信息论知识整理

范路遥 2021 信息工程 fanluyao@sjtu.edu.cn

更新时间：2024 年 4 月 20 日

本知识整理旨在帮助学习信息论的基本思想，侧重直觉上的理解而非数学上的严格推导，不会包含很多习题和证明。适用于 ICE4313, ICE2601, NIS2337, 不同课程范围各不相同，按需取用即可。内容可能不完整，也可能有错误，欢迎邮件反馈。

目录

信息熵(Entropy)

离散熵(Discrete Entropy)

$$H(X)=-\sum_{x\in\mathcal{X}}p_X(x)\log p_X(x)=\mathbb{E}_X[-\log p_X(X)]$$

X 为随机变量，$\mathcal{X}=\{x_1,\cdots,x_n\}$ 为 X 的符号集。

两处 $p_X(x)$ 意义完全不同：$-\log p_X(x)$ 为单个符号的信息量，概率越小信息量越大，且两个独立变量 X 和 Y 联合分布的符号概率 $-\log[p_X(x)p_Y(y)]=-\log p_X(x)-\log p_Y(y)$ 满足可加性，所以只能是负对数而不是别的函数形式；$-\log p_X(x)$ 前的系数 $p_X(x)$ 为期望，信息熵即一个随机变量提供的平均信息量，所以不是其中一个符号概率越小信息量就越大，而要考虑整体的总信息量。总信息量达到最大时，各符号概率均等。

设 $\forall x\in\mathcal{X},u(x)=\frac{1}{|\mathcal{X}|}$，则 $D(p||u)=\log|\mathcal{X}|-H(X)\geq 0$.

微分熵(Differential Entropy)

$$h(X)=-\int_X p_X(x)\log p_X(x)\mathrm{d}x=\mathbb{E}_X[-\log p_X(X)]$$

微分熵的定义可类比离散熵，但存在区别。

如果算连续分布变量的离散熵，则每个X取值$\frac{i}{n}$处的概率为$\left[p_X\left(\frac{i}{n}\right)\frac{1}{n}\right]$，

$$\begin{aligned}H(X)&=-\lim_{n\to\infty}\sum_{i=1}^{n}\left[p_X\left(\frac{i}{n}\right)\frac{1}{n}\right]\log\left[p_X\left(\frac{i}{n}\right)\frac{1}{n}\right]\\&=-\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}p_X\left(\frac{i}{n}\right)\left[\log p_X\left(\frac{i}{n}\right)+\log\frac{1}{n}\right]\\&=-\int_{-\infty}^{+\infty}p_X(x)\log p_X(x)\mathrm{d}x-\lim_{n\to\infty}\log\frac{1}{n}\\&=h(X)-\lim_{n\to\infty}\log\frac{1}{n}=h(X)+\infty\end{aligned}$$

任何连续分布变量都有无穷多个取值，所以离散熵都是无穷大，微分熵是离散熵统一减去一个“无穷大”的结果。所以微分熵完全可以小于0，毕竟都已经减掉过一个“无穷大”了。

如果不限制方差（平均功率），则微分熵可以无限大；如果限制方差（平均功率），则微分熵在正态分布下最大；如果限制定义域（峰值功率），则微分熵在均匀分布下最大。

联合熵(Joint Entropy)

$$\begin{aligned}H(X,Y)&=-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p_{XY}(x,y)\log p_{XY}(x,y)\\&=\mathbb{E}_{XY}[-\log p_{XY}(X,Y)]\end{aligned}$$

把随机变量X和Y看成一个整体的信息熵。

条件熵(Conditional Entropy)

$$\begin{aligned}H(Y|X)&=\sum_{x\in\mathcal{X}}p_X(x)H(Y|X=x)\\&=\mathbb{E}_X[H(Y|X=x)]\\&=-\sum_{x\in\mathcal{X}}p_X(x)\sum_{y\in\mathcal{Y}}p_{Y|X}(y|x)\log p_{Y|X}(y|x)\\&=-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p_{XY}(x,y)\log p_{Y|X}(y|x)\\&=\mathbb{E}_{XY}[\log p_{Y|X}(Y|X)]\end{aligned}$$

在X给定条件下Y的熵，再对X求期望。即已知X后再得到Y时，新增的信息量。

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

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

链式法则：

$$\begin{aligned}H(X_1,X_2,\cdots,X_n)&=H(X_1,X_2,\cdots,X_{n-1})+H(X_n|X_1,X_2,\cdots,X_{n-1})\\&=\sum_{i=1}^{n}H(X_i|X_1,X_2,\cdots,X_{i-1})\end{aligned}$$

带限信号(Bandlimited Signal)的熵

完全随机的波形相当于无穷多个随机的点，其微分熵一定无穷大；但如果限制频率，则随机性也降低，可以用可数多个$\mathrm{sinc}t=\frac{\sin\pi t}{\pi t}$来插值表示，连续波形就转化为离散序列了。根据奈奎斯特(Nyquist)采样定理，若波形$x(t)$带宽[1]为W，则其奈奎斯特频率$f_\mathrm{s}=2W$，以该频率采样，即$X_k=x\left(\frac{k}{f_\mathrm{s}}\right)$，离散序列$\boldsymbol{X}=\{\cdots,X_{-1},X_0,X_1,\cdots\}$即可完全还原本身的连续信号。

$$h(\boldsymbol{X})=h(\cdots,X_{-1},X_0,X_1,\cdots)$$

如果随手画一条曲线作为波形，那这种波形的频率几乎一定是无限的。但频率无限、几乎处处连续的波形可以用频率有限的波形逼近。

如果时间有限，则$\boldsymbol{X}$维数有限；如果方差（功率）也有限，则微分熵也就有限了。

$$\begin{aligned}h(\boldsymbol{X})&=h(X_1,X_2,\cdots,X_n)\\&=h(X_1)+h(X_2|X_1)+\cdots+h(X_n|X_1,X_2,\cdots,X_{n-1})\\&\leqslant h(X_1)+h(X_2)+\cdots+h(X_n)\end{aligned}$$

但傅里叶变换的性质表明，时间有限的信号频率一定无限[2]，所以这里也只能近似看作频率有限。

相对熵(Relative Entropy, Kullback–Leibler Divergence)

$$\begin{aligned}D(p||q)&=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}=\mathbb{E}_P\left[\log\frac{p(X)}{q(X)}\right]\\&=\mathbb{E}_P[\log p(X)-\log q(X)]\end{aligned}$$

p和q是定义在符号集$\mathcal{X}$上的两种分布。在该定义中，p和q的地位不对等。

相对熵为$[\log p(X)-\log q(X)]$在分布p下的期望，衡量p和q两个分布之间的差异。

可以证明，相对熵非负：

[1] 即最大频率，默认中心频率为 0

[2] 相当于某个函数与窗函数$\Pi(t)=u\left(t+\frac{1}{2}\right)-u\left(t-\frac{1}{2}\right)$卷积，所以频域长度不低于窗函数的无穷长度

$$
\begin{aligned}
-D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{q(x)}{p(x)} = \mathbb{E}_P\left[\log\frac{q(x)}{p(x)}\right] \\
&\leqslant \log\mathbb{E}_P\left[\frac{q(x)}{p(x)}\right] = \log\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)} \\
&= \log\sum_{x\in\mathcal{X}} q(x) = 0
\end{aligned}
$$

相对熵满足非负性但不满足对称性和三角不等式，所以不是距离。

$W:\mathcal{P}\to\mathcal{Q}, X\mapsto Y$ 是一个信道。那么对 $X,\tilde{X}\in\mathcal{P}$，有 $D\left(W\circ X||W\circ\tilde{X}\right)\leqslant D\left(X||\tilde{X}\right)$，即传输会减少分布间的差异(Transmission reduces divergence)。这是符合直觉的，因为同一种传输会使原先不同的分布趋同演变，从而使各种分布一起接近同一个稳定的状态，类似于马尔科夫链，每次迭代都会更接近稳态。所以两个不同的分布在传输后会越来越近。

$$
\begin{aligned}
D\left(W\circ X||W\circ\tilde{X}\right) &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_X(x)W(y|x)\log\frac{\sum_{x\in\mathcal{X}} p_X(x)W(y|x)}{\sum_{x\in\mathcal{X}}\tilde{p}_X(x)W(y|x)} \\
&\leqslant \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_X(x)W(y|x)\log\frac{p_X(x)W(y|x)}{\tilde{p}_X(x)W(y|x)} \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_X(x)W(y|x)\log\frac{p_X(x)}{\tilde{p}_X(x)} \\
&= \sum_{x\in\mathcal{X}} p_X(x)\log\frac{p_X(x)}{\tilde{p}_X(x)} \\
&= D\left(X||\tilde{X}\right)
\end{aligned}
$$

链式法则：

$$
D(p_{XY}(X,Y)||q_{XY}(X,Y)) = D(p_X(X)||q_X(X)) + D(p_{Y|X}(Y|X)||q_{Y|X}(Y|X))
$$

*交叉熵(Cross Entropy)

$$
\begin{aligned}
H(p,q) &= -\sum_{x\in\mathcal{X}} p(x)\log q(x) = \mathbb{E}_P[\log q(x)] \\
&= H(p) + D(p||q) \geqslant H(p)
\end{aligned}
$$

对比信息熵的定义，这是把另一个分布 q 的“虚假”信息量施加到真实的分布 p 上产生的平均“虚假”信息量。可以理解为把针对分布 q 的理想编码应用于分布时 p 的平均码长。

互信息(Mutual Information)

$$
\begin{aligned}
I(X;Y) &= D(p_{XY}(X,Y)||p_X(X)p_Y(Y)) \\
&= \mathbb{E}_{XY}\left[\log\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right]
\end{aligned}
$$

根据相对熵的定义，互信息衡量两个随机变量 X 和 Y 联合分布与独立分布的差异，即两个变量有多不独立。

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{XY}\left[\log \frac{p_{XY}(X,Y)}{p_X(X)\,p_Y(Y)}\right] \\ &= H(X)+H(Y)-H(X,Y) \\ &= H(X)-H(X|Y)=H(Y)-H(Y|X) \end{aligned}$$

互信息可以理解为同时观测到X和Y时，相比X和Y分别单独观察可以减少的信息量，或者已知一个变量时再观察另一个变量，相比单独观测该变量时可以减少的信息量。同样可理解为两者的相关性。

由相对熵的性质可得，互信息不小于0，所以$H(X|Y)\leqslant H(X)$，即给出其它条件可以减小信息量(Conditioning reduces entropy)。

如图 1，信息熵、条件熵、联合熵和互信息可以用 Venn 图表示。每个“集合”中“元素”的数量，即每个图形的面积，代表信息量。这种表示绝不只是用来方便记忆，而还是理解相关概念的重要方式。

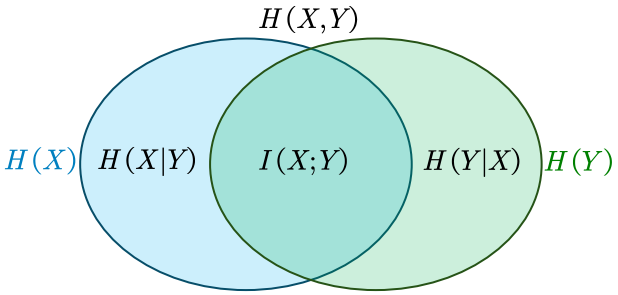

图 1 Venn 图表示

离散变量与连续变量的互信息在概念上完全相同，因为两者相差的“无穷大”在相减中可以抵消。也因此可以定义连续变量与离散变量之间的互信息，例如用高斯信道传输离散变量的情形。

条件互信息(Conditional Mutual Information)

$$\begin{aligned} I(X;Y|Z) &= H(X|Z)-H(X|Y,Z) \\ &= \mathbb{E}_{XYZ}\left[\log \frac{p_{XY|Z}(X,Y|Z)}{p_{X|Z}(X|Z)\,p_{Y|Z}(Y|Z)}\right] \\ &= \sum_{z\in\mathcal{Z}} p_Z(z)\,I(X;Y|Z=z) \\ &= \mathbb{E}_Z[I(X;Y|Z=z)] \end{aligned}$$

