p(x_i)\right]$$

$$= E\left[\log^2 p(x_i)\right] - H^2(X) = \sum_{i=1}^{q} p_i \log^2 p_i - H^2(X) \quad (5.2.11)$$

取 $\frac{\sigma^2}{N_0\delta^2} = \varepsilon$，则当 $N > N_0$ 时，有 $\frac{\sigma^2}{N\delta^2} < \frac{\sigma^2}{N_0\delta^2} = \varepsilon$

总结

- 对离散无记忆信源，给定 $\varepsilon,\delta>0$，令 $N_0=\dfrac{\sigma^2}{\varepsilon\delta^2}$ 取 $N\geq N_0$；那么对长度为N的信源序列，满足下式的为典型序列，否则为非典型序列。

$$\{\vec{x}:\left|\frac{N_i}{N}-p_i\right|<\varsigma,\ i=1,\cdots,q\}$$

- 定理说明，当N足够大时，典型序列 $\vec{x}$ 的 $\dfrac{-\log p(\vec{x})}{N}$ 的值接近信源的熵
- 对于有记忆的马氏源，定理5.2.1也成立

渐进均分特性(AEP)

✓ 典型序列的概率估计

设 $\vec{x} \in G_1 \quad \Rightarrow -\delta < \frac{\log p(\vec{x})}{N} + H(X) < \delta$

$$\Rightarrow -N[H(X)+\delta] < \log p(\vec{x}) < -N[H(X)-\delta]$$

设取2为底 $\Rightarrow 2^{-N[H(X)+\delta]} < p(\vec{x}) < 2^{-N[H(X)-\delta]}$

简记为：

$$p(\vec{x}) = 2^{-N[H(X)\pm\delta]}$$

- 当 δ 足够小时，每个典型序列的概率 $p(\vec{x})$ 接近 $2^{-NH(X)}$ 其偏差不大于 $2^{N\delta}$ ；
- 此时序列的长度需要很大

✓ 典型序列的个数估计

设 N_G 为 G_1 中序列的个数

先估计上界：

利用概率估计的下界 $\Rightarrow N_G \cdot 2^{-N[H(X)+\delta]} < N_G \cdot \min_{\vec{x}} p(\vec{x}) \le 1$

$$N_G < 2^{N(H(X)+\delta)}$$

再估计下界：

利用概率估计的上界 $\Rightarrow 1-\varepsilon \le N_G \cdot \max_{\vec{x}} p(\vec{x}) < N_G \cdot 2^{-N[H(X)-\delta]}$

$$N_G > (1-\varepsilon)2^{N[H(X)-\delta]}$$

$$(1-\varepsilon)2^{N[H(X)-\delta]} < N_G < 2^{N[H(X)+\delta]}$$

✓ 渐近均分特性

当ε、δ取值很小时（N要求很大），对于典型序列

$$N_G \approx 2^{NH(X)}, \qquad p(\vec{x}) \approx 2^{-NH(X)}$$

含意：

当长度N足够大时：

- 典型序列接近等概率 $2^{-NH(X)}$，数目近似于 $2^{NH(X)}$
- 非典型序列出现的概率接近为零
- $-\frac{1}{N}\log p(\vec{x}) \to H(X)$ （以概率收敛）

结论

长度为N的信源序列

✓ 设信源序列数为 q^N，编码序列数为 r^l。如果每个信源序列都至少要有一个码字，即需要 $r^l \geq q^N$。但是，随着信源序列长度的增加，基本上是典型序列出现，这样我们仅考虑对典型序列的编码，所以实际需要 $r^l \geq 2^{NH(X)}$ 个码字。而当信源的熵小于 $\log_2 q$ 时，就会使得码字的长度减小。

有剩余度

§5.2.3 定长码信源编码定理

定理5.2.2

离散无记忆信源的熵为H(X)，码符号集的符号数为r，将长度为N的信源序列编成长度为l的定长码序列。只要满足：

定长码的码率

$$\frac{l}{N}\log r \geq H(X)+\delta$$

则当N足够大时，译码差错可以任意小（$<\varepsilon$）；

若上述不等式不满足，肯定会出现译码差错。

证明思路

✓ 正定理：

典型序列的个数小于 $2^{N[H(X)+\delta]}$

$$\frac{l}{N}\log r \geq H(X)+\delta \quad \Rightarrow \quad r^{l} \geq 2^{N[H(X)+\delta]}$$

l：长度为1码的个数

在编码时，可以使所有典型序列都有对应的码字，而最坏的情况是所有的非典型序列无对应的码字。

证明思路

✓ 逆定理：若不满足上式 ⇒ $\frac{l}{N}\log r < H(X)$

$$I(X^N;Y^l) = H(X^N) - H(X^N/Y^l) \le H(Y^l) \le lH(Y) = l\log r$$

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

H(Y^l)-H(Y^l|X^N)

$$H(X^N/Y^l) \ge NH(X) - l\log r > 0$$

上式表明：在已知编码序列的条件下，信源序列仍有不确定性，即不可能无失真译码。

相关定义

✓ 定长码编码速率（码率）

$$R' = \frac{l \log r}{N} \quad (\text{比特} / \text{信源符号})$$

- 它表示编码后，一个信源符号平均所携带的最大信息量，也可以理解为传送一个信源符号平均所需的比特数。

- 压缩码率实际就是减小编码速率。

编码效率

$$\eta = \frac{H(X)}{R'} = \frac{NH(X)}{l \log r}$$

- NH(X)表示N长信源序列的所包含的信息量
- $l\log r$表示码序列所能携带的最大信息量。
- 由定理5.3.2可知，对定长无失真编码η总是小于1；当N足够大时，η可以接近1
- 由渐近均分特性，当 R' 减小时(极限为熵)，η 增加。
- 压缩码率和提高编码效率是同样的含义。

- 信息传输速率：每个传输符号（码）所含信息量

$$R = \frac{NH(X)}{l} \quad (\text{比特/码符号})$$

$$\eta = \frac{R}{\log r}$$

相关结论

- 无失真信源信源编码的另一种表述

如果编码速率 $R' > H(X)$ ，则存在无失真编码。

反之，肯定有失真。

该表述对定长码和变长码都成立。

编码效率与熵的关系

$$\eta=\frac{H(X)}{H(X)+\delta}\Rightarrow\eta\,\delta+\eta\,H(X)=H(X)\qquad\Rightarrow\delta=\frac{1-\eta}{\eta}H(X)$$

