《编码理论》课程

北京邮电大学
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

第一章 信息论与编码理论概述

李丽香，彭海朋

北京邮电大学网络空间安全学院

网络与交换技术国家重点实验室






课程群二维码

腾讯会议号75733089688

QQ课程群号码807193496

内容提要

- 数字通信系统、信道模型
- 差错控制方式、译码方法
- 信息、信源编码、信道编码定理

教学参考书

- 差错控制编码. 林舒, 科斯特洛著, 王育民, 王新梅译. 人民邮电出版社, 1986.
- 编码密码学. 杨义先著. 人民邮电出版社, 1992.
- 纠错码与差错控制. 王新梅编著. 人民邮电出版社, 1989.
- 纠错编码入门. 林舒著, 陈太一译.人民邮电出版社, 1976.
- 信息论与最优编码. 章照止, 林须端著. 上海科学技术出版社, 1993.
- 网络编码理论与技术.杨义先编. 国防工业出版社, 2009.
- 信息论、编码与密码学. Ranjan Bose著. 机械工业出版社, 2004.
- Introduction to Coding Theory. J.H. Van Lint. Springer-Verlag, 2003.
- 编码理论基础.陈鲁生, 沈世镒编. 高等教育出版社, 2005.
- 信息论与编码理论. Robert J. Mcelece著. 电子工业出版社, 2004.
- 编码理论.田丽华编. 西安电子科技大学出版社, 2003.
- 信息论与编码.吴伯修, 归绍升, 祝宗泰, 俞槐铨编. 电子工业出版社, 1987.
- 编码理论. 张鸣端, 邹世开编. 北京航空航天大学出版社, 1990.

数字通信系统模型

所有的数字通信系统，如雷达、遥控遥测、数字存储系统、计算机等，都可以归结为如图1所示的模型。

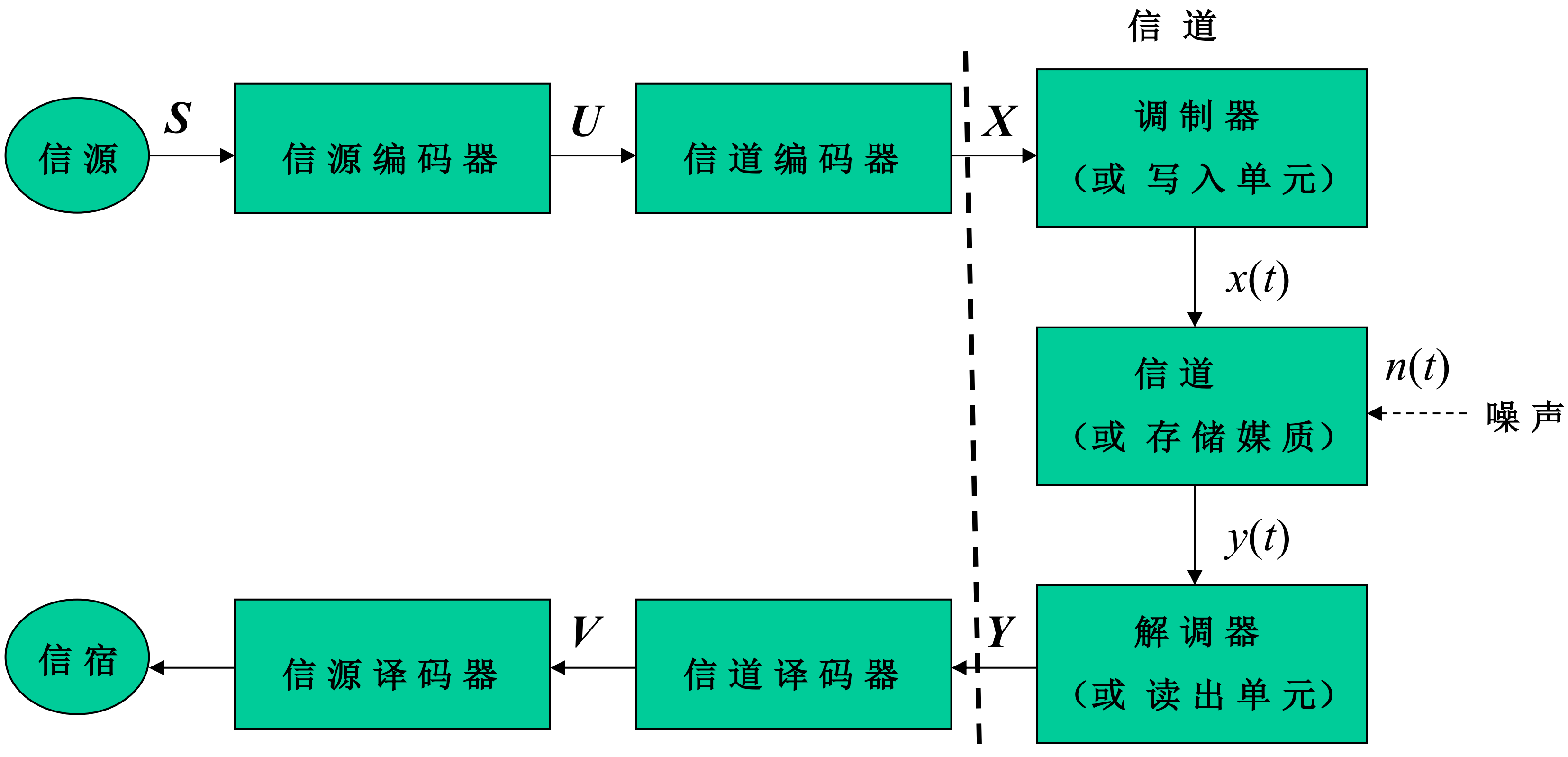


图1 数字通信系统模型

数字传输和存储有共同之处，两者都是将数据从信源传送到目的地。

数字通信系统的简化模型

编码理论关注的是信道编码器和信道译码器。为集中考虑信道编码与信道译码，将数据传输和存储模型简化为图2所示的数字通信系统模型。

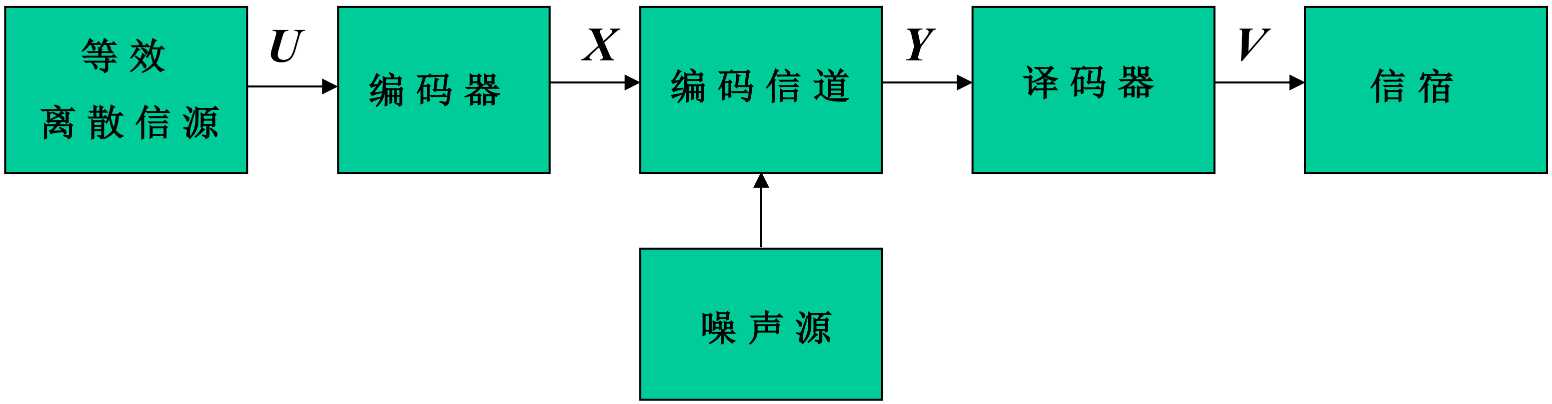


图2 简化的数字通信系统模型

$$\begin{cases} U = [u_1, u_2, \dots, u_K] & u_k \in u = \{a_1, a_2, \dots, a_Q\} \\ V = [v_1, v_2, \dots, v_K] & v_k \in v = \{a_1, a_2, \dots, a_Q\} \end{cases} \qquad \begin{cases} X = [x_1, x_2, \dots, x_N] & x_n \in x \\ Y = [y_1, y_2, \dots, y_N] & y_n \in y \end{cases}$$

- **离散信源：** 即图1中的信源和信源编码器
- **编码信道：** 调制器、实际信道和解调器
- **信宿：** 信源译码器和用户

$$\begin{cases} k = 1, 2, \dots, K \\ n = 1, 2, \dots, N \end{cases}$$

基本概念

- **纠错编码的目的：** 检测和纠正由于非人为(自然)因素引起的传输错误
- 纠错编码主要用于实现信道纠错，又叫做信道编码或者差错控制编码
- **设计通信系统的关键问题之一：** 如何解决可靠性和有效性(快速性)这对矛盾
- **矛盾：** 如要求有效性，则必然使每个数据符号所占的时间缩短、波形变窄、能量减少，这样，受到干扰后，产生错误的可能性就增加，传送消息的可靠性就降低；如要求可靠，则使传送消息的速率变慢

