

# 通信的数学理论

## (A Mathematical Theory of Communication)

C. E. SHANNON

### 引言

近来出现了许多以带宽换取信噪比的调制方法，比如 PCM 和 PPM，它们的出现进一步激发了人们对广义通信理论的兴趣。在奈奎斯特（Nyquist）<sup>1</sup>和哈特莱（Hartley）<sup>2</sup>发表的一些重要相关论文中，奠定了这一理论的基础。本论文将扩展该理论，增加一些新的因素，具体来说，就是信道中噪声的影响、由于原始消息的统计结构和最终信宿的本质而可能减省的内容。

通信的基本问题就是在一个地方复现在另一个地方选定的消息，这一复现可能是准确的，也可能是近似的。这些消息通常有**特定的含义**；也就是说，它们会根据某一系统，与特定的物理或概念实体关联在一起。通信的语义与工程问题无关。重要的是：实际消息是**从一个消息集合选出的**。所设计的系统必须能够处理任意选定的消息，而不是仅能处理实际选择的特定消息，因为在设计系统时，并不知道会实际选择哪条消息。

如果集合中的消息数目是有限的，而且选择每条消息的可能性相等，那就可以用这个消息数或者它的任意单调函数，来度量从集合中选择一条消息所生成的信息量。正如哈特莱所指出的那样，最自然的选择就是对数函数了。如果考虑消息统计信息的影响，如果消息的选取范围是连续的，那必须对其定义进行重要扩展，但在所有情况下，我们使用的度量在实质上都是对数函数。

对数度量之所以更为便利，其原因有多种：

1. 它在实践中更为有用。一些在工程上非常重要的参数，比如时间、带宽、延迟数，等等，往往与可能性的数量的对数值呈线性关系。例如，增加一个继电器会使继电器的可能状态数加倍。如果对这一数目求以 2 为底的对数，则增加一个继电器后，会使结果加 1。使时间加倍，会使可能消息数近似变为原来的平方，而其对数则是加倍，诸如此类。
2. 它更接近于人类对正确度量的直观认知。这一点与第 1 个原因密切相关，因为人们在对实体进行直觉度量时，通常是与公共标准进行线性比较。比如，人们认为，两张打孔卡存储信息的容量应当是一张打孔卡的两倍，两个相同信道的信息传输能力应当是一个信道的两倍。

<sup>1</sup> Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, v. 47, April 1928, p. 617.

<sup>2</sup> Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

3. 更适于数学运算。许多极限运算很容易用对数表示，如果采用可能性的数目表示，可能会需要进行冗繁、笨拙的重新表述。

对数底数的选择与信息度量单位的选择相对应。如果所用底数为 2，则所得到的结果可以称为二进制数位 (binary digit)，或者简称为**比特** (bit)，它是由 J. W. Tukey 提议采用的。一个具有两种稳定状态的器件，比如继电器或者触发电路，可以存储 1 比特信息。N 个此种器件可以存储 N 比特，因为可能状态的总数为  $2^N$ ，而  $\log_2 2^N = N$ 。如果所用底数为 10，则所得单位可以称为十进制数字 (decimal digit)。因为：

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2 \\ &= 3.32 \log_{10} M\end{aligned}$$

一个十进制数位大约为  $3\frac{1}{3}$  个比特。台式计算机上的一个数位轮有 10 个稳定状态，因此，其存储容量是一个十进制数位。在一些涉及积分和微分的分析中，底数 e 有时会很有用。所得到的信息单位称为自然单位。只需要乘以  $\log_b a$  就可以将底数 a 改为底数 b。

我们所说的通信系统，是指图 1 中示意给出的系统类型。它基本上由五部分组成：

1. **信源**，生成要传送给接收终端的消息或消息序列。消息可能是各种不同类型：(a) 字符序列，比如电传系统电报机中；(b) 单个时间函数  $f(t)$ ，比如无线电或电话通信中；(c) 时间及其他变量的函数，比如黑白电视机中——这里的消息可以看作是二个空间坐标与时间的函数  $f(x, y, t)$ ，也就是摄像管盘面上点  $(x, y)$  处在时刻 t 的光强度；(d) 时间的两个或更多个函数，比如  $f(t), g(t), h(t)$ ——“三维”声传送即属这一情景，如果通信系统要以多工方式为几个独立信道提供服务，则同属这一情景；(e) 几个变量的几个函数——在彩色电视机中，消息包含三个函数  $f(x, y, t), g(x, y, t), h(x, y, t)$ ，它们都定义在一个三维闭联集 (continuum) 上——我们还可以将这三个函数看作是定义在该区域上的一个向量场的分量——与此类似，几台黑白电视源所生成的“消息”由许多三变量函数组成；(f) 还会有各种组合情景，比如，在带有关联音频声道的电视中。
2. **发送器**，它以某种方式对消息进行处理，生成一个适于在信道中传送的信号。在电话通信中，这一处理就是将声压变换为比例变化的电流。在电报中采用一种编码操作，在信道中生成一系列与消息相对应的点、划和空。在多工 PCM 系统中，必须对不同的语音函数采样进行采样、压缩、量化和编码，最后进行恰当的交错，从而构造出信号。在声码器系统、电视、频率调制中，也都需要对消息进行一些复杂处理才能得到信号。
3. **信道**，就是供发送器向接收器传送信号的媒介。它可能是一对导线、一根同轴电缆、一个无线电频带、一道光束，等等。
4. **接收器**，通常是执行发送器所做处理的逆处理，由信号重构出消息。
5. **信宿**，意欲向其传送消息的人（或物）。

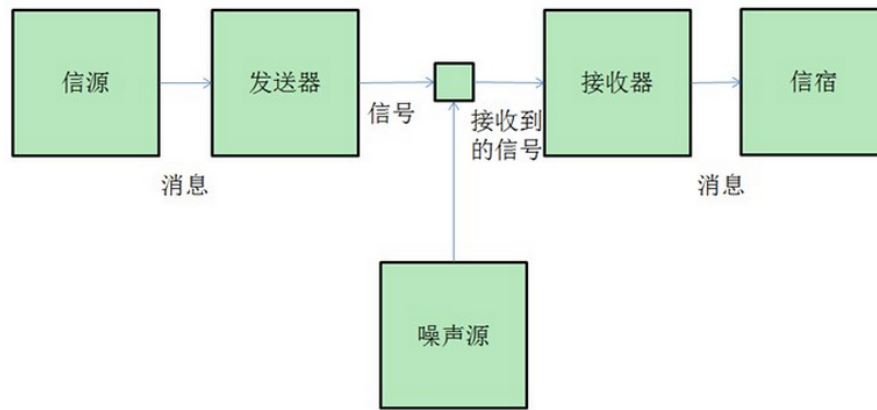


图 1 一般通信系统示意图

我们希望考虑涉及通信系统的某些一般性问题。为此，首先需要对所涉及的各个物理部分进行抽象，用数学方式表示出来。我们可以将通信系统粗略地分为三大类：离散系统，连续系统，混合系统。离散系统是指其中的消息和信号都是离散符号序列。电报是这种系统的一个典型例子，其中的消息是一个字符序列，信号是一个由点、划和空组成的序列。连续系统是指其中的消息和信号都可以看作连续函数，比如，无线广播或电视。混合系统是指离散变量与连续变量都可能出现的系统，比如语音的 PCM 传送。

我们首先考虑离散情景。这种情景不仅在通信理论中有应用，同样适用于计算机理论、电话交换设计及其他领域。此外，离散情景还为连续情景和混合情景奠定了基础，后两种情景将在本论文的第二部分讨论。

## 第 I 部分：离散无噪声系统

### 1. 离散无噪声信道

电传打字机和电报通讯是信息传送离散信道的两个简单例子。一般来说，离散信道意味着可以通过一个系统，从一点向另一点传送一个选择序列，而该序列选自一个由基本符号 $S_1, \dots, S_n$ 组成的有限集合。假定每个符号 $S_i$ 的特定持续时间为 $t_i$ 秒（对于不同的 $S_i$ ，此持续时间不一定相同，比如电报中使用的点和划）。并不要求在此系统中能够传送 $S_i$ 的所有可能序列；可以仅允许出现特定序列。这些特定序列就是可能出现在该信道中的信号。因此，在电报中，假定这些符号为：(1) 点，先将线路闭合一个时间单位，然后再断开一个时间单位；(2) 划，线路闭合三个时间单位，然后断开一个时间单位；(3) 字符空，比如将线路断开三个时间单位；(4) 字空，线路断开六个时间单位。我们可以对允许出现的序列设定限制：不允许两个空相邻（因此，如果两个字符空相邻，则与一个字空相同）。我们现在考虑的问题是，如何度量这样一个信道的信息传输能力。

在电传打字机中，所有符号的持续时间相同，允许出现任何由 32 个符号组成的序列，上面的问题很容易解答。每个符号表示 5 比特信息。如果系统每秒传送  $n$  个符号，那自然可以说该信道的容量为  $5n$  比特/秒。这并不是说电传信道总是以这一速度传送信息——这是最大可能速率，后面将会看到，实际速率能否达到这一最大值，取决于向信道馈送信息的信源。

在更一般的情况下，符号的长度不同，而且对允许序列设有限制，我们做出如下定义：

定义：离散信道的容量  $C$  给出如下：

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

式中， $N(T)$ 是指在允许出现的信号中，持续时间为  $T$  的信号数目。

容易看出，在电传情况下，这一公式简化为前面的结果。可以证明，在人们所关注的大多数情况下，上述极限值存在且有穷。假定允许出现信号 $S_1, \dots, S_n$ 的所有序列，而且这些符号的持续时间为 $t_1, \dots, t_n$ 。信道容量是多少呢？如果  $N(t)$  表示持续时间为  $t$  的序列数，则有：

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

该总数等于以 $S_1, S_2, \dots, S_n$ 结尾的序列数目之和，这些数目分别为 $N(t - t_1), N(t - t_2), \dots, N(t - t_n)$ 。由有限差分中一个众所周知的结果可知，对于大的 $t$ 值， $N(t)$ 趋近于 $X_0^t$ ，其中 $X_0$ 是以下特征方程的最大实数解：

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

因此，

$$C = \log X_0$$

在对允许出现的序列设定了限制时，仍然能够获得这一类型的差分方程，并由该特征方程求得  $C$ 。在前面提到的电报情景中，根据最后一个符号或者倒数第二个符号来计算符号序列的数目，可以得出：

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

因此， $C$  为  $-\log \mu_0$ ，其中  $\mu_0$  是  $1 = \mu^2 + \mu^4 + \mu^5 + \mu^7 + \mu^8 + \mu^{10}$  的正根。求解此方程后可得  $C=0.539$ 。

在对允许序列设定的限制中，有一种非常普通的类型：假设有大量可能状态  $a_1, a_2, \dots, a_m$ ，对于每种状态，只能传送集合中的特定符号  $S_1, \dots, S_n$ （不同状态对应的子集不同）。在传输一个序列后，系统状态改为一种新的状态，具体取决于原有状态和所传送的特定符号。电报是这种情景的一个简单示例。根据最后传送的符号是不是空格，共存在两种状态。如果是空格，则接下来只能传送一个点或一个划，状态总是发生改变。如果不是空格，则可以传送任意符号，如果发送的是空格，则状态发生变化，如果不是空格，则状态保持不变。这些条件可以用如图 2 所示的线性图表示。交点对应于状态，连线表示一种状态下可以传送的符号及传送符号后所得到的状态。在附录 1 中，如果可以用这种方式来描述对允许序列设定的条件，则  $C$  存在，并可计算如下：

**定理 1：** 设  $b_{ij}^{(s)}$  是指在状态  $i$  下允许出现并导致状态  $j$  的第  $s$  个符号的持续时间，则信道容量  $C$  等于  $\log W$ ，其中  $W$  为以下行列式方程的最大实根：

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

其中，当  $i = j$  时， $\delta_{ij} = 0$ ，否则，等于 0。

例如，在电报通讯中（图 2），该行列式为：

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

展开后，即可得到上文针对这一情景给出的方程。

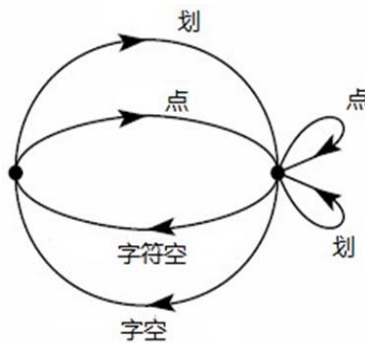


图 2 用图形表示针对电报符号设置的约束条件

## 2. 离散信源

我们已经看到，在非常一般的条件下，离散信道中可出现信号数的对数随时间线性增加。如果能给出这一增长速率，也就是每秒需要多少比特来表示所使用的特定信号，每秒钟所需要的比特数，就能给出信息传输容量。

我们现在考虑信源。如何用数学描述一个信源呢？一个给定信源每秒生成多少比特的信息呢？问题的要点在于，如何利用信源的相关统计知识，通过信息的正确编码，减少所需要的信道容量。比如，在电报通信中，要传送的消息由字符序列组成。但是，这些序列并不是完全随机的。一般情况下，它们会组成句子，具有某种语言的统计结构，比如英语。字符 E 的出现频率要高于 Q，序列 TH 的出现频率要高于 XP，等等。由于此种结构的存在，我们可以对消息序列进行适当编码，转换为信号序列，以节省时间（或信道容量）。其实在电报通讯中已经进行了一定程度的此种处理：为最常见的英文字母 E 使用最短的信道符号——点；而出现较少的 Q, X, Z 则使用较长的点、划序列来表示。这一思想一直沿用到一个特定的商用编码中，在这些编码中，常见的单词和短语用四字符或五字符代码组表示，大幅缩短了平均时间。现在使用的一些标准问候电报和周年纪念电报扩展了这一思想，将一个或两个句子编码为一个较短的数字序列。

我们可以认为离散信源是逐个字符地生成消息。它将会根据特定概率值选择相继符号，这些概率值通常取决于之前的选择和所考虑的特定符号。如果一个物理系统或者一个系统的数学模型，在一组概率的控制下生成符号序列，则这种系统或模型称为随机过程<sup>3</sup>。因此，我们可以考虑用随机过程表示离散信源。相反，任何一个随机过程，只要它生成的离散符号序列是从有限集合中选出的，则可以将其看作离散信源。它将包括类似以下的各种情况：

1. 自然书写的语言，比如英文，德文，中文。
2. 已经用某种量化过程变为离散的连续信源。比如，由 PCM 发送器输出的量化语音，或者是经过量化的电视信号。
3. 数学信源，在此类情况下，我们只是抽象地定义了一个生成符号序列的随机过程。下面是最后一种信源类型的示例。

(A) 假定我们有五个字母 A, B, C, D, E，各字母被选中的概率为 0.2，前后选择之间相互独立。这样会得到一个序列，下面是其中的一个典型示例。

B D C B C E C C C A D C B D D A A E C E E A

A B B D A E E C A C E E B A E E C B C E A D

这一序列是使用一个随机数字表生成的<sup>4</sup>。

(B) 使用相同的五个字母，设各概率为 0.4, 0.1, 0.2, 0.2, 0.1，连续选择之间互相独立。则由这一信源生成的典型消息为：

A A A C D C B D C E A A D A D A C E D A

E A D C A B E D A D D C E C A A A A A D

(C) 如果相邻符号的选择不是独立的，其概率取决于之前的字符，则会得到一种更为复杂的结构。在最简单的此种类型中，字符的选择仅取决于它前面的一个字母，而与再之前的字母无关。这种统计结构可以由一组

<sup>3</sup> 例如，请参阅 S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Reviews of Modern Physics*, v. 15, No. 1, January 1943, p. 1.

<sup>4</sup> Kendall and Smith, *Tables of Random Sampling Numbers*, Cambridge, 1939.

转换概率 $p_i(j)$ 来描述，该概率是指字母  $i$  之后跟有字母  $j$  的概率。下标  $i$  和  $j$  的取值范围为所有可能出现的符号。还有一种等价方式来指定该结构，即给出“连字(digram)”概率 $p(i, j)$ ，也就是连字 $i j$ 的相对频率。字母频率 $p(i)$ （即字母  $i$  的概率）、转换概率 $p_i(j)$ 和连字概率 $p(i, j)$ 之间的关系由以下公式给出：

$$p(i) = \sum_j p(i, j) = \sum_j p(j, i) = \sum_j p(j) p_j(i)$$

$$p(i, j) = p(i) p_i(j)$$

$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i, j} p(i, j) = 1$$

作为一个特定示例，假定共有三个字母 A, B, C，其概率表为：

$p_i(j)$		$j$		
		A	B	C
$i$	A	0	$\frac{4}{5}$	$\frac{1}{5}$
	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

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

这一信源给出的一条典型消息如下：

A B B A B A B A B A B A B B B A B B B B A B A B A B A B B B A C A C A B B A B B B B A B B  
A B A C B B B A B A。

接下来进一步提高复杂性，涉及三连字频率，但不涉及更多的连字频率。一个字母的选择取决于前面的两个字母，但与该点之前的消息无关。这里需要一组三连字频率 $p(i, j, k)$ ，或者一组等价的转换概率 $p_{ij}(k)$ 。以这种方式继续下去，可以持续得到更为复杂的随机过程。在一般的  $n$  连字情况下，需要用一组  $n$  连字概率 $p(i_1, i_2, \dots, i_n)$ 或者转换概率 $p_{i_1, i_2, \dots, i_{n-1}}(i_n)$ 来指定其统计结构。

(D) 还可以定义一些随机过程，它们生成由“单词”序列组成的文本。假定该语言中有五个字母 A, B, C, D, E 和 16 个“单词”，其相关概率为：

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

假定连续“单词”之间用空格分开，而且它们的选择相互独立。一条典型消息可能是：

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。

如果所有单词的长度有限，则这一过程与前一种类型等价，但用单词结构和概率进行描述时，可能更简单一些。我们也可以进行推广，引入单词之间的转换概率，等等。

要构造一些简单的问题和示例来说明各种概率，这些虚拟语言是很有用的。我们还可以利用一系列简单的虚拟语言来模拟一种自然语言。如果所有字母的选择频率相同，且相互独立，则得到零阶近似。如果连续字母的选择相互独立，但每个字母的选择概率和它在自然语言中的概率相同<sup>5</sup>，则得到一阶近似。于是，在英语的一阶近似中，选择 E 的概率为 0.12（也就是它在正常英语中的出现概率），选择 W 的概率为 0.02，但相邻字母之间没有影响，也不会倾向于生成自然语言中更频繁出现的二连字，比如 TH、ED 等。在二阶近似中，引入了两连字结构。在选定第一个字母后，根据各个字母在第一个字母之后的出现频率来选择下一个字母。这就是一个二连字频率  $p_i(j)$  表。在三阶近似中，引入了三连字结构。每个字母的选择概率依赖于它前面的两个字母。