码率不小于熵

$$N\geq\frac{\sigma^2}{\varepsilon\delta^2}=(\frac{\eta}{1-\eta})^2\cdot\frac{\sigma^2}{H^2(X)\varepsilon}$$

- 信源给定后，若要求编码效率越高，则信源序列长度 N 越大；同样，要求译码差错越低，N 值也越大。

长度为N的非典型序列概率小于

例 5.2.1 一离散无记忆信源的模型如下，要求用二元编码，如果 $\eta = 0.96$，$\varepsilon \le 10^{-5}$，估计信源序列的最小长度 N。

$$\begin{bmatrix} S \\ P(s) \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix}$$

解：信源的熵

$$H(S) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$$

自信息方差

$$\sigma^2 = \frac{3}{4}\log^2\frac{3}{4} + \frac{1}{4}\log^2\frac{1}{4} - 0.811^2 = 0.4715$$

$$N \ge \frac{\sigma^2}{\varepsilon\delta^2} = \left(\frac{\eta}{1-\eta}\right)^2 \cdot \frac{\sigma^2}{H^2(X)\varepsilon}$$

$$\Rightarrow N \ge \frac{0.96^2 \times 0.4715}{(1-0.96)^2 \times 0.811^2 \times 10^{-5}} = 4.13 \times 10^7$$

对于定长码，信源序列长的原因在于要求的编码效率过高

结论

对于定长码，要达到一定误码要求，信源序列长度需很长，所以编码器难于实现。

§5.3 变长码

本节主要内容

一、异前置码的性质

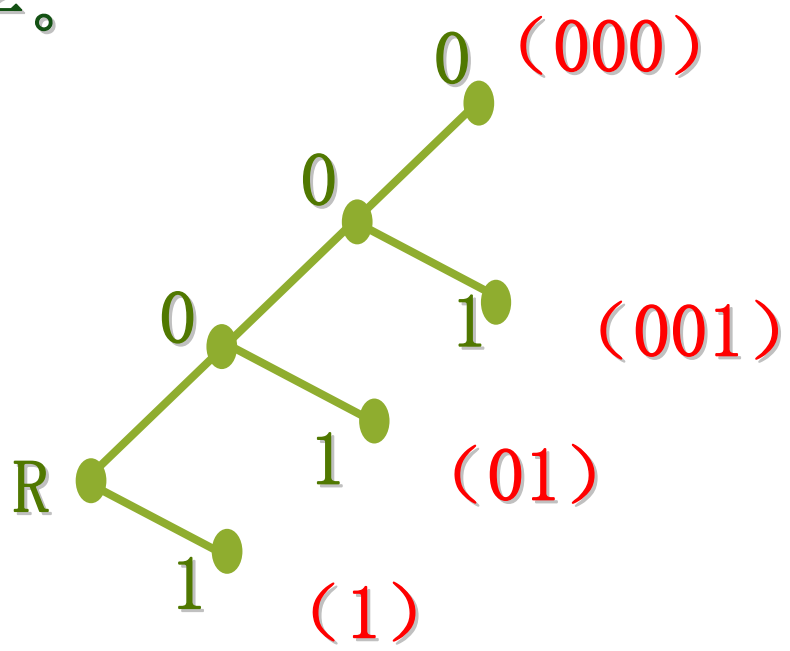
二、变长码信源编码定理

§5.3.1 异前置码的性质

✓ 变长码可用非全码树来描述。下图就是一个异前置码的码树。

✓ 只有端点（树叶）对应码字。

- ❖ 对应码字的端点与根之间不能有其它的节点作为码字，否则非异前置
- ❖ 端点不能向上延伸再构成新码字



定理5.3.1 (Kraft定理)

若信源符号数为q，码符号数为r，对信源符号进行编码，相应码长度为 $l_1, l_2, \cdots, l_q$，则异前置码存在的充要条件是：

$$\sum_{i=1}^{q} r^{-l_i} \leq 1 \quad \text{（克拉夫特不等式）}$$

只保证存在码长为l_1, l_2, ..., l_q的异前置码

证明思路

✓ 充分性： 做一个l_M阶全树，树叶总数 r^{l_M}

取 l_1 阶的任一节点作为第一个码字，去掉的树叶

$$r^{l_M-l_1} = r^{l_M} / r^{l_1}$$

$$r^{l_M-l_1}+r^{l_M-l_2}+\cdots+r^{l_M-l_q}$$

$$=r^{l_M}\left(r^{-l_1}+\cdots+r^{-l_q}\right)=r^{l_M}\sum_{i=1}^{q}r^{-l_i}\le r^{l_M}$$

1. 以第l_1阶的节点为根，对应的树叶为r^(l_M-l_1)，即去掉的树叶数目
2. 如果一种可行裁剪的非全码树和信源序列一一对应，则树叶对应的码字为异前置码
3. 即充分条件

以上的裁剪是为构造异前置码。去掉的总数叶不会超过全树叶数目，说明以上裁剪是可行的，或者指定1_1、1_2、...、1_q 后，树上有对应的树叶。

不等式成立—》说明存在对应树叶—》异前置码存在

证明思路

✓必要性：

构造一个码全树，最高阶为码字最大长度 l_M

对于阶为 l_k 的节点，占用的树叶数为 $r^{l_M - l_K}$

$$\sum_{K=1}^{q} r^{l_M - l_K} = r^{l_M} \sum_{K=1}^{q} r^{-l_K} \le r^{l_M}$$

1. 如果是异前置码，则对应码树是合法裁剪
2. 裁剪的树叶数目小于等于总数目

❖当码满足 Kraft 不等式时，未必就是异前置码

❖异前置码并不唯一，例如 0,1 交换。

只表明存在对应码长的异前置码

例 5.4.1 下表列出了3种变长码的编码，并给出了对应每个码的所有的码长 和具有同一码长的码字的个数，其中码符号集为{0, 1, 2, 3}。试问对每个码是否存在相应的异前置码？ r=4

