第十三章　量子信息论

§13.1，经典 *Shannon* 理论简介

1，*Shannon* 熵和数据压缩

i, *C.Shannon* 于 1948 年建立了两个定理。一条是“**无噪声编码定理**”。这涉及一条信息能被压缩多少？就是说，这条信息有多少冗余？另一条是“**噪声通道编码定理**”。涉及在一个噪声通道上，能以多大的速率来可靠地进行传递？就是说，需要添加多少冗余度到信息中去以防止误差？这两个问题都涉及“冗余度”——平均而言信息的相邻字符有多少是不需要的。*Shannon* 的关键观点之一是：**熵提供了量化冗余度的合适方法。**

一条信息是从共计 *K* 个字符中（可以重复）选出来的 *n* 个字符的一种序列。这里通常是 10 个阿拉伯数字序列、2 个二进制数序列、以及 26 个英文子母序列组成的文章等等。比如一条英语新闻报导，其 *K*=26 字母+空格+全部标点符号种类，*n* 则是这条报导所含字母、空格和标点符号的总数目。现在假定：1）在信息中各个字符的出现彼此统计无关；2）每个字符 $\left(a_x, x=1,2,\cdots,K\right)$ 以先验的（当然也就是已知的）概率分布 $p\left(a_x\right)$ 出现，这里 $\sum_{x=1}^{K} p\left(a_x\right)=1$。比如，双字符中 1 出现的先验概率为 p，而 0 出现的先验概率就为 $\left(1-p\right), 0\le p\le 1$。

【定义】 如果序列中各字符均以先验的概率出现，称这种序列为“典型序列”。

考虑一个很长的信息（$n>>1$）。现在提一个问题：可以将这个很长的信息压缩为一个较短的（字符序列较短而又实质上传递相同信息的）信息吗？

以 n 个双字符为例。典型序列是这样一类序列，其中 0 个数为 $n(1-p)$，1 个数为 np。不同典型序列的数目，它们数量级为二项式系数 $\binom{n}{np}$。下面为了二进制编码方便，规定对数 *log* 以 2 为底。当 n 很大时 *Stirling* 公式（$logn!=nlogn-n+O(logn)$）给出：

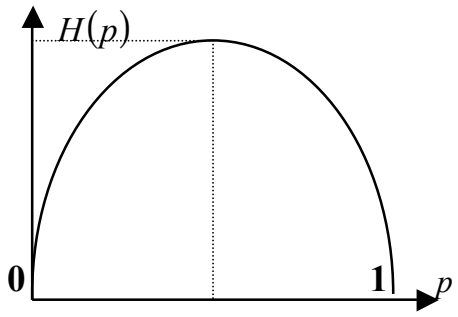
$$log\binom{n}{np}=log\left(\frac{n!}{(np)!\left[n(1-p)\right]!}\right)$$

$$\approx nlogn-n-\left\{nplog(np)-np\right\}-\left\{n(1-p)log\left[n(1-p)\right]-n(1-p)\right\} \quad (13.1)$$

$$=n\left\{-plogp-(1-p)log(1-p)\right\}\equiv nH(p)$$

$$\boxed{H(p)=-plogp-(1-p)log(1-p)} \quad (13.2)$$

$H(p)$ 称为 *Shannon* 熵函数。它满足 $0\le H(p)\le 1$，当 $H(p=1/2)=max=1$。它的图形如右图。



由此，对 n 个双字符情况，典型序列数目的量级是 $\binom{n}{np}=2^{nH(p)}$。于是，若只限于考虑字符出现概率符合先验概率分布的那一类序列，则 n 个双字符序列的总数大大下降，成为：

$$\boxed{2^n\to 2^{nH(p)}} \quad (13.3)$$

为了基本上传递全体（由 n 个二进制位的序列所荷载的）信息，并不需要对全部 n 位序列分别指定一个正整数的编码办法（码块），只需要选定一种对全部典型序列分别指定一个正

整数的编码办法就够了。这类码块（每个字符出现概率都符合先验概率分布）有 $2^{nH(p)}$ 个字符。于是可以用一个长度为 $nH(n)$ 的二进制序列来规定任一给定的长度为 n 字符的信息。

