信息论笔记

BY WHZECOMJM

Bar-Ilan University
December 27, 2019

这是我在 BIU 2019/2020 年度信息论课中的课程笔记。 本课的授课老师是 Reuven Cohen。该课的参考书籍如下:

1. Thomas M. Cover, Joy A. Thomas. Elements of information theory.

2. David JC MacKay. Information Theory, Inference and Learning Algorithms.

这篇笔记的大部分内容均可以在 Cover 和 Thomas 书中找到。我在下文引用的引理、 定理等如果有编号, 则其与这本书中的编号相同。我会尽量用中文介绍这么课程， 但是在描述引理、 定理及其证明时时可能会保留英文原文。

目录

1 信息熵

1.1 信息

> If we can compute the result based on our knowledge, then the result is not an information.

我们从一个简单的例子入手： 比如"圆周率π的第100万位数字是1", 当我们有圆周率的概念和知识以后, 这一数据就不再是信息, 即便我们不能马上判断上述内容是否为真。

关于信息 Information $i(p)$ 的公理定义:

定义 1.1. *$i(p)$ is the information of p with $0<p\leq 1$ satisfying:*

1. $i(p_1\,p_2)=i(p_1)+i(p_2)$*;*

2. $i(p)\geq 0$.

由此我们可以得到

引理 1.2. *$i(p)$ is non-increasing regarding to p.*

证明. If $p_1>p_2$, then

$$i(p_2)=i\left(p_1\cdot\frac{p_2}{p_1}\right)=i(p_1)+i\left(\frac{p_2}{p_1}\right)\geqslant i(p_1).$$

Note that $0<i\left(\frac{p_2}{p_1}\right)\leq 1$. □

进一步地, 我们可以得到信息的表达式:

命题 1.3. *$i(p)=-c\ln p$ for every fixed $c>0$.*

证明. $\forall n\in\mathbb{N}$, we have $i(p^n)=n\ i(p)$. For rational number, assume that $p_2=p_1^{a/b}$, then

$$p_2^b=p_1^a\Longrightarrow b\ i(p_2)=a\ i(p_1)\Longrightarrow i(p_2)=\frac{a}{b}\,i(p_1)=\frac{\ln p_2}{\ln p_1}\,i(p_1).$$

For $x\in R$, $p_2=p_1^x$, then there exist sequences of rational numbers $r_n\searrow x$ and $a_n\nearrow x$ such that

$$i(p_2)=\frac{\ln p_2}{\ln p_1}\,i(p_1),\quad \forall x$$

by the completion of $\mathbb{R}$.

Then, we have $\frac{i(p_1)}{\ln p_1}=\frac{i(p_2)}{\ln p_2}=-c$, which completes the proof. □

为了方便， 我们令$c=\frac{1}{\ln 2}$，则有 $i(p)=-\frac{1}{\ln 2}\ln(p)=-\log_2 p$. 为了简便，我们以后均使用这一定义，且将 $\log_2$ 简写为 log.

1.2 信息熵

接下来我们将讨论信息熵的概念， 首先我们给出信息熵（entropy）的定义：

定义 1.4. (信息熵) *The entropy $H(X)$ of a discrete random variable X is defined by*

$$H(X) = -\sum_{x\in\mathcal{X}} p(X)\log p(x).$$

我们有时也会写做 $H(p)$。容易验证，$H(X)\geq 0$。为了方便，我们约定 $0\log 0 = 0$，因为这能与$x\log x$ 的连续性相吻合（$x\log x \to 0$ as $x\to 0$）。

例 1.5. 投掷一次均匀硬币出现正反面的概率分布函数为 $p(x)$，则

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = -\log\frac{1}{2} = 1.$$

例 1.6. 对于两点分布（p,1-p），我们有

$$H(p) = H(p; 1-p) = -p\log p - (1-p)\log(1-p)$$

当 $p=\frac{1}{2}$，$H(p)$ 取最大值 1.

我们将**期望**记为 E，如果 $X\sim p(x)$, 则随机变量 $g(X)$ 的期望值记为

$$E_p\, g(X) = \sum_{x\in\mathcal{X}} g(x)\, p(x),$$

或者简记为 $E\,p(X)$ 当概率分布函数 p 不会产生混淆时。

1.2.1 联合熵和条件熵

我们先给出联合信息熵和条件信息熵的定义：

定义 1.7. *The joint entropy $H(X,Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x,y)$ is defined as*

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -E\log p(X,Y). \end{aligned}$$

注意到如果 X,Y 是相互独立的两个随机变量，则有

$$H(X,Y) = H(X) + H(Y).$$

定义 1.8. *If $(X,Y)\sim p(x,y)$, the conditional entropy $H(Y|X)$ is defined as*

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)\, H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -E\log p(Y|X). \end{aligned}$$

注意到上式地第三行公式中前面的p是联合概率。

联合熵和条件信息熵有如下关系， 我们称之为链式法则（Chain Rule）。

定理 1.9. (Chain Rule)

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y).$$

证明. Note that

$$\log p(X,Y)=\log p(X)\,p(Y|X)=\log p(X)+\log p(X|Y),$$

and take the expectation of both sides of the equation to obtain the theorem. □

1.2.2 相对熵与相互信息

定义 1.10. *The relative entropy or Kullback–Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D\left(p||\,q\right)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}=E_p\log\frac{p(X)}{q(X)}.$$

在上述定义中，我们需要用到约定 $0\log\frac{0}{0}=0\log\frac{0}{q}=0$ 以及 $p\log\frac{p}{0}=\infty$. 也就是说，如果存在任意 $x\in\mathcal{X}$ 使得 $p(x)>0$ 且 $q(x)=0$， 则 $D\left(p||\,q\right)=\infty$.

我们将会证明 $D\left(p||\,q\right)\geq 0$， 且等号成立当且仅当两个概率分布一致 $p=q$。虽然我们常常会用相对熵来表示分布之间的距离，但是注意到这一定义通常不是对称的且不满足三角不等式。

不过我们可以用相对熵定义一个对称的相互信息， 它用以衡量两个随机变量之间共同的信息。

定义 1.11. *Consider two random variables X and Y with a joint probability mass function $p(x,\,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)\,p(y)$:*

$$\begin{aligned}I(X;Y)=&\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}\\=&D\left(p(x,y)||\,p(x)\,p(y)\right)\\=&E_{p(x,y)}\log\frac{p(X,Y)}{p(X)\,p(Y)}.\end{aligned}$$

相互信息和信息熵的关系由如下定理给出：

定理 1.12. (Mutual information and entropy)

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X).
\end{aligned}
$$

注意到，相互信息有上界 $H(X)$ 和 $H(Y)$。这些关系可以很容易地由定义计算得出。我们可以把它们的关系表示为如下的文氏图。

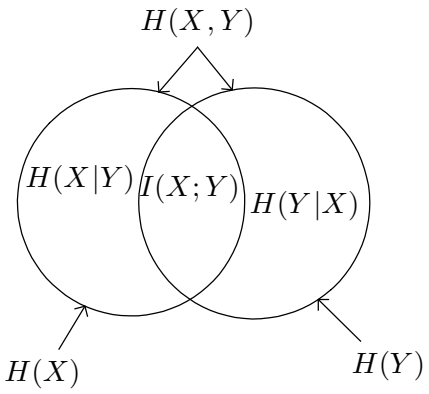


图 1.1. 相互信息和联合熵等的关系图

1.2.3 多个随机变量的信息熵

对于多个随机变量， 我们也有推广的熵的链式法则。

定理 1.13. (Chain rule for entropy) *Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, \ldots, x_n)$. Then*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1).$$

归纳法证明参见[1] 的 pp. 22-23.

由于在信息熵的多个随机变量中， 我们一般没有幂的概念， 所以常常用上下标表示初始和结束位置用以简写。比如上式可以简写为

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}).$$

另外， 我们约定公式右边第一项（即 $i=1$时）为 $H(X_1)$. 比如 n=2 时， 我们有 Theorem 2.2.1 [1]:

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1).$$

定义 1.14. *The conditional mutual information of random variable X and Y given Z is defined by*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)\,p(Y|Z)}.$$

相互信息因此也有链式法则。

定理 1.15. (Chain rule for information)

$$I(X_1^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1^{i-1}).$$

证明略， 参见 p24 [1].

