

《信息论与编码》第三章习题习题解答

本解答由两部分组成

PartA : (1 - 11 页) 包括:1,2,3,4,6,7,8,13,14,16

PartB: (12 - 16 页) 包括 : 5,9,10,12,15

《信息论与编码》第三章习题解答(PartA)

3.1 试证明长度为 N 的 D 元不等长码至多有 $D(D^N-1)/(D-1)$ 个码字。

[证明] 因为长度为 i 码字最多有 D^i 个, 所以长度不超过 N 的 D 元不等长码的码字数最多有

$$\begin{aligned}\sum_{i=1}^N D^i &= D + D^2 + \cdots + D^N \\ &= \frac{D(1-D^N)}{1-D}\end{aligned}$$

3.2 设一个 DMS, $U = \begin{pmatrix} a_1 & a_2 \\ p(a_1)=0.25 & p(a_2)=0.75 \end{pmatrix}$, 其熵为 $H(U)$ 。考察长度为 L 的

输出序列, 当 $L \geq L_0$ 时满足下式

$$P_r \left\{ \left| \frac{I(u^L)}{L} - H(U) \right| > \delta \right\} \leq \varepsilon$$

(a) 在 $\delta = 0.05, \varepsilon = 0.1$ 时求 L_0 值;

(b) 在 $\delta = 10^{-3}, \varepsilon = 10^{-6}$ 时求 L_0 值;

(c) $A = \left\{ u^L : \left| \frac{I(u^L)}{L} - H(U) \right| > \delta \right\}$

求在 (a), (b) 给定的 $L=L_0$ 情况下 A 中元素数目的上、下限。

[解] 由概率论中切比雪夫不等式

$$P\left\{ \left| \frac{I(U^L)}{L} - H(U) \right| > \delta \right\} \leq \frac{\sigma_I^2}{L\delta^2} = \varepsilon$$

其中 $H(U) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0.81 \text{ bit}$

$$\begin{aligned}\sigma_I^2 &= \sum_{i=1}^2 P(a_i) [I(a_i) - H(U)]^2 \\ &= \sum_{i=1}^2 P(a_i) \left[\log \frac{1}{p(a_i)} - H(U) \right]^2 \\ &= 0.471\end{aligned}$$

(a) 当 $\sigma = 0.05, \varepsilon = 0.1$ 时, 解出

$$L_0 = 1884$$

(b) 当 $\delta = 10^{-3}$, $\varepsilon = 10^{-4}$ 时, 解出

$$L_0 = 471 \times 10^7$$

(c) 记 $A_\delta^{(L)} = \left\{ u^L : \left| \frac{I(u^L)}{L} - H(U) \right| \leq \delta \right\}$

则对一切 $L \geq L_0$ 以及 $u^L \in A_\delta^{(L)}$ 具有

$$2^{-L[H(U)+\delta]} \leq P(u^L) \leq 2^{-L[H(U)-\delta]}$$

所以 $1 = \sum_{u^L \in u^L} p(u^L) \geq \sum_{u^L \in A_\delta^{(L)}} p(u^L) \geq 2^{-L[H(U)+\delta]} \cdot |A_\delta^{(L)}|$

因而 $|A_\delta^{(L)}| \leq 2^{L[H(U)-\delta]}$

又 $1 - \varepsilon \leq \sum_{u^L \in A_\delta^{(L)}} p(u^L) \leq \sum_{u^L \in A_\delta^{(L)}} 2^{-L[H(U)-\delta]} = 2^{-L[H(U)-\delta]} \cdot |A_\delta^{(L)}|$

即 $|A_\delta^{(L)}| \geq (1 - \varepsilon) \cdot 2^{L[H(U)-\delta]}$

对于 $\delta = 0.05$, $\varepsilon = 0.1$ 情况 $|A_\delta^{(L_0)}|$ 的上下界为

$$|A_\delta^{(L_0)}| \leq 2^{1884 \cdot [0.81+0.05]} = 2^{1884 \cdot 0.86} = 2^{1621}$$

$$|A_\delta^{(L_0)}| \geq (1 - \varepsilon) \cdot 2^{1884 \cdot [0.81-0.05]} = 0.9 \cdot 2^{1431}$$

对于 $\delta = 10^{-3}$, $\varepsilon = 10^{-4}$ 情况, $|A_\delta^{(L_0)}|$ 的上下界为

$$|A_\delta^{(L_0)}| \leq 2^{1471 \times 10^7 \cdot [0.81+0.05]} = 2^{406 \cdot 10^7}$$

$$|A_\delta^{(L_0)}| \geq 2^{471 \times 10^7 \cdot [0.81-0.05]} = 0.9 \cdot 2^{357 \cdot 10^7}$$