### 3. 英文的近似序列

为了让大家直观地感受这一系列过程是如何近似模拟一种语言的，我们构造了英文的一些典型近似序列，给出如下。在所有情况下，我们假定“字母表”中有 27 个符号——26 个字母和一个空格。

1. 零阶近似（符号的选择相互独立，概率相等）。

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD。

2. 一阶近似（符号的选择相互独立，但其概率与英文文本中相同）。

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

3. 二阶近似（英文中的二连字结构）。

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

4. 三阶近似（英文中的三连字结构）。

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

5. 一阶单词近似。我们接下来不再继续给出四连字结构，……，n 连字结构，而是直接由三连字跳到以单词为单位，这样更容易一些，也更好一些。这里的单词选择是相互独立的，但具有适当的各自频率。

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TOOF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE。

6. 二阶单词近似。单词转换概率与英文中一致，但没有包含其他结构。

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。

在上面给出的各个步骤中，每前进一步，与普通英文文本的相似度都大幅增加。注意，这些示例中的良好结构范围并不仅限于在其构造时考虑的范围，而是其大约两倍。比如，在(3)中，该统计过程确保可以对两字母序列生成符合常理的文本，但这个例子中给出的四字母序列通常也可以组合成很好的句子。在(6)中，不需要

<sup>5</sup> 字母、二连字和三连字频率由 Secret and Urgent by Fletcher Pratt, Blue Ribbon Books, 1939 给出。单词频率在 Relative Frequency of English Speech Sounds, G. Dewey, Harvard University Press, 1923 中列出。



绞尽脑汁进行什么不同寻常的构造过程，就可以将四个或更多个单词组成的序列放入句子中。其中有一个包括十个字音的特定序列“attack on English writer that the character of this”一点也不显得无厘头了。由此可以看出，用一个足够复杂的随机过程来表示一个离散信源，是可以让人满意的。

在构造前两个示例时，使用了一个随机数表和一个频率表（比如第 2 种情况）。因为有二连字频率、三连字频率和单词频率表可用，所以我们可以为第 (3)、(4)、(5) 种情景继续应用这一方法，不过，我们实际使用了一种更为简单的等价方法。比如，为了构造 (3)，我们可以随机翻开一本书，在该页上随机选择一个字母，并记下这个字母。然后翻到这本书的另一页，一直读下去，直到再次遇到前面记下的字母为止。然后记下跟在这个字母后面的字母。再翻到另一页，查找这个第二字母，并记下它后面的字母，以此类推。对 (4)、(5) 和 (6) 应用类似过程。进一步构造其他近似会是非常有趣的，但在下一阶段所涉及的劳动量是非常庞大的。

## 4. 马尔可夫过程的图形表示

上文所述类型的随机过程在数学上称为离散马尔可夫过程，在参考文献中有详尽研究<sup>6</sup>。一般情况可以描述如下：一个系统存在有限种可能“状态” $S_1, S_2, \dots, S_n$ 。此外，还有一组转换概率 $p_i(j)$ ，也就是当系统为状态 $S_i$ ，接下来进入状态 $S_j$ 的概率。为使此马尔可夫过程表示信源，只需要假定每次从一种状态转换到另一状态时，生成一个字符即可。这种状态对应于先前字符产生的“影响残余”。

此情景可以用图 3, 4, 5 表示。“状态”为图中的交点，转换概率和所生成的字符在相应线的旁边给出。图 3 表示第 2 节的例 B，图 4 对应于例 C。在图 3 中，由于连续字符相互独立，所以只存在一种状态。在图 4 中，所存在的状态数与字符相同。如果构造一个三连字示例，则最多存在 $n^2$ 种状态，对当前选定字符之前的可能字符对相对应。图 5 是例 D 中单词结构的对应图。这里的 S 表示“空(space)”符号。

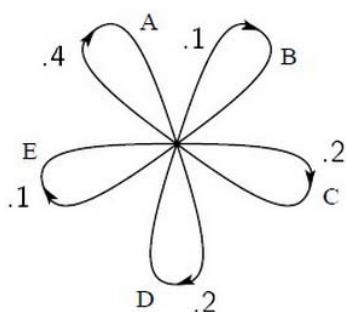


图 3 例 B 中信源的对应图

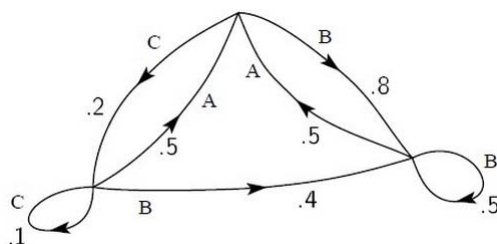


图 4 例 C 中信源的对应图

<sup>6</sup> 详细讨论请参阅 M. Fréchet, Méthode des fonctions arbitraires. Théorie des événements en chaîne dans le cas d'un nombre fini d'états possibles. Paris, Gauthier-Villars, 1938.

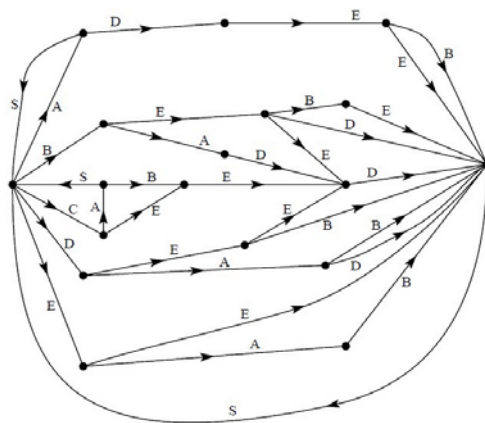


图 5 例 D 中信源的对应图

## 5. 各态历经与混合信源

上面已经指出，可以认为我们使用的离散信源能够用马尔可夫过程来表示。在各种可能存在的离散马尔可夫过程中，有一组过程拥有一些在通信理论中极为重要的特殊性质。这一特殊类别由“各态历经”过程组成，我们将相应的信源称为各态历经信源。尽管各态历经过程的严格定义有些复杂，但其一般思想却很简单。在各态历经过程中，该过程所生成的每个序列都具有相同的统计性质。因此，由特定序列获得的字符频率、二连字频率等数值，将会随着这些序列长度的增加，趋近于与该特定序列无关的确切极限。实际上，这一点并非对于所有序列都成立，但是，使之不成立的序列集的出现概率为 0。概略地说，各态历经性意味着统计意义上的均匀性。

上面给出的所有虚拟语言示例都是各态历经的。这一性质与相应图中的结构相关联。如果图形具有以下两个性质<sup>7</sup>，则相应的过程是各态历经的：

1. 图中不存在这样两个相互分离的 A、B 部分：根据箭头的方向，无法沿着图中的连线，由 A 部分的交点到达 B 部分中的交点，也无法由 B 部分的交点到达 A 部分的交点。
2. 如果图中有一系列连线闭合起来，而且连线上的所有箭头都指向相同方向，将这些闭合线称为“回路”。一个回路的“长度”就是其中连线的数目。因此，在图 5 中，BEBES 是一个长度为 5 的回路。需要满足的第二个性质是，图中所有回路长度的最大公约数为 1。

如果满足第一个条件，但不满足第二个条件，该最大公约数  $d$  大于 1，则此类序列会拥有某种特定类型的周期结构。各个序列会被划分为  $d$  个不同类别，除了相对于原点的一个偏移之外，这些类别在统计意义上是相同的（这里所说的“原点”，是指序列中被称为“字符 1”的那个字符）。任何一个序列，只要将其移动 0 到  $d - 1$  个位置，就可以使它与任意其他序列实现统计等价。下面是一个  $d = 2$  的简单示例：存在三个可能字符  $a, b, c$ 。字符  $a$  后面跟有  $b$  或  $c$  的概率分别为  $\frac{1}{3}$  和  $\frac{2}{3}$ 。 $b$  和  $c$  后面总是跟有字符  $a$ 。因此，一个典型序列为：

a b a c a c a c a b a c a b a b a c a c

这种情景在我们的研究中并不是特别重要。

<sup>7</sup> 这是针对 Fréchet 给出的条件图进行的重新表述。

如果不满足第一个条件，则可以将图形划分为两个分别满足该第一条件的子图。我们将假定对于每个子图都满足第二个条件。在此情况下，将会得到由一些单纯分量组成的所谓“混合”信源。这些分量与各个子图相对应。如果 $L_1, L_2, L_3, \dots$ 为分量信源，则可以记作：

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

其中， $p_i$ 是分量信源 $L_i$ 的概率。

在物理上，所述情景就是：存在几个不同信源 $L_1, L_2, L_3, \dots$ ，它们中的每一个都具有相同的统计结构（即，它们是各态历经的）。我们事先不知道将会使用哪个信源，但一旦在一个给定单纯分量 $L_i$ 中启动了一个序列，则它会根据该分量的统计结构无限持续下去。

作为一个例子，可以取得上面所定义过程中的两个，并假定 $p_1 = 0.2$ ， $p_2 = 0.8$ 。可以通过以下方式获得由混合信源

$$L = 0.2L_1 + 0.8L_2$$

给出的序列：首先以概率 0.2 和 0.8 选择 $L_1$ 或 $L_2$ ，在此选择之后，由所选定的任意信源生成一个序列。

除非明确给出相反表述，否则我们将假定信源是各态历经的。利用这一假定，我们可以认为一个序列的均值等于可能序列系集(ensemble)的均值（出现偏差的概率为 0）。例如，字符 A 在一个具体无限长序列中的相对频率，以概率 1 等于它在序列系集中的相对频率。

如果 $P_i$ 是状态 $i$ 的概率， $p_i(j)$ 是由状态  $i$  向状态  $j$  变换的转换概率，则对于平稳过程，显然可以得出， $P_i$ 必须满足平衡条件：

$$P_j = \sum_i P_i p_i(j)$$

在各态历经情况下，可以证明，无论起始条件如何，当 $N \rightarrow \infty$ 时，在 $N$ 个符号之后处于状态 $j$ 的概率 $P_j(N)$ 趋近于平衡值。

## 6. 选择，不确定性和熵

我们已经将离散信源表示为马尔可夫过程。我们能不能定义一个量，用来在某种意义上，度量这样一个过程“生成”多少信息？甚至更进一步，度量它以什么样的速率生成信息？

假定有一个可能事件集，这些事件的发生概率为 $p_1, p_2, \dots, p_n$ 。这些概率是已知的，但关于将会发生哪个事件，我们也就知道这么多了。我们能否找到一种度量，用来测量在选择事件时涉及多少种“选择”，或者输出中会有多少不确定性？

如果存在这样一种度量，比如说 $H(p_1, p_2, \dots, p_n)$ ，那要求它具有以下性质是合理的：

1.  $H$ 应当关于 $p_i$ 连续。
2. 如果所有 $p_i$ 都相等，即 $p_i = \frac{1}{n}$ ，则 $H$ 应当是 $n$ 的单调增函数。如果事件的可能性相等，那可能事件越多，选择或者说不确定性也更多。

3. 如果一项选择被分解为两个连续选择，则原来的 $H$ 应当是各个 $H$ 值的加权和。图 6 中展示了这一说法的含意。左侧的三个概率为 $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$ 。在右侧，我们首先以概率 $\frac{1}{2}$ 在两种可靠性之间做出选择，如果发生第二种情况，则以概率 $\frac{2}{3}, \frac{1}{3}$ 来进行另一次选择。最终结果将拥有和前面一样的概率值。在这一特例中，我们要求：

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

之所以存在系数 $\frac{1}{2}$ ，是因为这一第二选择仅在一半时间内出现。

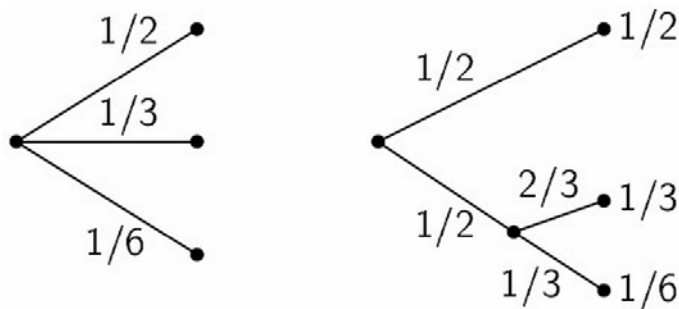


图 6 分解一个具有三种可能性的选择

在附录 2 中，将确定以下结果：

**定理 2：**唯一能够满足上述三条假定的 $H$ 具有如下形式：

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

其中 $K$ 是一个正常数。

在本文讨论的理论中，完全不需要这一定理，及其证明过程所做的一些假定。之所以给出这些内容，主要是为了提高后面一些定义的可信度。但是，这些定义的真正合理性体现在它们的推论中。

形如 $H = -\sum p_i \log p_i$ 的量（常数 $K$ 只不过是为了度量单位的选择）在信息论中扮演着核心角色，作为信息、选择和不确定性的度量。将会看出， $H$ 的这一形式就是统计力学中某些公式中定义的熵<sup>8</sup>，其中 $p_i$ 是一个系统处于其相位空间中单元格 $i$ 的概率。例如，这里的 $H$ 可以是著名的波尔兹曼 $H$ 定理中的 $H$ 。我们会将 $H = -\sum p_i \log p_i$ 称为概率集 $p_1, \dots, p_n$ 的熵。如果 $x$ 是一个随机变量，我将 $H(x)$ 记为它的熵；因此 $x$ 不是一个函数的参数，而是一个数值的记号，用于将它与 $H(y)$ 区分开， $H(y)$ 是随机变量 $y$ 的熵。

如果存在两种可能性，概率分别为 $p$ 和 $q = 1 - p$ ，则熵为：

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

将它作为 $p$ 的函数画在图 7 中。

<sup>8</sup> 比如，可参阅 R. C. Tolman, *Principles of Statistical Mechanics*, Oxford, Clarendon, 1938.

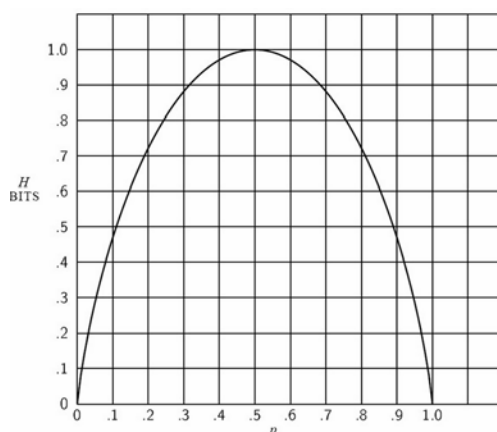


图 7 当存在两种可能性，概率分别为  $p$  和  $(1-p)$  时的熵

量  $H$  有许多重要性质，这些性质进一步表明了它作为选择度量或信息度量的合理性。

1. 当且仅当所有  $p_i$  中只有一个的取值为单位 1，其他均为 0 时， $H = 0$ 。仅当我们可以确定输出结果时， $H$  消失。否则， $H$  为正数。
2. 对于一个给定  $n$ ，当所有  $p_i$  都相等（即  $\frac{1}{n}$ ）时， $H$  达到最大值  $\log n$ 。这种情景也正是我们在直觉上感受的最不确定情景。
3. 假定要考虑两个事件  $x$  和  $y$ ，第一个事件有  $m$  种可能性，第二个事件有  $n$  种可能性。设  $p(i, j)$  是第一个事件为  $i$ 、第二个事件为  $j$  的联合概率。这一联合事件的熵为：

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

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：

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

仅当这些事件独立时（即  $p(i, j) = p(i)p(j)$ ），等号成立。一个联合事件的不确定性小于或等于各个不确定性之和。

4. 任何使概率  $p_1, p_2, \dots, p_n$  趋于相等的改变都会使  $H$  增大。因此，如果  $p_1 < p_2$ ，而且我们使  $p_1$  增大， $p_2$  减小一个相等量，使  $p_1$  和  $p_2$  更接近相等，则  $H$  增大。更一般地说，如果我们对  $p_i$  执行以下形式的任意“均化”操作：

$$p'_j = \sum_i a_{ij} p_i$$

式中， $\sum_i a_{ij} = \sum_j a_{ij} = 1$ ，且所有  $a_{ij} \geq 0$ ，则  $H$  增大（有一种特殊情景除外，在这种情景中，这种变化就是  $p_j$  的一个排列交换， $H$  当然保持不变了）。

5. 和 3 中一样，假定有两个随机事件  $x$  和  $y$ ，它们不一定相互独立。对于  $x$  可以取到的任意特定值  $i$ ，存在一个  $y$  取  $j$  值的条件概率  $p_i(j)$ 。此概率给出如下：

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

我们将  $y$  的**条件熵**  $H_x(y)$  定义为关于每个  $x$  值， $y$  的熵的加权平均，加权值为取得特定  $x$  值的概率。即：

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

这个量用来度量当我们知道  $x$  时， $y$  平均有多大的不确定性。代入  $p_i(j)$  的值，得：

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

或

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

联合事件  $x, y$  的不确定性（或熵）是  $x$  的不确定性再加上当  $x$  已知时  $y$  的不确定性。

6. 由 3 和 5 可得：

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

因此

$$H(y) \geq H_x(y)$$

$y$  的不确定性绝对不会因为知道了  $x$  值而增大。除非  $x$  和  $y$  是独立事件，否则  $y$  的不确定性会因为知道  $x$  值而下降，当两者独立时，该不确定性保持不变。

## 7. 信源的熵

考虑上面讨论的具有有限种状态的离散信源。对于每种可能存在的状态  $i$ ，都存在一组表示生成各种可能状态  $j$  的概率  $p_i(j)$ 。因此，对于每种状态都存在一个熵  $H_i$ 。信源的熵定义为这些  $H_i$  的加权平均，加权值为所考虑状态的发生概率：

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

这是信源关于每个文本符号的熵。如果马尔可夫过程以某一确定时间速率执行，则还存在一个关于每秒的熵：

$$H' = \sum_i f_i H_i$$

其中  $f_i$  是状态  $i$  的平均频率（每秒钟的发生次数）。显然：

$$H' = mH$$

其中  $m$  是平均每秒钟生成的符号数。 $H$  或  $H'$  度量该信源每个符号或每秒钟生成的信息数。如果对数底数为 2， $H$  和  $H'$  表示每个符号或每秒的比特数。

如果连续符号相互独立，则 $H$ 就是 $-\sum p_i \log p_i$ ，其中的 $p_i$ 是符号 $i$ 的概率。假定我们在此情况下考虑一个包含 $N$ 个符号的长消息。它将以很高的概率包含 $p_1 N$ 个第1符号， $p_2 N$ 个第2符号，以此类推。因此，这一特定消息的概率大约为：

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

或者，

$$\begin{aligned} \log p &\doteq N \sum_i p_i \log p_i \\ \log p &\doteq -NH \\ H &\doteq \frac{\log 1/p}{N} \end{aligned}$$