码字个数 / 码 / 码长	码1	码2	码3
1	3	2	1
2	3	7	7
3	3	3	3
4	3	3	7
5	4	5	4

解： 利用 Kraft 不等式来验证。

对于码1： $3\times4^{-1}+3\times4^{-2}+3\times4^{-3}+3\times4^{-4}+4\times4^{-5}$

$$=3\left[\frac{1}{4}+(\frac{1}{4})^2+(\frac{1}{4})^3+(\frac{1}{4})^4+(\frac{1}{4})^5\right]+(\frac{1}{4})^5$$

$$=1$$

$\Rightarrow$ 存在相应的异前置码

同理： 码2不存在相应的异前置码；码3存在相应的异前置码。

实际上，可以用码树来验证，方法更简单。

定理5.3.2

若一个码是唯一可译码且码字长为 $l_1, l_2, \cdots, l_q$ 则必满足Kraft不等式，即：

$$\sum_{i=1}^{q} r^{-l_i} \leq 1$$

q：信源符号数

r：码符号数

证明略

注意

满足kraft不等式并不一定唯一可译，因为奇异码可能满足kraft不等式。

推论5.3.1

任意唯一可译码都可用异前置码代替，而不改变码字的任一长度。

说明：唯一可译码—》满足Kraft不等式—》存在以 l_i 码长的异前置码

§5.3.2 变长码信源编码定理

✓ 单信源符号编码的平均码长：

$$\bar{l} = \sum_{k=1}^{q} p_k l_k$$

表示平均每个信源符号所需码符号的个数

✓ 对于N次扩展源编码，原信源符号平均码长为

$$\bar{l} = \frac{1}{N} \sum_{k=1}^{q} p_k l_k$$

注：这里的q为长度为N的信源序列个数

定理5.3.3 单符号信源变长码编码定理

给定熵为H(X)的离散无记忆信源X，用r元码符号集对单信源符号进行编码， 则存在唯一可译码，其平均码长满足：

$$\frac{H(X)}{\log r} \le \bar{l} < \frac{H(X)}{\log r} + 1$$

证明思路

(1)证明不等式前半部

变长码码率》熵

$$H(X)-\bar{l}\log r=-\sum_i p_i\log p_i-\sum_i p_i l_i\log r$$

$$=\sum_i p_i\log\frac{1}{p_i r^{l_i}}\le\sum_i p_i\left(\frac{1}{p_i r^{l_i}}-1\right)\log e$$

$$=(\log e)\left(\sum_{i=1}^{q} r^{-l_i}-\sum_i p_i\right)\le 0$$

等于1，再利用Kraft不等式

注意用到凸函数性质-
对于任意正实数 x
有：$1-1/x\le\ln x\le x-1$

等式成立条件

$$\frac{1}{p_i r^{l_i}}-1=0 \qquad \text{即} \quad p_i=r^{-l_i}$$

最优编码

证明思路

(2)证明不等式后半部

1. 按下式定义，l_i是不小于信源符号i信息量的最小整数
2. 故有符号i的信息量大于l_i -1

$$\frac{1}{r^{l_i}} \le p_i < \frac{1}{r^{l_i-1}}$$

存在——》 $l_i = \lceil \log_r(1/p_i) \rceil \Rightarrow \begin{cases} l_i \ge \log_r(1/p_i) \Rightarrow p_i \ge \dfrac{1}{r^{l_i}} \\ l_i - 1 < \log_r(1/p_i) \Rightarrow p_i < \dfrac{1}{r^{l_i-1}} \end{cases}$

$$\sum_{i=1}^{q} p_i \log p_i < \sum_{i=1}^{q} p_i \log \frac{1}{r^{l_i-1}} = (\log r)\left(\sum_{i=1}^{q} p_i - \sum_{i=1}^{q} p_i l_i\right) = (1-\bar{l})\log r$$

logr^(1-l_i)

$$-H(X) < (1-\bar{l})\log r \Rightarrow \bar{l} < \frac{H(X)}{\log r} + 1$$

定理5.3.4 有限序列信源变长码编码定理

若对长度为N的离散无记忆信源序列进行编码，则存在唯一可译码，且使每信源符号平均码长满足：

$$\frac{H(X)}{\log r} \le \bar{l} < \frac{H(X)}{\log r} + \frac{1}{N}$$

且对任何唯一可译码左边不等式都要满足。

证明要点：（1）H（X^N）=NH（X）；（2）编码码长为$\bar{l}$N

定理5.3.5

对于离散平稳遍历马氏源，有：

$$\frac{H_\infty(X)}{\log r} \le \bar{l} < \frac{H_\infty(X)}{\log r} + \frac{1}{N}$$

证明略

还记得吗？
q个信源符号遍历马氏源 - 当时间足够大时，可以在任意时间转移到任意符号

定理5.3.6 香农第一定理

若对信源X的N次扩展X^N进行编码，当N足够大时，总能找到唯一可译的r进制编码，使得X的平均码长任意接近信源的熵，即有：

$$H_r(X) = H(X)/\log r$$

说明：$H_r(X)$为平均码长，$H_r(X)\log r$ 编码码率。码率—》熵

相当于以比特表示的平均码长,代码码率

相关定义

- 编码速率：$$R' = \bar{l} \log r$$
- 编码效率：$$\eta = \frac{H(X)}{R'} = \frac{H(X)}{\bar{l} \log r}$$
- 信息传输速率：$$R = \frac{H(X)}{\bar{l}}$$
- 编码剩余度：$$\gamma = 1 - \eta$$

一些结论

✓ 平均码长的上、下界

- ❖ $\bar{l} \geq \frac{H(X)}{\log r}$ ⇒ 对所有唯一可译码都要满足
- ❖ $\bar{l} < \frac{H(X)}{\log r} + 1$ ⇒ 无需一定满足，但存在这种关系，通常希望越小越好

✓ $\bar{l} = \frac{H(X)}{\log r}$ 时，$\eta = 1$，此时：~~→~~ 最佳编码

- ❖ 各信源符号出现概率为 $p_i = (1/r)^{l_i}$ l_i为整数
- ❖ 每码元平均所带信息量为 $\frac{H(X)}{\bar{l}} = \log r$, 码元符号独立且等概

