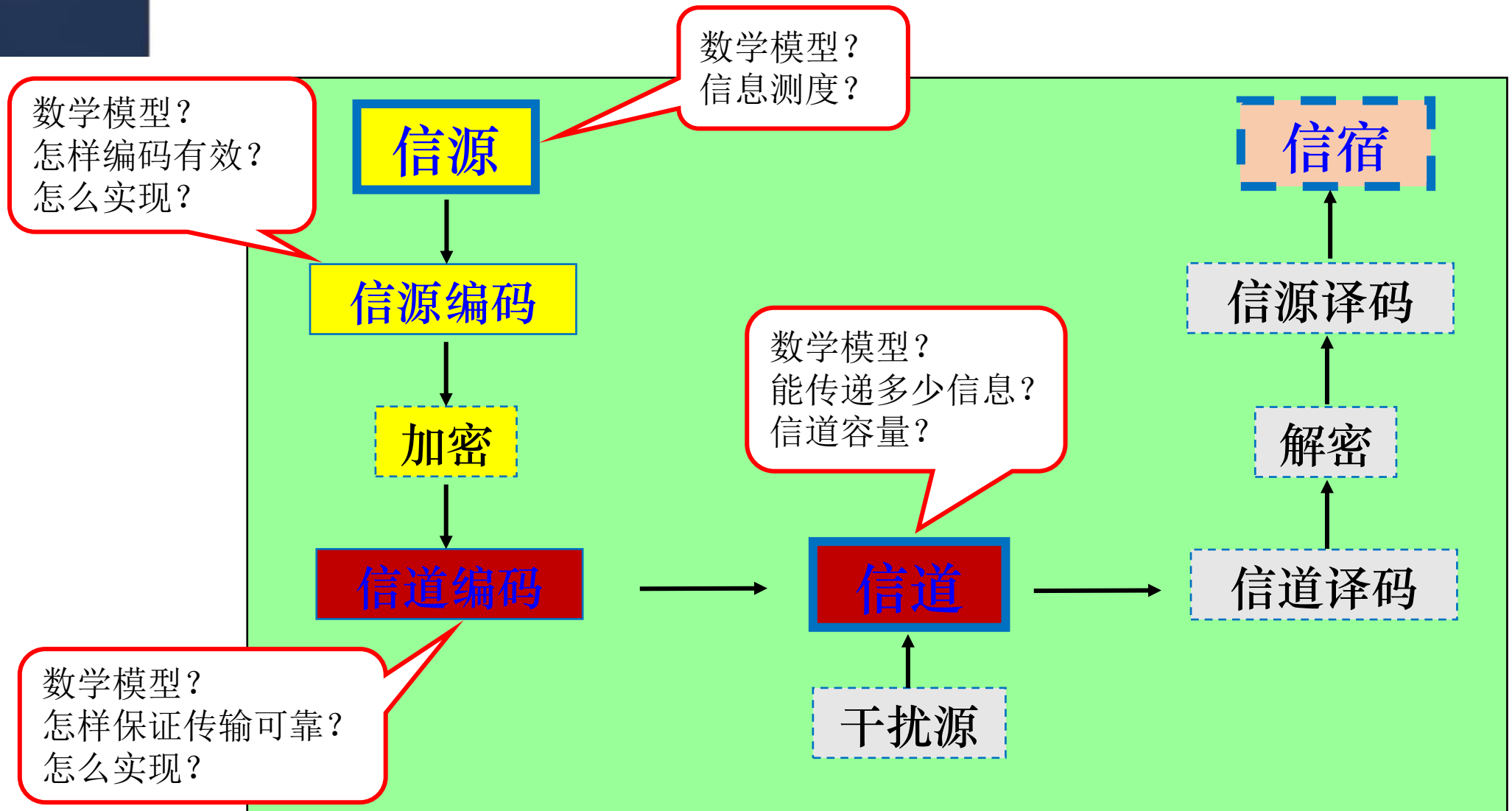
第 3 章　无失真离散信源编码

计算机科学与技术学院　孙伟平

基本内容

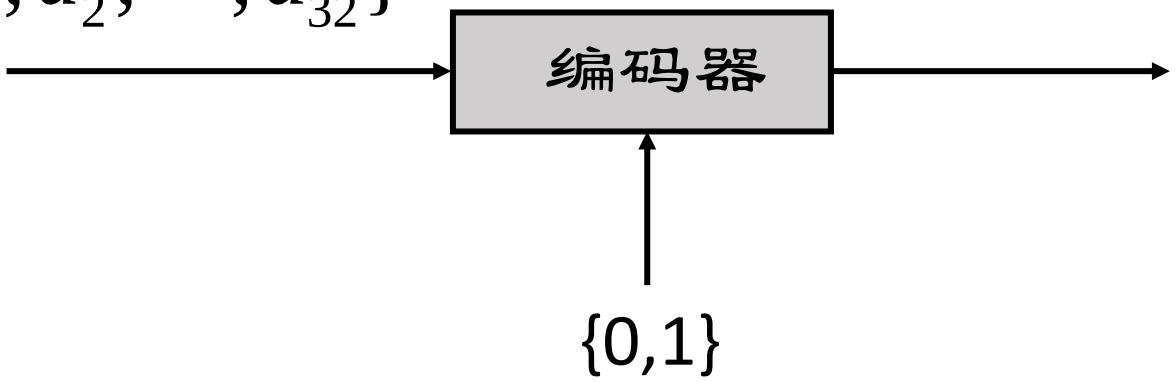
在信息可以量度的基础上，研究有效地和可靠地传递信息的科学。



第3章

例：

$$S=\{a_1,a_2,\cdots,a_{32}\}$$



编码一（电传码）： $A-11100, B-10011,...$

编码二（ASCII码）：$A-31\mathrm{H}, B-32\mathrm{H},...$

- **信源编码的任务：**
  - 使信源适合于信道的传输，用信道能传输的符号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。
- **信源编码的目的：提高通信的有效性**
  - 通常通过压缩信源的冗余度来实现。
  - 冗余度取决于符号间记忆的相关性与符号概率分布的非均匀性。
  - 压缩方式：

    概率匹配---统计编码（Huffman编码，算术编码）

    去除码符号间的相关性，再对各独立分量进行编码（变换编码）

    利用条件概率进行编码（预测编码）

    利用联合概率进行编码（无记忆信源的扩展编码）

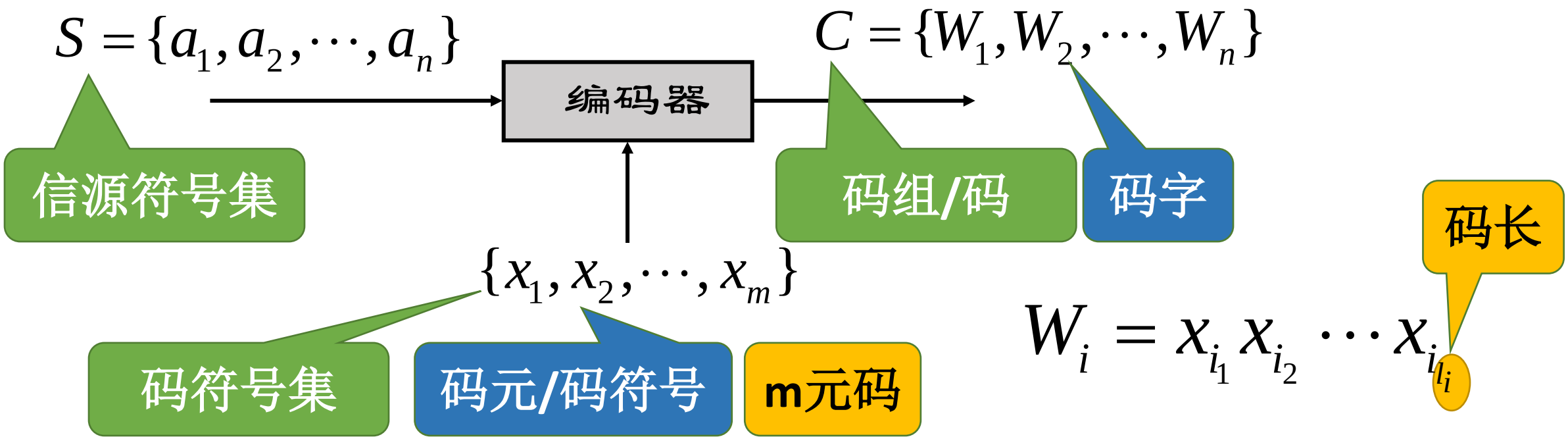
信源编码理论基础：

- 无失真信源编码定理
  - 统计编码：根据信源概率分布选用与之匹配的编码
- 限失真信源编码定理

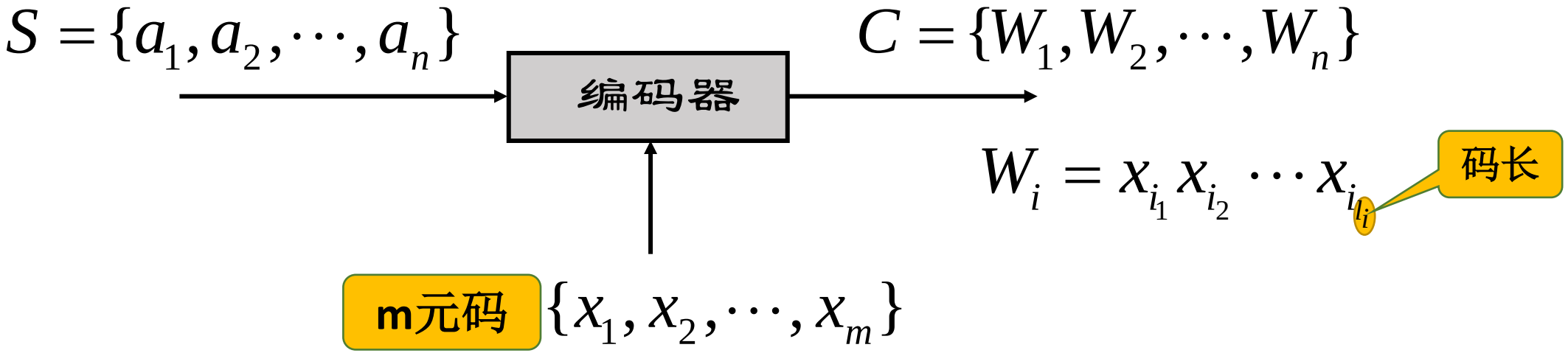
- 无失真信源编码：编码时没有信息丢失，译码器可以精确恢复编码之前的消息。又叫“无损压缩”。
- 无失真信源编码只适用于离散信源。
- 无失真信源编码的基本问题是研究如何用最少的比特数去表示离散信源的熵值，也就是如何找出最佳编码方案。

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。



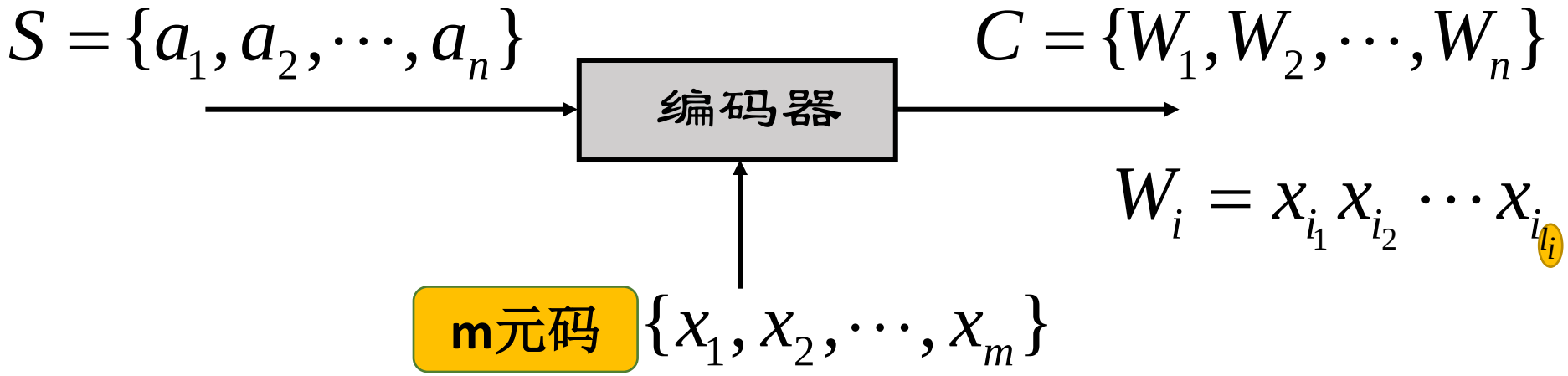
信源编码的过程是将信源符号集中的消息（符号）a_i（或长为N的符号序列 α_i）映射成由码符号 x_j 组成的长度为 l_i 的一个对应的码字 W_i。