于是， $H$ 近似等于对于一个典型长序列的概率的倒数求对数，再除以该序列中的符号数。同一结果对于任意信源都是成立的。我们可以更准确地表述为（见附录3）：

**定理 3：** 给定任意 $\varepsilon > 0$ 和 $\delta > 0$ ，我们可以求得一个 $N_0$ ，使得任意长度 $N \geq N_0$ 的序列都属于以下两个类别之一：

1. 一个总概率小于 $\varepsilon$ 的集合。
2. 不属于上述集合的所有序列，这些序列的概率都满足不等式：

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta$$

换言之，当 $N$ 很大时，我们几乎可以肯定 $\frac{\log p^{-1}}{N}$ 会非常接近于 $H$ 。

一个密切相关的结果讨论了不同概率的序列数目。再次考虑长度为 $N$ 的序列，并按频率的降序排列它们。假定我们要按照序列的概率大小，从这个序列集中取出一些序列（先取概率最大者）。对于一个给定的概率值 $q$ ，我们将 $n(q)$ 定义为：要使所取序列概率总和达到 $q$ 而必须从集合中取出的序列数目。

**定理 4：**

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

其中， $q$ 不等于0和1。

当我们仅考虑总概率为 $q$ 、最可能出现的序列时，我们可以将 $\log n(q)$ 解读为指定该序列所需要的比特数。因此， $\frac{\log n(q)}{N}$ 就是在这种指定方式中，每个符号的比特数。该定理说明，这一数值与 $q$ 无关，等于 $H$ 。对于具有合理可能性的序列，无论我们如何解读“具有合理可能性”，此种序列个数的对数增长率都是 $H$ 。根据这些结果（在附录3中进行证明），在大多数情况下，我们都可以认为这些长序列仅有 $2^{HN}$ 个，每个序列的概率为 $2^{-HN}$ 。

下面的两个定理表明， $H$ 和 $H'$ 可以通过求极限运算，直接由消息序列的统计量得出，不需要涉及状态，也不需要状态之间的转换概率。

**定理 5:** 设 $p(B_i)$ 是信源中发出的一个符号序列 $B_i$ 的概率。令

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

其中，该求和运算是针对所有包含  $N$  个符号的序列  $B_i$  执行。因此， $G_N$  是  $N$  的单调减函数，且：

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

**定理 6:** 设  $p(B_i, S_j)$  是序列  $B_i$  之后跟有符号  $S_j$  的概率， $p_{B_i}(S_j) = p(B_i, S_j) / p(B_i)$  是出现  $B_i$  之后出现  $B_j$  的条件概率。令：

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

其中，该求和运算是针对所有包括  $N - 1$  个符号的块  $B_i$  及针对所有符号  $S_j$  执行的。则  $F_N$  是  $N$  的单调减函数，

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

$$G_N = \frac{1}{N} \sum_{n=1}^N F_n$$

$$F_N \leq G_N$$

及  $\lim_{n \rightarrow \infty} F_N = H。$

这些结果在附录 3 中进行推导。它们表明，只需要考虑遍布  $1, 2, \dots, N$  个符号的序列的统计结构，就可以获得对  $H$  的级数近似。 $F_N$  是更佳近似。事实上， $F_N$  是上述信源类型的  $N$  阶近似。如果统计影响不会扩展到  $N$  个以上的符号，也就是说，如果在知道前  $(N - 1)$  个符号之后，出现下一个符号的条件概率不会因为知道了更早之前的符号而发生变化，则  $F_N = H$ 。 $F_N$  当然是在已知前  $(N - 1)$  个符号时，下一个符号的条件熵，而  $G_N$  则是包含  $N$  个符号的块中每个符号的熵。

一个信源的熵除以它在使用相同符号时所能达到的最大熵，所得的比值称为**相对熵**。这是我们利用相同字符表进行编码时，可能达到的最大压缩比。1 减去相对熵的差值称为**冗余度**。如果不考虑超过大约八个字符以上的统计结构，普通英语的冗余度大约为 50%。也就是说，在我们用英文写作时，所写内容中有一半由英语的结构决定，另一半是自由选择的。由几种相互独立的方法都可以求得数字 50%，这些方法给出的结果都在这一数字附近。一种方法是计算英文近似表达的熵。第二种方法是从一段英文文本样本中删除一定比例的字符，然后让别人尝试恢复原文。如果在删除 50% 的字符后，还能恢复出原文，则冗余度一定大于 50%。第三种方法依赖于密码学中的一些已知结果。

“基础英语”和 James Joyce 的 “Finnegans Wake” 一书代表着英文散文中冗余度的两个极端。基础英语的词汇不超过 850 个单词，冗余度非常高。在将一段文字翻译为基础英文时，其篇幅会扩展许多，这一现象反映了基础英语的高冗余度。而另一方面，Joyce 扩充了词汇，据说是为了实现语义内容的压缩。



一种语言的冗余度与纵横填字字谜的存在有关。如果冗余度为 0，则任何字符序列都是该语言中的合理文本，任何二维字符阵列都可以构成一个纵横字谜。如果冗余度太高，为使大型纵横字谜成为可能，该语言要设置太多的约束条件。一种更为详尽的分析表明，如果我们假定该语言设定的约束条件在实质上是相当混乱和随机的，那在冗余度为 50% 时，大型纵横字谜游戏刚好成为可能。如果冗余度为 33%，则三维交叉字谜应当有可能实现，以此类推。

## 8. 编解码处理的表示

我们还必须用数学方式来表示发送器和接收器在对信息进行编解码时所执行的处理。发送器和接收器都将被称为转换器(transducer)。转换器接收一个符号序列（称为输入符号序列），输出另外一个符号序列（称为输出符号序列）。转换器可能具有内部存储器，使其输出不仅依赖于当前的输入符号，还依赖于过去的历史输入。我们假定内部存储器是有限的，也就是说转换器存在  $m$  种可能状态（ $m$  为一个有限数），且其输出是当前状态和当前输入符号的函数。下一个状态将是这两个量的另一个函数。因此，转换器可以用两个函数来描述：

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

其中：

$x_n$  是第  $n$  个输入符号，

$\alpha_n$  是在送入第  $n$  个输入符号时，转换器的状态，

$y_n$  是在状态为  $\alpha_n$  时送入  $x_n$  时生成的输出符号（或者输出符号序列）。

如果一个转换器的输出符号与一个第二转换器的输入符号相同，可以将这两个转换器串联在一起，所得到的装置还是一个转换器。如果存在一个第二转换器，它对第一个转换器的输出进行处理，并恢复出原来的输入内容，则说第一个转换器是非奇异的，将第二个转换器称为它的逆。

**定理 7：** 一个有限状态转换器在由有限状态的统计信源驱动时，其输出是一个有限状态统计信源，其熵（每单位时间）小于或等于输入的熵。如果转换器是非奇异的，则两者相等。

令  $\alpha$  表示信源的状态，该信源生成一个符号序列  $x_i$ ；令  $\beta$  表示转换器的状态，该转换器在其输出中生成符号块  $y_j$ 。这个组合系统可以用  $(\alpha, \beta)$  对的“积状态空间”表示。如果  $\alpha_1$  可以生成一个  $x$ ，将  $\beta_1$  变为  $\beta_2$  则可以用一条线将该空间中的两个点  $(\alpha_1, \beta_1)$  和  $(\alpha_2, \beta_2)$  连在一起，并在线上给出在此情况下出现  $x$  的概率。在线上标出由转换器生成的符号块  $y_j$ 。可以针对各个状态求加权和来计算输出的熵。如果首先针对  $\beta$  求和，则所得到的每个项都小于或等于针对  $\alpha$  求出的相应项，因此，这个熵不会增大。如果转换器是非奇异的，则将其输出连接到逆转转换器。如果  $H'_1, H'_2$  和  $H'_3$  分别是信源、第一转换器和第二转换器的输出熵，则  $H'_1 \geq H'_2 \geq H'_3 = H'_1$ ，因此可得  $H'_1 = H'_2$ 。

假定我们对于可以出现的序列类型设定了一套约束条件，可以用图 2 中的线图加以表示。如果向各个将状态  $i$  连至状态  $j$  的线上指定概率  $p_{ij}^{(s)}$ ，它将会变成一个信源。存在一种特定的指定方式，使所得到的熵取最大值（见附录 4）。

**定理 8:** 将一套约束条件看作一个信道，设其容量为  $C = \log W$ 。如果我们指定：

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

其中， $l_{ij}^{(s)}$  是由状态  $i$  导致状态  $j$  的第  $s$  个符号的持续时间， $B_i$  满足：

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

则， $H$  为最大值，并等于  $C$ 。

通过正确地指定转换概率，一个信道上的符号的熵可以达到最大值，也就是信道容量。

## 9. 无噪声信道的基本定理

我们现在将证明  $H$  确定了在最高效编码方式下所需要的信道容量，从而表明我们将  $H$  解读为信息的生成速率是合理的。

**定理 9:** 设一个信源的熵为  $H$ （比特/符号），一个信道的容量为  $C$ （比特/秒）。有可能采用某种方式对该信源的输出进行编码，从而可以在该信道上以  $\frac{C}{H} - \epsilon$  符号/秒的平均速率上进行传送，其中  $\epsilon$  为任意小。不可能以大于  $\frac{C}{H}$  的平均速率进行传送。

反过来表述这一定理就是： $\frac{C}{H}$  不可能被超越。它可以这样证明：由于传送器必然是非奇异的，所以每秒钟输入给信道的内容的熵等于信源的熵，而这个熵不可能超过信道容量。因此， $H' \leq C$ ，每秒的符号数  $H'/H \leq C/H$ 。

定理的第一部分可以用两种不同方式证明。第一种方法是考虑信源生成的所有  $N$  符号序列的集合。当  $N$  很大时，我们可以将这些序列分为两组，一组内的成员数小于  $2^{(H+\eta)N}$ ，第二组的成员数小于  $2^{RN}$ （其中  $R$  是不同符号数的对数），而且其总概率小于  $\mu$ 。随着  $N$  的增大， $\eta$  和  $\mu$  趋近于 0。信道中持续时间为  $T$  的信号数大于  $2^{(C-\theta)T}$ ，当  $T$  很大时， $\theta$  很小。如果我们选择：

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

则当  $N$  和  $T$  足够大时（无论  $\lambda$  多么小）时，总有足够多的信道符号序列可供高概率组使用，而且还会有一些富余。可以采用任意一对一方式将高概率组编码到这个集合中。剩下的序列用较长的序列表示，这些较长序列的开头和结尾是一个未为高概率组使用的序列。这一特殊序列充当着不同代码的起始与结束标志。只要有足够的时间，就可以为所有低概率消息提供足够的不同序列。这就要求：

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

其中， $\varphi$  很小。由此得出，平均传送速率（单位：消息符号/秒）将大于：

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

随着 $N$ 的增大,  $\delta, \lambda, \varphi$ 趋近于 0, 该速率趋近于 $\frac{C}{H}$ 。

另一种执行这一编码过程并证明该定理的方法可以描述如下: 按概率的递减顺序排列长度为 $N$ 的消息, 并假定它们的概率为 $p_1 \geq p_2 \geq p_3 \dots \geq p_n$ 。令 $P_s = \sum_{i=1}^{s-1} p_i$ ; 即,  $P_s$ 是一直到 $p_s$  (但不包含) 的累积概率。我们首先将其编码为二进制。将 $P_s$ 扩展为一个二进制数, 即可得到消息  $s$  的二进制编码。这一扩展一直执行到 $m_s$ 位, 其中 $m_s$ 是满足下式的整数:

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

因此, 高概率消息用短代码表示, 低概率消息用长代码表示。由这些不等式可以得到:

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

$P_s$ 的代码与它前面的所有代码相比, 在 $m_s$ 位中的一个或多个位置有所不同。由于所有其余 $P_i$ 都至少要大出 $\frac{1}{2^{m_s}}$ , 因此, 它们的二进制扩展会在前 $m_s$ 位中有所不同。因此, 所有这些编码都是不同的, 从而有可能从消息的编码中恢复出消息。如果信道序列还不是二进制数位序列, 则可以采用任意一种方式将它们表示为二进制数, 然后再将二进制编码转换为适于信道传送的信号。

原消息中每个符号所使用的平均二进制数位数 $H'$ 很容易估计出来。有:

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

但是,

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

因此,

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

随着 $N$ 的增大,  $G_N$ 趋近于 $H$ , 即信源的熵,  $H'$ 也趋近于 $H$ 。

由此可以看出, 当仅使用 $N$ 个符号的有限延迟时, 编码中的低效性不一定大于 $\frac{1}{N}$ 加上 $H$ 与 $G_N$ 之差, 其中 $H$ 为真实熵,  $G_N$ 是为长度为 $N$ 的序列计算所得的熵。因此, 所需时间超出理想值的百分比小于:

$$\frac{G_N}{H} + \frac{1}{HN} - 1。$$

这一方法基本上与 R. M. Fano 独立发现的方法相同。他的方法是将长度为  $N$  的消息按照概率的降序排列。尽可能将这一系列消息划分为两个概率接近相等的组。如果消息属于第一组，则它的第 1 个二进制数位为 0，否则为 1。然后再采用类似方式，进一步将这些组分为概率近似相等的子集，具体的子集决定了第 2 个二进制数位。继续这一过程，直到每个子集中仅包含一条消息为止。容易看出，除了微小的差别之外（通常是在最后一位），这一方法与上述算术过程基本相同。

## 10. 讨论与示例

为从发电机向负载传送最大功率，通常必须引入变压器，使得从负载的角度来看时，发电机的阻抗与负载相同。这里的情景大体类似。执行编码的转换器应当使信源在统计意义上与信道相匹配。经过转换器，从信道的角度看，信源的统计结构应当与实现信道容量最大化的信源相同。定理 9 中的内容是说，尽管通常不可能实现完全匹配，但可以使它们达到任意接近程度。实际传送速率与容量  $C$  的比值可称为该编码系统的效率。它当然等于信道符号的实际熵除以最大可能熵的比值。

一般来说，理想编码或接近理想的编码都需要在发送器和接收器中有较长时间的延时。在我们正在考虑的无噪声情况下，这一延迟的主要功能是让序列的概率与其相应长度能够很好地匹配。通过好的编码方式，一条长消息的概率求倒数，再求对数，所得的结果必然与相应信号的持续时间成正比，事实上，对于大多数长消息而言，除了很少一部分之外，

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

必然很小。

如果一个信源只能生成一条特定的消息，则其熵为 0，根本就不需要信道。例如，如果一个计算机被设置用于计算  $\pi$  的连续数位，它会生成一个没有随机成分的确切序列。不需要信道就可以将这个序列由一点“传送”到另一点。人们可以在该点构造一个相同的第二机器来计算同一序列。但是，这可能不太现实。在这种情况下，我们可以选择忽略部分或全部有关信源的已知统计信息。既然我们要设计一个能够传送任意数位序列的系统，那可以将  $\pi$  的数位看作一个随机序列。采用类似方式，在设计代码时，可以选择使用有关英文的一部分已知统计信息，而不使用全部相关信息。在这种情况下，我们认为具有最大熵的信源满足我们希望保持的统计条件。这一信源的熵决定了充足、必须的信道容量。在  $\pi$  的例子中，唯一保留的信息就是所有数位都是从集合  $0, 1, \dots, 9$  中选出的。在英文中，可能希望利用由字符频率信息所带来的统计好处，但不再希望利用其他信息。因此，该最大熵信源就是英文的一阶近似，它的熵决定了所需要的信道容量。

作为一部分此类结果的简单例子，考虑一个信源，它生成一个序列，序列中的字符是以概率  $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$  从 A, B, C, D 中选出的，连续符号之间的选择相互独立。得：

$$\begin{aligned} H &= -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{2}{8}\log\frac{1}{8}\right) \\ &= \frac{7}{4} \text{ 比特/秒} \end{aligned}$$

因此，我们可以模拟一个编码系统，将来自这个信源的消息编码为二进制数位，每个符号平均占 $\frac{7}{4}$ 个二进制位。在此情况下，我们实际上可以利用以下编码达到极限值（这一编码方式是利用定理 9 第二种证明中的方法得到的）：

A	0
B	10
C	110
D	111

对一个 N 符号序列进行编码所使用的平均二进制位数等于：

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

容易看出，二进制数位的概率为 $\frac{1}{2}, \frac{1}{2}$ ，所以编码后所得序列的 H 为 1 比特/符号。由于平均来说，每个原始字符有 $\frac{7}{4}$ 个二进制符号，所以从时间上来说，这两个熵是相同的。原集合的最大可能熵为  $\log 4 = 2$ ，当 A, B, C, D 的概率为 $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ 时取此最大值。因此，相对熵为 $\frac{7}{8}$ 。我们可以利用下表，基于两对一的对应关系，将二进制序列转换为原来的符号集：

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

这个双重过程用相同符号对原来消息进行了编码，但其平均压缩比为 $\frac{7}{8}$ 。

作为第二个例子，考虑一个信源，它生成由 A 和 B 组成的一个序列，其中 A 的概率为 p，B 的概率为 q。如果  $p \ll q$ ，则有：

$$\begin{aligned} H &= -\log p^p (1-p)^{1-p} \\ &= -p \log p (1-p)^{(1-p)/p} \\ &\doteq p \log \frac{e}{p} \end{aligned}$$

在这种情况下，我们可以设计一种在 0, 1 信道上相当出色的消息编码方式：先为不经常出现的 A 发送一个特殊序列，比如说 0000，然后发送一个序列，表示在 A 之后有多少个 B。这一数值可以用其二进制来表示，但所使用的二进制数字中不能包含特殊序列。所有小于 16 的数字都用其通常的二进制数来表示；但对于 16，则用 16 之后第一个不包含四个 0 的二进制数来表示，即 17=10001，以此类推。

可以证明，只要正确调整特殊序列的长度，当  $p \rightarrow 0$  时，该编码方式就接近于理想编码。

## 第 II 部分：有噪声离散信道

### 11. 有噪声离散信道的表示

我们现在考虑信号在传输过程中或者在某终端处受到噪声干扰的情景。这意味着，接收信号不一定与发送器发出的信号相同。可以区分两种情景。如果一个特定的传送信号总是生成相同的接收信号，即接收信号是传送信号的确定函数，可以将这种影响称为“失真”。如果这一函数有逆函数——任何两个传送信号都不会生成相同的接收信号，只需要对接收信号执行逆函数运算，就可以校正该失真，至少在原理上如此。

这里关心的情景是：在传输过程中，信号变化并非总是一致的。在这种情况下，我们可以假定接收信号  $E$  是传送信号  $S$  及一个第二变量的函数，这个第二变量就是噪声  $N$ 。

