《信息论与编码》第三章习题解答

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

$$\sum_{i=1}^{N} D^i = D + D^2 + \cdots D^N$$

$$= \frac{D\left(1 - D^N\right)}{1 - D}$$

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 $L \geqslant L_0$ 时满足下式

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

(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

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

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

$$\sigma_I^2 = \sum_{i=1}^{2} P(a_i)[I(a_i) - H(U)]^2$$

$$= \sum_{i=1}^{2} P(a_i)\left[\log\frac{1}{p(a_i)} - H(U)\right]^2$$

=0.471

（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

$$L_0=471\times10^7$$

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

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

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

所以 $$1=\sum_{u^L\in u^L}p(u^L)\ge\sum_{u^L\in A_\delta^{(L)}}p(u^L)$$

$$\ge 2^{-L[H(U)+\delta]}\cdot|A_\delta^{(L)}|$$

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

又 $$1-\varepsilon\le\sum_{u^L\in A_\delta^{(L)}}p\left(u^L\right)\le\sum_{u^L\in A_\delta^{(L)}}2^{-L[H(U)-\delta]}$$

$$=2^{-L[H(U)-\delta]}\cdot\left|A_\delta^{(L)}\right|$$

即 $$\left|A_\delta^{(L)}\right|\ge\left(1-\varepsilon\right)\cdot2^{L[H(U)-\delta]}$$

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

$$\left|A_\delta^{(L_0)}\right|\le2^{1884\cdot[0.81+0.05]}=2^{1884\cdot0.86}=2^{1621}$$

$$\left|A_\delta^{(L_0)}\right|\ge(1-\varepsilon)\cdot2^{1884[0.81-0.05]}=0.9\cdot2^{1431}$$

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

$$\left|A_\delta^{(L_0)}\right|\le2^{1471\times10^7\cdot[0.81+0.05]}=2^{406\cdot10^7}$$

$$\left|A_\delta^{(L_0)}\right|\ge2^{471\times10^7\cdot[0.81-0.05]}=0.9\cdot2^{357\cdot10^7}$$

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	S_1	S_2
0	1	1
01		
11		

（3）$\{0,01,10\}$ 不是唯可译；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）$\{110,100,00,10\}$ 是唯可一译码，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_0	S_1	S_2	S_3	S_4	S_5	S_6
010	1	100	11	00	01	0
0001		1110		110	011	10
0110		01011			110	001
1100					0	110
00011						0011
00110						0110
11110						
101011						

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 $\overline{n}$ 和 η；

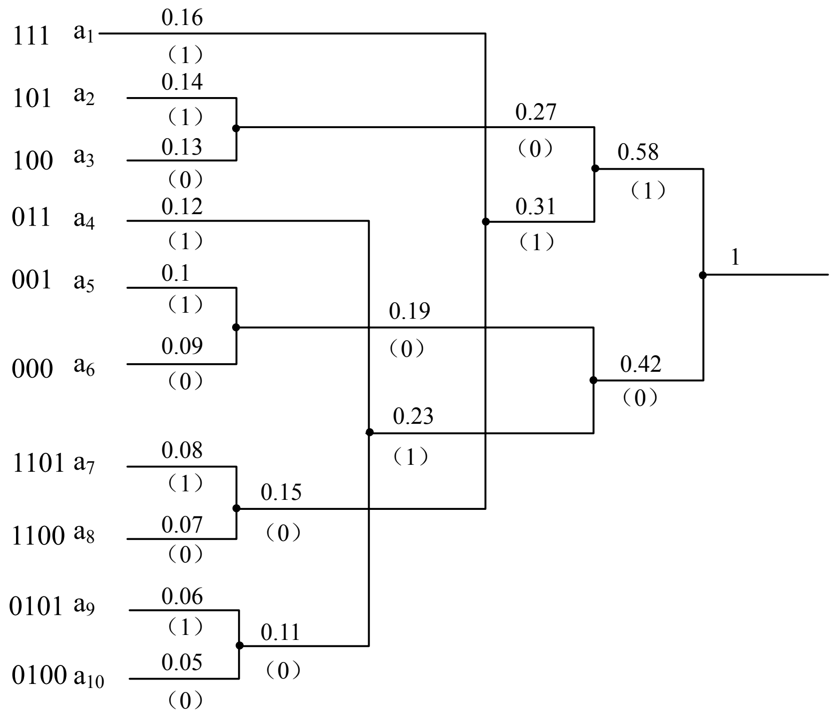
(b) 求三元 Huffman 码，计算 $\overline{n}$ 和 η；