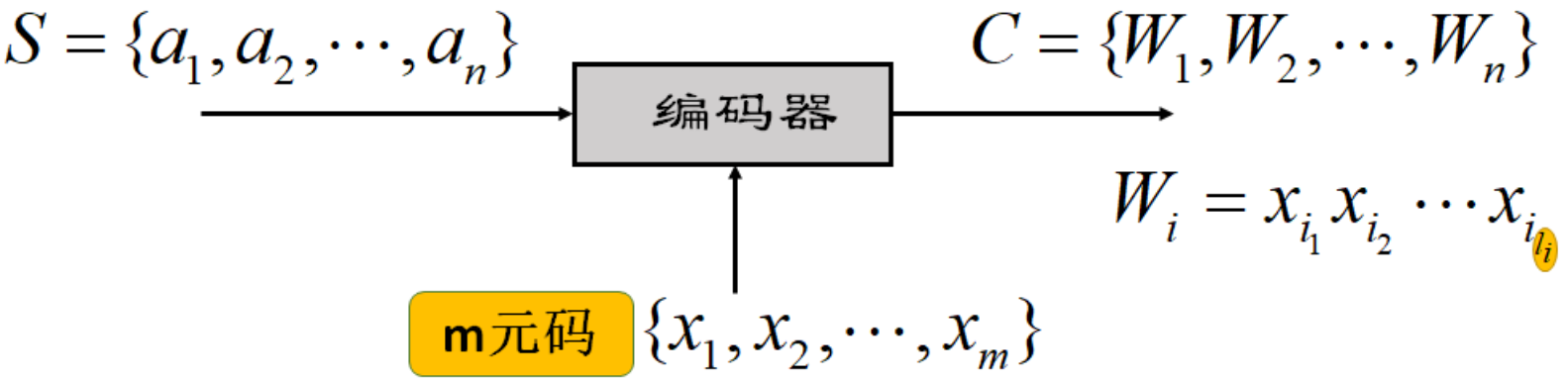


码的种类

- 定长码：码中所有码字的码长相同
- 变长码：码中码字的码长不相同

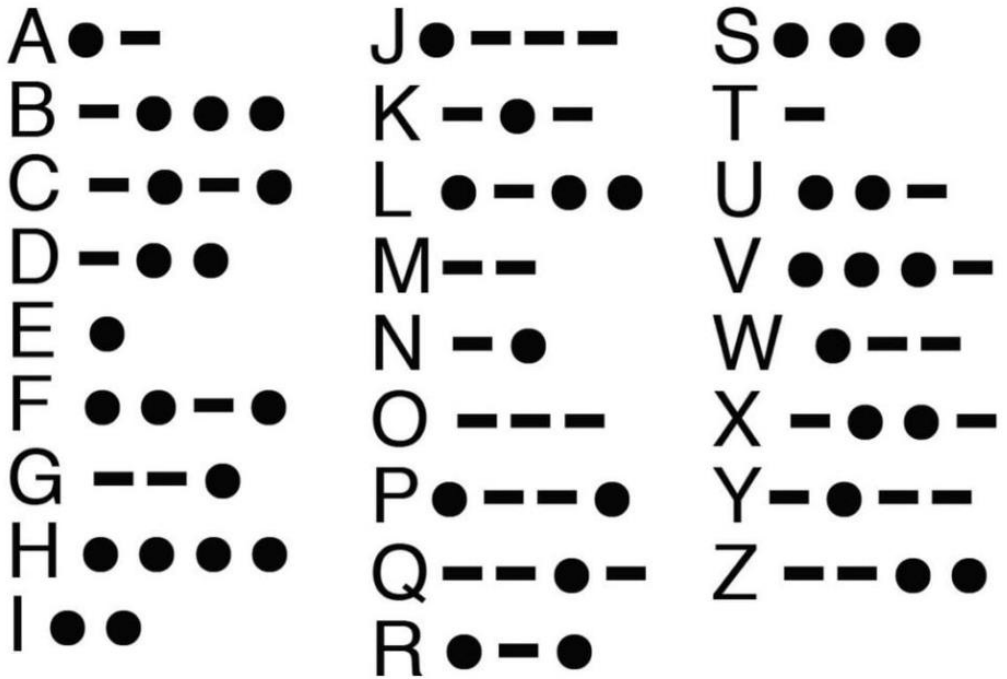
- 奇异码：码中存在相同的码字
- 非奇异码：码中所有码字都不相同


$S=\{a_1,a_2,\cdots,a_n\}$
编码器
$C=\{W_1,W_2,\cdots,W_n\}$
$W_i=x_{i_1}x_{i_2}\cdots x_{i_{l_i}}$
m元码
$\{x_1,x_2,\cdots,x_m\}$


$S=\{a_1,a_2,\cdots,a_n\}$
编码器
$C=\{W_1,W_2,\cdots,W_n\}$
$W_i=x_{i_1}x_{i_2}\cdots x_{i_{l_i}}$
m元码
$\{x_1,x_2,\cdots,x_m\}$

例　变长码-Moose电码

- 发明于1837年，是一种时通时断的信号代码，通过不同的排列顺序来表达不同的英文字母、数字和标点符号。
- 摩斯电码是一种早期的数字化通信形式。它不同于现代只使用0和1两种状态的二进制代码，它的代码包括五种：点、划、点和划之间的停顿、每个字符间短的停顿（在点和划之间）、每个词之间中等的停顿以及句子之间长的停顿。



信源编码的过程是将信源符号集中的消息（符号）a_i（或长为N的符号序列 α_i）映射成由码符号 x_j 组成的长度为 l_i 的一个对应的码字 W_i。

基本源编码：

信源符号集 $S=\{a_1,a_2,\cdots,a_n\}$ $\longleftrightarrow$ 码组 $C=\{W_1,W_2,\cdots,W_n\}$

$W_i=x_{i_1}x_{i_2}\cdots x_{i_{l_i}}$

扩展码源编码： ↓ N次扩展（无记忆）

$S^N=\{\alpha_1,\alpha_2,\cdots,\alpha_{n^N}\}$ $\longleftrightarrow$ $C=\{W_1,W_2,\cdots,W_{n^N}\}$

$\alpha_i=a_{i_1}a_{i_2}\cdots a_{i_N}$ $\qquad$ $W_i=x_{i_1}x_{i_2}\cdots x_{i_{l_i}}$

符号序列 $\qquad$ 码元序列

例　对二元离散无记忆信源（DMS）进行无失真编码：

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

用二元码元 $\{0,1\}$ **对其编码：** $s_1 \to 0, \quad s_2 \to 1$

对其进行二次扩展信源编码：

α_i	$p(\alpha_i)$	码字
$s_1 s_1$	$\frac{9}{16}$	0
$s_1 s_2$	$\frac{3}{16}$	10
$s_2 s_1$	$\frac{3}{16}$	110
$s_2 s_2$	$\frac{1}{16}$	111

例　某信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} a_1 \ , \quad a_2 \ , \quad a_3 \ , \quad a_4 \\ p(a_1), p(a_2), p(a_3), p(a_4) \end{Bmatrix}$$

对其进行编码，指出其中的定长码、变长码、奇异码、非奇异码。

信源符号	符号概率	码1	码2	码3
a_1	$p(a_1)=\frac{1}{2}$	00	0	0
a_2	$p(a_2)=\frac{1}{4}$	01	01	11
a_3	$p(a_3)=\frac{1}{8}$	10	001	00
a_4	$p(a_4)=\frac{1}{8}$	11	111	11

