XI'AN JIAOTONG UNIVERSITY

信息论与编码

第4章 信源编码

张建国

主要内容与基本要求

XI'AN JIAOTONG UNIVERSITY

- 主要内容
  - 序列通过信道传输
  - 定长信源编码
  - 变长信源编码，最佳变长编码定理及编码方法
- 基本要求
  - 理解信源编码的目的；了解序列传输的特点；
  - 掌握定长及变长的信源编码定理；
  - 理解唯一可解、异前缀码的概念，理解Kraft不等式；
  - 掌握二元及多元的最佳变长编码技术即Huffman编码方法，理解扩展编码方法，会计算编码效率。

本章目录

XI'AN JIAOTONG UNIVERSITY

4.0 引言

XI'AN JIAOTONG UNIVERSITY

信息的度量、信道和信道编码定理是信息论的三大基础。

前面已定义了消息所含的信息量和信道容量，如果这些定义都是正确的，则可以期望：**只要消息所含信息量不超过信道容量，总可以无误传输**。本章和下一章将围绕编码问题展开讨论，证明上述结论的正确性。

遗憾的是，目前只能证明这一结论是正确的，却无法指明如何才能达到这一目标。因此，编码定理只是给出了信息传输的理想目标或性能上界，而具体如何向这一目标靠近则形成了一个重要且庞大的分支编码技术。总的来说，编码的作用是将消息序列转换为适合于信道的形式，从而实现**快速**（**高效**）又**可靠**（**准确**、**无误**）的传输。

同时达到上述两个目标（去除冗余的同时保护信息）比较困难。一方面是这两个目标不统一；而另一方面信道是多种多样的，同一信源在通过不同信道时需要的保护不同。因此，为了实现方便，通常将编码过程分为信源编码和信道编码两部分。

4.0 引言

信源编码是在假设信道理想（或者说不考虑具体信道）的条件下，如何以最快的速度(或最少的次数)将给定信源序列送过信道。信源编码追求的是信息传输的高效性、快速性。

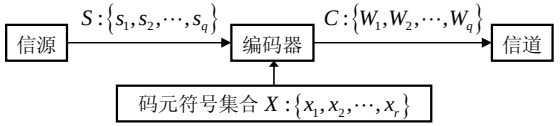
信道编码是对信源编码后的序列施加一定的保护措施，使其准确可靠地通过给定的实际信道（非理想信道）。信道编码追求的是信息传输的准确性、可靠性。

要充分利用信道容量是有条件的，要求信道的输入在某种程度上是可以分割组合的。前面提到的传输信息所用的单独二进制符号是无法再分割的。编码一般是将高进制符号转换为低进制符号，或将长序列变成短序列的转换，这样才能够分割与组合。

4.0 引言

编码器

编码实质上是对信源的原始符号按一定的数学规则进行的一种变换。



编码器将信源符号（或信源符号序列）转换为由码元符号（适合信道传输的符号）组成的码字。即

$$s_i\,(i=1,\cdots,q)\leftrightarrow W_i=\left(x_{i_1}x_{i_2}\cdots x_{i_{l_i}}\right),x_{i_k}\in X\,(k=1,2,\cdots,l_i)$$

或 $$\boldsymbol{\alpha}_i=\left(s_{i_1}s_{i_2}\cdots s_{i_N}\right)\leftrightarrow W_i=\left(x_{i_1}x_{i_2}\cdots x_{i_{l_i}}\right)$$

$$s_{i_k}\in S\,(k=1,\cdots,N);x_{i_k}\in X\,(k=1,2,\cdots,l_i)$$

4.0 引言

码的属性与分类

分组码与非分组码；定长（等长）码与变长码；同价码。

非奇异码与奇异码。如果不同的信源符号映射到不同的码符号序列，称为非奇异码；否则称为奇异码。

唯一可译码与非唯一可译码，若码的任意一串有限长的码符号序列只能被唯一地译成所对应的信源符号序列，则称该码为唯一可译码。

即时码（非延长码）与非即时码。如果接收端收到一个完整的码字后，不能立即译码，还需要等下一个码字开始接收后才能判断是否可以译码，这样的码称为非即时码。

第三章讨论的是单个符号通过信道或者说信道使用一次。而在编码时通常是以序列为单位的，因此接下来首先讨论序列传输的基本概念以及与符号传输之间的关系。

4.1 序列传输

设一离散信源发出数符序列 $u^L=(u_1,u_2,\cdots,u_L)$，其中 $u_l\in\{a_1,a_2,\ldots,a_K\}$，首先考察发出序列 u^L 的**概率**。信源的统计特性不随时间改变（时不变）。

$$\begin{aligned}p(u^L)&=p(u_1,u_2,\cdots,u_L)\\&=p(u_1)p(u_2,\cdots,u_L\,|\,u_1)\\&=p(u_1)p(u_2\,|\,u_1)p(u_3,\cdots,u_L\,|\,u_1u_2)\\&=\cdots\\&=p(u_1)p(u_2\,|\,u_1)\cdots p(u_L\,|\,u_1\cdots u_{L-1})\end{aligned}$$

如果信源无记忆，则有：

$$p(u^L)=p(u_1\cdots u_l\cdots u_L)\overset{\text{无记忆}}{=}p(u_1)p(u_2)\ldots p(u_L)=\prod_{l=1}^{L}p(u_l)$$

离散无记忆信源所产生序列的概率等于序列各符号发生的概率之积。

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

序列的信息量

$$I(u^L)=\log\frac{1}{p(u^L)}=\log\frac{1}{p(u_1,u_2,\cdots,u_L)}$$

$$=\log\frac{1}{p(u_1)p(u_2\,|\,u_1)\cdots p(u_L\,|\,u_1\cdots u_{L-1})}$$

$$=\log\frac{1}{p(u_1)}+\log\frac{1}{p(u_2|u_1)}+\cdots+\log\frac{1}{p(u_L|u_1u_2\cdots u_{L-1})}$$

$$=I(u_1)+I(u_2|u_1)+\cdots+I(u_L|u_1u_2\cdots u_{L-1})$$

若信源无记忆，则有：

$$I\left(u^L\right)=I(u_1)+I(u_2)+\cdots+I(u_L)=\sum_{l=1}^{L}I(u_l)$$

离散无记忆信源产生的序列的信息量等于序列各符号携带的信息量之和。

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

序列的熵

$$H(U^L)=\sum_{u^L}p(u^L)I(u^L)$$

$$\underline{\underline{\text{无记忆}}}\sum_{u^L}p(u_1,u_2,\cdots,u_L)\left[I(u_1)+I(u_2)+\cdots+I(u_L)\right]$$