由于熵函数 $H(n)$ 的性质，对任何 $p\neq\frac{1}{2}$ 情况，这类码块的码长都会短于原来的信息。这便是 *Shannon* 的结果。这里关键是：当 $n\to\infty$ 时，实际信息不是典型序列的概率将渐近地忽略不计。所以无须对字符所组成的每个顺序都赋予编码的字码，只需对典型顺序做这种编码。这正表明了常说的，信息就是对无知的某种否定。对某个系统的信息，就是对该系统不确定性的认知，同时也是存储该系统熵所需的比特数。例子见下。

ii) 推广到 k 个字符情况。

推广到多个字符的情况。设有系综 $X:\{x=(x_1,x_2,\cdots,x_k),p(x)\}$，$p(x)$ 是出现字符 x 的“先验”概率。在总长为 n 个字符的一个序列中，x 发生的概率是 $np(x)$ 次。典型序列的数目在量级上为（再次使用 *Stirling* 公式）

$$\frac{n!}{\prod_x(np(x))!}\approx 2^{nH(x)} \qquad (13.4)$$

这里

$$\boxed{H(x)\equiv\sum_x -p(x)logp(x)\equiv\langle -logp(x)\rangle} \qquad (13.5)$$

这就是系综 X 的 *Shannon* 熵，或简称为经典系综的经典熵【习题 1】。选定编码：给每个典型序列指定一个正整数。一个有 n 位字符序列的信息便可以压缩成 $nH(x)$ 位。在这个意义上，系综 X

中所用的一个字符 x 荷载着 $H(x)$ 位信息。

【例 1】太阳从东方升起，这是 $p=1,1-p=0$ 的确定分布。此“系统”没有不确定性，熵为零，所需存储比特数 $H(p=1)=0$ 为零。现在你告诉我信息：明天太阳从东方升起。此信息不再减少关于这件事的无知，存储此信息所需比特数也为零。【例 2】向地上抛钱币，这是 $p=1-p=1/2$ 完全随机的分布。此“系统”完全不确定，熵达最大，需用 $H(1/2)=1$ 比特来存储关于它的“无知”。现在你告诉我信息：这次是“花”朝上。此猜测消除了 1 比特的不确定性，相应信息含量为 1 比特内容。【例 3】掷骰子中 6 个点数等概率出现。此系综是 $X=\{x_i, p_i=1/6, i=1,2,\cdots,6\}$。按典型序列来考虑，此系综熵为 $H(X)=log6=2.42$ 比特。即，需用此比特数来存储关于投掷结果的“无知”。现在你猜说：这次结果是个 5。此猜测所提供的信息消除了此比特数的熵，信息含量也是此比特数。【例 4】已知 7 人中有一人要出差，但不知是哪一位。这需用 3 比特来存储此“系统”的无知。你告诉我信息：是张三出差。此信息消除了 3 比特的不确定性，相应的信息量含有 3 比特的内容，需用 3 比特来存储。这些例子说明，为什么熵既是系统不确定性的量度，又是信息含量的量度。

2，*Shannon* 无噪声编码定理

i) 将上面叙述换一种稍稍不同的等价方式来表述。

“由 k 种字符排列组成的一条 n 个字符的信息

$$x_1x_2\cdots x_n;\quad x\in X\equiv\{x,p(x);k\} \tag{13.6a}$$

但由于 n 是有限的，字符 x 在此条信息中出现的次数 $n(x)$ 并不正比于先验概率 $p(x)$，而体现作为频度 $p'_n(x)=n(x)/n$。但显然有，

$$\begin{aligned}\lim_{n\to\infty}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log p'_n(x_i)\right\}&=\lim_{n\to\infty}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)\right\}\\&=\sum_{j=1}^{k}-p(x_j)\log p(x_j)\equiv\langle-\log p(x)\rangle\end{aligned} \tag{13.7"}$$

这里 $p(x_i)$ 是字符 x_i 的先验概率

$$\begin{cases}P(x_1x_2\cdots x_n)=p(x_1)p(x_2)\cdots p(x_n)\\ \log P(x_1x_2\cdots x_n)=\sum_{i=1}^{n}\log p(x_i)\end{cases} \tag{13.6b}$$

由此得到两条结论：

a) 一方面，对单个典型序列，有

$$\lim_{n\to\infty}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)\right\}=H(X) \tag{13.8}$$

假如是足够长而有限的序列，可采用 δ,ε 极限语言，将之精确地表述为

$$\left|-\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)-H(X)\right|<\delta \tag{13.9}$$

