

# 通信的数学理论

克劳德·香农著<sup>1</sup>

近年来的多种调制方法，例如 PCM（脉冲编码调制）和 PPM（脉冲相位调制），它们都是通过带宽和信噪比之间的交换，增加了人们对通信普遍理论的兴趣。在奈奎斯特和哈特莱有关这方面的重要文献奠定了该理论。在本文中，我们将推广该理论，使它含有一些新的因素，特别是信道中噪声的影响，和利用原始消息的统计结构和最终受信者的性质来改善通信的可能性。

通信的基本问题是在一端精确地或者近似地复现另一端选择的消息，通常这些消息是有意义的。那就是说它们按照某一系统与特定的物质或概念的实体相互联系。通信的语义方面与工程问题是没有关系的，重要的方面是一个实际消息是从一组可能的消息集里面选择出来的，系统必须被设计成对所有可能的选择都能工作，而不是只适合工作于某一种选择，因为在设计时这是不知道的。

如果集合中消息的数目是有限的，则这个数目或这个数目的单调函数能被用来作为当一个消息被选出时所产生信息的度量，所有选择都是等概率的，正如哈特莱指出的，最自然的选择是取对数函数。肃然当我们考虑到消息统计特性的影响和当我们有一组连续的消息，这一定义必须大大的推广。但是我们在所有的情况下采用本质的对数度量。

对数度量更方便是因为有以下几个原因：

1. 实用性。工程上的重要参量，如时间，带宽，中继器的数目等，都趋于随可能数目的对数关系作线性变化。例如，在一组中继器中增加一个中继器则可能的状态就增加 1 倍。这个数目以 2 为底的对数加 1，时间加倍使得消息的数目成平方增加或是数目对数的 2 倍。

2. 相对于合适的度量，对数更直观。这与（1）密切相关，因为我们用与普通标准进行线性比较的方法来直观地测量事物。例如，我们感觉两张凿孔卡应该具有两倍于一张凿孔卡的信息量，两个完全相同的信道信息容量是一个信道的一倍。

3. 它在数学上更合适。很多极限运算在对数方面要简单的多，但如果用可能性的数目那就要求笨拙的重述。

对于对数基底的选择与信息度量的单位选择相一致。当基底是 2 时，所得到的单位可称为比特，这个字由 TUKEY 建议的，一个双稳态设备，如中继器或者触发器，能存储一个二进制单位的信息，N 个双稳态设备就可以存储 N 比特，因为可能状态的总数为  $2^N$ ，而  $\log_2 2^N = N$ 。如果取基底为 10，则单位被称为十进制。因为

$$\log_2 M = \log_{10} M / \log_{10} 2 = 3.32 \log_{10} M$$

故一个十进制单位约为  $3\frac{1}{3}$  个二进制单位。一架台式计算机有十个稳定状态，因此有一个十进制单位的信息存储量。在含有积分和微分的分析计算中，有时候取基底 e，所以所得的单位叫自然单位，把基底 a 换

---

<sup>1</sup>陈国伟、朱瀚敏、隋燕、徐孝芳译，应必娣校，浙江工商大学信电学院，浙江工商大学研究生培养课题（课外同时培养多种能力的模式）资助。

为基底  $b$  仅仅需要乘以  $\log_a b$  就可以。

通信系统可以用图 1 表示，它包含五个基本部分：

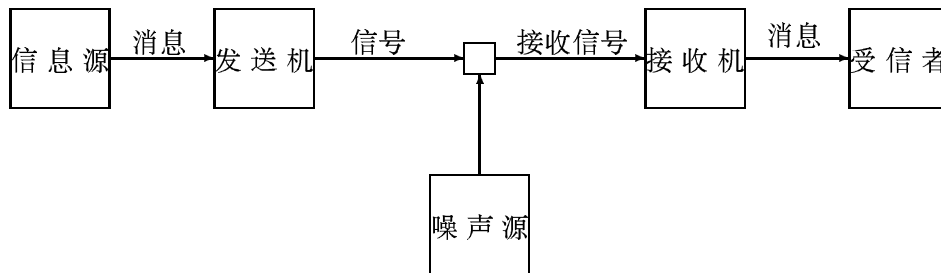


图 1：一般通信系统的示意图

1. 信息源。它产生将要传输给接收端的消息或消息序列。消息可以有各种类型：（a）在电报通信系统中的一系列字母。（b）如电话或无线电中的单独时间函数  $f(t)$ 。（c）如黑白电视中的时间函数和其它变量，在这里，消息可以看成是二维空间和时间函数  $f(x,y,t)$ ，即在  $t$  时刻摄像管上  $(x,y)$  点的光。（d）两个或更多个时间函数  $f(t)$ ， $g(t)$ ， $h(t)$ 。例如在彩色电视中消息是由三个所谓的三维连续函数  $f(x,y,t)$ ， $g(x,y,t)$ ， $h(x,y,t)$ ，组成的，我们可以把这三个函数定义为在这个区域上的矢量场的分量。同样，几个黑白电视源可以产生由几个三变量的函数所组成的消息。（f）各种情况的组合，例如在电视中配有声音信道。

2. 发送机。它是采用某种方法把消息变换为适合于信道上传输的信号。在电话中，这个工作就是把声压编程相应的电流。在电报中，这个工作就是把消息变换为点，划，间隔序列的编码工作。在多路脉冲编码调制系统中，不同的语言函数必须经过取样，压缩，量化和编码，而且最最后构成交叉信号。除此之外把消息变成相应信号的例子还有自动语音合成系统，电视和调频等。

3. 信道。它是发送机到接收机之间用以传输信号的媒质。它可以是一对导线，一条同轴电缆，一段射频的频带，一束光线等等。在传输过程中，或在某一个端点上，信号都可能被噪声所干扰，这种噪声干扰的作用可以看做一个噪声源作用在所传输的信号上构成接收机的信号，如图 1 所示。

4. 接收机。它通常完成与发送机相反的工作，把信号重新构成消息。

5. 消息收受者。是接受消息的人或物。

我们将研究关系到通信系统的某些一般问题。为此，首先必须通过理想化，把各单元用数学来表示。我们可以粗略地把通信系统划分为三个主要类型：离散的，连续的和混合的。离散系统指的是消息和信号都是离散符号的序列。典型的情况是电报，其中，消息是字母序列，而信号是点，划和间隔的序列。在连续系统中，信号和消息都是连续函数，例如无线电话和电视。在混个系统中，离散的和连续的变量都有，例如传输语言的脉冲编码调制系统。

我们首先研究离散情况。这种情况不但可用于信息论，而且也可用于计算机理论，电话交换的设计以及其他场合。此外离散情况也是研究连续和混合情况的基础，后者将在本文后半部分讨论。

# 第一部分无噪声的离散系统

## 1. 无噪声的离散信道

电报和电传打字机是用来传输离散信息的两个简单例子。通常指的离散信道是这样一种系统：它能从选自有限基本符号集合  $S_1, S_2, \dots, S_n$  的序列从一方传输到另一方。假设每个符号  $S_i$  持续时间为  $t_i$  秒（不同的  $S_i$ ，其  $t_i$  不一定相同，例如电报中的点和划）。其实，并不要求所以可能的符号序列都能在系统上传输，而只要求某些序列能够获得传输，这就是对信道的可能的信号。在电报中假设基本符号是：（1）点，它是由一个单位时间的线段和一个单位时间的间歇所组成；（2）划，它是由三个单位时间的线段和一个单位时间的间歇所组成；（3）字母间隔，它是由三个单位时间的间歇组成；（4）单词间隔，它是由六个单位时间的间歇组成。我们可以对序列加以限制，即不允许有间隔相连的情况，因为两个字母间隔连在一起时会变成一个单词间隔。现在我们要考虑的问题是，采用怎样的方法来度量这种信道的容量（或称为信道的传输能力）。

在打字电报，所有的符号都具有相同的持续时间，并且由 32 个符号所构成的任何序列都被允许传输。每个符号都代表五个二进制单位的信息。如果系统每秒能传输  $n$  个符号，则可以自然地认为此信道具有每秒  $5n$  个二进制单位的容量。这并不是说，打字电报信道经常能以这个速率传输信息，这是最大可能的速率，实际上未必能达到这个最大值，下面将谈到，它取决于信道输入端上的信息源。

在一般的情况下，各符号有不同的长度，并且允许的序列是有限制的，我们可以给出下面的定义：离散信道的容量  $C$  为

$$C = \lim_{T \rightarrow \infty} \log \frac{N(T)}{T}$$

其中  $N(T)$  是时间间隔  $T$  内允许信号的数目。

很容易看出，在打字电报中这将简化为前面的结果。可以证明，在大多数情况中，极限是存在的。假设符号  $S_1, S_2, \dots, S_n$  的所有序列都是允许的，并且这些符号的持续时间为  $t_1, \dots, t_n$ ，那么这种信道的容量是多少呢？如果  $N(T)$  表示  $t$  时间内序列的数目，则：

$$N(t) = N(t-t_1) + N(t-t_2) + \dots + N(t-t_n)$$

即这个总数等于终端符号为  $S_1, S_2, \dots, S_n$  的序列数目的综合，并且这些数分别为：

$$N(t-t_1), N(t-t_2), \dots, N(t-t_n)$$

根据有限差分运算，在  $t$  很大时， $N(t)$  就渐近于  $X_0^t$ ， $X_0$  是下列特征方程式的最大实数解：

$$X^{-t_1} + X^{-t_2} + \dots + X^{-t_n} = 1$$

故该信道的容量为：

$$C = \log X_0$$

当允许的符号序列有限制时，仍然经常可以得到这种形式的差分方程式，并可以从特征方程式求得  $C$ 。例如，在上述的电报情况下，则：

$$N(t) = N(t-2) + N(t-4) + N(t-5) + N(t-7) + N(t-8) + N(t-10)$$

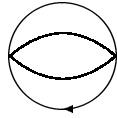
这正和按最后一个符号或最后第二个符号计算符号序列的结果一样。所以  $C$  为  $-\log \mu_0$ ，其中  $\mu_0$  是

$$\mu_2 + \mu_4 + \mu_5 + \mu_7 + \mu_8 + \mu_{10}$$

的正根。解上式，得  $C=0.539$ 。

下面是允许序列的一般限制形式。我们假想有一系列可能的状态  $a_1, a_2, \dots, a_m$ 。在每个状态下，只有  $S_1, S_2, \dots, S_n$  集合中的某些符号可以被传输（这些是不同状态的不同子集）。当其中某一符号被传输后，状

态就转变为新的状态。电报情况就是一个简单例子。它有两个状态，取决于终端符号是否是间隔。如果是间隔的话，那么下一个传输的只能是点或划这两个符号，并且状态总发生变化。如果不是间隔，那么任何符号都可以传输，而且状态只能传输间隔符号后才会转变，否则状态不变。所有这些都可用图 2 来表示。



图中结点代表状态，而线条则表示一状态中的可能符号和它即将转成的状态。在附录 I 中将证明，如果加在允许序列上的条件可以用这种形式来阐述时，那么信道容量  $C$  将存在，并可下下列定理来计算：

定理 1：设  $b_{ij}^{(s)}$  是第  $s$  个符号的长度，这个符号的  $i$  状态是可允许的，并且将转移到  $j$  状态，则信道容量  $C$  等于  $\log W$ ，其中  $W$  为下列行列式方程式中的最大实数根：

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

如果  $i=j$ ，则其中  $\delta_{ij} = 1$ ；如果  $i \neq j$ ，则  $\delta_{ij} = 0$ 。例如，在电报情况下（图 2），其行列式为：

$$\begin{vmatrix} -1 & (W^{-2}) \\ (W^{-3} + W^{-6}) & (W^{-2} + W^{-4} - 1) \end{vmatrix} = 0$$

展开这个行列式将可得到上面那个限制方程式。

## 2. 离散信息源

我们已经看到，在很普遍化的条件下，离散信道中的可能信号数目的对数将随时间作线性增长。因此传输信息的容量可由这个增长速率来确定，即对某一信号需要每秒多少个二进制单位数。

现在我们来讨论信息源。怎样利用数学来描述信息源？一个给定的信息源究竟能产生多少个二进制单位的信息？本文的要点在于研究在采用合理的编码以减少对信道容量要求方面，有关信息源的统计知识有什么作用。例如，在电报中。所传输的消息是由字母序列组成的，但是这些序列并不是完全随机的。通常，它们构成句子而且还是有统计结构例如英文统计结构的句子。字母 E 的出现要比 Q 经常的多，序列 TH 的出现要比 XP 出来的多等等。这种统计结构的存在，允许我们采用合理的编码来节省时间（或信道容量）。其实，这种措施在电报中，已经在一定程度上被采用了。它用最短的信道符号一点来代表最常用的字母 E，而不常用的字母 Q，X，Z，等则用较长的点划序列来表示。这种概念在某些商用电码中得到了进一步的改进，它采用四个到五个字母所组成的码组来表示最常用的单词和短语，因而大大地节省了平均时间。现用标准化的问候语和节日贺电中则更简化到整个一句话或两句话用很短的一个数字序列的编码来表示。

可以这样设想，离散源是一个符号接着一个符号地产生消息的。连续符号的选择是根据某些概率，通常这些概率取决于前面符号的选择及待选择的符号。任何一个能产生由一组概率控制的符号序列的物理系统或物理系统的数学模型都可以称为随机过程。因此，我们可用随机过程来表示离散源。反过来，任何从有限集中选择符号而产生离散符号序列的随机过程都可看成离散源。这包括下列一些情况：

1. 自然语言如英语，德语，汉语。

2. 经过某些量化处理而离散化的连续信息源。例如，在脉冲编码调制发送机中量化以后的语言，或量化以后的电视信号。

3. 在数学上抽象定义的随机过程，该随机过程能够产生序列。下面是最后一种信息源的例子。

( A ) 设有五个字母 A,B,C,D,E, 各以概率 0.2 独立选取，这样就会导致如下典型序列：

BDCBCECCACDCBDDAAECEE  
ABBDAEECACEEBAEECBCEAD.

这个例子是用随机数表构成的。

( B ) 采用同样五个字母，令概率依次为 0.4, 0.1, 0.2, 0.2, 0.1, 各个字母的选择仍是独立的，那么从这种源得到的典型消息为：

AAACDCBDCEAADADACEDA  
EADCABEDADDCECAAAAAD.

( C ) 如果前后字母的选择是不独立的，它们的概率还取决于前面的字母则就得到一个比较复杂的结构。这种类型的最见到情况是每一个字母的选择只与前一个字母有关，与更前面的字母无关。则统计结构可以用转移概率  $p_i(j)$  来描述， $p_i(j)$  表示字母 i 后出现字母 j 的概率。i 和 j 表示符号索引。另一种描述统计结构的等效方法是采用两个字母 ( i,j ) 的联合概率  $p(i,j)$ ，即两个字母一起出现的相对概率。字母出现的概率  $p(i)$ ，转移概率  $p_i(j)$ ，联合概率  $p(i,j)$  之间关系如下：

$$\begin{aligned} p(i) &= \sum_j p(i,j) = \sum_j p(j,i) = \sum_j p(j)p_j(i) \\ p(i,j) &= p(i)p_i(j) \\ p(i) &= \sum_j p_i(j) = \sum_i p(i) = \sum_{ij} p(i,j) = 1 \end{aligned}$$

举一个特例，假设有三个字母 A,B,C, 其概率表为：

$p_i(j)$		j			i	$p(i)$	$p(i,j)$	j		
		A	B	C				A	B	C
i	A	0	$\frac{4}{5}$	$\frac{1}{5}$	A	$\frac{9}{27}$	A	0	$\frac{4}{15}$	$\frac{1}{15}$
	B	$\frac{1}{2}$	$\frac{1}{2}$	0	B	$\frac{16}{27}$	i B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$	C	$\frac{2}{27}$	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

从这个源得到的消息序列为：

ABBABABABABABBBBABBBBABABABA  
BABBBACACABBABBBBABBBABACBBBABA

再复杂些时就是三个字母一起出现的频率。每个字母的选择将取决于前两个字母，但与更前面的字母无关。这时，应采用三个字母的联合概率  $p(i,j,k)$  或转移概率集  $p_{ij}(k)$  来描述。如果用这种方法不断的考虑下去，可以获得更复杂的随机过程。在一般的 n 个字母的情况下，必须采用 n 个字母的联合概率  $p(i_1, i_2, \dots, i_n)$  或转移概率集  $p_{i_1, i_2, \dots, i_{n-1}}(i_n)$  来描述它的统计结构。

( D ) 随机过程也可以定义为由“单词”序列组成文章的过程。假设在语言中有五个字母 A,B,C,D,E 和 16 个“单词”，其相应的概率为：

0.10A	0.16BEBE	0.11CABED	0.04DEB
0.04ADEB	0.04BED	0.05CEED	0.15DEED
0.05ADEE	0.02BEED	0.08DAB	0.01EAB
0.01BADD	0.05CA	0.04DAD	0.05EE

如果前后的“单词”是独立选择的，而且都用一定间隔分开，那么这样组成的典型消息为：

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE BEBE BEBE ADEE BED DEED  
DEED CEED ADEE A DEED DEED BEBE CABED BEBE BED DAB DEED ADEB.

如果所以单词的长度都为有限，则这个过程等效于前面一种类型的过程，不过采用单词结构及其概率来描述可更简单些。也可以普遍化，引入词间的转移概率等等。

这种人造语言在构造简单问题和例子来说明各种概率是很有用的。我们还能用一系列这种语言来近似自然语言。如果以相同的概率独立地选择前后的字母，可得零级近似。如果独立地选择相继的字母，但各个字母具有与它们在自然语言中相同的概率，则可得一级近似。因此，对英语进行以及近似时，E 被选择的概率为 0.12（就是通常英语中出现的概率），W 的概率为 0.02，但相邻字母间没有影响并且也没有使 TH,ED 等等两个字母优先在一起的趋势。在二级近似中，引入了两个字母的结构。当一个字母被选定后，下一个字母的选择就要按照它出现在前一个字母后的频率来选定。这就要求采用两个字母的频率  $p_i(j)$  表。在三级近似中，引入了三个字母的结构，这时第三个字母的选择概率就取决于前两个字母。

### 3. 逼近英语的序列

为了看清楚这个序列过程如何趋近于语言，我们将构造下列逼近英语的典型序列。在所有情况中，我们用一个含有 27 个符号的“字母表”，其中有 26 个字母和一个间隔。

#### 1. 0 级近似（符号独立概率相等）

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

#### 2. 一级近似（符号独立，但各个符号具有英文文字中的频率）

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH  
BRL.

#### 3. 二级近似（考虑到英语中两个字母在一起时的结构）

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE  
AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

#### 4. 三级近似（考虑到英语中三个字母在一起时的结构）

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES  
OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. 第一级单词近似。如果继续考虑英语中四个五个……字母在一起的结构的话，还不如直接跳到以单词为单位来得更好更容易些。这里单词是独立地选择的，但按它们的出现频率。

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL

HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE  
LINE MESSAGE HAD BE THESE.