$$\underline{\underline{\text{无记忆}}}\sum_{u_1}\sum_{u_2}\cdots\sum_{u_L}p(u_1)p(u_2)\cdots p(u_L)\left[I(u_1)+I(u_2)+\cdots+I(u_L)\right]$$

$$=\sum_{u_1}p(u_1)I(u_1)+\sum_{u_2}p(u_2)I(u_2)+\cdots+\sum_{u_L}p(u_L)I(u_L)$$

$$=H(U_1)+H(U_2)+\cdots+H(U_L)=\sum_{l=1}^{L}H(U_l)$$

$$\underline{\underline{\text{时不变}}}LH(U)$$

长为L的离散无记忆序列的熵等于一个符号熵的L倍。

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

回忆：有条件的信息熵不超过无条件的信息熵。

$$I(X;Y)=H(X)-H(X\,|\,Y)\geq 0 \Rightarrow H(X)\geq H(X\,|\,Y)$$

有记忆时序列的信息量

$$I(u^L)=I(u_1)+I(u_2\,|\,u_1)+\cdots+I(u_L\,|\,u_1\cdots u_{L-1})$$

求平均，得到熵

$$H(U^L)=H(U_1)+H(U_2\,|\,U_1)+\cdots+H(U_L\,|\,U_1\cdots U_{L-1})$$

无记忆时序列的信息量 $I(u^L)=I(u_1)+I(u_2)+\cdots+I(u_L)$

无记忆时序列的熵

$$H(U^L)=H(U_1)+H(U_2)+\cdots+H(U_L)$$

由于 $H(U_2\,|\,U_1)\leq H(U_2)\cdots\cdots H(U_L\,|\,U_1\cdots U_{L-1})\leq H(U_L)$

所以：有记忆时序列的熵小于无记忆时序列的熵。

记忆减少了信息量，有记忆就表示有冗余

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

要将长度为N的序列传过离散无记忆时不变信道，共要使用信道N次。

K进J出的无记忆信道传N次，输入记为X^N，输出记为Y^N，

$$X^N=(x_1,x_2,\cdots,x_N),\quad x_i\in\{a_1,a_2,\cdots,a_K\}$$

$$Y^N=(y_1,y_2,\cdots,y_N),\quad y_j\in\{b_1,b_2,\cdots,b_J\}$$

信道的输入有K^N种，信道的输出有J^N种。

此时可将整个传递的信道看作有K^N个输入，J^N个输出的信道用了一次，二者在传递的效果上是等价的。

等效信道的前向转移概率，即传送序列时的转移概率

$$p(y^N\,|\,x^N)\underline{\underline{\text{无记忆}}}\prod_{n=1}^{N}p(y_n\,|\,x_n)$$

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

例4.1.1：二元对称信道BSC用两次，即K=J=2, N=2，总的来看输入和输出序列均为00, 01, 10, 11共四个。相当于一个四进制数符传过信道。

显然，该信道的前向转移概率共16个：

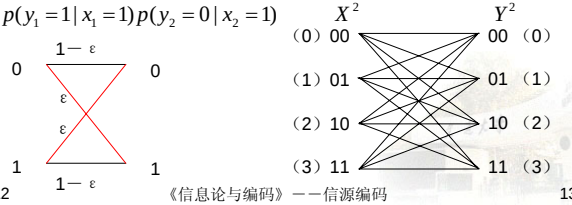
$$p(00|00),\ p(01|00),\ p(10|00),\cdots,\ p(11|11)$$

下面看两种信道的等价关系：

$p(y^2=2|x^2=3)$　　四元信道

$=p(y^2=10|x^2=11)$　　二元信道使用两次

$=p(y_1=1|x_1=1)p(y_2=0|x_2=1)$



4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

下面看序列通过信道的互信息熵。

对于符号的传递，信道传递一次的互信息熵为 $I(X;Y)$

传递一长度为N的序列时互信息熵为

$$I(X^N;Y^N)=H(X^N)-H(X^N|Y^N)$$

对时不变无记忆信道，有

$$H(X^N)=NH(X)$$

$$H(X^N|Y^N)=\sum_{x^N,y^N}p(x^N,y^N)\log\frac{1}{p(x^N|y^N)}$$

$$\underline{\underline{\text{无记忆}}}\sum_{x_1,y_1}p(x_1,y_1)\log\frac{1}{p(x_1|y_1)}+\cdots+\sum_{x_N,y_N}p(x_N,y_N)\log\frac{1}{p(x_N|y_N)}$$

$$=H(X_1|Y_1)+\cdots+H(X_N|Y_N)$$

$$\underline{\underline{\text{时不变}}}NH(X|Y)$$

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

所以：

$$\begin{aligned}I(X^N;Y^N)&=H(X^N)-H(X^N|Y^N)\\&=NH(X)-NH(X|Y)\\&=N[H(X)-H(X|Y)]\\&=NI(X;Y)\end{aligned}$$

小结：**时不变离散无记忆信道**，序列传输与符号传输的关系。

- 对于输入 X^N：　$P(x^N)=\prod_k P(x_k)$　　$H(X^N)=NH(X)$
- 对于输出 Y^N：　$P(y^N)=\prod_l P(y_j)$　　$H(Y^N)=NH(Y)$
- 对于信道的特性：　$P(y^N|x^N)=\prod_n P(y_n|x_n)$　　$H(Y^N|X^N)=NH(Y|X)$
- 对于信道传输的互信息量：　$I(X^N;Y^N)=NI(X;Y)$

4.1 序列传输

XI'AN JIAOTONG UNIVERSITY

讨论信源编码时，假设信道是理想的。根据理想信道模型，写出前向转移概率矩阵。再由对称信道的定义，可以判断理想信道是对称信道。

而对于对称信道，当输入等概分布时，信道传过的互信息量最大，达到信道容量。另一方面，信道理想时可以传过所有输入的信源信息，因此，信源熵最大时理想信道达到其信道容量，而离散信源在等概时的信源熵最大。所以**理想信道在输入等概时达到信道容量**。