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

和前面的消息一样，可以认为噪声是一个随机变量。一般情况下，它可以用一个适当的随机过程来表示。我们将要考虑的最一般有噪声离散信道类型是前面介绍的有限状态无噪声信道的一种推广。我们假定状态数目有限，概率集合为：

$$p_{\alpha,i}(\beta, j)。$$

这一概率是指，如果信道处于状态  $\alpha$  并传送符号  $i$ ，则接收到符号  $j$ ，且信道变为状态  $\beta$  的概率。因此， $\alpha$  和  $\beta$  的取值范围为所有可能状态， $i$  的取值范围为所有可能发送的信号， $j$  的取值范围是所有可能接收的信号。当连续信号受到噪声的干扰相互独立时，只有一种状态，可以用一组转换概率  $p_i(j)$  来描述，也就是将发送信号  $i$  接收为  $j$  的概率。

如果一个有噪声信道从一个信源那里接收内容，则有两个统计进程在发挥作用：信源和噪声。因此，可以计算许多种熵。首先，存在一个信源的熵  $H(x)$ ，或者是信道输入的熵（如果发送器是非奇异的，则这两个熵相等）。信道输出（也就是接收信号）的熵用  $H(y)$  表示。在无噪声情况下， $H(y) = H(x)$ 。输入与输出的联合熵为  $H(xy)$ 。最后，有两个条件熵  $H_x(y)$  和  $H_y(x)$ ，分别是在输入已知时输出的熵和输出已知时输入的熵。在这些量中，存在如下关系：

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

所有这些熵都可以基于“每秒”或“每符号”来计算。

### 12. 疑义度(equivocation)与信道容量

如果信道中存在噪声，一般情况下，无论对接收信号  $E$  执行什么处理，都无法**确定地**恢复出原始消息或发送信号。但是，有一些信息传送方式最适于对抗噪声。这就是我们现在要考虑的问题。

假定可能出现两个符号 0 和 1，我们正在以每秒 1000 个符号的速率传送符号，其概率为  $p_0 = p_1 = \frac{1}{2}$ 。因此，我们的信源正在以每秒 1000 比特的速率生成信息。在传输期间，噪声引入了误差，平均有百分之一的符号接收错误（发 0 收 1，或发 1 收 0）。信息的传送速率是多少呢？由于大约有 1% 的接收符号是错误的，所以信

息传送速率当然小于 1000 比特/秒。我们的第一冲动可能会说：该速率为 990 比特/秒，就是减去误差的期望值。这一结果是不能令人满意的，因为它没有考虑到接收方并不知道这些误差会发生在哪里。我们可以将它延伸到一种极限情景，假定噪声是如此之大，接收到的符号完全与发送的信号无关。无论传送的是什么符号，收到 1 的概率总是  $\frac{1}{2}$ ，收到 0 的情景也与此类似。因此，仅因为机缘巧合，大约有一半的接收符号是正确的，我们可能会说这个系统每秒传送 500 比特的信息，而实际上根本就没有传送任何信息。我们完全可以丢开这个信道，只需要在接收点抛掷硬币，就可以获得一个同样“出色”的传送过程。

显然，需要对发送信息量进行适当的修正，修正值就是接收信号中丢失的信息量，或者说，是当我们接收到实际发送内容的信号时存在的不确定性。在我们前面的讨论中，将熵作为不确定性的一种度量，在知道接收信号后，将消息的条件熵用作这一损失信息的度量看起来是合理的。后面将会看到，这实际上就是一个恰当的定义。根据这一思想，从生成速率（也就是信源的熵）中减去条件熵平均速率，就可以得到实际传送速率  $R$ 。

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

为方便起见，我们将条件熵  $H_y(x)$  称为“疑义度”。它度量了接收信号的平均模糊度。

在上面考虑的示例中，如果收到 0，则传送符号为 0 的后验概率为 0.99，传送符号为 1 的后验概率为 0.01。如果收到 1，则这些数值颠倒过来。因此：

$$\begin{aligned} H_y(x) &= -[0.99 \log 0.99 + 0.01 \log 0.01] \\ &= 0.081 \text{ 比特/符号} \end{aligned}$$

或者说 81 比特/秒。我们可以说，该系统是在以  $1000 - 81 = 919$  比特/秒的速率进行传送。在极端情况下，发送 0 时接收到 0 或 1 的概率相等，发送 1 时也是如此，则后验概率为  $\frac{1}{2}$ ， $\frac{1}{2}$ ，则

$$\begin{aligned} H_y(x) &= -\left[\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}\right] \\ &= 1 \text{ 比特/符号} \end{aligned}$$

或者说 1000 比特/秒。于是，传输速率为 0，和实际情况一致。

下面的定义给出了疑义度的直观解读，还用于证明它作为惟一适当度量的正当性。我们考虑一个通信系统和一位观察者（或者辅助设施），这位观察者既能看到发送内容，又能看到接收内容（接收内容中包含因为噪声导致的错误）。这位观察者记下接收消息中的错误，并通过一个“校正信道”向接收端传送数据，使接收器能够纠正这些错误。这一情景在图 8 中示意给出。

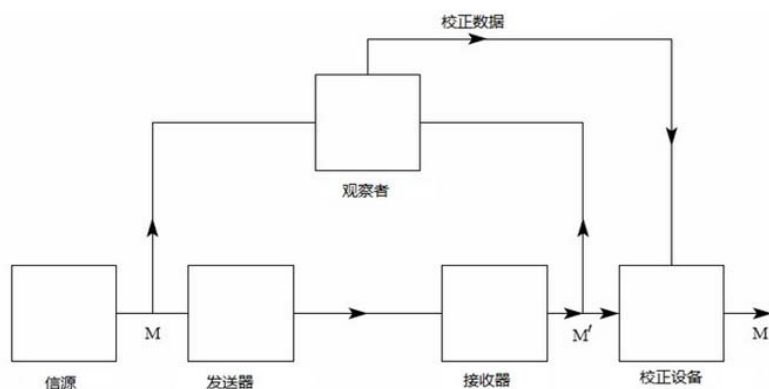


图 8 校正系统的示意图

**定理 10:** 如果该校正信道的容量为  $H_y(x)$ ，则有可能对校正数据进行适当编码，以通过这一信道发送出去，除了任意小的一部分错误  $\epsilon$  之外，所有其他错误都可以被纠正。如果校正信道的容量小于  $H_y(x)$ ，则不可能实现。

大体来说， $H_y(x)$  就是为了校正接收消息，在接收点每秒钟必须提供的附加信息量。

为了证明第一部分，考虑接收消息  $M'$  和相应原始消息  $M$  的长序列。以对数方式进行表示，有  $TH_y(x)$  个  $M$  可以合理地生成每个  $M'$ 。因此，每  $T$  秒中有  $TH_y(x)$  个二进制数位要传送。在容量为  $H_y(x)$  的信道上，能够以  $\epsilon$  的错误频率来完成这一任务。

第二部分的证明如下：首先注意到对于任意离散随机变量  $x, y, z$ ,

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

可以将左侧展开，得到：

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

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

如果我们把  $x$  看作信源的输出， $y$  看作接收信号， $z$  看作通过校正信道发送的信号，则右侧就等于疑义度减去校正信道的传送速率。如果这个信道的容量小于疑义度，则右侧将大于 0，且  $H_{yz}(x) > 0$ 。但这是在接收信号和校正数据都已知时，所发送的不确定性。如果它大于 0，则错误频率就不可能是任意小。

**示例：**

假定在一个二进制数位组成的序列中，随机出现错误：一个数位发生错误的概率为  $p$ ，正确的概率为  $q = 1 - p$ 。如果这些错误的位置已知，则可以进行纠正。因此，校正信道只需要发送有关这些错误位置的信息。这就相当于要传送一个信源生成的二进制数位，这些数据是 1 的概率为  $p$ （传送错误），是 0 的概率为  $q$ （传送正确）。这就一个容量为

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

这就是原系统的疑义度。



由于上面给出的等式，传送速率 $R$ 可以记为其他两种形式。有：

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

第一个定义表达式已经解读为传送的信息量减去所发送内容的不确定性。第二表达式度量的是接收内容减去由于噪声导致的部分。第三个表示式是两个数量之和减去联合熵，在某种意义上，就是两者每秒钟共有的比特数。因此，这三种表达式的意义都与直觉一致。

一个有噪声信道的容量 $C$ 应当是可能实现的最大传送速率，也就是说，在信源与信道恰好匹配时的速率。因此，我们将信道容量定义为：

$$C = \max(H(x) - H_y(x))$$

其中的最大值，是相对于所有可以用作信道输入的信源求得。如果信道是无噪声的，则 $H_y(x) = 0$ 。于是，这一定义与前面已经针对无噪声信道给出的定义等价，因为信道的最大熵就是它的容量。

## 13. 有噪声离散信道的基本定理

要定义一个有噪声信道的容量 $C$ ，这听起来似乎有些让人惊讶，因为在有噪声情况下，我们永远都无法确定地传送信息。但是，非常明确的是，如果以某种冗余方式来发送信息，那就可以减少出现错误的概率。例如，将消息重复发送许多次，并对所接收消息的不同版本进行统计研究，就可以将错误概率降到非常低。不过，有人可能希望让这一错误概率趋近于0。这绝对无法实现。如果能够实现，那就不存在一个定义严谨的容量了，只有相对于一个给定错误频率或者给定疑义度的容量；当错误标准越来越严格时，该容量就会下降。实际上，上面定义的容量 $C$ 有非常明确的重要性。通过正确编码，有可能通过该信道以速率 $C$ 传送信息，而**错误频率或疑义度可以小到令人满意**。而这一表述对于任何大于 $C$ 的速率都是不成立的。如果尝试以一个高于 $C$ 的速率进行传送，比如 $C + R_1$ ，则必然存在一个等于或大于 $R_1$ 的疑义度。而这是需要付出代价的，因为这样会具有更多的不确定性，我们实际上无法正确地获得大于 $C$ 的速率。

图9中描述了这一情景。输入信道的信息速率画在横轴上，疑义度画在纵轴。阴影区域中任何高于粗线的点都可以实现，而下面的点则无法实现。这条粗线上的点通常也是无法实现的，其中两个点通常是可以实现的。

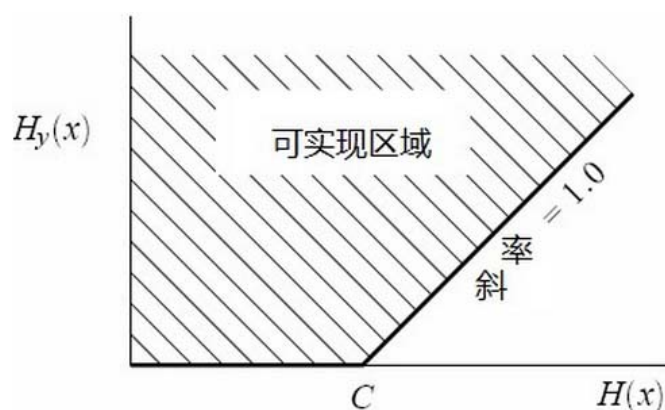


图 9 信道输入的熵给定时，可能实现的疑义度

这些结果是表明 $C$ 定义合理性的主要理由，现在我们将对其进行证明。

**定理 11:** 设一个离散信道的容量为  $C$ ，一个离散信源的熵（每秒）为  $H$ 。如果  $H \leq C$ ，则存在一个编码系统，可以通过该信道传送该信源的输出内容，而错误频率达到任意小（或者说，疑义度任意小）。如果  $H > C$ ，则有可能对信源进行编码，使得疑义度小于  $H - C + \epsilon$ ，其中  $\epsilon$  为任意小。不存在能够使疑义度小于  $H - C$  的编码方法。

为了证明这一定义的第一部分，并不是给出一个具有期望性质的编码方法，而是证明在特定的代码组中，必然存在这样一种编码。事实上，我们会将错误频率在这个组中求平均，从而证明这个平均值可以小于  $\epsilon$ 。如果一个数组的平均值小于  $\epsilon$ ，则这个集合中必然存在至少一个数字小于  $\epsilon$ 。这样就可以确定所需要的结果。

有噪声信道的容量  $C$  已经定义如下：

$$C = \max(H(x) - H_y(x))$$

其中， $x$  为输入， $y$  为输出。此最大值是针对所有可以用作信道输入的信源求取的。

设  $S_0$  是实现最大容量  $C$  的信源。如果任何信源都不能实际达到这个最大值，则设  $S_0$  是最接近能够给出最大速率的信源。假定  $S_0$  用作该信道的输入。我们考虑一个长持续时间  $T$  的传送与接收序列。以下表述是成立的：

1. 传送信号分为两组，一个高概率组，大约有  $2^{TH(x)}$  个成员，其余序列的总概率非常小。
2. 类似的，接收信号也有一个大约有  $2^{TH(y)}$  个成员的高概率集合，其余序列组成一个小概率集合。
3. 每个高概率输出可以由大约  $2^{TH_y(x)}$  个输入生成。所有其他情景的总概率很小。

这些表述中的“小”和“大约”等用词可以用  $\epsilon$  和  $\delta$  来表示，如果允许  $T$  增大， $S_0$  趋近于可实现最大值的信源，则所有  $\epsilon$  和  $\delta$  都趋近于 0。

这一情景总结于图 10 中，其中，输入序列为左侧的点，输出序列为右侧的点。交叉线组成的扇形表示可能一个典型输出的原因范围。

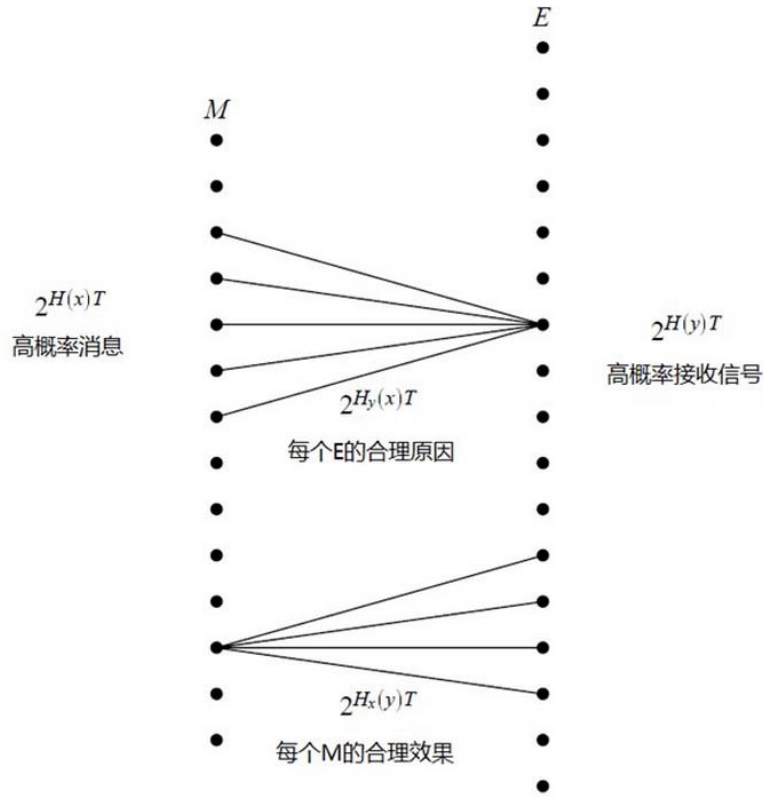


图 10 信号输入与输出关系的示意表示

现在假定我们有另外一个以速率 $R$ 生成信息的信源，其中 $R < C$ 。在周期 $T$ 中，这一信源有 $2^{TR}$ 条高概率消息。我们希望以某种方式将这些消息与可以作为信道输入的一组选择关联起来，以实现小的错误频率。我们将以所有可能方式来设定这一关联（但是，仅使用由信源 $S_0$ 确定的高概率输入组），并针对这一大类可能存在的编码系统求平均错误频率。这相当于将这些消息与持续时间为 $T$ 的信道输入随机关联起来时，计算误差频率。假定观察到一个特定的输出 $y_1$ 。在可能导致 $y_1$ 的原因集中，超过一条消息的概率为多少呢？在 $2^{TH(x)}$ 个点中随机分布着 $2^{TR}$ 条消息。因此，一个特定点被选作消息的概率为：

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

扇形中所有点都不是消息（真正的原始消息除外）的概率为：

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

现在， $R < H(x) - H_y(x)$ ，所以 $R - H(x) = -H_y(x) - \eta$ ，其中 $\eta$ 为正数。相应地，当 $T \rightarrow \infty$ 时：

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

趋近于

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

因此，错误概率趋近于 0，定理的第一部分得证。

注意到，我们只能以 $C$ 比特/秒的速率发送来自信源的内容，完全忽略所生成的其余信息，由此容易证明该定理的第二部分。在接收器处，被忽略部分给出的疑义度为 $H(x) - C$ ，所传送部分只需要增加 $\epsilon$ 。采用许多其他方法都可以达到这一极限，在考虑连续情景时将会看到这一点。

该定理的最后一种表述可以由 $C$ 的定义直接推出。假定我们可以采用某种方式对信源以 $H(x) = C + a$ 进行编码，以获得疑义度 $H_y(x) = a - \epsilon$ ，其中 $\epsilon$ 为正数。因此，有 $R = H(x) = C + a$ ，且：

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

其中 $\epsilon$ 为正数。而根据 $C$ 的定义，它是 $H(x) - H_y(x)$ 的最大值，矛盾。

实际上，我们所证明的内容并不仅限于定理中的所述内容。如果一个数字集合的平均值在其最大值的 $\epsilon$ 范围内，则这些数字中，最多有 $\sqrt{\epsilon}$ 部分与最大值之差大于 $\sqrt{\epsilon}$ 。由于 $\epsilon$ 是任意小，所以我们可以说几乎所有这些系统都任意接近于理想编码。

## 14. 讨论

定理 11 的证明方式尽管不是纯粹的存在性证明，但却具有此类证明的一些缺陷。如果尝试按照证明中的方法来很好地模拟理想编码，通常是不太现实的。事实上，除了一些相当少见的情景和特定的极限情景外，还没有发现一种明确的描述方式，能够给出理想编码的近似方式。这可能并不是一种意外，而是因为很难明确地构造出随机序列的良好近似。

一种理想随机序列的近似应当具有以下性质：如果噪声以一种合理方式改变了信号，则应当仍然能够恢复出原信号。换句话说，经过这一改变后，所得信号与其他合理信号的差别通常不应小于与原信号的差别。完成这一操作的代价是在编码中有一定数量的冗余。必须以正确方式引入冗余，以对抗所涉及的特定噪声结构。具体来说，如果该信源已经具有一定的冗余，而且没有尝试消除该冗余，以与信道匹配，则这一冗余度可以帮助对抗噪声。例如，在无噪声电报信道中，通过对消息的正确编码，可以节省大约 50% 的时间。人们并没有做这一项工作，英语中的大多数冗余仍然保留在信道符号中。但是，这样也有一个优点：允许信道中存在相当大的噪声。可能会错误收到相当数量的字符，因为英文的统计结构是相当棘手的，合乎情理的英文序列与随机选择的距离并不是特别遥远（这里所说的距离是指满足定理需求的程度）。