6. 第二级单词近似。这里考虑到单词间的转移概率，但不包含进一步的结构。

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER  
OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME  
OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

在上面的每个步骤中，可以看出，逼近英语文章的程度显著地增加了。必须指出，这些样品的合理结构，约比字母结构上考虑的好两倍。例如在（3）中的统计过程保证了两个字母序列构成较合理的文句。样品中的四个字母序列一般地说也能在良好的句子中适应。在（6）中，由四个或更多个单词构成的序列也很容易放进句子，而不会造成异常的或别扭的句法。由十个单词构成的特殊序列“Attack on an English writer that the character of this”并非完全不合理的。由此可见，一个足够复杂的随机过程能够满意地表示一个离散源。

前两个样品是利用一本任意字数的书和字母频率表（例2用）所构成的。这种方法也可以在（3），（4），（5）中使用，因为两个字母，三个字母和单词的频率表可以得到，但我们采用了更简单的等效的方法。例如，（3）是这样构成的，我们随便翻开一本书，并在该页上随便选择一个字母，把它记下。再把书翻到另外一页，并进行阅读，知道遇见这个字母为止，然后把这个字母后面的那个字母记下。我们再把书翻到另一页去寻找被记下的第二个字母，然后再记下它后面的那个字母……。 （4），（5），（6）也用这种方法。如果进一步逼近能够构成的话，那是很有意思的，但工作量是很大的。

#### 4. 马尔可夫过程的图解表示

上述的随机过程类型在数学上称为离散的马尔可夫过程，它在文献中已有广泛的研究。一般情况可以描述如下：系统存在有限个可能的状态， $S_1, S_2, \dots, S_n$ 。此外，有一组转移概率 $p_i(j)$ ，它是系统由状态 $S_i$ 转到状态 $S_j$ 的概率。为了能使这个马尔可夫过程称为一信息源，我们只需假设，当一个状态转到另一状态时产生一个字母。这些状态对英语前面字母的“影响残余”。

这种情况可以用图来表示（如图3，4和5）。图中结点表示“状态”，由一个状态转到另一个状态的转移概率及其所产生的字母注在相应的线旁。图3就是第二节中的例子B，图4就是例子C。在图3中，只有一个状态，因为前后字母是独立的。在图4中，字母和状态的数目是一样多的。在三个字母结构的情况下，最多有 $n^2$ 个状态与被选字母前一对字母相对应。图5是例D中单词结构的图解。其中S表示间隔符号。

画图：

#### 5. 遍历性源及混合源

如上所述，离散源可用马尔可夫过程来表示。在可能的离散马尔可夫过程中，有一组具有特别性质的并在通信理论中占有特殊意义的马尔可夫过程，这种特殊的马尔可夫过程是由“遍历”过程组成的，所以我们把相应的信息源称为遍历信息源。虽然，遍历过程的严格定义是相当复杂的，但一般概念是简单的。在一个遍历过程中，每一个由过程所产生的序列都有相同的统计性质。例如从某一个特定序列中得到的字母频率，两个字母在一起的频率等等，在增大序列长度时，将趋于某一个确定的极限，并与所取序列无关。其实，这

个并不是在所有的序列中都能成立，但不成立的那些序列其概率为零。粗略地说，遍历性就是统计均匀性。

所有上面讨论过的人造语言的例子都是遍历的。这种性质与相应线图的结构有关。如果线图具有下列两种性质，那么相应的过程将是遍历的：

1. 线图不能分割为两个隔离的部分 A 和 B，因而不可能从 A 部中的一个结点沿着图上的方向到达 B 部的结点，且也不可能从 B 部的结点到达 A 部的结点。

2. 对线图上那些封闭的线，如果在线上的各个箭头都指向同一个方向，我们称它为回路。回路的长度就是它里面的线段数目。例如在图 5 中，BEBES 就是一个长度为 5 的回路。而第二个性质就要求图中所有回路长度的最大公约数应当等于 1。

如果第一个条件满足，而第二个条件则由于最大公约数  $d > 1$  而被破坏，那么这种序列就具有某种周期结构的形式。各种不同的序列可以划分为  $d$  类，除了源点转移外（即在序列中称为 1 的字母），它们的统计结构都是一样的。任何序列，通过从 0 移到  $d - 1$ ，那么就可把它变成的统计上等效的另一序列。当  $d = 2$  时：设有三个可能的字母 a,b,c。跟在 a 后面的是 b 或者 c，它们的概率各位  $\frac{1}{3}$ ， $\frac{2}{3}$ 。在 b 和 c 后面总跟着字母 a。于是典型的序列为：

a b a c a c a c a b a c a b a b a c a c。

显然，这种情况对我们的工作是没有意义的。

如果第一个条件不满足，则线图可以分割为几个子线图，其中每个子线图都满足第一个条件。假设每个子线图满足第二个条件，那么这种情况将可称为混合源，它是由几个纯分量组成的。每个分量都对应着不同的子线图。如果  $L_1, L_2, L_3, \dots$  表示分量源，则可写成：

$$L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \dots$$

其中  $p_i$  是分量源  $L_i$  的概率。

此式的物理意义是：有几个不相同的源  $L_1, L_2, L_3, \dots$ ，其中每个源都有均匀的统计结构（即它们是遍历的）。虽然我们并不预先知道哪一个分量将被使用，但是一旦序列从某给定的纯分量  $L_i$  开始后，它将按照自己的统计结构无限地继续下去。

作为一个例子，可以取上面定义的两个过程，假设  $p_1 = 0.2$ ， $p_2 = 0.8$ 。于是从这个混合源可得一个序列：

$$L = 0.2L_1 + 0.8L_2$$

它是这样得到的：首先按概率 0.2 和 0.8 来选择  $L_1$  或  $L_2$ ，然后形成由这些选择所确定的序列。

除了特别注明的以外，我们将假设源是遍历的。这个假设允许我们把沿着序列的平均值和可能序列全体（总体）的平均值看做是相等的（差异的概率为零）。例如，在某无限序列中字母 A 的相对出现次数将（以概率 1）等于它在序列全体（总体）中的相对出现次数。

如果  $p_i$  是状态  $i$  的概率， $p_i(j)$  是转到状态  $j$  的转移概率，那么对于平稳过程来讲，显然  $p_i$  必须满足平衡方程式：

$$P_j = \sum_i P_i p_i(j)。$$

在遍历情况中，可以证明，在任何起始条件下，当  $N \rightarrow \infty$  时， $N$  个符号以后在  $j$  状态的概率  $p_j(N)$ ，将趋于平衡值。



## 6. 选择，不确定性和熵

我们已把离散信息源表示为马尔可夫过程。现在提出这样一个问题：能否定义一个量，这个量在某种意义上能度量这个过程所“产生”的信息是多少？或者更理想一点，所产生的信息速率是多少？

假设有一可能事件集，它们出现的概率为  $p_1, p_2, \dots, p_n$ 。这些概率是已知的，但是，我们所关心的是哪一个事件将会出现。现在我们能否找到一种测度来量度事件选择中含有多少“选择的可能性”，或者，找到一种测度，来量度选择的结果具有多大的不确定性呢？

如果这样的测度存在的话，我们用  $H(p_1, p_2, \dots, p_n)$  来表示，那它就应该具有下列的性质：

1.  $H$  对  $p_i$  应当是连续函数。

2. 如果所有的  $p_i$  相等， $p_i = \frac{1}{n}$ ，那么  $H$  将是  $n$  的单调递增函数。即对于等概率事件，当有更多可能事件时，则就有更多的选择可能性或更多的不确定性。

3. 如果选择分为前后两个步骤，那么原始的  $H$  将等于各个  $H$  值的加权和。它的意义可由图 6 来说明。

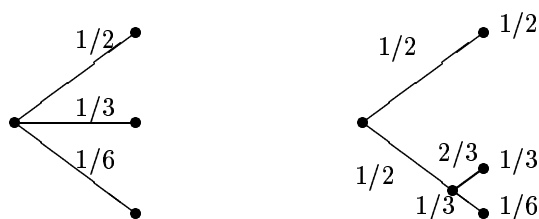


图 6：三种可能性选择的分解

在左图中有三种可能性，相应概率为  $p_1 = \frac{1}{2}$ ， $p_2 = \frac{1}{3}$ ， $p_3 = \frac{1}{6}$ 。在右图中首先在两个概率各为  $\frac{1}{2}$  的可能性中做出选择。如果第二种可能性出现，那么再以概率  $\frac{2}{3}$ ， $\frac{1}{3}$  作另一种选择。最后结果还是同前面的一样，在这个特殊情况中，我们要求：

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right)$$

其中系数  $\frac{1}{2}$  为加权因子，因为第二部选择只是在总数的一般的情况下进行的。

在附录 2 中得出了下列结果。

定理 2：唯一满足上述三个条件的  $H$  具有下列形式：

$$H = -K \sum_{i=1}^n p_i \log p_i$$

其中  $K$  是正常数。

这个定理以及证明它所要求的假设，对本理论来讲是次要的。这里谈它的目的主要是想给利用后面的定义加一些合理性。

度量  $H = -\sum p_i \log p_i$ （常数  $K$  仅等于度量单位的选择）在信息论中起着重要的作用，它作为信息，选择和不确定性的度量。 $H$  的公式与统计力学中所谓熵的公式是一样的。式中  $p_i$  表示一个系统处在它相空间中第  $i$  个元的概率。因此，这里的  $H$  就是波尔兹曼著名的  $H$  定理中的  $H$ 。我们将把  $H = -\sum p_i \log p_i$  称为概率集  $p_1, \dots, p_n$  的熵。如果  $x$  表示随机变量，那么我们可用  $H(x)$  表示它的熵，必须指出，这里的  $x$  并不是函数的变量，而仅仅是数的标记，使之与随机变量  $y$  的熵  $H(y)$  相区别。

在具有概率为  $p$  和  $q = 1 - p$  的两种可能性情况下，它的熵为：

$$H = -(p \log p + q \log q)$$

熵与  $p$  之间的关系表示在图 7 中。

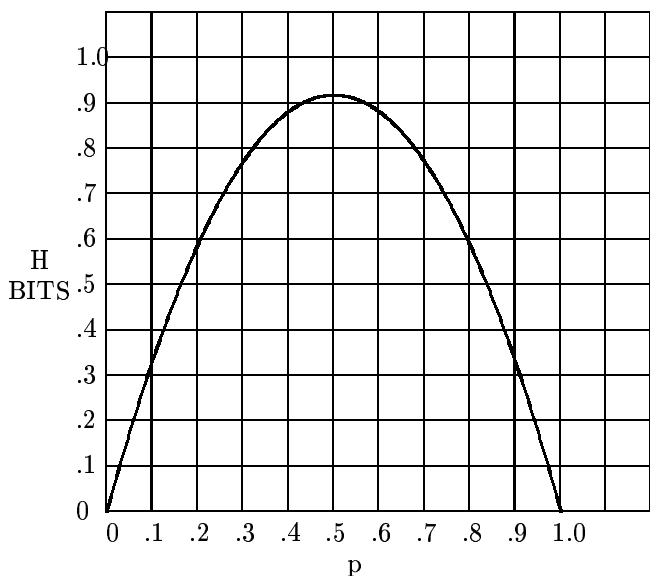


图  
(7)

$H$  具有许多有趣的性质，这些性质进一步正事它作为选择或信息的度量的合理性。

1. 当且仅当所有的概率  $p_i$  除了其中一个概率外其余概率为 0 时，则  $H = 0$ 。换句话说，仅仅在结果完全确定的情况下， $H$  才等于零。否则  $H$  将是正的。

2. 在给定  $n$  的条件下，若所有的  $p_i$  相等，即为  $\frac{1}{n}$  时，则  $H$  将达到最大值，并等于  $\log n$ 。这也是最不肯定的情况。

3. 设有两个事件  $x$  和  $y$ ，分别有  $m$  和  $n$  种可能性。令  $p(i, j)$  表示  $i$  (事件  $x$ ) 和  $j$  (事件  $y$ ) 的联合出现概率，那么该联合事件的熵为：

$$H(x, y) = - \sum_{ij} p(i, j) \log p(i, j)$$

而

$$H(x) = - \sum_{ij} p(i, j) \log \sum_j p(i, j)$$

$$H(y) = - \sum_{ij} p(i, j) \log \sum_i p(i, j)$$

容易证明：

$$H(x, y) \leq H(x) + H(y)$$

只有当两个事件为相互独立时，等式才会成立（即  $p(i, j) = p(i) \cdot p(j)$ ）。

4. 任何一种能使概率  $p_1, p_2, \dots, p_n$  趋于均等的变动，都会使  $H$  增加。例如，当  $p_1 < p_2$  时，如果我们增加  $p_1$ ，而  $p_2$  则减少相同的量，使  $p_1$  与  $p_2$  更接近相等时，那么  $H$  是增大的。更一般地说，如果我们对概率  $p_1$  做形式为

$$p'_i = \sum_j a_{ij} p_j$$

的“平均”运算，则  $H$  增大（除非在通常情况下，即这样的变换只不过是一种  $p_j$  排列，那么当然， $H$  将保持不变）。其中  $\sum_i a_{ij} = \sum_j a_{ij} = 1$ ，且所有  $a_{ij} \geq 0$ 。

5. 设有两个随机事件  $x$  和  $y$  如 3 中所述，但它们不一定独立。对  $x$  的任何可能值  $i$ ，有一个  $y$  有  $j$  值的条件概率  $p_i(j)$ ，它等于：

$$p_i(j) = \frac{p(i,j)}{\sum_j p(i,j)}$$

我们可以这样定义  $y$  的条件熵  $H_x(y)$ ：它是每一个  $x$  值所得的  $y$  的熵，根据获得  $x$  值的概率加权进行平均所得的平均值。即：

$$H_x(y) = - \sum_{i,j} p(i,j) \log p_i(j)$$

当  $x$  已知时，这个量可用来度量  $y$  的平均不确定性。将  $p_i(j)$  的值代入，得：

$$H_x(y) = - \sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} \log \sum_j p(i,j) p(i,j) = H(x,y) - H(x)$$

或者：

$$H(x,y) = H(x) + H_x(y)$$

因此联合事件的不确定性（或熵）等于  $x$  的不确定性加上当  $x$  为已知时  $y$  的不确定性。

6. 由 3 和 5 可得：

$$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y)$$

由此可见， $y$  的不确定性决不会由于  $x$  的熵的增加而有所增加。如果  $x$  和  $y$  不是独立的，那么它将减少，在独立的情况下，它不变。

## 7. 信息源的熵

现在讨论上面研究过的有限状态的离散源。对每个可能的状态  $i$  都有一产生各种可能符号  $j$  的概率集  $p_i(j)$ 。因此，每个状态都有一个熵  $H_i$ 。而信息源的熵将定义为这些  $H_i$  按照各个状态出现的概率加权而得到的平均值。

$$H = \sum_i p_i H_i = - \sum_{i,j} P_i p_i(j) \log p_i(j)$$

这是信息源中每个符号的熵。如果马尔可夫过程是按照一定速率进行的，那么信息源每秒的熵为

$$H' = \sum_i f_i H_i$$

其中  $f_i$  是状态  $i$  的平均出现次数（每秒出现次数）。显然

$$H' = mH$$

其中  $m$  是每秒产生的平均符号个数。 $H$  或  $H'$  度量了信源产生的信息量，即每个符号的信息量或每秒所产生的信息量。如果对数的底取 2，那么它将表示为每个符号多少二进制单位或每秒多少二进制单位。

如果顺序出现的符号是独立的，那么  $H$  就简单地变为  $-\sum p_i \log p_i$ ，其中  $p_i$  是符号  $i$  的概率。假设在这种情况下，我们讨论一个由  $N$  个符号组成的长消息。在这个长消息中，第一个符号出现  $p_1 N$  次，第二个符号出现  $p_2 N$  次，等等。因此这种消息的概率可以粗略地写为：

$$p = p_1^{p_1 N} \cdot p_2^{p_2 N} \cdot \dots \cdot p_n^{p_n N}$$

或者：

$$\log p = N \sum p_i \log p_i$$

$$\log p = -NH$$

$$H = \frac{\log 1/p}{N}$$

故  $H$  近似地等于一个典型长序列的概率倒数的对数初一序列中的符号个数。对任何信源都有同样的结果。更精确的公式，可参看附录 3。

定理 3 给定任何  $\epsilon > 0$  和  $\delta > 0$ ，我们可以找到  $N_0$ ，使得任何长度为  $N \geq N_0$  的序列都可分为两类：

1. 总概率小于  $\epsilon$  的一组。

2. 其他概率满足不等式:

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta \text{ 的为一组。}$$

换句话说, 当  $N$  很大时,  $\frac{\log p^{-1}}{N}$  肯定将接近于  $H$ 。

当长度为  $H$  的一些序列中, 把它们按概率递降的次序来排列。我们定义  $n(q)$  为这样一个数目, 即从概率最大的那个序列选取, 一直取到总概率等于  $q$  时所必需取的序列的数目。

定理 4 当  $q$  不为 0 或 1 时,

$$\lim_{N \rightarrow \infty} \frac{\log n(q)}{N} = H$$

其中  $\log n(q)$  表示具有总概率为  $q$  的那些最可能序列时为了描述序列所需要的二进制单位数。于是  $\frac{\log n(q)}{N}$  是说明每个符号所需要的二进制单位数。这个定理说明了当  $N$  很大时, 它就与  $q$  无关, 并等于  $H$ 。因此不管“比较可能”的术语如何解释, 但是, 比较可能的序列的个数的对数增长率将由  $H$  确定。由于这些结果 (在附录 3 中将有证明), 在大多数情况下, 可以将长序列看作  $2^{NH}$  个, 并且每个序列的概率为  $2^{-NH}$ 。

下面两个定理将证明  $H$  和  $h'$  可以直接从消息序列的统计特性的极限运算来确定, 而不必考虑这些状态之间的转移概率。

定理 5 令  $P(B_i)$  是信源输出端上出现符号序列  $B_i$  的概率。取

$$G_N = -\frac{1}{N} \sum_i p(B_i) \log p(B_i)$$

其中  $\Sigma$  是含有  $N$  个符号的所有序列  $P(B_i)$  的总和。因此,  $G_N$  是一个  $N$  的单调递减函数, 并且

$$\lim_{N \rightarrow \infty} G_N = H$$

定理 6 令  $p(B_i, S_j)$  表示被符号  $S_j$  所跟随的序列  $B_i$  的概率,  $p_{B_i}(S_j) = p(B_i, S_j)/p(B_i)$  表示  $B_i$  出现后出现  $S_j$  的条件概率。取

$$F_N = -\sum_{i,j} p(B_i, S_j) \log p_{B_i}(S_j)$$

其中  $\Sigma$  是对所以由  $N-1$  个符号组成的群 (区组)  $B_i$  以及所有符号  $S_j$  的总和。于是  $F_N$  是  $N$  的单调下降函数

$$F_N = NG_N - (N-1)G_{N-1}$$

$$G_N = \frac{1}{N} \sum_{i=1}^N F_N$$

$$F_N \leq G_N$$

并且

$$\lim_{N \rightarrow \infty} F_N = H$$

所有这些结果都在附录 3 中推导。它们指出, 只要考虑扩展到  $1, 2, \dots, N$  个符号的序列的统计结构, 就能得到一系列  $H$  的近似式。  $F_N$  是一种较好的近似。实际上,  $F_N$  是上面讨论过的源的第  $N$  级近似的熵。如果把序列扩展到多于  $N$  个符号而没有统计影响, 即如果知道了前面  $N-1$  个符号后, 下一个符号的条件概率并没有被任何前面的知识所改变, 于是  $F_N = H$ 。显然,  $F_N$  是已知前面 ( $N-1$ ) 个符号时, 下一个符号的条件熵, 而  $G_N$  是由  $N$  个符号所组成的群 (区组) 的每个符号的熵。

信源的熵与其最大值 (限于同样符号) 的比值称为相对熵。后面将看到, 这是当我们编成相同字母码时, 可以得到的最大可能的压缩。以 1 减去相对熵就是冗余度。普通英语在不考虑距离超过 8 个字母以上的统计结构时, 它的冗余度约为 50%。这就是说, 当我们写英语时, 其中一半是确定于语言的结构, 而另一半则可以自由选择的。用不同的方法都可求得与 50% 相近的结果。一种方法是计算近似英语的熵。第二种方法是从某一英语文中涂掉某些字母, 然后叫人去恢复它。如果涂掉 50% 以后仍能恢复, 那么冗余度必然大于 50%。第三种方法是利用密码技术中的某些已知的结果。