3.3 下面哪些码是唯一可译的

- (1) $\{0,10,11\}$, (2) $\{0,01,11\}$, (3) $\{0,01,10\}$, (4) $\{0,01\}$,
(5) $\{00,01,10,11\}$, (6) $\{110,11,10\}$, (7) $\{110,100,00,10\}$

【解】

(1) $\{0,10,11\}$ 是前缀码, 故唯一可译;

(2) $\{0,01,11\}$ 是唯一可译, 因为它的后缀分解集不含码字;

$$S_0 \quad S_1 \quad S_2$$

$$0 \quad 1 \quad 1$$

$$01$$

$$11$$

(3) $\{0,01,10\}$ 不是唯一可译；因为它的后缀分解集 S_2 含有码字“0”，于是例如“010”

有二种译码方法即“0, 10”和“01, 0”

(4) $\{0,01\}$ 是唯一可译，因为它的后缀分解集不含码字；

S_0	S_1	S_2
0	1	ϕ
01		

(5) $\{00,01,10,11\}$ 是唯一可译，它是前缀码；

(6) $\{110,11,10\}$ 是唯一可译，它的后缀分解集不含码字；

S_0	S_1	S_2
110	0	ϕ
11		
10		

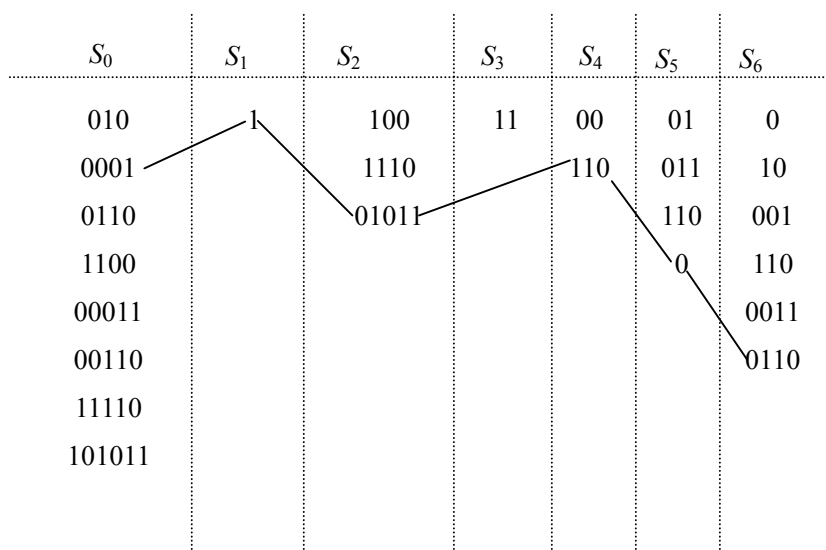
(7) $\{110,100,00,10\}$ 是唯一可译，它的后缀分解集不含码字；

S_0	S_1	S_2
110	0	0
100		
00		
10		

3.4 确定下面码是否唯一可译，若不是请构造一个模糊序列

a.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	010	0001	0110	1100	00011	00110	11110	101011
b.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	abc	abcd	e	dba	bace	ceac	ceab	eabd

[解] a. {010, 0001, 0110, 1100, 00011, 00110, 11110, 101011}的后缀分解为



S_6 中包含有码字“0110”，所以不是唯一可译。

例如序列

“000110101111000110”有二种译法，即

“0001, 101011, 1100, 0110”和“00011, 010, 11110, 00110”

(b) 编码{abc, abcd, e, dba, bace, ceac, ceab, eabd}的后缀分解为

S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7
abc	d	ba	ce	ac	c	eac	ac
abcd	abd			ab	cd	eab	ab
e							d
dba							
bace							
ceac							
cedb							
eabd							

因为 S_7 中元素在 S_1 和 S_4 中都出现过，所以 S_7 以后的后缀分解集中不会出现 $S_1 - S_7$ 中没有出现过的元素，所以从 $S_1 - S_7$ 可见后缀分解集中不含有码字，所以编码是唯一可译的。

3.6 令 DMS 为

$$U = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} \\ 0.16 & 0.14 & 0.13 & 0.12 & 0.1 & 0.09 & 0.08 & 0.07 & 0.06 & 0.05 \end{pmatrix}$$

(a) 求二元 Huffman 码, 计算 \bar{n} 和 η ;