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

如图 2，要算X和(Y,Z)这一联合分布之间的信息量，可以先算X与其一的信息量，再加上其一给定后X与其二的信息量。

与条件熵类似的链式法则：

$$
\begin{aligned}
I(X_1,X_2,\cdots,X_n;Y) &= H(X_1,X_2,\cdots,X_n) - H(X_1,X_2,\cdots,X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_1,X_2,\cdots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|Y,X_1,X_2,\cdots,X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i;Y|X_1,X_2,\cdots,X_{i-1})
\end{aligned}
$$

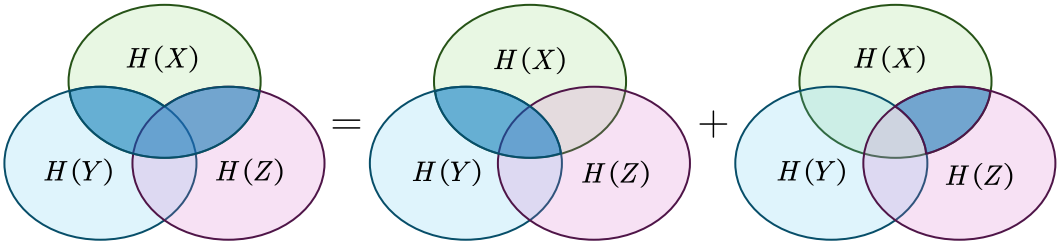

图 2 条件互信息等式的 Venn 图表示

凹凸性(Concavity and Convexity)

$$
\begin{aligned}
H&: \mathcal{P} \to \mathbb{R}^* \\
D&: \mathcal{P} \times \mathcal{P} \to \mathbb{R}^* \\
I&: \mathcal{P} \times \mathcal{Q} \to \mathbb{R}^*
\end{aligned}
$$

熵、相对熵和互信息都可以看成概率测度(Probability measure)映射到非负实数的泛函，所以可以在泛函层面定义凹凸性。

$$
\begin{aligned}
D(\lambda_1 p_1 + \lambda_2 p_2 || \lambda_1 q_1 + \lambda_2 q_2) &\leq \lambda_1 D(p_1||q_1) + \lambda_2 D(p_2||q_2) \\
H(p) &= \log|\mathcal{X}| - D(p||u)
\end{aligned}
$$

$I(X;Y)$ 在 $p_{Y|X}(y|x)$ 固定时对 $p_X(x)$ 凹，在 $p_X(x)$ 确定时对 $p_{Y|X}(y|x)$ 凸。

马尔科夫链(Markov Chain)

$$
\begin{aligned}
X - Y - Z \iff p_{XYZ}(x,y,z) &= p_X(x)\, p_{Y|X}(y|x)\, p_{Z|Y}(z|y) \\
p_{Z|XY}(z|x,y) &= p_{Z|Y}(z|y)
\end{aligned}
$$

后一个状态只于当前状态有关，与更早的状态无关。

$$
\begin{aligned}
p_{Z|XY}(z|x,y) &= p_{Z|Y}(z|y) \\
\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)} &= \frac{p_{YZ}(y,z)}{p_Y(y)} \\
\frac{p_{XYZ}(x,y,z)}{p_Y(y)} &= \frac{p_{XY}(x,y)\, p_{YZ}(y,z)}{p_Y^2(y)} \\
p_{XZ|Y}(x,z|y) &= p_{X|Y}(x|y)\, p_{Z|Y}(z|y)
\end{aligned}
$$

给定当前状态后，后一个状态与前一个状态独立（另一种表述）。这种形式表明马尔科夫链是对称的关系。

$$
X - Y - f(Y)
$$

$f(Y)$由Y唯一确定，所以在Y给定后显然与X独立。

$$X-Y-Z\Longrightarrow H(Z|X,Y)=H(Z|Y),I(X;Z|Y)=0$$

由条件独立性可得显然成立。

数据处理不等式(Data Processing Inequality)

$$X-Y-Z\Longrightarrow I(X;Y)\geqslant I(X;Z)$$

$$\begin{aligned}I(X;Y)&=I(X;Y)+I(X;Z|Y)=I(X;Y,Z)\\&=I(X;Z)+I(X;Y|Z)\geqslant I(X;Z)\end{aligned}$$

距离越远的变量分布差别越大。

$$I(X;Y)\geqslant I(X;f(Y))$$

$$X_1-X_2-X_3-X_4\Longrightarrow I(X_2;X_3)\geqslant I(X_1;X_4)$$

$$X-Y-Z\Longrightarrow I(X;Y)\geqslant I(X;Y|Z)$$

充分统计量(Sufficient Statistics)

设X的分布受参数θ控制，且T是X的函数，则一定有马尔科夫链$\theta-X-T(X)$，进而有$I(\theta;T(X))\leqslant I(\theta;X)$. 若取等，则未损失信息，即$\theta-T(X)-X$，称此时的$T(X)$为$\theta$的充分统计量。

注意这里的参数θ是一个分布未知的随机变量，$\theta-T(X)-X$说明给定$T(X)$后X可以有随机性，但一定与θ独立，所以$T(X)$可以不包含X本身的全部信息，但一定包含X能够提供的有关θ的全部信息。例如，二项分布中样本的和就是单次概率的充分统计量。

一致最小方差无偏估计(UMVUE)一定是充分统计量的函数。

费诺不等式(Fano's Inequality)

对任意估计量$X-Y-\hat{X}$，记错误率$P_e=\mathbb{P}\left(\hat{X}\neq X\right)$，则

$$H(P_e)+P_e\log|\mathcal{X}|\geqslant H\left(X|\hat{X}\right).$$

注意这里$H(P_e)$中的P_e是分布，而不是变量。

$$\begin{aligned}H\left(X|\hat{X}\right)&=H\left(X|\hat{X}\right)+H\left(X\oplus\hat{X}|X,\hat{X}\right)=H\left(X\oplus\hat{X},X|\hat{X}\right)\\&=H\left(X\oplus\hat{X}|\hat{X}\right)+H\left(X|X\oplus\hat{X},\hat{X}\right)\\&\leqslant H\left(X\oplus\hat{X}\right)+(1-P_e)H\left(X|\hat{X},X=\hat{X}\right)+P_eH\left(X|\hat{X},X\neq\hat{X}\right)\\&=H(P_e)+P_eH\left(X|\hat{X},X\neq\hat{X}\right)\\&\leqslant\begin{cases}H(P_e)+P_e\log\left(|\mathcal{X}|-1\right),\hat{\mathcal{X}}\subseteq\mathcal{X}\\H(P_e)+P_e\log|\mathcal{X}|\end{cases}\end{aligned}$$

不等式的两边可以放松。

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

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

既然误码率和X与Y之间的关系有关，为什么这里用的是条件熵而非互信息？因为误码率的分子只关心X和$\hat{X}=g(Y)$在相关的部分之外还有多少不相关的部分（条件熵），而不关心两者有多相关（互信息），分母则关心X的总信息量$H(X)$。证明过程的最后一步的$\log|\mathcal{X}|$可收紧到$H(X)$，故最后结论可收紧到$P_e\geq\frac{H(X|Y)-1}{H(X)}$. 如图 3，误码率下界刚好接近$X$中无法与$Y$耦合部分的占比。当$H(X|Y)$足够大后，分子的$-1$可忽略不计。

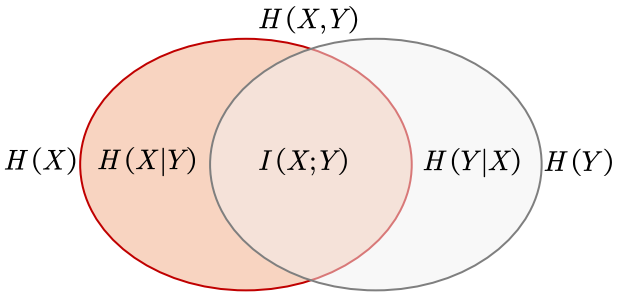

图 3 Fano 不等式的 Venn 图表示

矩阵表示

为方便使用矩阵表示形如$\{X_1,X_2,\cdots\}$的马尔可夫链，这里把定义限定得更严格，只考虑时不变(time-homogeneous)即平稳(stationary)的一阶马尔科夫链。

状态空间(State space)$\mathcal{S}$：可数多种状态$\mathcal{S}=\{1,2,\cdots,N\}$或$\mathcal{S}=\{1,2,\cdots\}$.

初态分布(Initial distribution)$\boldsymbol{\pi}_0$：$\forall i\in\mathcal{S}$，$\pi_0(i)=\mathbb{P}[X_0=i]\geq 0$，且$\sum_{i\in\mathcal{S}}\pi_0(i)=1$.

概率转移矩阵(Probability transition matrix)$\boldsymbol{P}_{n\times n}=(p_{ij})$：$p_{ij}=\mathbb{P}[X_{n+1}=j|X_n=i]$，需满足$\sum_{j=1}^{n}p_{ij}=1$，即每行之和为1.

有上述概念后，第n个状态就可用初态分布和转移矩阵表示为$\boldsymbol{\pi}_n=\boldsymbol{\pi}_{n-1}\boldsymbol{P}=\boldsymbol{\pi}_0\boldsymbol{P}^n$.

基本极限定理(Basic Limit Theorem)

稳态分布(Stationary distribution)：$\boldsymbol{\pi}$ s.t. $\boldsymbol{\pi}\boldsymbol{P}=\boldsymbol{\pi}$, $\sum_{i\in\mathcal{S}}\pi(i)=1$.

如果概率转移矩阵是双随机的(Doubly stochastic)，即每列之和也为1，那么稳态分布就是均匀分布。

基本极限定理：如果马尔科夫链$\{X_0,X_1,\cdots\}$不可约(irreducible)，非周期(aperiodic)

且具有一个稳态分布$\boldsymbol{\pi}$，则对任意初态分布$\boldsymbol{\pi}_0$，都有$\lim\limits_{n\to\infty}\boldsymbol{\pi}_n=\boldsymbol{\pi}$.