一种方法是计算近似英语的熵。第二种方法是从某一种英语文章中涂掉某些字母，然后叫人去恢复它。如果涂掉 50 % 以后仍能恢复，那么多余度必然大于 50

英文散文中多余度的两种极端情况可由“基础英语”和 James Joyce 的“Finnegans Wake”来代表。“基础英语”的基本词汇限制在 850 个字以内。它的多余度是很高。这反映在把一节文章译成“基础英语”时篇幅就得增大。另一方面，James Joyce 扩大了词汇，并被认为达到了语义内容的压缩。

语言的多余度与纵横字谜的存在是有联系的。如果多余度为零，则任何字母序列在语言中都是合理的句子，并且任何二维字母列都可构成一种纵横字谜。如果多余度太高，则语言的限制太多，大的纵横字谜的可能性就小了。更详细的分析指出，如果我们假定语言所加的约束是随机性质的，则大的纵横字谜只有在多余度为 50

#### 8. 编码和解码操作的表示法

现在我们从数学上来描述发送机和接收机编码和译码过程。它们都可称为离散变换器。在变换器输入端上加入输入符号序列，它的输出是输出符号序列。在一般情况下，变换器可以具有记忆力，因而它的输出不只决定于现在的输入符号，还决定于过去的符号。我们假设内部的记忆力是有限的，即变换器具有有限数目维  $m$  的可能状态，它的输出是现在状态和现在输入符号的函数。而下一个状态将是这两个量的函数。因此，一个变换器能用下列两个函数来描述：

$$\begin{aligned} y_n &= f(x_n, a_n) \\ a_{n+1} &= g(x_n, a_n) \end{aligned}$$

其中  $x_n$  是第  $n$  个输入符号。

$a_n$  是当第  $n$  个输入符号引入时，变换器所呈现的状态。

$y_n$  是在  $a_n$  状态下，引入  $x_n$  输入时产生的输出符号（或者输出符号序列）。

如果一个传感器的输出符号能与第二个变换器的输入符号相同，那么它们可以串联在一起构成一个变换器。如果接在第一个变换器输出端上的第二个变换器能够把第一个变换器上的输出符号序列恢复为原来的输入符号序列，那么这里第一个变换器将称为非奇异变换器，而第二个变换器成为反变换器。

定理 7：在限定状态的统计源激励下，有限状态变换器的输出也是一个有限状态的统计源，它的熵（每单位时间）小于或等于输入统计源的熵。如果变换器是非奇异变换器，他们的熵相等。

令  $a$  代表产生符号序列  $x_i$  的源的状态， $b$  表示在其输出端上产生符号群  $y_j$  的变换器的状态。则联合系统可以用序偶  $(a, b)$  在这空间  $(a_1, b_1)$  和  $(a_2, b_2)$  两个点可以用一条线连起来，假如  $a_1$  可以产生一个  $x$  并且把  $b_1$  变为  $b_2$ ，且该线就给出这个  $x$  在此情况下的概率。此线用变换器所产生  $y_j$  符号群来标记。输出的熵可以采用对整个状态的加权和来计算。如果我们先对  $b$  求和，结果每一项都小于或等于  $a$  的相应的各项，因此熵不会增加。如果传感器非奇异，那么就把它输出连接到一个反的变换器。如果  $H_1^1, H_2^1, H_3^1$  分别表示信源，第一和第二个变换器的输出熵，那么  $H_1^1 \geq H_2^1 \geq H_3^1$ ，因此  $H_1^1 = H_2^1$ 。

假设我们有一个对可能序列制约的系统，并且，这样的系统可以用图 2 的线图表示。如果概率  $p_{ij}^{(s)}$  指定由那些联接状态  $i$  到状态  $j$  的线来表示，则这样的系统将成为一个源。有一个特殊的指定方法能使得到的熵最大（见附录 4）。

定理 8：把制约系统看作一个具有容量  $C = \log^W$  的信道。如果我们令

$$p_{ij}^{(s)} = \frac{B_j}{B_i} W^{-l_{ij}^{(s)}}$$

其中  $l_{ij}^{(s)}$  是从状态  $i$  变成状态  $j$  的第  $s$  个符号长度，且  $B_i$  满足

$$B_i = \sum_{s,j} B_j W^{-l_{ij}^{(s)}}$$

那么熵  $H$  是将达到最大，并等于信道容量  $C$ 。

适当指定转移概率，可使信道上符号的熵达到最大，并等于信道容量。

### 9. 无噪声信道的基本原理

通过证明  $H$  确定所需的信道容量与最有效的编码，我们现在可以解释  $H$  产生信号的比率。

定理 9：假设信源的熵为  $H$ （每个符号的二进制单位数），信道容量为  $C$ （每秒的二进制单位数）。于是信源的输出可以进行这样的编码，使得在信道上传输的平均速率为每秒  $\frac{C}{H} - \epsilon$  个符号，其中  $\epsilon$  是任意小量，要使传输的平均速率大于  $\frac{C}{H}$  是不可行的。

定理的逆命题即速率不可能超过  $\frac{C}{H}$  可以这样来证明：因为发送机必须是是非奇异的，每秒输入信道的熵等于信源的熵。这个熵也不可能超过信道容量。故  $H^1 \leq C$ ，并且每秒符号的个数为  $= H^1/H \leq C/H$ 。

定理的第一部分可以用两种不同方法证明。第一种方法是考虑由信源产生的所有  $N$  个符号的序列集，当  $N$  很大时，可以把这些序列分为两组，一组包含少于  $2^{(H+\eta)N}$  个序列，第二组包含少于  $2^{RN}$  个序列（其中  $R$  是不同的符号数的对数）并且总的概率小于  $\mu$ 。当  $N$  增大时  $\eta$  和  $\mu$  趋向于 0。在信道中持续时间为  $T$  的信号数目将大于  $2^{(C-\theta)T}$  个，当  $\theta$  很小时， $T$  很大，如果选择

$$T = \left(\frac{H}{C} + \lambda\right)N$$

那么当  $N$  和  $T$  足够大时（无论  $\lambda$  多小概率高的一类总有足够数量的信道符号序列，当然还有某些附加的符号序列。高概率序列以任意一对一的形式编入这个序列集中。剩下的序列可用较长的序列来表示，以不同于高概率类的序列作为它的始端或终端。这种序列作为不同的码子的开始和终了信号。在两者之间有足够的的时间间隔，以便形成足够数量的不同低概率序列。这个要求

$$T_1 = \left(\frac{H}{C} + \varphi\right)N$$

其中  $\varphi$  很小。因此，在符号信息中平均每秒传输速率将会大于

$$\left[(1-\delta)\frac{T}{N} + \delta\frac{T_1}{N}\right]^{-1} = \left[(1-\delta)\left(\frac{H}{C} + \lambda\right) + \delta\left(\frac{R}{C} + \varphi\right)\right]^{-1}$$

当  $N$  增加了  $\delta$ ， $\lambda$  和  $\varphi$  接近于 0 并且比率接近  $\frac{C}{H}$ 。

第二种完成编码的方法是把长度为  $N$  的消息按概率递减的次序进行排列，并且假设它们的概率  $p_1 \geq p_2 \geq p_3 \dots \geq p_n$ 。令  $p_s = \sum_{i=1}^{s-1} p_i$ ；即累积概率一直积累到  $p_s$ ，但不包括  $p_s$ 。我们首先把它编成一个二进制系统。这个二进制编码通过把  $p_s$  扩大成一个二进制数字来获取信息。这种扩张执行到  $m_s$  的位置， $m_s$  是一个整数并且满足：

$$\log_2 \frac{1}{p_s} \leq m_s < 1 + \log_2 \frac{1}{p_s}$$

因此出现频率高的信息用短代码表示，出现频率低的信息用长代码表示。从这种不平均性我们可以得到

$$\frac{1}{2^{m_s}} \leq p_s < \frac{1}{2^{m_s-1}}$$

$p_s$  的代码和所以成功匹配的不相同或者高于  $m_s$  的位置，因为所有剩余的  $p_i$  至少有  $\frac{1}{2^{m_s}}$  那么大它们的二进制扩张因此和第一个  $m_s$  的位置不同。从而所有的代码都不相同并且从代码中恢复原来的信息是有可能的。

假如信道的序列不是二进制数字的顺序，它们可能以任意的方式归因到二进制数，并且二进制代码将会翻译成适合信道的信号。

原始消息中每个符号所用的二进制字的平均数  $H^1$  很容易确定：

$$H^1 = \frac{1}{N} \sum m_s p_s$$

但是，

$$\frac{1}{N} \sum \left( \log_2 \frac{1}{p_s} \right) \leq \frac{1}{N} \sum m_s p_s < \frac{1}{N} \sum \left( 1 + \log_2 \frac{1}{p_s} \right) p_s$$

因此，

$$G_N \leq H' < G_N + \frac{1}{N}$$

当  $N$  增加  $G_N$  接近  $H$ ，信源的熵和  $H^{A1/}$  接近  $H$ 。

从这里可以看出，当只使用  $N$  个符号的有限延迟时，编码的无效性不必大于  $\frac{1}{N}$  加上“真正的熵  $H$  与由长度为  $N$  的序列算得的  $G_N$  之差”。因此上述理想情况所要求的百分超额时间将小于

$$\frac{G_N}{H} + \frac{1}{HN} - 1.$$

这种编码方法和 *R.M.Fano*<sup>9</sup> 建立的方法实质上是一样的。他的方法是把长度为  $N$  的信息进行排序为了减少发生率。把这一系列分成两组，两组发生的概率是几乎相同的。假如信息在第一个组里那么它的第一个二进位数字是 0，否则是 1。这些组同样的分成概率相等的几个子集，特殊的子集决定了第二个二进位数字。这个过程直到每个子集仅包含一个信息时结束。很容易看到除了较小的差别（一般在最后一个数字）这个总数和先前描述的算法过程是一样的。

## 10. 讨论和举例

为了从发电机的负载获得最大的能量，变压器被引入了，所以发电机的负载有负载阻抗。这种情况在这里大体相似的。编码用的变换器在统计学意义上应该和信源到信道相匹配。通过变换器从信道而来的信源应该有相同的统计结构当作来源，它使在信道中的熵最大化。定理 9 的内容是，尽管精确的匹配一般是不可能的，我们可以把它近似得如我们所需要的，实际速率的比率传输到容量  $C$ ，可以被称为编码系统的效率。这个当然等于信道符号实际熵到最大可能熵的比率。

一般来说，理想或者接近理想的编码在发射机和接受机中要求一个很长的延时。在无噪声情况下我们已经知道，延迟的主要作用是允许合理好的概率匹配相当于排序的长度。有一个好的编码一个长信息的对数倒数的概率必须和通信信号的持续时间成比例，实际上

$$\left| \frac{\log p^{-1}}{T} - C \right|$$

必须对整个来说很小，但是对于长信息来说是一小部分。

<sup>9</sup>Technical Report No.65, The Research Laboratory of Electronics, M.I.T., March 17, 1949.

假如一个信源只能产生一个特殊的信息，它的熵是 0，并且没有所需的信道。一个计算机建立起来计算连续数字  $\pi$  时产生了一个无偶然元素的确定序列。没有信道是来要求把它传送给另一个指定的目标。在这一点上我们可以构造第二个机器来计算相同的序列。然而这个可能是不切实际的。在这种情况下我们可以

选择忽略一些或者全部关于那些信源的统计信息。我们可以认为数字  $\pi$  是一个随机的序列因为我们构造了可以发送任何数字序列的系统。用相似的方法我们可以选择使用一些我们关于英语统计的只是来构造一个代码，但不是全部。在这种情况下我们认为最大熵的信源附属我们所希望保留的统计情形。信源的熵决定了信道的能力哪个是必要和充分的。在这个  $\pi$  的例子中仅仅保留的信息是从集合  $0, 1, \dots, 9$  中选择的数字。在英文的情形下我们希望使字母来统计一切可能的字母频率，但不是其他的。最大的熵信源是第一个接近英文的并且它的熵决定了所需的信道容量。

举一个简单的例子来说明以上的结果。设有一个信源产生了一系列字母，分别是 A,B,C,D，它们的概率分别是  $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ ，并且相继符号的选择是独立的。

那么

$$H = - \left( \frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{2}{8} \log \frac{1}{8} \right) \\ = \frac{7}{4} \text{ 比特每符号}$$

因此我们可以近似地把一个从信源来的编码系统的编码信息转化为带有平均  $\frac{7}{4}$  二进制数字每字符的二进制数字。在这种情形下我们实际上可以达到下面代码的限值 (定理 9 中的第二种证明方法)：

A	0
B	10
C	110
D	111

N 个符号序列编码中使用的平均二进制数字值是

$$N \left( \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{2}{8} \times 3 \right) = \frac{7}{4} N.$$

很容易看到二进制数字 0.1 各自的概率都是  $\frac{1}{2}$ ，所以 H 编码序列是一个比特每符号。因为，平均我们有  $\frac{7}{4}$  二进制符号每原始字符，在同个时间基础上熵是相同的。对于原始集最大可能熵是  $\log 4 = 2$ ，当 A,B,C,D 各自的概率是  $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ 。因此相关的熵是  $\frac{7}{8}$ 。我们可以把二进制序列翻译成符号的原始集在 2 到 1 的基础上通过下表表示：

00	A'
01	B'
10	C'
11	D'

双处理把原始信息编码成相同的符号但是有一个平均的压缩比率  $\frac{7}{8}$ 。

第二个例子把信源认为是产生 A 和 B 的概率， $p$  代表 A 的概率， $q$  代表 B 的概率。假如  $p < q$  则

$$H = - \log p^p (1-p)^{1-p} \\ = -p \log p (1-p)^{(1-p)/p} \\ = p \log \frac{e}{p}$$

在这种情况下可以发送一个特殊序列，在 0 和 1 信道上建立良好的编码。例如，对不常用的符号 A 采用 0000，然后跟一个表示字母 B 的数目的序列。这个数目可以用但包含删去的特殊序列在内的所有二进制数表示。所有的数字包含了特殊删除的序列。所有数字达到 16 制后用平常的代表；16 用第二个二进制数字代表在 16 以后不包括 4 个 0，也就是  $17=10001$ ，等等。

可以证明当  $p \rightarrow 0$  时如果特殊序列的长度进行适当的调整，那么编码就趋于理想编码。

## 第二部分：有噪离散信道



### 11.A 离散噪声信道表示法

现在我们研究信号在传输过程中在信道的这一端或那一端受到噪声干扰的情况。这就是说接收到的信号并不一定与发送机发出的信号相同。这可以分成两种情况。如果特定的被传输的信号总是产生同样的接收信号，即接收到的信号是传输信号的确定的函数，那么，这是信道中的畸变。如果这个函数存在着反函数，即任何两个被传输的信号不会产生同样的接收信号，那么至少在原则上这种畸变是可以对接受的信号进行反函数运算得到校正。这里感兴趣的情况是信号在传输过程中并不总是受到同样的变化。于是我们可以认为接收到的信号

$E$  是传输信号  $S$  和第二个可变量  $N$  的函数。

$$E = f(S, N)$$

噪声和上述的消息一样。可以看成是一个随机变量。在一般情况下，噪声可以用一个合适的随机过程来表示。我们将研究噪声离散信道的最一般形式，它是前面所述的有限状态的无噪声信道的推广。我们假设状态数目是有限的，并有一概率集

$$p_{\alpha,i}(\beta, j)$$

这个概率是信道处于状态  $\alpha$ ，发送符号  $i$ ，接受符号  $j$ ，而信道处于状态  $\beta$  的概率。这里  $\alpha$  和  $\beta$  包括所有可能的状态， $i$  包括所有可能被传输的信号， $j$  是包括所有可能接收到的信号。如果相继的符号各自独立地受到噪声的干扰，则只有一个状态，信道可用转移概率  $p_i(j)$  来描述，它是传输的信号为  $i$  而收到的信号为  $j$  的概率。如果把信源馈给一不噪声信道，就有两个统计过程起作用：

信源和噪声。因此有大量的熵可以来计算。首先是信源的熵亦即输入信道的熵  $H(x)$  (如果发送机是非奇异的，它们是相等的)。其次是信道输出的熵，即接收到的信号的熵，它用将会由  $H(y)$  表示。在无噪声情况下  $H(y) = H(x)$ 。输入和输出的联合熵将会是  $H(xy)$ 。最后有两个条件熵  $H_x(y)$  和  $H_y(x)$ ，它们是当输入为已知时输出的熵以及输出为已知时输入的熵。这些量之间的关系是

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

所有的这些熵都可以通过每秒或每个符号来度量。

### 12. 模棱性和信道容量

如果信道中有噪声，一般地说，不可能对所接收到的信号进行任何运算来完全确定地重新构成原先的消息或所传输的信号  $E$ 。但是可以有某些传转信息的方法，这些方法在抗干扰方面是最佳的。这个是我们现在要讨论的问题。

假设有两个可能的符号 0 和 1，并且我们以每秒 1000 个字符的速率来传输，概率是  $p_0 = p_1 = \frac{1}{2}$ 。那么我们的信源以每秒 1000 二进制单位的速率产生信息。如果在传输过程中。由于噪声而引入了误差，从平均值来看，在 100 个接收到的符号中由一个符号是不正确的 (即当传输符号为 1 时接收到确实 0，或是当传输符号为 0 时接收到的符号却是 1)。那么这时在信道中的信息传输速率是多少？显然，它会小于每秒 1000 个二进制单位。因为在接收到的符号中大约有 1

显然对于应用传输信息的数量合适的修正是在接收信号中丢失的信息。或者当我们收到的信号是真正送出来的可选择的是不确定的。从我们先前关于熵的讨论中一个不确定的测量好像是很合理的使用有条件的信息熵，知道了接收的信号，作为丢失信息的量度，这个实际上是合适的定义，等我们会看到。这个思想后面是实际传输的比率， $R$ ，将会从产生条件熵的平均速率的比率中减去获得 (也就是，信源熵)。

$$R = H(x) - H_y(x)$$

这个条件熵  $H_y(x)$  方便一点可以叫做模糊理论。它是测量接收信号中的平均模糊信息。

在以上提到的信息中，假如 0 接收了后面的概率，那么 0 的传输是 0.99，并且 1 的传输是 0.01。假如 1 接收到的时候这些数字是相反的。因此

$$H_y(x) = -[.99 \log .99 + 0.01 \log 0.01] \\ = .081 \text{ 比特 / 符号}$$

或者 81 比特每秒。我们可以说系统是以  $1000-81=919$  比特每秒的速率传输。在极端情况下当 0 等于接收时的 0 或者 1 并且对于 1 也是一样的情况，它们的概率分别是  $\frac{1}{2}$ ， $\frac{1}{2}$  并且

$$H_y(x) = -\left[\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}\right] \\ = 1 \text{ 字节每符号}$$

或者 1000 字节每秒。传输速率就是 0。

下面的定理给出了关于模糊理论直接解释，同样证明了它是唯一的一个测量方法。我们认为一个通信系统和一个观测者(或者一个辅助设备)可以看到什么是发出的和什么是接收到的(由噪声产生的错误)。这个观测者注意到了在信息和传输数据复原时产生的错误在接收端有一个“改良的信道”使接收装置可以改正错误。这种情况在图 8 中说明。