和在无噪声情景中一样，通常需要有一定的延迟才能接近理想编码。延迟现在还有另外一项功能：在有很大的噪声取样对信号产生影响之后，仍然可能在接收点将接收信号判定为原消息。增大采样大小，总是可以强化可能存在的统计断言。

可以采用一种稍有不同的方式来描述定理 11 的内容及其证明，这种方法更清楚地呈现了它与无噪声情景的联系。考虑持续时间为  $T$  的可能信号，并假定选择其中的一个子集来使用。假设此子集中所有信号的使用概率相同，并假定设计一种接收机，在接收到受干扰信号时，从该子集中选择最可能导致该信号的项目作为原始信号。我们可以为该子集选择一些信号，使得错误判定的概率小于或等于  $q$ ，将  $N(T, q)$  定义为满足此条件的最大信号数目。

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

换言之，无论我们设定什么样的可靠性极限，只要 $T$ 足够大，我们就可以在时间 $T$ 内可靠地区分出足够多的消息，以与大约 $CT$ 比特相对应。可以将定理 12 与第 1 节给出的无噪声信道容量的定义进行对比。

## 15. 离散信道示例及其容量

图 11 给出了一个离散信道的简单示例。共存在三种可能符号。第一个符号永远不会受噪声影响。第二个和第三个符号正确传送的概率为  $p$ ，相互转换的概率为  $q$ 。（令  $\alpha = -[p \log p + q \log q]$ ， $P$  和  $Q$  分别为使用第一个和第二个符号的概率，）则有：

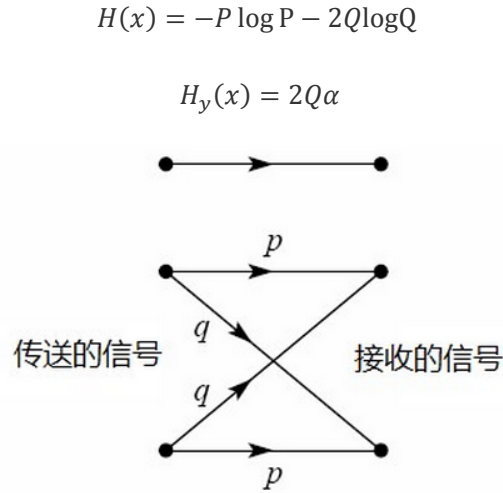


图 11 离散信道示例

我们希望以某种方式来选择 $P$ 和 $Q$ ，在 $P + 2Q = 1$ 的约束条件下，使 $H(x) - H_y(x)$ 达到最大。因此，我们考虑

$$U = -P \log P - 2Q \log Q - 2Q\alpha + \lambda(P + 2Q)$$

$$\frac{\partial U}{\partial P} = -1 - \log P + \lambda = 0$$

$$\frac{\partial U}{\partial Q} = -2 - 2 \log Q - 2\alpha + 2\lambda = 0$$

消去 $\lambda$

$$\log P = \log Q + \alpha$$

$$P = Qe^\alpha = Q^\beta$$

$$P = \frac{\beta}{\beta + 2}, \quad Q = \frac{1}{\beta + 2}$$

于是，信道容量为：

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

注意这一结果在 $P = 1$ 和 $P = \frac{1}{2}$ 的情况下与一眼即可看出的结果绝对吻合。在第一种情况下， $\beta = 1$ ， $C = \log 3$ ，这是正确的，因为这个信道是无干扰的，共有三个可能符号。如果 $P = \frac{1}{2}$ ，则 $\beta = 2$ ， $C = \log 2$ 。这时根本无

法区分第二个和第三个符号，它们好像是同一个符号。第一个符号的使用概率为 $P = \frac{1}{2}$ ，第二个和第三个加起来的概率为 $P = \frac{1}{2}$ 。可以采用任何期望方式在它们之间分布，仍然能够实现最大容量。

对于 $p$ 的中间值，信道容量将介于 $\log 2$ 和 $\log 3$ 之间。第二个符号和第三个符号之间的区别会传送一些信息，但不像无噪声情况下那么多。由于第一个符号不受噪声干扰，所以其使用频率要高于其他两个符号。

## 16. 某些特殊情况下的信道容量

如果噪声对连续信道符号的影响是独立的，则可以用一组转换概率 $p_{ij}$ 来描述。此概率就是发送符号 $i$ ，收到符号 $j$ 的概率。最大信道速率可用下式的最大值给出：

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

其中，我们改变 $P_i$ ，但保持 $\sum P_i = 1$ 。由拉格朗日方法可得到以下方程，

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

乘以 $P_s$ ，并针对 $s$ 求得，可以证明 $\mu = C$ 。设 $p_{sj}$ 的逆（如果存在的话）为 $h_{st}$ ，使得 $\sum_s h_{st} p_{sj} = \delta_{tj}$ 。则：

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

因此，

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

或：

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

这就是用于确定 $P_i$ 最大值的方程组，其中需要确定 $C$ ，使得 $\sum P_i = 1$ 。在完成这一工作后， $C$ 为信道容量， $P_i$ 是实现这一容量的信道符号的正确概率。

如果对于每个输入符号，由其引出的直线上都有相同的概率集，而且对于每个输出符号也是如此，就可以很轻松地计算出这一容量。图 12 中给出了一些示例。在这种情况下， $H_x(y)$ 与输入符号上的概率分布无关，由 $\sum p_i \log p_i$ 给出，其中 $p_i$ 是由任意输入符号发出的转换概率值。信道容量为：

$$\max[H(y) - H_x(y)] = \max H(y) + \sum p_i \log p_i$$

$H(y)$ 的最大值显然为 $\log m$ （其中 $m$ 是输出符号的数目），这是因为，如果使输入符号的可能性相等，则有可能使输出符号的概率也都相等。因此，信道容量为：

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

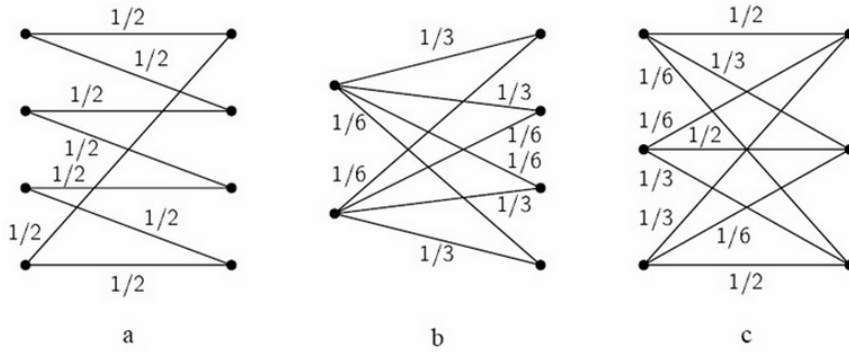


图 12 每个输入、每个输出的转换概率相同的离散信道示例

在图 12a 中，信道容量为：

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

仅使用第一、第三符号即可实现这一取值。在图 12b 中：

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

在图 12c 中，有：

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

假定这些符号被分为几组，噪声从来不会导致一个组中的符号被错认为另一个组中的符号。如果我们仅使用第  $n$  组中的符号，则该组的容量为  $C_n$ （单位为比特/秒）。因此，容易证明，为了达到整个集合的最佳使用效果，第  $n$  组中所有符号的总概率  $P_n$  应当是：

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

在一个组内，概率的分布与不存在其他符号时相同。信道容量为：

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

## 17. 高效编码举例

下面的例子尽管有些不太现实，但却有可能实现与有噪声信道的完全匹配。存在两个信道符号 0 和 1，噪声对七个该等符号组成的块产生影响。一组七个符号要么无错误传送，要么恰有一个符号错误。这八种可能性是等概率的。有：

$$\begin{aligned} C &= \max[H(y) - H_x(y)] \\ &= \frac{1}{7} \left[ 7 + \frac{8}{8} \log \frac{1}{8} \right] \\ &= \frac{4}{7} \text{ 比特/符号} \end{aligned}$$

下面给出一种高效编码，它能够完全纠正所有错误，并以速率  $C$  进行传送（此种编码方式是由 R. 汉明发明的一种方法发现的）：

设一组七个符号为  $X_1, X_2, \dots, X_7$ 。其中的  $X_3, X_5, X_6$  和  $X_7$  为消息符号，由信源任何选择。其他三个为冗余，其计算如下：

选择  $X_4$ ，使得  $\alpha = X_4 + X_5 + X_6 + X_7$  为偶数

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

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

在接收到一个七符号块后，计算  $\alpha, \beta$  和  $\gamma$ ，如果结果为偶数则记为 0，如果为奇数则记为 1。二进制数字  $\alpha\beta\gamma$  将给出错误  $X_i$  的下标（如果此二进制数为 0，则表示没有错误）。

## 附录 1：在有限状态条件下，符号块数目的增长率

设  $N_i(L)$  是长度为  $L$ 、最终以状态  $i$  结束的符号块的长度。于是，可得：

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

其中  $b_{ij}^1, b_{ij}^2, \dots, b_{ij}^m$  是在状态  $i$  下被选定并导致状态  $j$  的符号的长度。这是一些线性方程组，当  $L \rightarrow \infty$  时，其特性必然为以下类型：

$$N_j = A_j W^L$$

代入差分方程

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

或

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



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

为使其可能成立，行列式：

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

必须为 0，这就确定了  $W$ ，当然，它是  $D = 0$  的最大实根。

量  $C$  由下式给出：

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

我们还注意到，如果要求所有块的起始状态都相同（任意选定），则会得到相同的增长性质。

## 附录 2： $H = -\sum p_i \log p_i$ 的推导

设  $H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = A(n)$ 。根据条件 (3)，我们可以将在  $s^m$  中进行一次等概率选择，分解为在  $s$  中进行  $m$  次等概率选择，并得到：

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

类似的，

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

我们可以将  $n$  选择为任意大，并求得一个满足下式的  $m$ ：

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

然后，取对数，并除以  $n \log s$ ，得：

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

其中  $\epsilon$  为任意小。现在由  $A(n)$  的单调性质，得：

$$A(s^m) \leq A(t^n) \leq A(s^{m+1})$$

$$mA(s) \leq nA(t) \leq (m+1)A(s)$$

因此，除以  $nA(s)$ ，得：

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

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| < 2\epsilon \quad A(t) = K \log t$$

其中， $K$ 必须是满足(2)的正数。

现在假定我们要从  $n$  种可能选项中作一选择，其可测量概率为  $p_i = \frac{n_i}{\sum n_i}$ ，其中  $n_i$  为整数。我们可以把从  $\sum n_i$  种可能性进行一次选择，分解为在概率为  $p_1, \dots, p_n$  的  $n$  种可能性中进行一次选择，然后，如果选定了第  $i$  个，则以等概率从  $n_i$  中选择。再次利用条件(3)，使这两种方法计算得出的由  $\sum n_i$  做出的总选择相等：

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

因此，

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

如果  $P_i$  是不可测的，则可以用有理数进行近似，但根据我们的连续性假设，同一表达式必然成立。因此，该表达式通常均成立。系数  $K$  的选择是为了方便，用于针对所选择的度量单位进行调整。

### 附录 3：关于各态历经信源的定理

如果有可能沿一条概率为  $p > 0$  的路径，由任意  $p > 0$  的状态转入任意其他状态，则该系统是各态历经的，可以适用严格的大数定律。因此，该网络中一条给定路径  $p_{ij}$  在一条长度为  $N$  的长序列中的遍历次数正比于它在  $i$  处、然后选择这一路径的概率（比如说  $P_i$ ），即  $P_i p_{ij} N$ 。如果  $N$  足够大，使出现百分比误差  $\pm \delta$  的概率小于  $\epsilon$ ，使得除一个小概率集合之外，其他所有实际数值都位于下面的界限内：

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

因此，几乎所有序列的概率均为  $p$ ，给出如下：

$$p = \prod p_{ij}^{(P_i p_{ij} \pm \delta) N}$$

而  $\frac{\log P}{N}$  的范围为：

$$\frac{\log P}{N} = \sum (P_i p_{ij} \pm \delta) \log p_{ij}$$

或

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

这就证明了定理 3。

根据定理 3 中  $p$  值的可能范围，计算  $n(q)$  的上下限，立刻就可以推出定理 4。

在混合（非各态历经）情景中，如果

$$L = \sum p_i L_i$$

各个分量的熵为  $H_1 \geq H_2 \geq \dots \geq H_n$ ，我们有：

**定理：**  $\lim_{n \rightarrow \infty} \frac{\log n(q)}{N} = \varphi(q)$  是递减阶梯函数，

在区间  $\sum_{i=1}^{s-1} \alpha_i < q < \sum_{i=1}^s \alpha_i$  中有  $\varphi = H_s$

为了证明定理 5 和 6，首先注意到  $F_N$  是单调递减的，因为增大  $N$  值会向条件熵增加一个下标。将  $F_N$  的定义中对  $p_{B_i}(S_j)$  进行简单的代入，可以证明：

$$F_N = NG_N - (N-1)G_N$$

并针对所有  $N$  值进行求和，得出  $G_N = \frac{1}{N} \sum F_n$ 。因此， $G_N \geq F_N$ ， $G_N$  为单调递减。它们也必然趋近于相同极限。利用定理 3，可以看出  $\lim_{n \rightarrow \infty} G_N = H$ 。

## 附录 4：给定约束条件下信息生成速率的最大化

假定我们对符号序列有一组约束条件，这些符号为有限状态类型，因此，可以用线性图进行表示。设  $l_{ij}^{(s)}$  是在由状态  $i$  过渡到状态  $j$  时，可能出现的各个符号的长度。设  $P_i$  是不同状态的频率， $p_{ij}^{(s)}$  是在状态  $i$  下选择符号  $s$ ，并转换为状态  $j$  的概率，在这些约束条件下，如何选择  $P_i$  和  $p_{ij}^{(s)}$  的分布，使生成信息的速率达到最大？这些约束条件限定了一个离散信道，最大生成速率必须小于或等于这个信道的容量  $C$ ，这是因为，如果所有大长度的序列块都是等概率的，就会得到这一结果，如果可能的话，它就是最好的。我们将会证明，通过正确地选择  $P_i$  和  $p_{ij}^{(s)}$ ，就能实现这一速率。

所考虑的速率是：

$$\frac{-\sum P_i p_{ij}^{(s)} \log p_{ij}^{(s)}}{\sum P_i p_{ij}^{(s)} l_{ij}^{(s)}} = \frac{N}{M}$$

设  $l_{ij} = \sum_s l_{ij}^{(s)}$ 。显然，对于一个最大值， $p_{ij}^{(s)} = k \exp l_{ij}^{(s)}$ 。对最大值的约束条件是  $\sum P_i = 1$ ， $\sum_j p_{ij} = 1$ ， $\sum P_i (p_{ij} - \delta_{ij}) = 0$ 。因此，我们使下式达到最大：

$$U = \frac{-\sum P_i p_{ij} \log p_{ij}}{\sum P_i p_{ij} l_{ij}} + \lambda \sum_i P_i + \sum \mu_i p_{ij} + \sum \eta_i P_i (p_{ij} - \delta_{ij})$$

$$\frac{\partial U}{\partial p_{ij}} = -\frac{MP_i(1 + \log p_{ij}) + NP_i l_{ij}}{M^2} + \lambda + \mu_i + \eta_i P_i = 0$$

解出  $p_{ij}$

$$p_{ij} = A_i B_j D^{-l_{ij}}$$

由于：

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

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

$D$ 的正确值为容量 $C$ ， $B_j$ 是下式的解：

$$B_i = \sum_j B_j C^{-l_{ij}}$$

因为：

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

$$\sum_i p_{ij} \frac{B_j}{B_i} C^{-l_{ij}} = P_j$$

或者：

$$\sum_i \frac{P_i}{B_i} C^{-l_{ij}} = \frac{P_j}{B_j}$$

因此，如果 $\lambda_i$ 满足

$$\sum_i \gamma_i C^{-l_{ij}} = \gamma_j$$

$$P_i = B_i \gamma_i$$

由于  $C$  满足：

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

所以关于 $B_i$ 和 $\gamma_i$ 的方程组都成立。

在这种情况下，该速率为：

$$-\frac{\sum_i P_i p_{ij} \log \frac{B_j}{B_i} C^{-l_{ij}}}{\sum_i P_i p_{ij} l_{ij}} = C - \frac{\sum_i P_i p_{ij} \log \frac{B_j}{B_i}}{\sum_i P_i p_{ij} l_{ij}}$$

但是：

$$\sum_i P_i p_{ij} (\log B_j - \log B_i) = \sum_j P_j \log B_j - \sum_j P_j \log B_i = 0$$

因此，该速率为 $C$ ，而且它从来不会被超越，所以它就是最大值，从而表明上面所取解的合理性。

# 第 III 部分：数学预备知识

在本章的最后部分，我们考虑信号和消息之一或同时为连续变量的情景，与在此之前假定的离散类型形成对比。在一定程度上，连续情景可以通过求极限过程由离散情景获得：将消息和信号划分为大量但数目有限的区域，并基于离散情景计算所涉及的各个参数。随着这些区域大小的缩小，这些参数一般将会趋近于连续情景下的正确值。但是，是会出现一些新的影响，而且在由一般结果具体到特定情景时，强调重点也有所不同。

在连续情景中，我们不会尝试获得具有最大一般性的结果，也不会尝试以极度严格的纯数学方式给出结果，因为这样会涉及到大量的抽象测试理论，并模糊主要分析思路。但是，初步研究表明，这一理论可以用完全不证自明的、严格的方式加以表述，其中包括了连续情景、离散情景和许多其他情景。在本分析中的求极限过程中，会偶有违反约定的地方，但在所有具有实践意义的情景中，都有其正当理由。

## 18. 函数的集合与系集

在连续情景中，我们不得不处理函数的集合和函数的系集。由函数集的名字可以看出，它就是一组函数，通常是一个变量——时间的函数。为描述函数集，我们可以给出集合中各种函数的显式表达式，也可以给出只有集合中的函数才拥有的性质。下面是一些示例：

1. 由以下函数组成的集合：

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

$\theta$  的每个具体值确定了集合中的一个特定函数。

2. 一个由时间函数组成的集合，其中包含频率不超过  $W$  周期/秒的所有时间函数。
3. 一个由带宽局限于  $W$ 、幅度不超过  $A$  的所有函数组成的集合。
4. 将所有英文语音信号表示为时间函数组成的集合。

函数的**系集(ensemble)**是一个函数集合再加上一个概率度量，利用这一度量，我们可以确定集合中一个具有特定性质的函数的概率<sup>9</sup>。例如，对于以下集合：

