信息论

Leo Yan

2026 年 2 月 9 日

目录

第一章　熵，相对熵，互信息

1.1　基本概念

Definition 1.1.1 (熵). 设X是离散型随机变量，取值（样本空间）为$\mathcal{X}$，概率分布为$p(x)$，则X的熵(entropy)定义为

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}_p\left[\log\frac{1}{p(X)}\right]$$

单位为比特（bit）；若以自然对数为底，则单位为纳特（nat）

$$H(X)\geq 0;\quad H_b(X) = \log_b a\cdot H_a(X)$$

也记为$H(p)$

Remark 1.1.1. 熵描述随机变量的不确定度；给出了描述随机变量所需的信息量的下界

Definition 1.1.2 (联合熵). 设(X,Y)为联合分布的离散型随机变量，则(X,Y)的联合熵(joint entropy)定义为

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = \mathbb{E}_p\left[\log\frac{1}{p(X,Y)}\right]$$

也记为$H(XY)$

Definition 1.1.3 (条件熵). 设(X,Y)为联合分布的离散型随机变量，则在已知Y的条件下X的条件熵(conditional entropy)定义为

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) = \mathbb{E}_p\left[\log\frac{1}{p(X|Y)}\right]$$

Theorem 1.1.1 (链式法则). 设(X,Y)为联合分布的离散型随机变量，则有

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

Proof.

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

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

□

Definition 1.1.4 (相对熵). 设p和q为定义在同一样本空间$\mathcal{X}$上的两个离散概率分布，则p相对于q的相对熵(relative entropy)或Kullback-Leibler散度(KL散度)定义为

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right]$$

Remark 1.1.2. 描述两个概率分布之间的差异；真实分布为p，假设分布为q的无效性

Definition 1.1.5 (互信息). 设(X,Y)为联合分布的离散型随机变量，则X和Y的互信息(mutual information)定义为

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \mathbb{E}_p\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right]$$

Remark 1.1.3. 描述两个随机变量之间共享的信息量，或X包含Y的信息量

Proposition 1.1.2 (互信息的性质). *设(X,Y)为联合分布的离散型随机变量，则有*

1. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
2. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
3. $I(X;X) = H(X)$

Remark 1.1.4. 以上三条性质的直观：

- 给定X，Y的不确定性减少了$I(X;Y)$
- 容斥原理
- 熵又是自信息(self-information)

1.2 多维随机变量

符号说明

(1)$H(X,Y,Z)$实际上应该理解为$H((X,Y,Z))$，即H总为一元函数

(2)$I(X;Y,Z)$实际上应该理解为$I(X;(Y,Z))$，即$I(\cdot;\cdot)$总为二元函数

(3)$H(X,Y|Z)$实际上应该理解为$H((X,Y)|Z)$，即$H(\cdot|\cdot)$总为二元函数；同理$I(X;Y|Z)$实际上应该理解为$I(X;Y|Z)$，即$I(\cdot;\cdot|\cdot)$总为三元函数。应该认为“;”的优先级高于“|”

(4)D虽称为熵，但不是随机变量的函数，而是分布的函数。”||”类似于“;”，D总为二元函数

Definition 1.2.1 (条件互信息). X,Y在已知Z的条件下的条件互信息(conditional mutual information)定义为

$$I(X;Y|Z)=\sum_{z\in\mathcal{Z}}p(z)\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}=\mathbb{E}_p\left[\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]$$

Remark 1.2.1. 描述已知Z，给出Y引起的X的不确定度减少量

Definition 1.2.2 (条件相对熵). 对于联合概率质量函数$p(x,y)$和$q(x,y)$，X在已知Y的条件下的条件相对熵(conditional relative entropy)定义为

$$D(p(x|y)||q(x|y))=\sum_{y\in\mathcal{Y}}p(y)\sum_{x\in\mathcal{X}}p(x|y)\log\frac{p(x|y)}{q(x|y)}=\mathbb{E}_p\left[\log\frac{p(X|Y)}{q(X|Y)}\right]$$

