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. 香农码

编码步骤如下：

1. 将信源符号按概率从大到小顺序排列，为方便起见，令

$$p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$$

2. 按下式计算第*i*个符号对应的码字的码长（要取整）

$$-\log p(s_i) \leq l_i < -\log p(s_i) + 1$$

3. 计算第*i*个符号的累加概率

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

4. 将累加概率变换成二进制小数，取小数点后l_i位数作为第*i*个符号的码字。

例5.6　对如下信源编码:

$$\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{Bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{Bmatrix}$$

信源符号	符号概率	累加概率		码长	码字
s_1	0.20	0	2.34	3	000
s_2	0.19	0.2	2.41	3	001
s_3	0.18	0.39	2.48	3	011
s_4	**0.17**	**0.57**	**2.56**	**3**	**100**
s_5	0.15	0.74	2.74	3	101
s_6	0.10	0.59	3.34	4	1110
s_7	0.01	0.99	6.66	7	1111110

平均码长

$$\bar{L} = \sum_{i=1}^{q} p(s_i) l_i = 3.14 \quad \text{码元符号 / 信源符号}$$

信源熵

$$H(S) = -\sum_{i=1}^{q} p(s_i) \log p(s_i) = 2.61 \quad \text{比特 / 信源符号}$$

编码效率

$$\eta = \frac{2.61}{3.14} = 83.1\%$$

结论：

1) $H(S) \le \bar{L} < H(S) + 1$

2) 香农码是即时码，但冗余度稍大，不是最佳码。

2. 香农-费诺-埃利斯编码

1、不用对信源符号按概率大小排序。

2、直接计算修正的累加概率 $\overline{F}(s_i)=\sum_{k=1}^{i-1}p(s_k)+\frac{1}{2}p(s_i)$

3、计算码长 $l_i=\lceil -\log p(s_i)\rceil+1$

3. Huffman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，令

$$p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$$

2. 给两个概率最小的信源符号s_{n-1}和s_n各分配一个码元“0”和“1”，并将这两个信源符号合并成一个新符号，并用这两个最小的概率之和作为新符号的概率，结果得到一个只包含$(n-1)$个信源符号的新信源。称为信源的第一次缩减信源，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。

离散信源如下：

$$\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{Bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{Bmatrix}$$

解：Huffman编码结果如下：

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7
码字	10	11	000	001	010	0110	0111