熵率(极限熵，Entropy Rate)

$$H(\mathcal{X})=\lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots,X_n)$$

即足够长的字符串平均到每个字符的熵。

对平稳随机过程(stationary stochastic process)，有

$$\begin{aligned}H(\mathcal{X})&=\lim_{n\to\infty}H(X_n|X_1,X_2,\cdots,X_{n-1})\\&\leqslant H(X_k|X_1,X_2,\cdots,X_{k-1})\leqslant H(X_{k-1}|X_1,X_2,\cdots,X_{k-2})\leqslant\cdots\leqslant H(X_1)\end{aligned}$$

对平稳的马尔科夫链，有

$$\begin{aligned}H(\mathcal{X})=H_\infty&=\lim_{k\to\infty}H(X_k|X_1,X_2,\cdots,X_{k-1})\\&=\lim_{k\to\infty}H(X_k|X_{k-1})\\&=\lim_{k\to\infty}\sum_{i\in\mathcal{S}}\pi_{k-1}(i)H(X_k|X_{k-1}=i)\\&=\sum_{i\in\mathcal{S}}\pi(i)H(\boldsymbol{p}_i)=-\sum_{i\in\mathcal{S}}\pi(i)\sum_{j\in\mathcal{S}}p_{ij}\log p_{ij}\end{aligned}$$

渐近均分性质(Asymptotic Equipartition Property)

如果$X_1,X_2,\cdots,X_n$独立同分布，分布函数为$p_X(x)$，则有

$$-\frac{1}{n}\log p_{\boldsymbol{X}}(X_1,X_2,\cdots,X_n)\xrightarrow{\mathbb{P}}H(X)$$

$$-\frac{1}{n}\log p_{\boldsymbol{X}}(X_1,X_2,\cdots,X_n)=-\frac{1}{n}\sum_{i=1}^{n}\log p_X(X)\xrightarrow{\mathbb{P}}\mathbb{E}_X[\log p_X(X)]=H(X)$$

依概率收敛：$\forall\varepsilon>0,\exists n\in\mathbb{N}^+\text{s.t.}\,\mathbb{P}\left[\left|-\frac{1}{n}\sum_{i=1}^{n}\log p_X(X)-H(X)\right|<\varepsilon\right]$. 取到极限值的概率可能永远是$0$，但分布一定会越来越接近。

设$p_X(x)$的最小值是p_1，最大值是p_2，则随机变量$-\frac{1}{n}\log p_{\boldsymbol{X}}(X_1,X_2,\cdots,X_n)$[3]的最小值是$-\log p_2$，最大值是$-\log p_1$. 随着$n$的增大，取到这两个值的概率始终大于$0$，但会越来越小；取到$H(X)$附近的概率则会越来越大。

[3] 虽然里面包含了一个概率分布函数，但该概率分布函数在此处也只作为一个X^n的函数，随机性体现在$\boldsymbol{X}$上（注意这里都是大写，说明是变量而不是具体取值）而不在该函数上。

以二项分布为例，不失一般性，设取到0的概率p满足$0<p<\frac{1}{2}$，则取到1的概率为$(1-p)$。对n个随机变量X，其联合分布一共有2^n种，但变量$-\frac{1}{n}\log p_{\boldsymbol{X}}(X_1,X_2,\cdots,X_n)$的取值只有$n$种。$n$个$X$中共有$k$个0的情形有$\binom{n}{k}$种，每种的概率均为$p^k(1-p)^{n-k}$，则相应使得变量$-\frac{1}{n}\log p_{\boldsymbol{X}}(X_1,X_2,\cdots,X_n)$取到$-\frac{1}{n}\log p^k(1-p)^{n-k}$的概率为$\binom{n}{k}p^k(1-p)^{n-k}$.

图 4 为不同n的值对应的变量分布，可见随着n的增大，该变量会逐渐集中于熵$H(p)$.

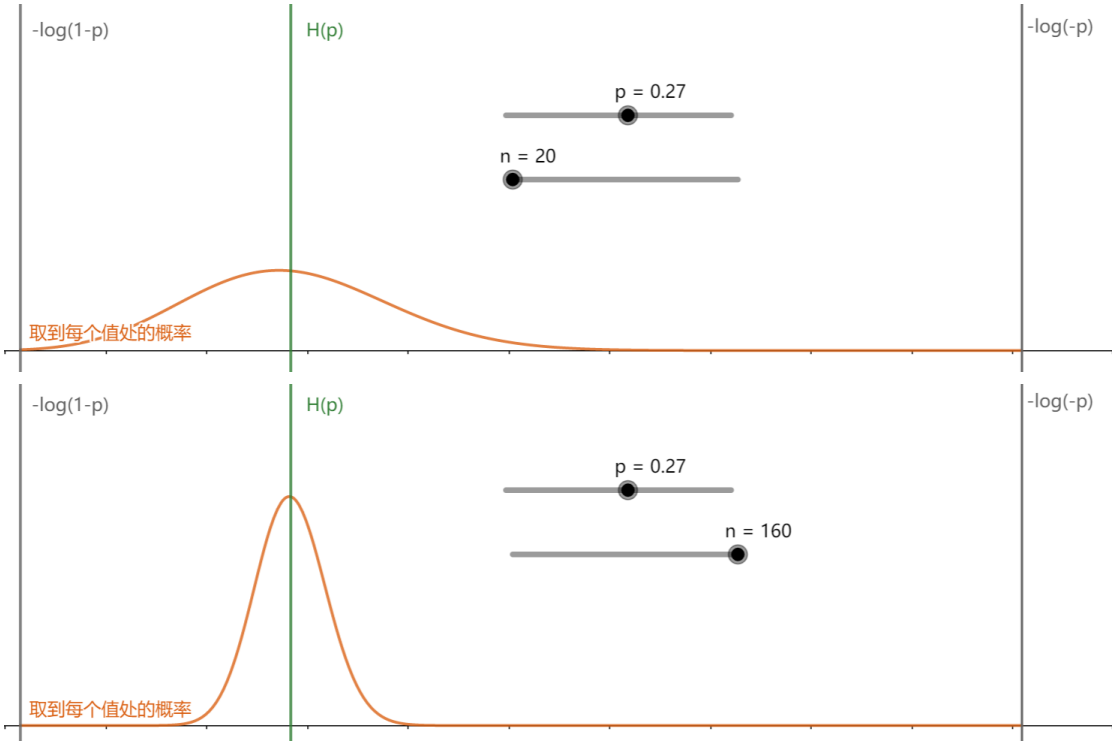

图 4 不同n的值对应的变量分布

典型集(Typical set)

$$A_\varepsilon^{(n)}=\{(x_1,x_2,\cdots,x_n)\in\mathcal{X}^n:2^{-n(H(X)+\varepsilon)}\leqslant p_{\boldsymbol{X}}(x_1,x_2,\cdots,x_n)\leqslant 2^{-n(H(X)-\varepsilon)}\}$$

条件$2^{-n(H(X)+\varepsilon)}\leqslant p_{\boldsymbol{X}}(x_1,x_2,\cdots,x_n)\leqslant 2^{-n(H(X)-\varepsilon)}$也可以写成上面出现过的形式

$$\left|-\frac{1}{n}\log p_{\boldsymbol{X}}(x_1,x_2,\cdots,x_n)-H(X)\right|\leqslant\varepsilon.$$

由定义得，$\forall(x_1,x_2,\cdots,x_n)\in A_\varepsilon^{(n)}$，$H(X)-\varepsilon\leqslant-\frac{1}{n}\log p_{\boldsymbol{X}}(x_1,x_2,\cdots,x_n)\leqslant H(X)+\varepsilon$.

根据渐近均分性，$\forall\varepsilon>0,\exists n\in\mathbb{N}^+\text{s.t.}\,\mathbb{P}[A_\varepsilon^{(n)}]\geqslant 1-\varepsilon$.

重要的数量性质：$(1-\varepsilon)2^{n(H(X)-\varepsilon)}\leqslant|A_\varepsilon^{(n)}|\leqslant 2^{n(H(X)+\varepsilon)}$（证明见课本）.

上述数量性质表明，如果变量个数n足够大，则大概率出现的变量组合（或字符串）只占2^n个中的$2^{nH(X)}$个左右。这就体现出概率分布和数量分布的区别。如图 5[4]，同样以二项分布为例，可以观察到熵越小，总概率分布与分布数量分布差距越大。单个字符串的概率足够大时，落在此区间的字符串总数太少，所以总概率不大。总概率最大的区间占了几乎所有的概率，但不一定占了几乎所有的字符串数量。熵越小，字符串概率分布的峰值离字符串数量分布的峰值越远。

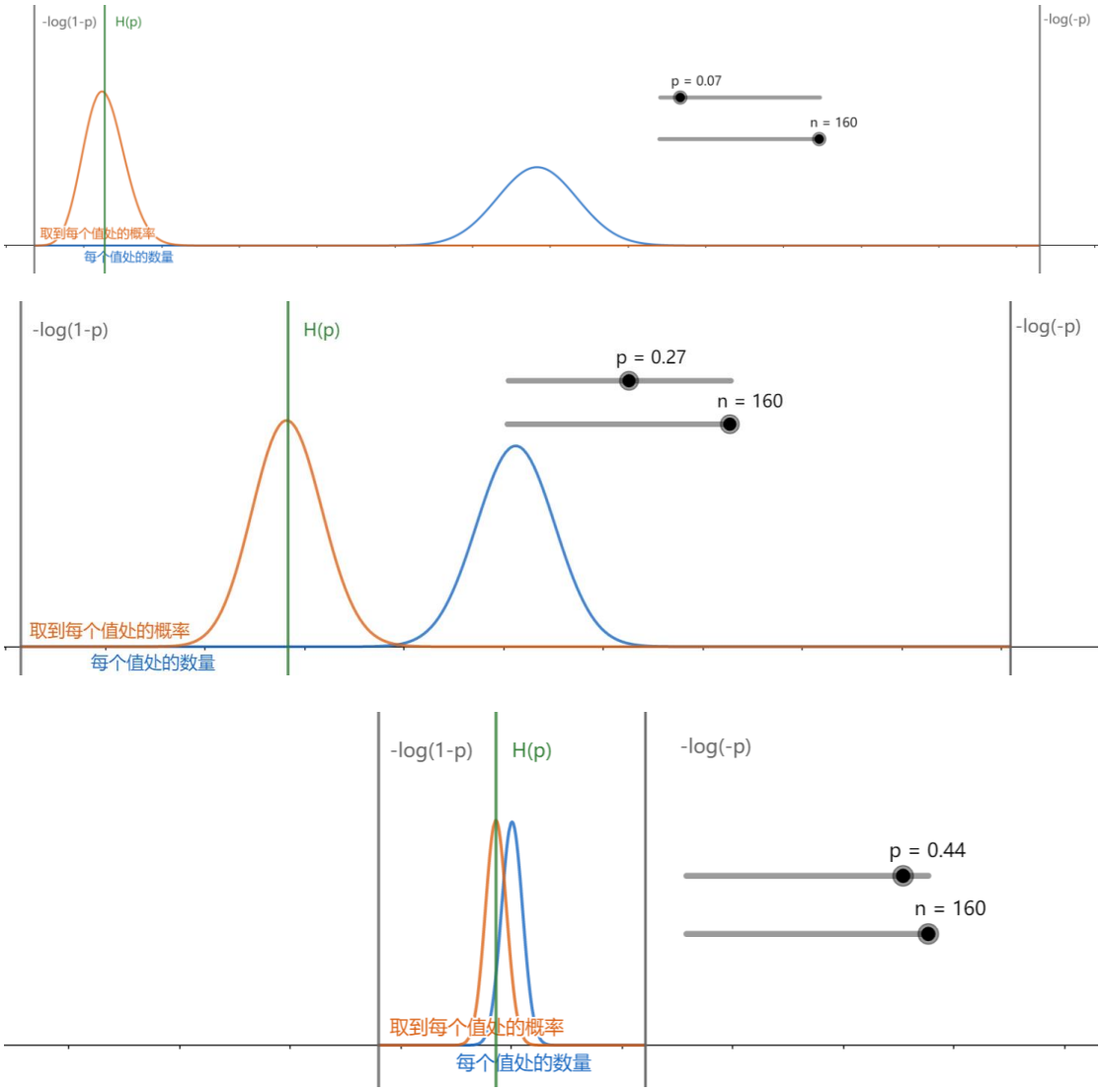

图 5 每个值处对应总概率和总数的分布

渐近均分性质指出了信息压缩和信息传输中的重要事实：如果熵不高，则看似可能传很多种字符串，但真正比较可能传到只有其中典型集内的部分，典型集外的字符串则几乎不可能出现。信息量的定义源于数据，自此又归于数据，完成了逻辑闭环，展现其外在价值[5]。相应的压缩或者传输方法就是只考虑出现概率大的（典型集内的）字符串，忽略那些相比之下几乎不可能出现的字符串。这种方法当然只在理论分析时有用，实际上字符串不会足够长，也不可能置典型集外的字符串于不顾。

[4] 横坐标随单个字符的概率单调递减，所以概率峰处于数量峰左边
[5] 内在价值就是满足的一系列自洽的性质，例如可加性、链式法则（均为个人观点）

信源编码(Source Coding)

