

第四章作业题答案

1. 有一信源，它有 6 个可能的输出，其概率分布如下表所示，表中给出了对应的码 A、B、C、D、E 和 F。

消息	$p(a_i)$	A	B	C	D	E	F
a_1	1/2	000	0	0	0	0	0
a_2	1/4	001	01	10	10	10	100
a_3	1/8	010	011	110	110	1100	101
a_4	1/16	011	0111	1110	1110	1101	110
a_5	1/32	100	01111	11110	1011	1110	111
a_6	1/32	101	011111	111110	1101	1111	011

Answer: 唯一可译码是ABCE

非延长码为ACE

A的平均码长: $\bar{n} = \sum_{i=1}^6 p(s_i)n_i = 3(1/2 + 1/4 + 1/16 + 1/16 + 1/16 + 1/16) = 3$ 码符号 / 信源符号

编码效率: $\eta = \frac{H(S)}{\bar{n} \log r} = \frac{2}{3} \times 100\% = 66.67\%$

B的平均码长 $\bar{n} = \sum_{i=1}^6 p(s_i)n_i = 1*1/2 + 2*1/4 + 3*1/16 + 4*1/16 + 5*1/16 + 6*1/16$
 $= 17/8 = 2.125$ 码符号 / 信源符号

编码效率: $\eta = \frac{H(S)}{\bar{n} \log r} = \frac{2}{2.125} \times 100\% = 94.12\%$

C的平均码长 $\bar{n} = \sum_{i=1}^6 p(s_i)n_i = 2.125$ 码符号 / 信源符号

编码效率: $\eta = \frac{H(S)}{\bar{n} \log r} = \frac{2}{2.125} \times 100\% = 94.12\%$

E的平均码长 $\bar{n} = \sum_{i=1}^6 p(s_i)n_i = 2*1/2 + 3*1/4 + 3*1/16 + 3*1/16 + 3*1/16 + 3*1/16$
 $= 10/4 = 2.5$ 码符号 / 信源符号

C4的编码效率: $\eta = \frac{H(S)}{\bar{n} \log r} = \frac{2}{2.5} \times 100\% = 80\%$

2. 有一个信源 X 如下:

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.32 & 0.22 & 0.18 & 0.16 & 0.08 & 0.04 \end{bmatrix}$$

(1)、求信源熵 $H(X)$;

(2)、用 Shannon 编码法编成二进制变长码，并计算其编码效率;

(3)、用 Fano 编码法编成二进制变长码，并计算其编码效率;

- (4)、用 Huffman 码编码成二进制变长码，并计算其编码效率；
 (5)、用 Huffman 码编码成三进制变长码，并计算其编码效率；
 (6)、比较三种编码方法的优缺点。
 (7)、如果对信源采用定长二元编码，要求编码效率 η 达到三进制 Huffman 编码的效率，同时允许错误概率 $P_e \leq 10^{-3}$ ，求所需要的信源符号序列长度 N 。

$$H(X) = \sum_{i=1}^6 p(x_i) \log \frac{1}{p(x_i)} = 2.3522 \text{ bit / 信源符号}$$

$$\log_r \frac{1}{p(x_i)} \leq n_i < \log_r \frac{1}{p(x_i)} + 1 \quad \text{由于：每个码长肯定是正整数，上式给出了选择}$$

范围

符号	$p(x_i)$	累加概率	$-\log p(x_i)$	码字长度		码字
x_1	0.32	0	1.6439	2	0000	00
x_2	0.22	0.32	2.1844	3	0100	010
x_3	0.18	0.54	2.4739	3	1000	100
x_4	0.16	0.72	2.6439	3	1011	101
x_5	0.08	0.88	3.6439	4	11100	11100
x_6	0.04	0.96	4.6439	5	111101	11110

平均码字长度 $\bar{n} = 2.84$ 码符号/信源符号

$$\text{编码效率 } \eta = \frac{H(X)}{\bar{n} \log r} = \frac{2.3522}{2.84 \log 2} \times 100\% = 82.82\%$$