两边乘 n，取 2 指数，即得

$$2^{-n(H(X)-\delta)}\geq P(x_1\cdots x_n)\geq 2^{-n(H(X)+\delta)} \tag{13.10}$$

这就是当 n 足够大（δ 就足够小）时，每个典型序列出现的概率范围。

b) 另一方面，就所有典型序列总体而言，当 $n\to\infty$ 时典型序列出现的总概率趋近于 1。由此结合（13.10）式，可知

n字符典型序列总概率$=P\left(x_1\cdots x_n\right)\times 2^{nH(X)}=2^{-nH(X)}\times 2^{nH(X)}\Rightarrow 1$

或

$$P\left(x_1\cdots x_n\right)=2^{-nH(X)}\left(1-\varepsilon\left(n\right)\right)$$

或者说，n字符全体典型序列出现的总概率为$\left(1-\varepsilon\left(n\right)\right)$，

$$\sum_{\text{all}}P\left(x_1\cdots x_n\right)=1-\varepsilon\left(n\right)\qquad\text{（13.11）}$$

ii) *Shannon* 无噪声编码定理

总之，上面说明了，当信息序列的字符数目$n\to\infty$时，信息序列呈现为非典型序列的概率$\to 0$，从而只需要考虑此时的典型序列。对每一种典型序列指定一个正整数的编码方案即可应付，而不必对所有序列都进行编码，即不必对非典型序列也编码。这时如此做法所带来的误差不大于$\varepsilon\left(n\right)\xrightarrow{n\to\infty}0$。

下面说明：对任何系综$X=\left\{x_i,p_i\right\}$，每个字符的信息不能被压缩到$H\left(X\right)$位以下。因为这时已不能对全部典型序列的每一个典型序列指定唯一的码符。

这就是【*Shannon* 第一定理——*Shannon* 无噪声编码定理】“如果想进一步将信息中每位字符的信息压缩到$H-\delta'$，即使δ'为任一给定的小量，则当$n\to\infty$时也不能达到足够小的误差$\varepsilon\left(n\right)$。最佳编码只可以渐近（$n\to\infty$）地将每个字符压缩到$H\left(X\right)$位。”

【论证如下】

这时对信息（包括非典型序列信息，添ε'，见下）成功编码的概率为：

$P_{Sucess}\le\left(H-\delta'\right)$决定的典型序列总数$\times$每个典型序列出现概率上

限 $+\varepsilon' = 2^{n(H-\delta')} \cdot 2^{-n(H-\delta)} + \varepsilon' = 2^{-n(\delta'-\delta)} + \varepsilon' \xrightarrow{n\to\infty} 0$ （13.12）

此式意思是，这时能够对 $2^{n(H-\delta')}$ 个典型序列信息进行编码，这里每个典型序列信息出现的概率要小于 $2^{-n(H-\delta)}$（见前面（13.10）式）。添加 ε' 是考虑到有限字符时也需要对非典型序列进行编码（$\varepsilon' \xrightarrow{n\to\infty} 0$）。于是，对任一选定的小量 δ'，当 $n\to\infty$ 时 δ 逐渐减小，终将有 $(\delta'-\delta)>0$，这使成功的概率趋于零。

3，互信息

在 $n\to\infty$ 的渐近过程中， ***Shannon*** 熵表示从系综 X 取出单个字符所传递的平均信息量。因为它告诉我们，为了对那个信息编码需要多少位数。

互信息 $I(X,Y)$ 表示两个信息之间有多少关联。即，我们从所获得的 Y^n 序列信息中，能推知多少关于 X^n 序列信息？

例如，假定从一个传送者发送一条信息给一个接受者。但讯道有噪声，以致于收到的信息 y 可能不同于送出的信息 x。噪声通道可以用条件概率 $p(y|x)$（发送 x 时收到 y 的概率）来表征。我们假设，字符 x 将以先验的概率 $p(x)$ 发送。下面来定量研究：当我们收到 y 时能对 x 知道多少，也即获得了多少信息？

如同已经看到的，熵 $H(X)$ 表示，所收到的每个字符能消除我们在收到信息之前的无知。就是说，传送者传递 $nH(X)$ 位（不带噪声的）字符给接受者，以完整地表达（渐近地）一个特定的 n 字符的信息。假如了解了这条噪声通道的特性，也即知道条件概率 $p(y|x)$，而先验概率分布也是事先知道的，于是接受者在