定理 10：假如改良的信道有  $H_y(x)$  的容量那么就可以经过编码纠正数据把它送到信道中并且改正任意小的错误  $\epsilon$ 。假如信道容量小于  $H_y(x)$  这个是不可能的。

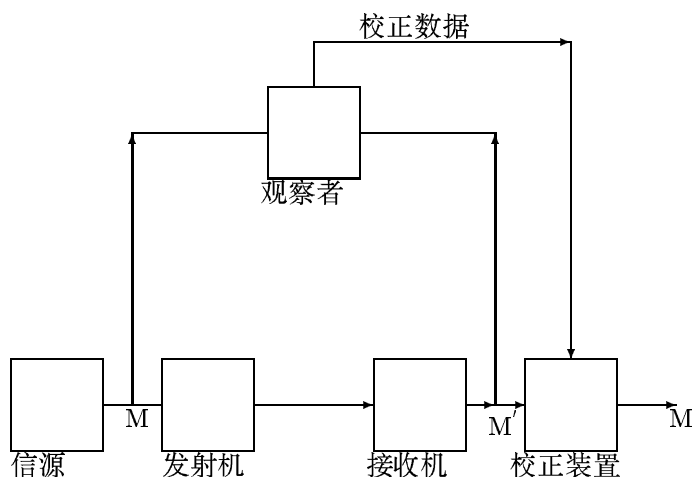


图 8- 校正系统示意图

大概来说， $H_y(x)$  是额外信息的总数所以必须在接收装置上来改正接收信息以每秒来供应。

为了证明第一部分，认为接收信息  $M'$  和原始通信信息  $M$  的长序列。这里将会有  $M's$  的对数  $TH_y(x)$  可以合理的产生每个  $M'$ 。因此我们有  $TH_y(x)$  二进制数字来发送每个  $T$  秒。这个将会在一个有  $H_y(x)$  的信道上带有错误频率  $\epsilon$  来完成。

第二部分不需要什么就可以证明，首先对于任意的离散机会变量  $x,y,z$

$$H_y(x, z) \geq H_y(x).$$

有

$$H_y(z) + H_{yz}(x) \geq H_y(x) \\ H_{yz}(x) \geq H_y(x) - H_y(z) \geq H_y(x) - H(z)$$

假如我们在改进的信道中把  $x$  当作信源的输出，把  $y$  当作接收信号把  $z$  当作发射信号，然后右边是比传输速率精确在校正信道上。假如信道的容量比右边模糊理论少，那么将会比 0 大并且  $H_{yz}(x) > 0$ 。但是发送

的是不确定的，知道了接收信号和改正的信号。假如这个比 0 大错误的频率将会任意小。

举个例子：

假设在一连串二进制数字中错误发生是随机的：概率  $p$  是发生错误的概率， $q = 1 - p$  是正确的概率。假如位置已知时错误会被纠正。因此正确的信道只需要发送信息到这些位置上。从信源中转移的数量产生了概率为  $p$  的二进制数字，是 1(错误)， $q$  是 0(正确)。于是校正信道所必要的信道容量为：

$$- [p \log p + q \log q]$$

它就是原来系统的模棱性。

传输  $R$  的比率可以用上面两种另外的恒等式来注解。我们有

$$\begin{aligned} R &= H(x) - H_y(x) \\ &= H(y) - H_x(y) \\ &= H(x) + H(y) - H(x, y). \end{aligned}$$

首先定义的表达已经可以被发送的信息所解释，少于不确定所发送的。其次是测量所接收到的哪个是噪声。第三步是两个数量的总和共同的熵因此在某种意义上比特每秒对于以上两个所共同的。因此这三种表达有着一定直觉的意义。

噪声信道中  $C$  的容量应该是传输比率中的可能最大值，也就是，当信源和信道相匹配时的比率。我们因此定义信道容量

$$C = \text{Max}(H(x) - H_y(x))$$

考虑到信源信息可能的最大值是使用到信道的输入。假如信道是无噪的， $H_y(x) = 0$ 。这个定义相当于已经给予无噪信道因为信道的最大熵是它的容量。

### 13. 离散有噪信道的基本原理

这个好像很令人感到奇怪我们本应该为有噪信道定义一个  $C$  容量因为没有把信息送到这种情况中。很清楚的是，然而，通过发送信息以一种冗余形式可以减少错误发生的概率。举个例子，通过多次重复信息和通过接收信息版本的不同统计研究错误发生概率将会很小。就像所预期的，然而，错误发生的概率接近于 0。编码冗余将会不确定地增加。传输速率因此接近于 0。这个毫无疑问是正确的。假如它是的话，将不会有好的定义容量。但是仅仅是给定错误频率的容量。或者是一个模糊的。当错误增加的时候容量将下降。实际上以上定义的容量  $C$  有一个非常明确的意义。通过合适的编码很有可能以一种错误频率很小或者所需的模糊的信道比率  $C$  来传送信息。这种情形只能对  $C$  适用。假设尝试比  $C$  的速率还大来传，是  $C + R_1$ ，然后将会有必要一种模糊等于或者高于过量的  $R_1$ 。自然的通过获取的花费是不确定的，以至于我们通过修正实际上并没有获得比  $C$  大。

这种情形在图 9 中说明。在信道中信息的比率是水平轴，模糊值是垂直轴。在这根粗线的上面的点阴影区域可以获得而在线的下方则不能。而在线上的点一般不能获得，但是线上如果有两点的话可以获得。

这些结论是对于  $C$  的解释，将会在下面证明。

定理 11：让一个离散信道有一个容量  $C$ ，一个离散信源熵是  $H$ 。假如  $H \leq C$  存在一个编码系统信源的输出可以通过任意小错误频率的信道传输 (或者一个任意小的模糊值)。假如  $H > C$  可以对信源进行编码以至于模糊值小于  $H - C + \epsilon$  当  $\epsilon$  是个任意小量。当模糊值小于  $H - C$  时不能编码。

证明定理第一部分的方法将不再展现所需特性的编码方法，但是通过展示这种编码必须存在于特定的编码组中。事实上我们把错误的频率在组上平均并且显示这个平均值小于  $\epsilon$ 。假如这而数集中的平均数比  $\epsilon$

小至少存在一个比  $\epsilon$  小的数。这个将会建立所需的结果。

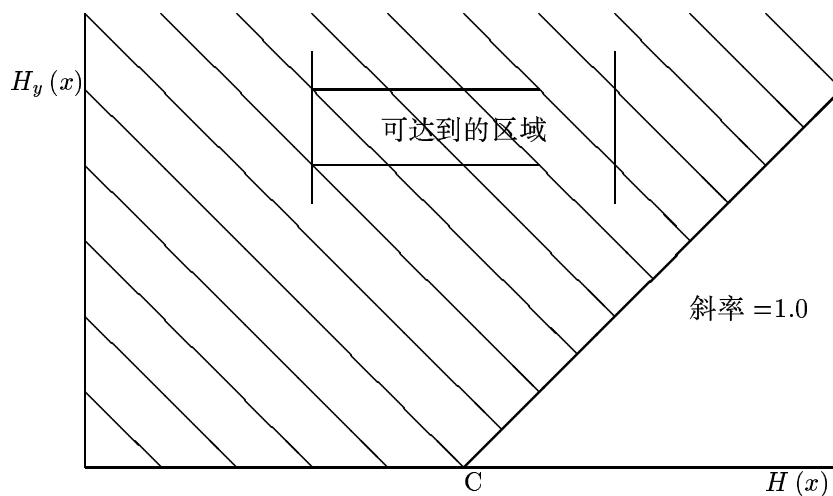


图 9- 已给的信道输入熵的可能模糊值

有噪信道 C 的容量已经定义为

$$C = \text{Max} (H(x) - H_y(x))$$

这里 x 是输入 y 是输出。信源的最大化可能被用作信道的输入。

让  $S_0$  作为信源达到 C 容量的最大值。假如最大值没有实际上达到任何信源，我们让  $S_0$  作为信源，接近于给出的最大比率。假设  $S_0$  是信源输入。我们认为可能的传输和接收的持续时间为 T 的序列中。下面将会成立：

1. 传送的顺序分为两种，高组频率  $2^{TH(x)}$  和剩余的小频率顺序。
- 2 相似地接收顺序有一个高频集  $2^{TH(y)}$  和一个剩余顺序的低频集。
3. 每个高频的输出通过  $2^{TH_y(x)}$  输入产生。而其他情形时是小概率。

所有  $\epsilon$ 's 和  $\delta$ 's 通过单词“小”和“关于”在当我们允许 T 增加和  $S_0$  接近最大信源这种情形时接近于 0。

这种情形在图 10 中表示，输入序列在左边，输出序列在右边表示。交叉线代表可能排列的典型输出排列。

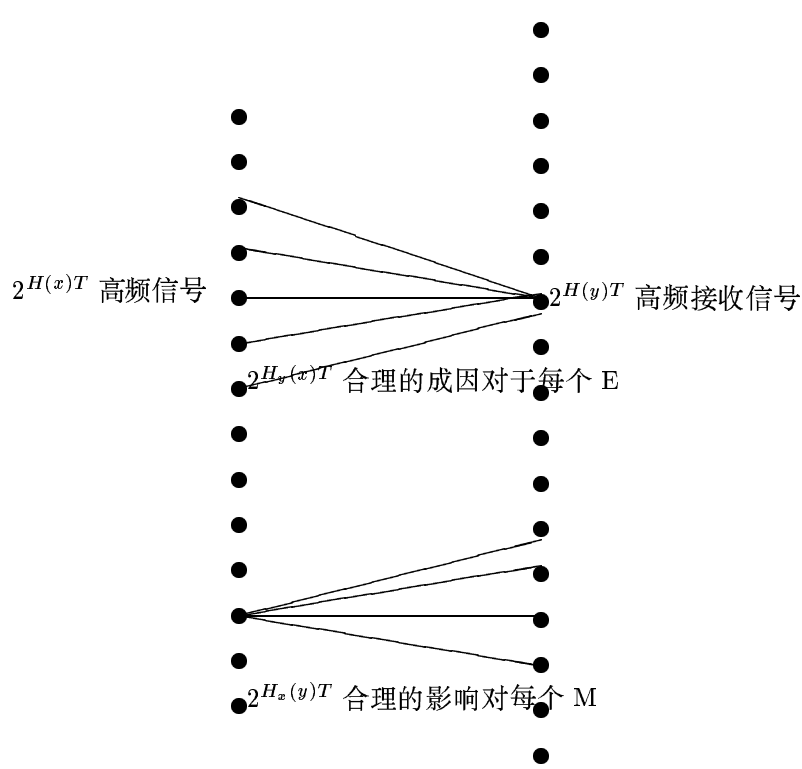


图 10- 信道中输入和输出关系的示意图

有另外一个信源它产生信息的速率是  $R$  并且  $R < C$ 。在时间  $T$  期间这个信源将会有  $2^{TR}$  高频信息。我们希望把它和有选择的可能的信道输入序列的选择联系起来, 以获得比较小的错误频率。我们将会用各种可能的方法建立一个联系 (然而通过信源  $S_0$  决定的只有高频输入组) 和对于庞大的编码系统错误频率的平均值。这个和信息输入信道的持续时间  $T$  的任意联合时计算错误的频率是一样的。假设一个特殊的输出信号  $y_1$  被观测。当不只一个信号在集合里可能形成  $y_1$  的概率是多少? 有  $2^{TR}$  个信息在任意  $2^{TH(x)}$  个点上分配。一个特殊点成为一条信息的概率是

$$2^{T(R-H(x))}$$

没有一点在扇形区的信息概率 (除了真正的原始信息以外) 是

$$P = [1 - 2^{T(R-H(x))}]^{2^{TH_y(x)}}$$

现在  $R < H(x) - H_y(x)$  所以  $R - H(x) = -H_y(x) - \eta$  并且  $\eta$  是正的。结论

$$P = [1 - 2^{-TH_y(x) - T\eta}]^{2^{TH_y(x)}}$$

接近 (当  $T \rightarrow \infty$ )

$$1 - 2^{-T\eta}$$

因此错误接近于 0 的概率和第一部分的定理证明了。

通过注意到我们可以从信源中仅仅发送  $C$  比特每秒, 定理的第二部分是很容易展现的, 完全忽略了信息产生的余数。在接收器上忽略的部分给出了一个模糊值  $H(x) - C$  传输的部分仅仅需要增加  $\epsilon A! \#$  这种限制可以用其他许多方法获得, 当我们认为持续的情形下将会展现。

定理的最后情况是  $C$  定义的一个简单结论。假设我们给  $H(x) = C + a$  的信源用这种方法进行编码以获得一个模糊值  $H_y(x) = a - \epsilon$ ,  $\epsilon$  是正的。然后  $R = H(x) = C + a$  并且

$$H(x) - H_y(x) = C + \epsilon$$

并且  $\epsilon$  是正的。当最大值是  $H(x) - H_y(x)$  时这和  $C$  的定义是相矛盾的。

实际上在定理中证明比声明更多。假如一个数集的平均数中  $\epsilon$  是最大值,  $\sqrt{\epsilon}$  中的部分比在最大以下的  $\sqrt{\epsilon}$  多。因为  $\epsilon$  是任意小的我们可以说系统是无穷接近理想状态。

#### 14. 讨论

定理 11 的证明, 并不是一个纯的证明形式, 这些证明有些不足。通过下面的证明方法获得一个好的接近理

想编码的是不切实际的。实际上，除了一些细小的情况和限制情形，还没有一系列接近理想的清楚的描述被发现。可能没有偶然时间但是关系到对于一个好的接近任意序列给出的清楚的构造是有困难的。

接近于理想的编码将有这样的性质：如果信号被噪声以一定的方式改变，那么原始信号将会恢复。换句话说这个改变一般不会像原始信号一样更接近于另一个理想信号，而仍然会更接近于原先那个信号。这一点是由于在编码中引入了某些多余度而得到的。多余度应以适当的方式加入，使它有利于抵抗具有特定结构的噪声在信道中的作用。只要在接收机中能够得到利用，那么任何信源的多余度都是有用的。特别是，如果信源已具有某些多余度，而且也不采用与信道匹配的方法来消除它，那么这种多余度将有利于噪声的克服。例如，在无噪声电报信道中，如果采用合理的编码，就可以节省大约 50% 的时间。但是没有这样做，英语的大部分多余度仍保留在信道符号中。因此它有一个优点：在传输时，信道中能允许存在很强的噪声，相当大的一部分接收的错误字母仍能根据文章间的相互联系而得到重新恢复。实际上在很多情况下这也许还是比较接近于理想的，因为英文的统计结构是比较复杂的，而合理的英语序列与随机选择（在定理要求的意义上）相差还不太远。

当在无噪情况下延迟一般需求接近于理想编码，现在有一个额外的功能允许大量的噪声在接收原始信息的接收点做任何判断之前影响信号。增加样本尺寸使统计声明更加锐化。

定理 11 的内容及其证明可以用另一种不同的方法简洁的陈述，这种方法能使它与无噪声情况的关系更为明显。我们讨论长度为  $T$  的可能信号，并假设选用它的子集。令子集中的所有信号都以等概率选用。并假设接收机在接收到被干扰的信号时，能选子集中最大可能的信号作为原先的信号。我们定义  $N(T, q)$  为我们能为子集选择的信号数目使得在这个最大数目下，错误复现概率小于或等于  $q$ 。

定理 12：  $\lim_{T \rightarrow \infty} \frac{\log N(T, q)}{T} = C, C$  信道容量，  $q$  不等于 0 或 1。

换句话说，不管我们如何地规定可靠性的极限，当  $T$  足够大时。我们可以在时间  $T$  内可靠地区别相当于  $CT$  个二进单位的消息。定理 12 可以和部分 1 给出的无噪信道容量的定义相比较

#### 15. 离散信道容量的例子

离散信道的简单例子在图 11 给出。这里有三种可能的符号。第一个不会被噪声干扰。第二个和第三个不被干扰的概率为  $p$ ，并且  $q$  变成了另外的一对。我们有（让  $a = -[p \log p + q \log q]$ ， $P$  和  $Q$  是第一种和第二种符号的概率）

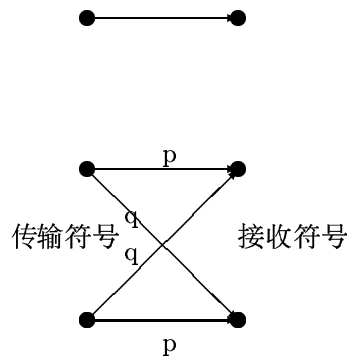


图 11 — 离散信道举例

$$H(x) = -P \log P - 2Q \log Q$$

$$H_y(x) = 2Qa$$

我们希望用这种方法选择  $P$  和  $Q$  使  $H(x) - H_y(x)$  最小化, 以  $P+2Q=1$  为条件。因此我们认为

$$\begin{aligned} U &= -P \log P - 2Q \log Q - 2Qa + \lambda(P + 2Q) \\ \frac{\partial U}{\partial P} &= -1 - \log P + \lambda = 0 \\ \frac{\partial U}{\partial Q} &= -2 - 2 \log Q - 2a + 2\lambda = 0 \end{aligned}$$

除去  $\lambda$

$$\begin{aligned} \log P &= \log Q + a \\ P &= Qe^a = Q\beta \\ P &= \frac{\beta}{\beta+2}, Q = \frac{1}{\beta+2} \end{aligned}$$

于是信道容量为:

$$C = \log \frac{\beta+2}{\beta}$$

$\beta=1$ , 则  $C=C = \log 3$ , 这是正确的, 因为信道对三个可能的符号都没有噪声干扰。如果  $p=\frac{1}{2}$ ,  $\beta=2$ , 则  $C=C = \log 2$ 。第二个和第三个符号就完全不能区分而作为一个符号。第一个符号的概率为  $p=\frac{1}{2}$ , 而第二个和第三个一起的概率为  $\frac{1}{2}$ 。不论它们之间采用任何所需方式来分配, 信道容量仍然达到最大值。

$p$  为中间值时, 信道容量将处在  $\log 2$  与  $\log 3$  之间。这时第二个和第三个符号间的区别带有某些信息量, 但没有象在无噪声情况下那样多。第一个符号因为它无噪声干扰, 因此比其他两个符号更常用。

#### 16. 某些特殊情况下的信道容量

如果噪声独立地干扰前后的信道符号, 那么它能够用转移概率  $p_{ij}$  来描述。  $p_{ij}$  是当发送符号  $i$ , 接收到符号  $j$  的概率。于是信道容量的最大值为

$$-\sum_{i,j} p_i p_{ij} \log \sum_i p_i p_{ij} + \sum_{i,j} p_i p_{ij} \log p_{ij}$$

其中改变  $p_i$  使满足  $\sum p_i = 1$ , 用拉格朗日法可导出下式:

$$\sum_j p_{sj} \log \frac{p_{sj}}{\sum_i p_i p_{ij}} = u \quad s=1,2,\dots$$

乘以  $p_s$ , 并对  $s$  求和, 可证得  $u=-C$ 。令  $p_{sj}$  的倒数 (如果它存在的话) 为  $h_{st}$ , 则  $\sum_s h_{st} p_{sj} = \delta_{tj}$ , 于是:

$$\sum_{s,j} h_{st} p_{sj} \log p_{sj} - \log \sum_i p_i p_{it} = -C \sum_s h_{st}$$

故:

$$\sum_i p_i p_{it} = \exp \left[ C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj} \right]$$

或:

$$P_i = \sum_t h_{it} \exp \left[ C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj} \right]$$

这些方程组就是确定  $C$  为最大值的那些  $p_i$ , 其中  $C$  确定于  $\sum p_i = 1$ 。当这样做后,  $C$  就是信道容量, 且  $p_i$  是得到这个容量时信道符号的概率。