我们可以定义条件概率版本的相对信息熵。

定义 1.16. *For joint probability mass functions $p(x,y)$ and $q(x,y)$, the conditional relative entropy $D\left(p(y|x)||\,q(y|x)\right)$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. Namely,*

$$D\left(p(y|x)||\,q(y|x)\right) = E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}.$$

相应地我们有相对熵的链式法则：

定理 1.17. (Chain rule for relative entropy)

$$D\left(p(x,y)||\,q(x,y)\right) = D\left(p(x)||\,q(x)\right) + D\left(p(y|x)||\,q(y|x)\right).$$

2 信息不等式

2.1 Jensen 不等式

在这里我们不再赘述凹凸函数的定义， 详细参见 Definitions in p.25 [1]. 存在二阶导数的函数的凹凸性判定也是已知内容， 不再列出。我们常用在信息论的Jensen不等式的表达如下：

定理 2.1. (Jensen's inequality) *If f is a convex function and X is a random variable, then*

$$\mathrm{E}\, f(X) \geqslant f\left(\mathrm{E}X\right).$$

Moreover, if f is strictly convex, the equality in above implies that $X = EX$ with probability 1 (i.e. X is a constant).

证明. The discrete distributions case can be proved by induction on the number of mass points. □

Jensen 不等式的一个重要的应用就是信息论中的信息不等式：

定理 2.2. (Information inequality) *Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. Then*

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

证明. Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned} -D(p||q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\ &\leqslant \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \qquad (2.1) \\ &= \log \sum_{x \in A} q(x) \\ &\leqslant \log \sum_{x \in \mathcal{X}} q(x) \qquad (2.2) \\ &= \log 1 = 0, \end{aligned}$$

where (2.1) follows from Jensen's inequality. Since $\log t$ is a strictly concave function of t , we have equality in (2.1) if and only if $q(x)/p(x)$is constant everywhere [i.e., $q(x) = c\,p(x)$ for all x]. Thus,

$$\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c.$$

We have equality in (2.2) only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x. □

上述信息不等式的直接推论是相互信息是非负的。

推论 2.3. (Nonnegativity of mutual information) *For any two random variables, X , Y,*

$$I(X;Y) \geq 0,$$

with equality if and only if X, Y are independent.

当然条件相互信息也是非负的。

推论 2.4.

$$I(X;Y|Z) \geq 0,$$

with equality if and only if X and Y are conditionally independent given Z.

我们将要看到均匀分布是最大信息熵的分布。注意到， 任何随机变量在其取值范围上的信息熵都不超过 $\log|\mathcal{X}|$.

定理 2.5. *$H(X)\leq\log|\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of $\mathcal{X}$, with equality if and only $\mathcal{X}$ has a uniform distribution over $\mathcal{X}$.*

证明. Let $u(x)=\frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over $\mathcal{X}$,and let $p(x)$ be the probability mass function for $\mathcal{X}$. Then

$$D(p\|u)=\sum p(x)\log\frac{p(x)}{u(x)}=\log|\mathcal{X}|-H(X).$$

Hence by the nonnegativity of relative entropy,

$$0\leq D(p||u)=\log|\mathcal{X}|-H(X).$$ □

定理 2.6. (Independent bound on entropy) *Let $X_1,\dots X_n$ be drawn according to $p(x_1,\dots,x_n)$. Then*

$$H(X_1^n)\leq\sum_{i=1}^{n}H(X_i)$$

with equality iff the X_i are independent.

证明. By the chain rule for entropies,

$$H(X_1^n)=\sum_{i=1}^{n}H(X_i|X_1^{i-1})\leq\sum_{i=1}^{n}H(X_i).$$

We have equality if and only if X_i is independent of $X_{i-1},\dots,X_1$ for all i (i.e., iff the X_i's are independent). □

2.2 数据处理不等式

数据处理不等式可以用来表明， 任何对数据的巧妙操纵就不能改善从数据中得出的推论。我们先给出 Markov 链的定义， 这一定义在随机过程的课程中已经学习过。

定义 2.7. *Random variables X,Y,Z are said to form a Markov chain in that order (denoted by $X\to Y\to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y , and Z form a Markov chain $X\to Y\to Z$ if the joint probability mass function can be written as*

$$p(x,y,z)=p(x)\,p(y|x)\,p(z|y).$$

上述定义的一些性质有：

1. $X\to Y\to Z\iff$ X 和 Z 是给定Y条件独立的。

2. $X \to Y \to Z \Longrightarrow Z \to Y \to X \Longrightarrow X \leftrightarrow Y \leftrightarrow Z$.

3. 如果 $Z = f(Y)$, 则有 $X \to Y \to Z$.

我们现在能证明非常重要的数据处理不等式：

定理 2.8. (Data-processing inequality) *If $X \to Y \to Z$, then*

$$I(X;Y) \geq I(X;Z).$$

我们可以这样理解： Y是一个信息传输过程， X是信息源头， Z是接收到的信息。也就是说， 源头和传输的相互信息是不小于源头和终点的相互信息。

证明. By the chain rule, we expand mutual information in two different ways:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y).$$

Since X and Z are conditionally independent given Y , we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, we have

$$I(X;Y) \geq I(X;Z).$$

We have equality iff $I(X;Y|Z) = 0$ (i.e., $X \to Z \to Y$ forms a Markov chain). Similarly, one can prove that $I(Y;Z) \geq I(X;Z)$. □

上述定理有如下推论：

1. $I(X;Y) \geq I(X;g(Y))$.

2. If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$.

2.3 Fano 不等式

根据之前的讨论, 如果存在两个随机变量 X (未知), Y(已知), 我们想要用 Y 来估计 X。容易知道， $H(X|Y) = 0$ 与用 Y 做0概率误差估计是等价的。现实的情况稍微比较复杂，条件熵是可能大于0的，因此用Y做估计也会有一定的误差，然而很巧合的是，这一误差被条件熵所限制下限。

假设我们想要用概率分布函数$p(x)$估计一个随机变量 X。我们观察一个与之相关的随机变量 Y 使得条件分布为 $p(y|x)$. 从 Y 中我们可以计算函数 $g(Y) = \hat{X}$, 其中 $\hat{X}$ 就是 X 的估计，且取值范围相应的为 $\hat{\mathcal{X}}$ （不要求 $\hat{\mathcal{X}} = \mathcal{X}$）。我们要求函数 $g(Y)$ 是随机的，我们想要给 $\hat{X} \neq X$ 的概率定界。注意到， $X \to Y \to \hat{X}$ 是一个 Markov 链， 定义误差概率

$$P_e = \Pr\{\hat{X} \neq X\}.$$

我们有如下定理：

定理 2.9. (Fano's inequality) *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr\{\hat{X} \neq X\}$, we have*

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

This inequality can be weakened to

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y)$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

证明. Define

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X, \\ 0 & \text{if } \hat{X} = X. \end{cases}$$

Then, using the chain rule for entropies to expand $H(E, X|\hat{X})$ in two ways, we have

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}). \end{aligned}$$

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Now since E is a function of X and $\hat{X}$, the conditional entropy $H(E|X, \hat{X})$ is equal to 0. Also, since E is a binary-valued random variable, $H(E) = H(P_e)$. The remaining term can be bounded as follows:

$$\begin{aligned} H(X|E, \hat{X}) &= \Pr(E=0)\, H(X|\hat{X}, E=0) + \Pr(E=1)\, H(X|\hat{X}, E=1) \\ &\leq (1 - P_e)\, 0 + P_e \log |\hat{X}|. \end{aligned}$$

Combining these results, we obtain

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}).$$

By the data-processing inequality, we have $I(X; \hat{X}) \leq I(X; Y)$, and therefore $H(X|\hat{X}) \geq H(X|Y)$. □

弱化形式是因为很明显有 $H(P_e) = H(E) \leq 1 = -2 \cdot \frac{1}{2} \log \frac{1}{2}$.

3 渐近等分性质

3.1 渐近等分性质

在信息论中， 渐近等分性质（ Asymptotic Equipartition Property, AEP）是大数定律的类比。它是弱大数定律的直接推论。

定义 3.1. (Convergence of random variables) *Given a sequence of random variables, $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$ converges to a random variable X:*

1. *In probability if for every $\epsilon > 0, \Pr\{|X_n - X| > \epsilon\} \to 0$.*
2. *In mean square if $E(X_n - X)^2 \to 0$.*
3. *With probability 1 (also called almost surely) if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$.*

渐近等分性质由下列定理给出：

定理 3.2. (AEP) *If $X_1, X_2, \ldots$ are i.i.d.$\sim p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \quad \textit{in probability.}$$

证明. Functions of independent random variables are also independent random variables. Thus, since the X_i are i.i.d., so are $\log p(X_i)$. Hence, by the weak law of large numbers,

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\sum_i \log p(X_i) \\ &\to -E\log p(X) \text{in probability} \\ &= H(X) \end{aligned}$$

which proves the theorem. □

3.2 典型集合

我们将引入一个重要概念：典型集合（typical set）。

定义 3.3. *The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ with the property*

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}.$$

根据 AEP， 我们可以得出 $A_\epsilon^{(n)}$ 的如下性质：

定理 3.4.