因此信源编码的任务是，通过适当编码，**使信源编码后的输出符号或符号序列（即信道的输入）等概或尽量接近等概**。只有这样才能达到充分利用（理想）信道容量的目的。

信源编码追求的目的是快速高效，给定一信源序列，以最快的速度或说最少的次数将其送过信道，因此，从这个角度讲，编码问题可等价为将信源序列编为尽可能短的信道输入序列。

信源编码要求输出尽可能等概、尽可能短。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

定长编码是指给源序列编出的码字长度是固定的。定理是对序列而不是对单个符号而言的。

设信源输出为序列 $u^L=(u_1,u_2,\cdots,u_L)$，每个符号的取值范围为 $u_l\in\{s_1,s_2,\cdots,s_K\}$；

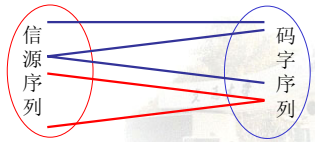
源序列被编成信道输入序列 $x^N=(x_1,x_2,\cdots,x_N)$，每个符号的取值范围 $x_n\in\{a_1,a_2,\cdots,a_D\}$；

可能的源序列个数为 K^L，可能的码序列个数为 D^N。

首先看正确解码（无误传输）的条件。这里无误传输是指信道接收端收到接收码字后能够正确进行解码。

码序列的个数不少于源序列的个数

$$D^N\geq K^L$$



4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

首先讨论无误传输的情况。

要无误传输，必须有 $D^N\geq K^L$

两边取对数，即 $N\log D\geq L\log K$

不等式的左边是 D元（进制）理想信道使用 N次的信道容量。理想信道是对称信道，输入等概时达到信道容量，求得信道容量为 $\log D$。该信道使用 N次的信道容量为 $C=N\log D$。

右边的 $\log K$是 K元随机变量的最大信息熵，所以 $L\log K$是 L 位长的信源序列具有的最大信息熵。输入符号的信息熵为 $H(U)$，L位长的信源序列的信息熵为 $LH(U)$。因此 $L\log K\geq LH(U)$。

所以 $C\geq L\log K\geq LH(U)$。

说明 L位长的信源序列所具有的最大信息量不超过信道容量时，源序列的个数就不会超过码序列的个数，在此情况下就能做到无误传输。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

上述情况只是完成了进制转换，起不到去除冗余的作用。在此情况下，信源编码也就没有太多研究的必要了。因此，我们有必要研究有编码损失的情况，即放弃一些出现概率很小的码的编码。当然，此时无法实现无误解码，但只要我们能保证这个损失足够小就可以了。

当码序列个数小于源序列个数时，需要对源序列进行有选择的编码，对**概率大的典型序列**编码传输，对概率小的、几乎不可能发生的不编码传输，这样就有可能使编码损失比较小（趋于零）。

由4.1节，对离散无记忆信源，L位长序列的信息量是各符号信息量之和。即：

$$I(u^L)\underline{\underline{\text{无记忆}}}\sum_{l=1}^{L}I(u_l)=I(u_1)+I(u_2)+\cdots+I(u_L)$$

序列中每个符号的取值范围为：$u_l\in\{s_1,s_2,\cdots,s_K\}$，设某序列中符号 s_k出现了L_k次，$k=1,2,\cdots K$，则有 $L_1+L_2+\cdots+L_K=L$。

上式可以改写为 $I(u^L)=L_1I(s_1)+L_2I(s_2)+\cdots+L_KI(s_K)$

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

所以：

$$I(u^L)=L\left[\frac{L_1}{L}I(s_1)+\frac{L_2}{L}I(s_2)+\cdots+\frac{L_K}{L}I(s_K)\right]$$

当序列长度 L趋于无穷时，各符号发生的频率 L_k/L趋于稳定，这就是符号发生的概率。因此，

$$I(u^L)\underline{\underline{L\to\infty}}L\cdot\sum_k p(s_k)I(s_k)=LH(U)$$

即有：

$$I(u^L)\to LH(U)\qquad L\to\infty$$

$$\frac{I(u^L)}{L}\to H(U)\qquad L\to\infty$$

可见，当序列长度趋于无穷时，序列每符号的平均信息量趋于符号的熵 $H(U)$。$I(u^L)/L$和 $H(U)$ 之间的关系在数学上可以用Chebyshev不等式精确描述。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

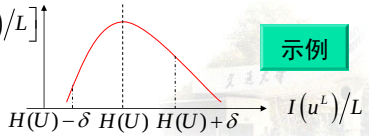
随机变量 $I(u^L)/L$ 的数学期望是 $H(U)$，方差为 σ^2，则由Chebyshev不等式：

$$\forall \delta > 0,\quad P\left[\left|\frac{I(u^L)}{L} - H(U)\right| > \delta\right] \le \frac{\sigma^2}{\delta^2} \triangleq \varepsilon \quad \sigma^2 = D\left[\frac{I(u^L)}{L}\right]$$

意义：序列每符号的平均信息量偏离符号熵 $H(U)$ 超过 δ 的概率不会超过 ε。如果把这部分序列看作是“**小概率**”序列，不对这些序列编码，而只对其余的“**大概率**”序列编码时，编码损失的上界为 ε。

当 $L \to \infty$ 时，$I(u^L)/L \to H(U)$，因此，$I(u^L)/L$ 的方差 σ^2 趋于零。此时，编码损失 $\varepsilon = \sigma^2/\delta^2 \to 0$，也就是编码损失会充分小。

可见编码损失会随着序列长度的增加越来越小，序列无限长时，编码损失会趋于0，即充分小。



示例

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

下面推导只对“大概率”的序列进行编码传输时，信道容量与信源熵之间的关系。

我们将满足 $\left|I(u^L)/L - H(U)\right| \le \delta$ 的序列记为 u_T^L，称为典型序列，而将典型序列的集合记为 T。根据Chebyshev不等式，有：

$$P(T) \ge 1-\varepsilon$$

$$P(T) = P(u_{T_1}^L) + P(u_{T_2}^L) + \cdots + P(u_{T_M}^L)$$

其中，M是典型序列的个数。下面估计M的大小。

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

$$L[H(U)-\delta] \le I(u_T^L) = \log\frac{1}{p(u_T^L)} \le L[H(U)+\delta]$$

$$2^{L[H(U)-\delta]} \le \frac{1}{p(u_T^L)} \le 2^{L[H(U)+\delta]}$$

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

取倒数，有 $2^{-L[H(U)+\delta]} \le P(u_T^L) \le 2^{-L[H(U)-\delta]}$

典型序列的概率有如下关系：

$$M \cdot \min P(u_T^L) \le \sum_T P(u_T^L) < 1$$

因此，

$$M \le \frac{1}{\min p(u_T^L)} = \frac{1}{2^{-L[H(U)+\delta]}} = 2^{L[H(U)+\delta]}$$

所以只要满足 $D^N \ge 2^{L[H(U)+\delta]}$

$$N \log D \ge L\left[H(U)+\delta\right]$$

总可以为每个典型序列找到一个码字。

意义：只要码序列的个数多于典型序列的个数，就能保证对这些典型序列都进行编码传输，从而使得编码损失的上界是非典型序列的概率，即 ε。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