信道模型

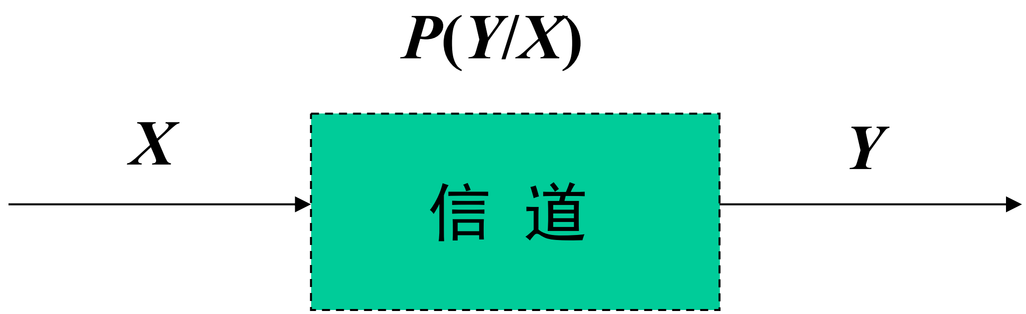


图3 信道模型

条件概率 $P(Y/X)$：似然函数

$$\begin{cases} Y=\left[y_1,y_2,\ldots,y_N\right] & y_n \in y \\ X=\left[x_1,x_2,\ldots,x_N\right] & x_n \in x \end{cases}$$

$$\begin{cases} k=1,2,\ldots,K \\ n=1,2,\ldots,N \end{cases}$$

- **无记忆信道：**如果在给定时间间隔上，检测器的输出只与在该时间间隔上传送的信号有关，而与任何前面时间的传送的信号无关，称此信道为无记忆信道

$$P(Y/X)=\prod_{n} p(y_n/x_n)$$

- **有记忆信道：**一种M元输入、Q元输出的信道模型

$$P\left(y_n \big/ x_{n-K+1},\dots,x_{n-1},x_n\right)$$

- **离散无记忆信道：** 是一种*M*元输入、*Q*元输出的信道模型
- **条件概率：** *P*(*Y*/*X*)是离散无记忆信道的最好描述方式，也叫做转移概率，*X*表示调制器输入符号，*Y*表示解调器输出符号，*P*(*Y*/*X*)表示发送为*X*接收为*Y*的概率

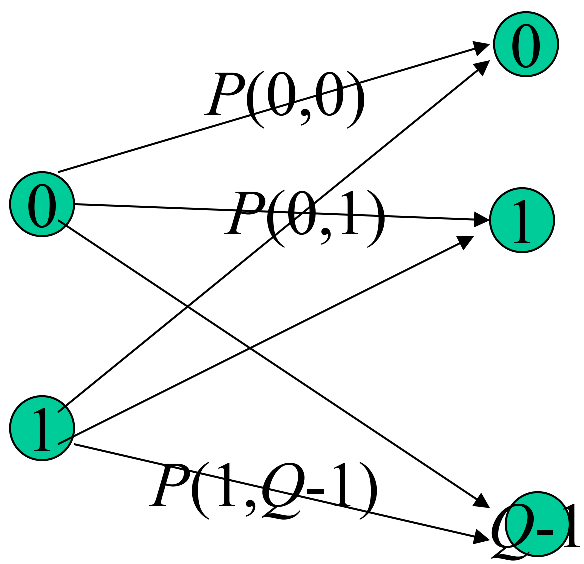
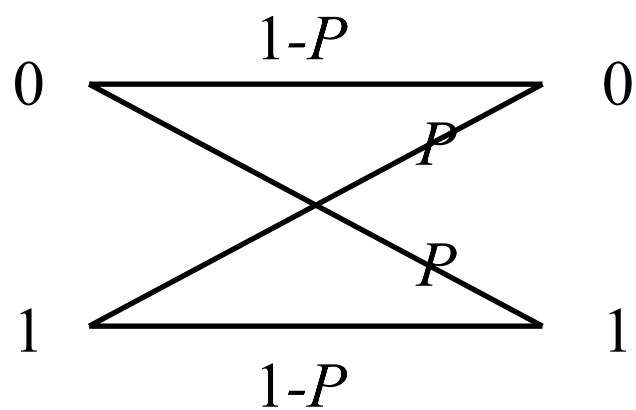


图4 二元输入Q元输出的离散无记忆信道模型

二元对称(BSC)信道

- **当Q=2时，称调制器采用硬判决。多数数字通信系统采用硬判决译码的二元编码，因为它比多元情形容易实现**
- **当Q>2时，称调制器采用软判决**



$P(0/0)=P(1/1)=1-P$

$P(0/1)=P(1/0)=P$

$0 \le P \le 1$

图5 二元对称信道模型

信道矩阵

- **一个信道，也可由它的传递概率组成的矩阵来表示**
- **例如二元BSC信道的信道矩阵为**

$$\begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} \end{array}$$

- **一般离散单符号信道的信道矩阵为**

$$\begin{array}{cc} & \begin{array}{cccc} b_1 & b_2 & \cdots & b_s \end{array} \\ \begin{array}{c} a_1 \\ a_2 \\ \vdots \\ a_r \end{array} & \begin{bmatrix} P(b_1|a_1) & P(b_2|a_1) & \cdots & P(b_s|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & \cdots & P(b_s|a_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(b_1|a_r) & P(b_2|a_r) & \cdots & P(b_s|a_r) \end{bmatrix} \end{array}$$

- **突发错误：**不同状态特性下，错误出现的概率不一样，这种错误为突发错误
- 由于噪声和干扰或读写头接触不良所引起的错误，往往不是随机错误，而是成群成串出现的错误
- 信道状态特性变化较大，进而有突发错误信道的概念。常出现在有记忆信道中，此时噪声对各次传输的影响是彼此相关的
- **突发错误信道：**这种产生比较密集错误的信道称为突发错误信道或者有记忆信道

双状态马尔可夫链模型

- **比较常用的一个突发信道模型：双状态马尔可夫链模型或Gilbert模型**
- **信道的状态有好（G）和坏（B）两种状态**
- **当处于G状态时，信道不产生错误**
- **当处于B状态时，以$1-P_1$的状态产生错误**
- **G和B两个状态之间分别以概率P_1和P_2转移**
- **信道停留在状态G和状态B的概率分别是$1\text{-}P_2$和$1\text{-}P_1$**

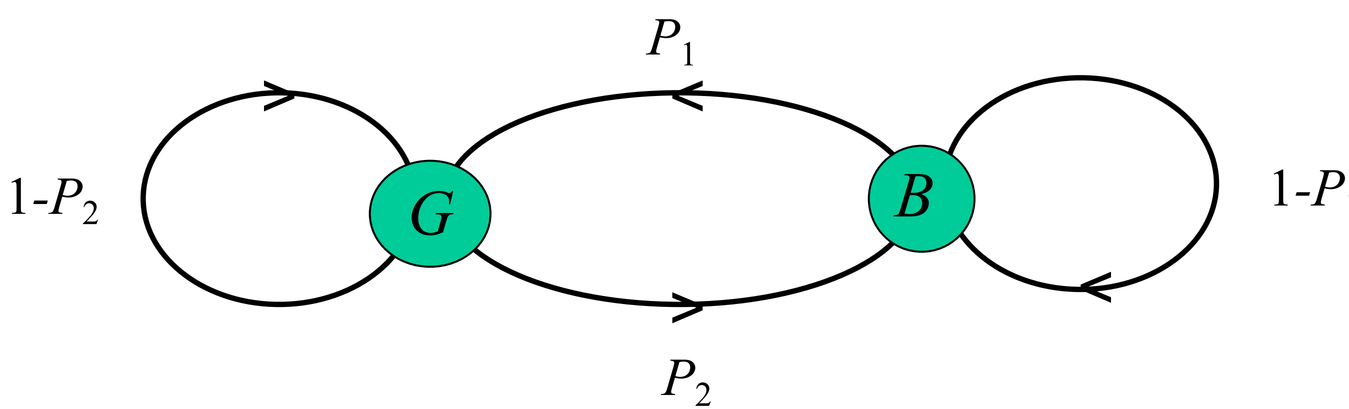


图6 双状态马尔可夫链模型

加性高斯白噪声(AWGN) 信道

- **加性高斯白噪声：Additive White Gaussian Noise**
- **加性噪声：叠加在信号上的一种噪声**
- **白噪声：噪声的功率谱密度在所有的频率上均为一常数**
- **高斯白噪声：白噪声取值的概率分布服从高斯分布，其自相关系数为无延时的冲击函数**

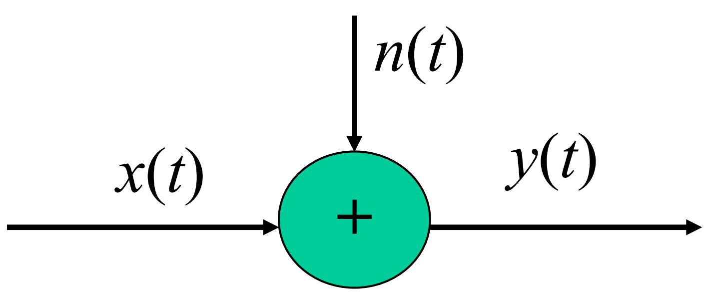


图7 AWGN信道模型

$n(t) \sim$ 高斯白噪声

$$E\{n(t)n(s)\} = \frac{N_0}{2}\delta(t-s)$$

差错控制方式

- **前向纠错(FEC)：**利用纠错码自动的纠正接收端检查出的错误
- 单向传输系统一般只能这样差错控制
- **优点：**不需要反馈信道，译码实时性好，控制电路简单，能进行一个用户对多个用户的同播通信
- **缺点：**译码设备较复杂，对信道适应性差，所选用的纠错码必须与信道的干扰情况相匹配，编码效率低
- **例子：**磁带存储系统(记录在磁带上的信息很久以后再重读)、深空通信系统(飞行体上的编码设备很简单，地面的译码设备可以很强大)、军事通信

- **当接收端检查到错误时，就自动要求发送端重新传送该消息**
- **优点：易于实现，成本和复杂性低**
- **缺点：必须有反馈信道，实现控制较复杂，难以用于同播系统，通信效率低，很难适合实时传输系统**
- **双向传输系统可使用这种方法进行差错控制**

- **ARQ包括：等待式ARQ和连续式ARQ**
  - **等待式ARQ：发送端发送一个码字，并等待接收的一个肯定回执（ACK）、或者否定回执（NAK）。收到ACK则发送下一个码字，否则重发**
  - **双向传输系统可使用这种方法进行差错控制**
- **连续式ARQ：有选择重传和退*N*步ARQ**
  - **退*N*步ARQ：若检测到有错的码字，重传该码字及其后续码字**
  - **有选择重传：只是重传那些有否定回执的码字**

