基础信息论

赫夫曼编码

华中科技大学电信学院

学习目标

- 编制赫夫曼码
- 评价赫夫曼码性能

赫夫曼编码

设有离散无记忆信源，$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1, & x_2, & \cdots, & x_i, & \cdots, & x_n \\ p(x_1), & p(x_2), & \cdots, & p(x_i), & \cdots, & p(x_n) \end{Bmatrix}$

二元码的编码步骤如下：

(1) 将信源符号按概率从大到小依次排列。设排序后的消息分别记为

$$x_1, x_2, \ldots, x_n$$

(2) 给两个概率最小的信源符号 $p(x_{n-1})$ 和 $p(x_n)$ 各分配一个码符号“0”和“1”，将这两个信源符号合并成一个新符号，并用 $p(x_{n-1}) + p(x_n)$ 作为新符号的概率，结果得到一个只包含 $n-1$ 个信源符号的新信源。将该信源称为第一次缩减信源，用 S_1 表示。

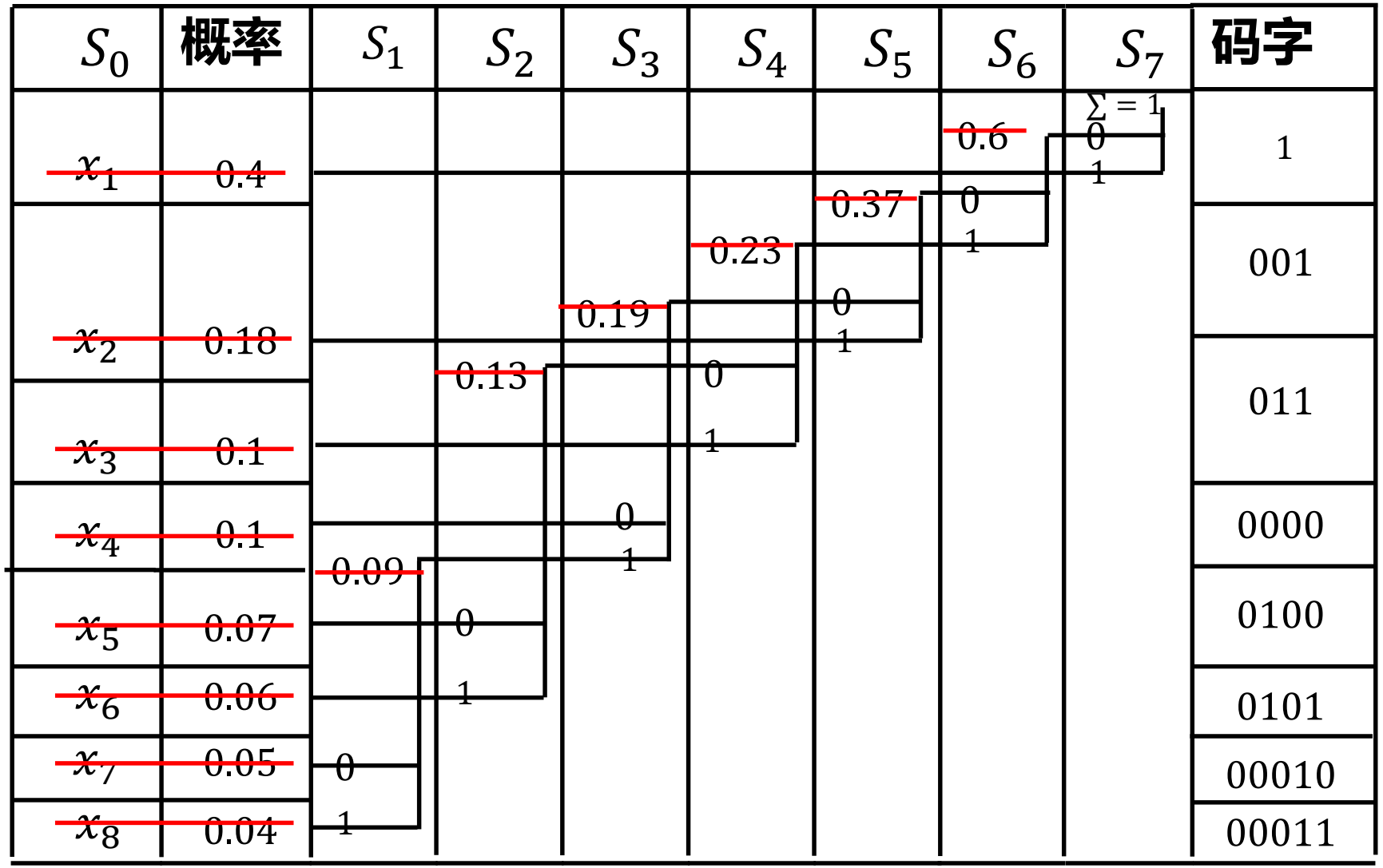
(3) 将缩减信源 S_1 的符号仍按概率从大到小的顺序排列，重复步骤2，得到只含 $(n-2)$ 个符号的缩减信源 S_2。