本质上，这是要求信道容量要比信息熵大任意一个正数 δ，即信息熵不能超过信道容量，哪怕就是小一点点，满足这个条件后，总可以通过让源序列趋于无穷长的办法，使编码损失趋于0，即充分小。

与无误传输时的情况进行比较。

$$N \log D \ge L \log K$$

$$N \log D \ge L\left[H(U)+\delta\right]$$

如果信息熵大于信道容量，此时一定会有编码损失，我们没有办法把编码损失做到尽可能的小，即无法让其趋于零。而且可以证明，增加序列长度只能把误码率越做越大，当序列无穷长时，编码损失趋于1。

编码实际上是通过增加每符号携带的信息量从而降低对信道的要求。

以上结论即香农第一定理，总结如下。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

定理4.2.1（定长信源编码定理）：令一个离散无记忆信源的符号熵为 $H(U)$，将 L 位长的信源序列 u^L 编为 N 位长的码字序列 x^N，其中 $u=\{s_1,s_2,\cdots,s_K\}$，$x=\{a_1,a_2,\cdots,a_D\}$。每个码序列只能对应一个源序列，并且令源序列无对应码序列的概率为 P_e，则对于任意的 $\delta>0$，如 $N\log D\geq L[H(U)+\delta]$，则可通过使 L 充分大使 P_e 任意小；反之，若 $N\log D\leq L[H(U)-\delta]$，则当 L 足够大时，$P_e\to 1$。

对编码条件变形可得：$\frac{N}{L}\log D\geq H(U)+\delta$。不等式左边是编码后平均每个信源符号能载荷的最大信息量，通常将其称为编码后信源的信息传输率，并用 R' 表示。编码定理说明，只有编码后信源的信息传输率大于信源的熵，才能实现几乎无失真的编码。

为了衡量各种实际编码方法的性能，定义如下的**编码效率**为：

$$\eta=\frac{H(U)}{R'}=\frac{H(U)}{N/L\bullet\log D}$$

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

因此，最佳等长编码的效率为：

$$\eta=\frac{H(U)}{H(U)+\delta}$$

例4.2.1：二进制信源 $H(U)=1\,\text{bit}$，$D=2$。

如果要求 $N\log D\geq L[H(U)+\delta]$，即 $N\geq L+L\delta,\ \forall\delta>0$

所以，$N>L$。此时只相当于增加了序列长度，没有编码效益。这是由于原信源已经达到最大熵，没有冗余，因此无需进行信源编码。

而如果有冗余 $H(U)<1\,\text{bit}$，此时要求 $N\geq L[H(U)+\delta]$。可以选取 $LH(U)\leq N<L$，从而在编码损失趋于零的同时将序列编为更短的码字序列。

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

例4.2.2：设有一离散无记忆信源，$K=4$ 概率空间为：

$$\begin{pmatrix}U\\P\end{pmatrix}=\begin{pmatrix}u_1 & u_2 & u_3 & u_4\\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8}\end{pmatrix}$$

对其进行近似于无失真的定长编码，若要求其编码效率为95%，差错率低于10^{-6}，试求符号联合编码长度 $L=?$

解：先求该信源的熵：

$$H(U)=-\sum_{i=1}^{4}p_i\log p_i=1.75\,\text{bit}$$

L 位长序列的每符号平均信息量的方差为：

$$D\left[\frac{I(u^L)}{L}\right]=\frac{1}{L^2}D\left[I(u^L)\right]=\frac{1}{L^2}D\left[\sum_l I(u_l)\right]=\frac{1}{L^2}L\bullet D[I(u)]=\frac{1}{L}D[I(u)]$$

$$D[I(u)]=E\left[I^2(u)\right]-E^2[I(u)]=1\times\frac{1}{2}+4\times\frac{1}{4}+2\times 9\times\frac{1}{8}-H^2(U)=\frac{11}{16}\,\text{bit}^2$$

4.2 定长编码

XI'AN JIAOTONG UNIVERSITY

由编码效率：$\frac{H(U)}{H(U)+\delta}=95\%$

解出，

$$\delta=H(U)/19=1.75/19\,\text{bit}$$

编码损失为：

$$P_e=D\left[\frac{I(u^L)}{L}\right]\Big/\delta^2=\frac{11}{16L}\Big/\left(\frac{1.75}{19}\right)^2\leq 10^{-6}$$

解得：$L\geq 8.1\times 10^7$

由此可见，需要8000多万个信源符号联合编码，才能达到上述要求，这显然是不现实的。

4.2 定长编码

定理4.2.1给出了信源编码的理论极限，但也仅具有理论上的意义，无法在实际编码时使用它，因此没有任何工程实用价值。因为，编码损失趋于零的条件是序列的长度趋于无穷，这会带来以下问题。

编码延时，要使编码损失无穷小，我们必须对无穷长的序列进行编码，需要长时间的等待。

存储空间，无穷长的序列需要无穷大的存储空间。

这些问题将通过变长编码的研究解决。

$$C \geq L\log K \geq LH(U) \qquad C \geq L[H(U)+\delta] \quad \forall \delta > 0$$

$$N\log D \geq L\log K \qquad N\log D \geq L[H(U)+\delta]$$

$$D^N \geq K^L \qquad M \leq D^N < K^L$$

$$P_e = 0 \qquad P_e \xrightarrow{L\to\infty} 0$$

4.3 变长编码

定长编码在实际中的可操作问题引出了变长编码问题的研究。

定长编码是要将源序列编成尽可能短的等长信道输入序列，而变长编码追求的是无误传输条件下的平均码长最短。

定义4.3.0（平均码长）：设一离散无记忆信源输出 $u=\{s_1,s_2,\cdots,s_K\}$，各符号概率分别为 $P(s_1),P(s_2),\cdots,P(s_K)$。每个源符号 s_k 用一个码字 x^{n_k} 表示，n_k 表示码字的长度。则该编码的平均码长为

$$\overline{n}=\sum_{k=1}^{K}P(s_k)n_k$$

从定义可以看出，要使平均码长最短，应该是给出现概率越大的符号编越短的码字，而给概率越小的符号编越长的码字才对。

讨论编码必须先讨论其解码问题。定长编码由于所有码字一样长，解码方法较为简单。先通过一个例子讨论变长编码的可解码问题。

4.3 变长编码

编码示例

信源字符	$P(S_k)$	1# 码	2# 码	3# 码	4# 码
S_1	0.5	0	0	0	0
S_2	0.25	0	1	10	01
S_3	0.125	1	00	110	011
S_4	0.125	10	11	111	0111