收到 y 值之后，可以利用条件概率的 ***Bayes* 定理**

$$p(x|y)=\frac{p(y|x)p(x)}{p(y)}\equiv\frac{p(x,y)}{p(y)} \qquad (13.13)$$

$p(x,y)$是 x,y 的联合概率，假如 x,y 各自独立，则有 $p(x,y)=p(x)p(y)$。按全概率公式（见习题 1）可得 $p(y)$如下

$$p(y)=\sum_x p(y|x)p(x) \qquad (13.14)$$

鉴于已经获得这些知识，比起以前来，现在对 x 的知识是增加了（“无知”减少了些）。假定在收到这些 y 后，现在再采用最理想的码，就可以用每字符如下比特数

$$H(X|Y)=\langle-\log p(x|y)\rangle=-\sum_{xy}p(x,y)\log p(x|y) \qquad (13.15)$$

发送某个 ***n*** 字符序列将关于 X 的信息再传给别人。$H(X|Y)$称为条件熵，其意义是：一旦知道 y 的数值，平均而言的 X 的不确定性；或者说，知道 Y 后，为规定 x 每个字符所需要的附加位数。

【定义】联合熵：

$$H(X,Y)=-\sum_{x,y}p(x,y)\log p(x,y) \qquad (13.16)$$

显然，联合熵关于 X 和 Y 为对称的。还可以

【定义】互熵——互信息：

$$\begin{cases}I(X;Y)=-\sum_{x,y}p(x,y)\log p(x;y)\\ p(x;y)=p(x)p(y)/p(x,y)\end{cases} \qquad (13.17)$$

这里 $p(x;y)$ 为互概率。若 X 和 Y 无关，联合概率独立 $p(x,y)=p(x)p(y)$，得 $p(x;y)=1$，于是 $I(X;Y)=0$。由于互概率 ≤ 1，所以 $I(X;Y)\ge 0$。$I(X;Y)$的物理意义是：当已知 Y 之后，为确定 X 所需每个字符的位数的减少数。或者说，通过已知 Y 所能获得的关

于 X 每个字符的比特数。又可以说，通过测量 Y 所得到的关于 X 的信息。

若有同一字符库的两种概率分布 $p(x)$和 $q(x)$，可以

【定义】 $p(x)$**相对于** $q(x)$**的相对熵：**

$$H\big(p(x)\|q(x)\big)=\sum_x p(x)\log\frac{p(x)}{q(x)}=-H(X)-\sum_x p(x)\log q(x) \qquad (13.18)$$

4， $H(X,Y),H(X|Y),I(X;Y)$性质总结

i, 联合熵的对称性 $H(X,Y)=H(Y,X)$;

ii, 条件熵非负性 $H(X|Y)\geq 0$。若 X,Y 相互独立，$H(X|Y)=H(X)$;

iii, 互信息的对称性 $I(X;Y)=I(Y;X)$。互信息的范围

$$0\leq I(X;Y)\leq \min(H(X),H(Y))$$

证明：由于 $0\leq p(x,y)\leq p(y)\leq 1$，所以

$$\log p(y)-\log p(x,y)\geq 0$$

$$\begin{aligned}\therefore\quad I(X;Y)&=-\sum_{x,y}p(x,y)\log\frac{p(x)p(y)}{p(x,y)}\\&=-\sum_{x,y}p(x,y)\{\log p(x)+\log p(y)-\log p(x,y)\}\\&\leq-\sum_{x,y}p(x,y)\log p(x)=-\sum_x p(x)\log p(x)=H(X)\end{aligned}$$

同样可得 $I(X;Y)\leq H(Y)$。

iv, 条件熵、联合熵、子系统熵的关系

$$\begin{cases}H(X|Y)=H(X,Y)-H(Y)\\H(Y|X)=H(X,Y)-H(X)\end{cases}$$

证明：$H(X|Y)=-\sum_{x,y}p(x,y)\log p(x|y)=-\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(y)}=H(X,Y)-H(Y)$。