[解]

(a) 由信源概率分布可知

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

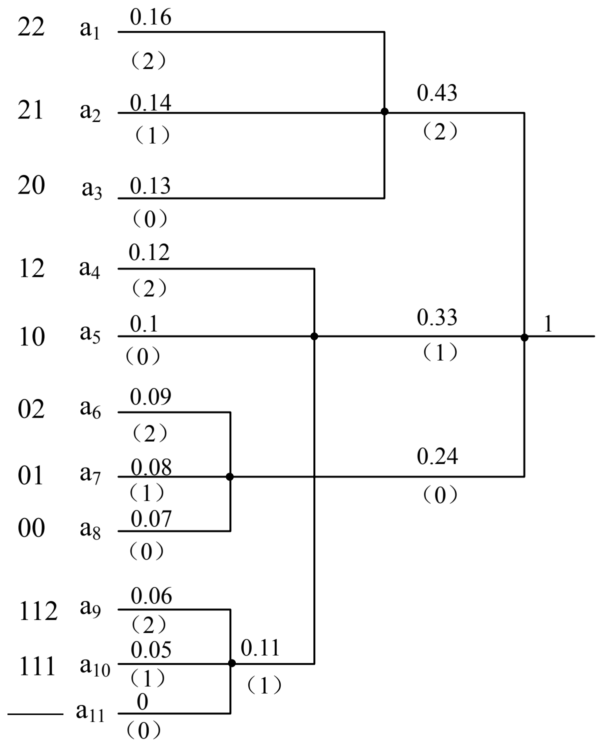
相应的 Huffman 编码过程如下图所示；



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

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

（b）三元 Huffman 码如下构成



$\overline{n} = 2.11$

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

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；

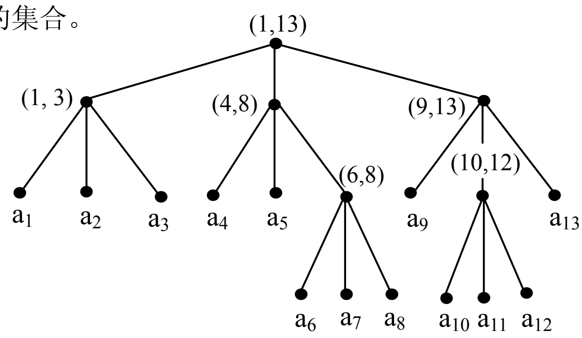
（c）{01，10}不可能成为 Huffman 码，因为显然{0，1}是平均码长更短的前缀码；

3.8 一个随机变量 X 的取值范围为 $\mathscr{X} = \{x_1, x_2 \cdots, x_m\}$,它的熵为 $H(X)$,若对这个源能找到一个平均码长为 $L = \dfrac{H(X)}{\log_2 3} \stackrel{\Delta}{=} H_3(X)$ 的三元即时码，试证

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

(2) 证明 m 为奇数；

[证明]（**a**）设 $\mathscr{D}$ 是平均码长 $L=\dfrac{H(X)}{\log 3}$ 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，使之为满树。现设消息总数为 m，消息集合为 $M=\{x_1,x_2,\cdots,x_m\}$，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 $\{x_i,x_{i+1},\cdots,x_j\}$，如果 $i=j$，则 (i,j) 表示该节点为树叶。