唯一可译性

译码：从接收到的码字序列得到信源符号序列的过程。

唯一可译码：任何一串有限长的码序列符号只能被唯一地译为对应的信源符号序列。

唯一可译码充要条件：编码任意次扩展均为非奇异码。

- **码字与信源符号一一对应**
- **码字符号序列与信源符号序列一一对应**

唯一可译码的判断

- 奇异码一定不是唯一可译码

a_1	0
a_2	11
a_3	00
a_4	11

- 非奇异码不一定是唯一可译码

a_1	0
a_2	10
a_3	00
a_4	11

- 等长非奇异码一定是唯一可译码
- 唯一可译码可分为即时码和非即时码

即时码：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。也称异前置码。

信源符号	码1	码2
a_1	1	1
a_2	10	01
a_3	100	001
a_4	1000	0001

给定码字　１１０１００１０００１

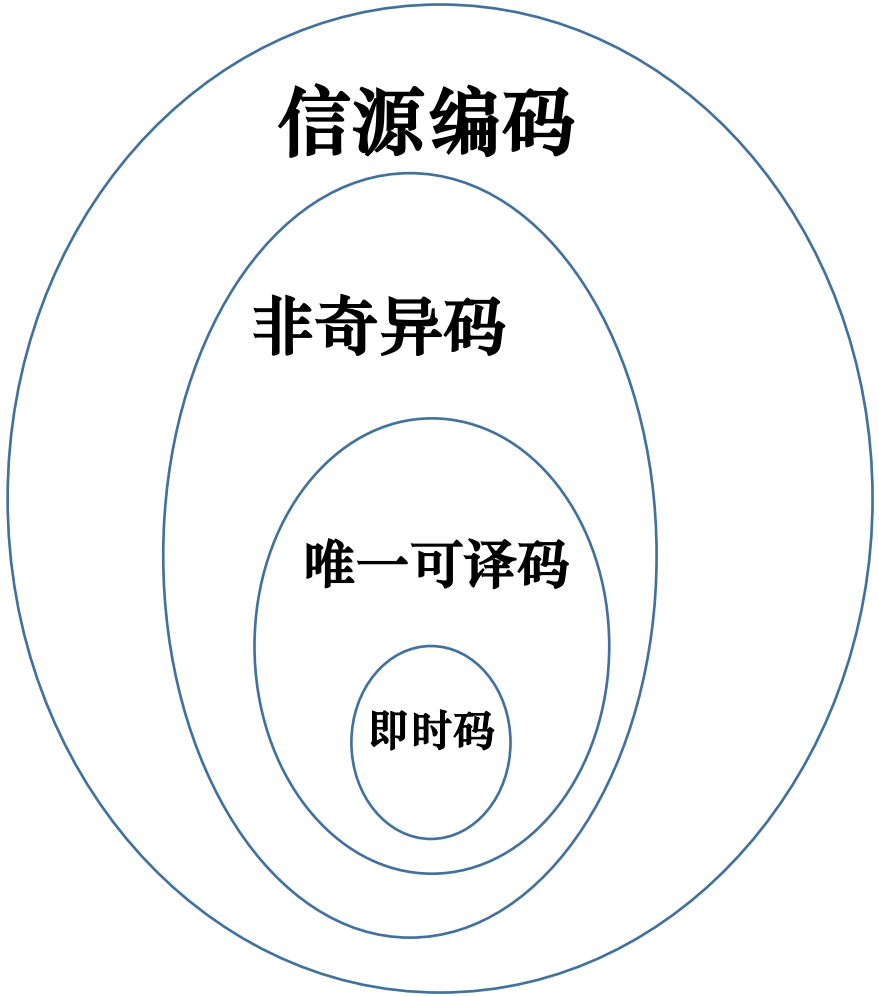
用码1译码　1 ……

用码2译码　1,1,０1,００1,０００1

即时码的充要条件：码组中任一码字都不是其他码字的前缀（前置）。

- **等长非奇异码一定是即时码。**

各类码之间的相互关系



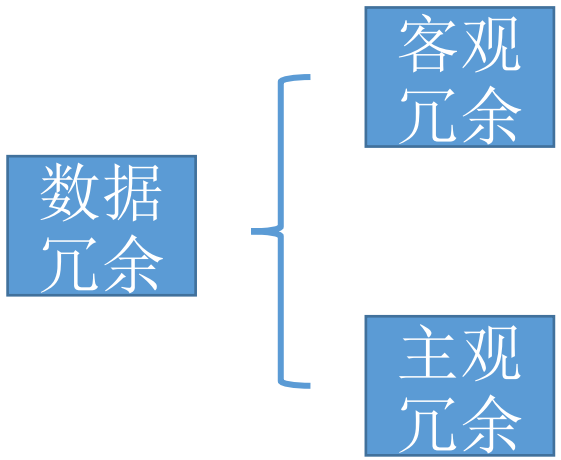
第3章

- 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 赫夫曼编码
- 3.6 游程组合编码

无失真定长编码定理

无失真变长编码定理

Shannon信息论对压缩编码的指导意义



- **数据压缩编码的基本途径**
- **数据压缩的理论极限**

Shannon信息论对压缩编码的指导意义

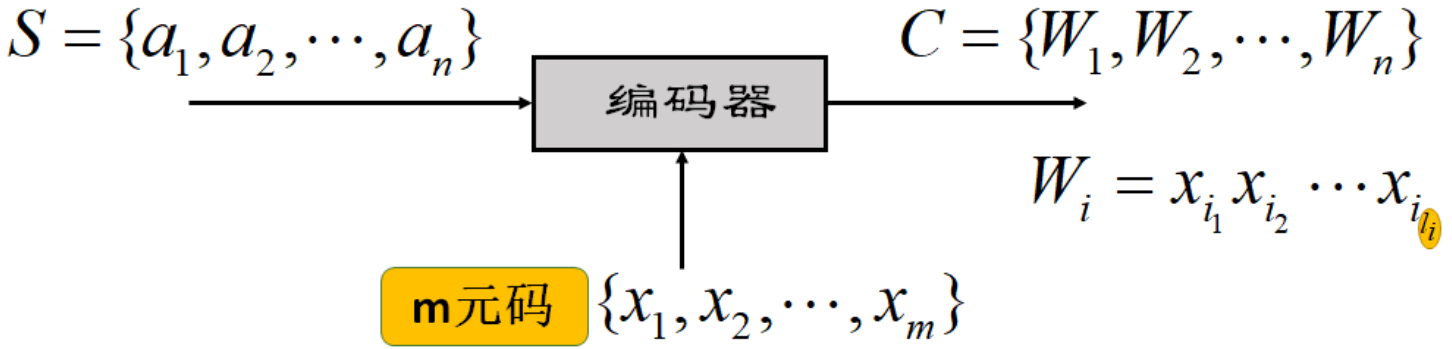
- 数据压缩编码的基本途径
- 数据压缩的理论极限

结论1：有记忆信源的冗余度寓于信源符号间的相关性中。去除它们之间的相关性，使之成为或几乎成为不相关的信源，其熵将增大。

结论2：离散无记忆信源的冗余度寓于信源符号概率的非均匀分布中。改变原信源的概率分布，使之成为或接近等概分布的信源，其熵将增大。

Shannon信息论对压缩编码的指导意义

- 数据压缩编码的基本途径
- 数据压缩的理论极限



编码的衡量指标

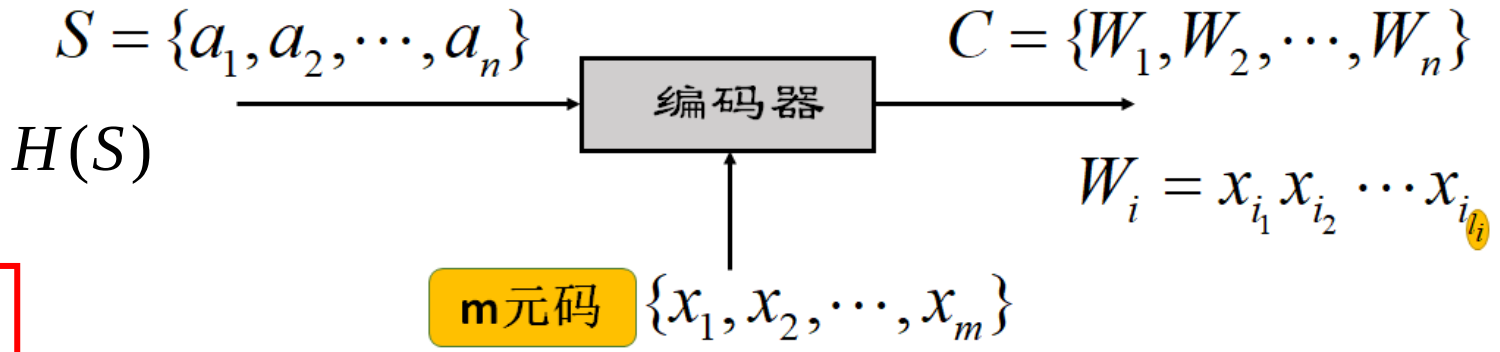
平均码长

对基本源编码：$\overline{L}=\sum_{i=1}^{n}p(a_i)l_i$ (码元/信源符号)

对N次扩展源编码：$\overline{L}_N=\sum_{i=1}^{n^N}p(\alpha_i)l_i$ (码元/N个信源符号)

$\overline{L}_N/N$ (码元/信源符号)

编码的衡量指标



编码后的信息传输率

对基本源编码：$R=\dfrac{H(S)}{\overline{L}}$ (比特 / 码元)

对**N**次扩展源编码：$R=\dfrac{H(S)}{\overline{L}_N/N}$ (比特 / 码元)

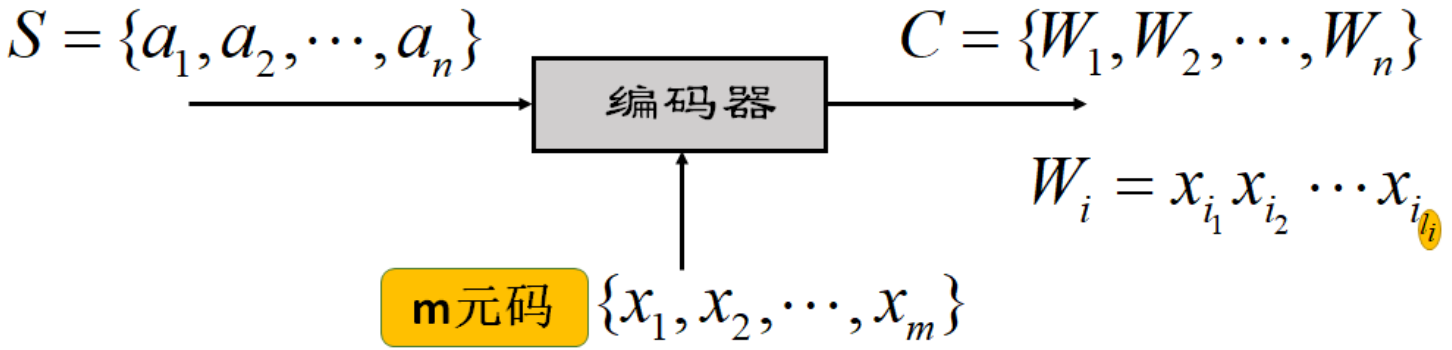
编码效率

对基本源编码：$\eta=\dfrac{R}{\log m}=\dfrac{H(S)}{\overline{L}\log m}$

对**N**次扩展源编码：$\eta=\dfrac{R}{\log m}=\dfrac{H(S)}{\dfrac{\overline{L}_N}{N}\log m}$

《信息论与编码》雷菁，黄英著，清华大学出版社

编码的衡量指标



编码后的信息率/编码速率

对基本源编码：$R=\overline{L}\log m$

对**N**次扩展源编码：$R=\dfrac{\overline{L}_N}{N}\log m$

编码效率

对基本源编码：$\eta=\dfrac{H(S)}{R}=\dfrac{H(S)}{\overline{L}\log m}$

对**N**次扩展源编码：$\eta=\dfrac{H(S)}{R}=\dfrac{H(S)}{\dfrac{\overline{L}_N}{N}\log m}$

例　对二元离散无记忆信源（DMS）进行无失真编码：

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

用二元码元 $\{0,1\}$ **对其编码：**　$s_1 \to 0,\quad s_2 \to 1$

信源熵为　$H(S) = H(\frac{3}{4}, \frac{1}{4}) = 0.811$(比特 / 信源符号)

平均码长　$\overline{L} = 1$　(码元 / 信源符号)

编码后的信息传输率　$R = \dfrac{H(S)}{\overline{L}} = 0.811$　(比特 / 码元)

编码效率　$\eta = \dfrac{R}{\log m} = 0.811$

对其进行二次扩展信源编码：

α_i	$p(\alpha_i)$	码字
s_1s_1	$\frac{9}{16}$	0
s_1s_2	$\frac{3}{16}$	10
s_2s_1	$\frac{3}{16}$	110
s_2s_2	$\frac{1}{16}$	111

信源熵为 $H(S)=H(\frac{3}{4},\frac{1}{4})=0.811$(比特 / 信源符号)

平均码长 $\overline{L}_2=\frac{9}{16}\times1+\frac{3}{16}\times2+\frac{3}{16}\times3+\frac{1}{16}\times3=1.688$ (码元 / 2个信源符号)

$\overline{L}_2/2=0.844$ (码元 / 信源符号)

编码后的信息传输率 $R=\frac{H(S)}{\overline{L}_2/2}=\frac{0.811}{0.844}=0.961$ (比特 / 码元)

编码效率 $\eta=\frac{R}{\log m}=0.961$

基本源编码 $R = 0.811$ (比特 / 码元) $\eta = 0.811$

二次扩展信源编码 $R = 0.961$ (比特 / 码元) $\eta = 0.961$

三次扩展信源编码 $R = 0.985$ (比特 / 码元) $\eta = 0.985$

四次扩展信源编码 $R = 0.991$ (比特 / 码元) $\eta = 0.991$

可见，随着信源扩展次数的增加，信息传输率越来越接近二元信源的最大熵，编码效率越来越接近**1**。

因此，提高信源的扩展次数可以有效地提高信源的编码效率，从而提高通信的有效性。

基本源编码 $R=0.811$ (比特 / 码元) $\eta=0.811$

二次扩展信源编码 $R=0.961$ (比特 / 码元) $\eta=0.961$

三次扩展信源编码 $R=0.985$ (比特 / 码元) $\eta=0.985$

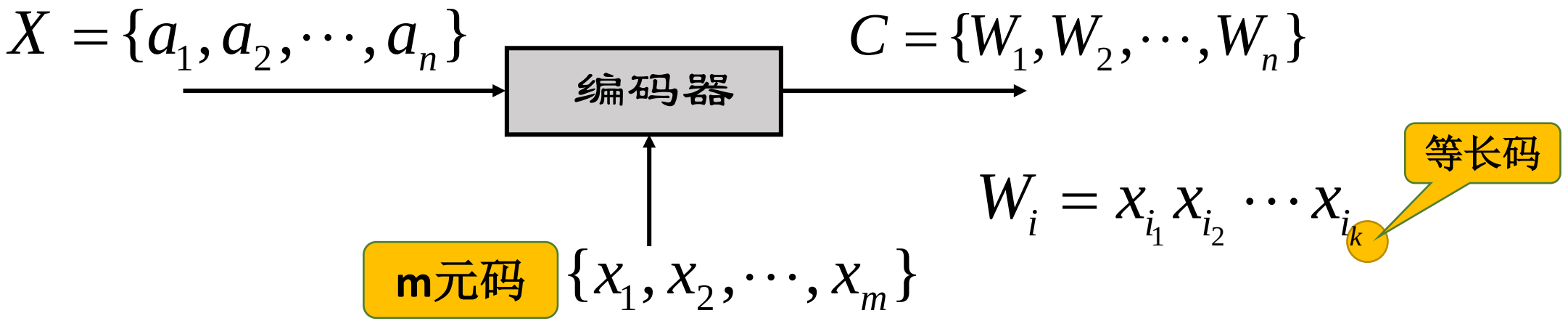
四次扩展信源编码 $R=0.991$ (比特 / 码元) $\eta=0.991$

本例采用了变长编码对扩展信源进行变换。若采用等长编码，可以计算得出：当编码效率达到**96%**时，需要对信源进行 4×10^7 次扩展。这张复杂度是难以实现的。

3.2.1 定长编码定理

定长码是所有码字的码长相同的码组。

前面讲到定长的非奇异码一定是唯一可译码，接下来我们讨论定长码的码长与非奇异性的关系，并分析码长的下界。



先讨论单符号无记忆离散信源。约定

- 信源符号集中共有 n 个符号（消息）$X=\{a_1,a_2,\cdots,a_n\}$
- 码符号是m 元码
- 编码后定长码的码长为K，则码字总数为 m^K

若满足非奇异性，则需要码字总数 $\geq$ 消息数，即 $m^K \geq n$

若对 L 次扩展信源进行定长编码，满足非奇异性，则需 $m^K \geq n^L$

例子　英文字母表中（不含空格），

1. 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?