由于$0 \le p(x,y) \le p(y)$，所以条件熵总是非负的$H(X|Y) \ge 0$。

v, 互熵、联合熵、子系统熵的关系

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

vi, 条件熵、互熵、子系统熵的关系

$$H(X) = H(X|Y) + I(X;Y); \quad H(Y) = H(Y|X) + I(X;Y)$$

vii, $H(X,Y) = H(X|Y) + H(Y|X) + I(X;Y)$

viii, 经典熵的基本不等式

$$Max(H(X), H(Y)) \le H(X,Y) \le H(X) + H(Y)$$

【证明】对于右边不等号$H(X,Y) \le H(X) + H(Y)$是因为，$H(X;Y) \ge 0$。对于左边不等号，是因为

$$-\log p(x,y) \ge -\log p(x)$$

$$\therefore \quad -\sum_{x,y} p(x,y)\log p(x,y) \ge -\sum_{x,y} p(x,y)\log p(x) = -\sum_{x} p(x)\log p(x) = H(X)$$

同样有$-\sum_{x,y} p(x,y)\log p(x,y) \ge H(Y)$。总之左边不等号成立。此式的物理意义：复合系统$X+Y$的熵$H(X,Y)$不可能小于任一子系统***X***或***Y***的熵。

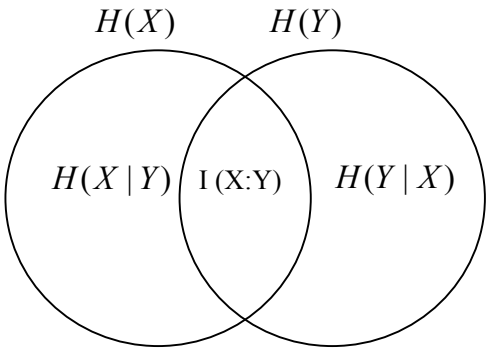
$H(X,Y)$的上限是子系统X和Y互相独立；下限相应于子系统X和Y之间有最大的经典关联。

xi, 相对熵$H(p(x)\|q(x))$的数值是非负的。注意对任意正数X有$\ln X \ge 1 - X^{-1}$，令$p(x)/q(x) = X$，换成以 **2** 为底的对数 **log**，得

$$H\left(p(x)\|q(x)\right)\equiv\sum_x p(x)\log\frac{p(x)}{q(x)}\geq\frac{1}{(\log e)}\sum_x p(x)\left(1-\frac{q(x)}{p(x)}\right)$$
$$=\frac{1}{(\log e)}\sum_x\left(p(x)-q(x)\right)=0$$

等号当且仅当 $p(x)=q(x)\ \forall x$ 时成立。

[关系示意图]



5，*Shannon* 噪声讯道编码定理

i, 若要在一条噪声讯道上传递信息，显然必须采用冗余度来增加传输的可靠性。比如，可以将每个比特多次发送，接收者可以采用多数选定制的办法去解比特。等等。

但是，给定一个噪声讯道之后，总能找到一种编码，当 $n\to\infty$ 时确保有任意好的可靠性吗？关于这种编码的比率——每一信息的每个字母需要多少比特的问题，能得出些什么结论呢？

事实上，***Shannon*** 证明了：只要在输入和输出之间有一定的关联，任何讯道都可以在一个有限比率之下作任意可靠的通讯。此外，他发现了对于能够达到的最佳比率的一个有用表达式。这些结果便是噪声讯道编码定理。下面分层次论述，并在最后给出这个定理。

为了具体，假定采用的是二进制字母表：**0** 和 **1**，各以一个

先验的概率发生。并且假定，讯道是“二进制对称通道”——它对每一个位的作用是独立的：以概率 p 翻转它的值，以 $(1-p)$ 的概率原封不动的保留它的值。就是说，条件概率是：

$$p(0|0)=1-p \quad p(0|1)=p(1|0)=p$$
$$p(1|1)=1-p$$

这时 $H(p)=-p\log p-(1-p)\log(1-p)$。现在打算针对增加着的块容积 n（因为信息长度在增加着）去构造、寻找各种编码方案，使得解码误差的概率当 $n\to\infty$ 时趋于零。假定编在块中的数据比特数是 $k(<n)$，于是码由 n 比特的 2^n 种可能弦中 2^k 个码符的某种选择所构成（即选择编码——如何用 2^n 弦荷载 2^k 弦，或用 n 位荷载 k 位的各种约定）。