混合差错控制(HEC)

- **是FEC和ARQ方式的结合，具有FEC和ARQ方式的优点**
- **发送端：发送有纠错和检错能力的码**
- **接收端：检查错误情况，如果错误在该码的纠错能力范围内，则自动进行纠正；如果信道干扰很严重，错误超过纠错能力，但是能检测出来，则经反馈信道请求发送端重发**
- **适用于环路时延大的高速传输系统中，如卫星通信**

信息反馈方式(IRQ)

- **接收端：** 把接收到数据，原封不动的通过反馈信道送回发送端
- **发送端：** 比较发送数据和反馈数据，如发现错误，则重新发送出错的消息，直到没有发信错误为止
- **优点：** 不需要检、纠错、编、译码器，控制设备和检错设备较简单
- **缺点：** 需要反馈信道，环路时延较大，发送端需要存储器存储已经发送的码组
- 仅适用于传输效率较低，数据信道错误率较低，有双向传输信道和控制简单的系统中

信息传输率

- **符号传输率：**为$1/T$，即每T秒传输一个编码符号
- **信息传输率（数据率）：**为R/T(bit/s)，在编码系统中，若码速率是$R=k/n$，即k个信息比特对应于传送n个符号，则信息传输率（数据率）为R/T

信道带宽

除了噪声的作用造成信号改变外，所有的通信系统都会因为带宽有限而造成信号失真。

- **最小信道带宽：** 粗略估计等于1/(2T)(HZ)，为基本保证信号不因带宽原因而失真
- **未编码系统：** 数据率1/T=2W，受到带宽影响
- **二元编码系统：** 数据率R/T=2RW。若要保持数据率不变，则要求带宽扩展1/R倍

AWGN信道的分组编码系统

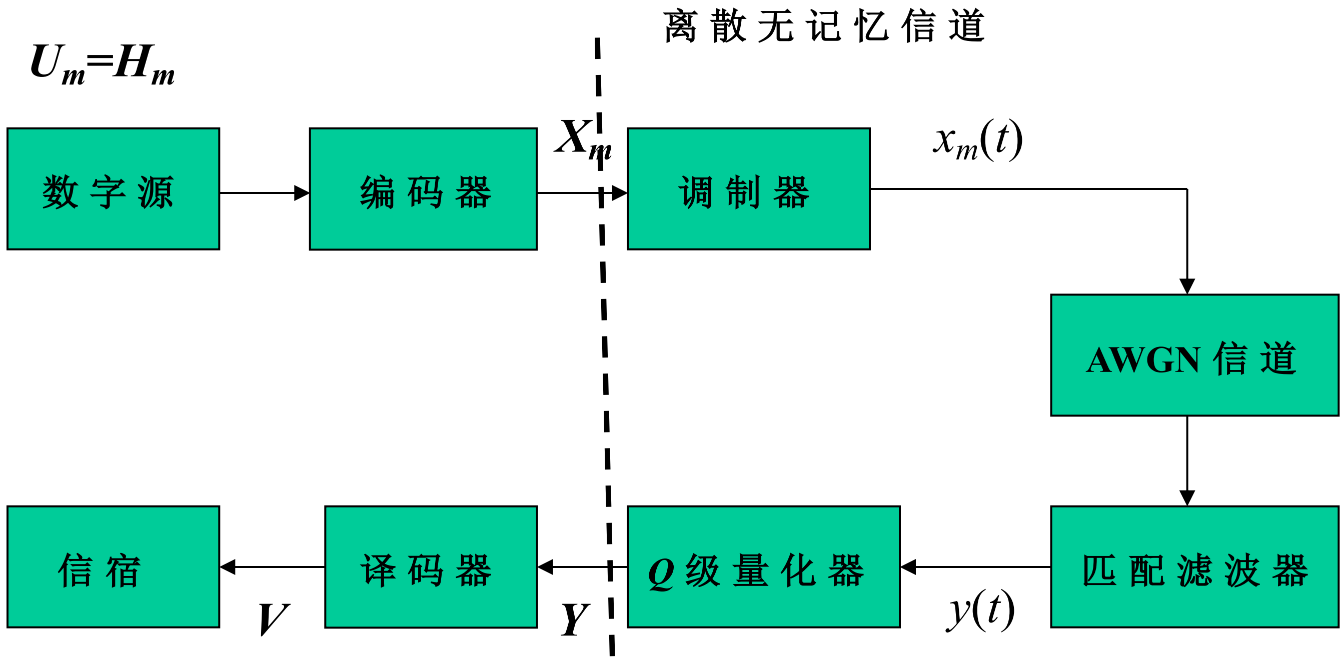


图8 AWGN信道的分组编码系统模型

$$m = 1,2,\ldots,M \quad M = Q^K$$

K ~ 分组长度

$$u_m = \left[u_{m1}, u_{m2}, \ldots, u_{mK}\right]$$

$\boldsymbol{H_m}$ ~ 第m个消息

译码器的基本任务

- **译码器的条件错误概率：已知接收序列*Y*时**

$$P(E/Y) \cong P(H_{\hat{m}} \neq H_m / Y) \cong P(Y \neq X / Y)$$

简记为：$P_e(H_m/Y), P_e(X_m/Y)$

- **译码器的错误概率：**

$$P_e(H_m) = P_e(\mathbf{X}_m) = \sum_{\mathbf{Y}} P_e(\mathbf{X}_m / \mathbf{Y}) P(\mathbf{Y})$$

- **最佳译码规则：** 能够使译码错误概率$P(E)$达到最小的译码规则
- **$P(Y)$与译码规则无关：** 由于接收序列Y是译码前产生的
- 最佳译码规则必须对所有Y，使得 $P\left(E/Y\right)\cong P\left(Y\neq X/Y\right)$ 最小，即 $P\left(Y=X/Y\right)$ 最大

- **最大后验概率译码：**对于每个输入Y，如果译码器能在码字集合中选择一个码字，作为发送码字的估值，并且使$P(X/Y)$最大，则这种译码规则一定能使译码器输出的错误概率最小

似然函数

- **Bayes公式：** $$P(X/Y)=\frac{P(Y/X)P(X)}{P(Y)}$$ 公式（1-1）
- **对于给定的*Y*，可选择能使公式(1-1)右边最大的向量作为*X*的估计值**
- **如果所有码字都是等可能的，即*P*(*X*)是常数，那么公式(1-1)左边最大，就推导出*P*(*Y*/X)最大**
- ***P*(*Y*/X)：叫做似然函数、信道转移概率**

最大似然译码

- **定义：**若能在码字集合选择合适的码字X，使得P(Y/X)最大，则这种译码规则被称为最大似然译码
- **MLD：**对应的译码器称为最大似然译码器
- **对数似然函数：**由于ln$x$与x是单调关系，故有ln$P(Y/X)$
- 对于离散无记忆信道，由于 $P(Y/X)=\prod_{n} p(y_i/x_i)$ 所以 $\ln P(Y/X)=\sum_{n}\ln p(y_i/x_i)$

最大似然译码(MLD)规则

- **MLD规则是最为可行的一种译码规则**
- **对于DMC和BSC信道，MLD规则是使译码错误概率最小的一种最佳译码准则**
- **在某些情况下，MLD规则并不是最佳的译码规则，如当发送端不是以等概率发送码字的时候**

BSC信道的最大似然译码

$$p(y_i / x_i) = \begin{cases} 1-p & y_i = x_i \\ p & y_i \neq x_i \end{cases}$$

- **Y为二元序列，用$d(Y, X)$表示Y和X之间的距离，即不同位数的个数**

$$\ln P(Y/X) = d(Y,X)\ln p + (n - d(Y,X))\ln(1-p)$$

$$= d(Y,X)\ln\frac{p}{1-p} + n\ln(1-p)$$

$$p < \frac{1}{2}, \quad \ln\frac{p}{1-p} < 0$$

即想要$P(Y/X)$最大，则需要$d(Y, X)$最小

- **最小距离译码器：在BSC中，MLD规则变成了选择能使Y和X之间的汉明距离为最小的向量，作为码字X的估计值**

仙农的观点

百度为您找到相关结果约410,000个　　搜索工具

他贡献大于爱因斯坦,玩股票能赢过巴菲特

2016年5月2日 除了参与协助美国军方在飞弹射控技术与密码学方面的研究,香农也被称为"信息理论之父"。有人将香农的贡献与爱因斯坦相提并论,但也有人认为这样的比喻还是委屈了香农。怎么说呢?因为...

个人图书馆　百度快照

曾经的世界3大科学家,为何其中一个没有另外2个出名? - 天...

2019年11月6日 香农在他一生成就完全值得同大众喜爱的爱因斯坦和图灵比较。正如科学作家威廉·庞德斯通所说,"在贝尔实验室和麻省理工学院有很多人拿香农的洞察力爱和爱因斯坦的作比较。其他人发现...

天文在线　百度快照

码农鼻祖-克劳德·艾尔伍德·香农,一个顽皮的科学巨匠,信...

2018年1月11日 有人认为认为香农对人类的贡献绝不亚于图灵,属于一个开创一个学科的伟人。 他对人类的贡献完全可以和牛顿、爱因斯坦比肩,比特(bit)这个单位,就是他发明的,用二进制存储信息也是他的...

搜狐网　百度快照

仙农的主要贡献 - 百度知道

2016年5月31日 回答: 仙农的主要贡献是创立了经典信息论。他在贝尔电话实验室从数学上和技术上研究"通信"、"信息"、"消息"等概念,其顶点则是1948年在《贝尔系统技术杂志》上发表"通信的数学...