- 平均码长为

$$\overline{L} = \sum_{i=1}^{7} p(s_i) l_i = 2.72 \quad \text{码元符号 / 信源符号}$$

- 信源熵为

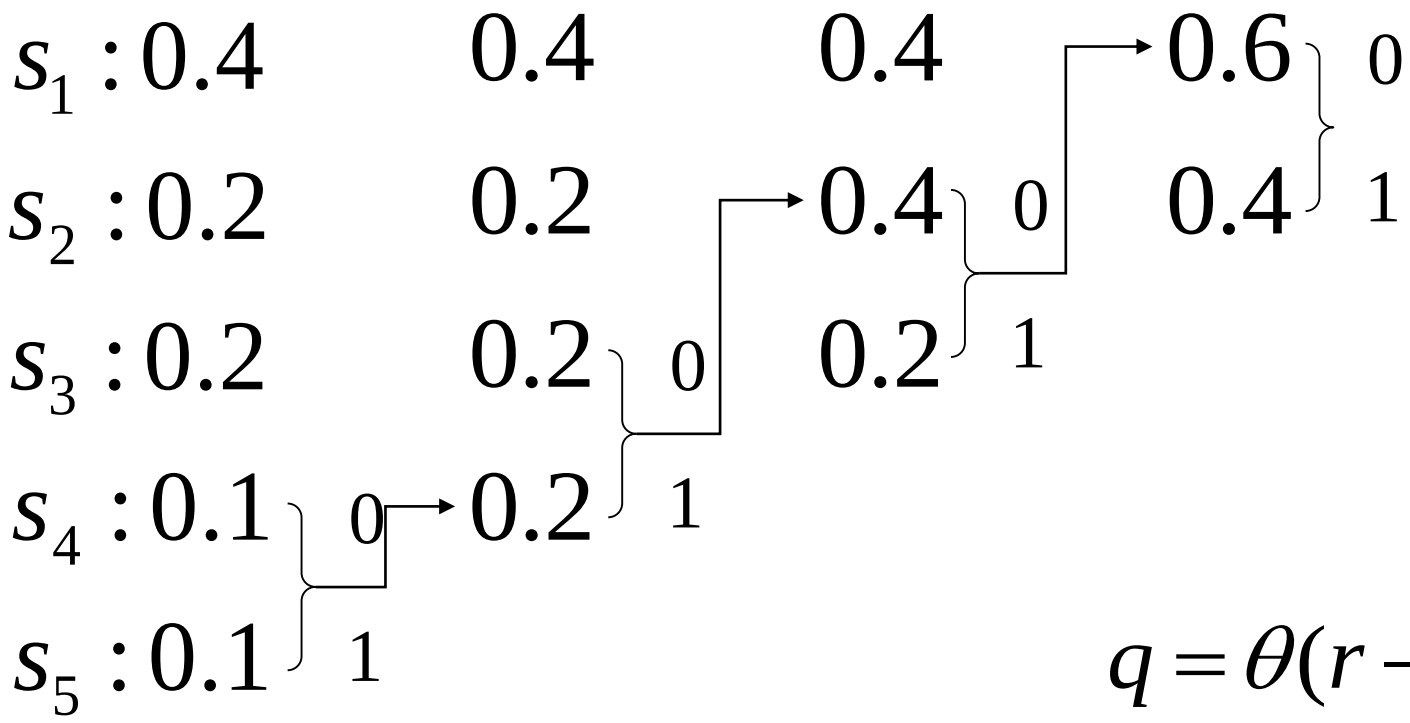
$$H(S) = -\sum_{i=1}^{7} p(s_i) \log p(s_i) = 2.61 \quad \text{比特 / 信源符号}$$