通信系统的常见如图 6[6]。信源编码是对数据进行压缩，以尽可能减少原始数据的冗余；信道编码是以特定的方式增加冗余，以降低传输的误码率（参考信道编码定理）。

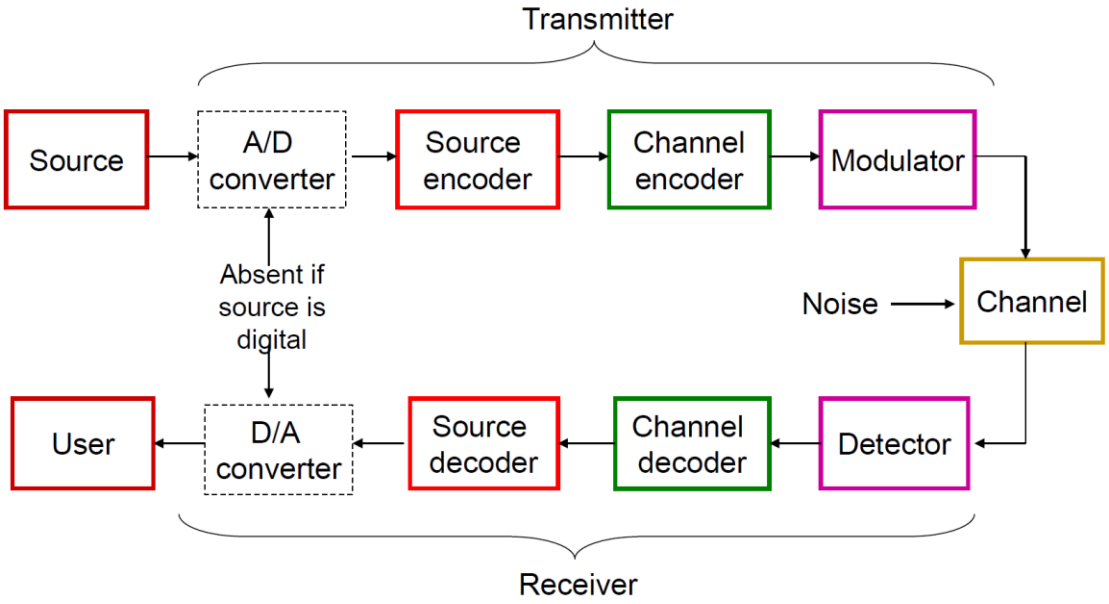

图 6 通信系统的常见结构

信源编码的定义

对随机变量X的信源编码C是从X的取值范围$\mathcal{X}$到$\mathcal{D}^*$的一个映射，其中$\mathcal{D}^*$为D元符号集$\mathcal{D}$上有限长度字符串构成的集合。$C(x)$表示x的码字，$l(x)$表示$C(x)$的长度。

信源编码$C(x)$的期望长度为$L(C)=\mathbb{E}[l(X)]$.

C若为单射，即$\forall x\neq x'\Rightarrow C(x)\neq C(x')$，则是非奇异的(nonsingular).

C的拓展(extension)为$C(x_1x_2\cdots x_n)=C(x_1)C(x_2)\cdots C(x_n)$.

编码的几种类型

若编码的拓展编码非奇异，则称其是唯一可译的(uniquely decodable). 虽然唯一可译，但有可能需要分析整个d^n字符串才能解码其中某个字符x_i.

表 1 各种编码示例

X	奇异	非奇异，不唯一可译	唯一可译，不即时	即时
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

若编码中无任何码字是其它码字的前缀，则称其为前缀码(prefix code)或即时码(instantaneous code). 这种码字不需要参考后面的字符就可以看出每个码字何时结束（即自我间断码，self-punctuating code），所以不需要参考后面的码字就可以完成译码。类似

[6] 图源为陶梅霞老师《通信原理》课件

的还有后缀码(suffix code).

奇异、唯一可译、即时码的示例如表 1，集合关系如图 7[7].

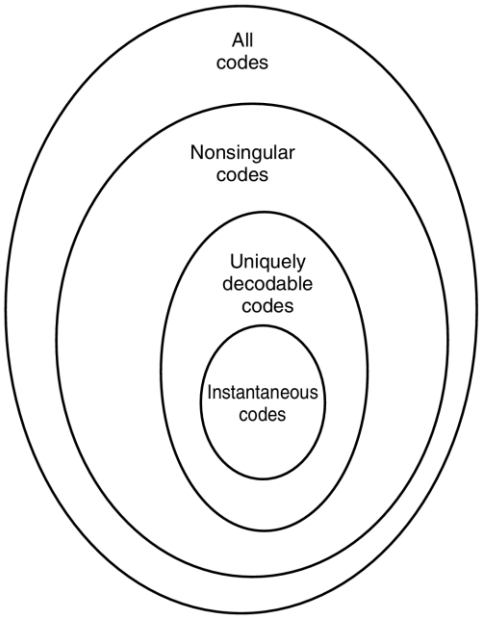

图 7 不同编码分类关系

Kraft 不等式

对D元符号集上的前缀码，码字长度$l_1, l_2, \cdots l_m$必满足不等式

$$\sum_{i=1}^{m} D^{-l_i} \leqslant 1$$

反之，若给定满足以上不等式的一组码字长度，则一定存在相应长度的前缀码。

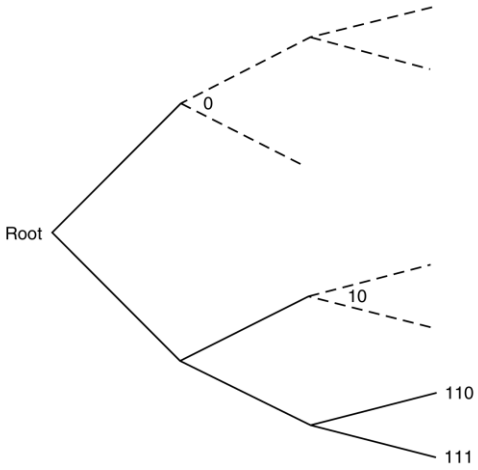

图 8 二叉树对应的前缀码

如图 8，每套前缀码的所有码字都可以和某个（非完全）D叉树上部分或全部的叶节点互相对应，每个码字的长度等于相应叶节点的深度。设$l_{\max}$为单个码字的最大长度。如

[7] Elements of Information Theory. Thomas M. Cover, Joy A. Thomas.–2nd ed.

果把D叉树补满到$l_{\max}$层，则此前的第l_i层在第$l_{\max}$层有$D^{l_{\max}-l_i}$个后代。在补满D叉树前，每层都有可能有叶节点，都可以分别对应到第$l_{\max}$层的后代。所有叶节点对应的第$l_{\max}$层的后代总数不超过第$l_{\max}$层的上限$D^{l_{\max}}$，即

$$\sum_{i=1}^{m} D^{l_{\max}-l_i} \leqslant D^{l_{\max}}$$

Kraft 不等式还可以推广到可数无限多种码字，和一切唯一可译码的情况。

最优码(Optimal Codes)

最优码的优化目标是在D元符号集上找到最小期望长度的前缀码，约束条件是 Kraft 不等式，即

$$\min_{l^m}\left[L=\sum_{i=1}^{m} p_i l_i\right] \text{ s.t. } \sum_{i=1}^{m} D^{-l_i} \leqslant 1$$

利用与并联高斯信道类似的求解法：观察到所有D^{-l_i}总和的上限固定，所以把期望码长求和里的每一部分对D^{-l_i}求偏导，得

$$\frac{\partial}{\partial D^{-l_i}} p_i l_i = -\frac{\partial}{\partial D^{-l_i}} p_i \log_D D^{-l_i} = -\frac{p_i}{D^{-l_i}\ln D}$$

若达到最优值，则必满足所有偏导相等（否则必然可以重新分配），即每个$\frac{p_i}{D^{-l_i}}$相等。

再把约束条件分配到每个D^{-l_i}，如果能取等，则$D^{-l_i}=p_i$，每个码字长度$l_i=-\log_D p_i$，总的期望长度为

$$L=\sum_{i=1}^{m} p_i l_i = -\sum_{i=1}^{m} p_i \log_D p_i = H_D(X).$$

这是期望码长的理论下限，但现实中往往达不到，因为码长只能取整数。考虑这一点，最优码的期望长度满足

$$H_D(X)=-\sum_{i=1}^{m} p_i \log_D p_i \leqslant L \leqslant \sum_{i=1}^{m} p_i \lceil -\log_D p_i \rceil < \sum_{i=1}^{m} p_i(-\log_D p_i+1) = H_D(X)+1$$

所以，对一个完整的字符串，每个字符的最小期望码长满足

$$\frac{H(X_1,X_2,\cdots,X_n)}{n} \leqslant L_n^* < \frac{H(X_1,X_2,\cdots,X_n)}{n}+\frac{1}{n}$$

若$X_1,X_2,\cdots,X_n$是平稳随机过程，则$L_n^* \to H(\mathcal{X})$，即最小期望长度趋近于熵率。

为此可以把多个字符组成一段字符串，当成一种新字符，对其整体编码，以尽量接近平均码长的下限。用单个字符编码，或者多个字符一起编码来表示字符串的方式统称为**分组码**(block code)。

构造最优码，就是构造从X到最接近D进制均匀分布的分布的映射。

霍夫曼码(Huffman Codes)

如图 9，不断合并概率最小的两个节点，构造出霍夫曼树，再为每个节点分配码字即可。在规定字符种类后，霍夫曼码是最优的。

Codeword Length	Codeword	X	Probability
2	01	1	0.25
2	10	2	0.25
2	11	3	0.2
3	000	4	0.15
3	001	5	0.15

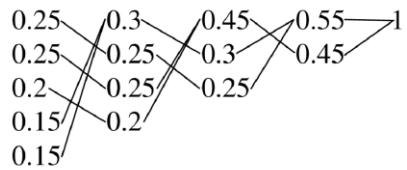

图 9 霍夫曼码的构造

香农码(Shannon Codes)

香农码满足每个码长$l_i = \lceil -\log_D p_i \rceil$. 香农码不一定是最优码，例如$D=2$，概率分别为0.9999 和 0.0001，其码长为 1bit 和 14bit，大量的浪费导致香农码无法达到最优。

把每个信源符号按概率从大到小依次分配到$[0,1)$，若累计概率$\mathbb{P}[X \leqslant i] = P_i$，则第$i$个符号分配到的区间为$[P_{i-1}, P_i)$，用$\lceil -\log_D p_i \rceil$位$D$进制小数（不包括小数点前的 0）表示区间当中的某个值，以此作为该符号的码字。

算术编码(Arithmetic Coding)

算术编码把整个数据文件当作一个长字符串X^n来统一编码，所以不属于分组码。其编码方式与香农码类似，但考虑整个字符串的联合分布，为每种X^n分配一个区间，再用D进制小数（不包括小数点前的 0）区间里的某个值。

LZ78 压缩算法(Abraham Lempel-Jacob Ziv)

设信源符号为$a_1, a_2, \cdots, a_K$共K个；

在第i步，从s_{i-1}短语后的第一个符号开始向后搜索此前未出现过的最短短语s_i，将该短语添入字典第i段s_i；

如果s_i去掉最后一个符号x后所得的前缀是在第j步出现的短语，那么对s_i的编码为(j,x)。对段号j，用$\lceil \log i \rceil$bit 来表示，符号x用$\lceil \log K \rceil$bit 来表示。

游程编码(RLE, Run-Length Encoding)

把连续不变序列用连续的长度来简便表示。这对前后明显正相关的序列压缩效果很好，如图像和视频。

信道容量(Channel Capacity)

离散无记忆信道(Discrete Memoryless Channel)定义

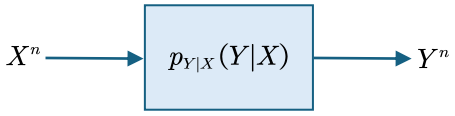

图 10 离散无记忆信道

$$p_{Y^n|X^n}(y^n|x^n)=\prod_{i=1}^{n}p_{Y|X}(y_i|x_i)$$

传输可能产生错误，离散接收机Y每一个取值分别可能对应多种离散信源X的取值，且概率确定，不随时间变化，所以每次收发都独立，不会受已发信号的影响。

高斯信道(Gaussian Channel)定义

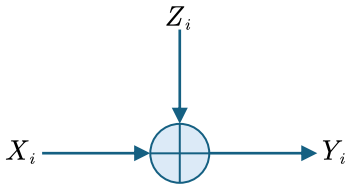

图 11 高斯信道

$$Y=X+Z,\ Z\sim\mathcal{N}(0,N)$$

$$p_{Y|X}(y|x)=\frac{1}{\sqrt{2\pi N}}\mathrm{e}^{-\frac{(y-x)^2}{2N}}$$

同样无记忆，输入X可以离散也可以连续，输出Y必然连续。因为连续输入无法无损复原，所以计算信道容量时考虑离散输入，计算率失真时通常考虑连续输入，噪声Z与输入X独立。

高斯信道通常对输入做限制，例如输入的总功率有限。对X的任意码字$(x_1,x_2,\cdots,x_n)$，需满足$\frac{1}{n}\sum_{i=1}^{n}x_i^2\leqslant P$.

