第四章无失真信源编码

第四节变长码

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

1. 引例子

内容提要

内容提要

内容提要

1. 引例子
2. 即时码
3. 最优码
4. 仙农码

例题 4.4.1：变长码引例

定长码的一个缺点是完全忽略了每个分组消息发生的概率可能不同这样一个事实，这时若使用定长码就可能起不到数据压缩的作用，若使用变长码常常会有更好的效果。

考虑下面单字符变长码：

字符	概率	码字	码长	等长码字
1	1/2	0	1	00
2	1/4	10	2	01
3	1/8	110	3	10
4	1/8	111	3	11

易知变长码是唯一可译码，并且平均码长为

$$\bar{L} = \sum_{x \in \mathcal{X}} p(x) l(x) = 1.75 \ \ bits = H(X).$$

例题 4.4.2：唯一可译码译码例

唯一可译码中有一种常用码称为即时码，在译码时只要发现是一个码字就可以立即翻译，没有延时。

已知三种二进编码

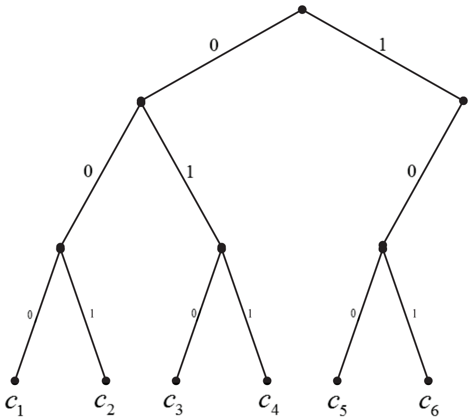
$$
\begin{aligned}
\mathcal{C}_1 &= \{000, 001, 010, 011, 100, 101\}\,, \\
\mathcal{C}_2 &= \{0, 10, 110, 1110, 11110, 111110\}\,, \\
\mathcal{C}_3 &= \{0, 01, 011, 0111, 01111, 011111\}\,.
\end{aligned}
$$

如果到达译码器码字序列是 $s = 010011101100111110$ *应当怎么译码（计时从左边开始）？*

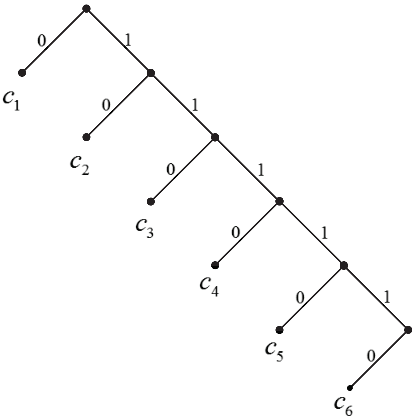
译码:

易检验这三个编码都是唯一可译码。因为 $\mathcal{C}_1$ 是定长码，译码器只要收到三个码符就立即作为码字来译码，不存在在延时，故 $s = 010.011.101.100.111.110$；若用 $\mathcal{C}_2$，也可以遇到码字就立即翻译，也不存在延时，故 $s = 0.10.0.1110.110.0.111110$；但若用 $\mathcal{C}_3$，当译码器收到码字序列 s 的第一个 0 时，不能确定它就是码字，因为还可能是另一个码字的前缀；当再收到第二个字符 1 时，也不确定 01 是一个码字，因为它可能是另一个码字的前缀；再收到第三个字符 0 时，因为 010 不是码字前缀，译码器才能确认 01 是码字，这时才可以将 01 翻译，因此有二个字符的延时。它们的码树如图 4-6，图 4-7。

码树图 4-6：



码树图 4-7：



定义 4.4.1：即时码

根据这个例子可知，在一个码字集中，若出现一个码字是另一个码字的前缀，译码时就会有延时。

*（即时码定义）如果唯一可译码的码字集中，没有一个码字是其它码字的前缀，则称这种编码为**即时码**。*

即时码码树特点：

为了弄清即时码的编码条件，需要借助码树来讨论。即时码的码树有如下特点：（1）如果某个结点已被用作码字，则该结点的后继结点就不能用作码字了；（2）不同码字结点的后继结点的集合是互不相交的；（3）如果某码字对应的码长为 l，则这个码字结点是 l 级结点，即在码树的第 l 层上。