- 编码效率为

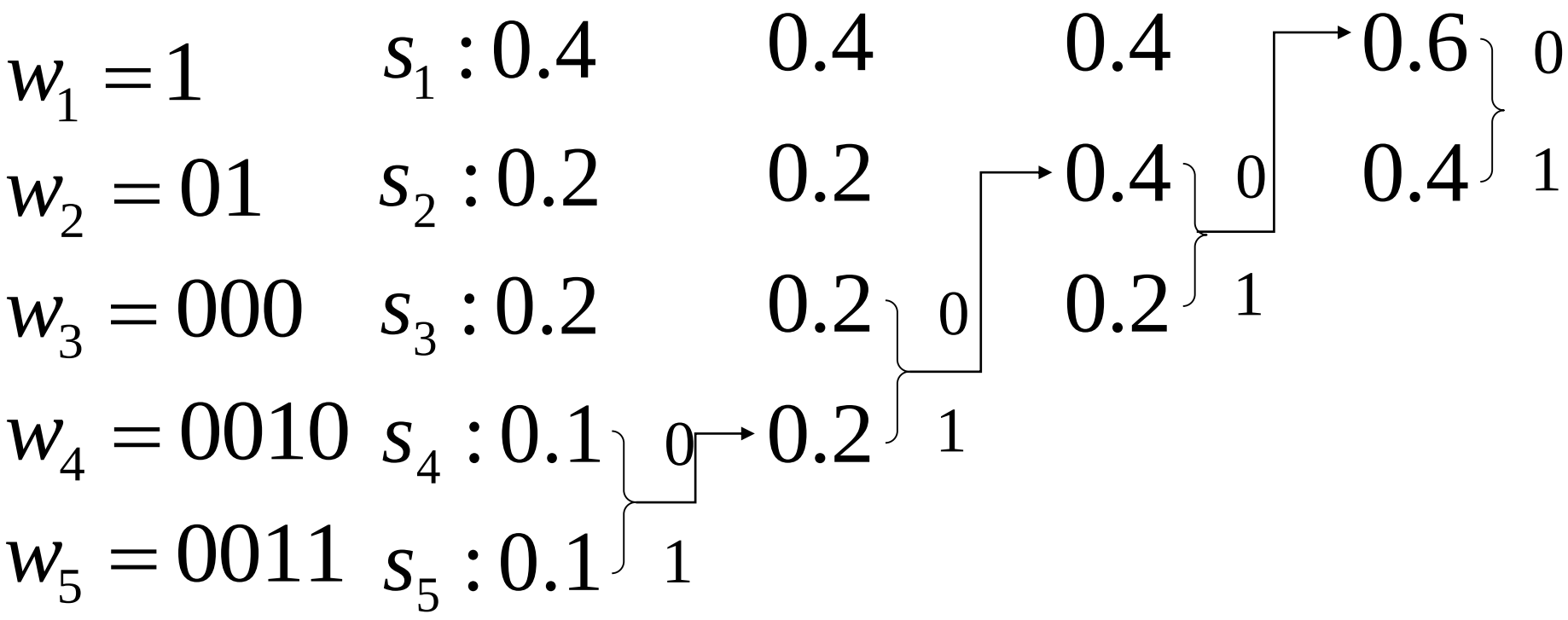
$$\eta = \frac{2.61}{2.72} = 96.0\%$$

例：$[S\cdot P]:\begin{cases} s_1 & s_2 & s_3 & s_4 & s_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{cases}$ $X:\{0,1\}$