信道编码(Channel Coding)的定义

离散无记忆信道$(\mathcal{X},p_{Y|X}(y|x),\mathcal{Y})$的$(M,n)$码包括：

- 下标集$\{1,2,\cdots,M\}$.
- 编码函数$X^n:\{1,2,\cdots,M\}\to\mathcal{X}^n$，生成码字$x^n(1),x^n(2),\cdots,x^n(M)$. 所有码字的集合为码簿(codebook).
- 译码函数$g:\mathcal{Y}^n\to\{1,2,\cdots,M\}$，这是固定的译码规则，函数本身没有随机性.

功率不超过P的高斯信道的(M,n)码包括：

- 下标集$\{1,2,\cdots,M\}$.
- 编码函数$X^n:\{1,2,\cdots,M\}\to\mathcal{X}^n$，生成码字$x^n(1),x^n(2),\cdots,x^n(M)$，且满足功率

限制 $\sum_{i=1}^{n} x_w^2(i) \leqslant nP,\ w=1,2,\cdots,M.$

- 译码函数 $g: \mathcal{Y}^n \to \{1,2,\cdots,M\}.$

条件误差概率(conditional probability of error):

$\lambda_i = \mathbb{P}[g(Y^n) \neq i | X^n = x^n(i)] = \sum_{y^n} p_{Y^n|X^n}(y^n|x^n(i)) I(g(y^n) \neq i)$，在发送 i 时译码的错误概率。

最大误差概率(maximum probability of error): $\lambda^{(n)} = \max_{i\in\{1,2,\cdots,M\}} \lambda_i.$

（算数）平均误差概率(average probability of error): $P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{n} \lambda_i.$

码率(rate): $R = \frac{\log_2 M}{n}$ bits per transmission.

如果存在一个 $(\lceil 2^{nR} \rceil, n)$ 码序列，满足 $\lim_{n\to\infty} \lambda^{(n)} \to 0$，则码率 R 是可达的(achievable).

传输信息通常需要经过信道编码，因为编码前的信息和直接用于收发的信息通常不是同一类[8]。编码前的信息与收发的信息甚至可能不是同一种进制。

信道编码的另一个作用是改变误码率。如果码率足够低，误码率总可以也达到任意小，除非输入和输出完全独立。

信道容量就是所有可达码率的上确界(supremum)。

信息信道容量(Information Channel Capacity)

信息信道容量为输入 X 和输出 Y 之间，在一定限制条件下可达的最大互信息。

对离散信道，容量为 $C = \max_{p_X(x)} I(X;Y)$，即任意可能的信源分布下可达[9]的最大互信息。

对高斯信道，容量为 $C = \max_{p_X(x):\mathbb{E}_X[X^2]\leqslant P} I(X;Y)$，即任意功率有限的输入分布下，可达的最大互信息。

信道编码定理(Channel Coding Theorem)

对离散无记忆信道或功率限制为 P 的高斯信道，小于信息信道容量的所有码率都可达。对任意码率 $R < C$，存在一个 $(2^{nR}, n)$ 码序列，其最大误差概率 $\lambda^{(n)} \to 0$.

反之，任何满足 $\lambda^{(n)} \to 0$ 的 $(2^{nR}, n)$ 码序列必定有 $R \leqslant C$.

这里不做严格证明，只简要说明其思想。根据渐近均分性质，如果码长 n 足够大，Y^n 序列自己会有大约 $2^{nH(Y)}$ 个容易出现的典型集元素，每个 X^n 序列比较可能对应的 Y^n 序列大约有 $2^{nH(Y|X)}$ 个，即 Y^n 相对于 X^n 的典型集。

---

[8] 默认信源已经经过信源编码，信道也已经完成调制(modulation)和解调(demodulation)

[9] 根据互信息的凹性，上确界一定可达

为了确保不同X^n对应的$\mathcal{Y}^n$中的典型集几乎不相交，根据抽屉原理，最多只能相应把$\mathcal{Y}^n$中比较可能出现的典型集码字再按每份$2^{nH(Y|X)}$个元素分割成$\frac{2^{nH(Y)}}{2^{nH(Y|X)}}=2^{nI(X;Y)}$个不相交集。

只发送$\mathcal{X}^n$中可与$\mathcal{Y}^n$的$2^{nI(X;Y)}$个不相交集对应的码字，就能保证每个发送码字对应的输出Y^n几乎不相交，所以传输就几乎不会出错。之所以只考虑拆分$\mathcal{Y}^n$的典型集而非所有的Y^n，是因为根据典型集元素几乎占全部概率的性质，Y^n相对于每个X^n的典型集几乎一定都在$\mathcal{Y}^n$的典型集里。

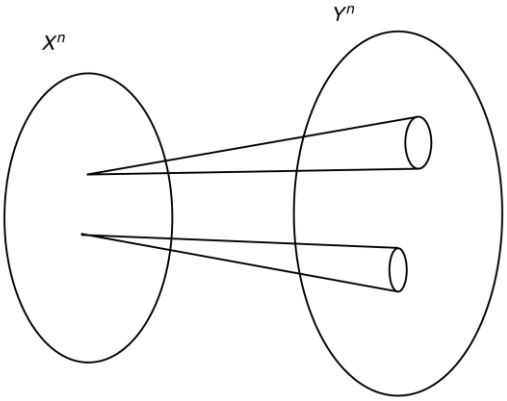

图 12 不相交的输出集合

信道编码定理表明对**这两类信道**而言，由最大互信息定义的信息信道容量就是由最大无失真码率定义的信道容量。这同样构建了从抽象的信息量到具体的数据之间的联系，其证明也需要用到渐近均分性质。

此外，信道编码定理还告诉我们一个重要事实：虽然信道的传输通常有随机性，即每一个信源X到接收端Y的映射服从概率分布而非确定的函数，但信道编码可以使得这种随机性几乎消失，在有限的码率下达到几乎为 0 的误码率。而且，从可达码率上看，无论本身的输入和输出数量或者概率分布如何、有多大随机性，最大互信息相同的信道都等价，只有编码方式的区别，这就为后面的率失真理论提供了基础。

无噪声二元信道(Noiseless Binary Channel)

$$I(X;Y)=H(X)-H(X;Y)=H(X)\leqslant 1\,\text{bit}$$

$$C=\max_{p_X(x)} I(X;Y)=1\,\text{bit}$$

达到容量时$p_X(x)=\left(\frac{1}{2}\quad\frac{1}{2}\right)$.

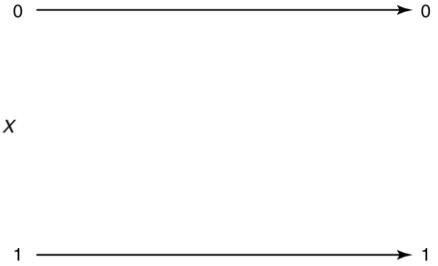

图 13 无噪声二元信道

非重叠输出噪声信道(Noisy Channel with Nonoverlapping Outputs)

$$I(X;Y)=H(X)-H(X|Y)=H(X)\leqslant 1\,\text{bit}$$
$$C=\max_{p_X(x)} I(X;Y)=1\,\text{bit}$$

达到容量时$p_X(x)=\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}$.虽然$X\to Y$有随机性，但$Y\to X$无随机性，所以没有损失信息。

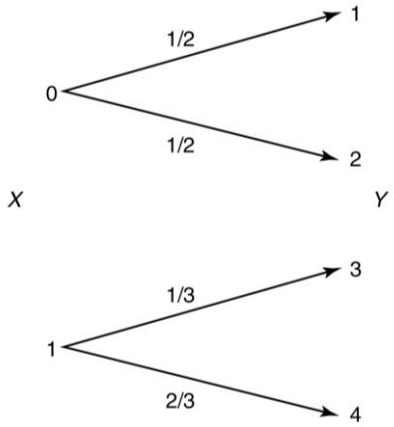

图 14 非重叠输出噪声信道

有噪声打字机信道(Noisy Typewriter)

$$I(X;Y)=H(Y)-H(Y|X)=H(Y)-1\leqslant \log_2|\mathcal{Y}|-1=\log_2 13$$
$$C=\max_{p_X(x)} I(X;Y)=\log_2 13$$

达到容量时$p_X(x)$的取值不唯一，只需使得Y为均匀分布即可。例如所有字母的均匀分布，或者仅奇数字母均匀分布，偶数字母不发送。

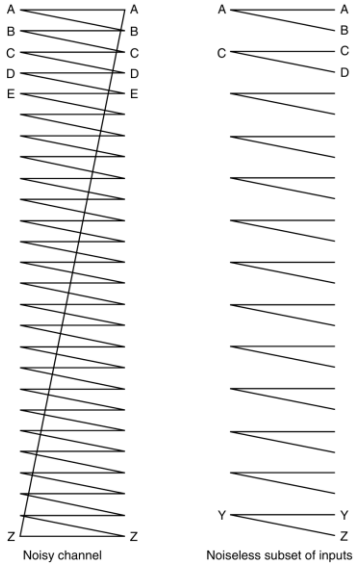

图 15 有噪声打字机信道

这也符合直觉。如图 16，传输能力的决定因素只有X和Y耦合部分的信息量$I(X;Y)$，

而与其它部分都无关。非重叠输出噪声信道（有噪无损信道）、有噪声打字机信道（无损有噪信道）分别只增加了$H(Y|X)$和$H(X|Y)$，其信道容量与相应的无噪无损信道无异。

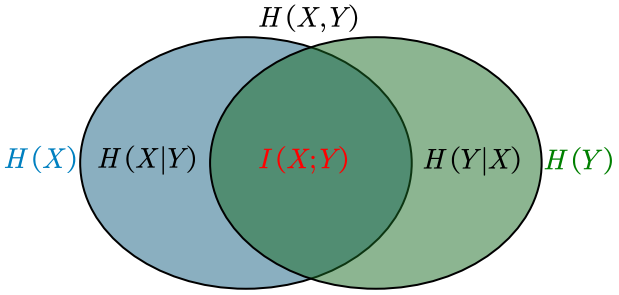

图 16 信道容量在 Venn 图中的位置

二元对称信道(Binary Symmetric Channel)

$$I(X;Y)=H(Y)-H(Y|X)=H(Y)-H(p)\leqslant 1-H(p)$$
$$C=\max_{p} I(X;Y)=1-H(p)$$

达到容量时$p_X(x)=\left(\frac{1}{2}\quad\frac{1}{2}\right)$.