从解码角度来看，为实现无误传输（即收到码字后能正确的解码，找到对应的源符号），任何编码须满足：

a）所有S_k对应的码字均不相同；

b）任意码字的解码结果不应与其它码字的解码结果相混淆。

显然，1#码不符合a)，2#码不符合b)；而3#码和4#码符合a)、b)，因此都可解。4#码必须整体解，不能把011中的0解为S_1。

4.3 变长编码

定义4.3.1(唯一可解码定义)：如每一个有限长源符序列所对应的码元序列，均不同于任何其它源符序列所对应的码元序列，则称该码是唯一可解的。

这个定义是条件b)的推广，条件b)是看一个码字，而这里是看一长串的序列。3#码和4#码都是唯一可解码。哪个更好？二者的平均码长为：

$$\overline{n}_{3\#}=0.5\times1+0.25\times2+0.125\times3+0.125\times3=1.75$$

$$\overline{n}_{4\#}=0.5\times1+0.25\times2+0.125\times3+0.125\times4=1.875$$

可见，对同一给定信源，唯一可解码可以有多种，但平均码长可能会不同。变长编码追求的是平均码长最短。所以，**对给定信源，我们希望找出其唯一可解码中平均码长最短的那种编码**。

唯一可解码要满足定义4.3.1的要求，但这个定义不够直接，不太好判断究竟是不是唯一可解码，也没有提供编唯一可解码的思路方法。

4.3 变长编码

下面介绍一种异前缀码，它一定满足定义4.3.1，即异前缀码一定是唯一可解码；且其中可以找到平均码长等于唯一可解码最短平均码长的码。**满足唯一可解且平均码长最短的不只是异前缀码**，但它概念直接且很容易编，所以在变长编码中的地位非常重要，是重点研究对象。

定义4.3.2(异前缀码定义)：设一编码的第 k 个码字为 $x_k^{n_k}=(x_{k,1},x_{k,2},\cdots,x_{k,n_k})$，则 $x_k^{n_k}$的前i个码元序列 $x_{k,1},x_{k,2},\cdots,x_{k,i}$，$i\le n_k$，称为 $x_k^{n_k}$的前缀；如一个编码的码字中，任何码字都不是其它码字的前缀，则该码称为异前缀码。

3#码是异前缀码，4#码不是异前缀码。

异前缀码的解码很简单。一旦接收序列中有码字出现，即可进行解码而无需等待后随符号的到来，所以也叫**即刻码**(instant code)。

例如，收到二进制码元序列1100011010，则按3#码可在接收时即刻地唯一解码为 $s_3s_1s_1s_3s_2$。

4.3 变长编码

码树是异前缀码的数学结构，它提供了一种表示异前缀码的编码过程和码字间关系的方便方法，是分析异前缀码的工具。

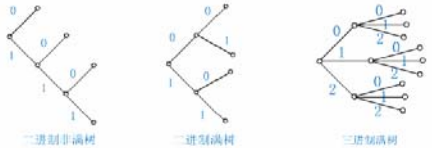
结点：码树的每一个结点代表一个码字，该码字是由从树根爬至该结点所经历的数字。通过码树可以很方便地写出编码的码字。

分叉：码树的分叉表示码元符号的进制，分几个叉就表示几进制。

级数：从码树的根到最远结点所经过的树枝（节）数。

全码树：又称满树，所有可能的分叉点都进行了分叉。

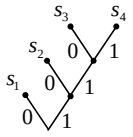
异前缀码和码树一一对应。爬树解码，遇到叶子就解出一个码。



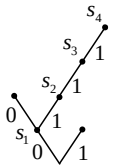
4.3 变长编码

异前缀码与码树的“一一对应”关系：

- 树根 ⟺ 码字起点
- 分叉数 ⟺ 码的进制
- 结点 ⟺ 码字或码字的一部分
- 终止结点 ⟺ 码字
- 节（树枝）数 ⟺ 码长
- 非满树 ⟺ 变长码
- 满树 ⟺ 等长码



3#码



4#码

全码树的结点个数：对于一个N级的D进制码树，共有 $\sum_{n=1}^{N}D^n$个结点。

异前缀码的码树特点：如取任一个结点为码字，则由该结点发出的所有结点所对应的码字不可再用。即如果将某个结点选作码字，则该结点就不能再分叉了。

4.3 变长编码

下面讨论在什么样的条件下能够编出异前缀码，异前缀码的设计方法（如何编），进而讨论变长信源编码定理。为此，需要先证明一个Kraft不等式，这是进行异前缀码的判别及设计的一个重要依据。

定理4.3.1(Kraft不等式)：设一编码的码元为 $x_n\in\{a_1,a_2,\cdots a_D\}$，对应于信源符号 $s_1,s_2,\cdots,s_K$ 的码字长度分别为 $n_1,n_2,\cdots n_K$，如果该码是异前缀码，则必须有 $\sum_{k=1}^{K}D^{-n_k}\le 1$。反之当上式成立时，总可找到一相应的异前缀码。

意义：如果是异前缀码，一定满足不等式；如果满足不等式，则肯定有异前缀码，但未必只有异前缀码。所以只要满足了不等式，就总可设计出一个异前缀码。

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

证明：先证如果是异前缀码，一定满足Kraft不等式。（必要性）

设码树共有N级，则第N层总码枝数为D^N个。

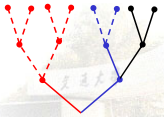
若某个长度为n_k的码枝被选用，则自该枝第n_k节点以后所有码枝不能再选用，第N级共有D^{N-n_k}码枝不能再选用。

由于n_k中k是任意的（$k=1,2,\cdots,K$），所以码树中最高一级不能再选用的总码枝数应为：$\sum_{k=1}^{K} D^{N-n_k}$。

显然其值一定要小于第N层的总码枝数D^N，

即有

$$\sum_{k=1}^{K} D^{N-n_k} \le D^N \Rightarrow \sum_{k=1}^{K} D^{-n_k} \le 1$$



4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

再证当Kraft不等式满足时，总存在异前缀码。（充分性）

从前面证明可知：当一个码字占用了一个n_k级节点时，将占用第N级所有结点的$1/D^{n_k} = D^{-n_k}$，所有码字占用了

$$\sum_{k=1}^{K} D^{-n_k}$$。

由于$n_1, n_2, \cdots n_K$满足Kraft不等式，即

$$\sum_{k=1}^{K} D^{-n_k} \le 1$$

这意味着码树第N级的结点未被用尽（小于），或恰好用尽（等于），因此只要码长分布满足Kraft不等式，就可以构造一个异前缀码，不会出现某个码字找不到结点的情况。 #