1. *If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \le -\frac{1}{n}\log p(x_1, \ldots, x_n) \le H(X) + \epsilon$.*
2. *$\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for n sufficiently large.*
3. *$|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in the set A.*
4. *$A_\epsilon^{(n)} \ge (1-\epsilon)\, 2^{n(H(X)-\epsilon)}$ for n sufficiently large.*

因此， typical 集的概率几乎为1（性质2）， 且其中所有的元素都几乎是一样的、 等概率的（性质1）， 集合中元素的个数将近是 2^{nH}（性质3-4）。

证明. The proof of property (1) is immediate from the definition of $A_\epsilon^{(n)}$. The second property follows directly from Theorem 3.1.1, since the probability of the event $(X_1, \ldots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. Thus, for any $\delta > 0$, there exists an n_0 such that for all $n \geq n_0$, we have

$$\Pr\left\{\left|-\frac{1}{n}\log p(X_1, \ldots, X_n) - H(X)\right| < \epsilon\right\} > 1 - \delta.$$

Setting $\delta = \epsilon$, we obtain the second part of the theorem. The identification of $\delta = \epsilon$ will conveniently simplify notation later. To prove property (3), we write

$$1 = \sum_{x \in \mathcal{X}^n} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}|,$$

where the second inequality follows from the definition of the typical set. Hence, we have

$$|A_\epsilon^{(n)}| \leq 2^{-n(H(X)+\epsilon)}.$$

Finally, for sufficiently large n, $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$, so that

$$1 - \epsilon < \Pr\{A_\epsilon^{(n)}\} \leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|,$$

where the second inequality follows from the definition of the typical set. Hence,

$$A_\epsilon^{(n)} \geq (1 - \epsilon)\, 2^{n(H(X)-\epsilon)},$$

which completes the proof of the properties of $A_\epsilon^{(n)}$. □

4 随机过程的熵率

一个随机过程的**熵率**或信源信息率是在一个随机过程的平均信息的时间密度。我们前文已经介绍Markov链， 我们将进一步地介绍一些随机过程的重要概念。

4.1 马尔科夫链

定义 4.1. (Stationary) *A stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index; that is,*

$$\Pr\{X_1 = x_1, \ldots, X_n = x_n\} = \Pr\{X_{l+1} = x_1, \ldots, X_{n+l} = x_n\}$$

for every n and every shift l and for all $x_1, \ldots, x_n \in \mathcal{X}$.

定义 4.2. (time invariant) *The Markov chain is said to be time invariant if the conditional probability* $p(x_{n+1}|x_n)$ *does not depend on n; that is, for n =1, 2, . . . ,*

$$\Pr\{X_{n+1}=b|X_n=a\}=\Pr\{X_2=b|X_1=a\} \qquad \textit{for all } a,b\in\mathcal{X}.$$

一个时不变Markov链由其初始状态和概率转移矩阵$P=[P_{ij}]$, $i,j\in\{1,2,\ldots,m\}$所确定，其中$P_{ij}=\Pr\{X_{n+1}=j|X_n=i\}$。

如果有可能以有限步长从马尔可夫链的任何状态到任何其他状态以正概率前进， 则认为马尔可夫链是**不可约的**（irreducible）。 如果从一个状态到其自身的不同路径的长度的最大公因数是1， 则马尔可夫链被称为**非周期的**（aperiodic）。

对于不可约非周期的马尔科夫链， 注意到

$$p(x_{n+1})=\sum_{x_n\in\mathcal{X}} p(x_n)\,p(x_{n+1}|x_n)=\sum_{x_n} p(x_n)\,P_{x_nx_{x+1}}.$$

于是我们有如下矩阵形式：

$$\overrightarrow{P_{n+1}}=\vec{P}_nP=\vec{P}_0P^n.$$

由此我们有唯一的平稳的分布。

4.2 熵率

如果我们有一个由 n个随机变量组成的序列， 自然要问的是： 序列的熵如何随 n 增长? 我们将熵率定义为该增长率。

定义 4.3. *The entropy of a stochastic process* $\{X_i\}$ *is defined by*

$$H(\mathcal{X})=\lim_{n\to\infty}\frac{1}{n}H(X_1,\ldots,X_n)$$

when the limit exists.

对于 i.i.d.， 我们有 $H(\mathcal{X})=H(X_1)$；对于打字机，我们有 $H(\mathcal{X})=\log m$。

另一种相关量为 (当极限存在时)

$$H'(\mathcal{X})=\lim_{n\to\infty}H(X_n|X_{n-1},\ldots,X_1)$$

上述是两种形式的熵率的定义。对于平稳的随机过程， 我们有

定理 4.4. *For a stationary stocastic process, the limits in the definitions of entropy rates exist and are equal:*

$$H(\mathcal{X})=H'(\mathcal{X}).$$

为了证明上述定理， 我们先证明$\lim H(X_n|X_1^{n-1})$存在且是随n非递增的。事实上， 我们有

$$H(X_{n+1}|X_1^n)\le H(X_{n+1}|X_2^n)=H(X_n|X_1^{n-1}),$$

也就是说，$H(X_n|X_1^{n-1})$是一个非负的递减序列， 因此存在极限， 即 $H'(\mathcal{X})$.

进一步地， 我们需要如下定理：

定理 4.5. (Cesaro mean) *If* $a_n \to a$ *and* $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, *then* $b_n \to a$.

由于序列$\{a_k\}$中的大多数项最终都趋近于a，因此前n个项的平均值b_n最终也趋近于a。详细的证明可以参见 p76 [1]. 接下来让我们证明定理 4.4.

证明. (定理 4.4) By the chain rule,

$$\frac{H(X_1^n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1^{i-1}).$$

The right conditional entropies tend to a limit H′. Hence, by Theorem 4.2.3, their running average has a limit, which is equal to the limit H′ of the terms. Thus,

$$H(\mathcal{X}) = \lim \frac{H(X_1^n)}{n} = \lim H(X_n|X_1^{n-1}) = H'(\mathcal{X}). \qquad \square$$

因此对于平稳的马尔科夫链， 其熵率为

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_1^{n-1}) = \lim H(X_n|X_{n-1}) = H(X_2|X_1).$$

进一步地， 我们有

定理 4.6. *Let* $\{X_i\}$ *be a stationry Markov chain with stationary distribution* μ *and transition matrix* P. *Let* $X_1 \sim \mu$. *Then the entropy rate is*

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$

证明. $H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij}\right)$. $\square$

5 数据压缩

5.1 码

现在， 我们通过建立信息压缩的基本限制， 将内容放入熵的定义中。可以通过为数据源的最频繁结果分配简短说明， 为不频繁的结果分配较长的说明来实现数据压缩。例如，在摩尔斯电码中， 最频繁的符号由单个点表示。在本章中， 我们找到随机变量的最短平均描述长度。

熵是数据压缩极限以及随机数生成所需的位数， 并且从许多角度来看， 实现H的码都是最佳的。

码的概念类似于函数， 是指一套转换信息的规则系统， 例如将一个字母、 单词、 声音、 图像或手势转换为另一种形式或表达， 有时还会缩短或加密以便通过某种信道或存储媒体通信。编码是把信息源转化为便于通信或存储的符号， 而解码则是将其逆向还原的过程， 将代码符号转化回收件人可以理解的形式。

定义 5.1. *A source code C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite-length strings of symbols from a D-ary alphabet. Let C(x) denote the codeword corresponding to x and let l(x) denote the length of C(x).*

比如 $C(red) = 00, C(blue) = 11$ 是一个$\mathcal{X} = \{red, blue\}$的以$\mathcal{D} = \{0, 1\}$为字母表的源代码。

定义 5.2. *The expected length L(C) of a source code C(x) for a random variable X with probability mass function p(x) is given by*

$$L(C) = El(X) = \sum_{x \in \mathcal{X}} p(x)\, l(x),$$

where l(x) is the length of the codeword associated with x.