例 5.3.2 用例5.2.1的信源模型，i）对单信源符号进行二元编码，即$s_1 \to 0$，$s_2 \to 1$，求平均码长和编码效率；ii）编成2次扩展码，信源序列与码序列的映射关系为：

$$s_1s_1 \to 0, s_1s_2 \to 10, s_2s_1 \to 110, s_2s_2 \to 111$$

求平均码长和编码效率。

解：1）$\overline{l} = 1$，$\eta = \dfrac{H(X)}{\overline{l}} = 0.811$

此时$\log_2 2=1$

2）信源序列的概率：

$$p(s_1s_1)=(3/4)\times(3/4)=9/16 \quad p(s_1s_2)=(3/4)\times(1/4)=3/16$$

$$p(s_2s_1)=(1/4)\times(3/4)=3/16 \quad p(s_2s_2)=(1/4)\times(1/4)=1/16$$

且：$\bar{l}=(1\times 9/16+2\times 3/16+3\times 3/16+3\times 1/16)/2$

$$=27/32$$

可以看出：随着信源符号扩展长度N的增加，平均码长趋小，即编码效率提高

$$\Rightarrow \eta=\frac{H(X)}{\bar{l}}=0.811/(27/32)=0.961$$

与例5.2.1相比，可以看出，为得到同样编码效率所用信源序列长度N比定长码小得多。 因此容易达到高的编码效率，是变长码的显著优点。

§5.4 哈夫曼编码

本节主要内容

- 一、二元哈夫曼编码
- 二、多元哈夫曼编码
- 三、马氏源的编码

§5.4.1 二元哈夫曼编码

若一个唯一可译码的平均码长小于所有其它唯一可译码，则称该码为最优码(或紧致码)。

应注意：最优是唯一可译码之间的比较，因此它的平均码长未必达到编码定理的下界。

条件： p_i=2^(-l_i)

定理5.4.1

这里的信源符号要广义理解 - 可以是单信源符号，也可以是信源符号序列

对于任意一个含q个符号的信源，存在最优的二进制码，其中有两个最长的码字有相同的长度且仅最后一个码位有别，即其中一个的最末尾是0，而另一个的最末尾是1（或者相反）

code —》 codewords

证明思路（1）：

- 首先证明对于最优码，概率小的符号对应长度长的码字。
  - 如果存在一对信源符号，其中较大概率符号对应较长的码字，将其码字对换，可以证明平均码长会减小。

证明思路（2）：

- 证明最长的码字有两个长度相同，且只有最后一位不同。

一个最优码唯一可译 ⇒ 满足Kraft不等式 ⇒ 存在与其同样码长的异前置码

假定x为最优异前置码中最长的码之一，则必存在一个长度相同但末位与其不同的码字；否则可去掉x最后一位，不违反异前置，但平均码长减小，和最优码矛盾

思考一下：是否可能出现长度相同，不止最后一位不同的2个码字？

二元最优异前置码的构造方法

- 设信源S为 $p(a_1) \geq \ldots \geq p(a_q)$ ，对应的码字为 $\vec{x}_1,\ldots,\vec{x}_q$
- 将概率最小的两个码符号 a_{q-1}, a_q 合并，从而产生新的信源S‘ $\{a'_1,\ldots,a'_{q-1}\}$
- 设 $\{a'_1,\ldots,a'_{q-1}\}$，对应的码字为 $\vec{x}'_1, \vec{x}'_2,\ldots,\vec{x}'_{q-1}$ 。对新信源编码后，按下面的关系就可恢复原来信源的码字：

$$\vec{x}_i = \vec{x}'_i, \quad i = 1,\ldots,q-2$$

$$\vec{x}_{q-1} = \vec{x}'_{q-1} + "0"$$

$$\vec{x}_q = \vec{x}'_{q-1} + "1"$$

若 $\vec{x}_i'$ 对信源 S' 是最优的异前置码，则 $\vec{x}_i$ 对信源 S 也是最优的异前置码

证明思路：

设 $S' \to l_1', \cdots, l_{q-1}'$　$S \to l_1, \cdots, l_q$　$\Rightarrow l_i = \begin{cases} l_i', & 1 \le i \le q\text{-}2 \\ l_{q-1}' + 1, & i = q-1, q \end{cases}$

对S，有 $\bar{l} = \sum_{i=1}^{q} p_i l_i = \sum_{i=1}^{q-2} p_i' l_i' + p_{q-1} l_{q-1} + p_q l_q$