(4) 重复上述步骤，直至缩减信源只剩两个符号为止。此时所剩两个符号的概率之和必为1。然后从最后一级缩减信源开始，依编码路径向前返回，就得到各信源符号所对应的码字。

例 对 $\begin{Bmatrix} x_1' & x_2' & x_3' & x_4' & x_5' & x_6' & x_7' & x_8' \\ 0.18 & 0.07 & 0.04 & 0.4 & 0.06 & 0.1 & 0.1 & 0.05 \end{Bmatrix}$ 编二元赫夫曼码

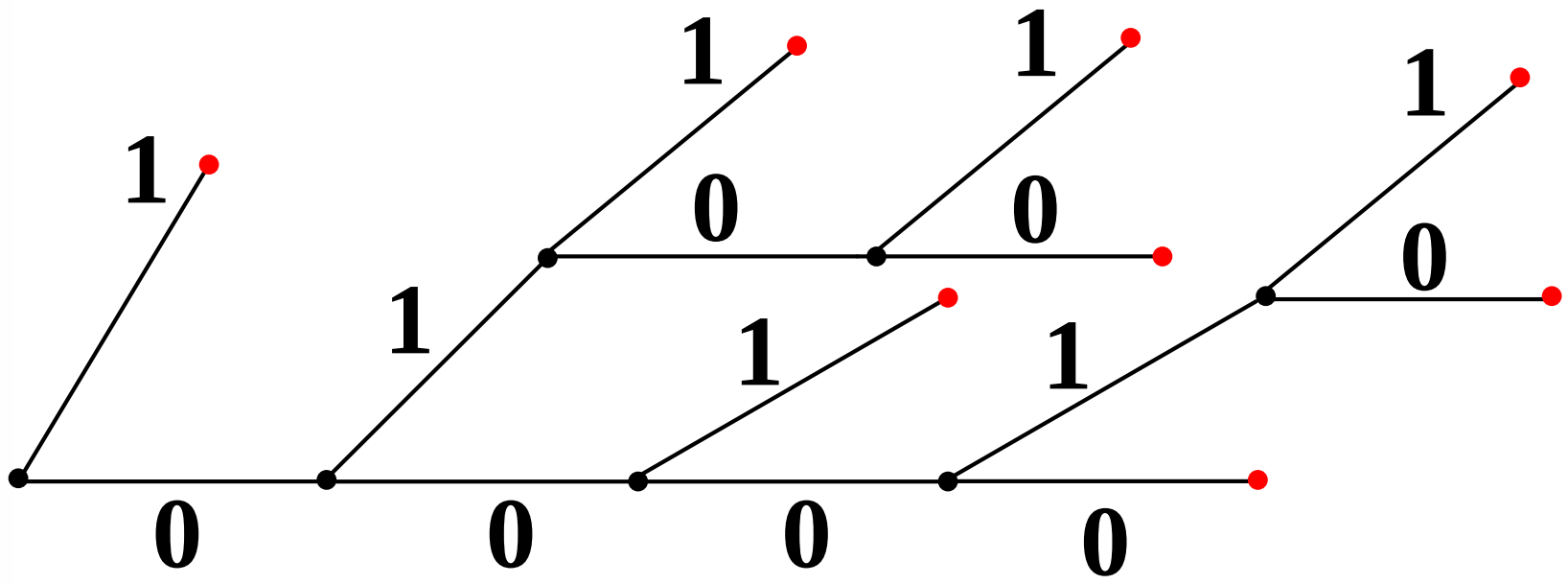
(1) 排序 $\begin{Bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{Bmatrix}$