百度知道　百度快照

1948年,香(仙)农在一篇具有历史意义的论文指出并证明了:对信息进行适当的编码,在不牺牲信息传输和存储速率的情况下,可以将有扰信道或存储介质引入的差错减到任意低的程度

- **仙农编码定理：如果系统的传输率小于信道容量，那么适当选择编码技术就能实现可靠通信，即可以将差错率减小到任意小的程度。**更确切地，每个信道都具有固定的信道容量C，对任何小于C的信息传输率R，存在一个码长为n码率为R的分组码，若用最大似然译码，则其译码错误概率为$Pe \le \Lambda e^{-nE_b(R)}$。对于码率为$R$约束长度为$n_e$的卷积码，其译码错误概率也有类似的关系，$Pe \le Be^{-n_eE_e(R)}$即，其中$A$和$B$都为大于0的数，$E_b(R)$和$E_e(R)$为正实函数，叫做误差指数。

科学哲学：过程比结果更重要

仙农编码定理解决了纠错码的存在性问题，虽然仙农当时未能构造出一个纠错码，但是从此之后纠错编码的研究与应用开始了前所未有的发展，人们一直在不停的努力解决构造性问题。为了接近信息传输率的上限，已经提出了很多的纠错编码技术。

我们从仙农(1916-2001)生平故事中得到一个启示：**富有创造性的富有成果的生活也可以是快乐的生活**

北京邮电大学 Beijing University of Posts and Telecommunications

信道纠错码的应用

- **通信系统**：如卫星、有线和无线的电话通信、军事通信等，利用纠错码来实现可靠通信和敌人的恶意干扰
- **计算机系统**：如计算机存储器、数字磁带、磁盘、光盘、数字逻辑电路中
- **商业领域**：如条形码，在运输、仓储、超级市场管理等物流行业获得了广泛的应用

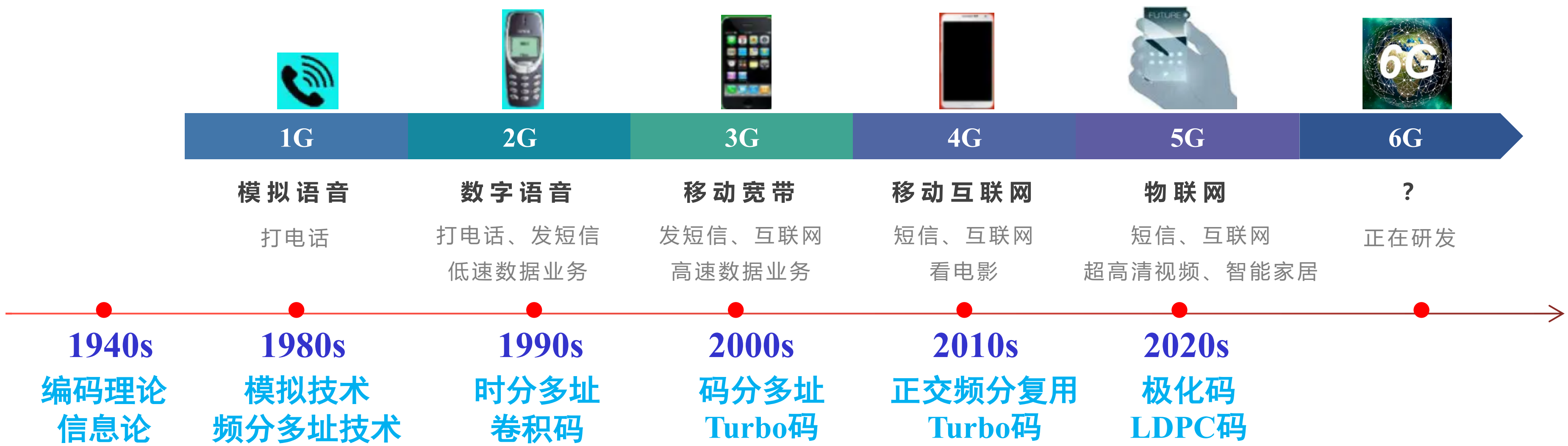


图9 1G-5G中使用的编码方法

信息的概念

- **1948年，仙农发表了一篇重要的论文《通信的数学理论》，研究通信系统的实质，对信息作了科学的定义，并进行了定量和定性的描述**
- **先请读几个句子：**
  - 明天，太阳从东方升起
  - 电话铃声将会在一个小时之内响
  - 这个冬天，哈尔滨将下雨
- **一个句子中所含信息的多少，同句子中所表达的事件出现的概率有关：呈现相反的关系**
- **仙农对于信息的定义：信息是事物运动状态或存在方式不确定性的描述**

自信息的概念

- **自信息**：考虑离散随机变量X，其样本空间为$\{x_i, i=1, 2,\ldots,n\}$，则事件$X=x_i$的自信息的定义为

$$I(x_i)=\log\left(\frac{1}{P(x_i)}\right)=-\log P(x_i)$$

- 自信息的**单位**由对数的底来决定，以2为底，单位就是比特(bits)，以e为底就是奈特(nats)
- 自信息**非负**

例子：考虑一个可以同时输出2个比特的离散无记忆的信源C。该信源有4种可能的输出{00，01，10，11}，其中每一种可能的输出都是独立不相关的，其出现的概率为$P(C)=0.25$。因此，信源C的每个可能的输出的自信息为

$I(C)=-\log_2 P(x_i)=-\log_2 0.25=2\text{bits}$

- **条件自信息量**：在给定$Y=y_i$的情况下，考虑离散随机变量X，其样本空间为$\{x_i,\ i=1,\ 2,\dots,n\}$，则事件$X=x_i$的条件自信息量的定义为

$$I(x_i|y_i)=\log\left(\frac{1}{P(x_i|y_i)}\right)=-\log P(x_i|y_i)$$

- 其中$P(x_i|y_i)$是x_i的条件概率，也称为后验概率
- $P(x_i)$称为先验概率
- 条件自信息量**非负**

- **互信息：** x_i和y_i之间的互信息定义为

$$I(x_i;y_i)=\log\left(\frac{P(x_i|y_i)}{P(x_i)}\right)$$

- **互信息的性质：**
  - **性质1：** 互易性， $I(x_i;y_i)=I(y_i;x_i)$
  - **性质2：** 当x_i和y_i是统计上独立的，即$P(x_i|y_i)=P(x_i)$，则$I(x_i;y_i)=0$
  - **性质3：** 互信息量可以是正的，也可以是负的

思考题： 计算二元对称(BSC)信道的互自信息量?

平均自信息量(信息熵)的概念

- **定义：** 离散随机变量X的平均自信息定义为，其样本空间为$\{x_i, i=1, 2,\ldots,n\}$，则事件$X=x_i$的平均自信息量的定义为

$$H(X)=E\left[\log\left(\frac{1}{P(x_i)}\right)\right]=\sum_{i=1}^{n}P(x_i)I(x_i)=-\sum_{i=1}^{n}P(x_i)\log P(x_i)$$

- H(X)表示每个信源符号的平均信息量
- **举例：** 变量X，$P(X=x_1)=0.99$, $P(X=x_2)=0.01$, 变量Y，$P(Y=y_1)=0.5$, $P(Y=y_2)=0.5$, 则信息熵为
- $H(X)=-0.99\log0.99-0.01\log0.01=0.08$(比特/符号)
- $H(Y)=-0.5\log0.5-0.5\log0.5=1$(比特/符号)

平均条件自信息量(条件熵)的概念

定义：平均条件自信息$H(X|Y)$定义为

$$H(X|Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_iy_i)\log\frac{1}{P(x_i|y_j)}$$

证明：

$$H(X|Y)=E[H(X|y_i)]=\sum_{j=1}^{m}P(y_i)H(X|y_i)$$

$$=\sum_{j=1}^{m}P(y_i)\sum_{i=1}^{n}P(x_i|y_j)\log\frac{1}{P(x_i|y_j)}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_iy_i)\log\frac{1}{P(x_i|y_j)}=\sum_{X,Y}P(xy)\log\frac{1}{P(x|y)}$$

思考题：计算二元对称(BSC)信道的平均条件自信息量(条件熵)?

平均互信息量的概念

思考： $H(X)$：接收到输出符号以前，关于输入变量X的平均不确定性。$H(X|Y)$：接收到输出符号以后，关于输入变量X的平均不确定性。可见，通过信道传输，消除了一些不确定性，获得了一定的信息。

- **定义：** X和Y之间的平均互信息$I(X;Y)$定义为 $I(X;Y)=H(X)-H(X|Y)$
- $I(X;Y)$代表接收到输出符号以后，平均每个符号获得关于输入变量X的信息量，还可表示输入和输出2个随机变量之间的统计约束程度

定理：一个熵为$H(X)$的离散无记忆信源，若对信源长为N的符号序列进行等长编码，设码字是从r个字母的码符号集合中，选取l个码元组成。对于任意的ε>0，只要满足

$$\frac{l}{N} \geq \frac{H(X)+\varepsilon}{\log r}$$

则当N足够大时，可实现几乎无失真编码，即译码错误概率可为任意小。

例子：

表1 译码不唯一的变长码例子

字母	码字	字母	码字
A	0	E	10
B	1	F	11
C	00	G	000
D	01	H	111

若对字母序列ABADCAB进行编码，则得到0 1 0 01 00 0 1

译码不唯一：可能被译为DADAAAB(01 0 01 0 0 0 1)或者AEDGB(0 10 01 000 1)

注：变长码必须是惟一可译码,才能实现无失真编码

定理：一个熵为$H(X)$的离散无记忆信源，并有r个码元的码符号集合，总可以找到一种无失真编码方法，构成惟一可译码，使其平均码长满足

$$\frac{H(X)}{\log r} \le \overline{L} < \frac{H(X)}{\log r} + 1$$

则当N足够大时，可实现几乎无失真编码，即译码错误概率可为任意小。

- 变长信源编码定理中的极限值$H(X)/\log r$同等长信源编码定理中的极限值是一致的
- **定义：**信道的信息传输率（码率）为

$$R = \frac{H(X)}{\log r}$$

- 因为 $\overline{L} \geq H(X)/\log r$，所以得到编码后信道的信息传输率为 $R \leq \overline{L}$

信源编码—霍夫曼(Huffman)编码

- 一个离散无记忆信源有7个符号x_i, i=1,..,7，$P(X=x_1)=0.37$, $P(X=x_2)=0.33$, $P(X=x_3)=0.16$, $P(X=x_4)=0.07$, $P(X=x_5)=0.04$, $P(X=x_6)=0.02$, $P(X=x_7)=0.01$, 将7个信源符号按照概率递减的顺序进行排序构造如图10所示的霍夫曼树
- 按照编码路径从后往前返回，就得到对应的码字x_7=111111, x_6=111110, x_5=11110, x_4=1110
- x_3=110, x_2=10
- x_1=0

思考题：如何编程序实现霍夫曼编码?