通常我们将 D-ary 的字母表定义为$\mathcal{D} = \{0, 1, \ldots, D-1\}$。

一般地， 我们要求源码是唯一可解码的。 一个码称之为**唯一可解码的** （uniquely decodable）当且仅当它的扩展是非奇异的。

定义 5.3. *A code is said to be nonsingular if every element of the range of X maps into a different string in $\mathcal{D}^*$; that is,*

$$x \neq x' \Longrightarrow C(x) \neq C(x').$$

定义 5.4. *The extension C^* of a code C is the mapping from finite-length strings of $\mathcal{X}$ to finite-length strings of $\mathcal{D}$, defined by*

$$C(x_1 x_2 \cdots, x_n) = C(x_1)\, C(x_2) \cdots C(x_n),$$

where the right hand side indicates concatenation of the corresponding codewords.

前缀码（或瞬时码）是指任何两个编码互不是对方的前缀。

定义 5.5. *A code is called a prefix code or an instantaneous code if no codeword is a prefix of any other codewords.*

一个前缀码也是自标点码（self-punctuating code）， 也就是说不用引入标点就可以唯一解码。因此， 对于不同的源码， 我们有如下的关系：

All codes > Nonsingular codes > Uniquely decodable codes > Instantaneous codes

Table 5.1 （p107 in [1]）给出了一个很好的例子来说明它们之间的关系。

5.2 Kraft 不等式

我们希望构造一个最小期望长度的瞬时代码来描述一个给定的源， 很明显， 我们不能为所有源符号分配短码字， 并且仍然没有前缀。瞬时代码可能使用的代码字长度集受以下不等式的限制。

定理 5.6. (Kraft inequality) *For any instantaneous code (prefix code) over an alphabet of size D, the codeword lengths* $l_1, \ldots, l_m$ *must satisfy the inequality*

$$\sum_i D^{-l_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

Proof omitted. Consider a D-ary tree, cf. pp. 107-109 [1].

事实上， 一个可数的瞬时码也满足 Kraft 不等式。

定理 5.7. (Extended Kraft Inequality) *For any countably infinite set of codewords that form a prefix code, the codeword lengths satisfy the extended Kraft inequality,*

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1.$$

Conversely, given any $l_1, l_2, \ldots$ *satisfying the extended Kraft inequality, we can construct a prefix code with these codeword lengths.*

5.3 最优码

在上节中， 我们证明了任何前缀码的字集都必须满足Kraft不等式， 并且Kraft不等式是存在具有指定码字长度集的源代码的充分条件。现在我们考虑寻找具有最小期望长度的前缀代码的问题。

这是一个典型的最优化问题：

$$\text{Minimize } L = \sum p_i l_i \text{ over integers } l_1, \ldots, l_m \text{ satisfying } \sum D^{-l_i} \leq 1.$$

在微积分中我们的方法是使用拉格朗日乘数法。令

$$J = \sum p_i l_i + \lambda \sum D^{-l_i}.$$

两边对l_i求偏导可得

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \ln D.$$

令这一偏导为0， 可得

$$D^{-l_i} = \frac{p_i}{\lambda \ln D}.$$

代入这一结果到限制（$\sum D^{-l_i} \leq 1$取等号）， 则有

$$\frac{\sum p_i}{\lambda \ln D} = 1 \Longrightarrow \lambda = \frac{1}{\ln D}.$$

于是， 我们有

$$p_i = D^{-l_i} \Longrightarrow l_i = -\log_D p_i.$$

这一（可能）非整数的最优解满足

$$L = \sum p_i l_i = -\sum p_i \log_D p_i = H_D(X).$$

事实上，因为 l_i 必须为整数，我们需要选择最接近最优解的整数。我们有如下定理：

定理 5.8. *The expected length L of any instantaneous D-ary code for a random variable X is greater than or equal to the entropy $H_D(X)$; that is,*

$$L \geq H_D(X),$$

with equality iff $D^{-l_i} = p_i$.

证明. We can write the difference between the expected length and the entropy as

$$L - H_D(X) = \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} = -\sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i.$$

Letting $r_i = D^{-l_i} / \sum_j D^{-l_j}$ and $c = \sum D^{-l_i}$, we obtain

$$L - H = \sum p_i \log_D \frac{p_i}{r_i} - \log_D c = D(p||r) + \log_D \frac{1}{c} \geq 0.$$

by the nonnegativity of relative entropy and the fact (Kraft inequality) that c ≤ 1. Hence, L ≥ H with equality if and only if $p_i = D^{-l_i}$ (i.e., if and only if $-\log_D p_i$ is an integer for all i). □

定义 5.9. *A probability distribution is called D-adic if each of the probabilities is equal to D^{-n} for some n. Thus, we have equality in the theorem if and only if the distribution of X is D-adic.*

更进一步地， 我们还能给出最优代码的最大期望长度。

定理 5.10. *Let $l_1^*, \ldots, l_m^*$ be optimal codeword lengths for a source distribution $\mathbb{p}$ and a D-ary alphabet, and let L^* be the associated expected length of an optimal code ($L^* = \sum p_i l_i^*$). Then*

$$H_D(X) \leq L^* < H_D(X) + 1.$$

Proof omitted. Let $l_i = \lceil \log_D \frac{1}{p_i} \rceil$, cf. pp. 112-113 [1].

对于一个随机变量系统（或者随机过程）， 我们有

定理 5.11. *The minimum expected codeword length per symbol satisfies*

$$\frac{H(X_1^n)}{n} \leq L_n^* < \frac{H(X_1^n)}{n} = \frac{1}{n}.$$

Moreover, if $X_1, X_2, \ldots, X_n$ is a stationary stochastic process,

$$L_n^* \to H(\mathcal{X}),$$

where H is the entropy rate of the process.

对于非瞬时码， 如果源码是唯一可解码的， 则也满足 Kraft 不等式。

定理 5.12. (McMillan) *The codeword lengths of any uniquely decodable D-ary code must satisfy the Kraft inequality. Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.*

证明参见 pp. 116-117 [1].

当然对应的可数的唯一可解码的源码也满足 Kraft 不等式， 在这里不再赘述， 参见 Corollary in p. 117[1].

6 算术编码与哈夫曼编码

算术编码是一种无损数据压缩方法， 也是一种熵编码的方法。和其它熵编码方法不同的地方在于， 其他的熵编码方法通常是把输入的消息分割为符号， 然后对每个符号进行编码， 而算术编码是直接把整个输入的消息编码为一个数， 一个满足（$0.0 \leq n < 1.0$）的小数n。

在给定符号集和符号概率的情况下， 算术编码可以给出接近最优的编码结果。使用算术编码的压缩算法通常先要对输入符号的概率进行估计， 然后再编码。这个估计越准， 编码结果就越接近最优的结果。

6.1 与哈夫曼编码关系

算术编码和哈夫曼编码的相似程度很高——事实上， 哈夫曼编码只是算术编码的一个特例。但是， 算术编码将整个消息翻译成一个表示为基数 b， 而不是将消息中的每个符号翻译成一系列的以 b 为基数的数字， 因此通常比哈夫曼编码更能达到最优熵编码。

因为算术编码不能一次压缩一个数据， 所以在压缩iid字符串时它可以任意接近熵。 相反， 使用霍夫曼编码（到字符串）的扩展不会达到熵， 除非字母符号的所有概率都是 2 的幂， 在这种情况下， 霍夫曼和算术编码都实现熵。

当霍夫曼编码二进制字符串时， 即使熵低（例如（{0, 1}）具有概率{0.95, 0.05}）， 也不可能进行压缩。霍夫曼编码为每个值分配 1 比特， 产生与输入长度相同的代码。相比之下， 算术编码可以更加地压缩比特， 接近最佳压缩比。

> 简而言之， 算术编码是一种整体编码， 不需要字母表； 而哈夫曼编码则需要一个字母表， 因此仍然会产生更多冗余的信息。

6.2 哈夫曼编码的最优化

引理 6.1. *For any distribution, there exists an optimal instantaneous code (with minimum expected length) that satisfies the following properties:*

1. *The lengths are ordered inversely with the probabilities (i.e., if $p_j > p_k$, then $l_j \leq l_k$).*
2. *The two longest codewords have the same length.*
3. *Two of the longest codewords differ only in the last bit and correspond to the two least likely symbols.*

Proof omitted. cf. pp. 123-125 [1].