(2) 第一次缩减信源 (3) 第二，三，…次缩减信源 (4) 最后一级



检验是否为即时码?

$$\begin{Bmatrix} X \\ P \\ \text{码字} \end{Bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7, & x_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \\ 1 & 001 & 011 & 0000 & 0100 & 0101 & 00010 & 00011 \end{Bmatrix}$$



$$\begin{Bmatrix} X \\ P \\ \text{码字} \end{Bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7, & x_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \\ 1 & 001 & 011 & 0000 & 0100 & 0101 & 00010 & 00011 \end{Bmatrix}$$

计算编码效率：

$$\eta = \frac{H(X)}{\frac{\bar{L} \cdot \log m}{N}} = \frac{-0.4\log 0.4 - 0.18\log 0.18 - \ldots - 0.04\log 0.0.04}{(0.4 \times 1 + 0.18 \times 3 + \ldots + 0.04 \times 5) \cdot \frac{\log 2}{1}}$$

$$= \frac{2.55}{2.61} = 97.7\%$$

从计算结果可看出，编码效率较高。

问题：赫夫曼方法所编的码字是否唯一?

习题

答案：不唯一。

1. 每次分配码字，0和1都是任意的。

2. 当新符号的概率与已有符号相等时，这些符号谁在前，谁在后，也会导致编码结果不同；而且所得编码方案性能也有差异。

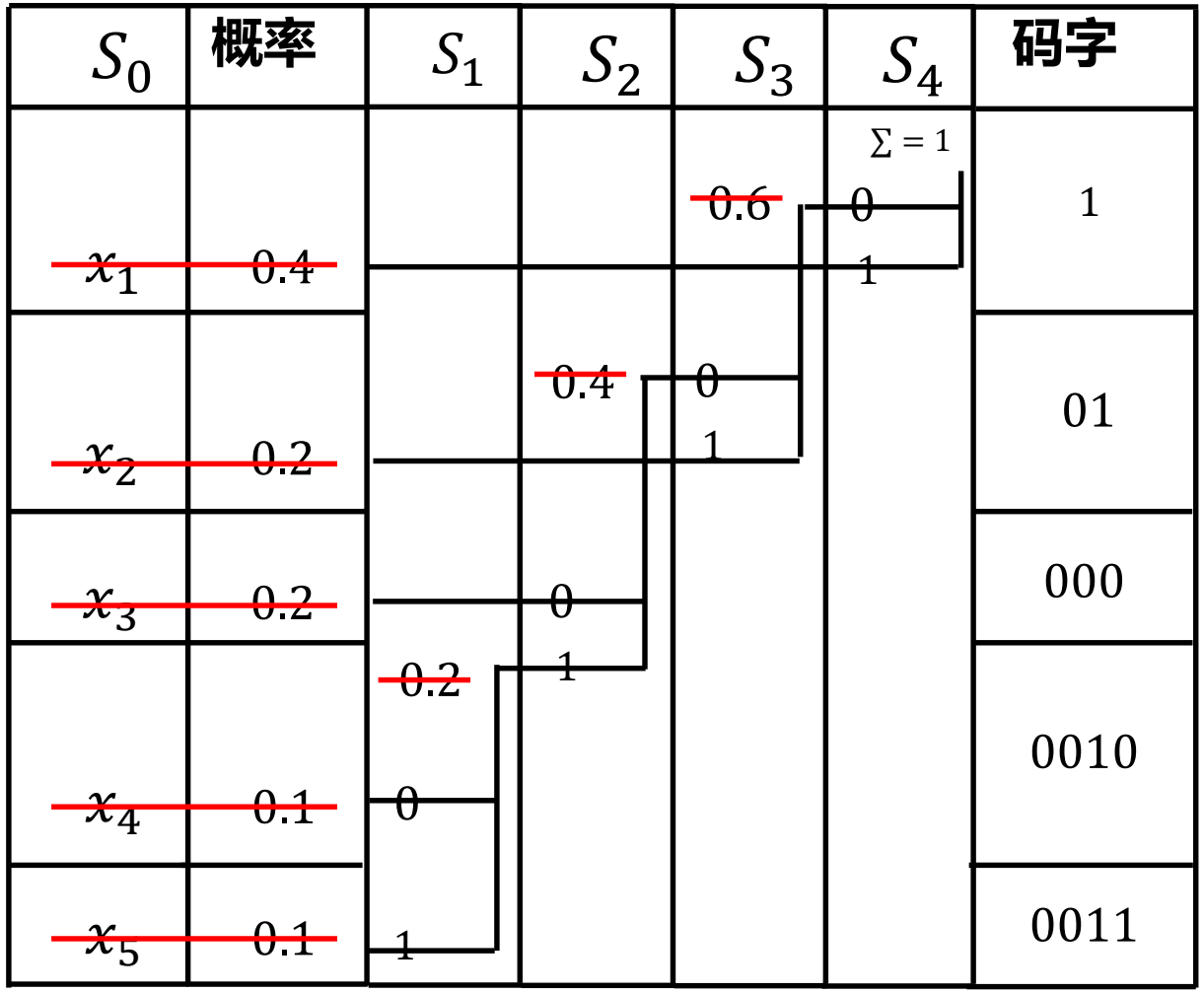
例 对 $\begin{Bmatrix} x_1' & x_2' & x_3' & x_4' & x_5' \\ 0.2 & 0.1 & 0.4 & 0.1 & 0.2 \end{Bmatrix}$ 编二元赫夫曼码。