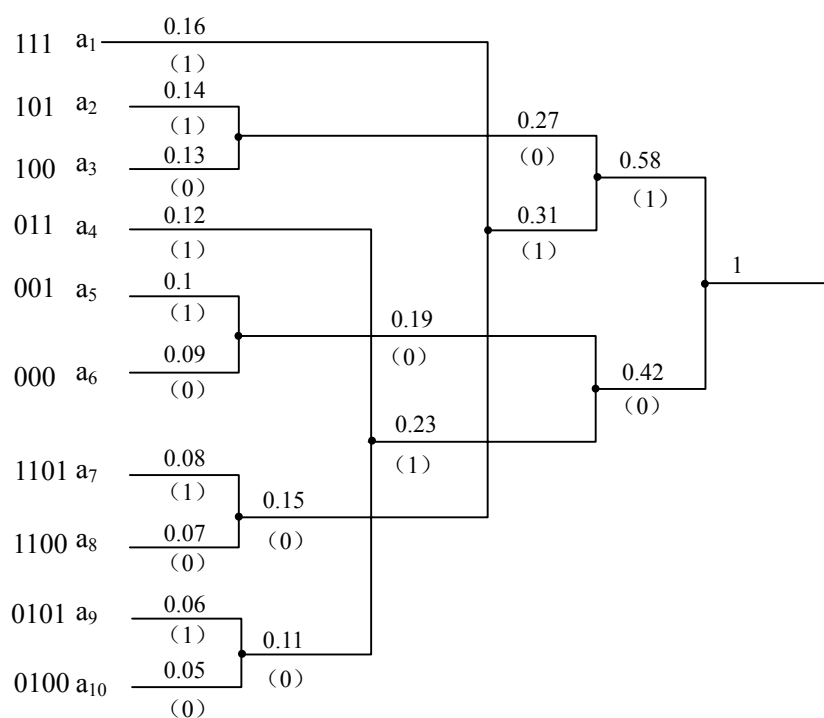
(b) 求三元 Huffman 码, 计算 \bar{n} 和 η ;

[解]

(a) 由信源概率分布可知

$$H(U) = -\sum_{i=1}^{10} p_i \log p_i = 3.234 \text{ bit}$$

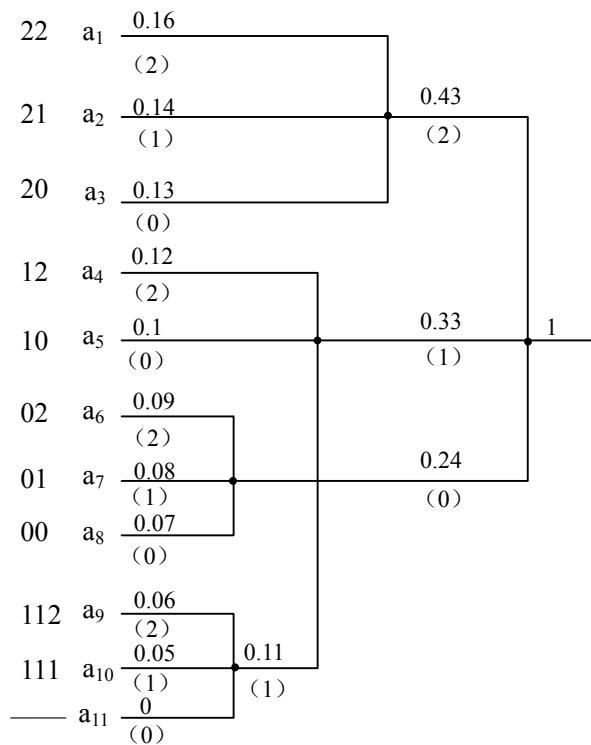
相应的 Huffman 编码过程如下图所示:



$$\bar{n} = \sum_{i=1}^{10} n_i \cdot p_i = 3.26$$

$$q = \frac{H(U)}{\bar{n}} = 99.2\%$$

(b) 三元 Huffman 码如下构成



$$\bar{n} = 2.11$$

$$n = \frac{3.234}{2.11 \cdot \log D} = 96.7\%$$

3.7 下面三个码中，哪些对任何概率分布都不可能成为是 Huffman 码？

(a) {0,10,11}

(b) {00,01,10,110}

(c) {01,10}

[解] (a) {0, 10, 11} 可能为 Huffman 码，因为它构成满树；

(b) {00, 01, 10, 110} 不可能为 Huffman 码，

因为码字“110”可以用更短的“11”代替，而保持前缀码条件；

(c) {01, 10} 不可能成为 Huffman 码，因为显然{0, 1}是平均码长更短的前缀码；

3.8 一个随机变量 X 的取值范围为 $\mathcal{X} = \{x_1, x_2, \dots, x_m\}$ ，它的熵为 $H(X)$ ，若对这个源能找到