Kraft不等式是异前缀码的必要条件，也是构造异前缀码的充分条件。

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

定理4.3.2：令一编码有码字长度$n_1, n_2, \cdots n_K$，且码元符号取值$x_n \in \{a_1, a_2, \cdots a_D\}$。如该码是唯一可解的，则Kraft不等式必成立。

证明

定理4.3.3（变长信源编码定理）：令一有限信源数符$u \in \{s_1, s_2, \cdots, s_K\}$的熵为$H(U)$，则总可找到一异前缀码，使其平均码字长度$\bar{n}$满足

$$\bar{n} < \frac{H(U)}{\log D} + 1$$；

另一方面，对于任何唯一可解编码，必有$\bar{n} \ge \frac{H(U)}{\log D}$。

即 $\frac{H(U)}{\log D} \le \bar{n} < \frac{H(U)}{\log D} + 1$。

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

证明：先证左边（下界）

$$H(U) - \bar{n}\log D = \sum_{k=1}^{K} p(s_k)\log\frac{1}{p(s_k)} - \sum_{k=1}^{K} p(s_k) n_k \log D$$

$$= \sum_{k=1}^{K} p(s_k)\log\frac{D^{-n_k}}{p(s_k)}$$

$$\le \log e \cdot \sum_{k=1}^{K} p(s_k)\left[\frac{D^{-n_k}}{p(s_k)} - 1\right]$$

$$= \log e \cdot \left[\sum_{k=1}^{K} D^{-n_k} - \sum_{k=1}^{K} p(s_k)\right] \le 0$$

所以：$\bar{n} \ge \frac{H(U)}{\log D}$

当$D^{-n_k} = p(s_k)$, $1 \le k \le K$时，上式取等号，$\bar{n}$达最小值。$\bar{n}_{\min} = \frac{H(U)}{\log D}$

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

再证右边（上界）

根据最佳编码的第一条原则，编码应使概率大的信源符号对应短码，而概率小的对应长码。为此来寻求直接由信源符号的先验概率 $P(s_k)$ 确切知道其对应码字长度 n_k 的方法。

将一个概率为 $P(s_k)$ 的信源符号编码为 D 进制码字，其长度 n_k（大于等于1的整数）通常应满足：

$$n_k \geq \log_D \frac{1}{P(s_k)} = -\log_D P(s_k)$$

若 $-\log_D P(s_k)$ 是整数，则 $n_k = -\log_D P(s_k)$ ；

如果不是整数，则向上取整，所以有：

$$-\log_D P(s_k) \leq n_k < -\log_D P(s_k) + 1 \qquad *$$

$$\log_D \left[1/P(s_k)\right] \leq n_k < \log_D \left[D/P(s_k)\right]$$

$$1/P(s_k) \leq D^{n_k} < D/P(s_k)$$

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

取倒数，$P(s_k)/D < D^{-n_k} \leq P(s_k)$ **

由于所有的信源符号都满足上式，对 k 求和，可得

$$\sum_{k=1}^{K} P(s_k)/D < \sum_{k=1}^{K} D^{-n_k} \leq \sum_{k=1}^{K} P(s_k)$$

即：

$$\frac{1}{D} < \sum_{k=1}^{K} D^{-n_k} \leq 1$$

不等式的右半部分即Kraft不等式，这说明用上述方法可以找到一个异前缀码。 下面推导用该方法得到的异前缀码的平均码长的上界。

对*式的右半边用换底公式，可得：$n_k < \log \frac{1}{P(s_k)} \Big/ \log D + 1$

两边同乘以 $P(s_k)$ 并对 k 求和，

$$\sum_{k=1}^{K} P(s_k) n_k < \frac{-\sum_{k=1}^{K} P(s_k) \log P(s_k)}{\log D} + 1 \Rightarrow \bar{n} < \frac{H(U)}{\log D} + 1 \qquad \#$$

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

结论：给定信源符号，如果按 $D^{-n_k} \leq P(s_k) < D^{-n_k+1}$ 来选择码长，它一定满足Kraft不等式，即一定存在异前缀码。

编异前缀码的方法：先根据概率计算各符号的最小码长，然后把码长按大小关系排序，构建码树，随后再从最短的码长开始，在码树上选择结点，依次编码。

扩展编码

对 L 长的离散无记忆信源序列进行异前缀码编码，用 u^L 代替 u，有

$$\frac{LH(U)}{\log D} \leq \bar{n}_L < \frac{LH(U)}{\log D} + 1$$

两边同除以 L，得

$$\frac{H(U)}{\log D} \leq \bar{n} < \frac{H(U)}{\log D} + \frac{1}{L}$$

4.3 变长编码

XI'AN JIAOTONG UNIVERSITY

当 $L \to \infty$ 时，$1/L \to 0$，$\bar{n} \to H(U)/\log D$。

这个结论与定长编码定理的结论实质上是一样的，都是说只要信息量不超过信道容量，就可使编码损失充分小。但二者在实现方法上完全不同。定长编码需要源序列达到无穷长，才能使编码损失充分小；而变长编码可以选择异前缀码，不需要源序列无穷长，容易实现，而且异前缀码对所有的信源符号都有编码，没有编码损失。

定义异前缀码的编码效率为：$\eta = \frac{H(U)}{\bar{n} \log D}$

由 $\frac{H(U)}{\log D} \leq \bar{n}$ 得 $\eta \leq 1$。

当 $P(s_k) = D^{-n_k}$ 时，$\eta = 1$。

如果不满足等号成立的关系，可通过序列编码提高效率。使得

当 $L \to \infty$ 时，$\eta \to 1$。

4.3 变长编码

由，$\frac{H(U)}{\log D} \le \bar{n} < \frac{H(U)}{\log D} + \frac{1}{L}$ ，分别除以 $\bar{n}$，整理可得：

$$1 - \frac{1}{\bar{n}L} < \eta \le 1$$

这说明，当 L 越大时，编码效率越高。

最后，并不是所有的异前缀码的平均码长都是最短的。

以3#码为例，平均码长下限为 $H(U)/\log D = 1.75$ 。

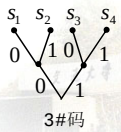
3#码的平均码长为：$\bar{n} = 0.5\times1 + 0.25\times2 + 0.125\times3 + 0.125\times3 = 1.75$