解: (1) 排序 $\begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$

(2) 第一次缩减信源 (3) 第二，三，…次缩减信源 (4) 最后一级

方案一：

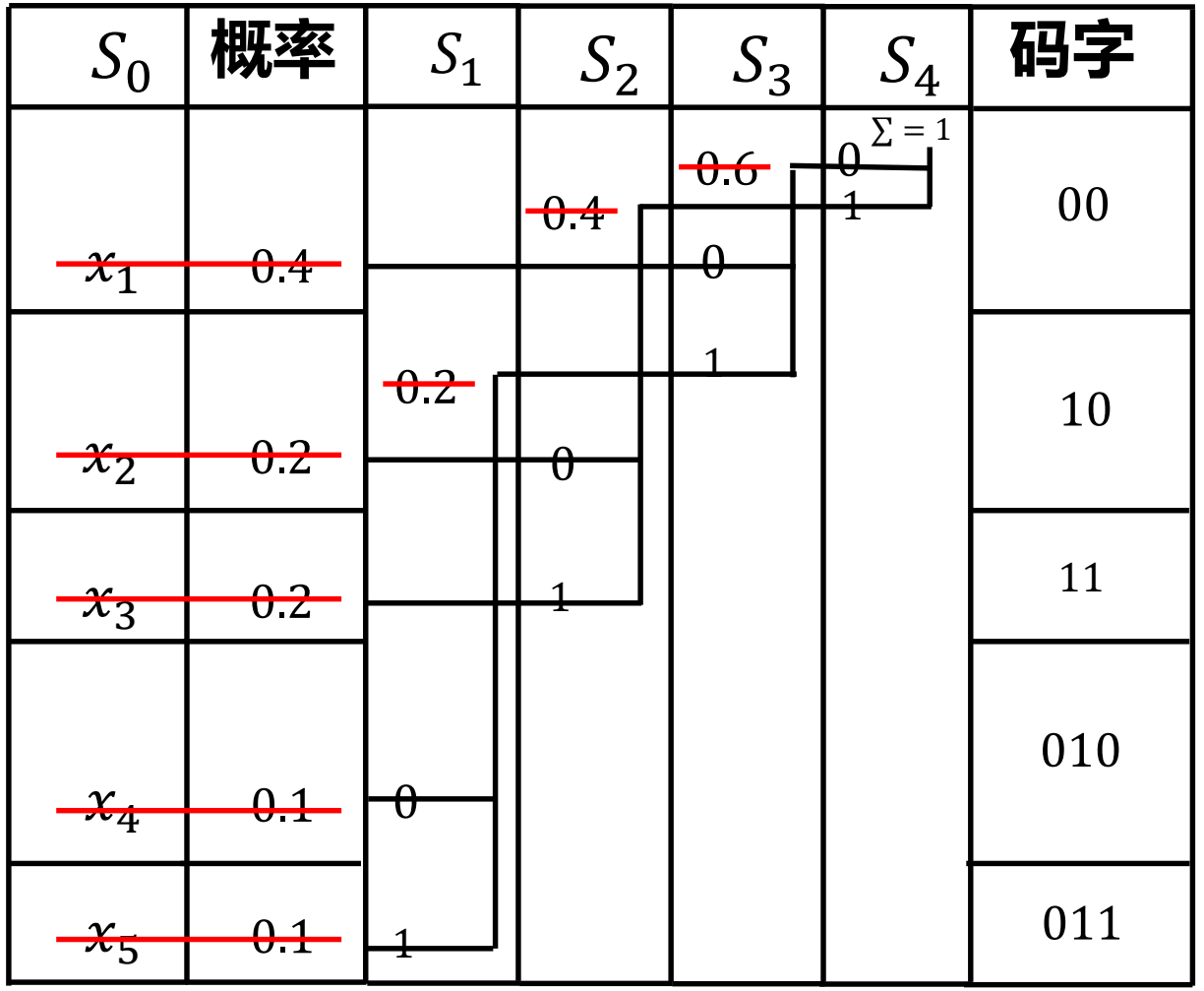
每次符号概率相等时，新码字都排在最后面。



$\bar{L} = 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 2 \times 0.1 \times 4 = 2.2$ **码元/符号**

方案二：

每次符号概率相等时，新码字都排在最上面。



$$\bar{L} = 0.4 \times 2 + 2 \times 0.2 \times 2 + 2 \times 0.1 \times 3 = 2.2$$ **码元/符号**

方案一

$$\begin{Bmatrix} X \\ P \\ \text{码字} \end{Bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \\ 1 & 01 & 000 & 0010 & 0011 \end{Bmatrix}$$

方案二

$$\begin{Bmatrix} X \\ P \\ \text{码字} \end{Bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \\ 00 & 10 & 11 & 010 & 011 \end{Bmatrix}$$

从直观上看，方案二的各码字之间，码字长度更均匀。

$$\sigma_1^2 = [(1-2.2)^2 \cdot 0.4 + (2-2.2)^2 \cdot 0.2 + \ldots + (4-2.2)^2 \cdot 0.1] = 1.36$$

$$\sigma_2^2 = [(2-2.2)^2 \cdot 0.4 + (2-2.2)^2 \cdot 0.2 + \ldots + (3-2.2)^2 \cdot 0.1] = 0.16$$