2. 若对英文字母信源的2次、4次扩展信源进行二元编码，则码字的最短码长又是多少?

$$m^K \geq n \qquad m^K \geq n^L$$

解:

1. 信源符号数为 n=26，码符号数是 m=2，

$$m^K \geq n \longrightarrow 2^K \geq 26 \longrightarrow K \geq \frac{\log 26}{\log 2} = 4.7$$

$$K_{\min} = 5 \text{ （码符号/信源符号）}$$

例子　英文字母表中（不含空格），

1. 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?

2. 若对英文字母信源的2次、4次扩展信源进行二元编码，则码字的最短码长又是多少?

$$m^K \geq n \qquad m^K \geq n^L$$

解：

2. 信源符号数为 n=26，码符号数是 m=2. 对二次扩展码，L=2

$$m^K \geq n^L \longrightarrow 2^K \geq 26^2 \longrightarrow K \geq \frac{\log 26^2}{\log 2} = 9.4$$

$K_{\min} = 10$（码符号/2个信源符号）

每个信源符号对应的平均码长为 $\overline{K} = \frac{K_{\min}}{2} = 5$（码符号/信源符号）

例子　英文字母表中（不含空格），

1. 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?

2. 若对英文字母信源的2次、4次扩展信源进行二元编码，则码字的最短码长又是多少?

$$m^K \geq n \qquad m^K \geq n^L$$

解：

2. 信源符号数为 n=26，码符号数是 m=2. 对四次扩展码，L=4

$$m^K \geq n^L \Rightarrow 2^K \geq 26^4 \Rightarrow K \geq \frac{\log 26^4}{\log 2} = 18.8$$

$K_{\min} = 19$ (码符号/4个信源符号)

每个信源符号对应的平均码长为 $\overline{K} = \frac{K_{\min}}{4} = 4.75$ (码符号/信源符号)

若对 L 次扩展信源 X^L 进行定长编码，如果编得的定长码是非奇异码，必须 $m^K \geq n^L$

- 信源符号集中共有 n 个符号 $X = \{a_1, a_2, \cdots, a_n\}$
- 码符号是m 元码
- 编码后定长码的码长为K

$$m^K \geq n^L \Rightarrow K \log m \geq L \log n \quad \text{或} \quad \frac{K}{L} \geq \log_m n$$

一个信源符号编码后的码长

$$m^K \geq n^L$$

$$K\log m \geq L\log n \quad \left(\frac{K}{L} \geq \frac{\log n}{\log m}\right)$$

满足此条件的定长编码虽然可以保证无失真的编码，但是编码效率低。

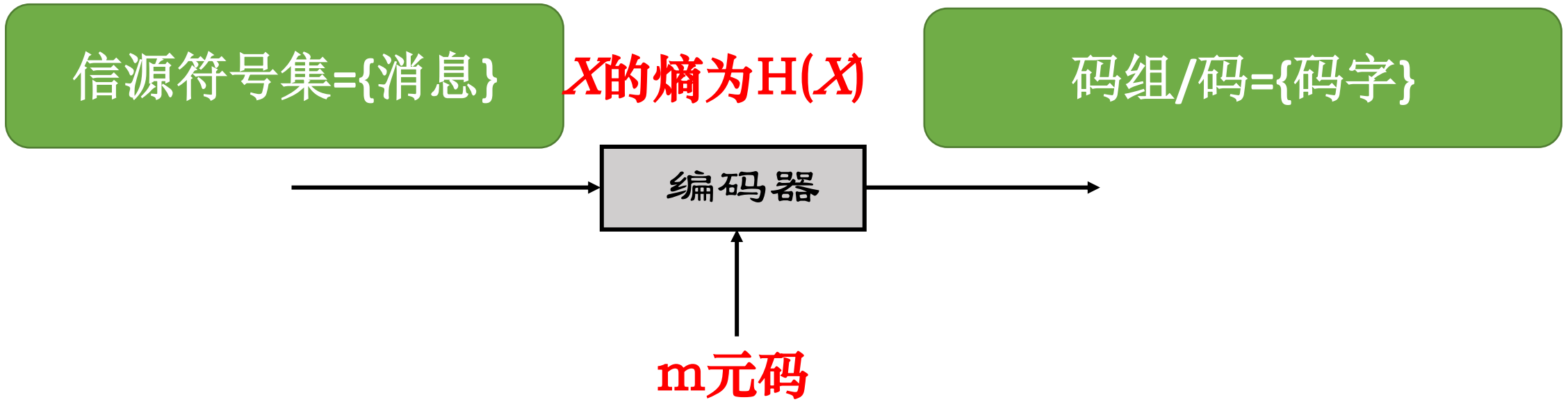
假定码元的取值是等概率的

$$LH(X) \leq L\log n$$

$$K\log m$$

扩展后消息的符号长度为L

码字的码长为K



3.2.1 定长编码定理

设离散无记忆信源 S 的熵为$H(X)$，若对其L次扩展信源 S^L 进行K定长编码（采用m元码），对任意 $\varepsilon > 0, \delta > 0$，只要

$$\frac{K}{L}\log m \geq H(X) + \varepsilon \qquad (3.2.1)$$

则当L足够大时，译码差错小于 δ 。

反之，若 $\frac{K}{L}\log m \leq H(X) - 2\varepsilon$

则当L足够大时，译码必定出错。

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

假定码元的取值是等概率的

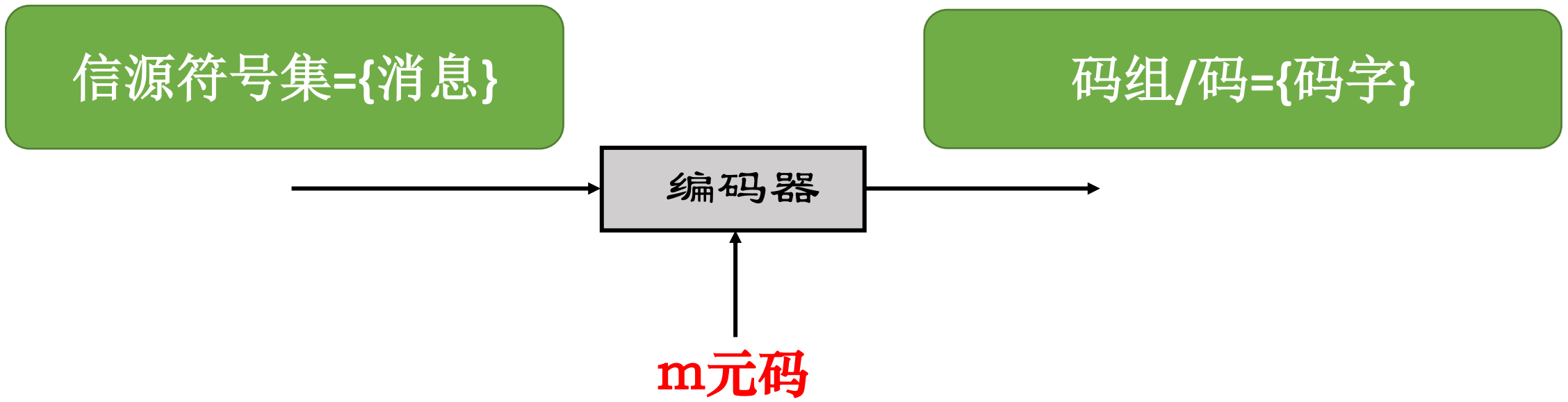
$LH(X)$

单个符号的熵为H(X)

消息的符号长度为L

$K\log m$

每个码字的码长均为K



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

编码后，平均每个信源符号携带的信息量

信息率*I*/编码速率*R*

$$\frac{K}{L}\log m \leq H(X)-2\varepsilon$$

定长编码正定理：信息率略大于单符号信源熵时，可做到几乎无失真译码，条件是L必须足够大。可以证明，只要

$$L \geq \frac{\sigma^2[I(a_i)]}{\varepsilon^2\delta} \qquad (3.2.2)$$

译码差错率一定小于任意正数 δ 。

编码后，平均每个信源符号携带的信息量

信息率*I*/编码速率*R*

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

$$\frac{K}{L}\log m \leq H(X)-2\varepsilon$$

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

$\frac{K}{L}$：一个信源符号编码后的码长

定长编码逆定理：信息率比单符号信源熵略小一点(ε)时，译码差错未必超过限定值 δ ；但若小于2ε ，则译码失真一定大于 δ ，$L \to \infty$ ，必定失真。

单符号信源熵**H(X)**是一个临界值。当编码器输出的信息率超过这个临界值时，就能无失真译码，否则就不能无失真译码。

信息率/编码速率： $R=\frac{K}{L}\log m$ （$\geq H(X)+\varepsilon$）

含义：编码后平均每个信源符号携带的信息量。

编码效率： $\eta=\frac{H(X)}{R}$ $\left(\leq \frac{H(X)}{H(X)+\varepsilon}\right)$

含义：编码前平均每个信源符号携带的信息量除以编码后平均每个信源符号携带的信息量，即效率。

例3.2.1 设某单符号信源模型为

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{Bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.4 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{Bmatrix}$$

计算得 $H(X) = 2.5525(bit/sign)$

$$\sigma^2\left[I(a_i)\right] = E\left[I^2(a_i)\right] - H^2(X) = 1.3082$$

若要求编码效率为90%，即 $\dfrac{H(X)}{H(X)+\varepsilon} = 0.90$

则 $\varepsilon = 0.2836$

$$L \geq \frac{\sigma^2[I(a_i)]}{\varepsilon^2 \delta}$$

设译码差错概率为10^{-6}，由式(3.2.2)得 $L \geq \dfrac{1.3082}{0.2836^2 \times 10^{-6}} = 1.6265 \times 10^7$

若用长度为3的二进制进行等长编码，$\eta = \dfrac{H(X)}{K \log m} = 2.5525/3 = 85.08\%$

3.2.2 变长编码定理

设离散无记忆信源 S 的熵为 $H(X)$，若对其 L 次扩展信源 S^L 进行变长编码（采用 m 元码），一定存在无失真的信源编码方法，使得

码字**平均长度** $\overline{K}$ 满足：$$1+\frac{LH(X)}{\log m}>\overline{K}\geq\frac{LH(X)}{\log m}$$

码字**平均信息率** R 满足：$$H(X)\leq R<H(X)+\varepsilon$$

对离散无记忆信源，消息长度为L，符号熵为$H(X)$,对信源进行m元变长编码，一定存在无失真的信源编码方法，使得

码字平均长度 $\overline{K}$ 满足：$1+\frac{LH(X)}{\log m}>\overline{K}\geq\frac{LH(X)}{\log m}$

证明：

设信源符号 $X=\{a_1,a_2,\ldots,a_i,\ldots,a_n\}$，概率为 $p(a_i)(i=1,2,...,n)$

若对a_i 用一个长度为 k_i 的码字，使得

$$1-\frac{\log p(a_i)}{\log m}>k_i\geq-\frac{\log p(a_i)}{\log m}$$

满足上式的整数一定存在。

码字**平均长度** $\overline{K}$ 满足：$1+\dfrac{LH(X)}{\log m}>\overline{K}\geq\dfrac{LH(X)}{\log m}$