如果每个输入符号在从它发出的线上有同样的概率集, 并且每个输出符号也如此, 则信道容量很容易算出。图 12 为其实例。

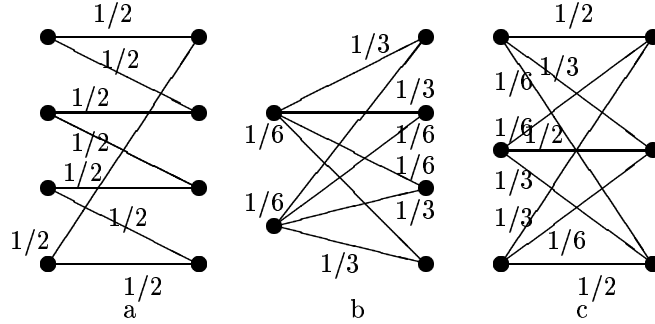


图 12. 作为每个输入和输出具有同样条件概率的离散信道例子

在这个情况下,  $H_x(y)$  与输入符号的概率分布无关, 并等于  $-\sum p_i \log p_i$ , 其中  $p_i$  是从任何输入符号得到的转移概率值。则信道容量为:

$$\text{Max}[H(y) - H_z(y)] = \text{Max} H(y) + \sum p_i \log p_i$$

$H(y)$  的最大值显然是  $\log m$ , 其中  $m$  是输出符号的个数。这是因为有可能通过使输入符号等概率而使输出符号构成等概率。故信道容量为:

$$C = \log m + \sum p_i \log p_i$$

图 12a 中, 它将等于:

$$C = \log 4 - \log 2 = \log 2$$

这可以只采用第一个和第三个符号来得到。

图 12b 中, 有:

$$\begin{aligned} C &= \log 4 - \frac{2}{3} \log 3 - \frac{1}{3} \log 6 \\ &= \log 4 - \log 3 - \frac{1}{3} \log 2 \\ &= \log \frac{1}{3} 2^{\frac{5}{3}}. \end{aligned}$$

图 12c 中, 有:

$$\begin{aligned} C &= \log 3 - \frac{1}{2} \log 2 - \frac{1}{3} \log 3 - \frac{1}{6} \log 6 \\ &= \log \frac{3}{2^{\frac{1}{2}} 3^{\frac{2}{3}} 6^{\frac{1}{6}}} \end{aligned}$$

假设这些符号分为这样几组, 噪声绝不至于使一组的符号被错认为另一组的符号。当我们只用这一组符号时, 可令第  $n$  组的容量为  $C_n$  (每秒多少个二进制单位), 容易证明, 为了最好地使用整个集合, 则在第  $n$  组中所有符号的总概率  $p_n$  应为

$$p_n = \frac{2^{C_n}}{\sum 2^{C_n}}$$

在一组内概率的分布恰如只利用这些符号一样。信道容量为

$$C = \log \sum 2^{C_n}$$

#### 17. 一个有效编码的例子

下面的例子是有可能对噪声信道正确匹配的例子。设有两种信道符号 0 和 1, 噪声作用于七个符号组成的群。这由七个符号组成的群或者无误差地得到传输, 或者, 有一个符号是错误的。八个概率完全相等。于是有:

$$\begin{aligned} C &= \text{Max}[H(y) - H_x(y)] \\ &= \frac{1}{7} \left[ 7 + \frac{8}{8} \log \frac{1}{8} \right] \end{aligned}$$



$$= \frac{4}{7} \text{ 二进单位 / 符号}$$

一个完全能校正误差并以速率  $c$  传送的有效码可用下列方法求得 (这个方法是 R.Hamming 找到的)。

令七个符号组成的群为:

$$X_1, X_2, X_3, \dots, X_7$$

其中  $X_3, X_5, X_6, X_7$  是信源中任意取出的消息符号, 其他三个符号是多余的。

选择  $X_4$  使  $a = X_4 + X_5 + X_6 + X_7$  为偶数

选择  $X_2$  使  $\beta = X_2 + X_3 + X_6 + X_7$  为偶数

选择  $X_1$  使  $\gamma = X_1 + X_3 + X_5 + X_7$  为偶数

当收到这个符号群后, 可以计算  $\alpha$ 、 $\beta$  和  $\gamma$ , 如果是偶数叫做 0, 如果是奇数叫做 1, 那么二进制数字  $\alpha$ 、 $\beta$  和  $\gamma$  将给出错误的  $X_i$  的下标 (如果下标为 0 就表示无误差)。

#### 附录 1

一些具有有限状态的符号区组数的增长

令  $N_j(L)$  是长度为  $L$ , 末状态为  $i$  的符号组的数目。则

$$N_j(L) = \sum_{i,s} N_i \left( L - b_{ij}^{(s)} \right)$$

其中  $b_{ij}^1, b_{ij}^2, \dots, b_{ij}^m$  是一些从状态  $i$  选出并引向状态  $j$  的符号的长度。这些是线性差分方程式, 当  $L \rightarrow \infty$  时必为

$$N_j = A_j W^L$$

形式, 把它代入差分方程式

$$A_j W^L = \sum_{i,s} A_i W^{L - b_{ij}^{(s)}}$$

或

$$A_j = \sum_{i,s} A_i W^{-b_{ij}^{(s)}} \\ \sum_i \left( \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right) A_i = 0$$

为使上式成立, 则行列式必须为零, 即:

$$D(W) = |a_{ij}| = \left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

由此可以解得  $W$ , 当然它等于  $D=0$  的最大实数根。

因此, 量  $C$  为:

$$C = \lim_{L \rightarrow \infty} \frac{\log \sum A_j W^L}{L} = \log W$$

并应指出, 如果我们要求所有群 (区组) 都以同样状态开始 (可以任意选择), 那么所得增长特性也是一样的。

#### 附录 2

公式  $H = -\sum p_i \log p_i$  的推导

令  $H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = A(n)$ 。有条件 (3), 我们可以把一个从  $s^m$  个等可能性中进行的选择, 分解为  $m$  个从  $s$  个等可能性中进行的选择, 得:

$$A(s^m) = mA(s)$$

同样

$$A(t^n) = nA(t)$$

我们可以任意选取  $n$  的大小并可以找到一个  $m$  满足

$$s^m \leq t^n < s^{m+1}$$

因此, 考虑对数并除以  $n \log s$

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \text{ 或者 } \left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon$$

其中  $\epsilon$  可以任意小。那么由  $A(n)$  的单调性可得,

$$\begin{aligned} A(s^m) &\leq A(t^n) \leq A(s^{m+1}) \\ mA(s^m) &\leq nA(t) \leq (m+1)A(s) \end{aligned}$$

因此, 除以  $nA(s)$ ,

$$\begin{aligned} \frac{m}{n} &\leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \text{ 或者 } \left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \epsilon \\ \left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| &< 2\epsilon \quad A(t) = K \log t \end{aligned}$$

其中  $K$  必须是正的, 以满足 (2).

假定可以从  $n$  个可能值中选取一个, 它们的可比概率为  $p_i = \frac{n_i}{\sum n_i}$  其中  $n_i$  是整数。也可以不从  $\sum n_i$  个值中选而从具有概率的  $n$  个值中选, 那么, 如果第  $i$  个值被选定, 这  $n_i$  个值具有相同的概率。使用条件 (3), 我们可以采用两种计算方法使从  $\sum n_i$  中选择的值相等。

$$K \log \sum n_i = H(p_1, \dots, p_n) + K \sum p_i \log n_i$$

因此

$$\begin{aligned} H &= K \left[ \sum p_i \log \sum n_i - \sum p_i \log n_i \right] \\ &= -K \sum p_i \log \sum \frac{n_i}{\sum n_i} = -K \sum p_i \log p_i \end{aligned}$$

如果  $p_i$  不可比, 它们有可能是有理数, 在连续性假设下表达式是一样的。因此这表达式是成立的。系数  $K$  的选择是基于测量的方便和数量大小的。

### 附录 3

#### 遍历定理来源

如果能够从任何状态  $p > 0$  沿着概率  $p > 0$  的路径到任何其它状态, 该系统遍历可以应用大数定律。因此网络中一个给定路径  $p_{ij}$  的次数是一个长为  $N$  的序列, 正比于  $i$  处的概率, 比如  $p_i$ , 然后选择这条路径,  $p_i p_{ij} N$ 。如果  $N$  足够大, 概率百分比误差  $\delta$  小于  $\epsilon$  因此除了小概率集对所有状态实际的序列长度是在限制范围内的

$$(p_i p_{ij} \pm \delta) N$$

因此几乎所有的序列具有概率  $p$ ，计算如下

$$p = \prod p_{ij}^{(p_i p_{ij} + \delta)N}$$

并且  $\frac{\log p}{N}$  是有限制的

$$\frac{\log p}{N} = \sum (p_i p_{ij} \pm \delta) \log p_{ij}$$

或者

$$\left| \frac{\log p}{N} - \sum p_i p_{ij} \log p_{ij} \right| < \eta$$

这证明定理 3。

基于定理 3 中  $p$  的可能值计算出  $n(q)$  的上下界就可以得到定理 4。

在混合（不遍历）的情况下，如果

$$L = \sum p_i L_i$$

并且组成元素的熵满足  $H_1 \geq H_2 \geq \dots \geq H_n$  可以得到

定理：  $\lim_{N \rightarrow \infty} \frac{\log(q)}{N} = \varphi(q)$  是一个单调递减的函数，

$$\varphi(q) = H_s \text{ 在间隔 } \sum_1^{s-1} \alpha_i < q < \sum_1^s \alpha_i.$$

要证明定理 5 和定理 6，首先注意到  $F_N$  是单调递减的函数是因为增加  $N$  就增加了条件熵的下标。可用  $p_{B_i}(s_j)$  代入，则

$$F_N = NG_N - (N-1)G_{N-1}$$

对所有的  $N$  求和，则  $G_N = \frac{1}{N} \sum F_n$ 。因此  $G_N \geq F_N$  并且  $G_N$  是单调递减的。当然它们应该达到相同的极限。应用定理 3 我们可以得到  $\lim_{N \rightarrow \infty} G_N = H$ 。

#### 附录 4

##### 受限信道的最大信息率

假如对符号序列有一系列的限制，因为这些序列是有限状态型的可由线状图表示。设  $\ell_{ij}^{(s)}$  是从状态  $i$  到状态  $j$  各种符号的长度，由于选择的符号  $s$  是在状态  $i$  到状态  $j$  在这种限制下增大了信息发生率，那么不同状态的概率  $p_i$  和  $p_{ij}^{(s)}$  的概率分布是怎样的？限制条件确定了一个离散信道，最大信息率必须小于等于信道容量  $C$ ，如果所有较长区段相匹配，那么这个速率可以达到，并且有可能这是最优的。通过选择合适的  $p_i$  和  $p_{ij}^{(s)}$  这个速率可以达到。

信息速率等式为

$$\frac{-\sum p_i p_{ij}^s \log p_{ij}^s}{\sum p_i p_{ij}^s \ell_{ij}^s} = \frac{N}{M}$$

设  $\ell_{ij} = \sum_s \ell_{ij}^{(s)}$ . 显然最大值  $p_{ij}^{(s)} = K \exp \ell_{ij}^{(s)}$ . 最大化的限制是  $\sum p_i = 1, \sum_j p_{ij} = 1, \sum p_i (p_{ij} - \delta_{ij}) = 0$ . 因此我们最大化

$$U = \frac{-\sum p_i p_{ij} \log p_{ij}}{\sum p_i p_{ij} \ell_{ij}} + \lambda \sum_i p_i + \sum \mu_i p_{ij} + \sum \eta_j p_i (p_{ij} - \delta_{ij})$$

$$\frac{\partial U}{\partial p_{ij}} = \frac{M p_i (1 + \log p_{ij}) + N p_i \ell_{ij}}{M^2} + \lambda + \mu_i + \eta_j p_i = 0$$

求解  $p_{ij}$

$$p_{ij} = A_i B_j D^{-\ell_{ij}}$$

因为

$$\sum_j p_{ij} = 1, A_i^{-1} = \sum_j B_j D^{-\ell_{ij}}$$

$$p_{ij} = \frac{B_j D^{-\ell_{ij}}}{\sum_s B_s D^{-\ell_{is}}}$$

$D$  最合适的值是信道容量  $C$  并且  $B_j$  是下式的解

$$B_i = \sum B_j C^{-\ell_{ij}}$$

因此

$$p_{ij} = \frac{B_j}{B_i} C^{-\ell_{ij}}$$

$$\sum p_i \frac{B_j}{B_i} C^{-\ell_{ij}} = p_j$$

或者

$$\sum \frac{p_i}{B_i} C^{-\ell_{ij}} = \frac{p_j}{B_j}$$

因此, 如果  $\lambda_i$  满足

$$\sum \gamma_i C^{-\ell_{ij}} = \gamma_j$$

$$p_i = B_i \gamma_i$$

关于  $B_i$  和  $\gamma_i$  的方程组是可以满足的因为  $C$  是这样的

$$|C^{-\ell_{ij}} - \delta_{ij}| = 0$$

在这种情况下, 速率为

$$-\frac{\sum p_i p_{ij} \log \frac{B_j}{B_i} C^{-\ell_{ij}}}{\sum p_i p_{ij} \ell_{ij}} = C - \frac{\sum p_i p_{ij} \log \frac{B_j}{B_i}}{\sum p_i p_{ij} \ell_{ij}}$$

但是

$$\sum p_i p_{ij} (\log B_j - \log B_i) = \sum_j p_j \log B_j - \sum_i p_i \log B_i = 0$$

因此速率是  $C$  并且不能超过它，这是最大值，证明了假定的解。

### 第三章 连续信息

最后考虑如下情况：信号或者信息或者二者全为连续变量与以前假设为离散变量作比较。在一定程度上连续的状态可以由离散状态通过一个有限处理过程得到，把信号和信息的统一体分为有限个小的区域并计算大量离散参数。随着区域的不断减小，这些参数以一定的方式限定了连续状态的值。然而表现出了新的功能并且不强调特定状态需具有特定结果。

在连续状态下，我们不期望获得数学推理严谨且最优的解，因为这需要大量的抽象测量理论并且会忽略分析重点。然而，初步研究表明，理论可以以十分公理化严谨的方式公式化，它包含了连续，离散还有其它变量。在目前的分析中，有限过程的随机选取在所有的实际应用中可以被证实。

#### 18. 集合和函数体

我们必须采用函数集和函数体来处理连续变量。函数集，是一类函数的集合，一般具有一个变量、时间。集合中的函数可以通过明确的符号表示或者是给定一个隐含的区别与其它函数的特性。下面是一些例子：

##### 1. 函数集：

$$f_\theta(t) = \sin(t + \theta)$$

每个特定的  $\theta$  值对应着集合中的一个函数。

2. 时间函数集的频率每秒  $W$  个周期。
3. 函数集的带宽被限制为  $W$ ，频谱为  $A$ 。
4. 集合中的英文符号为时间函数。

一个函数体是具有概率量的函数的集合<sup>1</sup>，通过概率度量可以决定具有特定功能的函数在集合中占的概率，举个例子：

<sup>1</sup> 在数学术语中，函数属于尺度空间，这个空间的所有尺度是一个集合。

$$f_\theta(t) = \sin(t + \theta)$$

我们可以给  $\theta$  指定一个概率分布  $p(\theta)$ 。那么这个集合就变为了函数体。

更多关于函数体的例子如下：

1. 一个有限函数集  $f_k(t) (k = 1, 2, \dots, n), f_k$  的概率为  $p_k$ 。
2. 一个有限的函数空间族

$$f(\alpha_1, \alpha_2, \dots, \alpha_n; t)$$

关于变量  $\alpha_i$  的概率分布为:

$$p(\alpha_1, \alpha_2, \dots, \alpha_n)$$

我们可以认为函数体定义为:

$$f(\alpha_1, \dots, \alpha_n, \theta_1, \dots, \theta_n; t) = \sum_{i=1}^n \alpha_i \sin i(\omega t + \theta)$$

幅度  $\alpha_i$  是均匀分布并且相互独立, 相角  $\theta_i$  是非均匀分布 (从 0 到  $2\pi$ ) 并且相互独立。

### 3. 函数体

$$f(\alpha_i, t) = \sum_{n=-\infty}^{+\infty} \alpha_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

$\alpha_i$  均匀分布并相互独立且具有相同的标准偏差  $\sqrt{N}$ 。这是白噪声的概率分布, 带宽被限制到 0 到  $W$ , 平均功率为  $N^2$ 。

4. 设点在  $t$  轴上服从泊松分布, 每一个点处的函数是确定的, 不同的函数相加就形成了函数体

$$\sum_{k=-\infty}^{+\infty} f(t + t_k)$$

其中  $t_k$  是泊松分布的点。这个函数体可以认为是一种脉冲或是散弹噪声, 其中所有的脉冲都是一样的。

5. 具有概率尺度的英语语音函数集是由日常应用中的发生频率产生的。

如果所有的函数平移固定的时间, 函数体不变, 那么函数  $f_\alpha(t)$  的函数体是静止的。函数体

$$f_\theta(t) = \sin(t + \theta)$$

是静止的如果  $\theta$  在 0 到  $2\pi$  上服从均匀分布。如果对每个函数平移  $t_1$ , 那么就可以得到

<sup>2</sup> 这种表示可以认为是对带限白噪声的一种定义。它具有一定的优势, 因为与过去采用的定义方法相比它涉及较少的限制操作。“白噪声”这个名字已经牢牢扎根于文献, 也许有点遗憾。光学中, 白光是指要么区别于点状谱的谱, 要么是具有很长的平坦谱。(与具有平坦频率的谱不一样)。

$$\begin{aligned} f(t + t_1) &= \sin(t + t_1 + \theta) \\ &= \sin(t + \varphi) \end{aligned}$$

其中  $\varphi$  在 0 到  $2\pi$  上服从均匀分布。每个函数都改变但是函数体作为一个整体不变。上面给出的其它例子也是静止的。

如果一个函数体是静止的那么它是各态遍历的, 并且静止的函数体任何子集的概率都不超出 0 到 1。函数体

$$\sin(t + \theta)$$

是各态遍历的。这个函数的任何子集的概率  $\neq 0,1$ , 在任何时间的变换下它都可以转换为它本身。另一方面, 函数体

$$a \sin(t + \theta)$$

其中  $a$  是均匀分布并且  $\theta$  是静止的但这个函数体不是遍历的。这个函数的子集如果  $a$  属于 0 到 1 就是静止的。

上面给出的例子, 3 和 4 是遍历的, 5 也有可能是。如果一个函数体是遍历的, 那么我们可以粗略的认为集合中的每个函数都是函数体的典型。更确切的说一个遍历函数体的统计平均与特定函数集的时间平均是相等的 (概率是 1) <sup>3</sup>。一般说来每个函数都是可以预测的, 随着时间的改变, 以一定的频率, 遍历集合中的所有函数。

正如我们可以对数字和函数实施各种操作获得新的数字和函数, 我们也可以对函数实施操作来获得新的函数体。举个例子, 假如有一个关于函数  $f_\alpha(t)$  的函数体和一个操作  $T$  它能使每个函数  $f_\alpha(t)$  产生一个函数  $g_\alpha(t)$ :

$$g_\alpha(t) = T f_\alpha(t)$$

$g_\alpha(t)$  的概率测度定义与  $f_\alpha(t)$  的概率测度定义方法一样。  $g_\alpha(t)$  函数即一个确定子集的概率与在  $T$  操作下产生  $g$  函数子集的  $f_\alpha(t)$  函数的子集的概率相等。物理上这相当于通过一些设备穿过函数体, 比如, 滤波器, 整流器或调制器。设备输出的函数就来自函数体  $g_\alpha(t)$ 。