一个平均码长为 $L = \frac{H(X)}{\log_2 3} = H_3(X)$ 的三元即时码，试证

(1) 对每个 $x_i \in \mathcal{X}$, $p(X = x_i) = 3^{-l_i}$ ，其中 l_i 为某个整数；

(2) 证明 m 为奇数；

[证明] (a) 设 \mathcal{D} 是平均码长 $L = \frac{H(X)}{\log 3}$ 的三元即时码，我们可以把它映射到一棵三岔树的

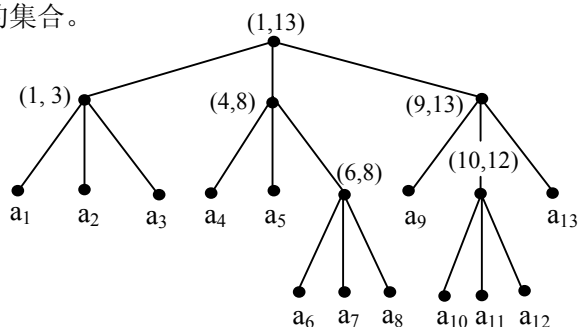
树上，不失一般性我们假定树为满树。因为若不然我们可以补上概率为零的消息，使之成为满树。现设消息总数为 m ，消息集合为 $M = \{x_1, x_2, \dots, x_m\}$ ，相应概率为 $p_i = P\{X = x_i\}$ 。编码树类似下图所示。不是树叶的节点称为内节点，内节点 (i, j) 表示从该节点长出的树叶为 $\{x_i, x_{i+1}, \dots, x_j\}$ ，如果 $i = j$ ，则 (i, j) 表示该节点为树叶。

\mathcal{J} 表示全体内节点 (i, j) ， $(i < j)$ 的集合。

内节点 (i, j) 上的累计概率定义为：

$$P(i, j) = \sum_{k=i}^j p_k$$

$$\begin{aligned} \text{于是} \quad L &= \sum_{i=1}^n l_i \cdot p_i \\ &= \sum_{(i,j) \in \mathcal{J}} p(i, j) \end{aligned}$$



又由熵的可加性

$$\begin{aligned} H(X) &= H(p_1, p_2, \dots, p_n) \\ &= \sum_{(i,j) \in \mathcal{J}} p(i, j) H\left(\frac{p(i, k_1)}{p(i, j)}, \frac{p(k_1 + 1, k_2)}{p(i, j)}, \frac{p(k_2 + 1, j)}{p(i, j)}\right) \end{aligned}$$

其中 (i, k_1) ， $(k_1 + 1, k_2)$ ， $(k_2 + 1, j)$ 是由内节点 (i, j) 分岔出去三个节点，所以 $p(i, k_1) + p(k_1 + 1, k_2) + p(k_2 + 1, j) = p(i, j)$ 。由于码 \mathcal{D} 的平均码长 $L = \frac{H(X)}{\log 3}$ ，所以

$$L \cdot \log 3 - H(X) = \sum_{(i,j) \in \mathcal{J}} p(i, j) \left[\log 3 - H\left(\frac{p(i, k_1)}{p(i, j)}, \frac{p(k_1 + 1, k_2)}{p(i, j)}, \frac{p(k_2 + 1, j)}{p(i, j)}\right) \right] = 0$$

因为 $p(i, j) > 0$ ， $\forall (i, j) \in \mathcal{J}$

所以要求对任何 $(i, j) \in \mathcal{J}$

$$H\left(\frac{p(i, k_1)}{p(i, j)}, \frac{p(k_1 + 1, k_2)}{p(i, j)}, \frac{p(k_2 + 1, j)}{p(i, j)}\right) = \log 3$$

也就是说要求

$$p(i, k_1) = p(k_1 + 1, k_2) = p(k_2 + 1, j) = \frac{1}{3} p(i, j)$$

所以从编码树每个内节点长出的三个分支都具有等概率，即第一层节点概率为 $\frac{1}{3}$ ，

第二层节点概率为 $\frac{1}{9}$ ， \dots 。从而任何一个消息（树叶）出现概率必定为 $\frac{1}{3}$ 的整数次幂。

(b) 由 (a), 我们知道补零概率消息是不必要的, 也就是说满足 $L = \frac{H(X)}{\log 3}$ 的即时

码对应一个满树。这时消息 m 满足

$$\begin{aligned} m &= (D-1) \cdot i + 1 \\ &= 2 \cdot i + 1 \end{aligned}$$

即是一个奇数。

3.13 设一个 DMS

$$U = \begin{pmatrix} a_1 & a_2 \\ 0.004 & 0.996 \end{pmatrix}$$

若对其输出长为 100 的序列中含有两个或更少个 a_1 的序列提供不同码字,

(a) 在等长编码中, 求二元码的最短长度;