FINO 编码

Huffman 编码								
符号	$P(x_i)$					码字	码字长度	
x_1	0.32	0.54	0			00	2	
x_2	0.22		1			01	2	
x_3	0.18	0.46	0.18	0		10	2	
x_4	0.16		0.28	0.16	0		110	3
x_5	0.08			0.12	1	0	1110	4
x_6	0.04				1	1	1111	4

平均码字长度 $\bar{n} = 2.4$ 码符号/信源符号

$$\text{编码效率 } \eta = \frac{H(X)}{\bar{n} \log r} = \frac{2.3522}{2.4 \log 2} \times 100\% = 98\%$$

二进制 Huffman 编码

符号	$p(x_i)$					码字	码字长度
x_1	0.32	0.32	0.32	0.40	0.60	00	2
x_2	0.22	0.22	0.28	0.32	0.40	10	2
x_3	0.18	0.18	0.22	0.28		11	2
x_4	0.16	0.16	0.18			010	3
x_5	0.08	0.12				0110	4
x_6	0.04					0111	4

平均码字长度 $\bar{n} = 2.4$ 码符号/信源符号

$$\text{编码效率 } \eta = \frac{H(X)}{\bar{n} \log r} = \frac{2.3522}{2.4 \log 2} \times 100\% = 98\%$$

三进制 Huffman 编码

判断 $q - (r-1)\alpha = r$?

首先, $6 - (3-1) \times 2 = 2 < 3$

选择 $m = r - [q - (r-1)\alpha] = 3 - 2 = 1$ 个虚假符号

符号	$p(x_i)$			码字	码字长度		
x_1	0.32	0.32	0.48	1	1		
x_2	0.22	0.22	0.32	2	1		
x_3	0.18	0.18	0.22	00	2		
x_4	0.16	0.16		01	2		
x_5	0.08	0.12		020	3		
x_6	0.04			021	3		
x^*	0			022	舍去		

平均码字长度 $\bar{n} = 1.58$ 码符号/信源符号

$$\text{编码效率 } \eta = \frac{H(X)}{\bar{n} \log r} = \frac{2.3522}{1.58 \log 3} \times 100\% = 93.93\%$$

$$H(X) = \sum_{i=1}^6 p(x_i) \log \frac{1}{p(x_i)} = 2.3522 \text{ bit / 信源符号}$$

$$\text{对于信源所有符号的自信息的方差 } D[I(s_i)] = \sum_{i=1}^6 p(s_i) [-\log_2 p(s_i)]^2 - H_2(S) = 0.5265$$

$$\varepsilon = \frac{1-\eta}{\eta} H(S) = 0.152$$

$$N \geq \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{0.5256}{0.0231 \times 10^{-3}} = 2.275 \times 10^4$$

3. 现有一幅已离散量化后的图像，图像的灰度量化分成 8 级，如下表所示。表中数字为相应像素上的灰度级。

另有一无噪无损二元信道，单位时间(秒)内传输 100 个二元符号。

(1)、现将图像通过给定的信道传输，不考虑图像的任何统计特性，并采用二元等长码，问需要多长时间才能传送完这幅图像？

(2)、若考虑图像的统计特性(不考虑图像的像素之间的依赖性)，求这幅图像的信源熵 $H(S)$ ，并对每个灰度级进行 Huffman 最佳二元编码，问平均每个像素需用多少二元码符号来表示？这时需多少时间才能传送完这幅图像？

(3)、从理论上简要说明这幅图像还可以压缩，而且平均每个像素所需的二元码符号数可以小于 $H(S)$ 比特。

解：

(1)一幅已离散化后的图象，其灰度划分成 8 级，
先不考虑图象的任何统计特性，采用二元等长码，因为 $q=8$ ，所以要满足

$$2^l \geq q = 8$$

故 $l = 3$ 二元码符号 / 灰度级
即每个灰度等级需采用三位二元符号来传输。

这一幅图象空间离散化后共有 $N=100$ 个像素，每个像素的灰度需用三个二元符号来编码，所以这幅图象采用二元等长码后共需 300 个二元符号来描述。所传输的信道是无噪无损信道，其每秒钟传输 100 个二元符号。因此，需 3 秒钟才能传送完这幅图象。