一个设备或操作  $T$  可被称为线性的, 如果改变输入也就改变了输出, 例如, 如果

3 这是著名的遍历定理的一方面, 这个定理曾用不同的方式得到证明, 例如由贝克荷夫、冯诺依曼和柯普曼以及最后由维纳、荷夫、胡利维茨等人所推广第定理。遍历定理的文献是很广泛的, 读者可以参看一些具有精确和普遍公式的书, 例如, E. Hopf, “Ergodentheorie,” *Ergebnisse der Mathematik und ihrer Grenzgebiete*, v. 5; “On Causality Statistics and Probability,” *Journal of Mathematics and Physics*, v. XIII, No. 1, 1934; N. Wiener, “The Ergodic Theorem,” *Duke Mathematical Journal*, v. 5, 1939.

$$g_\alpha(t) = T f_\alpha(t)$$

也就是

$$g_\alpha(t + t_1) = T f_\alpha(t + t_1)$$

对所有的  $f_\alpha(t)$  和所有的  $t_1$  都成立。它表明 (看附录 5) 如果  $T$  是线性的并且输入函数体是静止的那么输出函数体也是静止的。同样, 如果输入是遍历的那么输出也是遍历的。

滤波器和整流器在所有时间转变下都是线性的，调制器的操作不是线性的因为载波相位产生了一定的时间结构。尽管如此，调制器也可以是线性的当时间是载波周期的倍数时。

维纳指出物理设备的线性性与傅里叶定理有直接关系<sup>4</sup>。他指出，事实上，如果设备是线性的那么不变傅里叶分析是处理这个问题的最合适的数学工具。

一个函数体是对连续信息源（如，语音）产生的信息，发射器发出的信号以及扰动噪声最合适的数学表示。通信理论所关心的，正如维纳指出的，不是对特定函数的操作，而是对函数体的操作。一个通信系统不是为特定的语音函数设计的更谈不上为正弦波，但确实为语音函数体设计的。

## 19. 带限函数集

如果一个时间函数  $f(t)$  的带宽限制在 0 到  $2\pi$ ，它完全可以由间隔为  $\frac{1}{2W}$  秒的离散点上的坐标决定<sup>5</sup>。

定理 13：设  $f(t)$  的频率不超出  $W$ 。那么

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

其中

$$X_n = f(n/2W)$$

这个函数可以被认为限定在时间  $T$  内，如果所有的坐标  $X_n$  超出了这个时间间隔，那么这个函数为零。在这种情况下除了  $2TW$  坐标外，函数都为零。因此函数带宽限定为  $W$ ，持续时间为  $T$ ，对应与  $2TW$

<sup>4</sup> 通信理论的许多基本理念和理论是由维纳提出的。他的经典 NDRC 报告，插值，外推法和平滑的平稳时间序列（Wiley 出版社，1949 年），第一次明确指出通讯理论作为一个统计问题，研究时间序列。这本著作尽管主要关注线性预测和过滤问题，但却是现代论文的一个重要参考文献。我们也参考维纳的控制论（Wiley 出版社，1948 年），来处理通信和控制的一般问题。

<sup>5</sup> 为了进一步证明和讨论这个理论可以参考作者的论文“存在噪音的通信”出版于无线电工程学报，v. 37, No. 1, Jan., 1949, pp. 1021 扩张函数  $f(t)$  代表一系列正交函数。系数  $X_n$  可以认为是坐标为无穷维的“函数空间”。在这个空间中每个函数对应着一个确定的点，每个点对应着一个确定的函数。

空间的点。

函数子集的带宽和持续时间对应着空间的一块区域。例如，一个函数的能量小于或等于  $E$  对应空间  $2TW$  的球体，半径  $r = \sqrt{2WE}$ 。

一个限时限带的函数集可以由  $n$  维空间的概率分布  $p(X_1, \dots, X_n)$  表示。如果集合不是时间有限的，我们可以考虑用给定时间  $T$  的  $2WT$  坐标来代替间隔为  $T$  的函数部分并且概率分布  $p(X_1, \dots, X_n)$  给出了持续时间集合的统计结构。

## 20. 连续分布的熵

概率为  $p_1, \dots, p_n$  的离散集合的熵定义为：



$$H = - \sum p_i \log p_i$$

以同样的方式，我们可以定义具有密度分布函数  $p(x)$  的连续分布的熵：

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx$$

具有  $n$  维概率分布  $p(X_1, \dots, X_n)$  可得：

$$H = - \int \cdots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n) dx_1 \dots dx_n$$

如果有两个参数  $x$  和  $y$ （本身可能是多维的） $p(x, y)$  的联合和条件熵可由下式给出：

$$H(x, y) = - \int \int p(x, y) \log p(x, y) dx dy$$

并且

$$H_x(x, y) = - \int \int p(x, y) \log \frac{p(x, y)}{p(x)} dx dy$$

$$H_y(x, y) = - \int \int p(x, y) \log \frac{p(x, y)}{p(y)} dx dy$$

其中

$$p(x) = \int p(x, y) dy$$

$$p(y) = \int p(x, y) dx$$

连续分布熵具有大部分（不是全部）离散分布熵的特性，尤其我们可以得到下面几点：

1. 如果  $x$  在空间被限定为有限的体积  $\nu$ ，当  $p(x)$  是常数（ $1/\nu$ ），那么熵是最大的且等于  $\log \nu$
2. 对任何两个变量  $x, y$  我们可以得到

$$H(x, y) \leq H(x) + H(y)$$

同样，如果（且仅当） $x$  和  $y$  是相互独立，例如， $p(x, y) = p(x)p(y)$ （除了概率为零的集合）

3. 考虑如下的一个广义平稳过程：

$$p'(y) = \int a(x, y) p(x) dx$$

并且

$$\int a(x, y) p(x) dx = \int a(x, y) dy = 1, a(x, y) \geq 0.$$

那么平均分布  $p'(y)$  的熵大于或等于原始分布  $p(x)$  的熵。

4. 我们可以得到

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

并且

$$H_x(y) \leq H(y)$$

5. 设  $p(x)$  是一维分布,  $p(x)$  达到最大熵条件是  $x$  的标准差  $\sigma$  是高斯分布。为表明这一点, 我们必须最大化

$$H(x) = - \int p(x) \log p(x) dx$$

具有

$$\sigma^2 = \int p(x) x^2 dx. 1 = \int p(x) dx$$

作为制约因素。这就要求通过变量积分, 求最大值

$$\int [-p(x) \log p(x) + \lambda p(x) x^2 + \mu p(x)] dx$$

条件是

$$-1 - \log p(x) + \lambda x^2 + \mu = 0$$

因此, (调整常数, 以满足条件)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x^2}{2\sigma^2}\right)}$$

同样对于  $n$  维, 假定  $p(x_1, \dots, x_n)$  的二阶距为  $A_{ij}$ :

$$A_{ij} = \int \dots \int x_i x_j p(x_1, \dots, x_n) dx_1 \dots dx_n$$

这样最大熵产生了 (由一个类似的计算), 这时  $p(x_1, \dots, x_n)$  是  $n$  维高斯分布, 二阶距是  $A_{ij}$ 。

6. 标准差为  $\sigma$  的一维高斯分布的熵由下式给出

$$H(x) = \log \sqrt{2\pi e} \sigma$$

计算如下:

$$\begin{aligned}
 p(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x^2}{2\sigma^2}\right)} \\
 -\log p(x) &= \log \sqrt{2\pi}\sigma + \frac{x^2}{2\sigma^2} \\
 H &= - \int p(x) \log p(x) dx \\
 &= \int p(x) \log \sqrt{2\pi}\sigma dx + \int p(x) \frac{x^2}{2\sigma^2} dx \\
 &= \log \sqrt{2\pi}\sigma + \frac{\sigma^2}{2\sigma^2} \\
 &= \log \sqrt{2\pi}\sigma + \log \sqrt{e} \\
 &= \log \sqrt{2\pi e}\sigma
 \end{aligned}$$

同样 n 维高斯分布的相关二次型  $a_{ij}$  由下式给出

$$p(x_1, \dots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} \sum a_{ij} x_i x_j\right)$$

熵可以由下式计算

$$H = \log(2\pi e)^{\frac{n}{2}} |a_{ij}|^{-\frac{1}{2}}$$

其中  $|a_{ij}|$  是关键, 它的元素是  $a_{ij}$ 。

7. 如果 x 被限定为半线性的 (当  $x \leq 0$  时  $p(x) = 0$ ) 并且 x 的起始值限定为 a:

$$a = \int_0^{\infty} p(x) x dx$$

那么可以得到最大熵值, 当

$$p(x) = \frac{1}{a} e^{-\left(\frac{x}{a}\right)}$$

并且等于  $\log ea$ 。

8. 连续熵和离散熵具有一个重要的区别。在离散的情况下, 熵的度量是以一种绝对的方式, 随机性。在连续的情况下, 熵的度量与坐标系统有关。如果我们改变坐标熵会发生变化。实际上我们把坐标改为  $y_1, \dots, y_n$  新的熵可由下式给出

$$H(y) = \int \cdots \int p(x_1, \dots, x_n) J\left(\frac{x}{y}\right) \log p(x_1, \dots, x_n) J\left(\frac{x}{y}\right) dy_1 \cdots dy_n$$

其中  $J\left(\frac{x}{y}\right)$  雅可比坐标变换，扩大对数并把系数变为  $x_1, \dots, x_n$ ，我们可以得到：

$$H(y) = H(x) - \int \cdots \int p(x_1 \cdots, x_n) \log J\left(\frac{x}{y}\right) dx_1 \cdots dx_n$$

因此新熵是旧熵少于预期的雅可比对数。在连续情况下，熵可被视为衡量相对随机性假设的标准，即坐标系选择，每个小体积元  $dx_1 \cdots dx_n$  给予同样权重。当我们改变坐标系，新坐标系下的熵衡量随机性当体积元  $dy_1 \cdots dy_n$  具有同样权重系数。

依赖于坐标系统的熵的概念对连续情况和离散情况是同等重要的。这是由于信息率和信道容量的概念取决于两种不同的熵并且这种差异并不取决于坐标系，每个坐标都改变相同的量。

连续分布的熵可以是负值，根据每单位体积具有相同的分布，测量的范围可以任意设定零值。分布越局限，熵就越小并且还有可能是负的。然而，信息速率和信道容量不会是负的。

#### 9. 特定情况下坐标变换是线性变换

$$y_j = \sum_i a_{ij} x_i$$

在这种情况下雅可比仅仅是行列式  $|a_{ij}|^{-1}$  并且

$$H(y) = H(x) + \log |a_{ij}|$$

在坐标旋转的情况下（或其它改变方法） $J = 1$  并且  $H(y) = H(x)$ 。

#### 21. 函数集的熵

考虑一个带宽限于  $W$  的遍历函数集合。设

$$p(x_1, \dots, x_n)$$

是在  $n$  个采样点幅度  $x_1, \dots, x_n$  的密度分布函数。我们通过下式确定集合每自由度的熵：

$$H' = -\lim_{n \rightarrow \infty} \frac{1}{n} \int \cdots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n) dx_1 \cdots dx_n$$

我们还可以定义每秒的熵，不是由  $n$  来划分，而是由一定时间的  $n$  个样值。因为  $n = 2TW$ ， $H = 2WH'$ 。因为白噪声  $p$  是高斯分布并且

$$H' = \log \sqrt{2\pi e N}$$

$$H = W \log 2\pi e N \quad (76)$$

对于给定的平均功率  $N$ ，白噪声具有最大的熵。这是由上面提到的高斯分布的最大特性得到的。

连续随机过程 d 熵与离散过程有许多类似的性质。在离散的情况下，熵是与长序列的概率对数和数量有关。在连续情况下，同样与长序列样本的概率密度的对数和函数空间大概率的体积有关。

更确切的说，如果我们假定  $p(x_1, \dots, x_n)$  对于所有的  $n$  在  $x_i$  处连续，然后  $n$  足够大

$$\left| \frac{\log p}{n} - H \right| < \epsilon$$

对所有的选择  $(x_1, \dots, x_n)$  除了所有概率小于  $\delta$  的集合， $\delta$  和  $\epsilon$  任意小。如果我们把大的空间分成小的子空间那么遍历的特性就形成了。

$H$  与体积的关系可以描述为：在相同的假设下针对  $p(x_1, \dots, x_n)$  考虑  $n$  维空间。设  $V_n(p)$  是空间里的最小体积，在它内部包含了所有的概率  $q$ 。那么

$$\lim_{n \rightarrow \infty} \frac{\log V_n(p)}{n} = H'$$

假定  $q$  不等于 0 或 1。

这些结果表明，大  $n$  有相当明确的大概率的体积（至少在对数意义上），并且在这个体积内的概率密度是相对统一的（同样是在对数意义上）。

对于白噪声分布函数是

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi N)^n} \exp -\frac{1}{2N} \sum x_i^2$$

由于这个只取决于  $\sum x_i^2$ ，等概率密度的表面是球形并且整个分布具有球对称。大概率的区域是半径为  $\sqrt{nN}$  的球体。随着  $n \rightarrow \infty$ ，半径为  $\sqrt{n(N+\epsilon)}$  的球外的概率接近于零并且  $\frac{1}{n}$  倍球体积的对数接近于  $\log \sqrt{2\pi e N}$ 。

在连续的情况下，没有集合的熵  $H$  很容易计算但是却有一个叫做熵功率的量。当白噪声具有与原集合限于相同的频带和相同的熵时这被定义为功率。换句话说如果  $H'$  是集合的熵，它的熵功率为

$$N_1 = \frac{1}{2\pi e} \exp 2H'$$

在几何图形中这相当于通过具有相同体积球体的半径的平方测量大概率体积。既然白噪声对于给定的功率具有最大的熵，任何噪声的熵功率都小于它实际的功率。

## 22. 线性滤波器的熵损失

定理 14：如果一个集合在带宽  $W$  内每自由度的熵为  $H_1$ ，经过一个特征函数为  $Y(f)$  的滤波器，输出几集合的熵为

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 df$$

滤波器的操作实际是对坐标的线性变换。如果我们考虑一下原坐标系统的不同频率成分，新的频率成分仅仅是旧频率乘以系数。因此坐标变换矩阵实际上是这些坐标的对角化。雅可比变换是

$$J = \prod_{i=1}^n |Y(f_i)|^2$$

其中  $f_i$  均匀分布于带宽  $W$  内。这有个限制

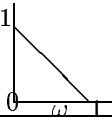
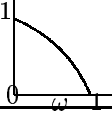
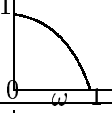
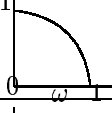
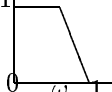
$$\exp \frac{1}{W} \int_W \log |Y(f)|^2 df$$

由于  $J$  是恒定的它的平均值不变并且应用了关于坐标改变熵也改变的定理，结果如下。我们也可以将它描述为熵功率。因此如果第一个集合的熵功率是  $N_1$ ，那么第二个为

$$N_1 \exp \frac{1}{W} \int_W \log |Y(f)|^2 df$$

最后熵功率是起始熵功率乘以滤波器的几何平均增益。如果增益是以分贝  $db$  衡量，那么输出熵功率将以算术平均分贝  $db$  增益增加超出  $W$ 。

表 1

增 益	熵功率系数	熵功率增益分贝数	脉 冲 响 应
 $1 - \omega \rightarrow$	$\frac{1}{e^2}$	-8.69	$\frac{\sin_2(t/2)}{t^2/2}$
 $1 - \omega^2 \rightarrow$	$(\frac{2}{e})^4$	-5.33	$2 [\frac{\sin t}{t^3} - \frac{\cos t}{t^2}]$
 $1 - \omega^3 \rightarrow$	0.411	-3.87	$6 [\frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3}]$
 $\sqrt{1 - \omega^2} \rightarrow$	$(\frac{2}{e})^2$	-2.67	$\frac{\pi}{2} \frac{J_1(1)}{t}$
	$\frac{1}{e^{2\alpha}}$	$-8.69\alpha$	$\frac{1}{\alpha t^2} [\cos(1 - \alpha)t - \cos t]$

表一中熵功率损耗计算(亦用分贝表示)是对一些理想增益特性。这些滤波器的脉冲响应是在  $W = 2\pi$ ，相位假定为 0。

对其它情况的熵损失可以通过这些结果得到。举个例子，第一种情况的熵损失系数  $1/e^2$  也可以应用于来自  $1 - \omega$  的增益特性，通过一个  $\omega$  轴的测量保存变换。尤其是一个线性增益  $G(\omega) = \omega$ ，或者“锯齿”特性在 0 到 1 之间，具有相同的熵损失。互惠增益具有互惠系数。因此  $1/\omega$  的系数为  $e^2$ 。增加增益到任何次

方也把系数增加到几次方。

### 23. 两个函数全体和的熵

如果有两个函数集  $f_\alpha(t)$  和  $g_\beta(t)$ ，我们可以通过“加”形成一个新函数全体。假定第一个函数全体具有概率密度函数  $p(x_1, \dots, x_n)$  第二个为  $q(x_1, \dots, x_n)$ 。那么两个集合的密度函数可通过积分得到：

$$r(x_1, \dots, x_n) = \int \dots \int p(x_1, \dots, x_n) q(x_1 - y_1, \dots, x_n - y_n) dy_1 \dots dy_n$$

实际上这相当于把原始函数全体表示的噪声和信号加起来。

在附录 6 中将导出下列结果：

定理 15：设两个集合的平均功率为  $N_1$  和  $N_2$ ，设两个熵功率为  $\bar{N}_1$  和  $\bar{N}_2$ 。那么总的熵功率， $\bar{N}_3$  位于

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2$$

高斯白噪声具有特殊性质，它可以吸收其它任何噪声和信号全体，这些噪声和信号集产生与两个集合相等的白噪声和信号功率（通过平均信号值算出，这些值通常为零）附加到高斯白噪声上，假定信号功率很小，在一定意义上说，相当于噪声。

假定这些集合的函数空间是  $n$  维的。白噪声在这个空间对应于球形高斯分布。信号全体对应于另一个概率分布，不一定是高斯或球形。令这个分布围绕它重心的二阶距为  $a_{ij}$ 。也就是，如果  $p(x_1, \dots, x_n)$  是密度分布函数，则

$$a_{ij} = \int \dots \int p(x_i - \alpha_i)(x_i - \alpha_j) dx_1 \dots dx_n$$

其中  $\alpha_i$  是重心的坐标。现在  $a_{ij}$  是正定二次型，并且我们可以旋转坐标系，使它与这种形式的主要方向一致，那么  $a_{ij}$  简化为对角线形式  $b_{ii}$ 。我们要求  $b_{ii}$  比球形分布的半径平方  $N$  来得小。

在这种情况下，卷积噪声和信号产生高斯分布，其相应的二次型为

$$N + b_{ii}$$

这个分布的熵功率为

$$\left[ \prod (N + b_{ii}) \right]^{1/n}$$

或近似为

$$\begin{aligned} &= \left[ (N)^n + \sum b_{ii} (N)^{n-1} \right]^{1/n} \\ &= N + \frac{1}{n} \sum b_{ii} \end{aligned}$$