定理 4.4.1：Kraft 不等式

（*Kraft1949*）*若有 m 个码字的 D 进即时码的码长序列为 $l_1, l_2, \cdots, l_m$，则它们满足* **Kraft 不等式**

$$\sum_{i=1}^{m} D^{-l_i} \leq 1. \tag{4.10}$$

~~(2.1)~~

反之，如果正整数列 $l_1, l_2, \cdots, l_m$ 满足 Kraft 不等式，则存在以 $l_1, l_2, \cdots, l_m$ 为码长的即时码。

证明：

设即时码的最大码长为 l_{max}，则 $l_{max} \geq 1$。在这个 D 进即时码树中第 l_{max} 层上的结点总数不超过 $D^{l_{max}}$ 个，可分为三类：
（1）恰好是一个码字结点；（2）是另外一个码字的后继结点；（3）既不是（1）也不是（2），即从根到叶子结点没有一个码字；

续证明：

如果一个码字结点在第 l_i 层上，则它的后继结点都不是码字结点，并且它在第 l_{max} 层上的后继结点总数是 $D^{l_{max}-l_i}$。但不同的码字结点它们的后继结点集是不相交的，故所有码字结点在第 l_{max} 层上的后继结点数之和不超过第 l_{max} 层上的总结点数，即

$$\sum_{i=1}^{m} D^{l_{max}-l_i} \leq D^{l_{max}},$$

从而

$$\sum_{i=1}^{m} D^{-l_i} \leq 1,$$

这就证明了 Kraft 不等式。

续证明：

如果给定一个码长序列 $l_1, l_2, \cdots, l_m$，它满足 Kraft 不等式，现在不妨设码长序列是单调递增的即 $l_1 \leq l_2 \leq \cdots \leq l_m$。先构造一个有 l_m 层的 D 进满树，分枝上权按由小到大从左到右排列。再从第 l_1 层的结点中选一个作为码字结点 c_1，并删除它的所有后继结点；然后假定已经找到了长度分别为 $l_1, \cdots, l_i$ 的码字结点，它们的后继结点也已经被删除；考虑是否还能取到码长为 l_{i+1} 的码字结点。对每个码长为 $l_j, 1 \leq j \leq i < m$ 的码字结点 c_j，它在第 l_{i+1} 层上的后继结点数为 $D^{l_{i+1}-l_j}$，不同的 c_j 它们后继结点集是不相交的，因此所有已知码字结点 $c_1, \cdots, c_i$ 在第 l_{i+1} 层上的后继结点总数为
$D^{l_{i+1}-l_1} + D^{l_{i+1}-l_2} + \cdots + D^{l_{i+1}-l_i}$，在第 l_{i+1} 层上还余下
$D^{l_{i+1}} - D_1^{l_{i+1}-l} - D^{l_{i+1}-l_2} - \cdots - D^{l_{i+1}-l_i}$ 个结点未被删除，第 $i+1$ 个码字结点 c_{i+1} 就要在这些余下结点中选择。

续证明：

利用 Kraft 不等式有

$$D^{-l_1}+D^{-l_2}+\cdots+D^{-l_i}<D^{-l_1}+D^{-l_2}+\cdots+D^{-l_m}\leq 1,$$

故两边同乘 $D^{l_{i+1}}$ 再移项得

$$D^{l_{i+1}}-D^{l_{i+1}-l_1}-D^{l_{i+1}-l_2}-\cdots-D^{l_{i+1}-l_i}\geq 1,$$

从而在余下的结点中选择一个码字结点是完全可能的。根据数学归纳法，这就证明了即时码的存在性。

定理 4.4.2：McMilan 定理

1956 年 McMilan 证明了 Kraft 不等式对唯一可译码也成立。

有 m 个码字的 D 进唯一可译码的码长序列 $l_1, l_2, \cdots, l_m$ 也满足 Kraft 不等式。

证明：自学

先作如下分析:

（1）设有唯一可译码 $\mathcal{C}=\{c_1,c_2,\cdots,c_m\}$，它们的码长序列为 $l_1,l_2,\cdots,l_m$。

（2）考虑由任意 n 个码字 $c_{i_1},c_{i_2},\cdots,c_{i_n}$ 所组成的码字序列 $M_i=c_{i_1}c_{i_2}\cdots c_{i_n}$，其长度 L_i 满足 $nl_{min}\le L_i=l_{i_1}+l_{i_2}+\cdots+l_{i_n}\le nl_{max}$，其中 l_{max},l_{min} 分别为最大与最小码长;