定理 6.2. *Huffman coding is optimal; that is, if C^* is a Huffman code and $C\prime$ is any other uniquely decodable code, $L(C^*) \leq L(C')$.*

香农构建出了一个唯一可解码的方法， 称之为 **Shannon-Fano-Elias Coding**， 参见课本[1]章节5.9。我们给出如下的例子就能明了的看出这一构造。

x	$p(x)$	$F(x)$	$\bar{F}(x)$	$\bar{F}(x)$ in Binary	$l(x) = \left\lceil \log\frac{1}{p(x)} \right\rceil + 1$	Codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.5	0.75	0.5	0.10	2	10
3	0.125	0.875	0.8125	0.1101	4	1101
4	0.125	1.0	0.9375	0.1111	4	1111

6.3 LZW 编码

蓝波-立夫-卫曲编码法（Lempel-Ziv-Welch， 缩写LZW）， 是亚伯拉罕·蓝波、杰可布·立夫与泰瑞·卫曲共同提出的一种无损数据压缩算法。

它在1984年由泰瑞·卫曲改良亚伯拉罕·蓝与杰可布·立夫在1978年发表的LZ78的版本而来（主要是基于蓝波、 立夫的压缩概念， 设计出一套具有可逆推的逻辑程序）。

与霍夫曼编码相比， 蓝波-立夫-卫曲编码法被视作将不同长度字符串以固定长的码编辑（霍夫曼编码将固定长度字符用不同长度的码编辑）。 其优点在于此方法只需**存储一个相当小的表格**， 即可存储数据还原时相对应的值， 所以所需成本相对地低； 然而，这种算法的设计着重在实现的速度， 由于它并没有对数据做任何分析， 所以并不一定是最好的算法（参考LZMA，LZ77）。

方法的主要关键是， 它会在将要压缩的文本中， 自动地创建一个先前见过字符串的字典。这些字典并不需要与这些压缩的文本一起被传输， 因为如果正确地编码， 解压缩器也能够依照压缩器一样的方法把它建出来， 将会有完全与压缩器字典在文本的同一点有同样之字符串。

其中LZ77是移动窗口编码(12.5.1), 而 LZ78就是我们熟知的字典编码(12.5.2)。我们集中讨论后者。因为我们的目的是使用字典去除长字符串的重复， 所以我们需要一个独特解析（distinct parsing）。

定义 6.3. *A parsing S of a binary string $x_1 x_2 \cdots x_n$ is a division of the string into phrases, separated by commas. A distinct parsing is a parsing such that no two phrases are identical. For example, 0,111,1 is a distinct parsing of 01111, but 0,11,11 is a parsing that is not distinct. Let $c(n)$ be the number of phrases in a distinct parsing of a sample of length n.*

定理 6.4. *Let $\{X_n\}$ be a binary stationary ergodic process with entropy rate $H(X)$, and let $c(n)$ be the number of phrases in a distinct parsing of a sample of length n from this process. Then*

$$\limsup_{n\to\infty} \frac{c(n)\log c(n)}{n} \le H(\mathcal{X})$$

with probability 1.

定理 6.5. *Let $\{X_i\}_{-\infty}^{\infty}$ be binary stationary ergodic stochastic process. Let $l(X_1,\ldots,X_n)$ be the LZ78 codeword length associated with $X_1,\ldots,X_n$. Then*

$$\limsup_{n\to\infty} \frac{1}{n} l(X_1,\ldots,X_n) \le H(\mathcal{X}) \quad \textit{with probability 1,}$$

where H is the entropy rate of the process.

7 柯氏复杂度

一个对象比如一段文字的**柯氏复杂度 (Kolmogorov complexity)** 是衡量描述这个对象所需要的信息量的一个尺度。

以下面的两个长度为64的字符串为例。

```
0101010101010101010101010101010101010101010101010101010101010101
1100100001100001110111101110110011111010010000100101011110010110
```

第一个字符串可以用中文简短地描述为“重复32个‘01’”。第二个字符串没有明显的简短描述。

一个字符串 s 的柯氏复杂度是这个**字符串的最短描述的长度**。换言之， 一个字符串s的柯氏复杂度是能够输出且仅输出这个字符串的最短计算机/图灵机程序的长度。

与康托尔的对角论证法、 哥德尔不完备定理和图灵的停机问题类似， 柯氏复杂度的概念可以用于阐述和证明不可能性。

令 x 为一个有限长度的二进制字符串，$\mathcal{U}$ 是一个通用计算机。令 $l(x)$ 记为字符串 x 的长度，$\mathcal{U}(p)$ 记为给出程序p的计算机输出。我们定义柯氏复杂度为最小描述长度。

定义 7.1. *The Kolmogorov complexity $K_{\mathcal{U}}(x)$ of a string x with respect to a universal computer $\mathcal{U}$ is defined as*

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p),$$

the minimum length over all programs that print x and halt.

上述定义没有提到任何关于 x 长度的信息， 如果我们已知 x 的长度， 我们可以定义条件柯氏复杂度

$$K_{\mathcal{U}}(x|l(x)) = \min_{p:\mathcal{U}(p,l(x))=x} l(p),$$

这是最短的描述长度。

定理 7.2. (Universality of Kolmogorov complexity) *If $\mathcal{U}$ is a universal computer, for any other computer $\mathcal{A}$ there exists a constant $c_{\mathcal{A}}$ such that*

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$$

for all strings $x \in \{0,1\}^*$, and the constant $c_{\mathcal{A}}$ does not depend on x.

证明. Assume that we have a program p_A for computer A to print x. Thus, $A(p_A) = x$. We can precede this program by a simulation program s_A which tells computer U how to simulate computer A. Computer U will then interpret the instructions in the program for A, perform the corresponding calculations and print out x. The program for U is $p = s_A\, p_A$ and its length is

$$l(p) = l(s_{\mathcal{A}}) + l(p_{\mathcal{A}}) = c_{\mathcal{A}} + l(p_{\mathcal{A}}),$$

where $c_{\mathcal{A}}$ is the length of the simulation program. Hence,

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p) \leq \min_{p:\mathcal{A}(p)=x} (l(p) + c_{\mathcal{A}}) = K_{\mathcal{A}}(x) + c_{\mathcal{A}}$$

for all string x. □

定理 7.3. (Conditional complexity is less than the length of the sequence)

$$K(x|l(x)) \leq l(x) + c.$$

证明. A program for printing x is

```
Print the following l-bit sequence: x1 x2 . . . xl(x).
```

Note that no bits are required to describe l since l is given. The program is self-delimiting because $l(x)$ is provided and the end of the program is thus clearly defined. The length of this program is $l(x)+c$. □

定理 7.4. (Upper bound on Kolmogorov complexity)

$$K(x) \leq K(x|l(x)) + 2\log l(x) + c.$$

证明. If the computer does not know $l(x)$, the method of Theorem 14.2.2 does not apply. We must have some way of informing the computer when it has come to the end of the string of bits that describes the sequence. We describe a simple but inefficient method that uses a sequence 01 as a "comma."

Suppose that $l(x)=n$. To describe $l(x)$, repeat every bit of the binary expansion of n twice; then end the description with a 01 so that the computer knows that it has come to the end of the description of n.

For example, the number 5 (binary 101) will be described as 11001101. This description requires $2\lceil \log n \rceil + 2$ bits. Thus, inclusion of the binary representation of l(x) does not add more than $2\log l(x) + c$ bits to the length of the program, and we have the bound in the theorem. □

定理 7.5. (Lower bound on Kolmogorov complexity) *The number of strings x with complexity $K(x)<k$ satisfies*

$$|\{x \in \{0,1\}^*: K(x) < k\}| < 2^k.$$

证明. There are not very many short programs. cf. p469 [1]. □

进一步地， 对于二进制字符串， 我们有

定理 7.6. *The Kolmogorov complexity of a binary string x is bounded by*

$$K\left(x_1 x_2 \cdots x_n | n\right) \leq n\, H_0\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{2}\log n + c.$$

证明. cf. Example 14.2.8 [1]. □

7.1 柯氏复杂度和熵

现在， 我们考虑一系列随机变量的Kolmogorov复杂度与其熵之间的关系。 总的来说，我们表明随机序列的Kolmogorov复杂度的期望值接近于Shannon熵。首先， 程序的长度满足 Kraft 不等式。

引理 7.7. *For any computer $\mathcal{U}$,*

$$\sum_{p:\mathcal{U}(p)halts} 2^{-l(p)} \leq 1.$$

证明. If the computer halts on any program, it does not look any further for input. Hence, there cannot be any other halting program with this program as a prefix. Thus, the halting programs form a prefix-free set, and their lengths satisfy the Kraft inequality. □

定理 7.8. (Relationship of Kolmogorov complexity and entropy) *Let the stochastic process $\{X_i\}$ be drawn i.i.d. according to the probability mass function $f(x)$, $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite alphabet. Let $f(x^n) = \prod_{i=1}^{n} f(x_i)$. Then there exists a constant c such that*

$$H(X) \leq \frac{1}{n} \sum_{x^n} f(x^n)\, K(x^n|n) \leq H(X) + \frac{(|\mathcal{X}|-1)\log n}{n} + \frac{c}{n}$$

for all n. Consequently,

$$E\frac{1}{n} K(X^n|n) \to H(X).$$

证明. Proof is omitted, cf. pp. 473-474 [1]. □

去除掉序列长度的条件， 根据同样的论证， 我们可以得到

$$H(X) \leq \frac{1}{n} \sum_{x^n} f(x^n)\, K(x^n) \leq H(X) + \frac{(|\mathcal{X}|+1)\log n}{n} + \frac{c}{n}.$$