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

我们可以给出  $\theta$  的一个概率分布， $P(\theta)$ 。该集合就变成一个系集。

函数系集的另外一些例子包括：

1. 一个有限函数集  $f_k(t)$  ( $k = 1, 2, \dots, n$ )，其中  $f_k$  的概率为  $p_k$ 。
2. 一个有限维的函数簇：

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

<sup>9</sup> 用数学术语来说，这些函数属于一个总测度为单位 1 的测度空间。

加上参数 $\alpha_i$ 的一个概率分布：

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

例如，我们可以考虑以下式定义的系集

$$f(a_1, \dots, a_n, \theta_1, \dots, \theta_n; t) = \sum_{i=1}^n a_i \sin i(\omega t + \theta_i)$$

其中的幅度 $a_i$ 独立、正态分布，相位 $\theta_i$ 在 0 到 $2\pi$ 之间均匀独立分布。

### 3. 系集

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

其中 $a_i$ 为正态独立分布，都具有相同的标准差 $\sqrt{N}$ 。这是一种“白”噪声的表示，其频带限于 0 到 $W$ 周期/秒，平均功率为 $N$ 。<sup>10</sup>

4. 设分布在  $t$  轴上的各个点服从泊松分布。在每个选定点上放置一个函数 $f(t)$ ，并添加不同的函数，给出系集：

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

其中 $t_k$ 是服从泊松分布的点。这一系集可以看作是一种脉冲或散粒类型的噪声，其中所有脉冲都相同。

5. 英文语音函数的集合，其概率度量由普通应用中的发生频率给出。

一个函数系集 $f_\alpha(t)$ ，如果将所有函数都平移任意固定时间量后，会得到同一系集，则说这个系集是**平稳的**。如果 $\theta$ 在 0 和 $2\pi$ 之间均匀分布，则系集：

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

是平稳的。如果我们将每个函数平移 $t_1$ ，则可以得到：

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

其中 $\varphi$ 在 0 和 $2\pi$ 之间均匀分布。每个函数都发生了变化，但系集作为整体在转换后是不变的。上面给出的其他示例也是平稳的。

如果一个系集是平稳的，而且集合中任何一个概率不是 0 和 1 的函数子集都不是平稳的，则说这个系集是**各态历经的**。系集：

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

是各态历经的。这些函数的任何一个概率不为 0 和 1 的子集，经过任何时间平移后，都不会转换为它自己。另一方面，如果  $a$  为均匀分布， $\theta$  为均匀分布，则系集

<sup>10</sup> 这一表示可以用作带限白噪声的定义。这种定义的好处是，它涉及的极限运算要少于过去的定义方法。尽管“白噪声”这个名字在参考文献中牢固地确立了自己的地位，但它可能还是有些不太恰当。在光学中，白光意味着要么是与点光谱相对的连续谱，要么是指波长上平坦的光谱（与在频率上平坦的频谱不同）。

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

是平稳的，但不是各态历经的。例如， $a$  介于 0 与 1 之间的函数子集是平移的。

在给出的示例中，3 和 4 是各态历经的，5 也可以看作是各态历经的。如果一个系集是各态历经的，我们可以大致地说，这个集合中的每个函数都是该系列的一个代表。更准确地说，我们知道对于一个各态历经系集，系集上任意统计值的平均值（以概率 1）等于对集合内一个具体函数各时间平移版本求得的平均值。<sup>11</sup>大致来说，随着时间的推移，可以预期每个函数都会以适当的频率遍历集合中任意函数的所有状态。

我们对大量函数执行各种操作可以获得新的函数，同样，对系集进行操作也可以获得新的系集。例如，假定有一个函数系集  $f_\alpha(t)$  和运算符  $T$ ，它对每个函数  $f_\alpha(t)$  进行运算后，得到函数  $g_\alpha(t)$ ：

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

利用集合  $f_\alpha(t)$  的概率度量，为集合  $g_\alpha(t)$  定义概率度量。 $g_\alpha(t)$  函数的一个特定子集的概率等于某个  $f_\alpha(t)$  函数子集的概率，这些  $f_\alpha(t)$  函数经过运算  $T$  后，生成给定  $g$  函数子集的成员。从物理上来说，这相当于将该系集通过某一设备，比如一个滤波器、整流器或调制器。该设备的输出函数构成系集  $g_\alpha(t)$ 。

如果一个设备或运算符  $T$  对输入移位后，只是让输出也同样移位，则称之为“时不变的”，也就是说，如果对于所有  $f_\alpha(t)$  和所有  $t_1$ ，可以由

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

可以推出：

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

则说它是时不变的。容易看出（见附录 5），如果  $T$  是时不变的，而且输入系集是平稳的，则输出系集也是平稳的。同样，如果输入是各态历经的，则输出也是各态历经的。

滤波器或整流器在所有时间平稳情况下，都是时不变的。而调制操作则不是，因为载波相位呈现一定的时间结构。但是，如果平移量是该载波周期的倍数，则调制操作也是时不变的。

维纳已经指出了物理设备在时间平稳情况下的不变性与傅里叶理论之间的密切关系。<sup>12</sup>他已经证明：如果一个设备是线性时不变的，则傅里叶分析是处理该问题的合适数学工具。

由连续信源（例如语音）生成的消息、发送器生成的信号，以及干扰噪声，都可以很准确地用函数系集进行数学表示。维纳已经强调过，通信理论应当关心的不是对特定函数的操作，而是对函数系集的操作。通信系统不是为特定的语音函数设计的，更不是为正弦函数设计的，而是为语音函数的系集设计的。

<sup>11</sup> 这是著名的各态历经定理，或者说是这一定理的一个方面，Birkoff, von Neumann 和 Koopman 在稍有不同的公式中进行了证明，后来由 Wiener, Hopf, Hurewicz 及其他人进行了推导。关于各态历经的文献非常广泛，读者可参阅这些作者的论文，获得精确的一般性公式；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。

<sup>12</sup> 通信理论在很大程度上要归功于维纳，大部分基础思想和理论都是由他给出的。在他的经典 NDRC 报告——*The Interpolation, Extrapolation and Smoothing of Stationary Time Series* (Wiley, 1949) 中，首次明确地将通信理论表述为统计问题，研究了对时间序列的处理。尽管这一著作主要讨论的是线性预测和滤波问题，但却是与本文相关的重要参考文献。我们还可以参考维纳的 *Cybernetics* (Wiley, 1948)，其中讨论了通信与控制的一般性问题。

## 19. 带限函数系集

如果一个时间函数 $f(t)$ 的带宽局限于0至 $W$ 个周期/秒，只需要给出该函数在一系列间距为 $\frac{1}{2W}$ 秒的离散点上的分量，就可以利用以下方式确定该函数。<sup>13</sup>

**定理 13:** 设 $f(t)$ 中没有超过 $W$ 的频率，则：

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

其中：

$$X_n = f\left(\frac{n}{2W}\right)$$

在这个展开式中， $f(t)$ 表示为正交函数之和。各项的系数 $X_n$ 可以看作是一个无限维“函数空间”中的坐标。在这个空间中，每个函数准确地对应一个点，每个点对应一个函数。

如果0到时刻 $T$ 之外的所有坐标 $X_n$ 都是0，则可以认为一个函数基本上局限于时刻 $T$ 。在这种情况下，除 $2TW$ 之外的所有坐标都是0。因此，带宽限制于 $W$ 、持续时间为 $T$ 的函数对应于一个 $2TW$ 维空间中的点。

带限为 $W$ 、持续时间为 $T$ 的函数子集对应于这个空间中的一个区域。例如，总能量小于或等于 $E$ 的函数对应于一个 $2TW$ 维球面内的点，其半径为 $r = \sqrt{2WE_r}$ 。

一个持续时间和带宽有限的函数系集可以用相应 $n$ 维空间中的一个概率分布 $p(x_1, \dots, x_n)$ 来表示。如果该系集不受时间限制，是可以用一个给定区间 $T$ 中的 $2TW$ 个坐标来大体表示该区间 $T$ 中的函数部分，用概率分布 $p(x_1, \dots, x_n)$ 给出该系集在该持续时间区间内的统计结构。

## 20. 连续分布的熵

离散概率集 $p_1, \dots, p_n$ 的熵已经定义为：

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

对于一个概率密度函数为 $p(x)$ 的连续分布，可以采用类似方式，将它的熵定义为：

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

对于一个 $n$ 维分布 $p(x_1, \dots, x_n)$ ，有：

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

如果有两个参数 $x$ 和 $y$ （它们本身可能是多维的），则 $p(x, y)$ 的联合熵和条件熵分别为：

<sup>13</sup> 关于这一定理的证明及深入讨论，请参阅作者的论文《有噪条件下的通信》（Communication in the Presence of Noise），发表在 *Proceedings of the Institute of Radio Engineers*, v. 37, No. 1, Jan., 1949, pp. 10–21。



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

和

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

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

其中:

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

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

连续分布的熵具有离散分布的大多数性质（但不是全部性质）。具体来说，有：

1. 如果 $x$ 限定于其空间中的一个特定空间 $v$ ，则 $H(x)$ 为最大值，当 $p(x)$ 在其空间内为常数 $(1/v)$ 时，等于 $\log v$ 。
2. 对于任意两个变量 $x, y$ ，有：

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

当（且仅当） $x$ 和 $y$ 独立时，即 $p(x, y) = p(x)p(y)$ 时（可能有一些概率为0的点除外），等式成立。

3. 考虑下面这种类型的广义求平均运算：

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

其中：

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

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

且

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

5. 设 $p(x)$ 为一维分布。如果要求 $x$ 的标准差固定为 $\sigma$ ，则当 $p(x)$ 为高斯分布时的熵最大。为进行证明，必须以：

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

为约束条件，使：

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

根据变分法，这要求下式取最大值：

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

其条件是：

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

相应地（调整常数，以满足约束条件）

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

类似地，在 $n$ 维中，假定 $p(x_1, \dots, x_n)$ 的二阶矩固定为 $A_{ij}$ ：

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

当 $p(x_1, \dots, x_n)$ 是二阶矩为 $A_{ij}$ 的 $n$ 维高斯分布时，（通过类似计算）得到最大熵。

6. 标准差为 $\sigma$ 的一维高斯分布的熵由下式给出：

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

其计算如下：

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

类似的，具有相关形式 $a_{ij}$ 的 $n$ 维高斯分布由下式给出：

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

它的熵可以计算如下：

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

其中 $|a_{ij}|$ 是元素为 $a_{ij}$ 的行列式。

7. 如果 $x$ 限制于半条线（对于 $x \leq 0$ ,  $p(x) = 0$ ），且 $x$ 的一阶矩固定为 $a$ ：

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

则当

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

时，得到最大熵，且等于 $\log ea$ 。

8. 连续熵和离散熵之间有一个重要差别。在离散情况下，这个熵是以**绝对**方式度量随机变量的随机性。在连续情况下，这个度量是**相对于坐标系统的**。如果改变了坐标系，这个熵通常也会改变。事实上，如果我们将坐标系改为 $y_1 \dots y_n$ ，则新的熵为：

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

其中， $J\left(\frac{x}{y}\right)$ 是雅可比坐标变换。展开对数，并将变量变为 $x_1 \dots x_n$ ，得：

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

因此，新的熵等于原熵减去雅可比行列式的期望值。在连续情况下，可以将熵看作是一种**相对于一种假定标准的**随机性度量，这个“标准”就是选择坐标系，使每个小的体积分量 $dx_1 \dots dx_n$ 具有给定权值。在

改变坐标系时，新坐标系下的熵度量的是在新系统下，体积分量 $dy_1 \dots dy_n$ 具有相等权值时的随机性。

除了与坐标系具有这样依赖关系之外，熵的概念在连续情况下的重要性与离散情景中完全相同。这是因为，由此推导得出的信息率概念和信道容量概念取决于两个熵之差，这个差值与坐标系无关，当坐标系变化时，这两个熵的改变值相同。

连续分布的熵可能是负值。度量刻度设定一个任意零点，对应于单位体积上的一个均匀分布。如果一个分布的范围小于这一单位体积，它的熵会减小，从而为负值。但是，信息率和信道容量总是非负值。

9. 坐标系变化的一种特定情况为线性变换：

$$y_i = \sum_j a_{ij} x_j$$

在这种情况下，雅可比变换行列式 $|a_{ij}|^{-1}$ 且

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

在坐标旋转情况下（或任何保测变换）， $J = 1$ 和 $H(y) = H(x)$ 。

## 21. 函数系集的熵

考虑一个各态历经的带限函数系集，最高频率为 $W$ 个周期/秒。设 $n$ 个连续采样点的幅度 $x_1, \dots, x_n$ 的分布密度函数为：

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

将该系集的熵（每自由度）定义为：

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

我们还定义一个熵 $H$ （每秒），这次除以的不是 $n$ ，而是 $n$ 个采样的持续时间 $T$ （单位为秒）。由于 $n = 2TW$ ，所以有 $H = 2WH'$ 。

当噪声为高斯白色热噪声时，有：

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

$$H = W \log 2\pi e N$$

对于给定平均功率 $N$ ，白噪声的熵最大。这一点是由上面给出的高斯分布的最大化属性得到的。

连续随机过程的熵有很多与离散情景类似的性质。在离散情况下，熵与长序列**概率**的对数相关联，与可能出现的长序列的**数目**相关联（这里所说的长，是在合理范围内）。在连续情况下，可以采用一种类似方式，将熵与一长串采样的**概率密度**的对数关联起来，与函数空间中的高概率**体积**相关联（这里所说的“高概率”，是指在合理范围内）。

更准确地说，如果假定对于所有 $n$ 值， $p(x_1, \dots, x_n)$ 在所有 $x_i$ 处都是连续的，则对于足够大的 $n$ 值，除了一个总概率小于 $\sigma$ 的集合之外，对于任意选择的 $(x_1, \dots, x_n)$ ，都有

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

其中 $\delta$ 和 $\epsilon$ 为任意小。如果将空间划分为大量小的单元格，则可以由各态历经性质推出以上结果。

$H$  与体积的关系可以表述如下：同样假定有一个 $n$ 维空间与 $p(x_1, \dots, x_n)$ 对应。设 $V_n(q)$ 是该空间中总概率为 $q$ 的最小体积。如果 $q$ 不等于0或1，则：

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

这些结果表明，对于大的 $n$ 值，存在一个定义相当明确的高概率体积（至少在对数意义上是这样的），在这一体积内部，概率密度相对均匀一些（同样是指在对数意义上）。

在白噪声情况下，分布函数为：

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

由于该函数仅取决于 $\sum x_i^2$ ，所以等概率密度的表面是一个球面，整个分布是球对称的。高密度区域是半径为 $\sqrt{nN}$ 。当 $n \rightarrow \infty$ 时，位于半径为 $\sqrt{n(N + \epsilon)}$ 的球体外部的概率趋近于0，该球体体积对数的 $\frac{1}{n}$ 倍趋近于 $\log \sqrt{2\pi e N}$ 。

在连续情况下，比较方便的做法不是处理系集的熵 $H$ ，而是处理一个推导出来的量，我们称之为熵功率。其定义为：与原系集的带宽相等且具有相同熵的白噪声的功率。换言之，如果 $H'$ 为一个系集的熵，则熵功率为：

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

在几何图形中，这相当于用一个等体积球体的半径平方来衡量一个高密度体积。对于给定功率，白噪声的熵最大，所以任何噪声的熵功率都小于或等于其实际功率。

## 22. 线性滤波器中的熵损失

**定理 14：**一个系集的带宽为 $W$ ，其熵为 $H_1$ （每自由度），如果使该系集通过一个特性函数为 $Y(f)$ 的滤波器，则输出系集的熵为：

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

此滤波器的运算实际上就是坐标系的线性变换。如果将不同的频率分量看作原坐标系，则新的频率分量实际上就是原频率分量乘以相应的因子。坐标变换矩阵的对角线元素实际上就是这些坐标。该变换的雅可比行列式为（对于 $n$ 个正弦分量和 $n$ 个余弦分量）：

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

其中 $f_i$ 在带宽 $W$ 内等间隔分布。因此，上式变为极限值：

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

由于 $J$ 是常数，所以它的均值是相同量，另外，根据在坐标变换时熵的变化定理，可以得出结果。我们还可以用熵功率来表达。因此，如果第一个系集的熵功率为 $N_1$ ，则第二个系集的熵功率为：

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

最终的熵功率等于原熵功率乘以滤波器的几何平均增益。如果该增益的单位为 $db$ ，则输出熵功率的增大值将等于频宽 $W$ 上的算术平均 $db$ 增益。

在表 1 中，对于几种理想增益特性曲线计算了熵功率损失（也以 $db$ 为单位）。这些滤波器的冲激响应也以 $W = 2\pi$ 给出，假定相位为 0。

其他许多情况下的熵损失都可以由这些结果计算得出。例如，第一种情况中的熵功率因子 $1/e^2$ 也适用于通过 $\omega$ 轴的保测变换，由 $1 - \omega$ 得到的任意其他增益特性。具体来说，线性增益 $G(\omega) = \omega$ 或介于 0 与 1 之间的“锯齿”特性都具有同样的熵损失。增益取倒数时，其因子也取倒数。因此， $1/\omega$ 的因子为 $e^2$ 。将任意功率增大该增益，都会使该功率乘以这一因子。

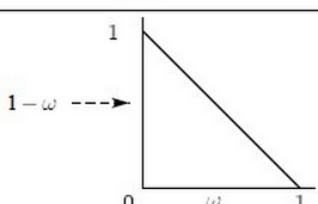
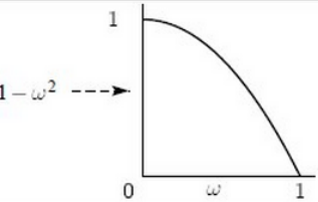
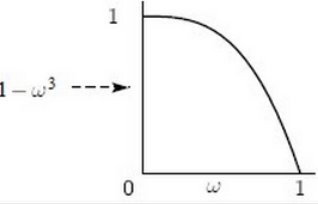
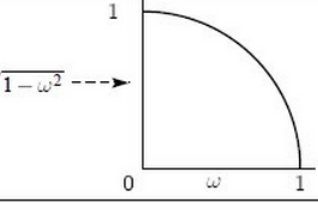
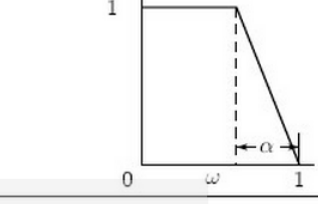
增益	熵功率因子	熵功率增益	冲激响应
	$\frac{1}{e^2}$	-8.69	$\frac{\sin^2(t/2)}{t^2/2}$
	$\left(\frac{2}{e}\right)^4$	-5.33	$2 \left[ \frac{\sin t}{t^3} - \frac{\cos t}{t^2} \right]$
	0.411	-3.87	$6 \left[ \frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3} \right]$
	$\left(\frac{2}{e}\right)^2$	-2.67	$\frac{\pi}{2} \frac{J_1(t)}{t}$
	$\frac{1}{e^{2\alpha}}$	-8.69 $\alpha$	$\frac{1}{\alpha t^2} [\cos(1-\alpha)t - \cos t]$