Theorem 1.2.1 (熵的链式法则).

$$H(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}H(X_i|X_1,X_2,\ldots,X_{i-1})$$

Proof. 反复使用

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

□

Theorem 1.2.2 (互信息的链式法则).

$$I(X;Y_1,Y_2,\ldots,Y_n)=\sum_{i=1}^{n}I(X;Y_i|Y_1,Y_2,\ldots,Y_{i-1})$$

Proof.

$$I(\mathbf{X};\mathbf{Y})=H(\mathbf{Y})-H(\mathbf{Y}|\mathbf{X})$$

□

Theorem 1.2.3 (相对熵的链式法则).

$$D(p(x,y)||q(x,y))=D(p(x)||q(x))+D(p(y|x)||q(y|x))$$

Proof.

$$\begin{aligned}D(p(x,y)||q(x,y))&=\mathbb{E}_{p(x,y)}\log\frac{p(X,Y)}{q(X,Y)}\\&=\mathbb{E}_{p(x,y)}\log\frac{p(X)p(Y|X)}{q(X)q(Y|X)}\\&=\mathbb{E}_{p(x)}\log\frac{p(X)}{q(X)}+\mathbb{E}_{p(x,y)}\log\frac{p(Y|X)}{q(Y|X)}\\&=D(p(x)||q(x))+D(p(y|x)||q(y|x))\end{aligned}$$

□

1.3 不等式

Theorem 1.3.1 (Jensen不等式). 设X为离散型随机变量，取值（样本空间）为$\mathcal{X}$，概率分布为$p(x)$，且f为凸函数，则有

$$\mathbb{E}_p[f(X)] \geq f(\mathbb{E}_p[X])$$

进一步，若f为严格凸函数，则等号成立 $\iff X = \mathbb{E}X$

Theorem 1.3.2 (信息不等式). 设p和q为定义在同一样本空间$\mathcal{X}$上的两个离散概率分布，则有

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

取等 $\iff p = q$

Proof. log在$(0, +\infty)$上为严格凹函数，故$-\log$为严格凸。考虑

$$A = \{x \in \mathcal{X} : p(x) > 0, q(x) > 0\}$$

则由Jensen不等式，

$$\begin{aligned} D(p||q) &= \sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \\ &= -\sum_{x\in A} p(x) \left[-\log \frac{q(x)}{p(x)}\right] \\ &\geq -\log\left(\sum_{x\in A} p(x)\frac{q(x)}{p(x)}\right) \\ &= -\log\left(\sum_{x\in A} q(x)\right) \\ &\geq 0 \end{aligned}$$

□

Corollary 1.3.3.

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

取等 $\iff X$与Y独立

条件相对熵、条件互信息也是非负的

Corollary 1.3.4.

$$H(X|Y) \leq H(X)$$

取等 $\iff X$与Y独立

Remark 1.3.1. 条件导致熵减小。但$H(X|Y = y)$可能大于$H(X)$，不等式仅描述平均性质

Theorem 1.3.5.

$$H(X) \leq \log |\mathcal{X}|$$

取等 $\Longleftrightarrow$ $X \sim U_{\mathcal{X}}$

Corollary 1.3.6 (熵的独立界)**.**

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

取等 $\Longleftrightarrow$ $X_1, X_2, \ldots, X_n$相互独立

Proof. 由链式法则，

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, X_2, \ldots, X_{i-1}) \leq \sum_{i=1}^{n} H(X_i)$$

□

1.4 其他背景下的不等式

Definition 1.4.1. 如果Z的条件分布$p(z|x, y)$仅依赖于y而与x条件独立，即

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

则称随机变量三元组(X,Y,Z)构成马尔可夫链(Markov chain)，记为$X \to Y \to Z$
条件独立的意思是

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

X,Z对称，即$X \to Y \to Z \iff Z \to Y \to X$，因此可记$X \leftrightarrow Y \leftrightarrow Z$

Theorem 1.4.1 (数据处理不等式)**.** 设随机变量三元组*(X,Y,Z)*构成马尔可夫链$X \to Y \to Z$，则有

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

取等 $\Longleftrightarrow$ $X \to Z \to Y$

Proof. 用互信息的链式法则，

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

给定Y，X与Z独立，即$I(X; Z|Y) = 0$，而$I(X; Y|Z) \geq 0$，故

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

□

Remark 1.4.1. 对Y的数据处理不能增加其包含X的信息

Corollary 1.4.2.