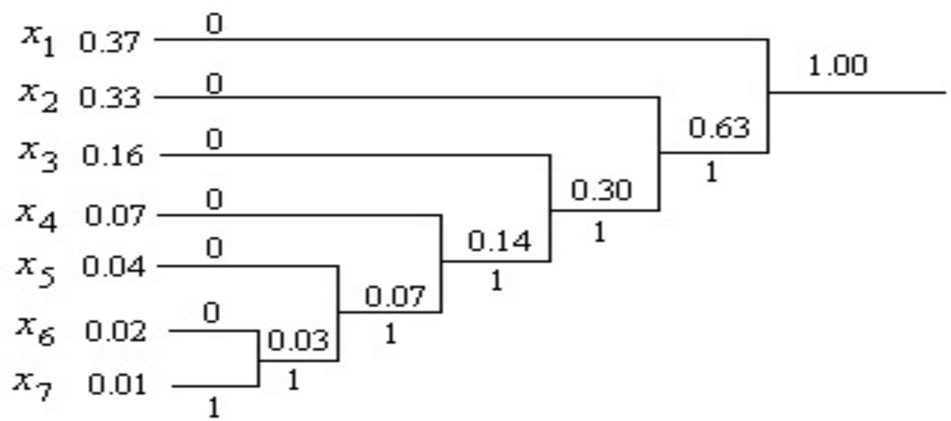


图10 霍夫曼树

信源编码—游程编码

- **游程：** 游程指的是信源输出的字符序列中，各种字符连续的重复出现的字符串的个数
- **游程编码：** 就是将这种字符序列映射成字符串的长度和字符串的位置的标志序列
- **举例：** 考虑比特序列111111111111111000000000000000000001
- 111，可以被表示成(15,1),(19,0),(4,1)，字符最长的重复的数目为19，因此，把该比特序列编码为(01111,1), (10011,0), (00100,1)，此时压缩率为18:38=1:2.11
- 游程编码常用作图形文件的编码方式，如.bmp和.tiff

思考题： 如何编程序实现游程编码?

信源编码—LZ(Lemple-Ziv)编码

- **分段的方法为：** (1)游程先取第一个符号作为第一段，然后再继续分段(2)若有出现与前面符号一样时，就再添加紧跟后面的一个符号一起组成一段(3)尽可能取最少个连着的符号并保证各段都不相同(4)以此类推，直至信源符号序列结束
- **编码方法为：** 首先去掉最后一个符号，然后看剩下的字符串在字典中的排序，这个排序值转换成二进制数作为指针*K*的值，最后一个信源符号作为码字第2项*d*的值，即得到码字(*X*, *d*)
- **举例：** 考虑比特序列10101101101010101011，根据上面的编码方法可把该比特序列分段为1, 0, 10, 11, 01, 101, 010, 1011
- **码字为：** (000,1),(000,0), (001,0), (001,1), (010,1), (011,1), (101,0), (110,1)

思考题： 如何编程序实现LZ编码?

定义： 考虑某种概率分布为$P(x)$的离散无记忆信源，对于一个固定的信道，信道容量被定义为最大的平均互信息，此时传输每个符号平均获得的信息量最大，即对于每个固定的信道可以达到最大的信息传输率，即

$$C=\max_{P(x_j)} I(X;Y)=\max_{P(x_j)}\sum_{j=1}^{q-1}\sum_{i=0}^{r-1}P(x_j)P(y_i|x_j)\log\frac{P(y_i|x_j)}{P(y_i)}$$

其约束条件为 $P(x_j)\geq 0,\sum_{j=0}^{q-1}P(x_j)=1$

思考题： 如何计算给定某种信道的信道容量?

信道编码定理

定理：假设DMS有信源字符集X，熵为每信源符号$H(X)$比特，而且信源每T_s秒产生一个符号，那么信源的平均信息率为每秒$H(X)/T_s$比特，假设信道可以每T_c秒使用一次，而信道容量为每次信道使用C比特，那么每单位时间的信道容量为每秒钟C/T_c比特。如果$H(X)/T_s \le C/T_c$，那么就存在编码方案使得在有噪声的信道上传输的信源消息，能够以任意小的错误概率进行恢复。

举例：考虑BSC信道，信道容量$C=1+p\log_2 P+(1-p)\log_2(1-p)$，如果转移概率$p=10^{-2}$，则可计算出$C=0.919$，那么就可以知道肯定存在信息传输率为$r \le 0.919$的编码方案，使得信息可以在有噪声的信道上传输的信源消息，能够以任意小的错误概率进行恢复。

谢谢聆听
敬请批评指正

图1模型各部分的功能

- **信源：** 信源可以是人或机器，信源发出的消息可以是语言、文字、图像、声音、传感器输出的数据等，信源的输出可以是连续波形，也可以是离散的符号序列

- **信源编码器：** 将信源输出变成二元数字(bit)序列，称为信息序列，在信源连续的情况下，还需要进行模/数(A/D)转换。理想信源编码器模型要满足(1)为表示信源输出所要求的单位时间的比特数要尽量小；(2)信源的输出S可从信息序列U中确切的重新构造

图1模型各部分的功能

- **信道编码器：** 将信息序列U变换成离散的有结构的编码序列X，这称为码字。即为了使传输有效，人为的增加一些冗余度，使其具有自动检错和纠错的能力。码字的结构主要用以对付传输或存储码字的有扰信道，码字的设计和实现是本课程的主题
- **调制器(或写入单元)：** 离散符号不适合于在实际信道上传输或记录在数字存储媒质上。调制器将信道编码器的每个输出的离散符号，通过调制变成适合传输(或存储)的持续时间为T的波形，此波型进入信道(或存储媒质)，并受噪声干扰

图1模型各部分的功能

- **典型的传输信道：** 有线信道、无线信道、电话线路、高频无线线路、遥测线路、微波线路、卫星线路、光纤信道、磁记录信道、大气光信道、水声信道等
- **典型的存储煤质：** 磁芯和半导体存储器、磁带、磁鼓、磁盘、光存储器、光盘等
- **典型的干扰：** 开关脉冲噪声、热噪声、串音、闪电、磁涂层缺损、光盘划痕等

图1模型各部分的功能

- **解调器(或读出单元）：** 处理接收到的每个持续时间为T的波形，并产生一个可能是离散的（量化的）或连续的（未量化的）输出。对应于编码序列X的解调器的输出序列Y称为接收序列
- **最佳解调器：** 总是含有一个匹配滤波器或者相关检测器后面再跟一个取样开关。例如，对于有相干检测的BPSK调制，其取样输出是一个实数

图1模型各部分的功能

- **信道译码器：** 将接收序列Y变换成二元序列V，称为估值序列。在理想的情况下，V与信息序列U完全一致，但是噪声会造成译码错误。译码方法根据信道编码规则和信道（或存储媒质）的噪声特性而定。但设计和实现译码错误概率最小的信道译码器也是本课程的主要论题之一
  - **模拟译码器：** 未量化的译码器输出序列可以直接送到信道译码器去处理。在这种情况下，信道译码器必须能够处理模拟输入
  - **数字译码器：** 将连续的检测器输出的实数量化为有限个离散输出符号集中的一个元素，然后再送到信道译码器。几乎所有的通信系统采用这种方式

图1模型各部分的功能

- **信源译码器：** 把估值序列*V*变成信源输出的估值(原来消息的估值)，并将此估值传送给用户。如果信源是连续的，需要进行数/模转换。在一个精心设计的系统中，除非信道(或存储媒质)的干扰太强，否则这个估值将是信源输出的准确重现
- **信宿：** 目的地或者用户

二元相移键控调制

- 在二元码的情况下，调制器必须产生两个信号中的一个
- $s_0(t)$对应编码0，$s_1(t)$对应编码1
- 对于宽带信道，最佳的信号选择为：

$$s_0(t)=\sqrt{\frac{2E}{T}}\sin\left(2\pi f_0 t+\frac{\pi}{2}\right),\quad 0\le t\le T$$

$$s_1(t)=\sqrt{\frac{2E}{T}}\sin\left(2\pi f_0 t-\frac{\pi}{2}\right),\quad 0\le t\le T$$

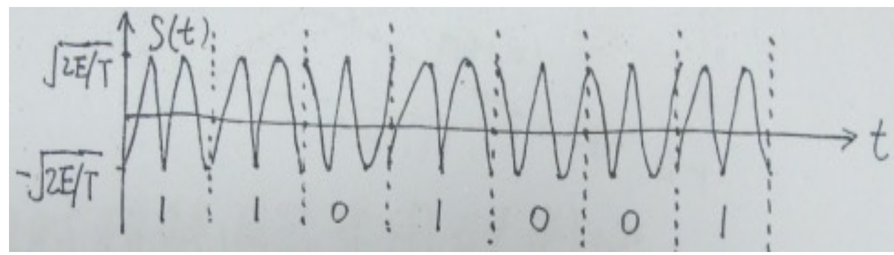


图11 二元相移键控调制的波形

- f_0是$1/T$的倍数, E是信号的能量
- 传送的信号是正弦脉冲，其相位由编码器输出决定，故称二元相移键控（BPSK）调制

判决方法

因为存在噪声或干扰，信号在信道中传输，经常会发生错误，所以接收端必须进行判决以确定发送端发送的是什么码元：

- **硬判决：** 当用二元码时，调制器仅有二元输入。类似的，解调器输出采用二元量化时，译码器只有二元输入，这种情况下，解调器采用硬判决
- **软判决：** 如解调器输出未量化或者量化门限大于2，则解调器采用软判决
- **删除信号：** 对没有把握做出正确判决的信号，就暂时搁置起来不做判决，并用“×”表示，称为删除符号