(b) 求错误概率 (误码字率)。

【解】 (a) 在输出长度为 $L=100$ 的序列中, 含有两个或更少个 a_1 的序列数共有

$$S = C_{100}^0 + C_{100}^1 + C_{100}^2 = 5051 \text{ 个}$$

所以若用二元码来表示这 5051 个序列, 最短码长为

$$N = \lceil \log_2 5051 \rceil = 13$$

(b) 当出现含有 3 个 a_1 或更多 a_1 的长度为 100 的序列时, 则出现译码错误, 所以误码率为:

$$\begin{aligned} P_E &= 1 - C_{100}^2 \cdot (0.004)^2 \cdot (0.996)^{98} - C_{100}^1 \cdot (0.004) \cdot (0.996)^{99} - (0.996)^{100} \\ &\approx 0.01 \end{aligned}$$

3.14 假定 DMS 为

$$X = \begin{pmatrix} a_1, & a_2 & \cdots, & a_m \\ p_1, & p_2, & \cdots, & p_m \end{pmatrix}$$

令 l_i 表示对应于消息 a_i 的二元码字的长度, C_i 表示消息 a_i 重要性的加权, 于是这个

码的平均代价为 $C = \sum_{i=1}^m p_i l_i \cdot C_i$;

(a) 在 $\sum_{i=1}^m 2^{-l_i} \leq 1$ 约束下最小化 C , 求出最小化 C 的值 C^* 和相应的 l_i^* , $i=1, 2, \dots, m$, (这里忽略对于 l_i 是整数的限制)

(b) 如何利用 Huffman 编码方法对所有唯一可译码来最小化 C , 这个最小化 C 记为

C_{Huffman} ;

(c) 证明

$$C^* \leq C_{\text{Huffman}} \leq C^* + \sum_{i=1}^m p_i c_i .$$

【证明】(a) 首先证明在约束 $\sum_{i=1}^m 2^{-l_i} \leq 1$ 条件下最小化 $C = \sum_{i=1}^m p_i l_i c_i$ 相当于在等式约束

$\sum_{i=1}^m 2^{-l_i} = 1$ 条件下最小化 C 。

若不然, 设在 $\sum_{i=1}^m 2^{-l'_i} = B < 1$, 使 $C' = \sum_{i=1}^m p_i \cdot l'_i \cdot c_i$ 最小; 则总可找到 $\alpha = \log \frac{1}{B} > 0$,

使得 $\sum_{i=1}^m 2^{-l_i} = \sum_{i=1}^m 2^{-l'_i + \alpha} = 1$ 但 $\sum_{i=1}^m p_i \cdot l_i \cdot c_i = \sum_{i=1}^m p_i \cdot l'_i \cdot c_i - \alpha \sum_{i=1}^m p_i c_i < C'$; 所以最小化必

定在等式约束下达到。

下面引入拉氏乘子 λ , 求下面无条件目标函数极小

$$J = \sum_{i=1}^m p_i \cdot c_i \cdot l_i + \lambda \sum_{i=1}^m 2^{-l_i}$$

令 $\frac{\partial J}{\partial l_i} = 0$ 得到

$$p_i c_i - \lambda 2^{-l_i} \cdot \ln 2 = 0$$

所以
$$2^{-l_i} = \frac{p_i c_i}{\lambda \ln 2}$$

由于
$$\sum_{i=1}^m 2^{-l_i} = 1$$

得到
$$\lambda = \frac{1}{\ln 2} \sum_{i=1}^m p_i c_i$$

所以当

$$l_i = l_i^* = -\log_2 \frac{p_i c_i}{\lambda \ln 2} = \log_2 \frac{\sum_{i=1}^m p_i c_i}{p_i c_i}$$

$$C^* = \sum_{i=1}^m p_i \cdot c_i \cdot \left[\log_2 \frac{\sum_{k=1}^m p_k c_k}{p_i c_i} \right]$$

(b) 记 $q_i = \frac{p_i c_i}{\sum_{k=1}^m c_k \cdot p_k}$ ，则 $\{q_i\}$ 构成一个概率分布，根据 $\{q_i\}$ 来构成 Huffman 码。

设这时的码长为 $\{l_i\}$ ，于是这时最小平均码长 \bar{n} 为 $\bar{n} = \sum_{i=1}^m \frac{p_i c_i l_i}{\sum_{k=1}^m c_k p_k}$ ，相应的

$$\begin{aligned} C_{Huffman} &= \sum_{i=1}^m p_i c_i l_i \\ &= \bar{n} \cdot \sum_{k=1}^m c_k p_k \end{aligned}$$