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

$$X \to Y \to Z \implies I(X;Y|Z) \leq I(X;Y)$$

Remark 1.4.2. 观察下游随机变量，X,Y的依赖程度可能降低

Definition 1.4.2 (充分统计量). 假定有一族概率质量函数$\{f_\theta(x)\}$，X是从其中一个分布$f_\theta(x)$中抽取的样本，T(X)是X的一个统计量，则

$$\theta \to X \to T(X)$$

由数据处理不等式，有

$$I(\theta;X) \geq I(\theta;T(X))$$

取等时统计量T(X)未损失X关于参数θ的信息，称T(X)为关于分布族$\{f_\theta(x)\}$的充分统计量(sufficient statistic)
等价定义：给定T(X)，X与θ条件独立，即$\theta \to T(X) \to X$

Definition 1.4.3 (最小充分统计量). 关于分布族$\{f_\theta(x)\}$的充分统计量T(X)是其他任何充分统计量$U(X)$的函数，则称T(X)为关于$\{f_\theta(x)\}$的最小充分统计量(minimal sufficient statistic)
定义蕴含$\theta \to T(X) \to U(X) \to X$

Remark 1.4.3. 最小充分统计量最大程度地压缩了样本X中关于θ的信息，而其他充分统计量可能包含冗余信息

由随机变量Y估计与之有关的X，X的估计值记为$\hat{X} = g(Y)$取值空间为$\hat{\mathcal{X}}$，则有马尔可夫链$X \to Y \to \hat{X}$。定义误差概率$P_e = P\{\hat{X} \neq X\}$。

Theorem 1.4.3 (Fano不等式). 设X为离散型随机变量，取值（样本空间）为$\mathcal{X}$，则有

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

可减弱为

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

Proof. 设错误指示变量$E = \mathbb{1}_{\{\hat{X} \neq X\}}$，则有马尔可夫链$X \to Y \to \hat{X} \to E$。由链式法则，

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

因为$H(E|X, \hat{X}) = 0$，所以

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X})$$

注意到

$$H(E|\hat{X}) \leq H(E) = H(P_e)$$

且

$$H(X|E,\hat{X}) = P_e H(X|\hat{X}, E=1) + (1-P_e)H(X|\hat{X}, E=0) \leq P_e \log(|\mathcal{X}|-1)$$

故

$$H(X|\hat{X}) \leq H(P_e) + P_e \log(|\mathcal{X}|-1)$$

另一方面，由数据处理不等式，

$$H(X|Y) \leq H(X|\hat{X})$$

□

Corollary 1.4.4. 令$\hat{X} = Y$，则有

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

若$\hat{X} = X$，结论变为

$$H(P_e) + P_e \log(|\mathcal{X}|-1) \geq 0$$

Proposition 1.4.5. 设X,X'独立同分布，则

$$P\{X = X'\} \geq 2^{-H(X)}$$

Corollary 1.4.6. 设X,Y独立，$X \sim p(x)$，$Y \sim q(y)$，取值空间均为$\mathcal{X}$，则

$$P\{X = Y\} \geq 2^{-H(p)-D(p\|q)}, P\{X = Y\} \geq 2^{-H(q)-D(q\|p)}$$

第二章　渐进均分性

Definition 2.0.1 (随机变量的收敛). 给定随机变量序列$\{X_n\}$和随机变量X，
①如果$\forall \epsilon > 0, P\{|X_n - X| \geq \epsilon\} \to 0 (n \to \infty)$，则称$X_n$依概率收敛于X，记为$X_n \xrightarrow{P} X$
②如果$P\{\lim_{n\to\infty} X_n = X\} = 1$，则称$X_n$几乎处处收敛（或以概率1收敛）于X，记为$X_n \xrightarrow{a.e.} X$
③如果$\mathbb{E}(X_n - X)^2 \to 0 (n \to \infty)$，则称$X_n$均方收敛于X，记为$X_n \xrightarrow{L_2} X$