$$= \sum_{i=1}^{q-2} p_i' l_i' + (p_{q-1} + p_q) l_{q-1}' + p_{q-1} + p_q$$

$$= \bar{l}' + p_{q-1} + p_q$$

注意：p_(q-1)+p_q为常数，l'最优，则最优l

一些结论

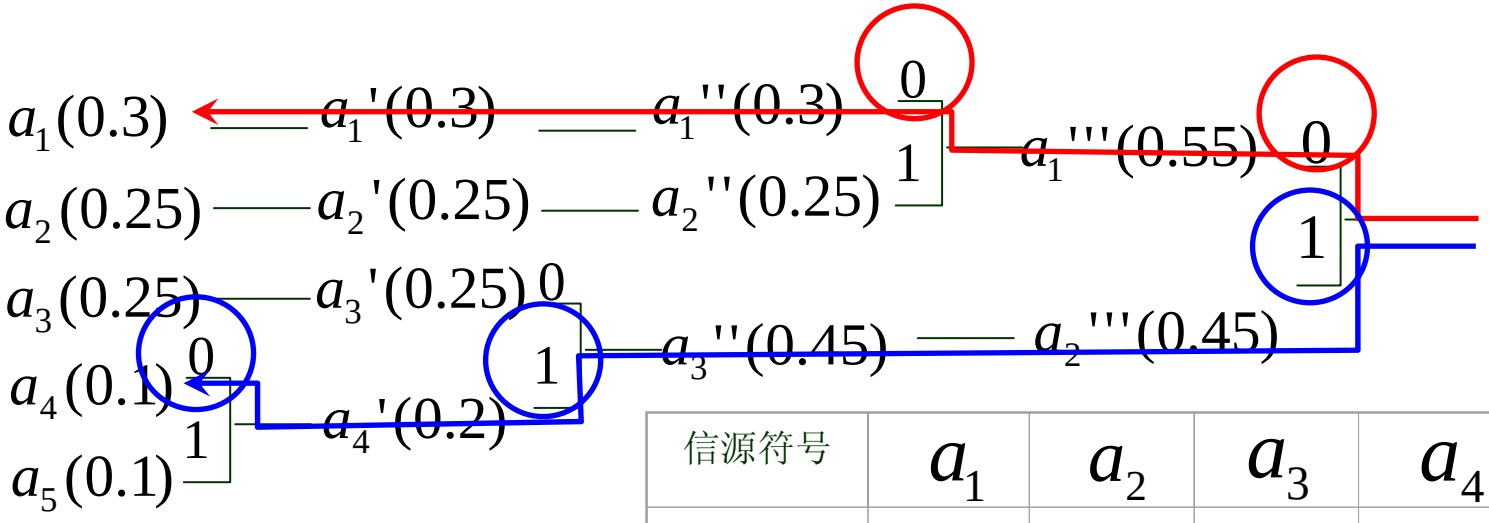
✓ 我们可以采用合并两个最小概率符号的方法，逐步地按这样的路线去编码：

$$S \rightarrow S' \rightarrow S'' \rightarrow \ldots \rightarrow 2\text{字母信源}$$

最后将2字母信源分配0、1符号；然后可逐步反推到原信源S，从而得到信源的最优编码。这种编码称做二元Huffman编码

例 5.4.1 一信源S的符号集A= $\{a_1, a_2, a_3, a_4, a_5\}$，概率分别为：0.3，0.25，0.25，0.1，0.1；试对信源符号进行二元Huffman编码

解：依次做信源S'S''S'''，最后将0、1符号分配给S'''，如下图：



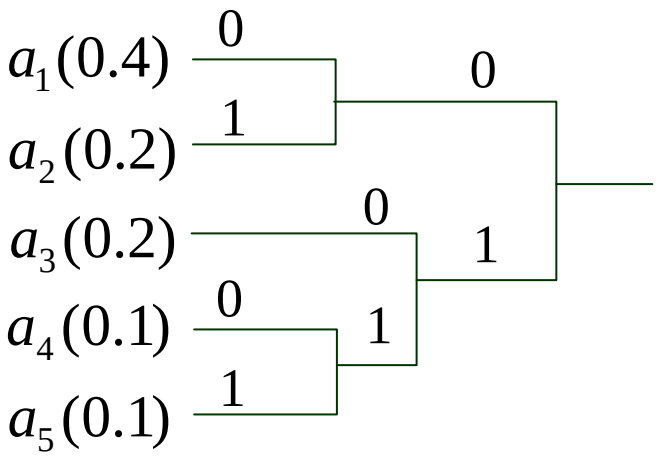
信源符号	a_1	a_2	a_3	a_4	a_5
码字	00	01	10	110	111

Huffman编码方法

- 将信源概率分布按大小依递减次序排列；合并两概率最小者，得到信新源；并分配 0,1 符号
- 新信源若包含两个以上符号返回（1）， 否则到（3）
- 从最后一级向前按顺序写出每信源符号所对应的码字

例 5.4.2 一信源S的符号集A= $\{a_1, a_2, a_3, a_4, a_5\}$，概率分别为：0.4，0.2，0.2，0.1，0.1；试对信源符号进行二元Huffman编码，并计算平均码长和编码效率

解：



a_1	a_2	a_3	a_4	a_5
00	01	10	110	111

$$\bar{l} = 0.4\times 2 + 0.2\times 2\times 2 + 0.1\times 3\times 2$$
$$= 2.2(\text{码元/信源符号})$$

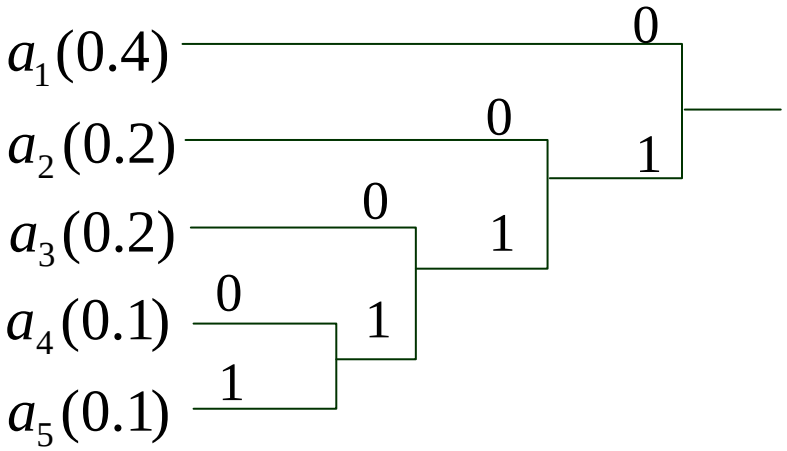
$$H(S) = -0.4\log 0.4 - (0.2\log 0.2)\times 2$$
$$-(0.1\log 0.1)\times 2 = 2.122$$

$$\eta = \frac{H(S)}{\bar{l}} = \frac{2.12}{2.2} = 0.965$$

一些结论

- ✓ Huffman编码是最优码（或紧致码），是异前置码
- ✓ 编码结果并不唯一，例如0、1可换，相同概率符号码字可换，但平均码长不变
- ✓ 不一定达到编码定理下界，达下界条件为 $p_i = 2^{-l_i}$
- ✓ 通常适用于多元信源，对于二元信源，必须采用合并符号的方法，才能得到较高的编码效率

例 例5.4.2还可以用以下的方法编码：



a_1	a_2	a_3	a_4	a_5
0	10	110	1110	1111

$$\bar{l} = 0.4\times1+0.2\times2+0.2\times3+0.1\times4\times2 = 2.2(\text{码元/信源符号})$$ 不变

但码长的方差改变了

$$\sigma_1^2 = \sum_i p_i l_i^2 - (\bar{l})^2 = 0.4\times2^2+0.2\times2^2+0.2\times2^2+0.1\times3^2+0.1\times3^2-2.2^2 = 0.16$$

$$\sigma_2^2 = 0.4\times1+0.2\times2^2+0.2\times3^2+0.1\times4^2+0.1\times4^2-2.2^2 = 1.36$$