最后一项是信号功率，第一项为噪声功率。

## 第四部分：连续信道

### 24. 连续信道的信道容量

在连续通道的输入或传输的信号是属于某个集合的连续时间函数  $f(t)$ ，而且输出和接收的信号是被干扰的输入。我们只考虑两种情况，传输和接收的信号具有有限带宽  $W$ 。那么它在时间  $T$  内就可用  $2TW$  个数来表示，并且它们的统计结构可用有限维的分布函数表示。因此传输信号的统计特性由下式决定

$$p(x_1, \dots, x_n) = p(x)$$

这些噪声的条件概率分布

$$p_{x_1, \dots, x_n}(y_1, \dots, y_n) = p_y(y)$$

连续信道的信息传输速率的定义类似于离散信道。即

$$R = H(x) - H_y(x)$$

其中  $H(x)$  是输入的熵， $H_y(x)$  为条件熵。当我们在所有集合中改变输入信道容量  $C$  定义为最大信息率  $R$ 。这意味着对有限维，我们必须改变  $p(x) = p_{x_1, \dots, x_n}(y_1, \dots, y_n)$  并且最大化

$$-\int p(x) \log p(x) dx + \int \int p(x, y) \log \frac{p(x, y)}{p(y)} dx dy$$

这可以写成

$$\int \int p(x, y) \log \frac{p(x, y)}{p(x) p(y)} dx dy$$

使用等式  $\int \int p(x, y) \log p(x) dx dy = \int p(x) \log p(x) dx$ 。信道容量可以表示为

$$C = \lim_{T \rightarrow \infty} \max_{p(x)} \frac{1}{T} \int \int p(x, y) \log \frac{p(x, y)}{p(x) p(y)} dx dy$$

很明显式中  $R$  和  $C$  的坐标系是相互独立的，因为当对  $x$  和  $y$  进行一一对应变换时，在  $\frac{p(x, y)}{p(x)}$  中分子和分母乘以相同的系数。 $C$  的积分表达式比  $H(x) - H_y(x)$  更为普遍。由附录 7 可知，这个式子总是成立，尽管有些情况  $H(x) - H_y(x)$  为一种不确定的形式  $\infty - \infty$ 。比如，如果  $x$  被限定为  $n$  维空间小于  $n$  维的表面，那么这种情况会发生。

如果计算  $H(x)$  和  $H_y(x)$  采用的对数的底为 2，即  $C$  就是每秒能在信道上以任意小的莫棱性传输的最大二进制数。这可以认为将信号空间划分为小的区域，足够小以至于使信号  $x$  被干扰成  $y$  的概率密度  $p_x(y)$  在一个体积元中保持常熟。如果这个区域为不同的点，这种情况与离散信道相同并且将引用离散的证明。但是，很明显，只要区域划分得足够小，把体积划分为点子是不会影响最后结果的。因此连续信道的容量将是离散子域的信道容量的极限。

在数学上首先可以证明（见附录 7），如果  $u$  为信息， $x$  为信号， $y$  是接受的信号（噪声干扰）， $v$  是接收的信息，那么

$$H(x) - H_y(x) \geq H(u) - H_v(u)$$



而与由  $u$  求  $x$  或由  $y$  求  $v$  的运算无关。因此不管我们如何将二进制数字编码为信号，或将接收的信号译码成信息，对应二进制的离散传输速率不会超出我们定义的信道容量。另一方面，在最一般的条件下可以找到一种编码系统，它能把二进制数字按速率  $C$  转换二进制数字以任意小的模棱性或误差性进行传输。这是正确的，例如，如果信号函数是一个有限维的空间， $p(x, y)$  对于  $x$  和  $y$  是连续的，除了在概率为 0 的集合的点。

一个重要的特殊情况是噪声附加到信号上并独立于信号（在概率意义上）。那么  $p_y(x)$  是矢量差  $n = (y - x)$  的函数，

$$p_x(y) = Q(y - x)$$

我们可以给噪声一个确定的熵（它与信号的统计性质无关），即分布  $Q(n)$  的熵。这个熵用  $H(n)$  表示。

定理 16：如果信号与噪声是独立的，并且所接收到的信号是被发射的信号与噪声之和，则传输速率为：

$$R = H(y) - H(n),$$

即接收到的信号与噪声的熵减去噪声的熵。则信道容量为：

$$C = \max_{P(x)} H(y) - H(n).$$

因为  $y = x + n$ ：

$$H(x, y) = H(x, n).$$

展开左边部分，由于  $x$  与  $n$  的相互独立，即

$$H(y) + H_y(y) = H(x) + H(n).$$

故

$$R = H(x) - H_y(y) = H(y) - H(n).$$

因为  $H(n)$  与  $P(x)$  无关，使  $R$  最大就要求  $H(y)$  最大，即接收到的信号的熵最大。如果对所传输的信号有某种限制，那么必须使它在满足限制条件下达到最大。

## 25. 有平均最大功率的信道容量

定理 16 的简单应用：当噪声为白热噪声，且发射信号功率为  $P$  时，接收信号平均功率为  $P + N$ ，其中  $N$  是噪声的平均功率。如果接收信号也构成一个白噪声，那么它将具有最大熵。因为这是功率  $P + N$  时的最大可能的熵，并且只要适当地选取被传输的信号全体，即如果发射的信号也构成功率  $P$  的白噪声，那么这个熵是可以得到的。于是接收信号的熵（每秒）为：

$$H(y) = W \log 2\pi e(P + N).$$

而噪声的熵为：

$$H(n) = W \log 2\pi eN.$$

于是信道容量为：

$$C = H(y) - H(n) = W \log \frac{P + N}{N}.$$

归纳起来，可得定理 17：当发射机的平均功率限于  $P$  时，频带为  $W$ 、受功率为  $N$  的白热噪声所干扰的信道的容量为：

$$C = W \log \frac{P+N}{N}.$$

这表明，只要采用足够复杂的编码系统，就能以任意小的误差频率，按  $2^{\frac{P+N}{N}}$  二进制每秒的速率传输二进制数字。在没有正误差频率下，任何编码系统都无法以更高的速率传输信号。为了接近这个极限传信率，被发射的信号必须在统计特性上接近白噪声。<sup>2</sup> 一个接近理想传信率的系统可以描述如下：设在每个持续时间  $T$  中，可以构成  $M=2^s$  个白噪声样品。这些样品可以从 0 到  $(M-1)$  的二进制数字。在发射机端，消息序列分为  $s$  个组。在每个组中，相应的噪声样品作为信号来传输。在接收机端， $M$  个取样是已知的，而将实际接收到的信号（被噪声所干扰的）同其中每一个进行比较。选择那些与接收信号相比均方根偏差最小的样品来作为被传输的信号，然后重构出相应的二进制数字。这种过程就等于选择最可能（后验的）信号。所用的噪声样品的数目  $M$  与所允许的误差频率  $\epsilon$  有关，但是，几乎对所有样品的选择都有：

$$\lim_{\epsilon \rightarrow 0} \lim_{T \rightarrow \infty} \frac{\log M(\epsilon, T)}{T} = W \log \frac{P+N}{N},$$

无论  $\epsilon$  取多小，只要  $T$  足够大，就可以在  $T$  秒内如所欲的接近于传输  $TW \log \frac{P+N}{N}$  个二进制数字。

对白噪声情况，其他作者亦曾独立地推得类于  $C = W \log \frac{P+N}{N}$  的公式，虽然某些解释是不同的 N.Wiener，<sup>3</sup> W.G.Tuller，<sup>4</sup> 和 H.Sullivan。

在任意干扰噪声情况下（不一定是白热噪声），没有发现确定信道容量中的求极大值问题可以得到清楚的解决。但是，用平均噪声功率  $N$  和噪声熵功率  $N_1$  来确定  $C$  的上下限是可以的。在大多数实际情况中，这些界限彼此很接近，因此可以为问题提供满意的结果。

定理 18：频带为  $W$ ，并受任意噪声干扰的信道其容量满足下面的不等式：

$$W \log \frac{P+N_1}{N_1} \leq C \leq W \log \frac{P+N}{N_1}$$

其中

$P$  = 发射机的平均功率

$N$  = 平均噪声功率

$N_1$  = 噪声熵功率

这里，被干扰信号的平均功率亦为  $P+N$ 。如果接收的信号为白噪声，这个功率可以得到最大的熵，即  $W \log 2\pi e(P+N)$ 。但这是不可能得到的，因为没有任何传送信号的全体在叠加了干扰噪声后可以在接收机端产生白热噪声。但是至少这是对  $H(y)$  的一个上限，故

$$\begin{aligned} C &= \text{Max} H(y) - H(n) \\ &\leq W \log 2\pi e (P+N) - W \log 2\pi e N_1 \end{aligned}$$

这就是定理中给出的上限。如果使发射信号是功率  $P$  的白噪声，则可以通过对传输速率的考虑而得到下限。在这种情况下，接收信号的熵功率至少和白噪声的功率  $P+N_1$  一样大，因为我们曾在定理 15 中证明两个函数全体的和的熵功率大于或等于各个熵功效之和。故：

$$\text{Max} H(y) \geq W \log 2\pi e (P+N_1)$$

<sup>2</sup> 从集合观点来讨论的白噪声情况的某些性质可参看 “Communication in the Presence of Noise”

<sup>3</sup> 如上所述控制论。

<sup>4</sup> “无线电信号传送速率的理论极限值”，无线电工程学报 1949 年 5 月刊，468 到 78 页

且

$$\begin{aligned} C &= W \log 2\pi e(P + N_1) - W \log 2\pi e N_1 \\ &\geq W \log \frac{P + N_1}{N_1} \end{aligned}$$

随着  $P$  的增加，上下限相互趋近，所以我们得到一个渐近速率

$$W \log \frac{P + N}{N_1}.$$

如果是白噪声， $N = N_1$  则前面的公式为

$$C = W \log \left(1 + \frac{P}{N}\right).$$

如果是高斯噪声，但其频谱不一定平坦，则  $N_1$  是频带  $W$  内各频率的噪声功率的几何平均值。即：

$$N_1 = \exp \frac{1}{W} \int_W \log N(f) df$$

这里  $N(f)$  是频率为  $f$  时的噪声功率。

定理 19：如果我们使功率为  $P$  的发射机的信道容量为

$$C = W \log \left( \frac{P + N - \eta}{N_1} \right)$$

当  $\eta$  减小， $P$  增大时，则其趋向于 0。

假设给定功率为  $P_1$ ，则信道容量为：

$$C = W \log \left( \frac{P_1 + N - \eta_1}{N_1} \right).$$

这就意味着，当最佳信号分布，例如  $p(x)$ ，叠加到噪声分布  $q(x)$  上时，则接收信号的分布为  $r(y)$ ，它的熵功率为  $(P_1 + N - \eta_1)$ 。假使把白噪声功率  $\Delta P$  加于信号使功率增加到  $P_1 + \Delta P$ ，那么接收信号的熵至少为：

$$H(y) = W \log 2\pi e(P_1 + N - \eta_1 + \Delta P)$$

这是应用和的最小熵功率的定理得到的。因而，由于可以得到所指定的  $H$ ，那么使  $H$  为极大的分布的熵至少应该一样大，而且  $\eta$  是必须是单调下降的。为了证明当  $P \rightarrow \infty$  时， $\eta \rightarrow 0$ ，我们考虑一个具有大功率  $P$  的白噪声信号。无论是怎样的干扰噪声，在熵功率接近  $P + N$  的意义上，只要  $P$  足够大，接收到的信号将接近于白噪声。

## 26. 峰值功率有限的信道容量

在某些应用上，并不限制发射机的平均输出功率，而是限制它的瞬时峰值功率。因此计算信道容量的问题就是在一定的条件下，即对所有的  $t$ ，全体中所有函数  $f(t)$  都小于或等于  $\sqrt{S}$  的条件下，这种类型的强迫不像平均功率极限那样数学化。我们最多得到的是对所有  $\frac{P}{N}$  的下限，一条上线渐近线（根据所有  $\frac{P}{N}$ ）还有一条  $C$  值渐近线亦根据  $\frac{P}{N}$ 。定理 20：频段  $W$ ，并受功率为  $N$  的白热噪声干扰的信道容量  $C$  范围是

$$C \geq W \log \frac{3}{\pi e^3} \frac{S}{N}$$

其中  $S$  是发送机允许的峰值功率。对足够大的  $\frac{S}{N}$ ,

$$C \leq W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon)$$

$\epsilon$  是任意最小值。当  $\frac{S}{N} \rightarrow 0$  (频段  $W$  从 0 开始)

$$C/W \log(1 + \frac{S}{N}) \rightarrow 1.$$

我们希望收到的信号的熵达到最大。如果  $\frac{S}{N}$  很大, 它将在发送信号全体之熵之前为极大时趋于最大。渐近上限可用放宽全体函数的条件来求得。假设功率不是在每个瞬时都限于  $S$ , 而是仅仅在取样点上限制在  $S$ , 那么在这些较弱的条件下被传输信号全体的最大熵肯定地大于或等于在原先条件下所得到的最大熵。这个改变了的问题是很易解决的。如果不同的取样是独立的, 并且分布函数在  $-\sqrt{S}$  到  $+\sqrt{S}$  内是常数, 则熵为最大, 并算得为:

$$W \log 4S.$$

那接收到的信号有一个熵小于

$$W \log(4S + 2\pi e N)(1 + \epsilon)$$

这里当  $\frac{S}{N} \rightarrow \infty$ ,  $\epsilon \rightarrow 0$ , 且信道容量是通过去掉白噪声的熵确定的为  $W \log 2\pi e N$

$$W \log(4S + 2\pi e N)(1 + \epsilon) - W \log(2\pi e N) = W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon).$$

这是为信道容量设定的上限。

为了求得下限, 我们考虑同一个函数全体。使这些函数通过一个具有三角形传输特性的理想滤波器。在频率为 0 时增益为 1, 频率增加时, 作线性下降, 频率为  $W$  时增益下降到 0。我们首先证明滤波器的输出函数在所有时刻 (不恰恰是取样点) 具有峰值功率限  $S$ 。首先注意, 脉冲  $\frac{\sin 2\pi W t}{2\pi W t}$  通过滤波器产生的输出为:

$$\frac{1}{2} \frac{\sin^2 \pi W t}{(\pi W t)^2}$$

这个函数是非负的。在一般情况下, 输入函数可以看成是一系列有位移的函数之和

$$a \frac{\sin 2\pi W t}{2\pi W t}$$

其中  $a$  是取样幅度, 它不大于  $\sqrt{S}$ 。

因此, 输出也是上述非负形式的有位移函数之和, 且位移函数的系数都相同。这由于位移函数值是非负的, 当所有的系数  $a$  取最大正值时, 就得到任何  $t$  时的最大正值, 即  $\sqrt{S}$ 。在这种情况下, 输入函数是幅度为  $\sqrt{S}$  的常数, 并且因为滤波器的直流增益为 1, 所以输出函数是相同的, 因而输出函数全体也有峰功率  $S$ 。

输出函数全体的熵可以借助于前面的定理根据输入函数全体的熵来求得。输出熵等于输入熵加滤波器的几何平均增益:

$$\int_0^W \log G^2 df = \int_0^W \log \left( \frac{W-f}{W} \right)^2 df = -2W.$$

故输出熵为:

$$W \log 4S - 2W = W \log \frac{4S}{e^2}.$$

并且信道容量将大于:

$$W \log \frac{2}{\pi e^3} \frac{S}{N}.$$

现在我们希望证明, 对小的  $\frac{S}{N}$  (信号峰功率与平均白噪声功率之比), 信道容量近似为:

$$C = W \log(1 + \frac{S}{N}).$$

更精确地说, 当  $\frac{S}{N} \rightarrow 0$  时

$$C/W \log(1 + \frac{S}{N}) \rightarrow 1$$

因为信号的平均功率  $P$  小于或等于峰值功率  $S$ , 对所有  $\frac{S}{N}$  可得:

$$C \leq W \log(1 + \frac{P}{N}) \leq W \log(1 + \frac{S}{N}).$$

所以, 如果我们找到一种传输速率接近于  $W \log(1 + \frac{S}{N})$  的函数全体, 并限于频带  $W$  和峰值功率  $S$ , 那么上述结果就可以得证。考虑下列形式的函数全体。一个由  $t$  个取样组成的系列具有同样的峰值  $+\frac{1}{2}\sqrt{S}$  或  $-\frac{1}{2}\sqrt{S}$ , 后面  $t$  个取样也有同样值等等。一个系列的值是随机选择的,  $+\frac{1}{2}\sqrt{S}$  的概率为  $\frac{1}{2}$ ,  $-\frac{1}{2}\sqrt{S}$  的概率也为  $\frac{1}{2}$ 。如果这个函数全体通过具有三角增益特性 (直流增益为 1) 的滤波器, 那么输出功率值限于  $\frac{S}{2}$ 。此外, 平均功率近于  $S$ , 并且当  $t$  足够大时, 这个值是可以趋近的。这个全体与噪声总和的熵可以用噪声与小信号总和的定理来求得, 如果  $\sqrt{t} \frac{S}{N}$  足够小, 定理就可以适用。这可以用  $\frac{S}{N}$  足够小 (选定  $t$  后) 来保证。熵功率将以所需的近似程度接近于  $S + N$ , 因此传信率也近于我们所希望的值

$$W \log(\frac{S + N}{N}).$$

## 第五部分: 连续信源产生信息的速率

### 27. 保真度的估值函数

在离散信源情况下, 我们能够确定一个确定的产生信息的速率, 即基本随机过程的熵。对连续信源, 情况就相当复杂。首先, 一个连续变量可以有无限个值, 因此为了正确表达它, 就要求有无限个二进数字。这就是说为了传输连续信源的输出, 并在接收端正确恢复, 通常就要求信道具有无限大容量。因为, 通常在信道中总有一定的噪声电平, 故容量是有限的, 要完全正确传输是不可能的。

但是这问题的实质。实际上, 在连续信源时, 我们要的不是精确的传输, 而是在一定误差范围内进行传输。问题在于当我们只要求一定的复现保真度 (用适当方法量度) 时, 我们能不能对连续信源规定一个确定的速率。当然, 当保真度的要求提高时, 信息产生率也将增大。可以证明, 在一般的情况下, 我们是能够确定这样的速率的, 只要采用合适的编码, 就能够在信道容量等于这个速率的信道上得到传输, 并满足所要求的保真度。容量比小的信道就不可能得到这种性质。

首先必须定出传输保真度概念的数学公式。我们考虑一组长度为  $T$  (秒) 的信息。信息源用选择信息的有关空间的概率密度来描述。一个给定的通信系统, (从外部看来) 可用条件概率  $P_x(y)$  (当信源产生的消息为  $x$  时, 接收端上复现  $y$  的概率) 来描述。整个系统 (包括信源和传输系统) 可以用消息  $x$  和最后输出  $y$  的概率函数  $P(x,y)$  来描述。如果这个函数已知, 在保真度观点上, 整个系统的特性就完全知道了。对保真度的任何估价在数学上必须相当于对函数  $P(x,y)$  进行运算, 这种运算至少应该能够将系统进行简单的比较。换句话说, 对于以  $P_1(x,y)$  和  $P_2(x,y)$  表示的两个系统, 根据我们的保真度标准, 至少能够说出究竟是:

(1) 第一个函数有较高的保真度, (2) 第二个函数有较高的保真度, 或者是 (3) 它们的保真度相等。这就是说保真度的标准能用数值上估价的估值函数来表达:

$$v(P(x, y))$$

其中总量的范围遍及所有可能的概率函数  $P(x, y)$ 。函数  $v(P(x, y))$  将系统保真度依次排列, 为了方便起见, 以后取  $v$  的较小的值对应于较高的保真度。