$\mathscr{I}$ 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：

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

于是

$$L=\sum_{i=1}^{n} l_i\cdot p_i=\sum_{(i,j)\in\mathscr{I}} p(i,j)$$

又由熵的可加性

$$H(X)=H(p_1,p_2,\cdots,p_n)=\sum_{(i,j)\in\mathscr{I}} p(i,j)H\left(\frac{p(i,k_1)}{p(i,j)},\frac{p(k_1+1,k_2)}{p(i,j)},\frac{p(k_2+1,j)}{p(i,j)}\right)$$

其中 (i,k_1)，(k_1+1,k_2)，(k_2+1,j) 是由内节点 (i,j) 分岔出去的三个节点，所以 $p(i,k_1)+p(k_1+1,k_2)+p(k_2+1,j)=p(i,j)$。由于码 $\mathscr{D}$ 的平均码长 $L=\dfrac{H(X)}{\log 3}$，所以

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

因为 $\quad p(i,j)>0 \quad , \quad \forall(i,j)\in\mathscr{I}$

所以要求对任何 $(i,j)\in\mathscr{I}$

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

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 $\dfrac{1}{3}$，第二层节点概率为 $\dfrac{1}{9}$，$\cdots$。**从而任何一个消息（树叶）出现概率必定为 $\dfrac{1}{3}$ 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 $L=\dfrac{H(X)}{\log 3}$ 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

$$N=[\log_2 5051]=13$$

（**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 $\sum_{i=1}^{m} 2^{-l_i}\le 1$ 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，…，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

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

[证明]（a）首先证明在约束 $\sum_{i=1}^{m} 2^{-l_i} \le 1$ 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 $q_i=\dfrac{p_i c_i}{\sum_{k=1}^{m} c_k \cdot p_k}$，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 $\bar{n}=\sum_{i=1}^{m}\dfrac{p_i c_i l_i}{\sum_{k=1}^{m} c_k p_k}$，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

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

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

3.16 设一个马尔可夫信源，其状态图如图所示。

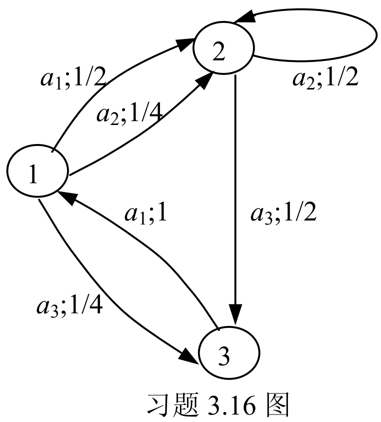
（a）求稳态下各状态概率 $q(i)$，以及各字母 a_i 的出现概率，i=1，2，3；

（b）求 $H(U\,|\,s_i)$，i=1，2，3；

（c）求 $H_\infty(U)$；

（d）对各状态 s_j 求最佳二元码；

（e）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

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

得到 $$q(1) = q(3) = \frac{2}{7}$$

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

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

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

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

（b） $$H(U \mid S_1) = 1.5bit$$

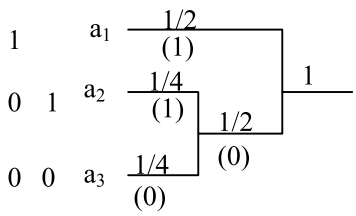
$$H(U \mid S_2) = 1bit$$

$$H(U \mid S_3) = 0bit$$

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

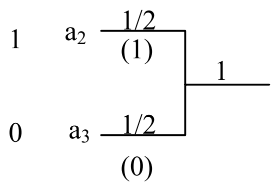
$$= \frac{2}{7}\cdot 1.5 + \frac{3}{7}\cdot 1 = \frac{6}{7}bit$$

（d）在 s_1 状态 Huffman 编码为：



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

在 s_2 状态 Huffman 编码为：



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

在 s_3 状态无需编码, $\overline{n}_3 = 0$

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