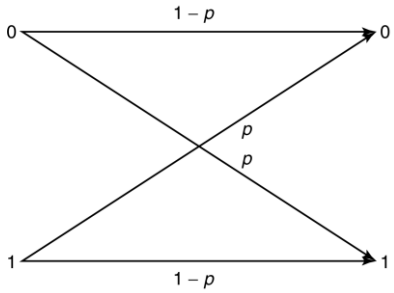

图 17 二元对称信道

二元擦除信道(Binary Erasure Channel)

$$\begin{aligned}I(X;Y)&=H(Y)-H(Y|X)\\&=-\alpha\log\alpha-p(1-\alpha)\log[p(1-\alpha)]-(1-p)(1-\alpha)\log[(1-p)(1-\alpha)]-H(\alpha)\\&=-\alpha\log\alpha-(1-\alpha)[p\log p+(1-p)\log(1-p)+\log(1-\alpha)]-H(\alpha)\\&=(1-\alpha)H(p)\leqslant 1-\alpha\end{aligned}$$

$$I(X;Y)=H(X)-H(X|Y)=H(p)-\alpha H(p)=(1-\alpha)H(p)\leqslant 1-\alpha$$

已知$p_X(x)$和$p_{Y|X}(y|x)$求解$p_{X|Y}(x|y)$可以用贝叶斯公式，即

$$p_{X|Y}(x|y)=\frac{p_{Y|X}(y|x)\,p_X(x)}{p_Y(y)}=\frac{p_{Y|X}(y|x)\,p_X(x)}{\sum_{x\in\mathcal{X}}p_{Y|X}(y|x)\,p_X(x)}.$$

$$C=\max_{p_X(x)} I(X;Y)=1-\alpha$$

对结构简单的信道，不同计算方法通常都可以得到答案，但计算量会有区别。达到容

量时 $p_X(x)=\left(\frac{1}{2}\quad\frac{1}{2}\right)$.

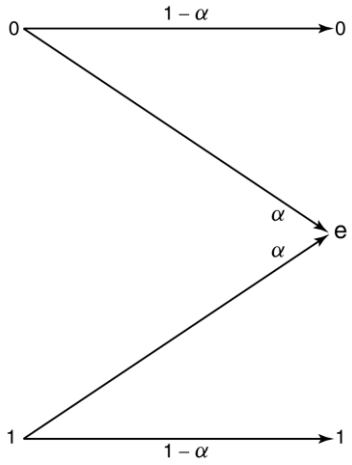

图 18 二元擦除信道

弱对称信道(Weakly Symmetric Channel)

如果信道的概率转移矩阵 $p_{Y|X}(y|x)$ 的任何两行互相置换，所有列的元素和相等，则为弱对称信道。例如，

$$p_{Y|X}(y|x)=\begin{pmatrix}\frac{1}{3}&\frac{1}{6}&\frac{1}{2}\\ \frac{1}{3}&\frac{1}{2}&\frac{1}{6}\end{pmatrix}.$$

$$I(X;Y)=H(Y)-H(Y|X)=H(Y)-H(\boldsymbol{p}_i)\leqslant\log|\mathcal{Y}|-H(\boldsymbol{p}_i)$$

$$C=\max_{p_X(x)}I(X;Y)=\log|\mathcal{Y}|-H(\boldsymbol{p}_i)$$

达到容量时 X 均匀分布，其中 $\boldsymbol{p}_i$ 为转移矩阵的任意一行。

如果任何两行互相置换，但列的元素和不全相等，就无法达到 $\log|\mathcal{Y}|-H(\boldsymbol{p}_i)$，因为 $H(Y)$ 无法达到 $\log|\mathcal{Y}|$.

矩阵分解法求信道容量

如果信道不是弱对称信道，但其概率转移矩阵可按列分割（允许改变顺序）成 r 个弱对称矩阵，即 $\boldsymbol{P}=[\boldsymbol{P}_1\ \cdots\ \boldsymbol{P}_r]$，那么互信息仍然在输入等概率分布时达到最大值，且为

$$C=\log|\mathcal{X}|-H(\boldsymbol{p}_i)-\sum_{k=1}^{r}N_k\log M_k$$

其中 N_k 是第 k 个矩阵中一行所有元素的和，M_k 是第 k 个矩阵中一列所有元素的和。

该结论是显然的，因为把信道划分为多个弱对称的部分后，每部分对应的互信息都在输入相等时达到最大值，所以总的最大值也必然在所有输入相等时取到。代入输入概率相等的条件即可得出上述表达式。

一般离散无记忆信道

如果输入和输出足够少，可以直接设一些概率为自变量，再代入计算求极值，例如二元擦除信道的推导。

如果输入和输出都很多，直接代入计算量太大，则需要用如下定理：

$$\begin{aligned}I(X;Y)&=\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}}p_{XY}(x_i,y_j)\log\frac{p_{XY}(x_i,y_j)}{p_X(x_i)p_Y(y_j)}\\&=\sum_{i\in\mathcal{X}}p_X(x_i)p_{Y|X}(y_j|x_i)\log\frac{p_{X|Y}(x_i|y_j)}{p_X(x_i)}\\&=\sum_{i\in\mathcal{X}}p_X(x_i)I(x_i;Y)\end{aligned}$$

$$\begin{aligned}&\forall x_i\text{ s.t. }p(x_i)=0,\ I(x_i;Y)\leqslant C;\\&\forall x_i\text{ s.t. }p(x_i)>0,\ I(x_i;Y)=C.\end{aligned}$$

求解方法：先假设两个$I(x_i;Y)$相等，计算其概率$p(x_i)$；如果另外某个$I(x_i;Y)$无法与之相等则设其概率$p(x_i)=0$. 如果导出矛盾则改变最初假设，直到不矛盾为止。

反馈信道容量(Feedback Capacity)

反馈信道容量等于相应的无反馈信道容量。这里不做严格证明，只举例说明其思想。

如果使用的码率R超过C，那么传输n次，在利用反馈之前，至少会有平均$\left(1-\frac{C}{R}\right)n$次误码。此时利用反馈，需要把误码的部分重新传一遍，又产生平均$\left(1-\frac{C}{R}\right)^2n$次误码。所以为了无损传输这段信息，总共需要传$\sum_{i=0}^{\infty}\left(1-\frac{C}{R}\right)^in=\frac{R}{C}n$次，平均码率为$\frac{R}{\frac{R}{C}}=C$，并没有真正增加信道容量。尽管如此，反馈信道在现实中仍有用，因为可以通过反复传输确保无误码，使误码率真正达到 0 而非仅仅趋于 0.

高斯信道

$$\begin{aligned}I(X;Y)&=h(Y)-h(Y|X)=h(Y)-h(X+Z|X)\\&=h(Y)-h(Z)=h(X+Z)-\frac{1}{2}\log2\pi eN\\&\leqslant\frac{1}{2}\log2\pi e(P+N)-\frac{1}{2}\log2\pi eN\\&=\frac{1}{2}\log\left(1+\frac{P}{N}\right)\end{aligned}$$

$$C=\max_{\mathbb{E}_XX^2\leqslant P}I(X;Y)=\frac{1}{2}\log\left(1+\frac{P}{N}\right)=\frac{1}{2}\log(1+\text{snr})$$

达到容量时X取正态分布，即$X\sim\mathcal{N}(0,P)$.

带宽有限信道(Bandlimited Channel)

根据采样定理，对带宽为W的信号，等价于对其以$2W$Hz采样得到的序列。如果高斯白噪声的功率谱密度为$\frac{N_0}{2}$ W/Hz，经过带宽为W的滤波后功率为N_0W W. 此时的带宽有限信道就转化成了功率不超过P，噪声功率为N_0W的高斯信道，且每秒传输$2W$次（这就改变了单位）。每次采样的信道容量是

$$C=\frac{1}{2}\log_2\left(1+\frac{P}{N_0W}\right)\text{bits/sample;}$$

每秒的信道容量就是

$$C=W\log_2\left(1+\frac{P}{N_0W}\right)\text{bits/s.}$$

如果带宽接近无限，则相应的信道容量为

$$C=\frac{P}{N_0}\log_2\mathrm{e}\,\text{bits/s.}$$

如果带宽很小，则相应的信道容量接近 0.

并联高斯信道(Parallel Gaussian Channel)

对每个高斯信道，其信道容量$C_i=\frac{1}{2}\log_2\left(1+\frac{P_i}{N_i}\right)$，其中噪声功率$N_i$固定，信号功率$P_i$可变且非负。为达到最大总容量，显然每个信道都需要独立以减少重复信息，所以总容量就是所有信道容量的加和。对每个信道容量求偏导得到$\frac{\partial C_i}{\partial P_i}=\frac{1}{2\ln 2}\frac{1}{N_i+P_i}$. 为使总信道容量达到最大，每个信道容量的偏导必须[10]相等（除非P_i出现负值），否则一定可以继续对偏导更大的信道增加功率并对偏导更小的信道减小功率以继续增加总信道容量。因此所有信道的噪声功率与信号功率之和相等，除非噪声过大。这在通信中称作注水(water-filling)法，即像注水一样给每个信道逐渐增加功率，始终保持每个信道的噪声与信号功率之和相等。

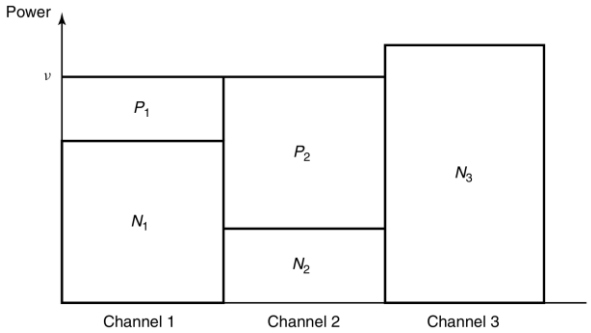

图 19 注水法

[10] 根据互信息的凹性，极大值一定就是最大值，所以相等是充要条件

率失真理论(Rate Distortion Theory)

率失真理论是研究码**率**和**失真**关系的理论。其定义本身不依赖信道（有随机性）的概念，而仅仅描述从信源X经过某种变换（无随机性）得到某个新的序列，再经译码（也无随机性）得到译码值（再生值）$\hat{X}$后，两者有多大的区别。这种理论可以用来衡量对模拟信号量化再复原的失真与量化位数的关系（模拟信号的量化永远无法完全精确）、对数据有损压缩再解压的损失与压缩比的关系，也可以用来衡量信道传输的失真与码率的关系（离散信道的信道容量就是刚好无失真的码率）。前面已经提到过，可达码率只与信道的最大互信息有关，而与信道具体形态、有多少随机性无关，所以信道的传输也可以用率失真理论描述。只要通过相应信道编码，信道传输就与数据压缩再解压无异。所以这里的信道都是一种“抽象信道”，其中的每一个信道都只给出了互信息的最大值，可以是一整类信道容量相同的信道，或者一整类最大互信息相同的压缩和解压的算法。

失真函数(Distortion)

$$d: \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^*$$

失真函数是从信源X的符号集和信源译码值$\hat{X}$的符号集映射到非负实数的函数。

对二进制信源，通常使用汉明距离(Hamming distance)表示失真：$d(x,\hat{x}) = x \oplus \hat{x}$；

对一般离散信源，可以使用克罗内克(Kronecker)记号表示失真，相等为0，不等为1：$d(x,\hat{x}) = 1 - \delta(x,\hat{x})$，也可以用失真矩阵直接规定每个信源和信源估计值的失真：$(d_{ij}) = (d(x_i,\hat{x}_j))$；

对连续信源，可以使用平方失真$d(x,\hat{x}) = (x-\hat{x})^2$，绝对失真$d(x,\hat{x}) = |x-\hat{x}|$和相对失真$d(x,\hat{x}) = \dfrac{|x-\hat{x}|}{|x|}$等。

x^n与$\hat{x}^n$序列间的失真为$d(x^n,\hat{x}^n) = \sum_{i=1}^{n} d(x_i,\hat{x}_i)$.