二元删除信道

当发现某位码元可信度不够，难以准确判定时，暂时将该位用符号“×”标出，分别将该位置按0和1进行最大似然译码，选择与接收序列最相似的码字作为发送码字

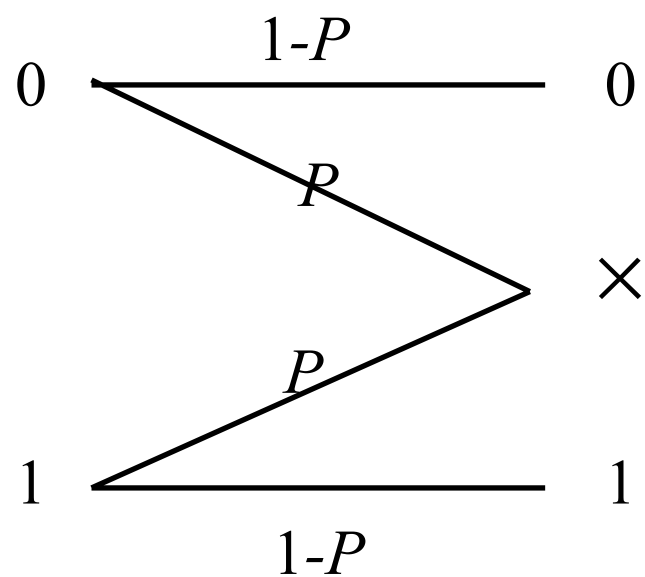


图12 二元删除信道模型

兼具二元对称信道和二元删除信道的性质

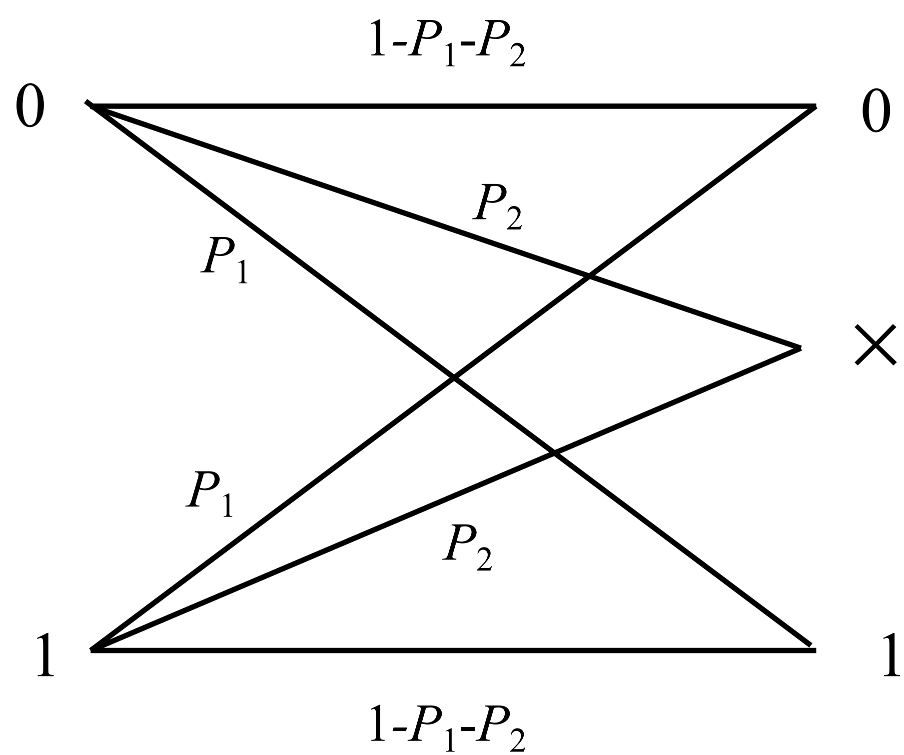


图13 二元混合信道模型

二元Z信道

- **单向错误：二元Z信道中，一种码元永远不会出错，另一种码元则可能出错**
- **大规模集成电路存储器、磁盘、磁带等缺陷所属的信道类型**

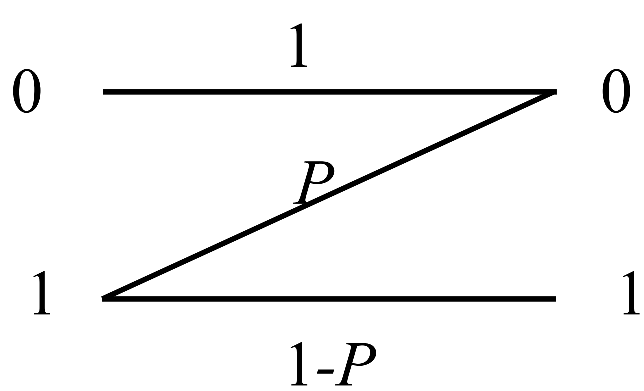


图14 二元Z信道模型

$$x_m = \left[x_{m1}, x_{m2}, \dots, x_{mN} \right]$$ **码字**

$$x_m(t) = \sum_{n=1}^{N} x_{mn} \varphi_n(t)$$

$\left\{ \varphi_n(t) \right\}$ **是时间段[0,*T*]上的完全正交集**

$$\int_0^T \varphi_n(t) \varphi_{n'}(t) dt = \delta_{nn'} \qquad n, n' = 1, 2, \dots, N$$

$$x_{mn} = \int_0^T x_m(t) \varphi_n(t) dt$$

二进制反相信号

$$x_1(t) = -x_2(t) = \sqrt{\frac{2\varepsilon_s}{T_s}} \cos 2\pi f_0 t \qquad 0 \le t \le T_s$$

T_s是信号长度

$$\varphi_1(t) = \sqrt{\frac{2}{T_s}} \cos 2\pi f_0 t \qquad N = 1, \quad M = 2$$

多相信号

$$x_m(t)=\sqrt{\frac{2\varepsilon_s}{T_s}}\cos\left[2\pi f_0 t+\frac{2\pi}{M}(m-1)\right] \qquad N=2$$

$$=\sqrt{\varepsilon_s}\cos\frac{2\pi}{M}(m-1)\sqrt{\frac{2}{T_s}}\cos 2\pi f_0 t-\sqrt{\varepsilon_s}\sin\frac{2\pi}{M}(m-1)\sqrt{\frac{2}{T_s}}\sin 2\pi f_0 t$$

$$\varphi_1(t)=\sqrt{\frac{2}{T_s}}\cos 2\pi f_0 t,\quad \varphi_2(t)=\sqrt{\frac{2}{T_s}}\sin 2\pi f_0 t$$

正交振幅(QAM)调制

$$x_m(t)=\sqrt{\varepsilon_s}A_{cm}\underbrace{\sqrt{\frac{2}{T_s}}\cos 2\pi f_0 t}_{\varphi_1(t)}-\sqrt{\varepsilon_s}A_{sm}\underbrace{\sqrt{\frac{2}{T_s}}\sin 2\pi f_0 t}_{\varphi_2(t)}$$

信道纠错码的分类

- **对消息序列处理方式：**分组码、卷积码
- **校验元与信息元之间的关系：**线性码、非线性码
- **纠正错误的类型：**纠随机错误码、纠突发错误码、纠同步错误码
- **码元的取值：**二进制码、多进制码
- **码字的循环结构：**循环码和非循环码
- **构造码的数学方法：**代数码、几何码、算术码
- **码字信息组的结构：**系统码和非系统码

随机错误

- **随机错误：**接收序列中的传输错误是随机出现的
- **随机错误信道：**这样的信道称为随机错误信道
- 常见于无记忆信道中，因为噪声随机独立的影响每个传输信号
- 深空通信和卫星通信信道都是典型的随机错误信道
- **纠随机错误码：**为纠正随机错误而设计的码称为纠随机错误码

突发错误

- **突发错误的例子：** 无线通信(由多径传输引起的信号衰落造成的错误)，有线和电缆传输(开关脉冲和串话的影响)，磁带记录(涂层缺损和灰尘引起的带脱落)
- **纠突发错误码：** 为纠正突发错误而设计的码称为纠突发错误码
- **组合错误信道：** 两种错误都包含的信道
- **纠突发和随机错误码：** 为纠正这类信道错误而设计的码称为纠突发和随机错误码

正交振幅(QAM)调制

- **调制的原理：**发送数据在编码器内被分成两路，各为原来信号的1/2，然后分别与一对正交调制分量相乘，求和后输出。接收端完成相反过程，正交解调出两个相反码流，均衡器补偿由信道引起的失真，判决器识别复数信号并映射回原来的二进制信号
- **幅度、相位联合调制：**在最小距离相同的条件下可实现更高的频带利用率
- 样点数目越多，传输效率越高。QAM最高已达到1024-QAM(1024个样点)，
- **举例：**具有16个样点的16-QAM信号，每个样点表示一种矢量状态，即有16态，规定了16种载波和相位的组合。由于每4位二进制数规定了16态中的一态，所以16-QAM的每个符号和周期传送4比特

汉明的编码方法

- **发送端：将输入数据每4bits分为一组，计算这些信息比特的线性组合，得到3个校验比特，之后将得到的7bits送入计算机**
- **接收端：按照一定的规则读取这些码字，采用一定的算法，不仅能够检测到是否有错误发生，而且能找到发生单个比特错误的位置**
- **缺点：编码效率较低，每4bits编码就需要3bits冗余校验信息；每个码组中只能纠正1bit的错误**

- **M. Golay研究了汉明码的缺点，并提出了两个以他自己名字命名的高性能码字**
- **二元Golay码：将信息比特每12个分为一组，编码生成11个冗余校验比特，相应的译码算法可以纠正3个错误**
- **三元Golay码：其操作对象是三元而非二元数字，将每6个三元符号分为一组，编码生成5个冗余校验三元符号，这样，由11个三元符号组成的三元Golay码码字可以纠正2个错误**