（3）在这些码序列中有些长度可能是相同的，记 N_i 为长度是 L_i 的码字序列的个数，所有长度为 L_i 的码字序列总数为 $D^{L_i}=D^{l_{i_1}+l_{i_2}+\cdots+l_{i_n}}$；由于是唯一可译码，故有所有长度为 L_i 的码字序列都是不同的，从而 $N_i\le D^{L_i}$。

（4）对每个由 n 个码字组成的码字序列都有一个重复排列 $i_1i_2\cdots i_n$ 相对应，所有由 n 个码字构成的码序列的总数为 m^n。

续证明：

考虑求和

$$\begin{aligned}
\sum_{i=1}^{m^n} D^{-L_i} &= \sum_{L_i=nl_{min}}^{nl_{max}} N_i D^{-L_i} \\
&\leq \sum_{L_i=nl_{min}}^{nl_{max}} D^{L_i} D^{-L_i} \\
&\leq \sum_{L_i=nl_{min}}^{nl_{max}} 1 \\
&= nl_{max} - nl_{min} + 1,
\end{aligned}$$

续证明：

又

$$
\begin{aligned}
\sum_{i=1}^{m^n} D^{-L_i} &= \sum_{i_1 i_2 \cdots i_n} D^{-L_i} = \sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m} D^{-(l_{i_1}+l_{i_2}+\cdots+l_{i_n})} \\
&= (\sum_{i_1=1}^{m} D^{-l_{i_1}})(\sum_{i_2=1}^{m} D^{-l_{i_2}})\cdots(\sum_{i_n=1}^{m} D^{-l_{i_n}}) \\
&= (\sum_{i=1}^{m} D^{-l_i})(\sum_{i=1}^{m} D^{-l_i})\cdots(\sum_{i=1}^{m} D^{-l_i}) \\
&= (D^{-l_1} + D^{-l_2} + \cdots + D^{-l_m})^n \\
&= (\sum_{i=1}^{m} D^{-l_i})^n,
\end{aligned}
$$

续证明：

于是

$$(\sum_{i=1}^{m} D^{-l_i})^n \le nl_{max} - nl_{min} + 1,$$

从而

$$\sum_{i=1}^{m} D^{-l_i} \le (nl_{max} - nl_{min} + 1)^{1/n} \le (nl_{max})^{1/n},$$

由 n 的任意性两边取极限得所证。
虽然 Kraft 不等式对唯一可译码也成立，但这个不等式不能用于判断编码是否唯一可译码或即时码，但可以用定义或码树来判断是否是即时码。

假定：

问题

设离散无记忆信源具有概率分布（4-2），即：

$$X \sim p(x) = \begin{pmatrix} a_1 & a_2 & \cdots & a_N \\ p_1 & p_2 & \cdots & p_N \end{pmatrix},$$

在字符空间 $\mathcal{X} = \{a_1, a_2, \cdots, a_N\}$ 所有 D 进唯一可译码或即时码中，是否存在平均码长最小的码？

最优码存在性

(引理 4.4.1)

设字符空间 $\mathcal{X}=\{a_1,a_2,\cdots,a_N\}$ 具有概率分布（4-2），则字符空间 $\mathcal{X}$ 的所有唯一可译码平均码长集合与所有即时码平均码长集合相同。

利用 Kraft 不等式易证明。

(定理 4.4.3)

*设字符空间 $\mathcal{X}=\{a_1,a_2,\cdots,a_N\}$ 具有概率分布（4-2），则字符空间 $\mathcal{X}$ 的所有 D 进即时码中一定存在平均码长最小的即时码（称为**最优码**）。*

证明:

因为总存在正整数 k 满足 $D^k > N$，故可以构造码字长度为 k 的等长码，所以最优码的平均码长不会超过 k。现在设 $\mathcal{D}$ 是字符空间 $\mathcal{X}$ 上任意一个 D 进即时码，并且平均码长 $L(\mathcal{D}) \le k$。如果它的码长序列为 $l_1, l_2, \cdots, l_N$，则有

$$p_i l_i \le \sum_i p_i l_i = L(\mathcal{D}) \le k,$$

据此得

$$l_i \le \frac{k}{p_i} \le \frac{k}{\min_i p_i},$$

这说明: 即时码 $\mathcal{D}$ 中每个码字的码长仅可以选择有限个可能的值，这也意味着满足 $L(\mathcal{D}) \le k$ 的即时码 $\mathcal{D}$ 仅有有限个，从这有限个即时码当中选择平均码长最小的即时码就是最优码。