率失真码(Rate Distortion Code)

一个$(2^{nR},n)$率失真码包括一个编码函数$f_n: \mathcal{X}^n \to \{1,2,\cdots,2^{nR}\}$，一个译码(decoding)函数$g_n: \{1,2,\cdots,2^{nR}\} \to \hat{\mathcal{X}}^n$，其失真定义为每个码字失真的期望

$$\mathbb{E}\left[d\left(X^n, g_n(f_n(X^n))\right)\right] = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n)\, d\left(x^n, g_n(f_n(x^n))\right).$$

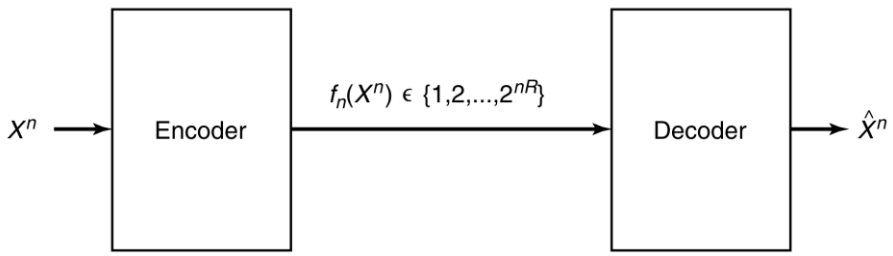

图 20 率失真码的结构

如果存在$(2^{nR},n)$率失真码序列(f_n,g_n)使得$\lim\limits_{n\to\infty}\mathbb{E}[d(X^n,g_n(f_n(X^n)))]\leqslant D$，则称率失真对(rate distortion pair)(R,D)可达。

信源的率失真区域是全体可达率失真对(R,D)组成的集合的闭包。

率失真函数是(R,D)中，每个失真D对应所有可行码率R的下确界；失真率函数是(R,D)中，每个码率R对应所有可行的失真D下确界。

量化(Quantization)

量化是用一系列整数的编码来表示实数，也就是连续的随机变量。在给定随机变量的分布后，求得给定码率下的最佳量化方式，也就是率失真理论的一种应用场景。

如果随机变量X的概率密度函数是$f(x)$，量化就是把其定义域分为n段，并给每一段赋予一个“量化值”；如果X的取值刚好落在这一段中，就用这个“量化值”来代替真实的X值。具体的“量化值”和“量化边界”都可以提前约定好，最终需要传输的就只有每一段的编号，也就是用n种可能，或者说$\log_2 n$ bit来表示一个连续的随机变量。

设n段的编号分别为$\{0,1,\cdots,n-1\}$，第i段的范围为$[a_i,a_{i+1}]$，量化值为b_i. 用均方误差表示的失真就是

$$D(\boldsymbol{a},\boldsymbol{b})=\sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}}f(x)\,(x-b_i)^2\mathrm{d}x$$

为使该失真达到最小，一个必要条件就是$\dfrac{\partial D}{\partial \boldsymbol{a}}=\boldsymbol{0}$, $\dfrac{\partial D}{\partial \boldsymbol{b}}=\boldsymbol{0}$，化简得

$$\begin{cases}b_i=\displaystyle\int_{a_i}^{a_{i+1}}xf(x)\mathrm{d}x\\ a_i=\dfrac{b_{i-1}+b_i}{2}\end{cases}$$

该条件也被称为最优量化的 Lloyd-Max 定理，即量化边界是相邻量化值的中点，量化值是量化区间的重心。这难以求出解析表达式，但可以快速迭代求出数值解：先生成一个随机的$\boldsymbol{a}$，代入得到相应的$\boldsymbol{b}$，再重新更新$\boldsymbol{a}$，直到两轮迭代的误差足够小。

信息率失真函数(Information Rate Distortion Function)

$$R^{(I)}(D)=\min_{p_{\hat{X}|X}:\mathbb{E}\left[d\left(X,\hat{X}\right)\right]\leqslant D}I\left(X;\hat{X}\right)$$

信息率失真函数是在限制最大失真的条件下，信源和译码值互信息的最小值。其中信源的分布$p_X(x)$是给定的，不可改变；其互信息和失真的自由度来源于$p_{\hat{X}|X}(\hat{x}|x)$的变化。

上文提到，最大互信息相等的信道都等价，而不用考虑信道的具体形态，所以率失真的定义里除了信源本身的分布外，没有用到任何概率。但信息率失真的定义里却出现了概率：可变的是$p_{\hat{X}|X}(\hat{x}|x)$，即X到$\hat{X}$并不是一个确定的映射，而有一定的概率。为什么这里要用到概率？因为信息率失真的定义里只考虑了单个信源和单个恢复值的关系，而没有取足够长度的信源序列和恢复值序列，对单个离散信源，考虑到离散失真的定义，恢复值的

集合通常与信源值相同，例如信源有 0 和 1，恢复值也将只有 0 和 1，失真函数的取值也只有 0 和 1，而不可能连续改变。失真如果不靠概率体现出来，就无法连续变化，例如对信源的 0,1,2 引入非零失真，如果把恢复值减少，变成 0,1, 这就需要把 0 映射到 0，把 1 和 2 映射到 1, 或者改变顺序。这样显然只能产生离散的率失真函数。为了使之连续变化，必须引入概率。对单个连续信源则没有这种限制，如果是进行量化，则可以改变量化区间，也可以随意改变恢复值的集合，这时哪怕不允许出现随机性也无妨，$p_{\hat{X}|X}(\hat{x}|x)$可以只取 0 和 1. 如果不是单个信源，而是足够长的离散信源序列和恢复值序列，则可以进一步编码和解码，在完全确定的情况下改变其中的错误个数（例如直接把几个字符置 0），使得失真仍然连续变化。总之，信息率失真函数里出现的概率只是一种表示方法，是不得不使单个离散信源产生连续失真而做出的妥协，其实并不一定需要产生随机性，还可以是长离散信源序列的错误个数占比，或者单个连续信源的量化区间和量化值。这种概率表示可以兼容离散信源序列和连续信源，所以引入概率。定义里真正重要的是最优化的目标函数互信息和约束条件平均失真，而不是这里的概率。

为什么在定义信息信道容量时，用的是互信息的最大值，这里用的却是最小值？因为信息信道容量是给定具体的信道后，该信道自身的“能力(capacity)”，只要能达到，自然是越大越好；信息率失真则是给定失真下，对信道的最低要求，这是在挑选信道，而不是已经选定了一个信道。出于减少浪费和增加选择的目的，对信道的需求越低越好，没有达到最小值就必然产生冗余。而且选择信道的标准既然是互信息，那必然已经达到了信道容量，即互信息的最大值，所以信息率失真函数是在一系列互信息最大值当中找最小值，即

$$R^{(I)}(D)=\min_{W:\mathbb{E}\left[d\left(X,\psi\circ h\left(W_{Y|\varphi(X)}(\varphi(X))\right)\right)\right]\leq D}\left[\max_{p_{\varphi(X)}(\varphi(x))}\left[I\left(X;W_{Y|\varphi(X)}(\varphi(X))\right)\right]\right].$$

该公式说明了取最小值和取最大值不矛盾，其中W指一系列信道，φ指把原X转换为信道输入的函数，因为分布原本确定的X必须先被转换成信道可接受的输入字符集，且还要能够根据需求改变每种字符出现的概率，否则信道无法达到真正的信息信道容量。h为由信道输出Y对输入$\varphi(X)$的最优估计，ψ为把$\varphi(X)$估计值转换回$\hat{X}$的函数。

原先信息率失真定义中可变的条件概率$p_{\hat{X}|X}(\hat{x}|x)$对应到此处的具体信道W，就相当于可变的信道$W_{Y|\varphi(X)}\left(h^{-1}\circ\psi^{-1}\left(\hat{X}\right)|\varphi(X)\right)$. 这种把失真和信道分开考虑的方法，由带失真的信源信道分离定理(Source–channel separation theorem with distortion)具体给出。

率失真定理(Rate Distortion Theorem)

对独立同分布的信源X，若分布为$p_X(x)$且失真函数$d(x,\hat{x})$有界，那么其率失真函数与对应的信息率失真函数相等。

于是，$R(D)=R^{(I)}(D)=\min_{p_{\hat{X}|X}:\mathbb{E}\left[d\left(X,\hat{X}\right)\right]\leq D} I\left(X;\hat{X}\right)$为失真$D$下的最小可达码率；反之，如果采用低于$R(D)$的码率来描述$X$，则不能达到比$D$小的失真。

这里不给出率失真定理的严格证明，但给出率失真定理的直观理解。

上文提到，信息率失真函数定义中的概率完全可以转化成从足够长序列X^n到恢复值$\hat{X}^n$的映射，用字符比例代替概率，这样除了信源本身的随机性以外，整个过程就没有了任何随机性。由这种确定性可得，

$$I\left(X;\hat{X}\right)=H\left(\hat{X}\right)-H\left(\hat{X}|X\right)=H\left(\hat{X}\right)$$

给定失真后有很多种$\hat{X}^n$，在该失真下需要的码率就是传输所需码率最低的$\hat{X}^n$，也就是熵最小的$\hat{X}^n$，也就是X^n和$\hat{X}^n$的最小互信息。

二元信源(Binary Source)

对二元信源，设$\mathbb{P}(X=0)=p$，则其在汉明度量下的率失真函数为

$$R(D)=\begin{cases}H(p)-H(D), & 0\leqslant D\leqslant \min\{p,1-p\}\\ 0, & D>\min\{p,1-p\}\end{cases}$$

证明该结论需要先给出互信息的下界，再证明该下界可达。

首先给出下界：

$$\begin{aligned}I\left(X;\hat{X}\right)&=H(X)-H\left(X|\hat{X}\right)\\&=H(p)-H\left(X\oplus\hat{X}|\hat{X}\right)\\&\geqslant H(p)-H\left(X\oplus\hat{X}\right)\\&=H(p)-H(D)\end{aligned}$$

证明可达可给出一个具体的$\left(X,\hat{X}\right)$联合分布，可以看成一个“反向测试信道”，如图21，可以验证该联合分布满足要求。

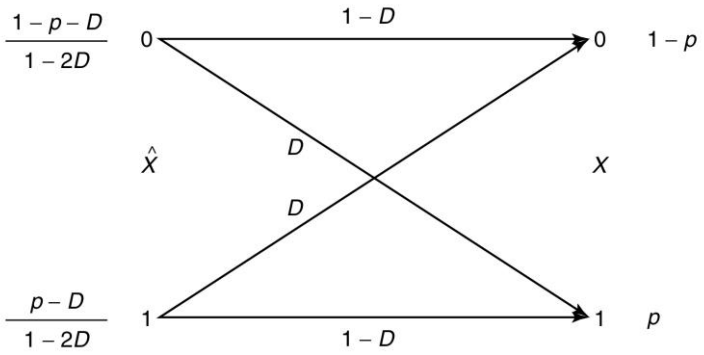

图 21 反向测试信道

接下来说明如何得到该分布。观察不等号的来源，可以发现，我们只需要使得不等式$H\left(X\oplus\hat{X}|\hat{X}\right)\geqslant H\left(X\oplus\hat{X}\right)$取等，即$X\oplus\hat{X}$独立于$\hat{X}$，所以构造从$\hat{X}$到$X$的映射，使得两个转移概率相等，也就是都为$D$.

如果不构造逆向的映射，而是直接正向思考，仍然可以得到相同的结论，但计算量显著增加。设$p_{\hat{X}|X}(0|0)=r$，$p_{\hat{X}|X}(1|1)=s$，则联合分布如表 2.