【定义】码的比率 R 等于每传送 1 比特所荷载的数据比特数，

$$R=\frac{k}{n} \quad (<1, \quad nR=k) \qquad (13.19)$$

k 不能太大，∵ k 上升 $\to R$ 上升 $\to$ 传送信息的冗余度下降 $\to$ 造成误译。

应当这样设计好的码，使得码弦之间彼此尽可能的分开。就是说，对于给定的 R，必须使从一个码符到另一个码符变化时，所必需翻转的位数尽可能的多（这个数被称作两个码符之间的 *Hamming* 距离）。

对于一条 n 位长的输入弦，一般性地说，误差将引起大约 np 位的误翻转——于是这一个输入一般将扩散到大约 $2^{nH(p)}$ 个典型输出弦中的一个。占据着以输入弦为中心的、以 *Hamming* 距离 np 为半径的一个“球”。∴ $2^k\to 2^k\cdot 2^{nH(p)}$ 个典型序列，每个被编在 n 位

中的k位码字符都扩散成为“球”。于是，要求各个球之间不重合就必须要求

$$2^k\text{个球的总体积}\le 2^n \qquad (13.20)$$

为了可靠地解码，就必须选择我们的输入码符，使得两个不同的码符的误差球不产生交叠。否则两个不同的输入会有同样的输出，解码误差就不可避免了。如果要避免这样的解码不确定性，（包含在总数为2^{nR}个误差球内的）弦的总数不能超过输出信息总数2^n。即需要

$$2^{nH(p)}\cdot 2^{nR}\le 2^n \qquad (13.21)$$

即：2^k码符数$\times n$序列压缩为典型序列数≤ ***n***序列总数。或为

$$0\le R\le 1-H(p)\equiv C(p) \qquad (13.22)$$

（若无噪声，则$p=0$或$p=1\to H(p)=0\to C(0)=C(1)=1$;若最大噪声，$p=\frac{1}{2}\to H\left(\frac{1}{2}\right)=1\to R=0$，不能传信息）。若要传送是高度可靠的，就不能指望码比率R超过$C(p)$。但比率$R=C(p)$在实际上是否能够渐近地达到的呢？就是说，平均而言，在可靠传送下，当$n\to\infty$时，R可以$\to C$吗？事实上，以任意接近C和任意小误差概率的R比率的传送是可能的。也许 ***Shannon*** 最富创造性的思想就是说明了：通过考虑在“随机码”上的平均，可以达到C。显然，随机地选择一个码并不是可能办法中最巧妙的。但结果却让人惊讶：随机编码和任何其它编码方案一样，在大n下也能渐近地获得高的比率。由于C是数据沿噪声通道可靠传送的最佳比率，所以被称作讯道的容量。

假设 2^{nR} 个码符用随机地从系综 X^n 中取样来选定。在传送一个信息（码符中的一个）时，为了解读这个信息码，画出一个围绕着所收到信息的“*Hamming* 球”，这个收到的信息将包含

$$2^{n(H(p)+\delta)}$$

个典型弦。假定包含在这个球中的这样一个码符存在并且是唯一的话，信息就被解码成为这个码符；如果这样的码符不存在，或是码符不唯一，就假设发生了一个解码错误。

一个解码错误多半是怎样的呢？已经选定解码球足够大，使得出现在球中的一个适当码符错误是非典型的，所以只需要关心占据球的多个适当的码符。因为总共存在 2^n 个可能的弦，围绕这个输出，一个 ***Hamming*** 球中的典型弦占全部弦中的份额为

$$\left(\frac{2^{n(H(p)+\delta)}}{2^n}=2^{-n(C(p)-\delta)}\right)$$

于是，2^{nR} 个随机选择码符中的一个码符以偶然的方式占据这个球的概率是

$$2^k\cdot 2^{-n(C(p)-\delta)}=2^{-n(C(p)-R-\delta)}$$

鉴于可以选择 δ 为任意小，R 可以被选为任意接近 C（但要在 C 的下方），于是这个失误概率当 $n\to\infty$ 时，将仍旧按指数减小。

迄今证明了，平均失误概率是小的。这里对所选择的随机码作平均，并且对每一个规定的码作平均，也对全部码作平均。于是必定存在一个具体的编码，这时对此码符平均的失误概率小于 ε。

ii, 进一步，愿意有一个更强的结果——失误概率对于每个码符都小。为了建立起这个更强的结果，令 p_i 表示传送码

符i时解码失误的概率。已经说明存在一个码，使得

$$\frac{1}{2^{nR}}\sum_{i=1}^{2^{nR}} p_i < \varepsilon$$

令$N_{2\varepsilon}$表示$p_i > 2\varepsilon$的码符的数目。于是我们推断

$$\frac{1}{2^{nR}}(N_{2\varepsilon})2\varepsilon < \varepsilon，\ 或\ N_{2\varepsilon} < 2^{nR-1} = 2^{n\left(R-\frac{1}{n}\right)}$$