$S:$	$S_1:$	$S_2:$	$S_{末}:$
q	$q-(r-1)$	$q-2(r-1)$	r

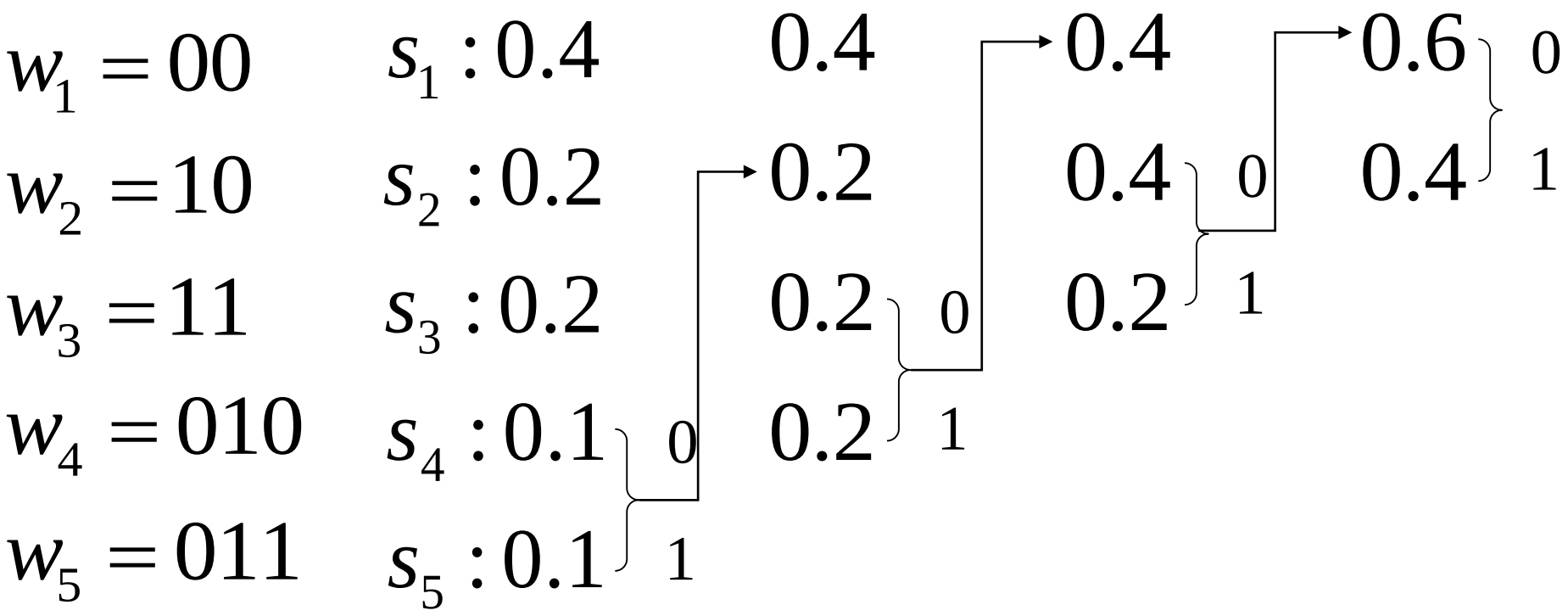


$$q=\theta(r-1)+r$$



$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i = 2.2$$

码符号/信源符号



$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i = 2.2$$

码符号/信源符号

讨论：

1) 两种方法平均码长相等。

2) 计算两种码的码长方差：

$$\sigma_1^{\ 2} = \sum_{i=1}^{5} p(s_i)(l_i - \overline{L})^2 = 1.36$$

$$\sigma_2^{\ 2} = \sum_{i=1}^{5} p(s_i)(l_i - \overline{L})^2 = 0.16$$

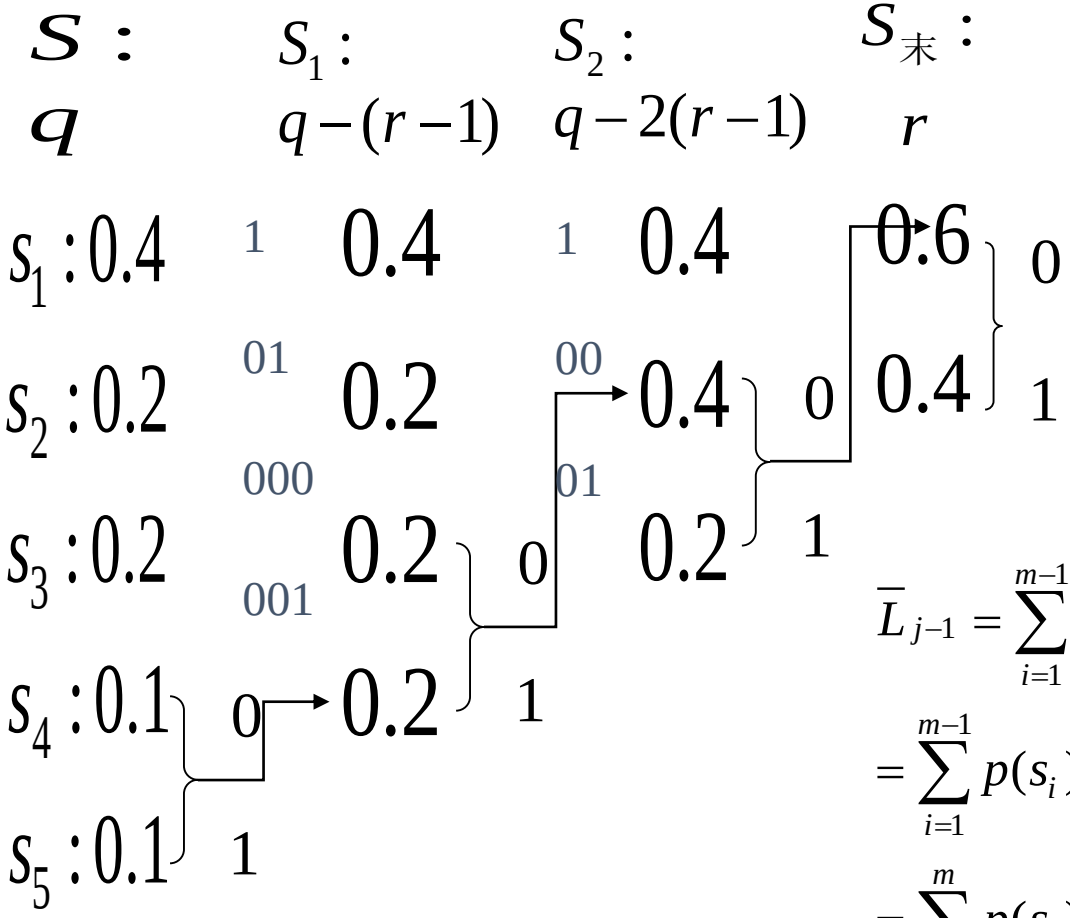
第二种方法编出的码码字长度变化较小，便于实现。

注意：霍夫曼编码后的码字不是惟一的。

1）每次对缩减信源两个概率最小的符号分配“0”或“1”码元是任意的，所以可得到不同的码字。不同的码元分配，得到的具体码字不同，但码长l_i不变，平均码长也不变，所以没有本质区别；

2）缩减信源时，若合并后的概率与其他概率相等，这几个概率的次序可任意排列，但得到的码字不相同，对应的码长也不相同，但平均码长也不变。

定理5.8　霍夫曼码是紧致码



假定缩减后信源为S_j共有m个元素。

缩减后信源为 S_{j-1}共有m+1个元素。

$$\bar{L}_j = \sum_{i=1}^{m} p(s_i)\cdot l_i$$　　最短

其中第k个元素码长 l_k，概率为

$$p(s_k) = p(s_{k_0}) + p(s_{k_1})$$

则缩减前

$$\bar{L}_{j-1} = \sum_{i=1}^{m-1} p(s_i)\cdot l_i + p(s_{k_0})(l_k+1) + p(s_{k_1})(l_k+1)$$

$$= \sum_{i=1}^{m-1} p(s_i)\cdot l_i + \left[p(s_{k_0}) + p(s_{k_1})\right] l_k + p(s_{k_0}) + p(s_{k_1})$$