下界是因为 $K(X^n)$ 是前缀码，上界是因为 $K(x^n) \leq K(x^n|n) + 2\log n + c$. 于是，

$$E\left[\frac{1}{n} K(x^n)\right] \to H(X).$$

7.2 整数的柯氏复杂度

我们可以定义一个任意整数的柯氏复杂度：

定义 7.9. *The Kolmogorov complexity of an integer n is defined as*

$$K(n) = \min_{p:\mathcal{U}(p)=n} l(p).$$

整数的柯氏复杂度的性质与位串的柯氏复杂度的属性非常相似。以下属性是字符串相应性质的直接结果。

定理 7.10. *For universal computers A and U,*

$$K_{\mathcal{U}}(n) \leq K_{\mathcal{A}}(n) + c_{\mathcal{A}}.$$

因为任意数可以表达为二进制，所以进一步地我们有如下结论：

$$K(n) \leq \log n + c.$$

定理 7.11. *There are an infinite number of integers n such that $K(n) > \log n$.*

证明. We know from Lemma 14.3.1 that $\sum_n 2^{-K(n)} \leq 1$ and $\sum_n 2^{-\log n} = \sum_n \frac{1}{n} = \infty$. But if $K(n) < \log(n)$ for all $n > n_0$, then

$$\sum_{n=n_0}^{\infty} 2^{-K(n)} > \sum_{n=n_0}^{\infty} 2^{-\log n} = \infty,$$

which is a contradiction. □

7.3 柯氏复杂度的不可计算性

定理7.11表明存在字符串， 拥有任意大的柯氏复杂度。更进一步地我们有，

K 不是一个可计算函数。

也就是说， 不存在一个程序， 可以把字符串 s 作为输入， 然后输出它的 $K(s)$。

这一结论证明可以参见维基百科：柯氏复杂性。更多关于柯氏复杂度的介绍可以参见上述词条。

8 信道容量

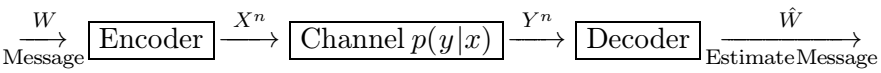
让我们先考虑一个问题： 因为任何传播的信道都会有噪音， 我们该如何避免噪音的存在?

一种很自然的编码方法是使用重复编码（repetition code）， 比如每个字符都重复奇数次， 比如3次。在这里我们先做一个计算， 如果传输错误的概率为 p=0.1。则一个长为n的字符串无错误传输的概率是 $p(\text{no error}) = 0.9\hat{}n$。而重复三次的无错误传输的概率是

$$p(\text{no error}) = (0.9^3 + 3 \cdot 0.9^2 \cdot 0.1)^n = 0.972^n.$$

当 n 很大时， 这一概率和未重复编码的概率类似， 因此并没有解决这个问题。为了解决信道编码的问题， 我们先引入如下一些概念。

我们定义一个离散信道（discrete channel）为一个包含输入字母表、 输出字母表和一个概率转移矩阵的系统。我们称一个离散信道是无记忆的（memoryless）， 当且仅当其输出只依赖于输入而与之前的信道输入输出条件独立。如下图所示， 这是一个典型的通信系统模型：



我们定义信息通道容量如下：

定义 8.1. *We define the "information" channel capacity of a discrete memoryless channel as*

$$C = \max_{p(x)} I(X;Y),$$

where the maximum is taken over all possible input distributions $p(x)$.

我们下面给出一些例子， 参见课本[1]的章节 7.1：

例 8.2.

1. $\begin{matrix} 0 & \to & 0 \\ 1 & \to & 1 \end{matrix}$, $C = 1$;

2. $\begin{matrix} 0 & & 1 \\ /\backslash & & /\backslash \\ 1\ 2 & & 3\ 4 \end{matrix}$ with $p_i = p = \frac{1}{2}$, Hence,

$$C = H(Y) - H(Y|X) = 2 - 1 = H(X) - H(X|Y) = 1 - 0.$$

3. Noisy typewriter. $C = \max_{p(x)} H(Y) - 1 = \log 26 - 1 = \log 13$.

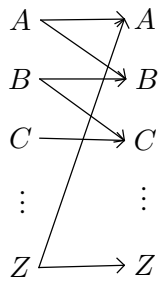


图 8.1. Noisy typewriter

4. Binary symmetric channel.

$$C = \max I(X,Y) = 1 - H(Y|X) = 1 - H(p, 1-p) = 1 - H(p).$$

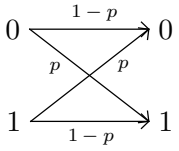


图 8.2. Binary symmetric channel

5. Binary asymmetric channel.

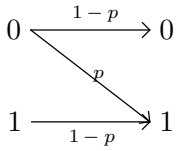


图 8.3. Binary asymmetric channel

$C = I(X;Y) = H(X) - H(Y|X)$. Let $p(X=0) = \alpha$, then $p(x=1) = 1-\alpha$. Thus,

$$I = -\alpha(1-p)\log(\alpha(1-p)) - (\alpha p + 1 - \alpha)\log(\alpha p + 1 - \alpha) - \alpha H(p).$$

Compute $\frac{dI}{d\alpha} = 0$, then find α.

6. Binary erasure channel.

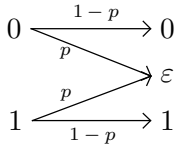


图 8.4. Binary erasure channel

$$H(Y) = -\frac{1}{2}(1-p)\log\left(\frac{1}{2}(1-p)\right) - \frac{1}{2}(1-p)\log\left(\frac{1}{2}(1-p)\right) - \frac{1}{2}\cdot 2p\log\left(\frac{1}{2}\cdot 2p\right)$$

$$H(Y|X) = 2\cdot\frac{1}{2}H(p;1-p)$$

$$I = 1 - p.$$

8.1 信道容量的性质

对于信道容量， 我们有如下的性质：

1. $C \geq 0$ since $I(X;Y) \geq 0$.
2. $C \leq \log|X|$ since $C = \max I(X;Y) \leq \max H(X) = \log|X|$.
3. $C \leq \log|Y|$ for the same reason.
4. $I(X;Y)$ is a continuous function of $p(x)$.
5. $I(X;Y)$ is a concave function of $p(x)$.

8.2 信道编码定理

有噪信道编码定理指出， 尽管噪声会干扰通信信道， 但还是有可能在**信息传输速率小于信道容量**的前提下， 以任意低的错误概率传送数据信息。这个令人惊讶的结果， 有时候被称为**信息原理基本定理**， 也叫做香农-哈特利定理或**香农定理**。

正如本节最开头定义的那样（cf p193[1]）：

定义 8.3. *A discrete Channel is* $(\mathcal{X}, p(y|x), \mathcal{Y})$ *with* $\sum_y p(y|x) = 1$. *The nth extension of the discrete memoryless channel (**DMC**) is the channel* $(X^n, p(y^n|x^n), \mathcal{Y}^n)$, *where*

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \quad k = 1, 2, \ldots, n.$$

如果信道是没有反馈的（without feedback）， 即$p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$, 则信道的n次扩展转移函数满足

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i).$$

为了论证香农定理， 我们引入如下的具体构造：

定义 8.4. *An (M, n) code for the channel (X, p(y|x), Y) consists of the following:*

1. An index set 1, 2, . . . , M.

2. An encoding function $X^n: \{1, \ldots, M\} \to \mathcal{X}^n$, *yielding codewords* $x^n(1), \ldots, x^n(M)$. *The set of codewords is called the codebook.*

3. A decoding function $g: \mathcal{Y}^n \to \{1, 2, \ldots, M\}$, *which is a deterministic rule that assigns a guess to each possible received vector.*

如果使用g解码(M, n)的字母表得到的结果和源码不同， 我们称之为误差的条件概率。我们也会关注其中的最大误差概率和平均误差概率。

定义 8.5. *Let*

$$\lambda_i = \Pr\{g(Y^n) \neq i | X^n = x^n(i)\} = \sum_{y^n} p(y^n|x^n(i))\, I\left(g(y^n) \neq i\right)$$

be the conditional probability of error given that index i was sent, where I is the indicator function. The maximal probability of error λ^n *for an (M, n) code is defined as*

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i.$$

The (arithmetic) average probability of error $P_e^{(n)}$ *for that code is defined as*

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.$$

我们定义 (M, n) 编码的速率（rate）如下：

定义 8.6. *The rate R of an (M, n) code is $R = \frac{\log M}{n}$ bits per transmission.*

一个速率是可达到的（achievable）当且仅当存在一个序列的编码 $(\lceil 2^{nR} \rceil, n)$， 使得最大误差概率 $\lambda^{(n)} \to 0$ 当 $n \to \infty$. 也就是说，这个速率大小对应的信道编码可以将误差控制到任意小。通常，为了简单书写，我们常常用 $(2^{nR}, n)$来代替 $(\lceil 2^{nR} \rceil, n)$。

由此我们有另一个信道容量的定义， 即信道容量是可达到的速率的极大值。

8.2.1 联合典型序列

类似于典型集合， 我们可以定义一个信息系统的联合典型序列（jointly typical sequences）。

定义 8.7. *The set $A_\epsilon^{(n)}$ of ==jointly typical sequences== $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:*

$$\begin{aligned} A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n: \\ \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon\}, \end{aligned}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$.