表 2 X与$\hat{X}$的联合分布

$p_{X\hat{X}}\left(X,\hat{X}\right)$	$\hat{X}=0$	$\hat{X}=1$	$p_X(X)$
$X=0$	pr	$p(1-r)$	p
$X=1$	$(1-p)(1-s)$	$(1-p)s$	$1-p$
$p_{\hat{X}}\left(\hat{X}\right)$	$pr+(1-p)(1-s)$	$p(1-r)+(1-p)s$	1

不等号取等，则等价于$X\oplus\hat{X}$与$\hat{X}$独立，即

$$\mathbb{P}\left[X\oplus\hat{X}=0\,|\hat{X}=0\right]=\mathbb{P}\left[X\oplus\hat{X}=0\,|\hat{X}=1\right]$$

$$\frac{pr}{pr+(1-p)(1-s)}=\frac{(1-p)s}{p(1-r)+(1-p)s}$$

$$\left(\frac{p}{1-p}\right)^2 r(1-r)=s(1-s)$$

$$\frac{\left(s-\frac{1}{2}\right)^2}{\frac{1}{4}\left(1-\left(\frac{p}{1-p}\right)^2\right)}-\frac{\left(r-\frac{1}{2}\right)^2}{\frac{1}{4}\left(\left(\frac{1-p}{p}\right)^2-1\right)}=1$$

这是一个中心为$\left(\frac{1}{2},\frac{1}{2}\right)$，且经过$(0,0)\,(0,1)\,(1,0)\,(1,1)$共 4 个点的双曲线方程，显然在$[0,1]\times[0,1]$有无穷多组解$(s,r)$. 相应地，失真$D=1-(1-p)s-pr$可取到的范围有$[0,\min\{p,1-p\}]\cup[\max\{p,1-p\},1]$. 这也符合结论，因为当$D>\max\{p,1-p\}$时，在这种联合分布下可以直接交换$\hat{X}$的 0 和 1，以重新并入$D<\min\{p,1-p\}$的情形。真正只需要$D>\min\{p,1-p\}$时，可以直接不传输，而直接在$\hat{X}$端取一个确定的值，毕竟这样已经足以满足失真要求。

r和s由一个双曲线方程约束，D又可以用r和s表示，所以可以使用D来唯一确定r和s. 代入方程得，在$D<\min\{p,1-p\}$时，有

$$r=\frac{(1-p-D)(1-D)}{(1-2D)(1-p)},s=\frac{(p-D)(1-D)}{(1-2D)p},$$

这就得到了与“反向测试信道”相同的联合分布.

图 22 为$p=\frac{1}{2}$时的率失真函数$R(D)$.

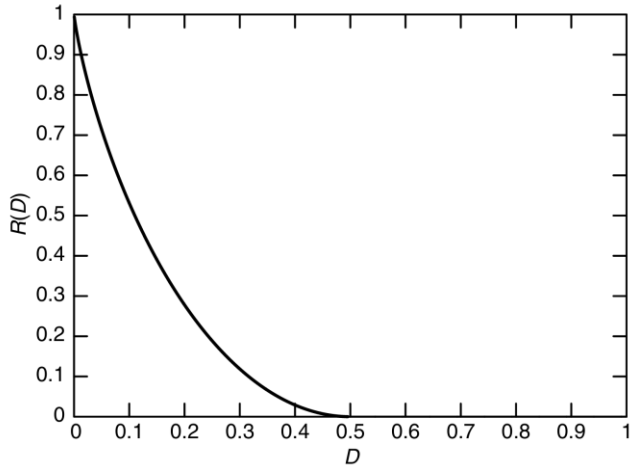

图 22 二元信源的率失真函数

高斯信源(Gaussian Source)

一个$\mathcal{N}(0,\sigma^2)$信源在平方误差失真度量下的率失真函数为

$$R(D)=\begin{cases}\frac{1}{2}\log\frac{\sigma^2}{D}, & 0\leqslant D\leqslant\sigma^2\\ 0, & D>\sigma^2\end{cases}$$

$$\begin{aligned}I\left(X;\hat{X}\right)&=h(X)-h\left(X|\hat{X}\right)\\&=\frac{1}{2}\log 2\pi\mathrm{e}\sigma^2-h\left(X-\hat{X}|\hat{X}\right)\\&\geqslant\frac{1}{2}\log 2\pi\mathrm{e}\sigma^2-h\left(X-\hat{X}\right)\\&\geqslant\frac{1}{2}\log 2\pi\mathrm{e}\sigma^2-\frac{1}{2}\log\left[2\pi\mathrm{e}\mathbb{E}\left[\left(X-\hat{X}\right)^2\right]\right]\\&=\frac{1}{2}\log 2\pi\mathrm{e}\sigma^2-\frac{1}{2}\log 2\pi\mathrm{e}D\\&=\frac{1}{2}\log\frac{\sigma^2}{D}\end{aligned}$$

为了证明该下界可达，同样可以给出一个具体的“测试信道”。

要使得第一个不等号取等，即$h\left(X-\hat{X}|\hat{X}\right)=h\left(X-\hat{X}\right)$，就需要使得$X-\hat{X}$与$\hat{X}$独立，所以同样构造从$\hat{X}$到$X$的反向加性噪声信道。

要使得第二个不等号取等，即$h\left(X-\hat{X}\right)=\frac{1}{2}\log\left[2\pi\mathrm{e}\mathbb{E}\left[\left(X-\hat{X}\right)^2\right]\right]$，就需要使得噪声为高斯噪声。

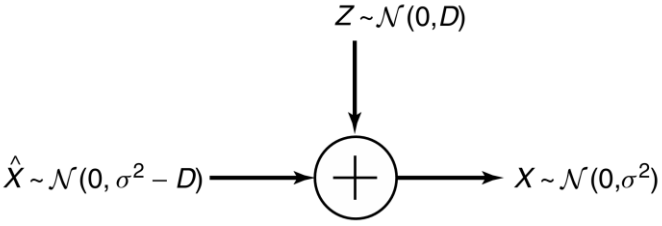

图 23 反向测试信道

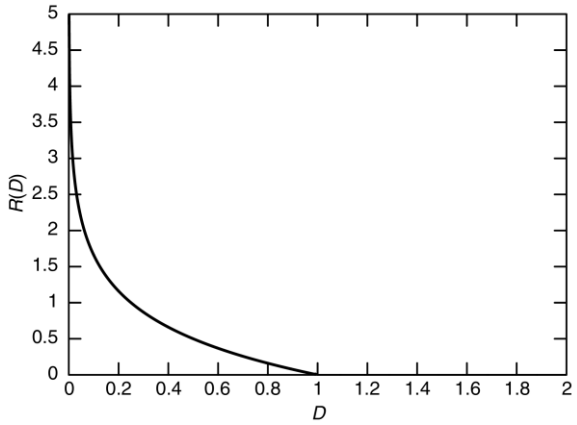

图 24 高斯信源的率失真函数

综上所述，构造如图 23 所示的反向测试信道，可以验证其满足等式。同样也可以用贝叶斯公式转换回常见的正向信道，只是更复杂且缺乏直观。

图 24 为$\sigma^2=1$时的率失真函数$R(D)$. 因为模拟信号永远无法完全复原，所以$D=0$时的码率趋近于无穷大。

独立高斯随机变量(Independent Gaussian Random Variables)

若有m个独立但不同分布的正态随机信源$X_1,X_2,\cdots,X_m$，其中$X_i\sim\mathcal{N}(0,\sigma_i^2)$，使用平方误差失真。若总失真$D=\sum_{i=0}^{m}D_i$的上限固定，应如何分配各个失真使得总码率需求最低？

注意到，对每个信源X_i，其率失真函数为$R_i(D_i)=\frac{1}{2}\log\frac{\sigma_i^2}{D_i}$. 总码率达到最小值的充要条件是每个码率对失真的偏导相等。又因为每个率失真函数对失真求偏导得到的都是相同的结果，与σ_i^2无关，所以除非σ_i^2足够小，不足以使得码率为正，否则就应该给每个信源分配相等的失真，即

$$D_i=\begin{cases}\lambda, & \lambda<\sigma_i^2\\ \sigma_i^2, & \lambda\geq\sigma_i^2\end{cases}$$

其中λ的选取满足$\sum_{i=1}^{m}D_i=D$.

如图 25，这被称作反注水法，即只考虑自身方差比较大的随机变量，而忽略自身方差足够小的随机变量。

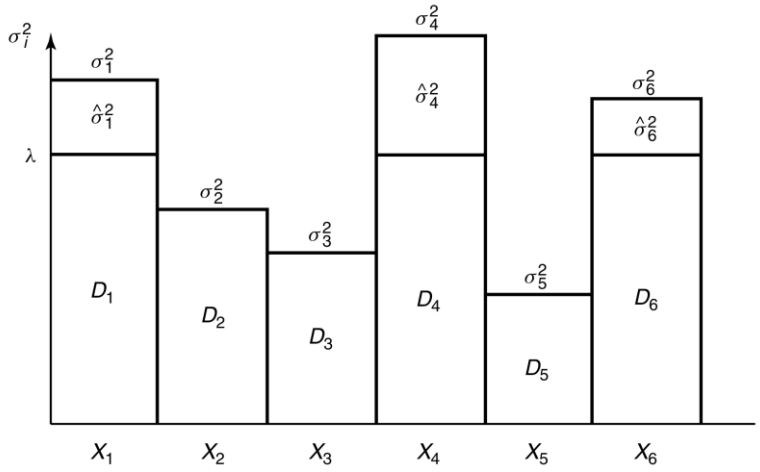

图 25 反注水法

信道编码(Channel Coding)

信道编码定理表明，对任意信道，都存在一种信道编码，使得在码率等于信道容量的情况下，误码率无限小。这种编码虽然存在，但实际应用中往往难以实现，因为对信源的分布有很高要求，还需要相当大的长度。所以这里介绍的信道编码并不是信道编码定理提

到的最优编码，而是工程中使用的具有一定检错能力的编码。

线性分组码(Linear Block Code)

(n,k)码在编码前的信息有k位，编码后共有n位码字。其编码效率为$R_c=\frac{k}{n}$.

(n,k)码的编码是一个从k维k重空间到n维k重码空间C的线性变换。写成矩阵形式是$\boldsymbol{c}_{1\times n}=\boldsymbol{m}_{1\times k}\boldsymbol{G}_{k\times n}$，其中$\boldsymbol{m}_{1\times k}$为编码前的信息，$\boldsymbol{c}_{1\times n}$位编码后的码字，$\boldsymbol{G}_{k\times n}$为生成矩阵，由$C$的基底组成。为了确保有$k$重，生成矩阵$\boldsymbol{G}_{k\times n}$的每行必须线性无关，否则编码之后反而丢失了信息。在线性无关的条件下，每行仍然有不同取法，有可能对应相同的码集。任何生成矩阵都能经过初等行变换和列置换得到一个系统形式。码集相同的系统码对应的生成矩阵等价，可以得到相同的系统形式$\boldsymbol{G}=[\boldsymbol{I}_k\vdots\boldsymbol{P}]$，其对应码字前$k$位与编码前信息保持不变，称为信息位，后$n-k$位为校验位。

校验矩阵$\boldsymbol{H}=[\boldsymbol{P}^\top\vdots\boldsymbol{I}_{n-k}]$由$n$维$k$重空间$C$的对偶空间$H$的基底组成。如果无误码，则$\boldsymbol{c}\boldsymbol{H}^\top=\boldsymbol{0}$；如果误码为$\boldsymbol{e}$，则$\boldsymbol{R}\boldsymbol{H}^\top=(\boldsymbol{c}+\boldsymbol{e})\boldsymbol{H}^\top=\boldsymbol{e}\boldsymbol{H}^\top=\boldsymbol{S}$. 伴随式$\boldsymbol{S}$通常不为$\boldsymbol{0}$，除非$\boldsymbol{e}$刚好也落在码空间$C$中。$n$位码字中，至少错几位才有可能使得$\boldsymbol{S}=\boldsymbol{e}\boldsymbol{H}^\top=\boldsymbol{0}$，刚好看不出错误？这就是线性分组码最小距离$d_{\min}$的定义。相应的检错能力为$d_{\min}-1$，也就是即使出现$d_{\min}-1$个错误，$\boldsymbol{S}$也可以不为$\boldsymbol{0}$，即误码也不会落在码空间里。纠错能力要求不仅检查出错误，还要得出错在哪，即确定$\boldsymbol{e}$的值。同一个$\boldsymbol{S}$可以对应多个$\boldsymbol{e}$，对此只能相信更少误码出现的概率最大，所以选择错误较少的那个。因为对多种可能的错误，我们只能选择误码数量较少的那个，所以纠错能力只有检错能力的一半并取整。