Theorem 2.0.1 (渐进均分性(Asymptotic Equipartition Property, AEP)). 以X记信源随机变量，它生成的序列$X_1, X_2, \ldots, X_n i.i.d. \sim p(x)$，则有

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{P} H(X)$$

Proof.

$$X_k \ i.i.d. \sim p(x) \implies -\log p(X_k) \ i.i.d. \sim -\log p(x)$$

由弱大数定律，

$$\frac{1}{n}\sum_{k=1}^{n} -\log p(X_k) \xrightarrow{P} \mathbb{E}[-\log p(X)] = H(X)$$

□

Definition 2.0.2 (典型集). 关于p(x)的典型集(typical set)定义为

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) \middle| -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) - H(X) < \epsilon\right\}$$

性质：
①

$$\forall \mathbf{x} \in A_\epsilon^{(n)}, H(X) - \epsilon < -\frac{1}{n}\log p(\mathbf{x}) < H(X) + \epsilon$$

② n充分大时，$P\{A_\epsilon^{(n)}\} > 1 - \epsilon$
③ $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$（概率和不超过1）
④ （第一条的推论）n充分大时，$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$

Remark 2.0.1. 直观：
① $\implies$ 典型集中的元素在数量级意义下是几乎等可能的；
② $\implies$ 典型集出现概率随n增大而趋近于1（渐进）；
③④ $\implies$ 典型集的元素个数近似等于$2^{nH(X)}$（均分）

设X_n $i.i.d.\sim p(x)$，存在一个编码将长为n的序列映射为比特串，且映射为双射（从而可逆），其码字长度$l(x_n)$满足

$$\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}l(x_n)\right]=H(X)$$

因而理论上用$nH(X)$比特即可表示序列$x_1,x_2,\ldots,x_n$（等长码需要$n\log|\mathcal{X}|$比特）

码字长度：信源编码时某个符号x使用的比特数为$l(x)$；x^n使用的比特数为$l(x^n)$

Proof. 考虑给大概率的典型集较短的编码。

$$A_\epsilon^{(n)}\le 2^{n(H(X)+\epsilon)}\implies\lceil n(H(X)+\epsilon)\rceil\le n(H(X)+\epsilon)+1(bit)$$

可表示$A_\epsilon^{(n)}$中的每个序列，同理$n\log|\mathcal{X}|+1$bit可表示$A_\epsilon^{(n)c}$

在典型集序列前标0而在非典型集序列前标1作为表示为，则码字长度

$$l(x^n)\le\begin{cases}n(H(X)+\epsilon)+2, & x^n\in A_\epsilon^{(n)}\\ n\log|\mathcal{X}|+2, & x^n\in A_\epsilon^{(n)c}\end{cases}$$

取n充分大使$P\{A_\epsilon^{(n)}\}>1-\epsilon$，则

$$\begin{aligned}\mathbb{E}[l(x^n)]&=\sum_{x^n\in\mathcal{X}^n}p(x^n)l(x^n)\\&\le P\{A_\epsilon^{(n)}\}[n(H(X)+\epsilon)+2]+P\{A_\epsilon^{(n)c}\}[n\log|\mathcal{X}|+2]\\&\le(1-\epsilon)[n(H(X)+\epsilon)+2]+\epsilon[n\log|\mathcal{X}|+2]=n(H(X)+\epsilon')\end{aligned}$$

其中$\epsilon'=\epsilon\log|\mathcal{X}|-\epsilon H(X)+2/n$

$$\begin{aligned}\mathbb{E}[l(x^n)]&\ge P\{A_\epsilon^{(n)}\}[n(H(X)+\epsilon)+1]+P\{A_\epsilon^{(n)c}\}[n\log|\mathcal{X}|+1]\\&\ge(1-\epsilon)[n(H(X)+\epsilon)+1]+\epsilon[n\log|\mathcal{X}|+1]=n[(1-\epsilon)H(X)+\epsilon'']\end{aligned}$$