一些结论

当码长的方差小时，编码器所需缓冲器容量小。因此要尽量减小码长的方差。

方法是：在编码时，应使合并后的信源符号位于缩减信源符号尽可能高的位置上（减少合并次数）。

§5.4.2 多元哈夫曼编码

通过观察可知，要使编码的平均码长最短，对应的码树要构成满树是必要条件。对于r元哈霍夫曼编码，从第n阶的1个节点到n+1阶节点，增加的数目为r-1。因此，达到满树时，总的树叶数为：

$$s = r + (r-1)m$$

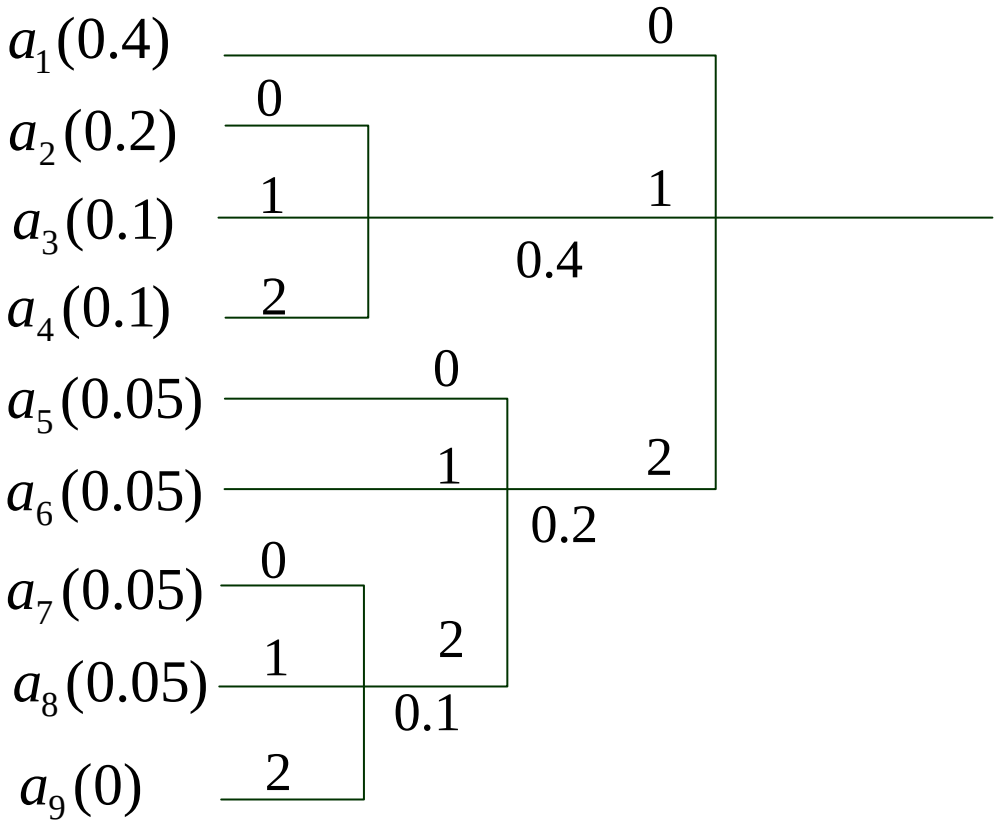
→ 非负整数

否则，就利用上式计算出大于q的最小正整数s。然后给信源增补零概率符号，使增补后的信源符号总数为s。编码后，去掉这些零概率符号所对应的码字，其余码字为所需码字

例 5.4.3 一信源S的符号集A= $\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$ 概率分别为：0.4, 0.2, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05 ；试对信源符号进行3元哈夫曼编码，并计算平均码长和编码效率

解：$\left.\begin{array}{l}\bar{l}=1.7(\text{码元/信源符号})\\ H(X)=2.522\text{比特/符号}\end{array}\right\}\Rightarrow \eta=\dfrac{H(X)}{\bar{l}\log r}=\dfrac{2.522}{1.7\times\log 3}=0.936$

$\left.\begin{array}{l}r=3\\ m=3\end{array}\right\}\Rightarrow s=3+2m=9\left.\begin{array}{r}\\ q=8\end{array}\right\}\Rightarrow$ 信源要增加1个零概率符号