证明（续）：

若对a_i 用一个长度为 k_i 的码字，使得

$$1-\frac{\log p(a_i)}{\log m}>k_i\geq-\frac{\log p(a_i)}{\log m}$$

不等式两端同时乘以 $p(a_i)$ 并对 i 求和，得

$$1-\sum_{i=1}^{n}p(a_i)\frac{\log p(a_i)}{\log m}>\sum_{i=1}^{n}p(a_i)k_i\geq-\sum_{i=1}^{n}p(a_i)\frac{\log p(a_i)}{\log m}$$

$$1+\frac{H(X)}{\log m}>\overline{K}\geq\frac{H(X)}{\log m}$$

码字**平均长度** $\overline{K}$ 满足：$1+\frac{LH(X)}{\log m}>\overline{K}\geq\frac{LH(X)}{\log m}$

证明（续）：

若对a_i用一个长度为k_i的码字，使得

$$1+\frac{H(X)}{\log m}>\overline{K}\geq\frac{H(X)}{\log m}$$

对于平稳无记忆信源来说，当信源输出长度为L的消息序列时，

$$1+\frac{LH(X)}{\log m}>\overline{K}\geq\frac{LH(X)}{\log m}$$

变长编码定理的含义：

$$1+\frac{LH(X)}{\log m}>\overline{K}\geq\frac{LH(X)}{\log m}$$

$$H(X)\leq R<H(X)+\varepsilon$$

$$\Rightarrow\quad\frac{\overline{K}}{L}\geq\frac{H(X)}{\log m}$$

是无失真信源编码
平均码长的下界

香农第一定理给出了本课程的第一个理论极限：平均码长的压缩下限，它和信源的熵有关。

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

编码后每个码符号实际载荷的信息量

编码前每个信源符号实际载荷的信息量

编码信源符号所需的平均码长下界

编码信源符号的实际平均码长

编码效率η最大是100%

编码后每个码符号最大能载荷的信息量

编码后每个信源符号能够载荷的最大信息量

- 对信源进行变长编码，所要求的信源长度L一般来说比定长编码小得多。

- 变长码编码效率存在下界

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

对例3.2.1编变长码：$H(X) = 2.5525(bit/sign)$

$$\eta = \frac{2.5525}{2.5525 + \frac{1}{L}} = 0.90 \Rightarrow L = \frac{1}{0.2836} = 3.5261$$

3.2.3 码字唯一可译条件

译码：从接收到的码字序列得到信源符号序列的过程。

对不等长编码，如何分离码字?

$$\underline{0_{?}1_{?}}10011\cdots$$

如果0，01都是码字，译码时如何分离?

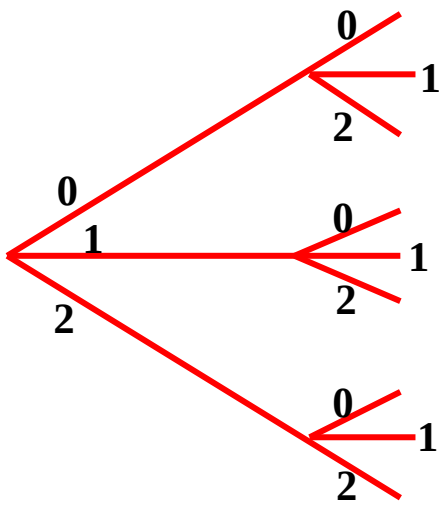
- 码
  - 奇异码
  - 非奇异码
    - 非唯一可译码
    - 唯一可译码
      - 即时码/异前置码
      - 非即时码

即时码的充要条件：码组中任一码字都不是其他码字的前缀（前置）。

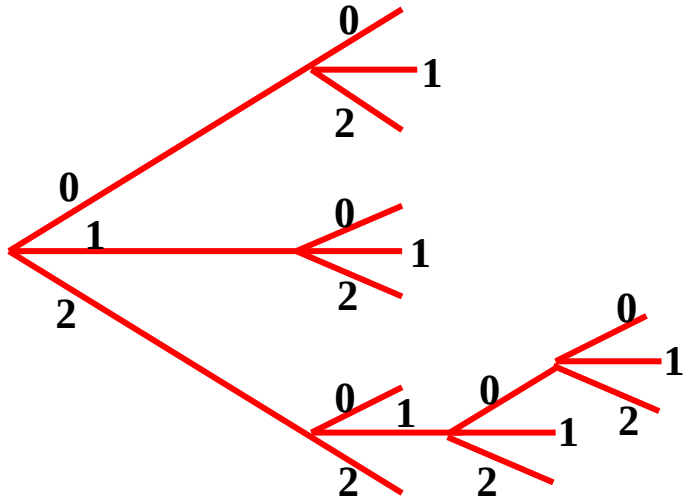
例3.2.2

消息	概率	码A	码B	码C	码D
a_1	0.5	0	0	0	0
a_2	0.25	0	1	01	10
a_3	0.125	1	00	011	110
a_4	0.125	10	11	0111	1110
		非唯一可译码	非唯一可译码	唯一可译有延时	唯一可译即时码异前置码

即时码可用树图法来构造：



(a) 满树



(b) 非满树

树根，树枝，根节点，中间节点（一级节点，二级节点,…），终节点

图3.2.1 3元码树图

唯一可译码和即时码的判别

克拉夫特（Kraft）不等式：对于码长分别为$k_1, k_2,\ldots, k_n$的m元码，存在即时码的充要条件是 $\sum_{i=1}^{n} m^{-k_i} \le 1$

麦克米伦（McMillan）不等式：对于码长分别为$k_1, k_2,\ldots, k_n$的m元码，存在唯一可译码的充要条件是 $\sum_{i=1}^{n} m^{-k_i} \le 1$

说明：n-码字的个数；m-m元码；k_i-第i个码字的码长

- 这表明在码长选择的条件上，即时码与唯一可译码是一致的，唯一可译码并不比即时码有更宽松的条件。

例子：

符号	码1	码2	码3	码4	码5
a_1	00	0	0	1	1
a_2	01	10	11	01	10
a_3	10	00	00	001	100
a_4	11	01	11	0001	1000

$m = 2$

$n = 4$

码1： $\sum_{i=1}^{n} m^{-k_i} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-2} = 1 \leq 1$

码2、3： $\sum_{i=1}^{n} m^{-k_i} = 2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} = \frac{5}{4} > 1$

码4、5： $\sum_{i=1}^{n} m^{-k_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = \frac{15}{16} \leq 1$

符号	码1	码2	码3	码4	码5
a_1	00	0	0	1	1
a_2	01	10	11	01	10
a_3	10	00	00	001	100
a_4	11	01	11	0001	1000

码1：$\sum_{i=1}^{n} m^{-k_i} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-2} = 1 \leq 1$

码2、3：$\sum_{i=1}^{n} m^{-k_i} = 2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} = \frac{5}{4} > 1$

码4、5：$\sum_{i=1}^{n} m^{-k_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = \frac{15}{16} \leq 1$

- 奇异码：至少两个符号的编码相同（码3）
- 非奇异码：所有码字均不相同（码1、码2、码4、码5）
  - 非唯一可译码：译码时会产生歧义 （码2）
  - 唯一可译码：译码时不会产生歧义（码1、码4、码5）
    - 即时码：不需要知道下一个码子的符号就能译码（码1、码4）
    - 非即时码：需要知道下一个码子的符号才能译码（码5）

唯一可译码和即时码的判别

克拉夫特（Kraft）不等式：对于码长分别为$k_1, k_2,\ldots, k_n$的m元码，存在即时码的充要条件是 $\sum_{i=1}^{n} m^{-k_i} \le 1$

麦克米伦（McMillan）不等式：对于码长分别为$k_1, k_2,\ldots, k_n$的m元码，存在唯一可译码的充要条件是 $\sum_{i=1}^{n} m^{-k_i} \le 1$

说明：n-码字的个数；m-m元码；k_i-第i个码字的码长

- 上述不等式给出的是存在性定理。

即：当满足 *Kraft*（或 *McMillan* 不等式），必然可构造出满足其码长要求的即时码（或唯一可译码），否则不可，为存在性的验证。两不等式可作为判断一种码不是即时码（或唯一可译码）的依据，而不能作为判断判断一种码是即时码（唯一可译码）的依据。

第3章

- 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 赫夫曼编码
- 3.6 游程组合编码

信源编码的基本途径有两个：

使序列中的各个符号尽可能地互相独立，即解除相关性。

使编码中各个符号出现的概率尽可能地相等，即概率均匀化。

信源编码方案：

- 定码长，定码字
- 衡量性能：平均码长，编码效率等

设有离散无记忆信源 $\begin{bmatrix} a_1 & a_2 & \cdots\cdots & a_n \\ p(a_1) & p(a_2) & \cdots\cdots & p(a_n) \end{bmatrix}, \sum_{i=1}^{n} p(a_i) = 1$

二元香农编码方法步骤如下：

1. 按信源符号的概率从大到小的顺序排队，不妨设

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

2. 令$p(a_0) = 0$，用 $p_a(a_j)$ 表示第 j 个码字的累加概率：

$$p_a(a_j) = \sum_{i=0}^{j-1} p(a_i), \quad j = 1, 2, \cdots, n$$

3. 确定每个码字的码长：$-\log_2 p(a_i) \leq k_i < 1 - \log_2 p(a_i)$

4. 把$p_a(a_j)$用二进制表示，用小数点后的 k_i 位作为 a_i 的码字。

小数的二进制：

$$0.625*2=1.25$$
$$0.25*2=0.5$$
$$0.5*2=1$$

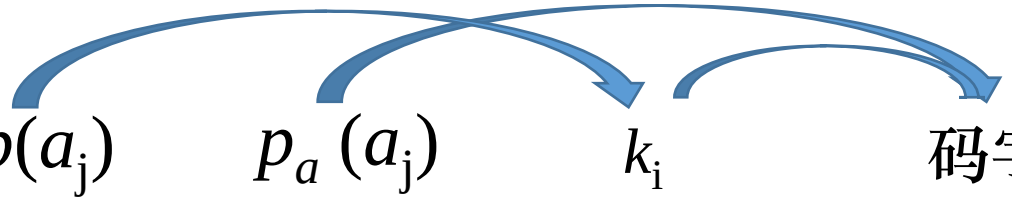
$$(0.625)_{10}=(0.101)_2$$

例3.3.1 设有一单符号离散无记忆信源

$$\begin{pmatrix} X \\ P(X) \end{pmatrix}=\begin{Bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ 0.2 & 0.25 & 0.25 & 0.1 & 0.15 & 0.05 \end{Bmatrix}$$

试对该信源编二进制香农码，并求平均码长、信息率和编码效率。

解：编码过程：



	$p(a_j)$	$p_a(a_j)$	k_i	码字
a_0	0			
a_2	0.25	0.00	2	00
a_3	0.25	0.25	2	01
a_1	0.2	0.50	3	100
a_5	0.15	0.70	3	101
a_4	0.1	0.85	4	1101
a_6	0.05	0.95	5	11110

$$-\log_2 p(a_i) \le k_i < 1-\log_2 p(a_i)$$

$$H(X) = 2.4233(bit/\text{信源符号})$$

平均码长：$\bar{K} = \sum_{i=1}^{6} p(a_i)k_i = 2.7(bit\,/\,\text{信源符号})$

信息率：$R = \frac{\bar{K}}{L}\log m = 2.7$

编码效率：$\eta = \frac{H(X)}{R} = 89.75\%$

第3章

费诺编码步骤如下：

1. 按信源符号的概率从大到小的顺序排队

   不妨设 $p(a_1) \geq p(a_2) \geq \ldots\ldots \geq p(a_n)$

2. 对概率按m进行分组，使每组概率和尽可能相等。

3. 给每个分组分配一个码元。

4. 对每个分组重复2、3步，直到不可分为止。

例

设有一单符号离散无记忆信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ 0.32 & 0.22 & 0.18 & 0.16 & 0.08 & 0.04 \end{Bmatrix}$$