其中$\epsilon''=\epsilon(1-\epsilon)+\frac{1-\epsilon}{n}$ □

Remark 2.0.2. 熵是无损压缩的下限

第三章　熵率与Markov链

3.1　熵率

Definition 3.1.1 (熵率). 设$\{X_n\}$为随机过程，则$\{X_n\}$的熵率(entropy rate)在极限存在时定义为

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

Theorem 3.1.1. 定义

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_1, X_2, \ldots, X_{n-1})$$

对于平稳过程，两种极限均存在且$H(\mathcal{X}) = H'(\mathcal{X})$

Proof.

$$H(X_{n+1}|X_1, X_2, \ldots, X_n) \le H(X_{n+1}|X_2, X_3, \ldots, X_{n+1}) = H(X_n|X_1, X_2, \ldots, X_{n-1})$$

因此$H(X_n|X_1, X_2, \ldots, X_{n-1})$非负递减，$H'(\mathcal{X})$存在。

由链式法则，

$$\frac{1}{n} H(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_1, X_2, \ldots, X_{i-1})$$

即熵率为条件熵的时间平均值。$H_n(\mathcal{X})$为$H'(\mathcal{X})$的Cesàro和，故$H(\mathcal{X}) = H'(\mathcal{X})$ □

3.2　Markov链

Theorem 3.2.1. 对于平稳的*Markov*链$\{X_n\}$，设平稳分布为μ，转移概率矩阵为P，则熵率为

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i \sum_j P_{ij} \log \frac{1}{P_{ij}}$$

Proof.

$$\begin{aligned} H'(\mathcal{X}) &= \lim_{n\to\infty} H(X_n|X_1, X_2, \ldots, X_{n-1}) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}) \\ &= H(X_2|X_1) \end{aligned}$$

$$= \sum_i \mu_i \sum_j P_{ij} \log \frac{1}{P_{ij}}$$

□

Theorem 3.2.2. 设μ_n, μ'_n为n时刻同一*Markov*链的两条轨迹的分布，则

$$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}||\mu'_{n+1})$$

特别地，

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu)$$

Proof. 记μ_n, μ'_n对应两盒概率分布为p, q，$r(\cdot|\cdot)$为转移概率分布，则

$$p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n), q(x_n, x_{n+1}) = q(x_n)r(x_{n+1}|x_n)$$

由相对熵的链式法则，

$$\begin{aligned} & D(p(x_n)||q(x_n)) \\ =& D(p(x_n)||q(x_n)) + D(r(x_{n+1}|x_n)||r(x_{n+1}|x_n)) \\ =& D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})) \end{aligned}$$

由于$D(r||r) = 0, D(p||q) \geq 0$，因此$D(p(x_n)||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1}))$，即$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}||\mu'_{n+1})$。 □

Remark 3.2.1. 相同转移概率下，不同初始分布趋同于平稳分布

Corollary 3.2.3. 若平稳分布μ为均匀分布，则熵增大，即$H(\mu_n) \leq H(\mu_{n+1})$

Proof.

$$\begin{aligned} D(\mu_n||\mu) &= \sum_{x_n} \mu_n(x_n) \log \frac{\mu_n(x_n)}{1/|\mathcal{X}|} \\ &= \log|\mathcal{X}| - H(\mu_n) \\ &= \log|\mathcal{X}| - H(X_n) \end{aligned}$$

□

Remark 3.2.2. 热学中的熵$S = \log\Omega$在各状态等可能的前提下与信息熵一致，这解释了热二律“熵增大”

Definition 3.2.1. 转移概率矩阵P的行和为1：$\sum_j P_{ij} = 1$，且$P_{ij} \geq 0$。若P列和也为1：$\sum_i P_{ij} = 1$，则P是双随机矩阵(doubly stochastic matrix)
均匀分布是平稳分布 $\Longleftrightarrow$ P双随机

Theorem 3.2.4. 对于平稳的*Markov*链$\{X_n\}$，$H(X_{n+1}|X_1) \geq H(X_n|X_1)$

Proof.