(2)考虑图象的统计特性(不考虑图象的像素之间的依赖性)时，根据此图象进行统计，把像素的灰度值作为信源 S ，可得

$$\begin{bmatrix} S \\ P(s_i) \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \frac{40}{100} & \frac{17}{100} & \frac{10}{100} & \frac{10}{100} & \frac{7}{100} & \frac{6}{100} & \frac{5}{100} & \frac{5}{100} \end{bmatrix}$$

$$\text{所以 } H(S) = -\sum_{i=1}^8 p(s_i) \log p(s_i) \approx 2.572 \text{ 比特 / 灰度级}$$

对此灰度进行哈夫曼最佳二元编码：

码长 l_i	码字	灰度级	概率 $P(s_i)$					
1	1	1	0.40	0.40	0.40	0.40	0.40	0.60 0
3	001	2	0.17	0.17	0.20	0.23	0.37 0	0.40 1
4	0000	3	0.10	0.13	0.17	0.20 0	0.23 1	
4	0001	4	0.10	0.10	0.13 0	0.17 1		
4	0100	5	0.07	0.10 0	0.10 1			
4	0101	6	0.06	0.10 1				
4	0110	7	0.05					
4	0111	8	0.05					

得
$$\bar{L} = \sum_{i=1}^8 P(s_i) l_i = 2.63 \quad \text{二元符号 / 灰度级}$$

通过哈夫曼最佳二元编码后，每个像素平均需要用 2.63 个二元符号，则此图象平均共需要用 263 个二元符号来表示。因此，需 2.63 秒才能传送完这幅图象。

(3) 在 (2) 题中计算时没有考虑图象的像素之间的依赖关系，但实际此图象的像素之间是有依赖的。例如，若考虑像素前后之间灰度的依赖关系，就有灰度“1”后面只可能出现灰度“1”或“2”；灰度“2”后只可能出现“2”或“3”，等等。这时，此图象灰度值信源 S 可以看成一阶马尔可夫信源。还可以进一步看成为 m 阶马尔可夫信源。因此，在考虑了这些依赖关系后，像素的灰度值信源 S 的实际信息熵 $H_{\infty} < H(S)$ 。根据香农第一理，总可以找到一种编码，使每个灰度级的平均码长 $\bar{L} \rightarrow H_{\infty}$ (极限熵)。所以，这幅图象还可以进一步压缩，平均每个像素 (灰度) 所需要的二数码符号数 $\bar{L} < H(S)$ 。

4. 一个离散无记忆信源，它的样本空间为 $\{W, B\}$ ，符号 W 出现概率为 0.99，符号 B 出现的概率为 0.01。

- (1)、对此信源的二次扩展信源，求出信源符号序列的概率分布，找出与之相应的二元 Huffman 编码，并求出平均码长；
- (2)、对此信源的三次扩展信源，重复上一问；
- (3)、计算信源的单符号熵，并于以上两个结果进行比较；
- (4)、要想使得单符号平均码长 \bar{L}_n / n 只比单符号信源熵大 10%，请确定信源最小的扩展次数 n 。

解：

(1) 二次扩展信源的符号序列、概率分布及码字如表 5 11 所示。

表 5 11

符号序列	概 率	码 字
WW	$0.99 \times 0.99 = 0.9801$	0
WB	$0.99 \times 0.01 = 0.0099$	11
BW	$0.01 \times 0.99 = 0.0099$	100
BB	$0.01 \times 0.01 = 0.0001$	101

平均码长为

$$\overline{L}_2 = 0.9801 \times 1 + 0.0099 \times 2 + 0.0099 \times 3 + 0.0001 \times 3 = 1.0299$$

(2) 三次扩展信源的符号序列、概率分布及码字如表 5 12 所示。

(2) 三次扩展信源的符号序列、概率分布及码字如表 5 12 所示。

表 5 12

符号序列	概 率	码 字
WWW	$0.99 \times 0.99 \times 0.99 = 0.970299$	0
WWB	$0.99 \times 0.99 \times 0.01 = 0.009801$	100
WBW	$0.99 \times 0.01 \times 0.99 = 0.009801$	101
BWW	$0.01 \times 0.99 \times 0.99 = 0.009801$	110
WBB	$0.99 \times 0.01 \times 0.01 = 0.000099$	11100
BWB	$0.01 \times 0.99 \times 0.01 = 0.000099$	11101
BBW	$0.01 \times 0.01 \times 0.99 = 0.000099$	11110
BBB	$0.01 \times 0.01 \times 0.01 = 0.000001$	11111