我们看到，至多能抛弃掉码符的一半，以达到对每一个码符有$p_i < 2\varepsilon$。构造出的新码有比率，

$$rate = R - \frac{1}{n}; \qquad \left(\frac{1}{2}2^k = 2^{nR-1} = 2^{n\left(R-\frac{1}{n}\right)}\right)$$

它当$n \to \infty$时趋于R。

于是看到，

$$C(p) = 1 - H(p)$$

这是能渐近达到的、有任意小失误概率的最大传输比率。

iii)现在考虑怎样将这些论证推广到更一般的字符表和噪声道。给定一个由$p(y|x)$等等所规定的通道，并为输入字符规定一个概率分布$X = \{x, p(x)\}$。设发送n个字符的一个弦，并假定通道对每个字符独立的作用（一个按如此方式工作的通道称为无记忆通道）。自然，一旦$p(y|x)$和$X = \{x, p(x)\}$给定了，则$p(x|y)$和$Y = \{y, p(y)\}$也就确定了。

为了建立一个可达到的比率，再次考虑对随机码的平均。那里码符是以先验概率（由X^n支配的）选择的。于是，这些码符将以高的概率从字符弦的典型弦序列中选定。那里大约有$2^{nH(X)}$个这样的典型弦。

对于Y^n中一个典型的接受信息，存在着大约$2^{nH(X|Y)}$个能够送出来的信息。可以将收到的信息和一个包含$2^{n(H(X|Y)+\delta)}$个可能输入的一个球相联系的办法来解码。如果在这个球中存在唯一的码符，就从那个码符中解码出信息。

和前面一样，球中不可能不存在码符，但必须排除多于一个码符的概率。每个解码球包含典型输入的一个份额

$$\frac{2^{n(H(X|Y)+\delta)}}{2^{nH(X)}}=2^{-n[H(X)-H(X|Y)-\delta]}=2^{-n[I(X;Y)-\delta]}$$

这里$I(X;Y)=H(X)-H(X|Y)$为前面的互信息。假如存在2^{nR}个码符，任一个偶然落入解码球中的概率是

$$2^{nR}2^{-n[I(X;Y)-\delta]}=2^{-n[I(X;Y)-R-\delta]}$$

由于δ可选作任意小，可以选择R尽量接近I（但要小于I），并仍然使解码失误概率当$n\to\infty$时呈指数减少。

这个论证表明，对随机码和码符作平均时，一次失误的概率对于任意比率$R<I$来说变得很小。和前面同样的理由，说明对于每个码符以失误概率$<\varepsilon$的一个特殊码的存在性。这是一个满意的结果，因为它与对I的解释是一致的，将I解释作为这样的信息：当获得信号Y时，所得到的输入X的信息——就是说，I是能够沿通道传送的每个字母的信息。

互信息$I(X;Y)$不仅依赖于讯道的条件概率$p(y|x)$，而且也依赖于字符的先验概率分布$p(x)$。上面的随机编码论证能应用于$p(x)$的任意选择，所以就说明了，对于任何小于

$$C\equiv\underset{\{p(x)\}}{Max}\,I(X;Y) \qquad (13,23)$$

比率 R 作无失误的传送都是可能的。C 被称作讯道的容量，它现在只决定于讯道的条件概率 $p(y|x)$。

iv)至此已经证明了，任何 $R<C$ 的比率是可以达到的。但是，还必须回答：当 R 超过 C 的情况下，当 $n\to\infty$ 时失误概率仍然趋于零吗？在一般情况下 C 是比率 R 的上限的证明，考虑到对不同字符失误概率会有不同，比起利用双对称通道来作的证明看来需要更细緻一些，在设计码的时候并没有考察这一点，现在推究如下：