最优码长应当怎么选取

事实上它们应当是下面整数最优化问题

$$\bar{L} = \sum_{i=1}^{N} p_i l_i = \min,$$
$$\sum_{i=1}^{N} D^{-l_i} \le 1,$$
$$l_1, l_2, \cdots, l_N \text{是正整数}.$$

的解。这时可以利用 Lagrange 乘数法，对函数

$$F(l_1, \cdots, l_N, \lambda) = \sum_{i=1}^{N} p_i l_i + \lambda \left(\sum_{i=1}^{N} D^{-l_i} - 1 \right)$$

最优码长：

求驻点后可得

$$p_i = D^{-l_i}, i = 1, 2, \cdots, N.$$

于是最优码长可能就是

$$l_i = \log_D \frac{1}{p_i}, i = 1, 2, \cdots, N.$$

上式确定的 l_i 不一定是整数，但是它很接近最优码长 $l_1^*, l_2^*, \cdots, l_N^*$。

定理 4.4.4：平均码长

一般地可以证明最小平均码长就是信源的熵率。

设离散无记忆信源有概率分布（4-2）。（1）对字符集 $\mathcal{X}$ 的任何一个 D 进唯一可译码的平均码长成立

$$\bar{L} = \sum_{i=1}^{N} p_i l_i \geq H_D(X), \qquad 4.11 \quad \cancel{(3.1)}$$

其中等号成立充要条件是 $p_i = D^{-l_i}, i = 1, 2, \cdots, N$。
（2）总存在字符集 $\mathcal{X}$ 的一个 D 进唯一可译码使得平均码长满足不等式：

$$H_D(X) \leq \bar{L} = \sum_{i=1}^{N} p_i l_i < H_D(X) + 1. \qquad 4.12 \quad \cancel{(3.2)}$$

证明:

第（1）条证明：设随机变量 X 的一个 D 进唯一可译码的码长为 $l_1, l_2, \cdots, l_N$，记 $C = D^{-l_1} + D^{-l_2} + \cdots + D^{-l_N}$，则由 Kraft 不等式得

$$0 < C \leq 1, \log \frac{1}{C} \geq 0.$$

令 $\pi_i = D^{-l_i}/C, i = 1, 2, \cdots, N$，则 $\pi_1, \pi_2, \cdots, \pi_N$ 构成一个概率分布。又

续证明：

$$
\begin{aligned}
\bar{L} - H_D(X) &= \sum_i p_i l_i - \sum_i p_i \log_D \frac{1}{p_i} \\
&= \sum_i p_i \log_D \frac{p_i}{D^{-l_i}} = \sum_i p_i \log_D \frac{p_i}{C\pi_i} \\
&= \sum_i p_i \log_D \frac{p_i}{\pi_i} + \log \frac{1}{C} \geq 0,
\end{aligned}
$$

因为最后两项均非负，所以 $\bar{L} \geq H_D(X)$，并且等号成立条件为

$$
\sum_i p_i \log_D \frac{p_i}{\pi_i} = 0, \log \frac{1}{C} = 0,
$$

从而得

$$
p_i = \pi_i, C = 1, i = 1, 2, \cdots, N,
$$

此即 $p_i = D^{-l_i}, i = 1, 2, \cdots, N$。

续证明：

第（2）条证明：可以选择正整数 l_i 使得：

$$\frac{1}{D^{l_i}} \leq p_i < \frac{1}{D^{l_i-1}},$$

对这样选择的一列数 $l_1, l_2, \cdots, l_N$ 成立 Kraft 不等式：

$$\sum_{i=1}^{N} D^{-l_i} \leq \sum_{i=1}^{N} p_i = 1,$$

续证明:

故存在以 $l_1, l_2, \cdots, l_N$ 为码长的即时码，并且这种码使

$$
\begin{aligned}
\sum_{i=1}^{N} p_i \log_D p_i &< \sum_{i=1}^{N} p_i \log_D \frac{1}{D^{l_i-1}} \\
&= \sum_{i=1}^{N} [p_i(1-l_i)] = \sum_{i=1}^{N} p_i - \sum_{i=1}^{N}(p_i l_i) = 1 - \bar{L},
\end{aligned}
$$