$a_1(0.4) \rightarrow 0$

$a_2(0.2) \rightarrow 10$

$a_3(0.1) \rightarrow 11$

$a_4(0.1) \rightarrow 12$

$a_5(0.05) \rightarrow 20$

$a_6(0.05) \rightarrow 21$

$a_7(0.05) \rightarrow 220$

$a_8(0.05) \rightarrow 221$

Huffman编码的应用

决策树

如果有n个互斥随机事件，概率分别为p_i，现用某种测试方法分步对所选择的目标事件进行识别，要求具有最小的决策平均次数，相当于对这些事件进行Huffman编码。

决策树举例

例如，甲手中有4张纸牌，点数分别为1、2、3、4，要求乙猜：乙可以向甲提问题，甲只能用是否来回答。求乙平均最少问几个问题可以猜到纸牌的点数和相应的策略。

(1)1、2、3、4的概率均为1/4的决策树；

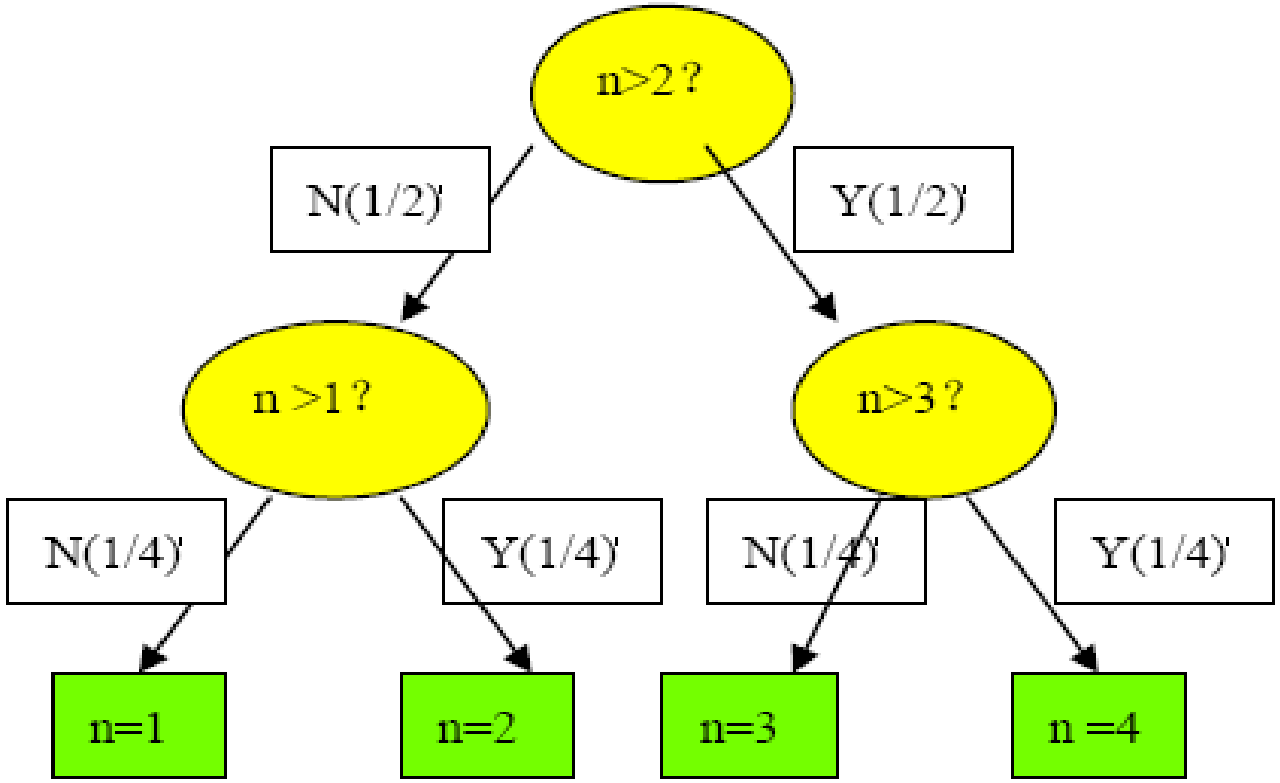
(2)1、2、3、4的概率分别为1/2、1/4、1/8、1/8的决策树.

❖首先Huffman编码；

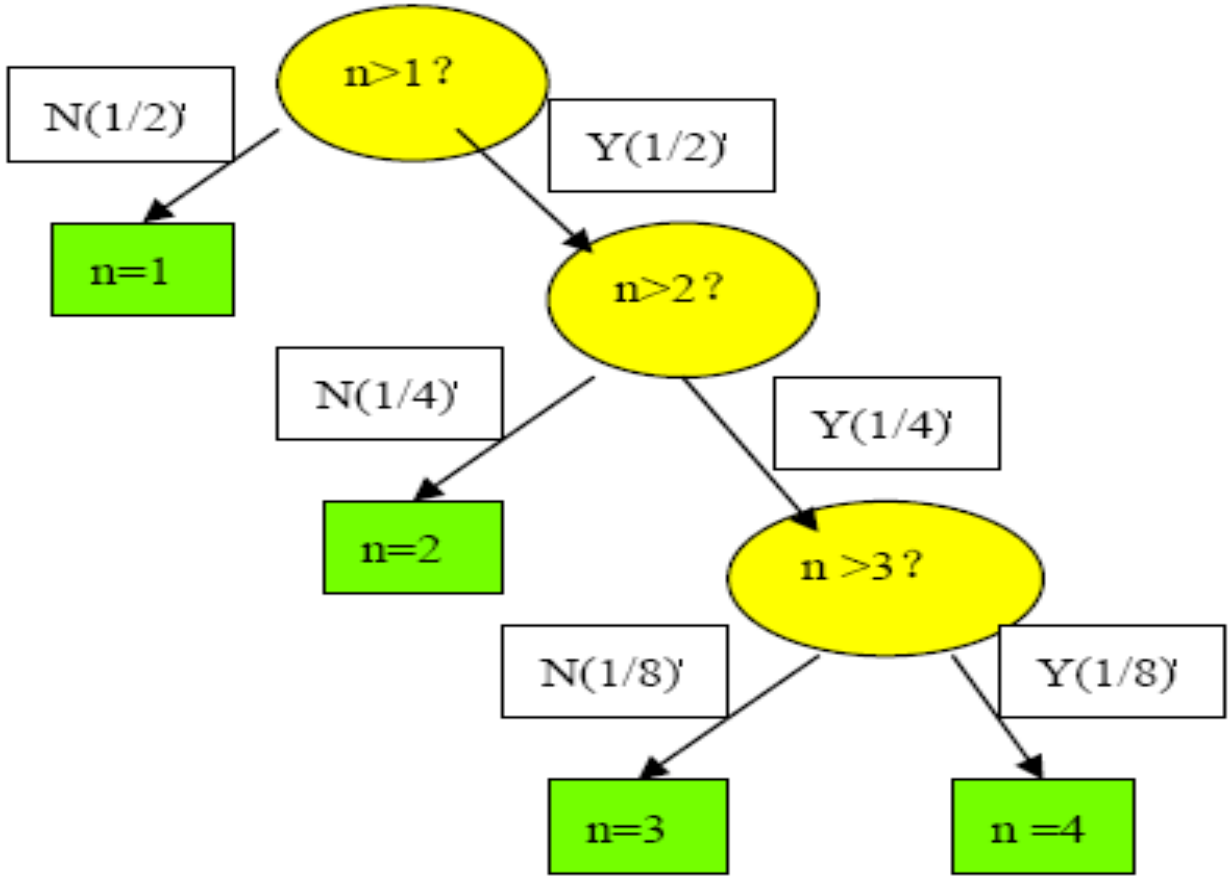
❖Huffman编码码树变成决策树。

❖决策的设计：每步决策结果应该与节点分支的概率匹配。

第(1)问



第(2)问



§5.4.3 马氏源的编码

马氏源可以采用按状态编码和多个符号合并编码

按状态编码

根据马氏源的特性，当前发出的符号所含信息量取决于当前的状态。这个信息量可能很大也可能很小。例如，一个马氏源包含3个状态{a，b，c}，每个状态代表一个输出符号，状态转移矩阵如下：

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

下一个字母b. c出现等概

包含的信息量最大

下一个字母必然出现b

包含的信息量为0

因此，采用一个符号用一个码字代替的方法会使编码效率降低

按状态编码方法

1. 给定一个初始状态 S_0

2. 对每个状态 S, 根据转移概率 $p(a_i \mid s=j), i=0,1,...,q-1$

进行最优编码，例如 Huffman编码.