试对该信源编二进制费诺码，并求平均码长、信息率和编码效率。

解：编码过程

a_1	0.32	0	0			00
a_2	0.22		1			01
a_3	0.18	1	0			10
a_4	0.16		1	0		110
a_5	0.08			1	0	1110
a_6	0.04				1	1111

$$H(X) = 2.35(\text{比特/符号})$$

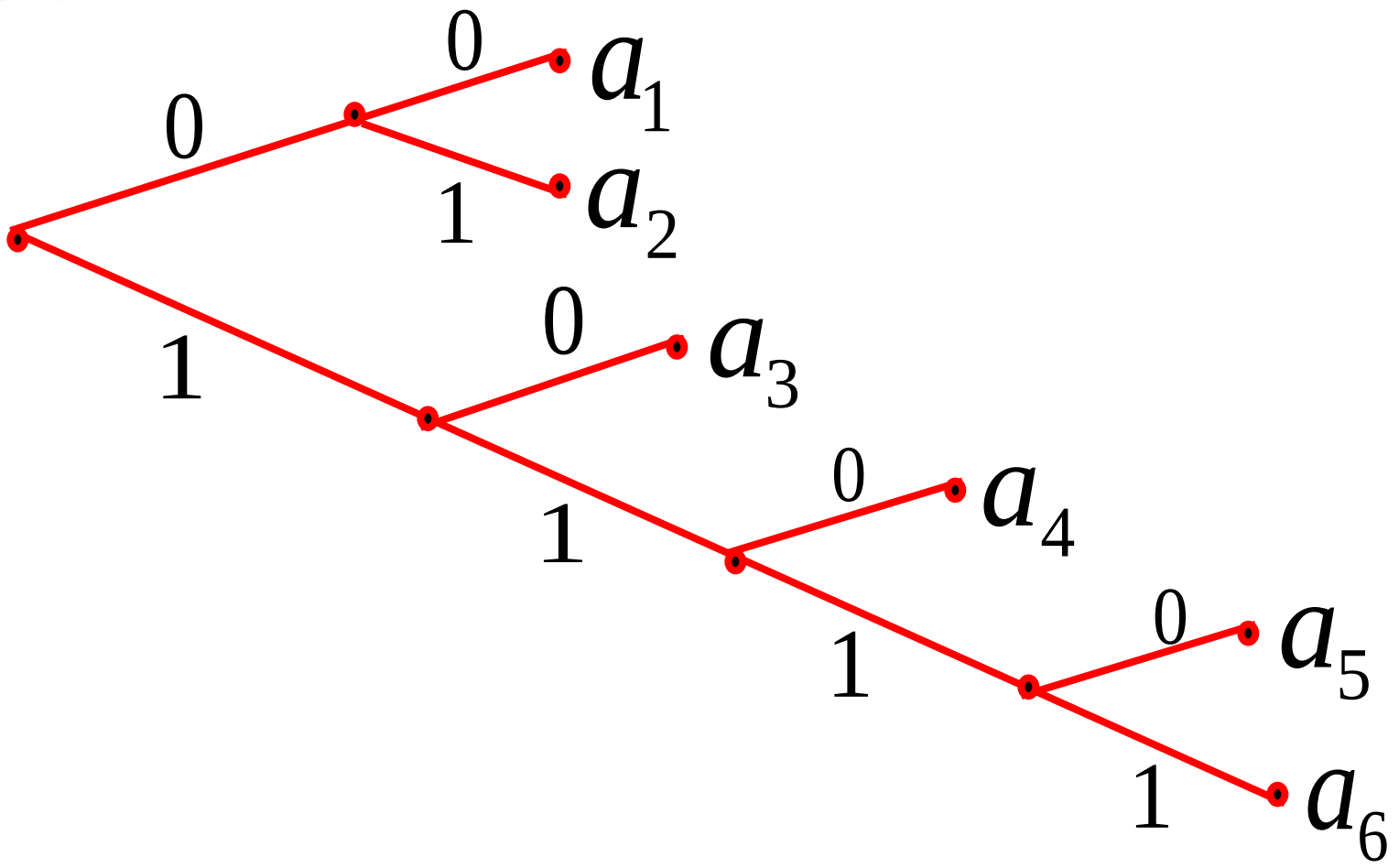
$$\overline{K} = \sum_{i=1}^{6} p(a_i)k_i = 2.4(\text{比特/符号})$$

$$R = \frac{\overline{K}}{L}\log_2 m = 2.4$$

$$\eta = \frac{H(X)}{R} = 97.92\%$$

- **本例中费诺码有较高的编码效率。**
- **费诺码比较适合于每次分组概率都很接近的信源。**

树图：



第3章

赫夫曼（Huffman）编码步骤如下：

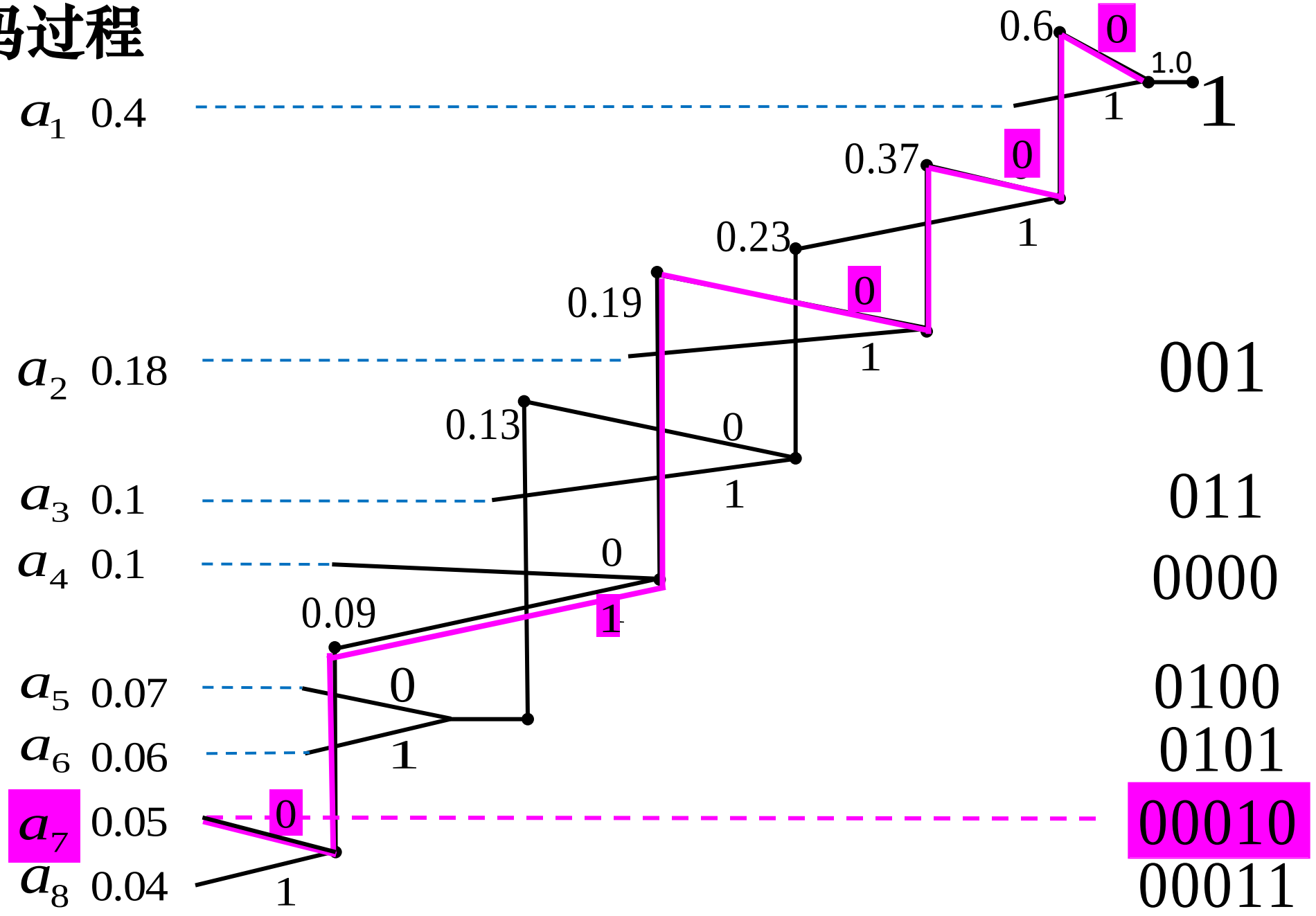
1. 将信源符号按概率由大到小顺序排队。
2. 给两个概率最小的符号各分配一个码位，将其概率相加后合并作为一个新的符号，与剩下的符号一起，再重新排队。
3. 给缩减信源中概率最小的符号各分配一个码位。
4. 重复步骤2、3直至概率和为1。

例3.5.3

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。

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

编码过程


a_1 0.4
a_2 0.18
a_3 0.1
a_4 0.1
a_5 0.07
a_6 0.06
a_7 0.05
a_8 0.04
0.6
1.0
0.37
0.23
0.19
0.13
0.09
0
1
001
011
0000
0100
0101
00010
00011