$$= \sum_{i=1}^{m} p(s_i)\cdot l_i + p(s_{k_0}) + p(s_{k_1})$$

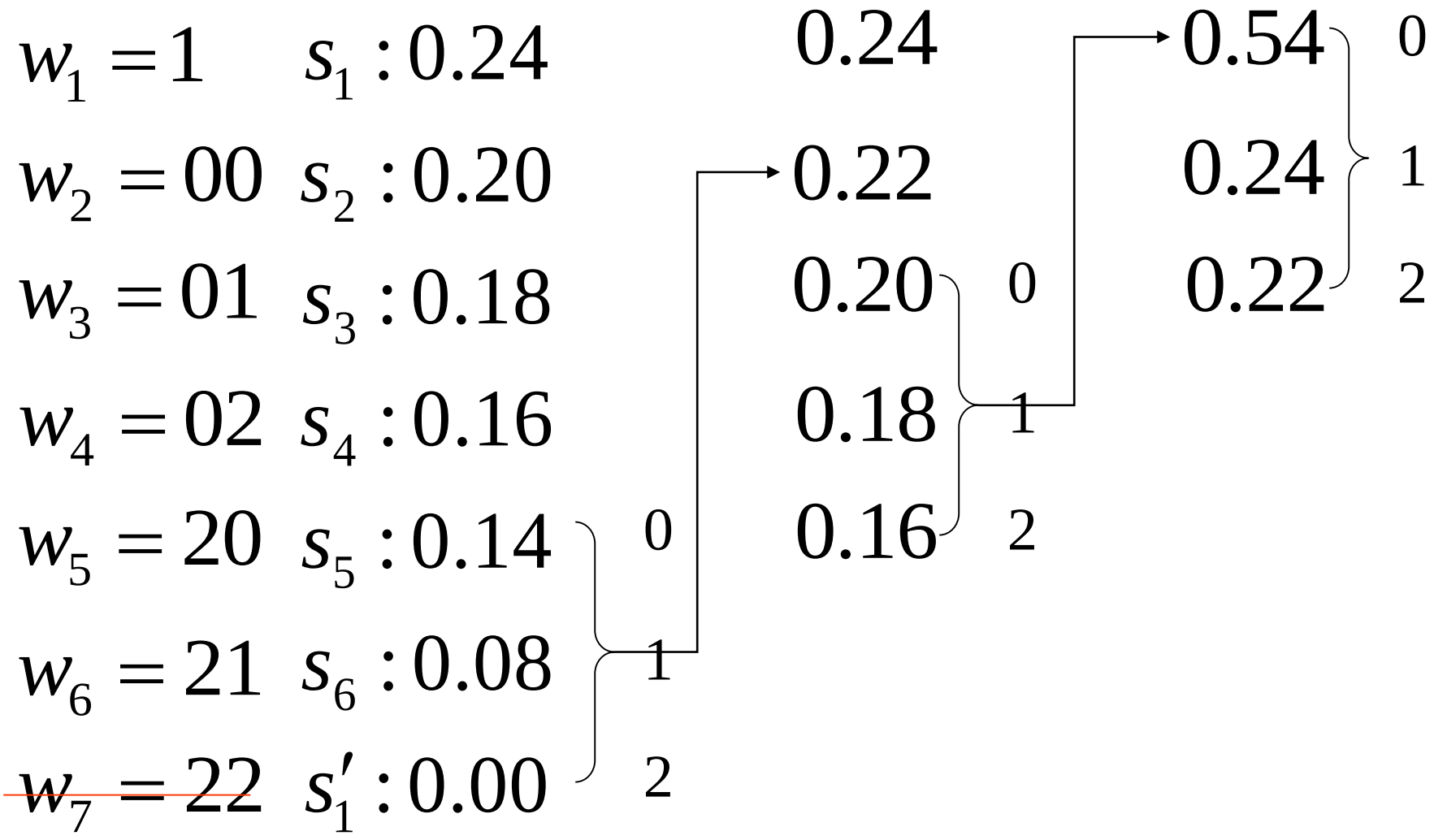
4. r 元霍夫曼码

$$[S \cdot P]: \begin{cases} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ 0.24 & 0.20 & 0.18 & 0.16 & 0.14 & 0.08 \end{cases}$$

$$X: \{0,1,2\}$$

$$q + i = \theta(r-1) + r$$

$$q = \theta(r-1) + r \qquad i = 1$$


$w_1 = 1$ $s_1 : 0.24$
$w_2 = 00$ $s_2 : 0.20$
$w_3 = 01$ $s_3 : 0.18$
$w_4 = 02$ $s_4 : 0.16$
$w_5 = 20$ $s_5 : 0.14$
$w_6 = 21$ $s_6 : 0.08$
$w_7 = 22$ $s_1' : 0.00$
0
1
2
0.24
0.22
0.20
0.18
0.16
0
1
2
0.54
0.24
0.22
0
1
2

$$[S \cdot P]: \begin{cases} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.2 & 0.1 & 0.1 & 0.05 & 0.05 & 0.05 & 0.05 \end{cases}$$

$$X: \{0,1,2\}$$

5. Fano码

编码步骤如下：

1. 将概率按从大到小的顺序排列，令

$$p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$$

2. 将依次排列的信源符号按概率分成两组，使每组概率和尽可能接近或相等。
3. 给每一组分配一位码元“0”或“1”。
4. 将每一分组再按同样方法划分，重复步骤2和3，直至概率不再可分为止。