(c) 对于 Huffman 码，平均码长满足

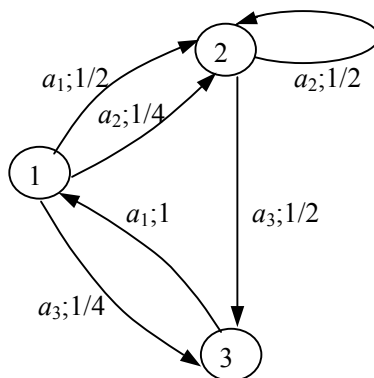
$$H(\{q_i\}) \leq \bar{n} < H(\{q_i\}) + 1$$

其中
$$H(\{q_i\}) = \sum_{i=1}^m \frac{p_i c_i}{\sum_{k=1}^m c_k p_k} \left[\log \frac{\sum_{k=1}^m p_k c_k}{p_i c_i} \right]$$

所以
$$C^* \leq C_{Huffman} = \bar{n} \cdot \sum_{k=1}^m c_k p_k < C^* + \sum_{k=1}^m c_k p_k$$

3.16 设一个马尔可夫信源，其状态图如图所示。

- 求稳态下各状态概率 $q(i)$ ，以及各字母 a_i 的出现概率， $i=1, 2, 3$ ；
- 求 $H(U|s_i)$ ， $i=1, 2, 3$ ；
- 求 $H_\infty(U)$ ；
- 对各状态 s_j 求最佳二源码；
- 计算平均码长。



习题 3.16 图

[解]

(a) 稳态状态概率 $q(1)$, $q(2)$, $q(3)$ 满足

$$\begin{cases} q(1) = q(3) \\ q(2) = \frac{3}{4}q(1) + \frac{1}{2}q(2) \\ q(3) = \frac{1}{4}q(1) + \frac{1}{2}q(2) \\ q(1) + q(2) + q(3) = 1 \end{cases}$$

得到 $q(1) = q(3) = \frac{2}{7}$

$$q(2) = \frac{3}{7}$$

$$p(a_1) = q(1) \cdot p(a_1 | s_1) + q(3) \cdot p(a_1 | s_3) = \frac{6}{14}$$

$$p(a_1) = q(1) \cdot p(a_2 | s_1) + q(2) \cdot p(a_2 | s_2) = \frac{4}{14}$$

$$p(a_3) = q(1) \cdot p(a_3 | s_1) + q(2) \cdot p(a_3 | s_2) = \frac{4}{14}$$

(b) $H(U | S_1) = 1.5 \text{ bit}$

$$H(U | S_2) = 1 \text{ bit}$$

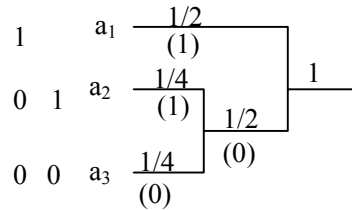
$$H(U | S_3) = 0 \text{ bit}$$

(c) $H_\infty(U) = q(1) \cdot H(U | S_1) + q(2)H(U | S_2) + q(3)H(U | S_3)$

$$= \frac{2}{7} \cdot 1.5 + \frac{3}{7} \cdot 1 = \frac{6}{7} \text{ bit}$$

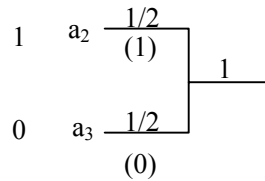
(d) 在 s_1 状态 Huffman 编码为:

$$\bar{n}_1 = 1 \frac{1}{2}$$



在 s_2 状态 Huffman 编码为:

$$\bar{n}_2 = 1$$



在 s_3 状态无需编码, $\bar{n}_3 = 0$

(e) $\bar{n} = \bar{n}_1 \cdot q(1) + \bar{n}_2 \cdot q(2) = \frac{3}{7} + \frac{3}{7} = \frac{6}{7}$

3.5 对于下列码，令 $N(k)$ 表示长度为 k 的码字序列，如下列码 (a) 中 $N(1) = 1$ ， $N(2) = 3$

(x_1x_1 , x_2 , x_3), $N(3) = 5$ ($x_1x_1x_1$, x_1x_2 , x_1x_3 , x_2x_1 , x_3x_1) 等，求 $N(k)$, $k = 1$, 2 , ...