如右图所示的另一种异前缀码，其平均码长为：

$$\bar{n} = 2$$

显然，不是最短的。

3#码即是例4. 2. 2中的信源。



4.4 最佳变长编码技术

最佳，是指该方案给出了唯一可解且具有最小平均长度的编码。

同一信源可能有多种异前缀码，如何编出平均码长最短的异前缀码？下面先讨论二进制编码。

引理4. 4. 1：设源符号 $u \in \{s_1, s_2 \cdots s_k\}$ ，有概率 $P(s_1), P(s_2), \cdots, P(s_K)$ 为便于讨论，设 $P(s_1) \ge P(s_2) \ge \cdots \ge P(s_K)$ 。如果 $K \ge 2$ 且 $D = 2$ ，则最佳异前缀码必须满足可能性最小（即概率最小）的两个源符所编的码字长度相等，且 $n_K = n_{K-1} \ge n_k$， $k = 1,2,\cdots K-2$ 。

结论：概率越小，码长越长，概率最小的要拥有最大码长，而且要有两个最大码长。

证明：先证当最小概率的码不是最大长度时，就不是最佳码。即需证明 $n_K \ge n_k$， $k = 1,2,\cdots K-1$。

用反证法。假设有某个码 n_k 比 n_K 还长，即 $n_K < n_k$ 。

4.4 最佳变长编码技术

由平均码长的定义：

$$\bar{n} = \sum_{k=1}^{K} p(s_k)n_k = p(s_1)n_1 + p(s_2)n_2 + \cdots + p(s_K)n_K$$

将 s_K 的码字与 s_k 的码字互换，形成的新编码的平均码长为：

$$\bar{n}' = \sum_{k=1}^{K} p(s_k)n_k = p(s_1)n_1 + p(s_2)n_2 + \cdots + p(s_k)n_K + p(s_K)n_k$$

平均码长的变化量：

$$\begin{aligned}\Delta n = \bar{n}' - \bar{n} &= p(s_K)n_k + p(s_k)n_K - p(s_K)n_K - p(s_k)n_k \\ &= p(s_K)(n_k - n_K) - p(s_k)(n_k - n_K) \\ &= [p(s_K) - p(s_k)](n_k - n_K) < 0\end{aligned}$$

这说明互换后平均码长变短了，与已知矛盾。所以最佳异前缀码必须有：$n_K \ge n_k$， $k = 1,2,\cdots K-1$，同理：$n_{K-1} \ge n_k$， $k = 1,2,\cdots K-2$ 。

综合两式，可得：$n_K \ge n_{K-1} \ge n_k$， $k = 1,2,\cdots K-2$ 。

4.4 最佳变长编码技术

再证最小概率的两个码长要相等，即 $n_K = n_{K-1}$ ，同样用反证法。

假设 $n_K \ne n_{K-1}$ ，由于 $n_K \ge n_{K-1}$ ，则必有 $n_K > n_{K-1}$。

由于其它码字长度均小于 n_K，这意味着该码字是唯一的 n_K 级结点，该级的其它结点均未被使用，即该码字是一个孤枝。

由于是异前缀码，该码字在 $n_K - 1$ 级上的结点也未被使用。因此，该码字可以使用 $n_K - 1$ 级上的那个结点，即去掉孤枝。

这样就降低了平均码长，与已知条件矛盾。

综上，最佳异前缀编码必须满足：

$$n_K = n_{K-1} \ge n_k, \quad k = 1,2,\cdots K-2 \qquad \#$$

可能的编码方法：

- 当 $K = 2$ 时，根据引理4. 4. 1，取 $n_1 = n_2 = 1$。
- 当 $K > 2$ 时，根据引理4. 4. 1，对概率最小的两个源符，取它们第 n_K

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

个的码元不同，而前n_K-1个码元相同，这样可以保证长度相等且为最大码长。将二者看为一个符号，则可以得到一个新的信源符号集合。

设信源符号集合为 $U=\{s_1,s_2,\cdots,s_{K-2},s_{K-1},s_K\}$，则新的符号集合共有$K-1$种符号，记为 $U'=\{s_1,s_2,\cdots,s_{K-2},s_{K-1}\cup s_K\}$，各符号对应的概率为：

$$p(s_k')=\begin{cases}p(s_k) & k=1,2,\cdots,K-2\\ p(s_{K-1})+p(s_K) & k=K-1\end{cases}$$

对新的符合集合，再次将最小概率的两个符号组合，进一步减小信源符号数，直至只剩两个信源符号(两个源符时，码长均取1)，即可完成编码。

只要我们能证明当对 U'的编码达到最佳时，U的编码也达到最佳，那么我们就可以按上述的递归（迭代）编码方法得到最佳的异前缀编码。

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

引理4.4.2：当U'的编码为最佳时，U的编码同时达到最佳。

证明：由于是异前缀码，U'和U的码长满足：$n_k'=n_k \quad k\le K-2$

U'和U的平均码长之间有如下关系：$n_{K-1}'=n_{K-1}-1=n_K-1$

$$\begin{aligned}\bar{n}&=\sum_{k=1}^{K}p(s_k)n_k=\sum_{k=1}^{K-2}p(s_k)n_k+p(s_{K-1})n_{K-1}+p(s_K)n_K\\&=\sum_{k=1}^{K-2}p(s_k')n_k'+p(s_{K-1})(n_{K-1}'+1)+p(s_K)(n_{K-1}'+1)\\&=\sum_{k=1}^{K-2}p(s_k')n_k'+\left[p(s_{K-1})+p(s_K)\right](n_{K-1}'+1)\\&=\sum_{k=1}^{K-2}p(s_k')n_k'+p(s_{K-1}')(n_{K-1}'+1)\\&=\sum_{k=1}^{K-1}p(s_k')n_k'+p(s_{K-1}')=\bar{n}'+p(s_{K-1}')\end{aligned}$$

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

可以看出，两者只差了一个常数。也就是说，当$\bar{n}'$达到最小时，$\bar{n}$也达到最小。 #

通过这两个引理，就得到了最佳异前缀码的编码方法，即Huffman编码技术。

1）将信源发出的K个符号，按其概率递减顺序进行排列。

2）将概率最小的二个符号分别编码为“1”和“0”，再对这两个符号求概率之和。

3）将上述概率之和作为一新符号的概率，与其余的符号一起组成一新的信源，再按概率递减顺序重新排列。（如果概率之和与原信源的某个或某几个概率相等，则把概率之和排在上面。）

4）如此一直进行下去，直到两个合并消息的概率之和为1。

5）从最后一步开始，沿编码逆过程取各步骤编出的码元符号组成码元符号序列即为对应信源符号的码字。

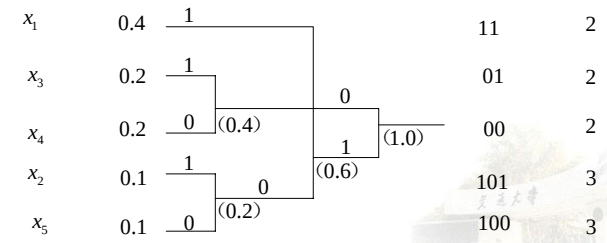
4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