- **在Golay码提出之后，最主要的一类分组码就是R-M码**
- **首先由Muller于1954年提出，此后Reed在Muller的基础上得到一种新的分组码，称作R-M码**
- **1969年−1977年，R-M码在火星探测方面得到了广泛的引用**
- **现在，R-M码也具有很大的研究价值，其快速译码算法非常适合于光纤通信系统**

循环码

- 在R-M码提出之后，循环码被提出来
- 也是分组码，但码字具有循环移位特性
- 也称为循环冗余校验码，特点如下：
  - 循环结构使码字设计范围大大增加，大大简化了编译码结构
  - 可用一个幂次为n-k的生成多项式来表示
  - 可用Meggitt译码器来实现译码
  - 由于该译码器的复杂性随纠错能力t的增加而呈指数形式的增加，所以一般的，循环码用于纠正只有单个错误的应用情况，常作校验码，而非纠错码

- 循环码一个非常重要的子集就是分别由霍昆格姆Hocquenghem在1959年，博斯Bose和查德胡里Ray-Chaudhuri研究组在1960年，几乎同时提出的能够纠正多个随机错误的BCH码
- 是迄今为止发现的一类很好的线性纠错码，纠错能力很强，特别是在较短和中等码长下，性能接近于理论限，具有严格的代数结构，构造方便，编/译码简单，编/译码设备也不太复杂
- 目前用的最为广泛的码类之一

分组码的缺陷

- 虽然分组码的理论分析和数学描述已非常成熟，在实际通信系统中也已得到广泛的应用，但其固有的缺陷大大限制了它的进一步发展：
  - 面向数据块的，数据块较大时，引入的系统时延是非常大的
  - 要求精确的帧同步，即需要对接收码字或帧的起始符号时间和相位精确同步
  - 大多数基于代数的分组编码的译码算法都是硬判决算法，而不是对解调器输出未量化信息的软译码，从而造成了一定程度的增益损失

- 可以改善分组码的固有缺陷
- 1995年由P. Elias等提出
- 同分组码的区别：
  - 编/译码过程充分利用了各个信息块之间的相关性
  - 其编/译码过程都是连续进行的，可获得相对比较小的编/译码延时
  - 在系统条件相同的条件下，要达到相同的译码性能，卷积码的信息块和码字长度较小，所以译码的复杂性也小些

卷积码的译码算法

- **1961年, Wozencraft和Reiffen提出了序贯译码算法，Fano和Jelinek分别在1963年和1969年改进了该算法，目前统称为Fano算法，该算法是基于码字树图结构的一种次优概率译码算法**
- **1963年，Massey提出门限译码算法，该算法利用码字的代数结构进行代数译码**
- **1967年，Viterbi提出Viterbi算法，该算法是基于码字格图结构的一种最大似然译码算法，是一种最优译码算法**
- **之后，卷积码在通信系统中得到了广泛的应用，如GSM、3G、商业卫星通信系统等**

编码理论的发展历史

- 近年来，在信道编码定理的指引下，人们一直致力于寻找能满足现代通信业务要求、结构简单、性能优越的好码
- 在分组码、卷积码等基本编码方法和最大似然译码算法的基础上提出了许多好码及简化译码复杂性的方法
- 提出了乘积码、代数几何码、低密度校验码(LDPC)、分组-卷积级联码等编码方法
- 提出了逐组最佳译码、软判决译码等译码方法及编码与调制相结合的网格编码调制(TCM)技术
- 其中对纠错码发展贡献较大的有级联码、软判决译码和TCM技术等

- **1966年，Forney将分组码和卷积码结合起来，首先提出级联码，其研究表明，级联码的性能具有明显的改善，但其解码复杂度却没有明显的增加**
- **在随后的应用中，多采用卷积码作为内码，RS码作为外码，内外码之间加入交织器**
- **在1993年Turbo码出现之前，级联码是AWGN信道下性能最好的编码方法**

- **1982年，Ungerboeck将编码和调制技术有机的结合起来，提出带限信道下的网格编码调制(TCM)的思想，可带来3-4dB的编码调制增益**
- **虽然软判决译码、级联码和编码调制技术都对信道编码的设计和发展产生了重大的影响，但是其增益与仙农理论极限始终都存在2-3dB的差距**

- **20世纪90年代以后，以迭代译码为基础的高效纠错码成为主要研究对象**
- **人们从一直研究以代数为基础的代数码上，转向寻找新的纠错码的认识方式**
- **其中有以Tanner为基础发展起来的编译码的可视化方法，如因子图，以及基于图中双边上信息传递的和积算法，还有用于评估码字性能限的数值化方法，如密度进化、典型几何界理论等**

- 最大似然译码性能最好，但复杂度随码长指数增加，因此必须研究新的编译码方案，期望在性能和复杂度之间取得平衡
- 如串行级联码、乘积码等编码方案的目标都是为了构造出具有较大等效分组长度的纠错码，并允许将最大似然译码分为几个较简单的译码步骤，这是一种次优但可实现的译码策略
- 另一种次优方法是迭代译码，如Turbo码、LDPC码等
- 总之，为了实现高效纠错，不是采用级联方式构造随机长码，就是采用迭代译码，或者两者都采用，以逼近仙农限

- **1993年，在瑞士日内瓦召开的国际通信会议上，两位任教于法国不列颠通信大学的教授C. Berrou, A. Glavieux和他们的缅甸籍博士生P. Thitimajshima首次提出了Turbo码**
- **Turbo码很好的应用了仙农信道编码定理中的随机性编/译码条件，从而获得了几乎接近仙农理论极限的译码性能**
- **在采用长度为65536的随机交织器并译码迭代18次的情况下，在信噪比0.7dB并采用二元相移键控(BPSK)调制时，码率为1/2的Turbo码在加性高斯白噪声信道上的错误比特率小于等于10^{-5}，达到了与仙农极限仅相差0.7dB的优异性能**

LDPC码

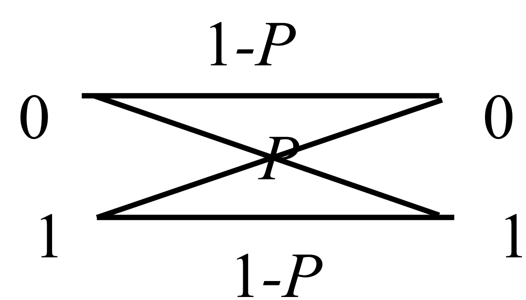
- 1962年，Gallager在其博士学位论文中提出了低密度奇偶校验(LDPC)码, 给出了该码的构造方法、迭代解码算法及其性能在理论上的证明。由于当时计算机发展水平有限，该码硬件实现上困难，被遗忘了30多年
- 1995年，Mackay和Neal重新发现LDPC码和Turbo码有着相似的性能，而且在某些方面已经超过了Turbo码，随后， LDPC码得到广泛的关注，成为新的研究热点
- 目前，LDPC码已经得到广泛的应用，如CMMB(中国移动多媒体广播)、WiMAX(无线城域网)、Wi-Fi(无线局域网)、第二代DVB(数字视频广播)等

- 1997年，Host, Johannesson, Ablov提出了编织卷积码(WCC)的概念，随后，该码得到了发展，它是一种组合码，其系统结构可完全包容传统分组码、卷积码及各类Turbo码，从而开创了编码领域的一个新天地
- 编织码的结构综合了并行级联卷积码(Turbo码)和串行级联卷积码的结构特点，当外编码器个数足够多时，该码型完全拥有了仙农编码定理中随机长码的特性，因此其纠错性能理论上要比Turbo码优异
- 但编织码的编码结构复杂性较高，编码效率也不高
- 目前研究的最多的是1/3编织卷积码，译码采用BCJR算法的迭代译码

互自信息的计算

例子：考虑二元对称(BSC)信道

$$P(Y=0)=P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(X=1)$$
$$=0.5(1-p)+0.5p=0.5$$
$$P(Y=1)=P(Y=1|X=0)P(X=0)+P(Y=1|X=1)P(X=1)$$
$$=0.5p+0.5(1-p)=0.5$$
$$I(x_0;y_1)=\log_2\left(\frac{P(y_1|x_0)}{P(y_1)}\right)=\log_2\left(\frac{p}{0.5}\right)=\log_2 2p$$



$$p=0.1,$$

$$I(x_0;y_1)=I(0;1)=\log_2 2p=\log_2 0.2=-2.322bits$$

信息熵的物理含义

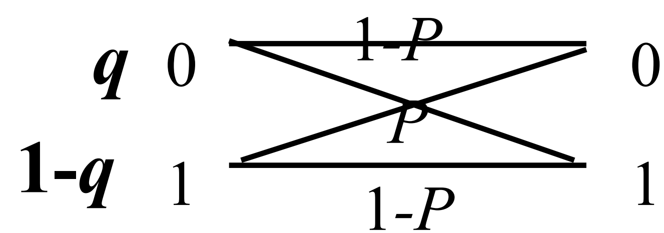
- $H(X)$表示信源输出后，每个消息(或符号)提供的平均信息量
- $H(X)$表示信源输出前，信源的平均不确定性
- 例如：第38页例子中， $H(Y)>H(X)$，对于信源X，两个输出消息不是等概率的，事先大致可以猜测消息x_1会出现，故信源X的不确定性要小
- $H(X)$表示变量X的随机性
- 当信源符号是等概率出现的时候，信息熵可以达到最大值

平均条件自信息量(条件熵)的计算

例子：考虑二元对称(BSC)信道

$$P(Y=0)=P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(X=1)$$
$$=q(1-p)+(1-q)p=p+q-2pq$$

$$P(Y=1)=P(Y=1|X=0)P(X=0)+P(Y=1|X=1)P(X=1)$$
$$=qp+(1-q)(1-p)=2pq-p-q+1$$



$$P(x_i,y_j)=P(y_j|x_i)P(x_i)$$

$$H(X|Y)=\sum_{i=1}^{2}\sum_{j=1}^{2}P(x_i,y_j)\log\left(\frac{P(x_i,y_j)}{P(x_i)P(y_j)}\right)$$