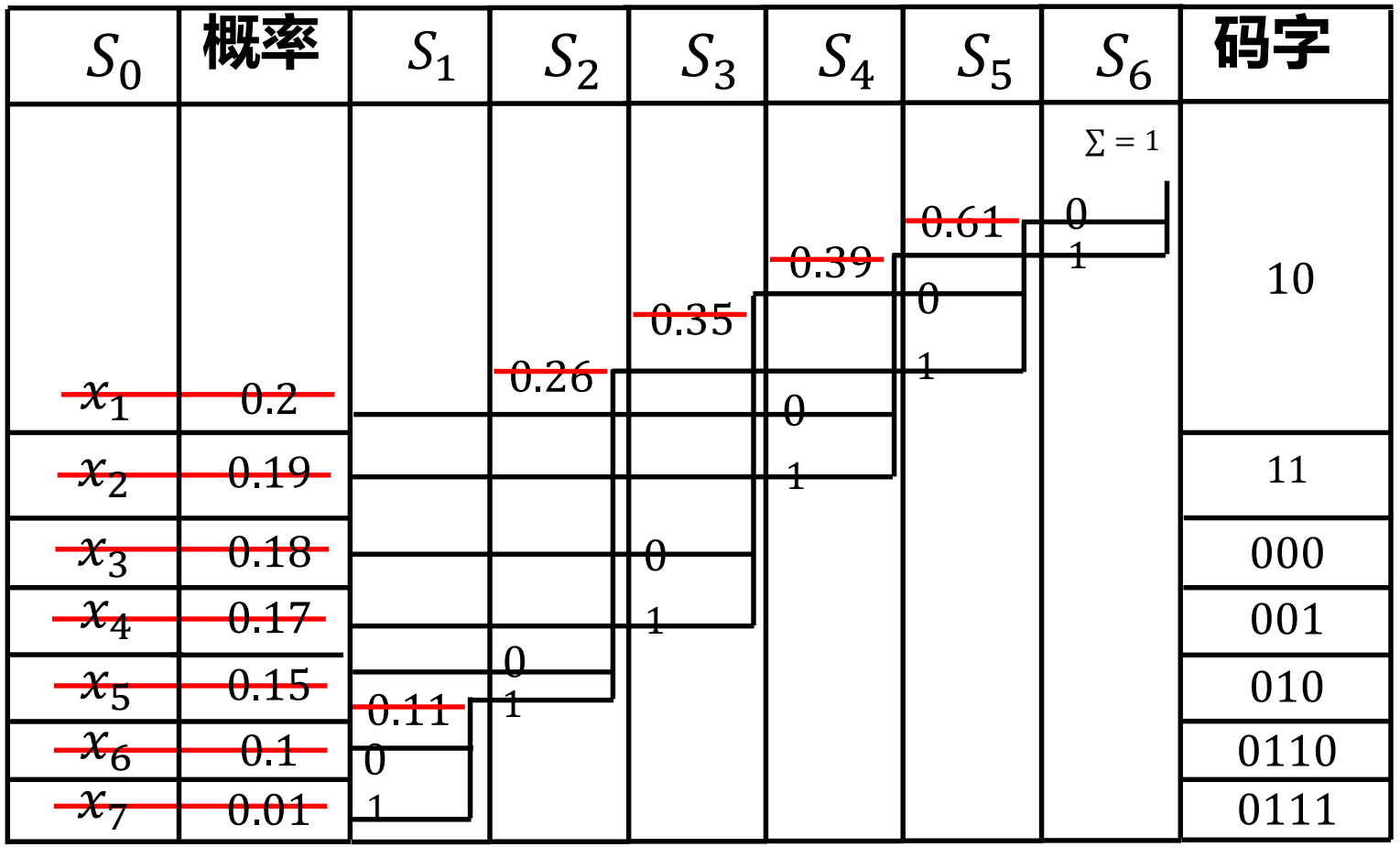
由此得出结论：

在赫夫曼编码过程中，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上的位置，这样可使合并后的新符号重复编码次数减少，码字间长度更加均匀。

练习：设有信源 $\begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7 \\ 0.2 & 0.19 & 0.18 & 0.17 & 0.15 & 0.1 & 0.01 \end{Bmatrix}$

(1) 编二进制赫夫曼码　　(2) 计算平均码长及编码效率

(1)



(2)计算平均码长及编码效率

$$\begin{Bmatrix} X \\ P \\ \text{码字} \end{Bmatrix} = \begin{Bmatrix} x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7 \\ 0.2 & 0.19 & 0.18 & 0.17 & 0.15 & 0.1 & 0.01 \\ 10 & 11 & 000 & 001 & 010 & 0110 & 0111 \end{Bmatrix}$$

$$\bar{L} = 0.2 \times 3 + 0.19 \times 2 + \ldots + 0.01 \times 4 = 2.72$$

比特/符号

$$\eta = \frac{H(X)}{\frac{\bar{L} \cdot \log m}{N}} = \frac{2.609}{\frac{2.72 \cdot \log 2}{1}} \approx 95.9\%$$