从而得到 $\bar{L} < 1 + H_D(X)$。不等式另一半就是第（1）条。如果信源字符的概率分布 $p_1, p_2, \cdots, p_N$ 已知，则可以选择码长 l_i 使 $D^{-l_i} = p_i$，这样就可以使平均码长最小，意即使每个信源字符平均占用的码符数最少，达到最佳压缩效果。

定理 4.4.5：长消息编码

对于 n 长分组消息有如下编码定理。

对离散无记忆信源，总存在 n 长消息集合 $\mathcal{X}^n$ 上的一个 D 进唯一可译码使得码率 平均码长

$$\bar{l}_n = \frac{\sum_{x^{(n)} \in \mathcal{X}^n} p(x^{(n)}) l(x^{(n)})}{n}$$

满足不等式

4.13

$$H_D(X) \le \bar{l}_n < H_D(X) + \frac{1}{n}. \qquad \cancel{(3.3)}$$

推广：

可以将上面结果推广到一般离散信源上。

*对离散信源 $X_1, X_2, \cdots$，总存在 n 长分组消息集合 $\mathcal{X}^n$ 上的一个 D 进唯一可译码使得码率*平均码长

$$\bar{l}_n = \frac{\sum_{x^{(n)} \in \mathcal{X}^n} p(x^{(n)}) l(x^{(n)})}{n},$$

满足不等式

$$H_D(X_1, X_2, \cdots, X_n) \leq \bar{l}_n < H_D(X_1, X_2, \cdots, X_n) + \frac{1}{n}. \qquad \cancel{(3.4)}$$

特别如果这个离散信源又是平稳的，则 $\bar{l}_n \to H_{D\infty}(X)$（熵率）。

总结：

对离散无记忆信源的 n 长消息集进行 D 进变长编码，只要所给消息长度 n 充分大，总可以找到一种唯一可译码，使得这个唯一可译码的码率充分接近信源的 D 进熵。对有记忆的信源也成立类似的编码定理，可参考 [5,20]。

平均码长

求码长：

设离散无记忆信源 X 具有概率分布（4.2），考虑对它的 n 长消息进行即时码编码。为此要先求出 n 长消息的概率分布 $p(x^{(n)}), x^{(n)} \in \mathcal{X}^n$，再根据定理4.1.4选取码长 $l(x^{(n)})$ 使 $p(x^{(n)}) = D^{-l(x^{(n)})}$，就可以达到最小平均码长，于是取

$$l(x^{(n)}) = \log_D \frac{1}{p(x^{(n)})},$$

但它未必是整数，可取

$$l(x^{(n)}) = \left\lceil \log_D \frac{1}{p(x^{(n)})} \right\rceil, \tag{4.14}$$

它表示满足

$$\log_D \frac{1}{p(x^{(n)})} \leq l(x^{(n)}) < \log_D \frac{1}{p(x^{(n)})} + 1, \qquad \overset{4.15}{\cancel{(4.2)}}$$

最小整数。

仙农码的编码算法:

（1）求出所有 n 长消息的概率

$$r_k = p(\alpha_k), \alpha_k \in \mathcal{X}^n, k = 1, 2, \cdots, N^n,$$

并将所有 N^n 个 n 长消息按它的概率从大到小降序排列

$$\begin{array}{ccccccccc} \alpha_1 & & \alpha_2 & & \alpha_3 & & \cdots & & \alpha_{N^n} \\ r_1 & \geq & r_2 & \geq & r_3 & \geq & \cdots & \geq & r_{N^n} \end{array}.$$

（2）] 求第 k 个 n 长消息 α_k 的码长

$$l_k = l(\alpha_k) = \left\lceil \log_D \frac{1}{r_k} \right\rceil, k = 1, 2, \cdots, N^n.$$

续：

（3）求第 k 个消息的累积概率 P_k

$$P_1 = 0, P_k = \sum_{i=1}^{k-1} r_i, k = 2, 3, \cdots, N^n.$$

（4）将每个累积概率 $P_1, P_2, P_3, \cdots, P_{N^n}$ 用 D 进制小数表示，并取小数点后的 $l_1, l_2, \cdots, l_{N^n}$ 个 D 进数字作为消息 $\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_{N^n}$ 的 D 进编码。

定理 4.4.6:

