信息论

Leo Yan

2026 年 2 月 28 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.

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

取等 $\iff X \sim U_{\mathcal{X}}$

Corollary 1.3.6 (熵的独立界)**.**

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

取等 $\iff X_1, X_2, \ldots, X_n$相互独立

Proof. 由链式法则，

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

□

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*

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

取等 $\iff X \to Z \to Y$

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$I(X;Y|Z) \ge 0$，故

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

□

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

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

故

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

另一方面，由数据处理不等式，

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

□

Corollary 1.4.4. 令$\hat{X}=Y$，则有

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

若$\hat{X}=X$，结论变为

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

Proposition 1.4.5. 设X,X'独立同分布，则

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

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

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

第二章 渐进均分性

2.1 渐进均分性

Definition 2.1.1 (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 2.1.1 (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 2.1.2 (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 $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$（概率和不超过1）
④ （第一条的推论）n充分大时，$|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$

Remark 2.1.1. 直观：

① $\Longrightarrow$ 典型集中的元素在数量级意义下是几乎等可能的；
② $\Longrightarrow$ 典型集出现概率随n增大而趋近于1（渐进）；
③④ $\Longrightarrow$ 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 2.1.2. 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

Definition 4.1.1. 字母表 设X是取值空间为$\mathcal{X}$的随机变量，Σ为D元字母表(alphabet)，Σ^*表示Σ上有限长字符串构成的集合。
总假设D元字母表为$\{0, 1, \cdots, D-1\}$。

Definition 4.1.2 (信源编码). 信源编码(source code)为映射$C : \mathcal{X} \to \Sigma^*$。
x的码字(codeword)记为$C(x)$，其长度记为$l(x)$
若C是单射，称该编码是非奇异的(nonsingular)。

Definition 4.1.3 (期望长度). 设$X \sim p(x)$，信源编码C的期望长度(expected length)为

$$L(C) = \mathbb{E}_p l(X) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Definition 4.1.4 (前缀码). C是前缀码(prefix code)或即时码(instantaneous code)，如果在序列中任一字符x_i的码字结束时，可以不依赖于后续码字而立即确定x_i。
等价定义：没有任何码字是其他码字的前缀，即

$$\forall c_1, c_2 \in C, c_1 \neq c_2, !\exists s \in \Sigma^*, c_1 \cdot s = c_2$$

Theorem 4.1.1 (Kraft不等式). *D元字母表上的即时码，码字长度$l_1, \cdots, l_m$满足*

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

给定一组满足上述不等式的码字长度，比存在相应的即时码。

Proof. 考虑D叉树，第k层为长为k的字符串，树枝表述字符。
前缀条件表明，**没有码字是其他码字的祖先，即码字必然对应叶子**.
令$l_{\max} = \max\{l_i\}$，考虑$l_{\max}$层中码字节点及其后代数M：
一方面，$M \leq D^{l_{\max}}$；
另一方面，位于l_i层的码字在$l_{\max}$层恰有$D^{l_{\max}-l_i}$个后代，且这些后代胡补充和，因此

$$M = \sum_i D^{l_{\max}-l_i}$$

综上，$\sum_{i=1}^{m} D^{-l_i} \le 1$.

反之，依字典序将第一个深度为l_1的节点标为码字1并去除其后代。依次对l_i重复以上步骤。不等式保证直至最后一步，必然存在所需层的节点。 □

Theorem 4.1.2 (推广的Kraft不等式). *任意可数的即时码字集，码字长度$\{l_i\}_{i=1}^{\infty}$满足*

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

反之也有相应即时码。

Proof. 以D进制小数

$$0.y_1 y_2 \cdots y_{l_i} = \sum_{j=1}^{l_i} y_j D^{-j}$$

表示码字，则在区间

$$[0.y_1 \cdots y_{l_i}, 0.y_1 \cdots y_{l_i} + D^{-l_i})$$

没有其他码字。区间长度之和≤ 1即Kraft不等式。反之构造同上。 □

Definition 4.1.5 (D进制概率分布). 如果概率质量函数p的每个概率值形如$D^{-n}, n \in \mathbb{N}$，则称p是D进制的(D-adic)

使用D元字母表，信源编码的期望长度不小于以D为底的信源分布的熵：

Theorem 4.1.3 (期望长度的估计). *设信源分布为$p(x_i) = p_i$，对随机变量X的任一D元即时码C，令$d = \sum D^{-l_i} \le 1$，D进制分布$q_i = D^{-l_i}/d$，则*

(1)

$$L(C) = H_D(X) + D(p\|q) - \log_D c \ge H_D(X)$$

取等时必须满足：分配最优码长$l_i^* = -\log_D p_i \in \mathbb{N}^*$.