例

$$\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{Bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{Bmatrix}$$

解：

信源符号	符号概率	第一次分组	第一次分组	第一次分组	第一次分组	码字	码长
s_1	0.20	0	0			00	2
s_2	0.19		1	0		010	3
s_3	0.18			1		011	3
s_4	0.17	1	0			10	2
s_5	0.15		1	0		110	3
s_6	0.10			1	0	1110	4
s_7	0.01				1	1111	4

- 平均码长为

$$\overline{L} = \sum_{i=1}^{7} p(s_i) l_i = 2.74 \quad \text{码元符号 / 信源符号}$$

- 信源熵为

$$H(S) = -\sum_{i=1}^{7} p(s_i) \log p(s_i) = 2.61 \quad \text{比特 / 信源符号}$$

- 编码效率为

$$\eta = \frac{2.61}{2.74} = 95.3\%$$

例

$$[S \cdot P]: \begin{cases} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ 0.32 & 0.22 & 0.18 & 0.16 & 0.08 & 0.04 \end{cases}$$

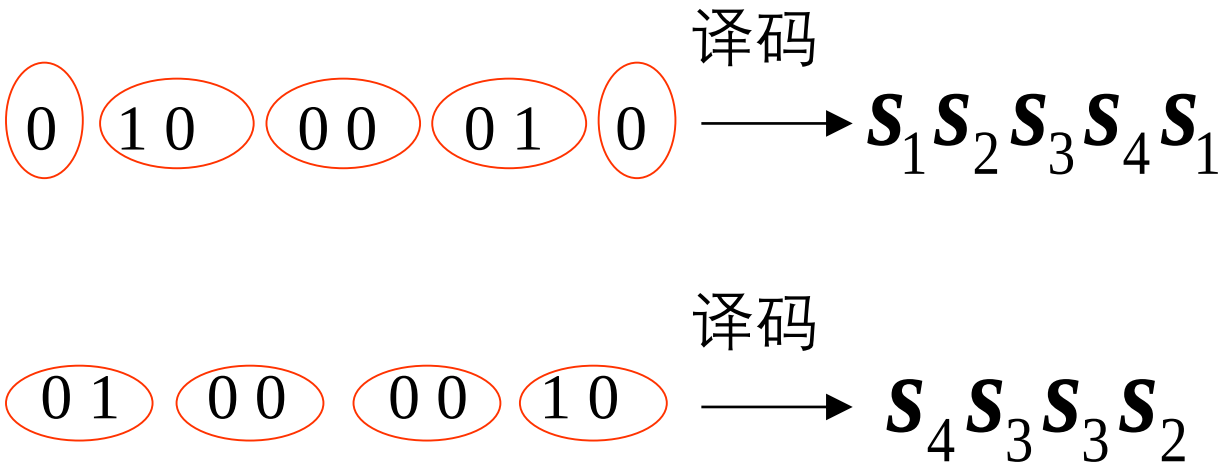
香农码、Huffman码、Fano码总结

- 香农码、费诺码、霍夫曼码都考虑了信源的统计特性，使经常出现的信源符号对应较短的码字，使信源的平均码长缩短，从而实现了对信源的压缩。
- 香农码编码结果唯一，但在很多情况下编码效率不是很高。
- 费诺码和霍夫曼码的编码方法都不唯一。
- 费诺码比较适合于对分组概率相等或接近的信源编码。
- 霍夫曼码对信源的统计特性没有特殊要求，编码效率比较高，对编码设备的要求也比较简单，因此综合性能优于香农码和费诺码。

Huffman码编码应用中的一些问题

首先是速率匹配问题

其次是差错扩散问题



第三是霍夫曼码需要查表来进行编译码。

信源统计特性未知时，怎么办？可采用所谓通用编码的方法。