假设已选 2^{nR} 个字符的弦作为码符。考虑一个概率分布（记为 $\tilde{X}^n$），在这个分布中每个码符以等概率（$=2^{-nR}$）发生。显然有

$$H(\tilde{X}^n)=nR\ （左边=-\sum_{i=1}^{2^{nR}}(2^{-nR})\log(2^{-nR})=\frac{1}{2^{nR}}\cdot nR\cdot 2^{nR}=nR）$$

通过讯道送出码符时，得到输出态的一个概率分布 $\tilde{Y}^n$。

因为假设讯道对每个字符的作用是独立的，一条 n 字符的弦的条件概率可以因式化为

$$p(y_1y_2\cdots y_n|x_1x_2\cdots x_n)=p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n)$$

于是可得条件熵满足

$$H(\tilde{Y}^n|\tilde{X}^n)=\langle -\log p(y^n|x^n)\rangle=\sum_i\langle -\log p(y_i|x_i)\rangle=\sum_i H(\tilde{Y}_i|\tilde{X}_i)$$

这里 $\tilde{X}_i$ 和 $\tilde{Y}_i$ 是在码符上的分布所决定的第 i 个字符的边缘概率（***marginal probability***）分布。回忆起 $H(X,Y)\le H(X)+H(Y)$，或

$$\boldsymbol{H(\tilde{Y}^n)\le\sum_i H(\tilde{Y}_i)}$$

可得

$$I(\tilde{Y}^n;\tilde{X}^n)=H(\tilde{Y}^n)-H(\tilde{Y}^n|\tilde{X}^n)$$
$$=\sum_i\{H(\tilde{Y}_i)-H(\tilde{Y}_i|\tilde{X}_i)\}=\sum_i I(\tilde{Y}_i;\tilde{X}_i)\le nC$$

送出和收到（n个字母的）一条信息的互信息的上限不超过每个字母的互信息之和；而每个字符的互信息上限不超过容量C（$\because C$被定义为$I(X;Y)$的最大值）。

忆及互信息的对称性，有

$$I(\tilde{X}^n;\tilde{Y}^n)=H(\tilde{X}^n)-H(\tilde{X}^n|\tilde{Y}^n)=nR-H(\tilde{X}^n|\tilde{Y}^n)\le nC$$

现在，如果无失误传送是可能的，也即我们能当$n\to\infty$时可靠地解码，就意味着，输入的码符完全由所收到的信号所决定，或是每个字符的输入条件熵必定很小

$$\frac{1}{n}H(\tilde{X}^n|\tilde{Y}^n)\to 0$$

于是由上面$I(\tilde{X}^n;\tilde{Y}^n)$第二式即得

$$R\le C \qquad (13.24)$$

等号当$n\to\infty$极限时成立。比率不能超过容量（回忆起条件熵是不对称的，不像互信息）。（实际上，$\frac{1}{n}H(\tilde{Y}^n|\tilde{X}^n)$并不变小，因为讯道引入了对于将接受的信息的不确定性。但是假如我们能精确地解码，只要信息被接收到了，就不存在关于被传送码符的不确定性。）

到此，我们就证明了 *Shannon* 噪声讯道编码定理——*Shannon* 第二定理：“容量$C\equiv\underset{\{p(x)\}}{Max}I(X;Y)$是通过噪声讯道通讯时所能达到的最高比率。这时，当信息字符数趋于无穷时，失误概率趋于零”。也可以换一种说法：“设有噪声离散无记忆

讯道，其容量为C。只要信息传输率$R<C$，则总存在编码方法，可使译码错误概率渐近地任意小；若$R>C$，则不存在这样的编码方法”。

自然，上面用以证明$R=C$的方法是在渐近下获得的（沿随机码平均）。这并不是一个十分肯定和积极的证明，因为一个随机码没有结构和型式，编码和解码会是相当笨重和不合用的（显然它需要一本指数增大的码书）。尽管如此，定理是十分重要和有用的。因为它告诉我们：什么是原则上可以达到的，什么是原则上不可能达到的。并且，因为$I(X;Y)$是$X=\{x,p(x)\}$（$p(y|x)$固定）的一个凸函数（*concave function*——实际含意为向外凸性，但按数学定义是如此用字）。它有唯一的一个局域极大值。于是对感兴趣的讯道，C是可以算出来的（至少可以数值计算出来）。