平均码长为

$$\overline{L}_3 = 0.970299 \times 1 + 0.009801 \times (3 + 3 + 3) + 0.000099 \times (5 + 5 + 5) + 0.000001 \times 5 = 1.06$$

(3) 信源的单符号熵为

$$H(S) = 0.99 \log_2 \frac{1}{0.99} + 0.01 \log_2 \frac{1}{0.01} = 0.080179 \text{ 比特/符号}$$

二次扩展信源和三次扩展信源的单符号平均码长分别为

$$\frac{\overline{L}_2}{2} = 0.51495, \quad \frac{\overline{L}_3}{3} = 0.35333$$

都远大于信源的单符号熵。

(4) 单符号信源熵的 110% 是 $1.1 \times 0.080179 = 0.08819$, 而 \overline{L}_n 至少为 1, 即单符号平均码长至少为 $1/n$, 因此, 根据题意有

$$\frac{1}{n} \leq 0.08819 \quad n \geq 11.339 \quad n = 12$$

5.假设有一页传真文件，其中有三条扫描线上的像素点如下图所示，分别为第一行、第二行以及最后一行。

- 1) 请编制这些扫描行的 MH (Modified Huffman) 编码；
- 2) 计算第三行的数据压缩比。

686 黑	455 白	355 黑	155 白	13 黑	5 白	14 黑	45 白
85 白	720 黑	108 白	3 黑	64 白	6 黑	30 白	712 黑
.....							
832 白	728 黑	68 白	64 黑	36 白			

原则：

- 1) 0-63 像素的游程，用对应的结尾码表示；
- 2) 64-1728 像素的游程，用一个构造码+结尾码表示；
- 3) 为了保证收发双方颜色同步，数据行的起点都是白游程开始(如果其中开始游程为黑游程，那么先增加一个白游程长度为 0 编码字为 00110101)；
- 4) 行终码 EOL (000000000001 即 11 个 0 后加一个 1) 出现在每一行扫描线之最后；行终码也出现在每一页的第一行数据之前（如果出现阵发性差错后可以重新建立同步）。
- 5) 页面结尾码：一页面结束后，发送控制规程码 (RTC，连续发送 6 个 EOL 码)；

参考 MH 编码表：

第一行：

第一行	编码	长度
插入 EOL；	<u>000000000001</u>	12
加入长度为零的白色游程	<u>00110101</u>	8
686 黑游程：640+46	0000001001010+ <u>000001010110</u>	13+12=25
455 白游程：448+7	01100101+ <u>1111</u>	8+4=12
355 黑游程：320+35	000000110011+ <u>000011010011</u>	12+12=24
155 白游程：128+27	10010+ <u>0100100</u>	5+7=12
13 黑游程：	00000100	8
5 白游程：	1100	4
14 黑游程：	00000111	8
45 白游程：	00000100	8
EOL	<u>000000000001</u>	12
合计		

第 2 行	编码	长度
85 白游程：64+21	11011+ <u>0010111</u>	8+4=12
720 黑游程：704+16	0000001001011+ <u>0000010111</u>	10+13=23
108 白游程：64+44	11011+ <u>00101101</u>	5+8=13
3 黑游程：	10	2
64 白游程：加入长度为零	11011+ <u>00110101</u>	5+8=13

6 黑游程:	0010	4
30 白游程:	00000011	8
712 黑游程: 704+8	0000001001011+ <u>000101</u>	13+6=19
EOL	<u>0000000000001</u>	12
合计		

最后一行	编码	长度
832 白游程: 加入长度为零	011010010+ <u>00110101</u>	9+8=17
728 黑游程: 704+24	0000001001011+ <u>00000010111</u>	10+11=21
68 白游程: 64+4	11011+ <u>1011</u>	5+4=9
64 黑游程: 加入长度为零	0000001111+ <u>0000110111</u>	10+10=20
36 白游程:	00010101	8
1 个控制规程码 RTC	6 个 EOL	6*12=72
合计		147

计算第三行的数据压缩比:
 $1728/147=11.76$