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

□

Lemma 3.2.5. 设$\{X_n\}$为平稳的*Markov*链，$\{Y_n\}$为$\{X_n\}$的对应项的函数，熵率为$H(\mathcal{Y})$，则

$$H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) \leq H(\mathcal{Y})$$

Proof.

$$\begin{aligned} H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) &\leq H(Y_n|Y_{n-1}, \cdots, Y_1, X_1, X_0, \cdots, X_{-k}) && \text{(Markov性)} \\ &= H(Y_n|Y_{n-1}, \cdots, Y_1, \cdots, Y_{-k}, X_1, X_0, \cdots, X_{-k}) && (Y_k = \phi_k(X_k)) \\ &\leq H(Y_n|Y_{n-1}, \cdots, Y_1, \cdots, Y_{-k}) \\ &= H(Y_{n+k+1}|Y_{n+k}, \cdots, Y_1) && \text{(平稳性))} \end{aligned}$$

令$k \to \infty$，则

$$H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) \leq \lim_{k\to\infty} H(Y_{n+k+1}|Y_{n+k}, \cdots, Y_1) = H(\mathcal{Y})$$

□

Lemma 3.2.6.

$$HY_n|Y_{n-1}, \cdots, Y_1 - H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) \to 0, n \to \infty$$

Proof.

$$\begin{aligned} H(X_1) &\geq \lim_{n\to\infty} I(X_1; Y_n|Y_{n-1}, \cdots, Y_1) \\ &= \sum_{n=1}^{\infty} I(X_1; Y_n|Y_{n-1}, \cdots, Y_1) && \text{(链式法则)} \\ &= \sum_{n=1}^{\infty} [H(Y_n|Y_{n-1}, \cdots, Y_1) - H(Y_n|Y_{n-1}, \cdots, Y_1, X_1)] \end{aligned}$$

□

又由3.1.1，

$$\{X_n\}\text{平稳} \implies H(X_n|X_1, X_2, \ldots, X_{n-1}) \downarrow H(\mathcal{X})$$

Theorem 3.2.7. 设$\{X_n\}$为平稳的*Markov*链，$\{Y_n\}$为$\{X_n\}$的对应项的函数，熵率为$H(\mathcal{Y})$，则

$$H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_{n-1}, \cdots, Y_1)$$

且

$$\lim_{n\to\infty} H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n|Y_{n-1}, \cdots, Y_1)$$

第四章　数据压缩

4.1　编码理论

4.2　Huffman编码

第五章　信道容量

5.1　通信系统模型

Definition 5.1.1 (信源模型). (1)根据信源输出信号所对应的随机过程是否平稳，分为稳恒（平稳）信源和非稳恒（非平稳）信源
(2)根据特殊的随机过程类型，分为高斯信源、Markov信源等
(3)信源字母表离散，信号取值时刻离散的稳恒信源称为离散稳恒信源

Definition 5.1.2 (信道模型). (1)按输入输出信号在幅值和时间上的取值分为离散信道（数字信道）、连续信道等
离散信道(discrete channel)是至多可数的输入字母表$\mathcal{X}$和输出字母表$\mathcal{Y}$，及$\mathcal{X}$到$\mathcal{Y}$的转移概率模型构成的系统
(2)如果信道输出值域信道在该时刻的输入有关，二与先前的输入输出等无关，则信道是无记忆信道(memoryless channel)；否则为有记忆信道
离散无记忆信道的转移概率模型可以用转移概率分布$p(y|x)$描述
(3)按输入输出信号之间的关系是否确定，分为有噪信道和无噪信道等

Definition 5.1.3 (信息信道容量). 离散无记忆信道的信息信道容量(information channel capacity)定义为

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

香农第二定理将指出信息信道容量=信道容量=信道最高码率。