仙农码一定是即时码。

由仙农码的编码方法，设消息 α_i, α_j 对应的 D 进码字分别为

$$x_1x_2\cdots x_{l_i} \quad , \quad y_1y_2\cdots y_{l_j},$$

它们分别对应两个 D 进小数

$$\tilde{\alpha}_i = 0.x_1x_2\cdots x_{l_i} \quad , \quad \tilde{\alpha}_j = 0.y_1y_2\cdots y_{l_j},$$

并且对应的累积概率为

$$P_i = 0.x_1x_2\cdots x_{l_i}x_{l_i+1}\cdots \quad , \quad P_j = 0.y_1y_2\cdots y_{l_j}y_{l_j+1}\cdots \ ,$$

续证明：

不仿设 $i < j$ 则 $r_i \geq r_j$，从而 $l_i \leq l_j$。如果消息 α_i 的码字是消息 α_j 的码字的前缀，则有

$$P_j = 0.x_1 x_2 \cdots x_{l_i} y_{l_i+1} \cdots y_{l_j} y_{l_j+1} \cdots ,$$

从而

$$P_i, P_j \in \left[\tilde{\alpha}_i, \tilde{\alpha}_i + D^{-l_i}\right),$$

故有

$$r_i \leq P_j - P_i = r_i + \cdots + r_{j-1} < D^{-l_i},$$

导致

$$l_i < \log_D \frac{1}{r_i},$$

这与码长的选择矛盾！当然也可以不对所有 n 长消息序列编码，而只对弱典型序列进行仙农编码。

定理 4.4.7：平均码长

对离散无记忆信源 X 的 n 长消息进行仙农码编码，则码率满足不等式（4-13）即：

$$H_D(X) \leq R < H_D(X) + \frac{1}{n}.$$

证明：

由不等式（4-15）可得：

$$p(x^{(n)})\log_D \frac{1}{p(x^{(n)})} \le p(x^{(n)})l(x^{(n)}) < p(x^{(n)})\log_D \frac{1}{p(x^{(n)})} + p(x^{(n)}),$$

$$\sum p(x^{(n)})\log_D \frac{1}{p(x^{(n)})} \le \sum p(x^{(n)})l(x^{(n)}) < \sum p(x^{(n)})\log_D \frac{1}{p(x^{(n)})} +$$

$$H_D(X_1, X_2, \cdots, X_n) \le \bar{L} < H_D(X_1, X_2, \cdots, X_n) + 1,$$

$$nH_D(X) \le \bar{L} < nH_D(X) + 1,$$

从而得

$$H_D(X) \le \frac{\bar{L}}{n} < H_D(X) + \frac{1}{n},$$

注意到按定义 4.1.2，$\bar{L}/n$ 就是码率，如果 n 充分大，码率 R 趋于熵率 $H_D(X)$。

定理 4.4.8:

这个定理可以推广到更一般的离散信源。

对离散平稳信源 X 的 n 长消息 $(X_1, X_2, \cdots, X_n)$ 进行仙农码编码，则它的平均码长满足

$$\frac{H_D(X_1, X_2, \cdots, X_n)}{n} \leq \frac{\bar{L}}{n} \leq \frac{H_D(X_1, X_2, \cdots, X_n)}{n} + \frac{1}{n}.$$

假定离散平稳信源 X 熵率存在，则有极限

$$\lim_{n \to \infty} \frac{\bar{L}}{n} = H_\infty(X),$$

即码率趋于信源的熵率。

例题4.4.3

设有离散无记忆信源字符空间中概率分布为

$$X \sim p(x) = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 1/4 & 1/8 & 1/8 & 6/16 & 1/16 & 1/16 \end{pmatrix}.$$

（1）对给定的码长序列 $l_1 = l_2 = 1, l_3 = 2, l_4 = l_5 = 4, l_6 = 5$ 求每个字符对应的二进及三进即时码。（2）求每个字符对应的二进与三进仙农码。

解：求具有指定码长的即时码

（1）易计算

$$\sum_{i=1}^{6} 2^{-l_i} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^4} + \frac{1}{2^6} = \frac{45}{32} > 1,$$

故不存在二进即时码。但是

$$\sum_{i=1}^{6} 3^{-l_i} = \frac{1}{3}+\frac{1}{3}+\frac{1}{3^2}+\frac{1}{3^4}\frac{1}{3^4}\frac{1}{3^5} = \frac{3^4+3^4+3^3+3+3+1}{3^5} = \frac{196}{243} < 1,$$