多元赫夫曼编码

多(m)元码的编码步骤：

(1) 按概率从大到小排序。

(2) 挑出概率最小的m个信源符号，分别赋予码符号$0,1,\ldots,m-1$，并将概率的和赋予合并后的新符号， 得到S_1。

(3) 按相同步骤计算$S_2,S_3,\ldots$。

(4) 直到最后一级，倒序读出码字。

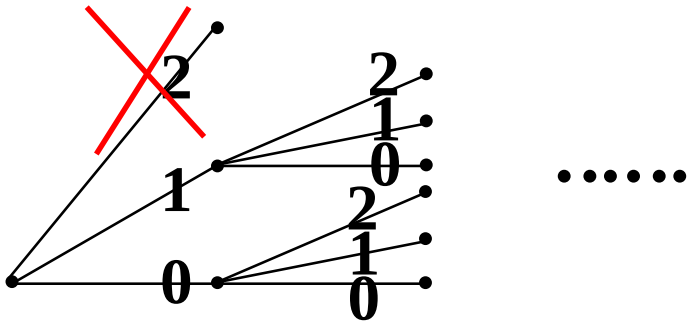
多元码的编码过程与二元码基本类似，但可能会遇到如下问题：

二元码		S_0	S_1	S_2	…	最后一级
	一般情况　信源符号数	n	n-1	n-2	…	2
	举例	8	7	6	…	2

m元码		S_0	S_1	S_2	…	最后一级
	一般情况　信源符号数	n	n-(m-1)	n-2(m-1)	…	$\leq m$
	举例(m=3)	8	6	4	…	2

可能出现最后一级信源，信源符号数不足m个的情况。

S_0	概率	S_1	S_2	S_3	S_4	码字
x_1	0.4			0.6	$\Sigma=1$ 0 1	1
~~x_2~~	~~0.18~~		~~0.27~~	0 1 2		01
~~x_3~~	~~0.1~~	~~0.15~~	0			000
~~x_4~~	~~0.1~~		1			001
~~x_5~~	~~0.07~~		2			002
~~x_6~~	~~0.06~~	0				020
~~x_7~~	~~0.05~~	1				021
~~x_8~~	~~0.04~~	2				022



在码树图中，某些分枝从第一级就被砍掉，从而造成平均码长偏长的情况。

解决办法：从最后一级缩减信源倒着排，保证最后一级有m个符号。

要求最后一级缩减信源保证有m个符号，则有：

倒数第二级 → 倒数第三级 …… 第一级

$m+(m-1)$　　$m+2\cdot(m-1)$　　$m+k\cdot(m-1)\geq n$

为保证每次缩减均为m个变1个，第一级符号所缺的个数：　　$[m+k\cdot(m-1)]-n$

第一级符号保留的个数：　　$m-[m+k\cdot(m-1)-n]$

【保留+欠缺=m】　　$=n-k\cdot(m-1)$

其中：$k\geq(n-m)/(m-1)$ *

例：$n=8,m=3$，求：第一级信源符号所需保留个数。

$k\geq\dfrac{n-m}{m-1}=\dfrac{8-3}{3-1}=2.5$　　$\therefore\quad k=3$

第一级保留　　$8-3\cdot(3-1)=2$

S_0	概率	S_1	S_2	S_3	S_4	码字
x_1	0.4				$\Sigma=1$	0
				0.38	0	
			0.22		1	
~~x_2~~	~~0.18~~			0	2	10
~~x_3~~	~~0.1~~			1		11
~~x_4~~	~~0.1~~			2		12
		~~0.09~~	0			
~~x_5~~	~~0.07~~		1			21
~~x_6~~	~~0.06~~		2			22
~~x_7~~	~~0.05~~	0				200
~~x_8~~	~~0.04~~	1				201

方案1：

$$\bar{L}_1 = 0.4\times 1 + 0.18\times 2 + (0.1+0.1+0.07+0.06+0.05+0.04)\times 3$$

$$= 2.02 \quad \textbf{码元/符号}$$

方案2：

$$\bar{L}_2 = 0.4\times 1 + (0.18+0.1+0.1+0.07+0.06)\times 2 + (0.05+0.04)\times 3$$

$$= 1.69 \quad \textbf{码元/符号}$$

计算方案2的编码效率：

$$\eta=\frac{H(X)}{\frac{\overline{L}\cdot\log m}{N}}=\frac{-0.4\log0.4-0.18\log0.18-\ldots-0.04\log0.04}{1.69\cdot\frac{\log 3}{1}}=\frac{2.55}{1.69\cdot\frac{\log 3}{1}}=95.2\%$$

练习：

x_1,	x_2,	x_3,	x_4,	x_5,	x_6,	x_7,	x_8,	x_9,	x_{10}
0.16	0.14	0.13	0.12	0.1	0.09	0.08	0.07	0.06	0.05