例4.4.1：已知一信源的概率空间为 $\begin{matrix}\mathbf{X}\\P(\mathbf{X})\end{matrix}:\begin{cases}x_1,\ x_2,\ x_3,\ x_4,\ x_5\\0.4,\ 0.1,\ 0.2,\ 0.2,\ 0.1\end{cases}$

对其进行Huffman编码并求编码效率。

解：



4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

由编码结果可求得其平均码长为：

$$\bar{n}=\sum_{k=1}^{K}P(s_k)n_k=0.4\times2+0.2\times2\times2+0.1\times3\times2=2.2$$

其信源熵为：

$$H(\mathbf{X})=-\sum_{k=1}^{K}P(s_k)\log P(s_k)$$

$$=-0.4\log0.4-2\times0.2\log0.2-2\times0.1\log0.1=2.122 \text{ 比特}$$

编码效率：

$$\eta=\frac{H(\mathbf{X})}{\bar{n}\log D}=\frac{2.122}{2.2\log2}=0.964=96.4\%$$

达到平均码长最短的异前缀码也可能会有多种。

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

L次扩展编码，即将K进制L个符号看成一个K^L进制的联合符号，对扩展后的联合符号进行Huffman编码。扩展次数越大，编码效率越高。可以对例4.4.1的信源进行二次扩展编码。

下面讨论 $D>2$ 时的Huffman编码，即多进制Huffman编码。

首先考察异前缀码树的端枝数目（叶子结点）。

引理4.4.3：任一D进制码树的端枝数为 $M=D+m(D-1)$，其中m为分叉次数 $m=0,1,\cdots$ 。

证明：当不分叉即 $m=0$ 时，一级结点有D个。

当任意一个结点分叉生成D个结点时，增加了端枝数 $D-1$个。

依此类推，分叉m次时，总的端枝数为 $M=D+m(D-1)$。 #

多进制编码时的空枝问题。（二进制不存在空枝问题）

例如：当 $K=4, D=3$，一层不够，一次分叉后两层又多出一个。

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

显然分配码字时，空枝只能在最高层，否则平均码长不是最短的。

由于Huffman编码从最高层开始分配码元符号，所以有必要计算最高层的空枝数，据此决定第一次给几个符号编码。

编码时：K、D已知，设空枝数为B，则有 $B+K=D+m(D-1)$ 。

由异前缀条件，$0\le B\le D-2$ 。

所以：$0\le D-2-B\le D-2<D-1$

整理，得 $K=m(D-1)+D-B$

$$K-2=m(D-1)+(D-B-2)$$

可将 $(D-B-2)$ 看作是$(K-2)/(D-1)$的余因子。

即 $(D-B-2)=R_{(D-1)}(K-2)$

所以：$B=D-2-R_{(D-1)}(K-2)$

B只与K、D有关，对给定信源和信道是确定的，且在最高层。

4.4 最佳变长编码技术

XI'AN JIAOTONG UNIVERSITY

多进制Huffman编码的步骤如下：

1）先计算最高层的空枝数 $B=D-2-R_{(D-1)}(K-2)$

2）对概率最小的 $D-B$ 个信源符号编码，合并概率，重排新字符集合的概率。

3）然后，每次对D个概率最小的符号进行编码，直至编码完成。

例4.4.2：一离散信源的概率空间为

$$\begin{bmatrix}\mathbf{X}\\P(\mathbf{X})\end{bmatrix}=\begin{bmatrix}x_1&x_2&x_3&x_4&x_5&x_6\\0.3&0.2&0.15&0.15&0.1&0.1\end{bmatrix}$$

求 $D=3$ 时的Huffman编码。

解：先计算空枝数，

$$B=D-2-R_{(D-1)}(K-2)=3-2-R_2(6-2)=1$$

4.4 最佳变长编码技术

Huffman编码过程如下：

消息	概率	编码过程	三进制代码组	b_i
x_1	0.3	1	1	1
x_2	0.2	2	22	2
x_3	0.15	1 2 (0.5) (1.0)	21	2
x_4	0.15	0	20	2
x_5	0.1	1 0 (0.2)	01	2
x_6	0.1	0	00	2

4.5 其它变长编码方法

4.5.1 费诺（Fano）码

费诺编码属于概率匹配编码。虽然有时可以得到最佳码的性能，但它不是最佳编码。

二元Fano码的编码过程为：

1）将信源符号按概率递减的次序依次排列；

2）将排列好的信源符号划分为两大组，使两个组的概率之和近似相等，然后对各组赋予一个二进制码元“0”和“1”。

3）将每个组的信源符号在分成两个小组，并使每个小组的概率和近似相等，并对各组赋予一个二进制码元“0”和“1”。

4）如此重复，直至每个小组只剩下一个信源符号为止。

5）对每个信源符号，由前向后读取码符号序列作为码字。

对多元码，每次分成相应的多组即可。

4.5 其它变长编码方法

例：离散无记忆信源的概率空间如下，对其进行Fano编码并求平均码长。

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.2 & 0.19 & 0.18 & 0.17 & 0.15 & 0.1 & 0.01 \end{bmatrix}$$

表 5-5 费诺码编码过程

消息符号 a_i	各个消息概率 $p(a_i)$	第一次分组	第二次分组	第三次分组	第四次分组	二元码字	码长 K_i
a_1	0.20	0	0			00	2
a_2	0.19		1	0		010	3
a_3	0.18			1		011	3
a_4	0.17	1	0			10	2
a_5	0.15		1	0		110	3
a_6	0.10			1	0	1110	4
a_7	0.01				1	1111	4

平均码长为： $\bar{n} = 2.72$ 码元/符号

4.5 本章小结

信源编码与信道编码；信源编码追求高效、快速；输入可以分割。

序列传输

序列的概率、信息量、熵；序列通过信道(互信息熵)；等效多元信道

定长编码

无误传输 $N\log D \geq L\log K$；有损编码 $N\log D \geq L[H(U)+\delta]$

变长编码

平均码长、唯一可解码（无误传输）、异前缀码、码树；Kraft不等式；变长编码定理；扩展编码；编码效率。

最佳变长编码技术（Huffman编码）

平均码长最短的异前缀编码技术。二进制编码，最高层不能有孤枝；扩展编码；多进制Huffman编码，空枝数的计算。