因此存在三进即时码，可以借助码树图 **??** 来编码，注意要擦除所有码字结点的后继结点。

x_1	x_2	x_3	x_4	x_5	x_6
0	1	20	2201	2202	22000

需要注意的是，这种具有同样码长的 3 进即时码可以不止一个。

三进即时码码树：

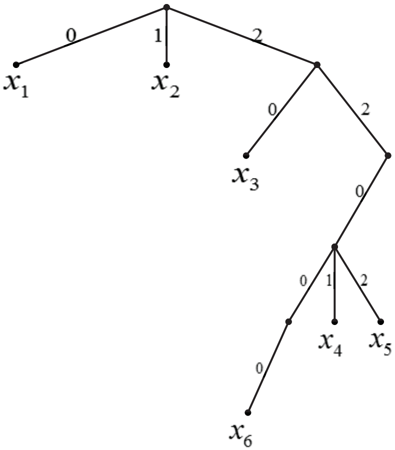


Figure: 图 4-8

续解：求二进仙农码码长

（2）现在取二进仙农码的码长：$l(x_i) = \left\lceil \log_2 \frac{1}{p(x_i)} \right\rceil$，容易求得：

$$l_1 = l(x_1) = \left\lceil \log_2 \frac{1}{p(x_1)} \right\rceil = \lceil \log_2 4 \rceil = 2,$$

$$l_2 = l(x_2) = \left\lceil \log_2 \frac{1}{p(x_2)} \right\rceil = \lceil \log_2 8 \rceil = 3,$$

$$l_3 = l(x_3) = \left\lceil \log_2 \frac{1}{p(x_3)} \right\rceil = \lceil \log_2 8 \rceil = 3,$$

2.666

$$l_4 = l(x_4) = \left\lceil \log_2 \frac{1}{p(x_4)} \right\rceil = \left\lceil \log_2 \frac{16}{6} \right\rceil = 2,$$

$$l_5 = l(x_5) = \left\lceil \log_2 \frac{1}{p(x_5)} \right\rceil = \lceil \log_2 16 \rceil = 4,$$

$$l_6 = l(x_6) = \left\lceil \log_2 \frac{1}{p(x_6)} \right\rceil = \lceil \log_2 16 \rceil = 4,$$

续解：求二进仙农码

据此可以构造一个二进仙农码，编码过程如下表：

符号	概率	累积概率	二进制小数	仙农码长	仙农码
x_4	6/16	$P_1 = 0$	0.00	2	00
x_1	1/4	$P_2 = 6/16$	0.011	2	01
x_2	1/8	$P_3 = 10/16$	0.101	3	101
x_3	1/8	$P_4 = 12/16$	0.11	3	110
x_5	1/16	$P_5 = 14/16$	0.111	4	1110
x_6	1/16	$P_6 = 15/16$	0.1111	4	1111

二进仙农码平均码长：

其平均码长为

$$\bar{L}=\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3+\frac{6}{16}\times 2+\frac{1}{16}\times 4+\frac{1}{16}\times 4=40/16=2.5\text{ bits},$$

而二进熵 $H_2(X)=2.28$ bits ， 正好有
$H_2(X)\le\bar{L}<H_2(X)+1$。
但还可以有这样的二进即时码

符号	x_1	x_2	x_3	x_4	x_5	x_6
码字	100	101	110	0	1110	1111

其平均码长为 $\bar{L}=38/16=2.375$ bits，这说明还有比仙农码更好的即时码。

续解：求三进仙农码

（3）现在取三进仙农码的码长：$l(x_i)=\left\lceil \log_3 \frac{1}{p(x_i)} \right\rceil$，容易求得

$$l_1=l(x_1)=\left\lceil \log_3 \frac{1}{p(x_1)} \right\rceil = l(x_i) = \lceil \log_3 4 \rceil = 2,$$

$$l_2=l(x_2)=\left\lceil \log_3 \frac{1}{p(x_2)} \right\rceil = l(x_i) = \lceil \log_3 8 \rceil = 2,$$

$$l_3=l(x_3)=\left\lceil \log_3 \frac{1}{p(x_3)} \right\rceil = l(x_i) = \lceil \log_3 8 \rceil = 2,$$