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

$$\bar{K} = 2.61(bit\ /\ sign)$$

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

若采用定长编码，码长$K = 3$,则编码效率

$$\eta = \frac{2.55}{3} = 85\%$$

可见，哈夫曼编码的效率提高了12.7%。

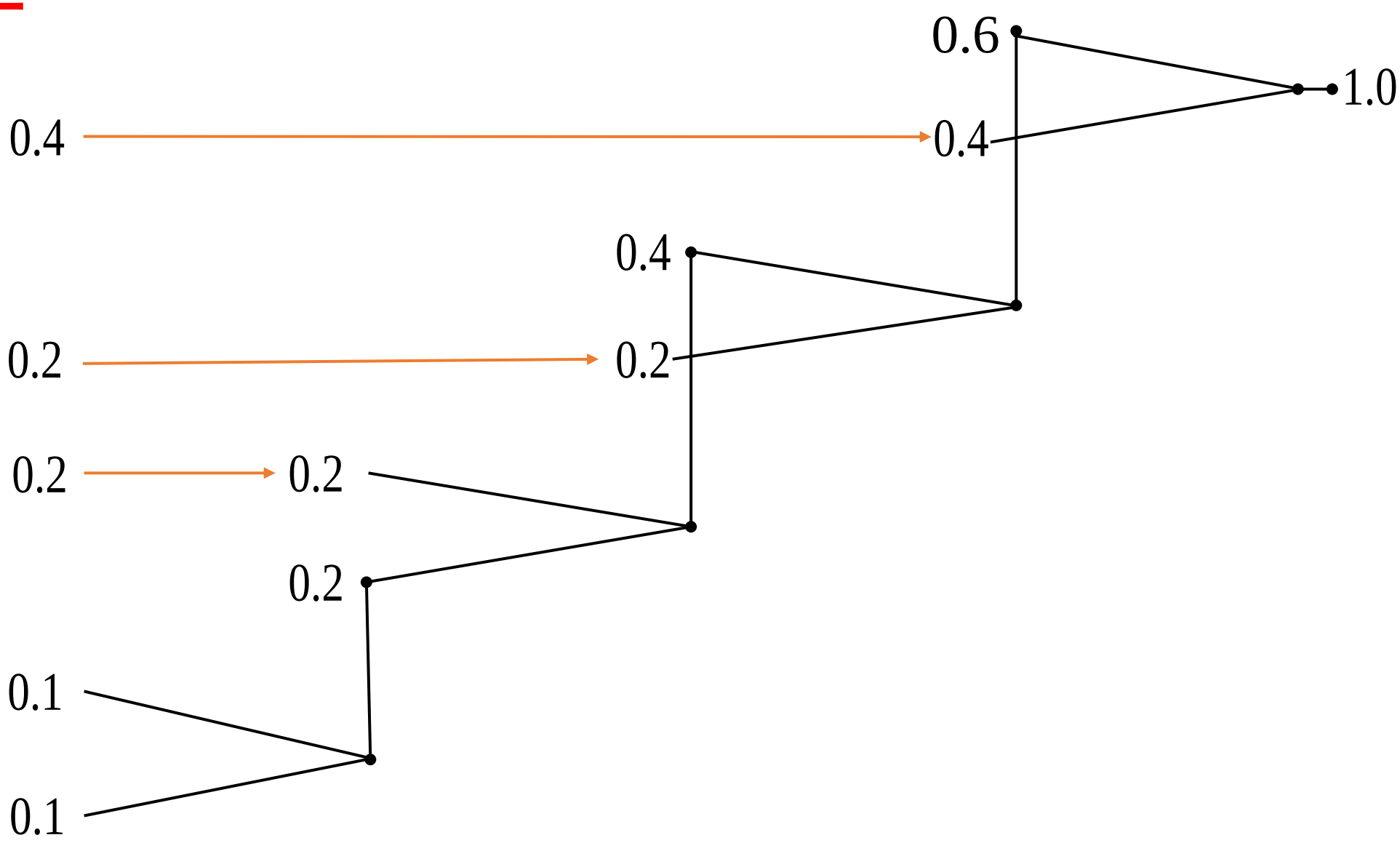
Huffman码的编码方法不是唯一的。

例3.5.2 设有离散无记忆信源

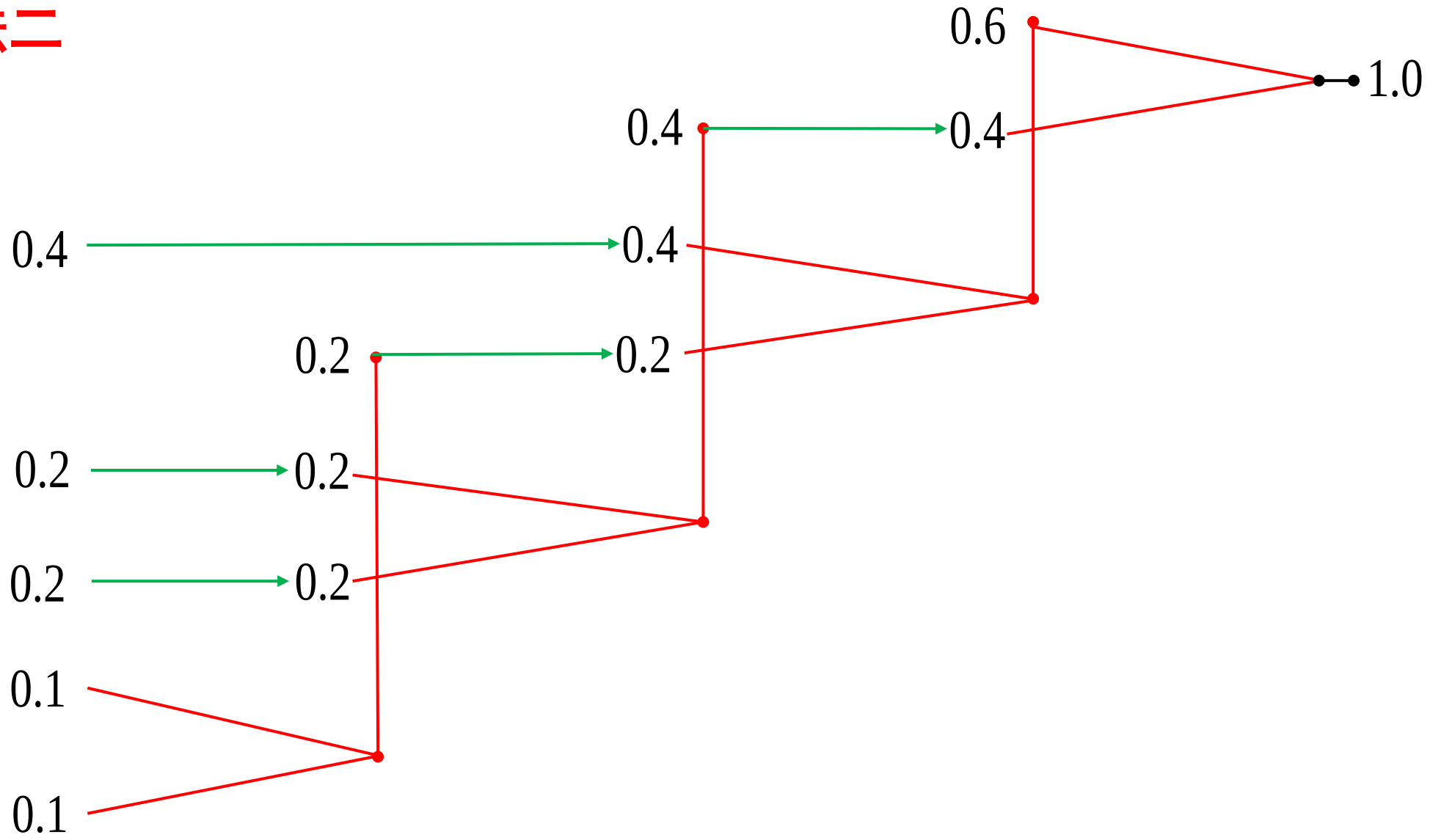
$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{bmatrix}$$

用两种不同的方法对其编二进制Huffman码。

方法一



方法二



两种不同编码方法的对比

信源符号a_i	概率$p(a_i)$	码字W_{i1}	码长K_{i1}	码字W_{i2}	码长K'_{i2}
a_1	0.4	1	1	00	2
a_2	0.2	01	2	10	2
a_3	0.2	000	3	11	2
a_4	0.1	0010	4	010	3
a_5	0.1	0011	4	011	3

平均码长和编码效率

$$\bar{K}=\sum_{i=1}^{7}p(a_i)k_i=2.2(bit/sign),\quad \eta=\frac{H(X)}{\bar{K}}=0.965$$

两种编码方法的码长方差比较

$$\sigma^2=E\left[(k_i-\bar{K})^2\right]=\sum_{i=1}^{n}p(a_i)(k_i-\bar{K})^2,\ \sigma_1^2=1.36,\quad \sigma_2^2=0.16$$

？ 码方差小意味着什么？

进行赫夫曼编码时，如果有几个符号概率相等，应使合并后的信源符号**尽可能排在缩减信源序列的前面**，以减少再次合并的次数，充分利用短码。

扩展：m元赫夫曼编码

- 构造m元赫夫曼编码，每次将概率最小的m个进行合并

- 为充分利用短码，必须最后一次合并有m个符号

  信源符号数 $n = m + k(m-1)$，可完全合并

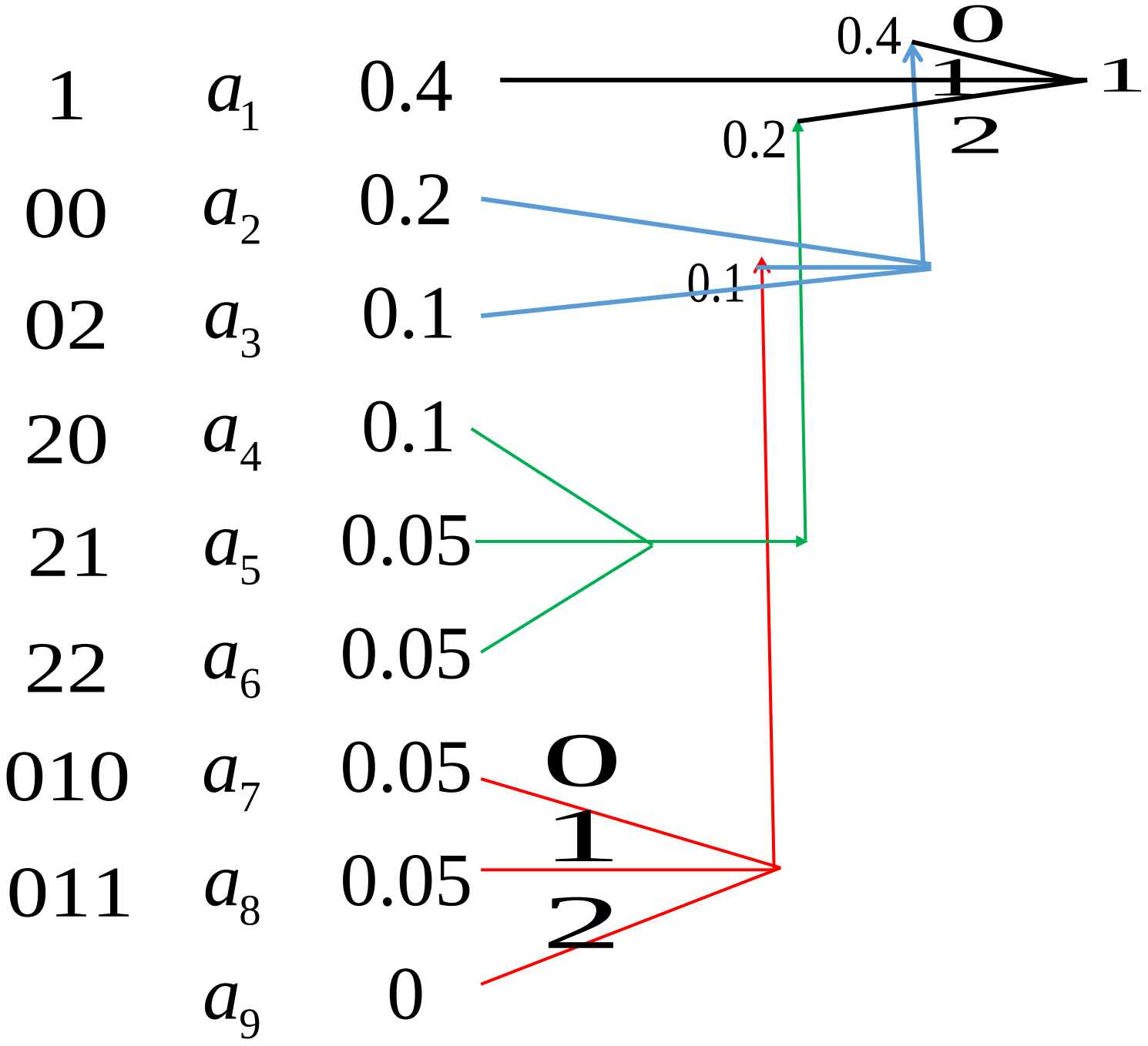
  若信源符号数不满足 $n = m + k(m-1)$，则补充一些概率为0的符号，使得符号总数满足 $n = m + k(m-1)$ 。

例 设有离散无记忆信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.4 & 0.2 & 0.1 & 0.1 & 0.05 & 0.05 & 0.05 & 0.05 \end{Bmatrix}$$

码符号为$\{0,1,2\}$，对其编三元Huffman码。

$$n = m + k(m-1) = 3 + 2k \quad \Rightarrow \quad \text{取} n = 9$$


1 a_1 0.4
00 a_2 0.2
02 a_3 0.1
20 a_4 0.1
21 a_5 0.05
22 a_6 0.05
010 a_7 0.05
011 a_8 0.05
a_9 0
0.4
0
1
1
2
0.2
0.1
0
1
2

补充：赫夫曼编码是最佳编码。

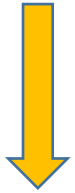
定理1　对于给定的信源，存在最佳唯一可译二元码，其最小概率的两个码字的长度最长且相等，它们之间仅最后一位码元取值不同（一个为0，另一个为1）。

定理2　对缩减信源为最佳码，则对原始信源也是最佳码。

定理1　对于给定的信源，存在最佳唯一可译二元码，其最小概率的两个码字的长度最长且相等，它们之间仅最后一位码元取值不同（一个为0，另一个为1）。

证明：设

$$\begin{cases} s_1, s_2, \ldots, s_{K-1}, s_K \\ p_1, p_2, \ldots, p_{K-1}, p_K \end{cases}$$



$$\begin{cases} c_1, c_2, \ldots, c_{K-1}, c_K \\ l_1, l_2, \ldots, l_{K-1}, l_K \end{cases}$$