(a) x_1 x_2 x_3
 0 10 11

(b) x_1 x_2 x_3 x_4
 0 10 110 111

[解] (a) $N(k) = N(k-1) + N(k-2) \times 2$

特征根为 2 , -1

令 a, b 为待定系数

$$N(k) = a(-1)^k + b(2)^k$$

代入初始条件 $N(1) = 1, N(2) = 3$ 得到

$$\begin{cases} a = \frac{1}{3} \\ b = \frac{2}{3} \end{cases}$$

$$\text{于是: } N(k) = \frac{1}{3}(-1)^k + \frac{2}{3}(2)^k$$

$$(b) \quad N(k) = N(k-1) + N(k-2) + N(k-3) \times 2$$

$$\text{特征根为: } 2, \frac{-1 \pm \sqrt{3}i}{2}$$

$$\text{同样: } N(k) = a\left(\frac{-1 + \sqrt{3}i}{2}\right)^k + b\left(\frac{-1 - \sqrt{3}i}{2}\right)^k + c(2)^k$$

代入初始条件: $N(1) = 1, N(2) = 2, N(3) = 5$ 得:

$$\begin{cases} a = \frac{\sqrt{3}}{21}i \\ b = -\frac{\sqrt{3}}{21}i \\ c = \frac{4}{7} \end{cases}$$

$$\text{所以: } N(k) = \frac{\sqrt{3}}{21} i \left(\frac{-1 + \sqrt{3}i}{2} \right)^k + \left(-\frac{\sqrt{3}}{21} i \right) \left(\frac{-1 - \sqrt{3}i}{2} \right)^k + \frac{4}{7} (2)^k$$

3.9 对于一个码, 如果没有一个码字是其他码字的后缀, 则称该码为后缀码, 试证后缀码是唯一可译的; 对于任何随机变量, 最短后缀码的平均码长等于相应 Huffman 码的平均码长。

[证明] (1) 如果一个码为后缀码, 即没有一个码字是任何一个其他码字的后缀, 则必定这些码的后缀分解集中不包含有这些码字, 这样就满足唯一可译码的充要条件, 即后缀码必定唯一可译。

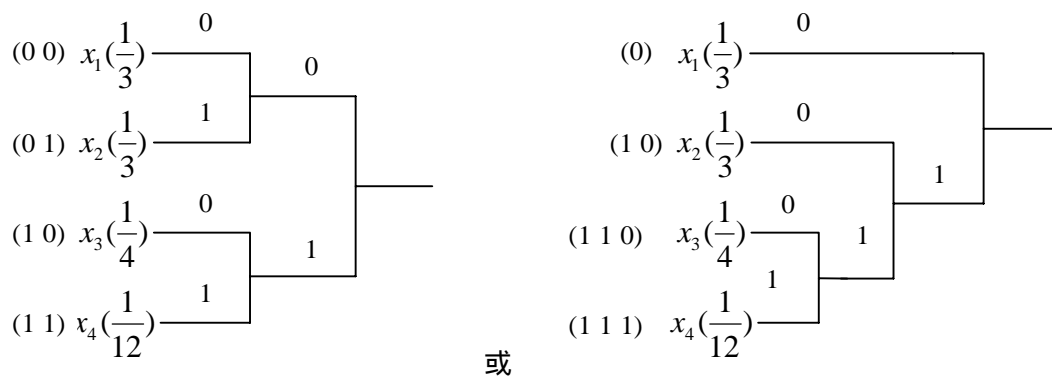
(2) 又 Huffman 码可以使唯一可译码的平均码长最短, 则将 Huffman 码应用于后缀码的源序列, 必可以得到平均码长最短的后缀码。这也说明最短后缀码的平均码长等于相应的 Huffman 码的平均码长。

3.10 考虑一个概率分布为 $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$ 的随机变量

- (a) 对该随机变量构成一个 Huffman 码
 (b) 证明存在 2 个不同的最佳码长集合, 即证明码长集合为 (1, 2, 3) 和 (2, 2, 2, 2) 都是最佳的

- (c) 证明某些符号的最佳码长长度可能超出 shannon 编码码长 $\left\lceil \log \frac{1}{p(x)} \right\rceil$

[解] (a) 构成的 2 种 Huffman 编码方式如下:



(b) 上面所示的 2 个码字集合都满足最佳不等长编码的条件, 即最小概率的 2 个消息对应的码字等长, 且长度最长, 同时这两个码字仅仅最后 1 位码元取值不同。

上面两种码字集合的平均码长分别为:

$$\bar{n}_1 = 2 \times \left(\frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{12} \right) = 2$$

$$\bar{n}_2 = \frac{1}{3} + \frac{2}{3} + 3 \times \left(\frac{1}{4} + \frac{1}{12} \right) = 2$$

即平均码长相等, 即对于此信源, 存在 2 个不同的最佳码字集合。

(c) 对于给定信源计算 shannon 编码得到各符号码字长度为：

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ 2 & 2 & 2 & 4 \end{bmatrix}$$