2.666

$$l_4=l(x_4)=\left\lceil \log_3 \frac{1}{p(x_4)} \right\rceil = l(x_i) = \left\lceil \log_3 \frac{16}{6} \right\rceil = 1,$$

$$l_5=l(x_5)=\left\lceil \log_3 \frac{1}{p(x_5)} \right\rceil = l(x_i) = \lceil \log_3 16 \rceil = 3,$$

$$l_6=l(x_6)=\left\lceil \log_3 \frac{1}{p(x_6)} \right\rceil = l(x_i) = \lceil \log_3 16 \rceil = 3,$$

续解：求三进仙农码

据此可以构造一个三进仙农码，编码过程如下表：

符号	概率	累积概率	三进制小数	仙农码长	仙农码
x_4	6/16	$P_1 = 0$	0.00	1	0
x_1	1/4	$P_2 = 6/16$	0.101010...	2	10
x_2	1/8	$P_3 = 10/16$	0.121212...	2	12
x_3	1/8	$P_4 = 12/16$	0.202020...	2	20
x_5	1/16	$P_5 = 14/16$	0.212121...	3	212
x_6	1/16	$P_6 = 15/16$	0.22102210...	3	221

平均码长：

其平均码长为

$$\bar{L} = \frac{1}{4}\times 2+\frac{1}{8}\times 2+\frac{1}{8}\times 2+\frac{6}{16}\times 1+\frac{1}{16}\times 3+\frac{1}{16}\times 3 = 28/16 = 1.75\text{三进制单位}$$

而熵 $H_3(X) = 1.4389$三进制单位 正好有
$H_3(X) \leq \bar{L} < H_3(X) + 1$。
但还可以有这样的三进即时码

符号	x_1	x_2	x_3	x_4	x_5	x_6
码字	1	20	21	0	220	221

其平均码长为 $\bar{L} = 24/16 = 1.5$ 三进制单位，这说明还有比仙农码更好的即时码。

例题 4.4.4：

设二进离散无记忆信源字符分布为 $p(0)=2/3, p(1)=1/3$，如果对 n 长消息 $x^{(n)}=x_1x_2\cdots x_n$ 进行二进仙农码编码，记 L_n 是平均码长，证明

$$\lim_{n\to\infty}\frac{L_n}{n}=H(X).$$

证明：（1）易求得信源的熵率 $H(X)=\log_2 3-2/3$ bits。

续证明：

（2）指定 k 个位置恰好为 0 的 n 长消息 $x^{(n)} = x_1 x_2 \cdots x_n$ 发生概率为

$$p(x^{(n)}) = \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{n-k} = \frac{2^k}{3^n},$$

这种形式的消息总共有 C_n^k 个，每个消息对应着相同的仙农码码长

$$\begin{aligned} l_k &= \left\lceil \log_2 \frac{1}{p(x^{(n)})} \right\rceil = \left\lceil \log_2 \frac{3^n}{2^k} \right\rceil \\ &= \lceil n \log_2 3 - k \rceil = \lceil n \log_2 3 \rceil - k. \end{aligned}$$

续证明：

（3）平均码长为

$$
\begin{aligned}
\bar{L}_n &= \sum_{k=0}^{n} C_n^k p(x^{(n)}) l_k = \sum_{k=0}^{n} C_n^k p(x^{(n)}) \left[\lceil n\log_2 3\rceil - k\right] \\
&= \frac{1}{3^n}\left[\lceil n\log_2 3\rceil \sum_{k=0}^{n} C_n^k 2^k - \sum_{k=0}^{n} C_n^k k 2^k\right] \\
&= \frac{1}{3^n}\left[\lceil n\log_2 3\rceil 3^n - 2n3^{n-1}\right] \\
&= \lceil n\log_2 3\rceil - \frac{2n}{3}.
\end{aligned}
$$

续证明：

（4）再由于

$$n\log_2 3 \le \lceil n\log_2 3\rceil < n\log_2 3 + 1,$$

$$\log_2 3 - \frac{2}{3} < \frac{L_n}{n} < \log_2 3 - \frac{2}{3} + \frac{1}{n},$$

故得

$$\lim_{n\to\infty}\frac{L_n}{n} = \log_2 3 - \frac{2}{3}.$$

这个例子说明，当采用仙农码编码时随着消息长度的增大，平均每个信源字符占用的码符数接近信源的熵。