- l_K **最大**
- **存在另外一个码字长度也为** l_K，**并且与** c_K **仅最后一位码元取值不同（一个为0，另一个为1）**
- **满足上个条件的码字是** c_{K-1}

$$S:\begin{cases} s_1, s_2, \ldots, s_{K-1}, s_K \\ p_1, p_2, \ldots, p_{K-1}, p_K \end{cases}$$

$$S^{(1)}:\begin{cases} s_1^{(1)}, s_2^{(1)}, \ldots, s_{K-1}^{(1)} \\ p_1^{(1)}, p_2^{(1)}, \ldots, p_{K-1}^{(1)} \end{cases}$$

$$S^{(K-2)}:\begin{cases} s_1^{(K-2)}, s_2^{(K-2)} \\ p_1^{(K-2)}, p_2^{(K-2)} \end{cases}$$

$$S:\begin{cases} s_1, s_2, \ldots, s_{K-1}, s_K \\ p_1, p_2, \ldots, p_{K-1}, p_K \end{cases} \qquad S':\begin{cases} s_1', s_2', \ldots, s_{K-1}' \\ p_1', p_2', \ldots, p_{K-1}' \end{cases}$$

$$\begin{cases} c_1, c_2, \ldots, c_{K-1}, c_K \\ l_1, l_2, \ldots, l_{K-1}, l_K \end{cases} \qquad \begin{cases} c_1', c_2', \ldots, c_{K-1}' \\ l_1', l_2', \ldots, l_{K-1}' \end{cases}$$

$$c_1 = c_1' \qquad l_1 = l_1' \qquad p_1' = p_1$$

$$\ldots \qquad \ldots \qquad \ldots$$

$$c_{K-2} = c_{K-2}' \qquad l_{K-2} = l_{K-2}' \qquad p_{K-2}' = p_{K-2}$$

$$c_{K-1} = (c_{K-1}' 0) \qquad l_{K-1} = l_{K-1}' + 1 \qquad p_{K-1}' = p_{K-1} + p_k$$

$$c_K = (c_{K-1}' 1) \qquad l_K = l_{K-1}' + 1$$

$$l_1 = l_1^{'}$$

$$\cdots$$

$$l_{K-2} = l_{K-2}^{'}$$

$$l_{K-1} = l_{K-1}^{'} + 1$$

$$l_K = l_{K-1}^{'} + 1$$

$$p_1^{'} = p_1$$

$$\cdots$$

$$p_{K-2}^{'} = p_{K-2}$$

$$p_{K-1}^{'} = p_{K-1} + p_k$$

$$\bar{L} = \sum_{k=1}^{K} p_k l_k$$

最小

$$= \sum_{k=1}^{K-2} p_k^{'} l_k^{'} + p_{K-1}(l_{K-1}^{'} + 1) + p_K(l_{K-1}^{'} + 1)$$

$$= \sum_{k=1}^{K-2} p_k^{'} l_k^{'} + (p_{K-1} + p_K) l_{K-1}^{'} + (p_{K-1} + p_K)$$

$$= \sum_{k=1}^{K-1} p_k^{'} l_k^{'} + (p_{K-1} + p_K)$$

最小

第3章

- 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 赫夫曼编码
- 3.6 游程组合编码

前面几种编码方法主要针对无记忆信源，当信源有记忆时，这些编码方法可以用，但编码效率不高。游程编码是针对有记忆信源的编码方法，对相关信源的编码更有效。

游程

指数字序列中连续出现相同符号的一段。在二元信源中，连续的一段‘0’称为一个‘0’游程，‘0’的个数称为此游程的长度，同样，也有‘1’游程。

100001：长度为4的“0”游程

01110：长度为3的“1”游程

游程序列

用交替出现的‘0’游程、‘1’游程的长度，来表示任意二元序列而产生的一个新序列。它和二元序列是一个一一对应的变换。

二元序列：000101110010001……

游程序列：31132131……

游程编码只适用于二元序列

- 若已知二元序列以0起始，从游程序列很容易恢复出原来的二元序列。
- 游程变换是一一对应的可逆变换，所以游程变换后熵不变。
- 游程序列是多元序列，各长度可按赫夫曼编码或其它方法处理以达到压缩码率的目的。

游程的概率

设“0”、“1”的概率分别为 p_0、p_1，则

长度 i 的“0”游程概率：$p[l_i^0] = p_0^{i-1} p_1$

长度 j 的“1”游程概率：$p[l_j^1] = p_1^{j-1} p_0$

$$\sum_{l_i^0=1}^{\infty} p[l_i^0] = \frac{p_1}{1-p_0} = 1 \qquad \sum_{l_j^1=1}^{\infty} p[l_j^1] = \frac{p_0}{1-p_1} = 1$$

$$\begin{pmatrix} L_0 \\ P(L_0) \end{pmatrix} = \begin{Bmatrix} l_1^0 = 1 & l_2^0 = 2 & l_3^0 = 3 \cdots & l_i^0 = i & \cdots \\ p_1 & p_0 p_1 & p_0^2 p_1 \cdots & p^{i-1} p_1 & \cdots \end{Bmatrix},$$

$$0 \le p(l_i^0) \le 1, \sum_{i=1}^{\infty} p(l_i^0) = 1$$

$$\begin{pmatrix} L_1 \\ P(L_1) \end{pmatrix} = \begin{Bmatrix} l_1^1 = 1 & l_2^1 = 2 & l_3^1 = 3 \cdots & l_j^1 = j & \cdots \\ p_0 & p_1 p_0 & p_1^2 p_0 \cdots & p_1^{j-1} p_0 & \cdots \end{Bmatrix},$$

$$0 \le p(l_j^1) \le 1, \sum_{j=1}^{\infty} p(l_j^1) = 1$$

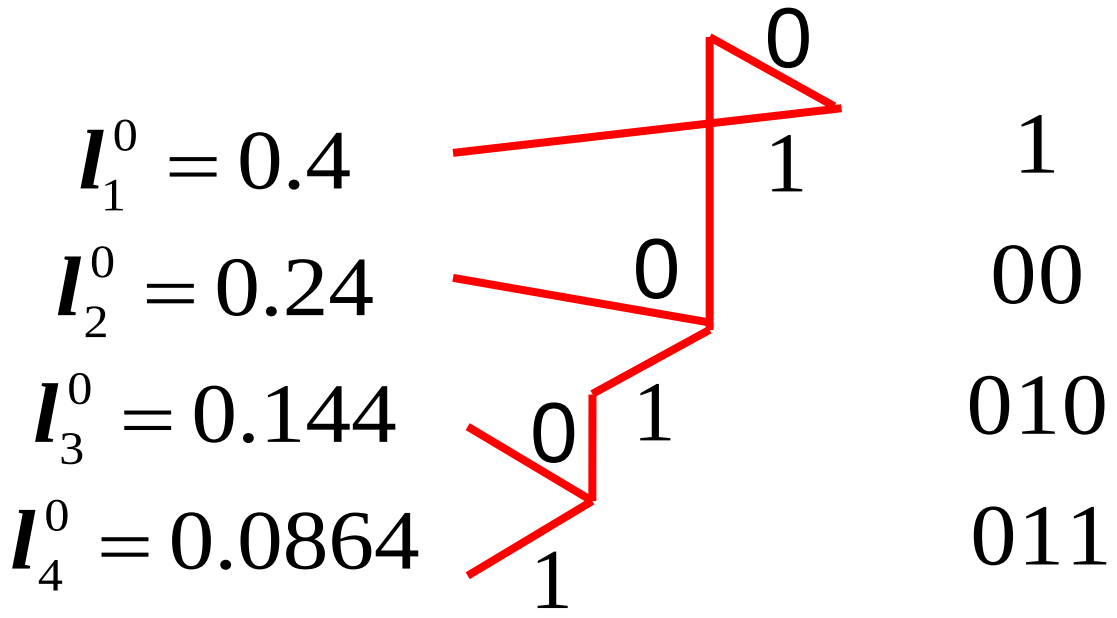
例 二元离散无记忆信源中，“0”和“1”发生的概率分别为

$$p_0 = 0.6, p_1 = 0.4$$

求二元序列的游程组合编码：00010111001000011110011000101110

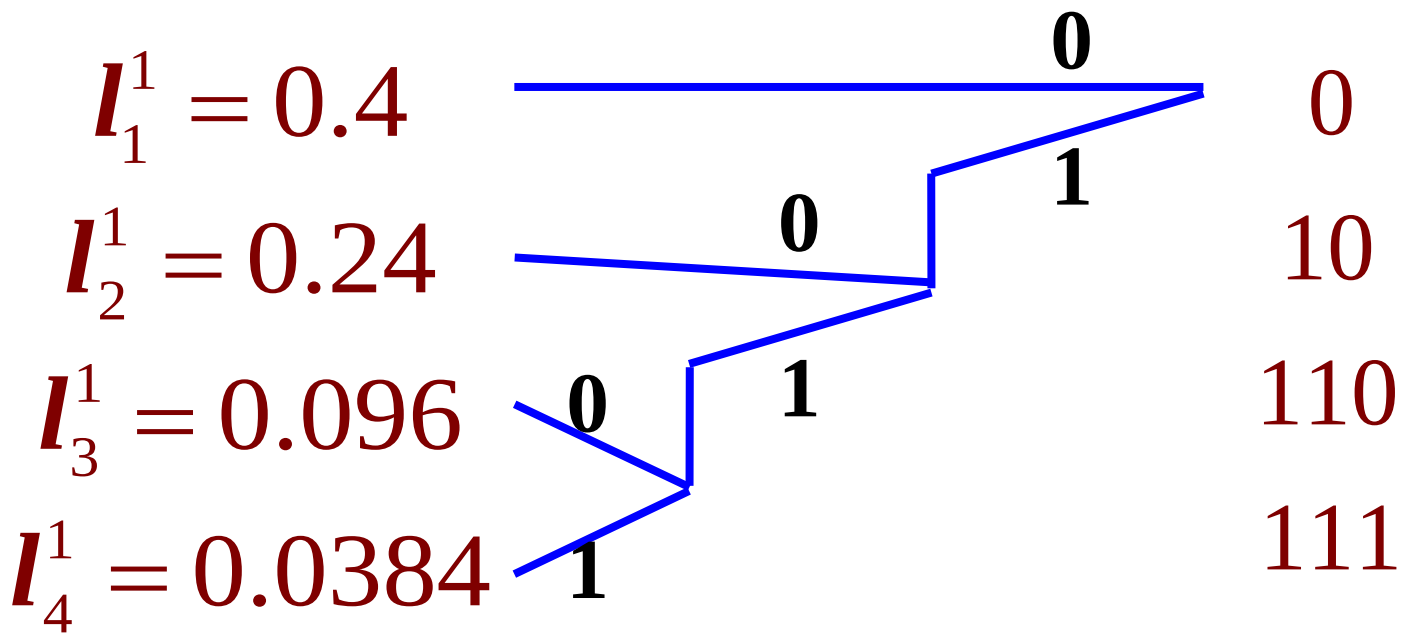
解：“0”游程长度信源

$$\begin{pmatrix} L_0 \\ P(L_0) \end{pmatrix} = \begin{Bmatrix} 1 & 2 & 3 & 4 & \cdots \\ 0.4 & 0.24 & 0.144 & 0.0864 & \cdots \end{Bmatrix}$$



“1”游程长度信源

$$\begin{pmatrix} L_1 \\ P(L_1) \end{pmatrix} = \begin{Bmatrix} 1 & 2 & 3 & 4 & \cdots \\ 0.6 & 0.24 & 0.096 & 0.0384 & \cdots \end{Bmatrix}$$



$$0 \rightarrow 1 \quad , \quad 1 \rightarrow 0$$

$$00 \rightarrow 00 \quad , \quad 11 \rightarrow 10$$

$$000 \rightarrow 010 \quad , \quad 111 \rightarrow 110$$

$$0000 \rightarrow 011 \quad , \quad 1111 \rightarrow 111$$

信号序列 0001011100100001111001100010101110

码序列 0100111000001111100100100111101

第3章 小 结

1. 掌握信源编码的基本概念

   分组码、非奇异码、唯一可译码、即时码

2. 掌握无失真编码定理

   定长编码定理、变长编码定理

   信息率、编码效率

3. 掌握三种编码方法

   香农、费诺、赫夫曼

   衡量指标：平均码长、信息率、编码效率