表 1

## 23. 两个系集之和的熵

如果有两个函数系集 $f_\alpha(t)$ 和 $g_\beta(t)$ ，可以通过“加法”构成一个新系集。假定第一个系集的概率密度函数 $p(x_1, \dots, x_n)$ ，第二个的为 $q(x_1, \dots, x_n)$ 。因此，两者之和的概率密度由以下卷积给出：

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

在物理上，这种情况对应于将由原函数系集表示的噪声或信号求和。

在附录 6 中推导了以下结果。

**定理 15：** 设两个系集的平均功率为 $N_1$ 和 $N_2$ ，并设其熵功率为 $\bar{N}_1$ 和 $\bar{N}_2$ 。则两者之和的熵 $\bar{N}_3$ 的上下限为：

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

高斯白噪声有一种特殊的性质：如果在特定意义下，信号功率相对于噪声很低，则它可以吸收任何向其增加的噪声或信号系集，所得到的熵功率近似等于白噪声功率和信号功率之和（从平均信号值的角度来测量，信号功率通常为 0）。

考虑与这些系集相关联的函数空间，它有  $n$  个维度。白噪声与这个空间中的球形高斯分布相对应。信号系集对应于另一种概率分布，不一定是高斯分布或球形分布。设这一分布关于其重心的二阶矩为 $a_{ij}$ 。因此，如果 $p(x_1, \dots, x_n)$ 为密度分布函数，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，我们可以旋转坐标系，使它对准这一形式的主要方向。 $a_{ij}$ 于是化简为对角形式 $b_{ii}$ 。我们要求每个 $b_{ii}$ 相对于  $N$  都很小（ $N$  为该球形分布半径的平方）。

在这种情况下，噪声与信号的卷积近似生成一个高斯分布，其对应的二次型为：

$$N + b_{ii}$$

这一分布的熵功率为：

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

或近似为：

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

最后一项为信号功率，第一项为噪声功率。

## 第 IV 部分 连续信道

### 24. 连续信道的容量

在连续信道中，输入信号或所传送的信号是某一特定集合中的连续时间函数 $f(t)$ ，输出信号或所接收的信号是前述信号受到干扰后的结果。我们仅考虑传送信号和接收信号的带宽都不超过 $W$ 的情况。因此，对于某一时间 $T$ ，可以用 $2TW$ 个数字来指定这些信号，用有限维分布函数来指明其统计结构。因此，传送信号的统计信息由下式给出：

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

噪声的统计信息由下面的条件概率分布给出：

$$P_{x_1, \dots, x_n}(y_1, \dots, y_n) = P_x(y)。$$

可以采用一种类似于离散信道的方式来定义连续信道的信息传送速率，即：

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

式中， $H(x)$ 是输入的熵， $H_y(x)$ 是疑义度。信道容量 $C$ 定义为：在所有可能选择的系集范围内改变输入， $R$ 取得的最大值。这意味着，采用一种有限维近似，我们必须改变 $P(x) = P(x_1, \dots, x_n)$ ，使下式取最大值：

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

事实上， $\iint P(x, y) \log P(x) dx dy = \int P(x) \log P(x) dx$ ，由此可将上式记为：

$$\iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy$$

因此，信道容量可表示如下：

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

由这一形式容易看出， $R$ 和 $C$ 与坐标系无关，因为，当 $x$ 和 $y$ 以任意一对一方式进行变换时， $\log \frac{P(x, y)}{P(x)P(y)}$ 中的分子和分母将乘以相同因子。 $C$ 的这一积分表达式要比 $H(x) - H_y(x)$ 更具一般性。只要正确解读，这一表述式的结果总是存在的，而在某些情况下， $H(x) - H_y(x)$ 可能会取未定式 $\infty - \infty$ 。例如，在其 $n$ 维近似中，如果 $x$ 局限于一个维度小于 $n$ 的表面，就会出现上述情况。

如果在计算 $H(x)$ 和 $H_y(x)$ 时的对数底数为2，则和离散情况中一样， $C$ 表示每秒钟以任意小的疑义度通过该信道传送的最大二进制数位个数。从物理上来说，将信道空间划分为足够多的小单元格，使得在一个单元格内，信号 $x$ 受干扰后变到点 $y$ 的概率密度 $P_x(y)$ 基本为恒定。如果将这些单元格看作不同点，则此情况下离散信道中基本相同，当时的证明过程在这里同样适用。但从物理上容易看出，在任意实际情况下，只要这些区域足够小，这种将体积量化为独立点的过程不会显著改变最终的答案。因此，这一容量就是离散细分区域的容量的极限，也就是上述定义的连续容量。

在数学上，可以首先证明（见附录 7）：如果  $u$  为消息， $x$  是信号， $y$  是接收信号（受到了噪声的干扰）， $v$  是恢复得出的消息，则有：无论对  $u$  执行什么操作以得到  $x$ ，或者对  $y$  执行什么操作以得到  $v$ ，都可以得到

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

因此，无论如何对二进制数位进行编码，以得到信号，也无论如何对接收信号进行解码以恢复出消息，二进制数位的离散速率都不会超出前面定义的信道容量。另一方面，在非常一般的情况下，有可能找出一种编码系统，以所期望的小疑义度或错误频率，以速率  $C$  传送二进制数位。例如，如果为信道函数取一个有限维近似空间，除了一些概率为零的点集之外， $P(x, y)$  在  $x$  和  $y$  上都是连续的，则上述结果成立。

在向信号添加一个与信号独立的噪声时（在概率意义上独立），会出现一种重要的特例。 $P_x(y)$  只是差值  $n = (y - x)$  的函数，

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

我们可以为噪声指定一个确定的熵（与信号的统计信息相独立），即分布  $Q(n)$  的熵。这个熵用  $H(n)$  表示。

**定理 16：**如果信号与噪声独立，且接收信号是传送信号与噪声的和，则传送速率为：

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

即接收信号的熵减去噪声的熵。信道容量为：

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

由于  $y = x + n$ ，所以有：

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

展开左侧，并利用  $x$  与  $n$  相独立这一事实，有：

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

因此，

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

由于  $H(n)$  与  $P(x)$  独立，要使取最大值，则需要使  $H(y)$ （接收信号的熵）取最大值。如果对传送信号的系集具有特定的约束条件，则必然在遵守这些约束条件的情况下，使接收信号的熵最大化。

## 25. 平均功率受限时的信道容量

定理 16 的一种简单应用情景是：噪声为白色热噪声，传送信号的平均功率不超过  $P$ 。则接收信号的平均功率为  $P + N$ ，其中  $N$  是平均噪声功率。当接收信号也构成一个白噪声系集时，接收信号的熵最大，因为这是功率为  $P + N$  时的最大可能熵，通过适当选择传送信号，也就是在使其构成一个功率为  $P$  的白噪声系集时，可以获得这一最大值。接收系集的熵（每秒）为：



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

噪声熵为：

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

信道容量为：

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

综述如下：

**定理 17：**当平均传送功率不超过 $P$ ，带宽为 $W$ 的信道受功率为 $N$ 的白色热噪声干扰时，其信道容量为：

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

这意味着，通过足够复杂的编码系统，能够以 $W \log_2 \frac{P+N}{N}$ 比特/秒的速率传送二进制数位，使错误频率为任意小。任何编码系统在以更高速率传送信息时，都会使错误频率高于某个确定的正值。

为了接近这一极限传送速率，传送信号的统计性质必须接近于白噪声。<sup>14</sup>下面介绍一种可以趋近于理想速率的系统：假设有 $M = 2^s$ 个白噪声样本，每个持续时间为 $T$ 。为它们分配 $0$ 至 $M-1$ 的二进制数字。在传送机将这些消息序列分为 $s$ 个组，对于每一组为说，相应的噪声样本被作为信号传送。在接收端，已知 $M$ 个样本，并将实际接收信号（受噪声干扰）与所有这些样本对比。与接收信号的 RMS 最小的样本被选作传送信号，并重建相应的二进制数字。这一过程用于选择可能性最大的信号（后验概率）。所使用的噪声样本数 $M$ 将取决于可容忍的错误频率 $\epsilon$ ，但几乎所有不同的样本选择，都有：

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

使得无论选择多么小的 $\epsilon$ ，总是可以选择足够大的 $T$ ，在时间 $T$ 内传送的二进制任意接近 $T W \log \frac{P+N}{N}$ 。

对于白噪声情况，其他几位作者已经独立地推导出类似于 $C = W \log \frac{P+N}{N}$ 的公式，只是其解释有所不同。我们会提到 N. Wiener<sup>15</sup>、W. G. Tuller<sup>16</sup>和 H. Sullivan 在这方面的工作。

当干扰噪声为任意噪声时（不一定为白色热噪声），似乎无法显式解决信道容量 $C$ 的最大化问题。但是，可以用平均噪声功率 $N$ 和噪声熵功率 $N_1$ 为 $C$ 设定上、下限。在大多数实际情况下，这些限制足够接近，从而给出该问题的一个令人满意的答案。

**定理 18：**一个信道的带宽为 $W$ ，受任意噪声的干扰，其容量满足以下不等式：

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

---

<sup>14</sup> 在前述引文中提到的《有噪条件下的通信》一文中，从几何的观点讨论了白噪声的这一性质及其他性质。

<sup>15</sup> 前述引文中提到的 *Cybernetics*

<sup>16</sup> “Theoretical Limitations on the Rate of Transmission of Information,” *Proceedings of the Institute of Radio Engineers*, v. 37, No. 5, May, 1949, pp. 468–78.

其中：

$P$ =平均传送功率

$N$ =平均噪声功率

$N_1$ =噪声的熵功率。

受干扰信号的平均功率同样为  $P + N$ 。当接收信号为白噪声时，得到这一功率的最大熵，即  $W \log 2\pi e(P + N)$ 。这是不可能实现的；也就是说，对于传送信号的任意系集，与干扰噪声叠加在一起后，都不会在接收端生成一个白色热噪声，但至少它为  $H(y)$  设定了上限。因此可得：

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

这就是定理中给出的上限。如果传送信号是功率为  $P$  的白噪声，考虑其传送速率就可以得到下限。在这种情况下，接收信号的熵功率至少和功率为  $P + N_1$  的白噪声的熵功率一样大，这是因为，我们在前面的定理中已经证明，两个系集之和的熵功率大于或等于各个熵功率之和。因此：

$$\max H(y) \geq W \log 2\pi e(P + N_1)$$

及

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

当  $P$  增大时，上下限相互接近，所以可以得到一个近似速率：

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

如果噪声本身是白色的， $N = N_1$ ，则该结果简化为前面证明的公式：

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

如果噪声是高斯噪声，但其频谱不一定是平坦的，则  $N_1$  是该噪声功率在频带  $W$  中各频率上的几何均值。因此，

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

其中， $N(f)$  是在频率  $f$  处的噪声功率。

**定理 19：** 如果对于给定传送功率  $P$  的信道，设定其容量为：

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

则  $\eta$  随着  $P$  的增大而单调递减，并趋近于极限 0。

假定对于一个给定功率 $P_1$ ，信道容量为：

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

这意味着在将最佳信号分布（比如说 $p(x)$ ）加到噪声分布 $q(x)$ 上时，会得出熵功率为 $P_1 + N - \eta_1$ 的接收信号分布 $r(y)$ 。假定我们向信号加上一个功率为 $\Delta P$ 的白噪声，从而将该功率增加到 $P_1 + \Delta P$ 。针对和的最小熵功率应用该定理，可以得出接收信号的熵至少为：

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

因此，由于我们可以得到上述 $H$ ，所以最大化分布的熵必然至少一样大，而且 $\eta$ 必然为单调递减的。要证明当 $P \rightarrow \infty$ 时 $\eta \rightarrow 0$ ，可考虑一个信号，它是功率 $P$ 很大的白噪声。如果 $P$ 足够大，无论干扰噪声如何，接收到信号都近似为一个白噪声，也就是说他的熵功率接近于 $P + N$ 。

## 26. 峰值功率受限时的信道容量

在某些应用中，传送器受到的限制不是平均功率而是瞬时峰值功率。信道容量的计算问题于是就变成了：系集中的所有函数 $f(t)$ 对于所有 $t$ 都小于或等于 $\sqrt{S}$ ，在此条件下，如何（通过改变所传送符号的系集）使下式取最大值：

$$H(y) - H(n)$$

这种约束条件无法像平均功率限制那样很好地以数学方式给出。对于这种情景，我们最多获得一个对于所有 $\frac{S}{N}$ 有效的下限，一个“渐近”上限（对于大的 $\frac{S}{N}$ 有效）和当 $\frac{S}{N}$ 较小时的渐近 $C$ 值。

**定理 20：**一个信道的带宽为 $W$ ，受功率为 $N$ 的白色热噪声干扰，则其信道容量的下限为：

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

其中， $S$  是允许传送的峰值功率。对于足够大的 $\frac{S}{N}$ ，有：

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

其中 $\epsilon$ 为任意小。当 $\frac{S}{N} \rightarrow 0$ 时（假定频带 $W$ 的起点频率为0）

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

我们希望使接收信号的熵取最大值。如果 $\frac{S}{N}$ 很大，则近似在使传送系集的熵最大时，接收信号的熵取最大值。

放松对系集的条件，可以获得渐近的上限。假定该功率不是在每个时刻受限于 $S$ ，而只是在样本点处受限。在这一弱化条件下，所传送系集的最大熵当然大于或等于在原条件下的最大熵。这个经过修改的问题很容易求解。如果不同的样本点相互独立的，而且有一个在从 $-\sqrt{S}$ 到 $+\sqrt{S}$ 范围内为常数的分布，即可得到最大熵。这个熵值可计算如下：

$$W \log 4S$$

于是接收信号的熵将小于：

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

其中，当  $\frac{S}{N} \rightarrow 0$  时， $\epsilon \rightarrow 0$ ，减去白噪声的熵（ $W \log 2\pi eN$ ），即可得到信道容量：

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

这就是我们想要的信道容量上限。

为获得下限，考虑同一函数系集。让这些函数传送一个具有三角传输特性的理想滤波器。其增益在频率 0 处为单位 1，然后线性下降，在频率  $W$  处的增益为 0。我们首先证明滤波器的输出函数在所有时间内（而不是取样点）的峰值功率上限为  $S$ 。首先注意到脉冲  $\frac{\sin 2\pi Wt}{2\pi Wt}$  进入滤波器后，在输出端得到：

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

此函数永远不会为负值。一般情况下的输入函数可以看作是一系列如下函数之和：

$$a \frac{\sin 2\pi Wt}{2\pi Wt}$$

其中  $a$  为样本的函数，不大于  $\sqrt{S}$ 。因此，输出也是上述非负形式的移位函数之和，且系数相同。这些函数是非负的，所以当所有系数  $a$  取其最大正值（即  $\sqrt{S}$ ）时，获得对于任意时间  $t$  的最大正值。在这种情况下，输入函数为一个幅度为  $\sqrt{S}$  的常数，由于对于直流而言，该滤波器的增益为单位 1，所以输出也相同。因此，输出系集的峰值功率为  $S$ 。

利用适用于这一情景的定理，由输入系集的熵可以计算出输出系集的熵。输出熵等于输入熵加上滤波器的几何平均值：

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

因此，输出熵为：

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

信道容量大于：

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

我们现在希望证明：对于小的  $\frac{S}{N}$ （峰值信号功率除以平均白噪声功率），则信道容量近似为：

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

更准确地说，当  $\frac{S}{N} \rightarrow 0$  时， $C/W \log\left(1 + \frac{S}{N}\right) \rightarrow 1$ 。由于平均信号功率  $P$  小于或等于峰值功率  $S$ ，则可以得出，对于所有  $S/N$ ：

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

因此，如果可以找到一个函数系集，近似对应于  $W \log\left(1 + \frac{S}{N}\right)$ ，且受限于带宽  $W$  和峰值功率  $S$ ，则可以证得结果。考虑以下类型的函数系集。 $t$  个采样的值相同，或者为  $+\sqrt{S}$  或  $-\sqrt{S}$ ，接下来的  $t$  个采样也具有相同值，以此类推。一个系列的取值为随机选定，以概率  $\frac{1}{2}$  取  $+\sqrt{S}$ ，以概率  $\frac{1}{2}$  取  $-\sqrt{S}$ 。如果将这一系集通过具有三角增益特性的滤波器（对于直流为单位增益），则输入的峰值功率限于  $\pm\sqrt{S}$ 。此外，平均功率近似为  $S$ ，取足够大的  $t$  值，即可趋近于该值。应用关于噪声与小信号之和的定理，可以求出这一系集与热噪声之和的熵。如果

$$\sqrt{t} \frac{S}{N}$$

足够小，这一定理将适用。（在选定  $t$  之后）取  $\frac{S}{N}$  足够小，可以确保满足这一条件。熵功率将任意接近于  $S + N$ ，因此，传送速率将任意接近于

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

## 第 V 部分：连续信源的信息产生速率

### 27. 保真度评价函数

对于一个离散信源，我们能够算出一个确定的信息生成速率，也就是其随机过程的熵。对于连续信源，情况就要复杂得多。首先，连续变化量可以取无数个值，需要无数个二进制数位才能准确表示。这意味着要将一个连续信源的输出传送到接收端，而且能够**完全恢复**，一般情况下需要有一个具有无穷大容量（单位为比特/秒）的信道。由于在一般情况下，信道中总会有一定数量的噪声，而且容量是有限的，所以不可能实现精确传送。

但这种说法模糊了真正的问题。实际上，对于连续信源，我们感兴趣的并不是精确传送，而是在特定容差范围内的传送。问题是，当我们仅需要特定的恢复保真度（采用适当的形式度量）时，能否为连续信源指定一个确定的生成速率。当然，随着保真度要求的增大，该速率会下降。可以证明，在非常一般的情况下，我们能够定义这样一个具有如下性质的速率：通过正确的信息编码，有可能通过一个容量等于该速率的信道来传送该信息，同时满足保真度要求。当信道容量更小时，则不足以正确传送该信息。