对于符号 x_4 我们比较上述得到的 shannon 编码长度和 (1, 2, 3, 3) Hffman 编码长度

可以看到： $L_{shannon}(x_4) > L_{Huffman}(x_4)$

因此某些符号的最佳码长可能超出 shannon 编码的码长。

3.12 令 DMS 为

$$U = \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_K \\ p(a_1) \geq p(a_2) \geq p(a_3) \geq \dots \geq p(a_K) \end{bmatrix}$$

$$\text{定义} \begin{cases} Q_i = \sum_{k=1}^{i-1} p(a_k) & i > 1 \\ Q_1 = 0 \end{cases}$$

按如下的方法进行编码，消息 a_k 用 Q_i 的二进制小数展开的截短来表示，截短的长度为

保留小数点后 l_k 位， $l_k = \left\lceil \log \frac{1}{p(a_k)} \right\rceil$ ， $\lceil x \rceil$ 表示大于，等于 x 的最小整数。

(a) 对 DMS

$$U = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \end{bmatrix}$$

进行上述编码。

(b) 证明上述编码满足异字头条件，且平均码长 \bar{n} 满足 $H(U) \leq \bar{n} \leq H(U) + 1$

[解] (a) 计算 Q 得：

$$Q = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{5}{8} & \frac{3}{4} & \frac{13}{16} & \frac{7}{8} & \frac{15}{16} \end{bmatrix}$$

于是通过保留小数点后位数得到编码长度为：

$$L = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 2 & 2 & 3 & 3 & 4 & 4 & 4 & 4 \end{bmatrix}$$

则得到编码为：

$$C = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 00 & 01 & 100 & 101 & 1100 & 1101 & 1110 & 1111 \end{bmatrix}$$

(b) 首先用反证法证明上述编码是异字头码：

[证明] 若上述编码不是异字头码，则总存在 a_i 是 a_j 的码字前缀

即 $i < j, n_i < n_j$ ，且 a_i 是 a_j 的码字前缀，

则 $a_j = a_i d_1 d_2 d_3 \dots d_n$

这样必有 $p_j - p_i = 00000\dots d_{n_i+1} \dots d_{n_i+j} < 2^{-n_i} < p_i$

当同时有： $p_j - p_i = \sum_{l=0}^{j-1} p_l - \sum_{l=1}^{i-1} p_l = p_i + p_{i+1} + \dots + p_{j-1} > p_i$

因此前后矛盾，假设不成立，因此上述编码是异字头码。

证明平均码长 \bar{n} 满足 $H(U) \leq \bar{n} \leq H(U) + 1$

[证明] $\bar{n} = 2 \times \frac{1}{2} + 3 \times \frac{1}{4} + 4 \times \frac{1}{4} = 2.75$

又计算 $H(U) = 2.75$

则平均码长 \bar{n} 满足 $H(U) \leq \bar{n} \leq H(U) + 1$

3.15 已知二元信源 $\{0,1\}$ ， $p_0 = \frac{1}{8}, p_1 = \frac{7}{8}$ ，试用算术码对以下序列进行编码

11111110111110

[解] 编码过程为：

输入	符号概率	保留长度	累加概率	编码码字
初始	$p(\lambda) = 1$	0	$F(\lambda) = 0$	-
1	0.111	1	(bin)0.001	0.1
1	$(\frac{7}{8})^2$	1	(bin)0.001111	0.1
1	$(\frac{7}{8})^3$	1	(bin)0.010101001	0.1
1	$(\frac{7}{8})^4$	1	(bin)0.011010011111	0.1
1	$(\frac{7}{8})^5$	1	(bin)0.011111001011001	0.1
1	$(\frac{7}{8})^6$	2	(bin)0.100011010001101111	0.11

1	$(\frac{7}{8})^7$	2	0.607304	0.11
0	$(\frac{7}{8})^7(\frac{1}{8})$	5	0.607304	0.10100
1	$(\frac{7}{8})^8(\frac{1}{8})$	5	0.613440	0.10100
1	$(\frac{7}{8})^9(\frac{1}{8})$	5	0.618801	0.10100
1	$(\frac{7}{8})^{10}(\frac{1}{8})$	5	0.623499	0.10100
1	$(\frac{7}{8})^{11}(\frac{1}{8})$	6	0.627610	0.101001
1	$(\frac{7}{8})^{12}(\frac{1}{8})$	6	0.631207	0.101001
0	$(\frac{7}{8})^{12}(\frac{1}{8})^2$	9	0.631207	0.101000100

于是总终的编码结果为：101000100