平均互信息量的公式

$$I(X;Y)=\sum_{X}P(x)\log\frac{1}{P(x)}-\sum_{X,Y}P(xy)\log\frac{1}{P(x|y)}$$

$$=\sum_{X,Y}P(xy)\log\frac{1}{P(x)}-\sum_{X,Y}P(xy)\log\frac{1}{P(x|y)}$$

$$=\sum_{X,Y}P(xy)\log\frac{P(x|y)}{P(x)}=\sum_{X,Y}P(xy)\log\frac{P(xy)}{P(x)P(y)}$$

$$=\sum_{X,Y}P(xy)\log\frac{P(y|x)}{P(y)}$$

$$I(X;Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)}$$

等长码

- **定义：** 一组码中，所有码字的码长都相同
- **举例：** 若要用比特位来表示26个英文字母，观察到$2^5=32>26$，即每个英文字母需要5位二元符号编码才可以

- **定义：** 若码C中，没有任何完整的码字是其他码字的前缀，称此码为前缀码，也称即时码或非延长码
- **前缀码和即时码的定义是一致的：** 如果没有一个码字是其他码字的前缀，则在译码过程中，当收到一个完整码字的码符号序列时，就能直接把它译成对应的信源符号，无需等待下一个信号到达后才作判断，这就是即时码

注：前缀码是惟一可译码的一类子码，即前缀码一定是惟一可译码，但是惟一可译码不一定是前缀码

变长信源编码定理

- 信源的熵越高，则码字的平均长度越大
- 码字的平均长度不能小于极限值$H(X)/\log r$，否则惟一可译码不存在
- 这个极限值$H(X)/\log r$同等长信源编码定理中的极限值是一致的
- 该定理给出了平均码长的上界，但并不是说大于这上界就不能构成惟一可译码，而是希望平均码长尽可能短

变长码的编码效率

- **定义：编码效率为平均码长与极限值*H*(*X*)/log*r*之比**

$$\eta = \frac{\dfrac{H(X)}{\log r}}{\overline{L}} = \frac{H(X)}{\overline{L}\log r}$$

- **编码效率一定是小于或等于1的一个数**
- **信源编码主要用作图像或者数据的压缩**
- **对同一信源来说，若码的平均码长越小，编码效率越高**

变长码编码效率的计算

- 有一个离散无记忆信源，$P(X=x_1)=3/4$, $P(X=x_2)=1/4$，其熵为$H(X)=-(3/4)\log(3/4)-(1/4)\log(1/4)=0.811$(比特/信源符号)
- 下面用二元码 (0,1)来构造一个即时码，该码平均长度为
- 在二元无噪无损信道中$r=2$, $\log r=1$
- 编码效率为

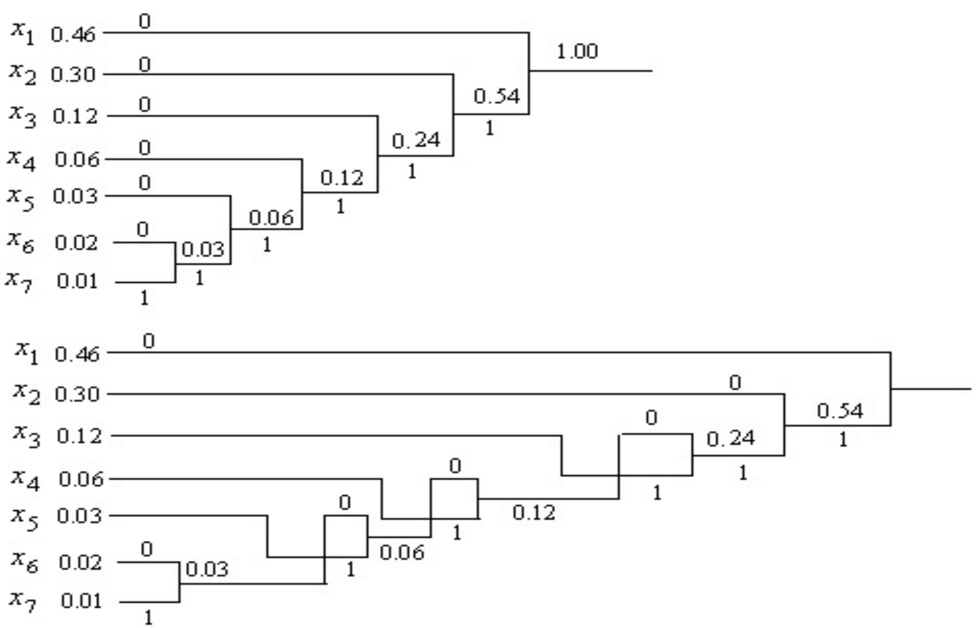
$$\eta=\frac{H(X)}{\bar{L}\log r}=0.811$$

霍夫曼编码的步骤

- 将信源符号按照概率递减的顺序进行排序
- 将0和1符号分别分配给概率最小的两个信源符号，并将这两个概率最小的符号合并成一个新符号，用这两个信源符号的概率之和作为这个新符号的概率
- 以此类推继续这个过程，直到只剩下2个符号为止，从而完成霍夫曼树的构造
- 从树的最后一个节点，依编码路径从后往前返回，读出每个分支上对应的符号标示，即可得到对应的码字

霍夫曼编码不唯一

例子：一个离散无记忆信源有7个符号x_i, i=1,..,7，$P(X=x_1)=0.46$, $P(X=x_2)=0.30$, $P(X=x_3)=0.12$, $P(X=x_4)=0.06$, $P(X=x_5)=0.03$, $P(X=x_6)=0.02$, $P(X=x_7)=0.01$, 将7个信源符号按照概率递减的顺序进行排序构造如下图所示的2种霍夫曼树

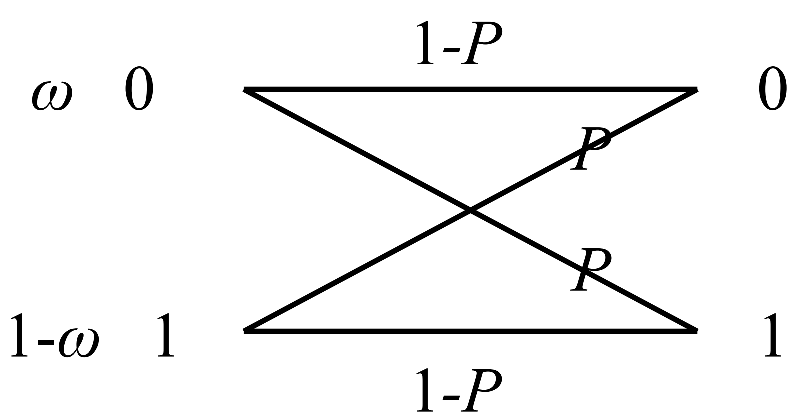


按照编码路径从后往前返回，就得到对应的码字(1)x_7=111111, x_6=111110, x_5=11110, x_4=1110, x_3=110, x_2=10, x_1=0; (2) x_7=110001, x_6=110000, x_5=11001, x_4=1101, x_3=111, x_2=10, x_1=0

- **定义：** 设离散无记忆信源变量为$X=\{x_1,x_2,\ldots,x_N\}$，信源符号通过信道传到接收端，接收端的接收变量为$Y=\{y_1,y_2,\ldots,y_N\}$，对于每一对(x,y)，指定一个非负的函数$d(x_i,y_i)\geq 0$，成为单个符号的失真度或失真函数
- **方差失真度：** $d(x_i,y_i)=(x_i-y_i)^2$
- 一般的，失真度还可定义为： $d(x_i,y_i)=|x_i-y_i|^p$
- 序列的失真度

$$d\left(X,Y\right)=\sum_{i=1}^{N}d\left(x_i,y_i\right)$$

举例： 考虑一个BSC信道，其信道转移概率为$P(0|1)=P(1|0)=p$，如果$\omega=0.5$，用信道容量的公式，可以获得BSC的信道容量为$C=1+p\log_2 p+(1-p)\log_2(1-p)$，熵函数为$H(p)=-p\log_2 p-(1-p)\log_2(1-p)$，因此得到$C=1-H(p)$



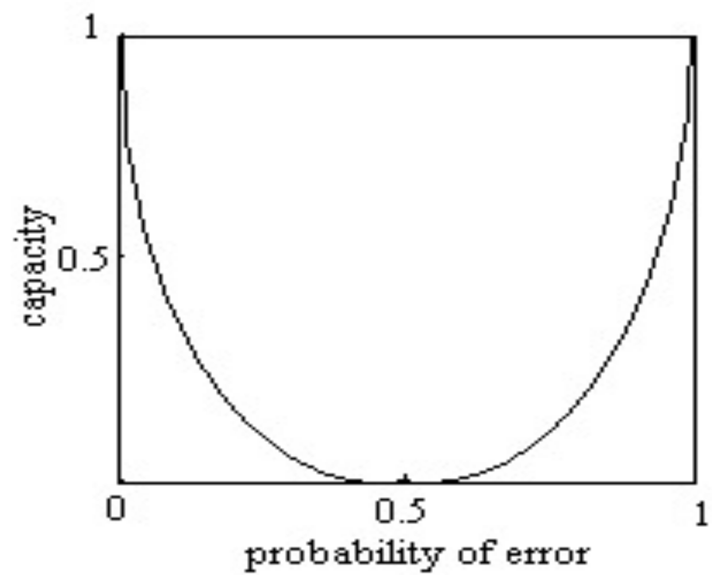


表2 不同应用领域信道编码误差概率的例子

应用领域	误差概率
语音电话	10^{-4}
声乐数据	10^{-6}
电子邮件	10^{-6}
因特网访问	10^{-6}
可视电话，高速计算	10^{-7}

为了信息传输能获得这么高的可靠性，就需要设计信道编码，使得数字通信系统有更高的抗噪能力，即，在信息流中增加冗余来实现的

欢迎大家积极参与爱课堂“交流互动”论坛讨论：