首先，对于传输保真度的概念，需要给出一种一般性的数学描述。考虑一组持续时间很长的消息，比如为  $T$  秒。一个信源可以描述如下：给出该信源在相关空间内选择特定消息的概率密度  $P(x)$ 。一个给定通信系统可以描述如下（从外部视角来看）：给出信源生成消息  $x$  而接收端恢复出消息  $y$  的条件概率  $P_x(y)$ 。该系统作为一个整体（包括信源和传送系统），可以用发出消息  $x$  而最终输出为  $y$  的概率函数  $P(x, y)$  来描述。如果已知此函数，那么从保真度的角度来看，该系统的完整特性就是已知的。对保真度的任意评价在数学上必须对应于对  $P(x, y)$  应用的一种运算。这种运算至少要为这些系统建立一种简单的顺序关系；即，对于两个分别用  $P_1(x, y)$  和  $P_2(x, y)$  表示的系统，必须能够根据我们的保真度准则做出以下判定之一：(1) 第一个系统的保真度较高，(2) 第二个系统的保真度较高，或者 (3) 它们的保真度相同。这意味着保真度准则可以用一个数值函数表示：

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

其参数的变化范围是所有可能存在的概率函数  $P(x, y)$ 。

我们现在将证明，在非常一般且合理的假定条件下，函数  $v(P(x, y))$  可以写为一种非常专业的形式，即在  $x$  和  $y$  的取值范围内，对函数  $\rho(x, y)$  求均值：

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

为得到这一结果，只需要假定：(1) 信源和系统是各态历经的，这样，一个非常长的样本成为该系集典型代表的概率接近于 1，(2) 该评价是“合理”的，也就是说，通过观察典型输入  $x_1$  和  $y_1$ ，有可能基于这些样本给出一个尝试性评价；如果这些样本的持续时间增大，则这一尝试性评价将以概率 1 趋近于根据  $P(x, y)$  的全部知识所获得的精确评价。设该尝试性评价为  $\rho(x, y)$ 。在与该系统相对应的高概率区域中的几乎所有  $(x, y)$ ，当  $T \rightarrow \infty$  时，函数  $\rho(x, y)$  趋近于一个常数：

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

另外，由于：

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

所以可写出：

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

这就给出了所要的结果。

函数 $\rho(x, y)$ 具有 $x$ 与 $y$ 之间“距离”的一般特性。<sup>17</sup>它衡量了在发送 $x$ ，收到 $y$ 时的意外程度（根据我们的保真度准则）。上面给出的一般性结果可以重新表述如下：只要使消息的持续时间 $T$ 足够大，任何合理的评价都可以表示为一个距离函数的均值，这里所说的均值，是指根据发送消息 $x$ 、接收消息 $y$ 的概率 $P(x, y)$ ，在 $x$ 和 $y$ 的集合上进行加权平均。

下面是几个简单的评价函数例子：

1. RMS 准则。

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

在这种很常用的保真度量中，距离函数 $\rho(x, y)$ 就是相关联的函数空间中点 $x$ 和点 $y$ 之间普通欧氏距离的平方（只相当一个常数因子）。

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

2. 频率加权 RMS 准则。更一般的情况下，在应用保真度的 RMS 度量之前，可以对不同频率分量应用不同权值。这相当于将 $x(t) - y(t)$ 通过一个赋形滤波器，以决定输出的平均功率。因此，设：

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

及

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

则

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

3. 绝对误差准则。

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

4. 耳朵和大脑的结构隐含了确定了一种评价或相当大量的评价，适用于语音或音乐传输的情况。例如，存在一种“可懂度”准则，在这一准则中， $\rho(x, y)$ 等于错误解读单词的频率，也就是发送消息 $x(t)$ ，被接收为 $y(t)$ 。尽管在这些情况下无法给出 $\rho(x, y)$ 的显式表示，则在原理上，可以通过足够多的试验确定这些情况。它的一些性质是从一些众所周知的听力试验结果中获得的，例如，耳朵对相位相对不够敏感，对振幅和频率的敏感度近似为对数特性。

---

<sup>17</sup> 不过，它并不是严格意义上的“度量”，因为它既不满足 $\rho(x, y) = \rho(y, x)$ ，也不满足 $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$ 。

5. 离散情景可以看作是一种特殊化，在此情况下，我们根据错误频率来假设评价准则。因此，函数 $\rho(x, y)$ 定义为：序列 $y$ 中的符号不同于 $x$ 中相应符号的数目，除以 $x$ 中的总符号数。

## 28. 信源在给定保真度评价函数下的信息生成速率

我们现在要确定一个连续信源生成信息的速率了。已知信源的 $P(x)$ 和由距离函数 $\rho(x, y)$ 确定的评价 $v$ ，假定该距离函数在 $x$ 和 $y$ 上都是连续的。对于一个具体的系统 $P(x, y)$ ，其质量可以衡量如下：

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

此外，与 $P(x, y)$ 对应的二进制数位流的速率为：

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

对于给定的消息复现质量 $v_1$ ，信息生成速率 $R_1$ 的定义为：使 $v$ 固定为 $v_1$ ，改变 $P_x(y)$ 时得到的最小 $R$ 值。即：

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

其服从以下约束条件：

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

这意味着，我们实际上考虑了所有可以使用、以所需保真度进行传送的通信系统。对于每一个系统计算传送速率（单位为比特/秒），并选择速率最低的那个系统。这个最低速率就是在所讨论保真度下为信源指定的速率。

由以下结果可以看出这一定义的合理性：

**定理 21：**如果对于一个评价 $v_1$ ，一个信源的信息生成速率为 $R_1$ ，一个信道的容量为 $C$ ，只要 $R_1 \leq C$ ，就有可能对信源的输出进行编码，通过该信道，以任意接近 $v_1$ 的保真度进行传送。如果 $R_1 \geq C$ ，则不可能实现。

通过 $R_1$ 的定义和前面给出的结果，马上可以得出定理的第二部分。如果该表述错误，那就可以通过一个容量为 $C$ 的信道以超过 $C$ 比特/秒的速率传送了。定理的第一部分可以通过一种用于证明定理 11 的类似方法来证明。首先，我们将 $(x, y)$ 空间分为大量的小单元格，并像离散情景中一样进行表示。当单元格非常小时，由于假定 $\rho(x, y)$ 是连续的，所以这一过程使评价函数的变化不会超过一个任意小的量。假定 $P_1(x, y)$ 是一个具体系统，它使该速率取最小值 $R_1$ 。我们从高概率的 $y$ 中随机选择一个集合，其中包含

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

个成员，式中，当 $T \rightarrow \infty$ 时， $\epsilon \rightarrow 0$ 。当 $T$ 很大时，每个选定点都会通过一条出现概率很高的直线（如图 10 中所示）连接到 $x$ 的一个集合。采用定理 11 证明过程中的类似计算过程，可以证明：当 $T$ 很大时，对于 $y$ 的几乎所有可能选择，几乎所有 $x$ 都会被一个以所选 $y$ 点为起点的扇形所覆盖。通信系统的操作如下：为选定点指定二进制数。如果发送一条消息 $x$ ，当 $T \rightarrow \infty$ 时， $x$ 位于至少一个扇形中的概率趋近于 1。以适当的编码



方式，通过该信道传送对应的二进制数，实现很低的错误概率（如果存在几个对应的二进制数，则传送其中任意选定的一个）。由于  $R_1 \leq C$ ，所以这是可能实现的。在接收端，重构出相应的  $y$ ，并当作恢复出来的消息。

只要使  $T$  足够大，就可以使这一系统的评价  $\nu'_1$  任意接近  $\nu_1$ 。其原因是：对于每个长的消息样本  $x(t)$  和恢复出来的消息  $y(t)$ ，该评价（以概率 1）接近于  $\nu_1$ 。

值得注意的是，在这个系统中，接收消息中的噪声实际上是传送器中的一般性量化过程产生的，而不是由信道中的噪声产生的。它多少类似于 PCM 中的量化噪声。

## 29. 信息产生速率的计算

信息产生速率的定义在许多方面都类似于信道容量的定义。在信息产生速率的定义中，

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

其中  $P(x)$  和  $\nu_1 = \iint P(x, y) \rho(x, y) dx dy$  为固定值。而在信道容量的定义中，

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

其中  $P_x(y)$  固定，可能还有一个或多个  $K = \iint P(x, y) \lambda(x, y) dx dy$  形式的其他约束条件（例如，平均功率限制）。

对于确定信源信息产生速率的一般性最大化问题，可以给出部分解。利用拉格朗日方法，考虑：

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

当对  $P(x, y)$  取一次变分时，由变分方程可以得出：

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

其中，通过确定  $\lambda$  给出所需要的保真度，通过选择  $B(x)$  以满足：

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

这表明，在采用最佳编码方式时，对于接收到的各个  $y$ ，其发送消息的条件概率  $P_y(x)$  将随着  $x$  与  $y$  之间的距离函数  $\rho(x, y)$  指数下降。

在某些特殊情况下，距离函数  $\rho(x, y)$  仅取决于  $x$  与  $y$  之间的（向量）差，

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

在此情况下，有：

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

因此,  $B(x)$  为常数, 比如  $\alpha$ , 且

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

遗憾的是, 这些正式解在特定情况下很难计算, 似乎没有多少价值。事实上, 仅在有限几种非常简单的情况下, 才能实际计算出信息产生速率。

如果距离函数  $\rho(x, y)$  是  $x$  与  $y$  之间的均方差, 且消息系集为白噪声, 则可以确定该速率。在这种情况下, 有:

$$R = \min[H(x) - H_y(x)] = H(x) - \max H_y(x)$$

其中,  $N = \overline{(x-y)^2}$ 。但当  $y-x$  为白噪声时, 得到  $\max H_y(x)$ , 它等于  $W_1 \log 2\pi eN$ , 其中  $W_1$  是消息系集的带宽。因此,

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

其中  $Q$  是平均消息功率。它证明了以下定理:

**定理 22:** 在采用保真度的 RMS 度量时, 一个功率为  $Q$ 、带宽为  $W_1$  的白噪声信源, 其信息产生速率为:

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

其中,  $N$  是允许原消息与接收消息之间出现的均方误差。

更一般地, 对于任意信源, 我们可以针对均方误差准则, 给出该速率的上下限不等式。

**定理 23:** 对于带宽为  $W_1$  的任意信源, 其信息产生速率的上下限为:

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

其中  $Q$  是信源的平均功率,  $Q_1$  是熵功率,  $N$  是允许的均方误差。

在给定  $\overline{(x-y)^2} = N$  时, 在白噪声情况下得到  $H_y(x)$  的最大值, 由此推出不等式中的下限。如果我们不是采用最佳方式来放置在证明定理 21 时使用的各个点, 而是将它们随机放在半径为  $\sqrt{Q-N}$  的球体中, 则可以得出上限。

## 致谢

本文作者衷心地感谢实验室的各位同事, 特别是 H. W. Bode 博士、J. R. Pierce 博士、B. McMillan 博士和 B. M. Oliver 博士, 他们在此论文的撰写期间提出了许多富有裨益的建议和批评。还要感谢维纳教授, 他对平稳系集的滤波与预测问题的精彩解法深刻地影响了本文作者在这一领域的思想。

## 附录 5

设 $S_1$ 是 $g$ 系集的任意可测子集， $S_2$ 是 $f$ 系集的一个子集，对其执行运算 $T$ 之后，可以得出 $S_1$ 。于是：

$$S_1 = TS_2$$

设 $H^\lambda$ 是一个运算符，对将一个集合中的所有函数都平移时间 $\lambda$ 。由于 $T$ 是时不变的，从而可以与 $H^\lambda$ 交换，于是，

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

因此，如果 $m[S]$ 是集合 $S$ 的概率度量，则：

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

其中第二个等式是根据 $g$ 空间中度量的定义得出，第三个是因为 $f$ 系集是平稳的，最后一个是根据 $g$ 度量的定义得出。

为了证明在时不变运算下，各态历经性质得以保持，设 $S_1$ 是 $g$ 系集的一个子集，对其进行 $H^\lambda$ 运算时是时不变的，设 $S_2$ 是由所有将会变换为 $S_1$ 的函数 $f$ 组成的集合。于是，

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

使得对于所有 $\lambda$ ， $H^\lambda S_2$ 都包含在 $S_2$ 中。现在，由于：

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

这意味着，对于所有 $\lambda$ 且 $m[S_2] \neq 0, 1$ ：

$$H^\lambda S_2 = S_2$$

这一矛盾表明， $S_1$ 不存在。

## 附录 6

上限 $\bar{N}_3 \leq N_1 + N_2$ 可由以下事实推导得出：当功率为 $N_1 + N_2$ 时，白噪声的熵最大。在这种情况下，熵功率为 $N_1 + N_2$ 。

为了得到下限，假定有两个 $n$ 维分布 $p(x_i)$ 和 $q(x_i)$ ，其熵功率分别为 $\bar{N}_1$ 和 $\bar{N}_2$ 。 $p$ 和 $q$ 的卷积为：

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

为了使该卷积的熵功率 $\bar{N}_3$ 最小， $p$ 和 $q$ 应当取什么形式？

$r$  的熵  $H_3$  由下式给出:

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

我们希望在以下约束条件下, 使上式的取值最小:

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

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

于是, 我们考虑:

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

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

若  $p(x)$  在一个特定的参数  $x_i = s_i$  处变化, 则  $r(x)$  中的变化为:

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

且:

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

$q$  变化时的情况与此类似。因此, 取最小值的条件为:

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

$$\int p(x_i - s_i) \log r(x_i) dx_i = -\mu \log q(s_i)$$

如果将第一个等式乘以  $p(s_i)$ , 第二个等式乘以  $q(s_i)$ , 并对  $s_i$  积分, 则得到:

$$H_3 = -\lambda H_1$$

$$H_3 = -\mu H_2$$

或者, 对  $\lambda$  和  $\mu$  求解, 并代入方程,

$$H_1 \int q(x_i - s_i) \log r(x_i) dx_i = -H_3 \log p(s_i)$$

$$H_2 \int p(x_i - s_i) \log r(x_i) dx_i = -H_3 \log q(s_i)$$

现在假定  $p(x_i)$  和  $q(x_i)$  是正态的

$$p(x_i) = \frac{|A_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \sum A_{ij} x_i x_j$$

$$q(x_i) = \frac{|B_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \sum B_{ij} x_i x_j$$

则 $r(x_i)$ 将具有正态二次型 $C_{ij}$ 。如果这些形式的逆为 $a_{ij}, b_{ij}, c_{ij}$ ，则：

$$c_{ij} = a_{ij} + b_{ij}$$

我们希望证明，当且仅当 $a_{ij} = K b_{ij}$ 时，这些函数满足最小化条件，因此，在这些约束条件下给出最小值 $H_3$ 。首先，有：

$$\log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \sum C_{ij} x_i x_j$$

$$\int q(x_i - s_i) \log r(x_i) dx_i = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \sum C_{ij} s_i s_j - \frac{1}{2} \sum C_{ij} b_{ij}$$

它应当等于：

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

这要求 $A_{ij} = \frac{H_1}{H_3} C_{ij}$ 。在这种情况下， $A_{ij} = \frac{H_1}{H_2} B_{ij}$ ，两个等式简化为恒等式。

## 附录 7

下面将给出一种更一般、更严格的方法，引出通信理论的核心定义。考虑一个概率度量空间，它的元素是有序对 $(x, y)$ 。变量  $x, y$  被看作是可能被传送和接收的信号，它们具有较长的持续时间 $T$ 。一些点的 $x$ 属于一个由点 $x$ 组成的子集 $S_1$ ，我们将所有这些点组成的集合称为 $S_1$ 上的条带(strip)，类似的，如果一些点的 $y$ 都属于子集 $S_2$ ，则将所有此类点组成的集合称为 $S_2$ 上的条带。我们将 $x$ 和 $y$ 划分为非重叠可测子集 $X_i$ 和 $Y_i$ 的一个集合，它们以下式近似表示传送速率 $R$ ：

$$R_1 = \frac{1}{T} \sum_i P(X_i, Y_i) \log \frac{P(X_i, Y_i)}{P(X_i)P(Y_i)}$$

其中

$P(X_i)$ 是 $X_i$ 上条带的概率度量

$P(Y_i)$ 是 $Y_i$ 上条带的概率度量

$P(X_i, Y_i)$ 是这些条带交集的概率度量

无论再如何细分，都不会使 $R_1$ 减小。将 $X_1$ 划分为 $X_1 = X_1' + X_1''$ ，并令：

$$\begin{array}{ll} P(Y_1) = a & P(X_1) = b + c \\ P(X_1') = b & P(X_1', Y_1) = d \\ P(X_1'') = c & P(X_1'', Y_1) = e \end{array}$$

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

因此，对于 $X_1$ 和 $Y_1$ 的交集，我们已经用

$$d \log \frac{d}{ab} + e \log \frac{e}{ac} \text{ 代替了 } (d + e) \log \frac{d+e}{a(b+c)}$$

容易证明，考虑到对 $b, c, d, e$ 设定的限制，

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

结果，该和值增大。因此，各种可能出现的细分构成了一个有向集合，随着细分的深入， $R$ 单调递增。我们可以明确地将 $R$ 定义为 $R_1$ 的最小上限，并将其记为：

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

根据上述意义，这个积分包含了连续和离散两种情况，当然还有许多其他不能用这两种形式来表示的情景。如果 $x$ 和 $u$ 具有一一对应关系，从 $u$ 到 $y$ 的速率等于从 $x$ 到 $y$ 的速率，则它在这一公式中是平凡的。如果 $v$ 是 $y$ 的任意函数（该函数不一定可逆），则从 $x$ 到 $y$ 的速率大于从 $x$ 到 $v$ 的速率，这是因为在计算近似时， $y$ 的细分实质上是对 $v$ 更精细再分。更一般地说，如果 $y$ 和 $v$ 没有函数关系，而是统计相关，也就是说有一个概率测度空间 $(y, \nu)$ ，则 $R(x, \nu) \leq R(x, y)$ 。这意味着对接收信号应用的任何操作，即使涉及到统计分量，也不会增大 $R$ 。

还有另外一个概念，应当用一个抽象的理论公式来准确定义，那就是“维度速率”，也就是为了指定系集中的一个成员，平均每秒需要多少维度。在带限情况下，每秒种有 $2W$ 个数就足够了。一般定义可构造如下。设 $f_\alpha(t)$ 是一个函数系集，并设 $\rho_T[f_\alpha(t), f_\beta(t)]$ 是一个度量，用来测量在时间 $T$ 上从 $f_\alpha$ 到 $f_\beta$ 的“距离”（例如，在 $T$ 时间内的RMS差）。通过选择一定数量的元素 $f$ 可以使系集中的所有元素（除去一个测度为 $\delta$ 的集合之外）至少与一个所选元素 $f$ 的距离不大于 $\epsilon$ 。设 $N(\epsilon, \delta, T)$ 是满足上述条件的最少 $f$ 数。于是，除了一个小测度 $\delta$ 的集合之外，我们将整个空间涵盖在 $\epsilon$ 范围内。我们用下面的三重极限来定义该系集的维度速率 $\lambda$ ：

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

这是拓扑学中维度测度类型定义的推广。在一些简单的系集中，所期望的结果是明显成立的，而上式与这些简单系集的直观维度速率也是一致的。