(1) 编三进制赫夫曼码　　(2) 计算平均码长及编码效率

解：(1) 关键是求第一级信源应保留的符号个数

$$k\geq\frac{n-m}{m-1}=\frac{10-3}{3-1}=3.5 \quad \therefore \quad k=4$$

第一级应保留 $n-k\cdot(m-1)=10-4\cdot(3-1)=2$

S_0	概率	S_1	S_2	S_3	S_4	S_5	码字
					0.43	$\Sigma=1$ 0	00
				0.33		1	
			0.24			2	
x_1	0.16				0		
x_2	0.14				1		01
x_3	0.13				2		02
x_4	0.12			0			10
		0.11		1			
x_5	0.1			2			12
x_6	0.09		0				20
x_7	0.08		1				21
x_8	0.07		2				22
x_9	0.06	0					110
x_{10}	0.05	1					111

(2) 计算平均码长及编码效率

$$\bar{L} = (0.16 + 0.14 + 0.13 + 0.12 + 0.1 + 0.09 + 0.08 + 0.07) \times 2 + (0.06 + 0.05) \times 3$$

$= 2.11$ **码元/符号**

$$H(X) = -0.16 \log 0.16 - \ldots\ldots - 0.05 \log 0.05$$

$= 3.23$ **比特/符号**

$$\eta = \frac{H(X)}{\bar{L} \cdot \log m} \approx 96.6\%$$

定理：赫夫曼码是紧致码。

说明：这里只证明二元赫夫曼码是紧致码，其结论可推广到多元赫夫曼码。

思路：采用类似数学归纳法的证明方法。

证明最后一级缩减信源是紧致码 → **假设 S_j 级缩减信源是紧致码** → **证明 S_{j-1} 级缩减信源是紧致码**

证明：

对于二元码，最后一级缩减信源只有2个信源符号，因此一定是紧致码。

假设 S_j 级缩减信源对应的编码 C_j 是紧致码，因此有：

消息数

$$\bar{K}_j = \sum_{i=1}^{k} p(x_{j\_i}) \cdot l_{j\_i}$$

第 i 个消息的码长

S_j 级缩减信源的平均码长

第 i 个消息的概率

在所有可能的唯一可译码编码方案中，$\bar{K}_j$ 的长度最短。

对于 S_{j-1} 级缩减信源，一定有 $k+1$ 条消息，而且其中某两个消息的概率的和一定为 S_j 级某消息的概率。

$S_{j-1}(C_{j-1})$			$S_j(C_j)$		
x_{j-1_1}	$p(x_{j-1\_1})$	l_{j-1_1}	x_{j_1}	$p(x_{j\_1})$	l_{j_1}
x_{j-1_2}	$p(x_{j-1\_2})$	l_{j-1_2}	x_{j_2}	$p(x_{j\_2})$	l_{j_2}
⋮	⋮	⋮	⋮	⋮	⋮
			x_{j_m}	$p(x_{j\_m})$	l_{j_m}
x_{j-1_k}	$p(x_{j-1\_k})$	$l_{j\_m}+1$	⋮	⋮	⋮
x_{j-1_k+1}	$p(x_{j-1\_k+1})$	$l_{j\_m}+1$	x_{j_k}	$p(x_{j\_k})$	l_{j_k}

+= 与码长无关的常数

$$\bar{K}_{j-1} = \sum_{i=1}^{k+1} p(x_{j-1\_i}) \cdot l_{j-1\_i} = \sum_{i=1}^{k-1} p(x_{j-1\_i}) \cdot l_{j-1\_i} + [p(x_{j-1\_k}) \cdot (l_{j\_m} + 1)$$

$$+p(x_{j-1\_k+1}) \cdot (l_{j\_m} + 1)] = \sum_{i=1}^{k} p(x_{j\_i}) \cdot l_{j\_i} + p(x_{j\_m}) = \bar{K}_j + p(x_{j\_m})$$

∴ 只要 C_j 是紧致码，则缩减前信源的编码 C_{j-1} 也是紧致码

谢谢！

黑晓军
华中科技大学
电子信息与通信学院
Email: heixj@hust.edu.cn
网址： http://eic.hust.edu.cn/aprofessor/heixiaojun