现在我们将在一般和合理的假设下证明函数  $v(P(x, y))$  能够写成看来颇为特殊的形式, 即函数  $\rho(x, y)$  在  $x$  和  $y$  的可能值的集合中的平均值:

$$v(P(x, y)) = \int \int P(x, y) \rho(x, y) dx dy.$$

为了得到这个结果, 我们只要假设: (1) 信源和系统都是遍历的, 故一个很长的样品将是函数全体的代表 (概率近于 1); (2) 所谓估价是“合理的”, 是指有可能通过对典型的输入和输出样品  $x_1$  和  $y_1$  的观察, 在这些样品基础上构成试验性的估值。如果增长这些样品的长度时, 试验性的估值将以概率 1 接近于在  $P(x, y)$  完全知道的基础上得到的正确的估值。令试验性的估值是  $\rho(x, y)$ 。那么函数  $\rho(x, y)$  (当  $T \rightarrow \infty$  时) 几乎对于所有对应于系统高概率区域的  $(x, y)$  都趋近常数。

$$\rho(x, y) \rightarrow v(P(x, y))$$

我们也可写成

$$\rho(x, y) \rightarrow \int \int P(x, y) \rho(x, y) dx dy$$

因为

$$\int \int P(x, y) dx dy = 1$$

这个建立了所期望的结果。

函数  $\rho(x, y)$  具有  $x$  与  $y$  之间“距离”的一般性质<sup>5</sup>

它度量了当  $x$  被传输时, 接收  $y$  的拒绝程度 (根据我们的保真度要求)。上面所得到的结果可以重新叙述如下: 任何合理的估值都可以距离函数在消息  $x$  和复现消息  $y$  集合上根据得到它们的联合概率  $P(x, y)$  加权的平均值来表示。但该消息的持续时间  $T$  应取得足够大。

下面是估值函数的简单例子:

1. 均方根标准:

$$v = \overline{(x(t) - y(t))^2}.$$

在这个很常用的保真度的测度中, 距离函数  $\rho(x, y)$  是 (除了一个常数因子外) 在有关函数空间中  $x$  与  $y$  点间的欧几里得距离的平方

$$\rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 dt.$$

2. 频率加权均方根标准:

更普遍地, 在采用均方根标准度量保真度之前, 可对不同的频率分量进行不同的加权。这就相当于使差值  $x(t) - y(t)$  通过一个成形滤波器, 而后在其输出端上求平均功率。令

$$e(t) = x(t) - y(t)$$

---

<sup>5</sup> 这不是严格的用尺来量度的距离, 因为它通常不满足条件  $\rho(x, y) = \rho(y, x)$  或者  $\rho(x, y) + \rho(y, x) \geq \rho(x, z)$ 。

及

$$f(t) = \int_{-\infty}^{+\infty} e(\tau)k(t-\tau)d\tau$$

于是

$$\rho(x, y) = \frac{1}{T} \int_0^T |x(t) - y(t)| dt.$$

3. 绝对误差标准:

$$\rho(x, y) = \frac{1}{T} \int_0^T f(t)^2 dt.$$

4. 人耳和大脑的结构决定了隐隐的确定了一些适用于语言或音乐传输的感觉标准。例如，有一个可懂度标准  $\rho(x, y)$  是当  $x(t)$  消息被接受成为  $y(t)$  时误解字的相对频率。虽然对这种场合我们不能给出  $\rho(x, y)$  的明确的表达式，但在原则上，可由足够多的实验来确定。它的一些性质可从熟知的听觉实验结果得出。例如，耳朵对相位是不灵敏的，而耳朵对幅度和频率的灵敏度大约是对数关系。
5. 离散情况可以被认为我们已假定了估值是建立在误差频率基础上的特殊情况。于是函数  $\rho(x, y)$  的定义是在序列  $y$  中与  $x$  序列中相应符号不同的符号数除以  $x$  中的符号总数。

## 28. 信源速率（相对于保真度估值）

现在我们定义连续信源的产生信息的速率。我们已给定信源的  $P(x)$  和由距离函数  $\rho(x, y)$  确定的估值，这个距离函数假定对  $x, y$  都是连续的。对以特定的  $P(x, y)$  系统，其质量可用下式来度量

$$v = \int \int \rho(x, y) P(x, y) dx dy.$$

此外，对应于  $P(x, y)$  的二进制数字流速为：

$$R = \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy.$$

我们把复现质量为  $v_1$  的信息产生率  $R_1$  定义为当使  $v$  固定在  $v_1$  而改变  $P_x(y)$  时  $R$  的最小值：

$$R_1 = \min_{P_x(y)} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy.$$

约束条件为：

$$v_1 = \int \int P(x, y) \rho(x, y) dx dy.$$

实际上，这意味着我们研究了所有能够使用的并能保证所需保真度的通信系统。对每一系统可用每秒二进单位数来计算传输速度，并且我们取其最小的一个。这个最小值就是为所要求的保真度选定的信息源产生速率。

定理 21：如果信源对估值  $v_1$  有速率  $R_1$ ，则可以将信源的输出进行编码，并以任意接近于  $v_1$  的保真度在容量  $C$  的信道上传输，条件是  $R_1 \leq C$ 。如果  $R_1 > C$ ，则不可能。

定理的最后部分可以直接从  $R_1$  的定义和前面的结果得到。如果它不成立，我们就以大于二进单位每秒的速率在容量为  $C$  的信道上进行。定理的第一部分，可以用类似于定理 11 中采用的方法证明。首先我们可以将  $(x, y)$  空间分为大量的小体积元而将连续信息离散化。因为假定  $\rho(x, y)$  是连续的，所以这不会使估值函

数发生大于以任意小量的变化（如果体积元分得很小）。假设  $P_1(x,y)$  是给出最小速率  $R_1$  的一个系统。我们在从高概率  $y$  中随意选取包括

$$2^{(R_1+\epsilon)T}$$

个元素的集，当  $T \rightarrow \infty$  时  $\epsilon \rightarrow 0$ 。

在大  $T$  时（如图 10）连接到  $x$  点集。类似于用来证明定理 11 的计算指出，在  $T$  很大时几乎对所有  $y$  的选择，从选择点  $y$  出发的扇形线几乎包括了所有  $x$  点。所有通信系统工作如下：被选点给以二进制数字。一个消息  $x$  出发后它（当  $T \rightarrow \infty$  时发生概率接近为 1）至少将处于一个扇形之中。于是相应的二进制数字（如果有好几个时任意取其中一个）用适当的编码以小的误差概率在信道上传输。因为  $R_1 \leq C$  所以这是可能的。在接收端熵，相应的  $y$  可以重新构成作为复现消息。

这个系统的估值  $v'_1$ ，当  $T$  足够大时，能任意地接近于  $v_1$ 。这是由于对每个长信息样本  $x(t)$  和复现的信息  $y(t)$ ，估值趋近于  $v_1$ （概率为 1）。

有指出下列事实是意义；在这个系统中，复现消息中的噪声实际上是发送机中的一般量化（分层）所产生而不是由信道中的噪声产生的。它或多或少地类似于脉冲编码调制系统中的噪声量化。

## 29. 信息产生率的计算

信息产生率的定义在很多方面与信道容量的定义相类似。

$$R = \min_{P_x(y)} \int \int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dx dy$$

$P(x)$  和  $v_1 = \int \int P(x,y) \rho(x,y) dx dy$  是确定的。下面

$$C = \max_{P(x)} \int \int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dx dy$$

$P_x(y)$  是确定的，并且可能有一种或更多种其他限制（如平均功率限制），其形式为： $K = \int \int P(x,y) \lambda(x,y) dx dy$ 。

求资源速率问题中的最大值问题的一种方法可给出。用拉格朗日方法我们有

$$\int \int [P(x,y) \log \frac{P(x,y)}{P(x)P(y)} + \mu P(x,y) \rho(x,y) + \nu(x) P(x,y)] dx dy.$$

方程的变化（我们先对  $P(x,y)$  变化）

$$P_y(x) = B(x) e^{-\lambda \rho(x,y)}$$

这里  $\lambda$  确定于所要求的保真度， $B(x)$  需要满足

$$\int B(x) e^{-\lambda \rho(x,y)} dx = 1.$$

这证明，在最佳编码时，引起不同接收消息  $y$  的原因的条件概率  $P_y(x)$ ，将随着  $x$  与  $y$  之间的距离函数  $\rho(x,y)$  作指数下降。在距离函数  $\rho(x,y)$  只取决于  $x$  和  $y$  之间的矢量的特殊情况下，

$$\rho(x,y) = \rho(x-y).$$

那么将得到

$$\int B(x) e^{-\lambda \rho(x-y)} dx = 1.$$



因此  $B(x)$  是一个常数, 我们记为  $\alpha$ , 得到

$$P_y(x) = \alpha e^{-\lambda \rho(x-y)}.$$

遗憾的是, 这些正式解在某些情况下很难估值, 因此似乎它的价值不大。实际上, 只有在一些很简单情况下进行了信息产生率的具体计算。

如果距离函数  $\rho(x, y)$  是  $x$  和  $y$  间的均方差, 而消息全体是白噪声, 那么信息产生率是可以确定的。在这种场合,

$$R = \text{Min}[H(x) - H_y(x)] = H(x) - \text{Max} H_y(x)$$

其中  $N = \overline{(x-y)^2}$ , 但是  $H_y(x)$  的最大值当且仅当  $y-x$  为白噪声时发生, 且等于  $W_1 \log 2\pi e N$ ,  $W_1$  为消息全体的带宽。因此

$$\begin{aligned} R &= W_1 \log 2\pi e Q - W_1 \log 2\pi e N \\ &= W_1 \log \frac{Q}{N} \end{aligned}$$

其中  $Q$  为平均信息能量, 证明如下:

定理 22: 对功率  $Q$ , 频带  $W_1$  的白噪声信源, 在用均方根标准度量标准度时的信息产生率为:

$$R = W_1 \log \frac{Q}{N}$$

其中  $N$  为原先消息和复现消息间允许的均方误差。

更一般的说, 任何消息源, 在均方误差标准下, 信息产生率在两个界限之间。

定理 23: 任何频带为  $W_1$  的信源其信息产生率满足:

$$W_1 \log \frac{Q_1}{N} \leq R \leq W_1 \log \frac{Q}{N}$$

其中  $Q$  为信源的平均功率,  $Q_1$  为熵功率,  $N$  为允许的平方误差。

下限是根据这样的事实, 即在白噪声情况下, 对于给定的  $N = \overline{(x-y)^2}$ , 出现  $\text{Max} H_y(x)$ 。当我们不是以最佳方法来安排各个点子 (用在定理 21 的证明中的), 而是随机的放在半径为  $\sqrt{Q-N}$  的球上, 就可得出上限。

## 致谢

作者写此书受惠于实验室的同事, 特别是 Dr.H.W.Bode, Dr.J.R.Pierce, Dr.B.McMillan 和 Dr.B.M.Oliver, 感谢他们在著书过程中所提供的宝贵意见和批评。此荣誉同时必须给予 N.Wiener 教授, 他完美的解答了滤波器的问题, 和对静态总体的预测, 这些都极好的影响了作者对这个领域的思考。

## 附录 5

令  $S_1$  为集合  $g$  的任一个子集,  $S_2$  为集合  $f$  中的任一个子集, 且  $S_1$  和  $S_2$  的关系是

$$S_1 = TS_2.$$

令  $H^\lambda$  为一个运转器, 在时间  $\lambda$  内可转换所有函数到一个设置中。因此

$$H^\lambda S_1 = H^\lambda TS_2 = TH^\lambda S_2$$

因为  $T$  是常量，从而将改变  $H^\lambda$ 。因此如果  $m[s]$  为关于  $S$  的可能性测量

$$\begin{aligned} m[H^\lambda S_1] &= m[H^\lambda S_2] = m[H^\lambda S_2] \\ &= m[S_2] = m[S_1] \end{aligned}$$

其中第二个等式是在集合  $g$  中定义的测量。第三个等式因为集合  $f$  所以是固定的。最后一个等式同样是在集合  $g$  中定义。

下面证明静态过程下的遍历特性，设  $S_1$  为集合  $g$  的一个子集，其中  $g$  在  $H^\lambda$  下是不变的，设  $S_2$  为所有函数  $f$  中的一个，它将转化为  $S_1$ ，则

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2 = S_1$$

以便  $H^\lambda S_2$  对于所有  $\lambda$  都包括在  $S_2$  中。则

$$m[H^\lambda S_2] = m[S_1]$$

即

$$H^\lambda S_2 = S_2$$

其中对所有  $\lambda m[S_2] \neq 0, 1$ 。这个矛盾说明  $S_1$  不存在。

## 附录 6

由于在白噪声下的最大熵为  $N_1 + N_2$ ，则上限带宽  $\overline{N_3} \leq N_1 + N_2$ 。

假设有两个容量为  $n$ ，且熵权分别为  $\overline{N_1}$  和  $\overline{N_2}$  的  $n$  维分布函数  $p(x_i)$  和  $q(x_i)$ ， $p$  和  $q$  将采取何种形式才能使它们的旋转速率的熵权  $\overline{N_3}$  取得最小：

$$r(x_i) = \int p(y_i) q(x_i - y_i) dy_i.$$

对  $r$  的  $H_3$  的熵权是：

$$H_3 = - \int r(x_i) \log r(x_i) dx_i.$$

我们希望最小化  $H_3$ ，而给出下面这两个限制

$$H_1 = - \int p(x_i) \log p(x_i) dx_i$$

$$H_2 = - \int q(x_i) \log q(x_i) dx_i.$$

那我们考虑

$$U = -[r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)] d_x$$

$$\delta U = -[[1 + \log r(x)] \delta r(x) + \lambda [1 + \log p(x)] \delta p(x) + \mu [1 + \log q(x)] \delta q(x)] d_x.$$

如果  $p(x)$  在一个特殊的自变量  $x_i = s_i$  变化，那么在  $r(x)$  里的变化是

$$\delta r(x) = q(x_i - s_i)$$

和

$$\delta U = - \int q(x_i - s_i) \log r(x_i) dx_i - \lambda \log p(s_i) = 0$$

相似的当  $q$  变化时。因此最小值的条件是

$$\begin{aligned} \int q(x_i - s_i) \log r(x_i) dx_i &= -\lambda \log p(s_i) \\ \int p(x_i - s_i) \log r(x_i) dx_i &= -\mu \log q(s_i). \end{aligned}$$

如果我们将第一个乘以  $p(s_i)$ ，第二个乘以  $q(s_i)$  与  $s_i$  相结合则

$$H_3 = \lambda H_1$$

$$H_3 = \mu H_2.$$

或者将  $\lambda$  和  $\mu$  进行替换，等式变为

$$\begin{aligned} H_1 \int q(x_i - s_i) \log r(x_i) dx_i &= -H_3 \log p(s_i) \\ H_2 \int p(x_i - s_i) \log r(x_i) dx_i &= -H_3 \log q(s_i). \end{aligned}$$

现在假定  $p(x_i)$  和  $q(x_i)$  相互垂直

$$\begin{aligned} p(x_i) &= \frac{|A_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \sum A_{ij} x_i x_j \\ q(x_i) &= \frac{|B_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \sum B_{ij} x_i x_j. \end{aligned}$$

然后  $r(x_i)$  也二次齐次正态分布。那么  $c_{ij}$  为

$$\begin{aligned} c_{ij} &= a_{ij} + b_{ij}. \\ \log r(x_i) &= \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \sum C_{ij} x_i x_j \\ \int q(x_i - s_i) \log r(x_i) dx_i &= \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \sum C_{ij} s_i s_j - \frac{1}{2} \sum C_{ij} b_{ij}. \end{aligned}$$

这个应该等于

$$\frac{H_3}{H_1} \left[ \frac{n}{2} \log \frac{1}{2\pi} |A_{ij}| - \frac{1}{2} \sum A_{ij} s_i s_j \right]$$

在这里要求  $A_{ij} = \frac{H_3}{H_1} C_{ij}$ 。在这种情况下  $A_{ij} = \frac{H_2}{H_1} B_{ij}$  两个公式相等。

## 附录 7

下面将指出一个更普遍更严格的方法来研究通信理论中的主要定义。假定有一个概率测度空间，该空间的元素是序偶  $(x, y)$ 。变量  $x, y$  看做是长度为  $T$  的发射和接收信号。称  $x$  属于点子集  $S_1$  的所有点的点集为  $S_1$  的带，称  $y$  属于  $S_2$  的集为  $S_2$  的带。将  $x$  及  $y$  分成为不相重叠的可测子集  $X_i$  和  $Y_i$  的集合，则

$$R_1 = \frac{1}{T}$$

这里

$P(X_i)$  是  $X_i$  上的带的概率测度。

$P(Y_i)$  是  $Y_i$  上的带的概率测度。

$P(X_i, Y_i)$  是带交叉处的概率测度。

进一步分割绝不会减小  $R_1$ 。如果  $X_1$  分解为  $X_1 = X'_1 + X''_1$  和使

$$P(Y_1) = a \quad P(X_1) = b + c$$

$$P(X'_1) = b \quad P(X'_1, Y_1) = d$$

$$P(X''_1) = c \quad P(X''_1, Y_1) = e$$

$$P(X_1, Y_1) = d + e$$

然后我们利用  $d \log \frac{d}{ab} + e \log \frac{e}{ac}$  替换 (对  $X_1, Y_1$  交叉)

$$(d + e) \log \frac{d + e}{a(b + c)}$$

很容易证明 b,c,d,e 有极限,

$$\left[ \frac{d + e}{b + c} \right]^{d+e} \leq \frac{d^d e^e}{b^d c^e}$$

对各种可能的情况再分割构成一个有向集,  $R$  随着分割的继续进行而单调递增。令  $R_1$  的上限为  $R$ , 则

$$R = \frac{1}{T} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy$$

这个积分, 在上面的意义熵来理解, 能包括连续和离散以及许多不能用连续或离散形式表示的情况。在此式中, 如果  $x$  和  $u$  是一一对应, 那么从  $u$  到  $y$  的传输速率等于从  $x$  到  $y$  的传输速率。如果  $v$  是  $y$  的任何函数 (不一定有反函数), 那么从  $x$  到  $y$  的传信率将大于等于从  $x$  到  $v$  的传输速率。因为在近似计算中,  $y$  的分割是  $v$  的更细分割。更一般的, 如果  $y$  和  $v$  是统计上的联系而不是函数上的, 即有一个概率测度空间  $(y, v)$  那么  $R(x, v) \leq R(x, y)$ 。这意味着对任何接收信号可以作任何运算, 即使它包含有统计元素, 也不能增加  $R$ 。

另一个在抽象理论中应该精确的定义的概念是所谓“维速率”, 它是描述函数全体中的一个元素每秒所需要的维的平均数。在频带限制情况中, 每秒  $2W$  割维数就足够了。更普遍的定义可构造如下: 让  $f_\alpha(t)$  作为一个函数全体, 让  $\rho_T[f_\alpha(t), f_\beta(t)]$  作为时间  $T$  内从  $f_\alpha(t)$  到  $f_\beta(t)$  距离的测量 (例如这个区间的均方根差)。令  $N(\epsilon, \delta, T)$  为元素  $f$  的最小数, 除一个测度为  $\delta$  的集合外, 函数全体的所有元素至少都处在一个被选元素的距离  $\epsilon$  以内。于是, 除一个小测度  $\delta$  的集合, 可以普及到  $\epsilon$  以内的空间。用三重极限来定义函数全体的“维速率”  $\lambda$

$$\lambda = \lim_{\delta \rightarrow 0} \lim_{\epsilon \rightarrow 0} \lim_{T \rightarrow \infty} \frac{\log N(\epsilon, \delta, T)}{T \log \epsilon}.$$

这是拓扑学中维的测度定义推广, 而且与结果很明显的简单的函数全体的比较直观“维速率”相一致。