于是我们有联合的AEP定理。

定理 8.8. (Joint AEP) *Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then*

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \to 1$ *as* $n \to \infty$.

2. $|A_\epsilon^{(n)}| \le 2^{n(H(X,Y)+\epsilon)}$.

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)\, p(y^n)$ (i.e. $\tilde{X}^n, \tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$), then

$$\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \le 2^{-n(I(X;Y)-3\epsilon)}.$$

Also, for sufficiently large n,

$$\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \ge (1-\epsilon)\, 2^{-n(I(X;Y)+3\epsilon)}.$$

证明. Proof. cf. pp. 196-198 [1]. □

8.2.2 信道编码定理

定理 8.9. (Channel coding theorem) *For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda(n) \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.*

证明. cf. [1] pp. 200-204 for achievability. cf. pp. 206-208 for converse to the coding theorem. □

8.3 网络编码

网络编码(Network Coding)是一种通过中继节点对接收到的信息进行编码来达到提高多播网络容量的技术。Rudolf Ahlswede, Ning Cai, Shuo-Yen Robert Li, Raymond W. Yeung在2000年首次提出网络编码的概念。

在传统的数据传输技术中， 中继节点只负责数据的存储转发， 而基于网络编码技术的网络的中继节点在具备传统中继功能的基础上， 会根据网络编码规则将接收到的信息进行线性或非线性处理再进行传播， 这种做法最直观的优势是减少了传输次数。利用图论中**最大流最小割原理** (从源点到目标点的最大的流量等于最小割的每一条边的和) 论证了网络编码可以达到网络最大信息流。

8.3.1 蝶形网络

一个经典的问题是蝶形网络（Butterfly Network）， 如下图所示。它通常用于说明线性网络编码如何胜过路由。两个源节点（在图片的顶部）具有信息A和B， 这些信息必须传输到两个目的地节点（在底部）， 它们每个都想知道A和B。每边只能携带一个值（边的传输效率是1b/s）。

如果仅允许路由， 则中央链路将只能承载A或B， 但不能同时承载两者。假设我们通过中心发送A； 那么离开目的地将收到两次A， 而根本不知道B。对于正确的目的地，发送B也会带来类似的问题。所以我们说路由是不够的， 因为没有路由方案可以同时将A和B传输到两个目的地。也就是说我们至少需要两秒（两次传输才能达到目的）。

如图所示， 使用一个简单的代码， 即可通过中心发送符号之和将A和B同时发送到两个目的地。换句话说， 我们使用公式“ A + B”对A和B进行编码。左目的地接收A和A + B， 并且可以通过减去两个值来计算B。同样， 正确的目的地将收到B和A + B， 并且也将能够确定A和B。

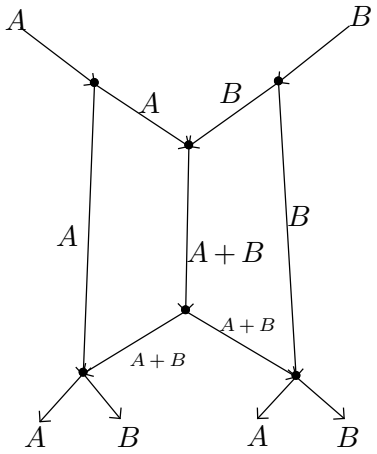


图 8.5. Butterfly Network

8.3.2 线性网络编码

更进一步地， 我们可以推广这一方案。假设网络是有向的， 执行**线性网络编码**时每个节点收到所有连入线路的数据后， 再执行编码， 然后把数据从连出线路发出。新的数据包括执行线性编码所用的系数以及合成后的数据。

例如组播源发送三条封包， $p_1=1, p_2=2, p_3=3$。封包经过一系列的中间节点， 目标节点收到的封包是(5,8,1),24),((2,3,7),29),((9,6,5),36)。目标节点对下列矩阵求解， 可得p_1, p_2, p_3的值。

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 5 & 8 & 1 \\ 2 & 3 & 7 \\ 9 & 6 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 24 \\ 29 \\ 36 \end{bmatrix}.$$

8.3.3 随机线性网络编码

随机线性网络编码可以取得更好的组播传输速率， 较为实用。在实际网络中， 节点会将来自连入线路的封包缓存起来， 当节点需要发送封包时再将缓存的封包执行网络编码，然后发出。

9 微分熵

微分熵是消息理论中的一个概念， 是从以离散随机变数所计算出的香农熵推广， 以连续型随机变数计算所得之熵， 微分熵与离散随机变数所计算出之香农熵， 皆可代表描述一信息所需码长的下界， 然而， 微分熵与香农熵仍存在着某些相异的性质。

定义 9.1. *The differential entropy h(X) of a continuous random variable X with density f (x) is defined as*

$$h(X) = -\int_S f(x)\log f(x)\,dx,$$

where S is the support set of the random variable.

以下是一些例子:

1. 均匀分布 $X\sim U[0,a]$: $h(X)=-\int_0^a \frac{1}{a}\log\frac{1}{a}\,dx=\log a$. 注意，如果 0<a<1，则对应的微分熵是负的。

2. 标准正态分布 $X\sim N(0,\sigma^2)$: $h(X)=\frac{1}{2}\log(2\,\pi\,e\,\sigma^2)$.

注记 9.2. 例子2中是**给定方差最大的微分熵**。期望值为 0，方差为 σ^2且值域为R之随机变数 X的微分熵，其上界为正态分布$N(0,\sigma^2)$ 的微分熵。(cf. Theorem 8.6.5 in [1])

容易验证，我们有如下性质（Theorem 8.6.6 & 8.6.7 in [1]）：

1. $h(X+c)=h(X)$ Translation does not change the differential entropy.

2. $h(\mathrm{aX})=h(X)+\log|a|$. （使用变量替换容易证明）

对于估计误差， 我们有

定理 9.3. (Estimation error and differential entropy) *For any random variable X and estimator $\hat{X}$,*

$$E\,(X-\hat{X})^2 \geq \frac{1}{2\,\pi\,e}\,e^{2h(X)},$$

with equality if and only if X is Gaussian and $\hat{X}$ is the mean of X.

证明. $E\,(X-\hat{X})^2 \geq \min_{\hat{X}} E\,(X-\hat{X})^2 = E\,(X-E(X))^2 = \mathrm{var}(X) \geq \frac{1}{2\,\pi\,e}\,e^{2h(X)}.$ □

9.1 与离散熵的关系

我们将连续变量分割为宽度为Δ的量化。则有 $X^{\Delta}=x_i$ 对应的概率为 $p_i=\int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(x_i)\,\Delta$. 于是我们有:

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{-\infty}^{+\infty} p_i \log p_i = -\sum f(x_i)\,\Delta \log\left(f(x_i)\,\Delta\right) \\ &= -\sum \Delta\, f(x_i) \log f(x_i) - \log \Delta \end{aligned}$$

如果 $f(x)\log f(x)$ 是黎曼可积的，则上述离散熵存在极限，也就是说我们有如下定理:

定理 9.4. *If the density $f\,(x)$ of the random variable X is Riemann integrable, then*

$$H(X^{\Delta}) + \log\Delta \to h(f) = h(X),\ as\ \Delta \to 0.$$

Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X)+n$.