3. 设 $C_j(j=0,1,..,J-1)$ 为对应的码表, 其中规定

信源符号 a_i 和码字 $y_i^{(j)}$ 的对应关系，记为

$C_j(a_i, y_i^{(j)})$

编码过程

1. 给定一信源序列 $x_0x_1\cdots x_n\cdots$ ，设初始状态 s_0

2. 用 C_{s_0} 码表，查出 $x_0 = a_{i_0}$ 对应的码字 $y_{i_0}^{(s_0)}$ 作为编码器输出，同时根据 s_0, x_0 得到下一个状态 s_1

3. 如此重复，直到处理完最后一个信源符号 x_n

4. 编码器输出为 $y_{i_0}^{(s_0)}, y_{i_1}^{(s_1)}, \ldots, y_{i_n}^{(s_n)}$

译码过程

1. 根据译码器初始状态 s_0 ， 用 C_{s_0} 码表查出其中的码字与序列 $b_0 b_1 \ldots b_m$ 的前缀的相同部分，设 $b_0 b_1 \ldots b_{k_0} = y_{i_0}^{(s_0)}$，则 $y_{i_0}^{(s_0)}$ 对应的 a_{i_0} 为译码器的输出

2. 根据 s_0 和 a_{i_0} 确定下一个状态，设为 s_1 ，则找到 C_{s_1} 码表中的码字与序列 $b_{k_0+1} b_{k_0+2} \ldots b_m$ 中的前缀相同的部分，设 $b_{k_0+1} b_{k_0+2} \ldots b_{k_1} = y_{i_1}^{(s_1)}$，则 $y_{i_1}^{(s_1)}$ 对应的 a_{i_1} 为译码器的输出

3. 如此重复，直到最后一个序列符号处理完。

例 5.4.4 对状态转移矩阵如下的马氏源进行哈夫曼编码，并计算编码效率。

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

解：在3个状态下的Huffman编码如下

编码 / 状态 / 符号	a	b	c
a	——	10	——
b	0	0	——
c	1	11	——

先求平稳分布 $[\pi_a \quad \pi_b \quad \pi_c]\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}=[\pi_a \quad \pi_b \quad \pi_c]$

得到 $[\pi_a \pi_b \pi_c] = \left[\frac{2}{13} \frac{8}{13} \frac{3}{13}\right]$

平均码长 $\bar{l} = \sum_i \pi_i \bar{l}_i$，$\pi_i$为平稳分布的概率，$\bar{l}_i$ 为在

每一个状态编 码的平均码长

$$\bar{l}_a = 1,\ \bar{l}_b = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 \times 2 = 1.5,\ \bar{l}_c = 0$$

$$\bar{l} = \frac{2}{13} \times 1 + \frac{8}{13} \times 1.5 = \frac{14}{13}$$

信源的熵 $H(X) = \frac{2}{13} \times 1 + \frac{8}{13} \times 1.5 = \frac{14}{13}$ 比特

编码效率 $\eta = \frac{H(X)}{\bar{l}} = \frac{14/13}{14/13} = 1$

如果利用平稳分布编码结果为：a:11， b:0， c:10

$$\bar{l} = \frac{2}{13} \times 2 + \frac{8}{13} \times 1 + \frac{3}{13} \times 2 = \frac{18}{13}$$

$$\eta = \frac{H(X)}{\bar{l}} = \frac{14/13}{18/13} = \frac{7}{9} = 78\%$$

用状态编码比利用平稳分布编码效率高。

- 把信源的 N个符号合并成一个新符号再编码，会使编码效率提高。游程编码可使一阶马氏链的游程序列变为独立序列。

§本章小结

✓ 信源编码的主要目的是提高信息传输的有效性，分为如下几类

- ❖ 概率匹配编码（信源符号概率已知）
  - 分组码： 定长码， 变长码
  - 非分组码
- ❖ 通用编码（信源符号概率未知）

§本章小结

✓ 信源序列渐近均分特性

- ❖典型序列的概率 $p(\vec{x}) = 2^{-N[H(X)\pm\delta]}$，个数 $N_G \approx 2^{NH(X)}$，
- ❖当序列长度N足够大时，有 $\left|\frac{1}{N}\log p(\vec{x}) + H(X)\right| < \delta$

✓ 唯一可译码必须满足Kraft不等式

$$\sum_{i=1}^{q} r^{-l_i} \leq 1$$

§本章小结

✓ 无失真信源编码定理（香农第一定理）

若对信源X的N次扩展源 X^N 进行编码，当N足够大时，总能找到唯一可译的r进编码，使得X的平均码长任意接近信源的熵

§ 本章小结

✓ 关于信源编码定理的另一种描述

只要编码后，信息传输速率不大于无噪声信道的容量，就可实现无失真信源编码。

$$\eta = \frac{H}{\bar{l}\log r} = \frac{R}{C} \le 1$$

- $R = \frac{H}{\bar{l}}$ 为编码信道信息传输速率
- $C = \log r$ 为无噪信道的容量

✓编码序列的特性

R的最大限制：$R \le \log r$

$R = \log r \Rightarrow$ 编码符号独立且编码符号等概率

✓无失真信源编码所采取的主要措施

(1)概率匹配（Huffman编码等）使编码符号等概率

(2)解除相关性，使信源变成无记忆

✓无失真信源编码的限制

◆典型序列个数估计 $N_G = 2^{NH(X)}$ ，若 $H(X) = \log_2 q$ 则 $N_G = q^N$ ，即每个序列都是典型序列。要实现无失真，必须有 $r^l \geq q^N$ 。与无编码情况一样。

◆当信源的熵接近 $\log_2 q$ 时，无失真信源编码的意义不大；此时信源冗余度 $r = 1 - \eta = 1 - \frac{H(X)}{\log q} = 0$ ，没有压缩的余地。

信源符号独立等概

✓ 信源压缩编码的下限

◆不采用信源编码时每信源符号的码长为 $\log q$ ，而通过压缩编码后的平均码长会减小，但大于等于

$$H(X)/\log r = H_r(X)$$

◆压缩编码的目的就是尽量降低传送每个信源符号时所需的比特数，而信源的熵H(X)为无失真压缩码长的下界

✓ 几种重要编码技术

- Huffman编码
- 算术编码
- 游程编码
- L-Z编码

课后习题

- **P.102**

  5.1, 5.4, 5.6, 5.7

Thank you!