9.2 连续变量的 AEP

首先同离散情况一样， 经过连续变量的大数定律我们可以得到:

定义 9.5. *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f\,(x)$. Then*

$$-\frac{1}{n}\log f(X_1, X_2, \ldots, X_n) \to E\left[-\log f(X)\right] = h(X) in\ prob.$$

进一步地， 我们需要定义连续变量的典型集（Typical set）。

定义 9.6. *For $\epsilon > 0$ and any n, we define the typical set $A_{\epsilon}^{(n)}$ with respect to $f(x)$ as follows:*

$$A_{\epsilon}^{(n)} = \left\{(x_1, \ldots, x_n) \in S^n : \left| -\frac{1}{n}\log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\},$$

where $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.

我们需要定义一个体积的概念用以推广离散情况的集合元素个数。

定义 9.7. *The Volume Vol(A) of a set $A \subset R^n$ is defined as*

$$\mathrm{Vol}(A) = \int_A d\,x_1\, d\,x_2 \cdots d\,x_n.$$

于是我们类似于离散情况的进一步的估计：

定理 9.8. *The typical set $A_\epsilon^{(n)}$ has the following properties:*

1. $\Pr(A_\epsilon^{(n)}) > 1-\epsilon$ *for n sufficiently large.*

2. $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)}$ *for all n.*

3. $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon)\, 2^{n(h(X)-\epsilon)}$ *for n sufficiently large.*

证明. Proof. cf. p 246 [1]. □

这说明，比离散情况更加直观，$A_\epsilon^{(n)}$ 集中分布在一个非常薄的球壳上，球壳的半径大约是 $2^{nh(X)}$，并且发生的概率接近于1。

进一步地， 定理 8.2.3 [1] 表明典型集是拥有概率 $\ge 1-\epsilon$ 的最小体积的集合。

9.3 联合、 条件微分熵与相对熵和相互信息

联合微分熵的定义与联合离散熵类似

定义 9.9. *The differential entropy of a set $X_1, \ldots, X_n$ of random variables with density $f(x_1, \ldots, x_n)$ is defined as*

$$h(X_1^n) = -\int f(x^n) \log f(x^n)\, d\,x^n.$$

条件微分熵

定义 9.10. *If X, Y have a joint density function f(x,y), we can define the conditional differential entropy h(X|Y) as*

$$h(X\,|Y) = -\int f(x,y) \log f(x|y)\, d\,x\, d\,y = h(X\,,Y) - h(Y).$$

相互熵（Kullback–Leibler 距离）

定义 9.11. *The relative entropy $D\,(f\,||\,g)$ between two densities $f\,, g$ is defined by*

$$D\,(f\,||\,g) = \int f \log \frac{f}{g}.$$

注意， $D(f||g)$ 是有限的当且仅当f的支撑集包含在g的支撑集中。上述定义中我们要求 $0\log 0/0=0$.

相互信息

定义 9.12. *The mutual information I(X;Y) between two random variables with joint density f(x,y) is defined as*

$$I(X;Y)=\int f(x,y)\log\frac{f(x,y)}{f(x)\,f(y)}\,dx\,dy.$$

根据定义我们很容易得到如下等式：

$$\begin{aligned} I(X;Y) &= h(X)-h(X|Y)=h(Y)-h(Y|X)=h(X)+h(Y)-h(X,Y) \\ &= D(f(x,y)||f(x)\,f(y)) \end{aligned}$$

注意， 离散量化的相互信息和连续的相互信息是一致的。

$$I(X^\Delta;Y^\Delta)=H(X^\Delta)-H(X^\Delta|Y^\Delta)\approx h(X)-\log\Delta-h(X|Y)+\log\Delta=I(X;Y).$$

9.3.1 相对熵和相互信息的性质

定理 9.13. $D(f||g)\geq 0$ *with equality iff f=g almost everywhere (a.e.).*

证明. Let S be the support of f. Then

$$-D(f||g)=\int_S f\log\frac{g}{f}\leq\log\int_S f\frac{g}{f}=\log\int_s g\leq\log 1=0. \qquad \square$$

根据上述定理我们有如下推论：

1. $I(X;Y)\geq 0$ 等号当且仅当 X, Y 是相互独立的；
2. $h(X|Y)\leq h(X)$ 等号当且仅当 X, Y是相互独立的。

根据条件熵的定义我们可以得到微分熵的链式法则。

定理 9.14. (Chain rule for differential entropy)

$$h(X_1^n)=\sum_{i=1}^n h(X_i|X_1^{i-1}).$$

由此可以得到一个重要推论：

$$h(X_1^n)\leq\sum h(X_i).$$

等号当且仅当 $X_1, \ldots, X_n$ 相互独立。

这一定理的一个直接应用就是 Hadamard 不等式，参见[1], 253页。

10 高斯信道

加性高斯白噪声（Additive white Gaussian noise， AWGN）在通信领域中指的是一种功率谱函数是常数（即白噪声）， 且幅度服从高斯分布的噪声信号。因其可加性、 幅度服从高斯分布且为白噪声的一种而得名。

该噪声信号为一种便于分析的理想噪声信号， 实际的噪声信号往往只在某一频段内可以用高斯白噪声的特性来进行近似处理。由于AWGN信号易于分析、 近似， 因此在信号处理领域， 对信号处理系统（如滤波器、 低噪音高频放大器、 无线信号传输等）的噪声性能的简单分析(如： 信噪比分析)中， 一般可假设系统所产生的噪音或受到的噪音信号干扰在某频段或限制条件之下是高斯白噪声。

10.1 高斯信道的容量

AWGN 信道由一系列的 Y_i （输出） 来表示，其中的 i 表示离散的时间事件索引。Y_i 是 X_i (输入)和噪音 Z_i 的数值和。其中Z_i 是独立恒等分布(iid)的随机变量并来自于均值为 0，方差为 N (噪声) 的正态分布。

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N).$$

信道的容量通常是无穷的， 除非噪声N非零且X_i有足够的约束。输入中最常见的约束被叫做**功率约束**， 这要求码字通过信道传送。我们有：

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leqslant P,$$

其中 P 代表信道功率的最大值。因此信道容量的功率约束（也称为**功率约束的高斯信道的信道容量**）可以通过以下公式给出：

$$C = \max_{f(x):E(X^2)\leq P} I(X;Y)$$

这里的 f(x) 是 X 的分布， I(X;Y) 可以扩展为微分熵的形式。因为X, Z相互独立， 所以我们有

$$I(X;Y) = h(Y) - h\left(X+Z|X\right) = h(Y) - h(Z|X) = h(Y) - h(Z).$$

而 h(Z) 通过计算是 $h(Z) = \frac{1}{2}\log\left(2\,\pi\,e\,N\right)$.

另外 $E(Y^2) = E\left(X+Z\right)^2 = E(X^2) + 2\,E(X)\,E(Z) + E(Z^2) = P + N$, 从此约束中我们可以得到 （固定方差的微分熵最大为正态分布的微分熵）

$$h(Y) \leq \frac{1}{2}\log\left(2\,\pi\,e\,(P+N)\right).$$

由此可以得到信道容量的约束为

$$I(X;Y) \leq \frac{1}{2}\log\left(2\,\pi\,e\,(P+N)\right) - \frac{1}{2}\log\left(2\,\pi\,e\,N\right)$$

其中 I(X;Y) 再 X~N(0,P) 时最大。由此得到信道容量为

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right).$$

误差的衡量由下式给出：

$$P_e = \frac{1}{2}\Pr(Y<0|X=\sqrt{P}) + \frac{1}{2}\Pr(Y>0|X=-\sqrt{P}) = \Pr(Z>\sqrt{P}) = 1-\Phi(\sqrt{P/N}).$$

其中 Φ 是累积正态函数。

10.2 压缩感知

压缩感知（Compressed sensing），也被称为压缩采样（Compressive sampling）或稀疏采样（Sparse sampling），是一种寻找欠定线性系统的稀疏解的技术。压缩感知被应用于电子工程尤其是信号处理中，用于获取和重构稀疏或可压缩的信号。这个方法利用信号稀疏的特性，相较于奈奎斯特理论，得以**从较少的测量值还原出原来整个欲得知的信号**。核磁共振就是一个可能使用此方法的应用。这一方法至少已经存在了四十年，由于David Donoho、Emmanuel Candès和陶哲轩的工作，最近这个领域有了长足的发展。近几年，为了因应即将来临的第五代移动通信系统（5G），压缩感知技术也被大量应用在无线通信系统之中，获得了大量的关注以及研究。

更多参考：

1. 维基百科. 采样定理
2. 维基百科. 压缩感知
3. compressed sensing sub-Nyquist sampling (Terence Tao)
4. Compressive Sensing Resources