(2)$-\log_D p_i \in \mathbb{N}^*$的必要条件是$p$是$D$进制分布，此时$p$=$q$.

Proof.

$$\begin{aligned} L(C) - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} \\ &= \sum p_i \log_D p_i D^{-l_i} \\ &= \sum p_i \log_D \frac{p_i}{q_i} \cdot \frac{1}{c} \\ &= \sum p_i \log_D \frac{p_i}{q_i} - \log_D c \\ &= D(p\|q) - \log_D c \ge 0 \qquad \text{(相对熵非负)} \end{aligned}$$

下面说明q_i及最优码长的由来，即最优化问题：

在$l_i \in \mathbb{N}, \sum D^{-l_i} \le 1$的条件下，求$\min L(C)$.

不考虑$l_i \in \mathbb{N}.\Omega = \{\mathbf{x} \in \mathbb{R}^m : \sum D^{-l_i} \le 1\}$是凸集，因为$\forall \mathbf{x}, \mathbf{y} \in \Omega$，由Hölder不等式，

$$sum_i D^{-\lambda x_i - (1-\lambda) y_i} \le \left(\sum_i D^{-x_i} \right)^{\lambda} \left(\sum_i D^{-y_i} \right)^{1-\lambda} \le 1, \forall 0 < \lambda < 1$$

用拉乘法求$\min_{\mathbf{x}\in\Omega} L(C)$.令

$$J(l_1,\cdots,l_m,\lambda)=\sum_i p_i l_i+\lambda\left(\sum_i D^{-l_i}-1\right)$$

再令

$$\frac{\partial J}{\partial l_i}=p_i-\lambda D^{-l_i}\log D=0$$

得

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

代回条件$\mathbf{x}\in\partial\Omega$得$\lambda=1/\log D$，从而最优码长为

$$l_i^*=-\log_D p_i$$

□

以上过程提供了**寻找最优编码**（在$D(p||q)$最小的意义下）的流程：寻找最接近$X\sim p$的D进制分布q，按q选取码字长度，再用Kraft不等式证明中的方法构造编码。按此流程构造的编码是香农码：

Definition 4.1.6 (香农码). 满足$l_i=-\log_D p_i$的编码称为香农码。

以下定理从期望长度的角度阐释了香农码的有效性：

Theorem 4.1.4. 设$\{l_i^*\}_{i=1}^m$是关于信源分布p和D元字母表的最优码长，$L^*=\sum p_i l_i^*$为期望长度，则

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

Proof. 令

$$l_i=\left\lceil\log_D\frac{1}{p_i}\right\rceil\in\left[\log_D\frac{1}{p_i},\log\frac{1}{p_i}+1\right)$$

它满足Kraft不等式：

$$\sum D^{l_i}\le\sum D^{-\log\frac{1}{p_i}}=1$$

这样就构造出在所需区间中的期望长度。 □

下面总令D=2。

通过对字符进行**分组编码**，可以减少每字符的附加位，从而降低L:

Theorem 4.1.5. 考虑将$\mathbf{x}=(x_1,\cdots,x_n)$视为$\mathcal{X}^n$中的超字符，$l(\mathbf{x})$为$\mathbf{x}$的码字长度，则每输入字符的最小期望码字长度满足

$$\frac{1}{n}H(\mathbf{X})\le L_n^*<\frac{1}{n}H(\mathbf{X})+\frac{1}{n}$$

进一步，若$\{X_j\}$是平稳过程，则

$$\lim_{n\to\infty}L_n^*=H(\mathcal{X})$$

Proof.

$$L_n = \frac{1}{n}\sum p(\mathbf{x})l(\mathbf{x}) = \frac{1}{n}\mathbb{E}l(\mathbf{X})$$

按

$$l(\mathbf{x}_i) = \left\lceil \log\frac{1}{p(\mathbf{x}_i)} \right\rceil$$

分配码长，则构造出所需区间中的L_n。后一命题得自夹逼定理。 □

以下定理估计了考虑偏码(wrong code)时的期望码长:

Theorem 4.1.6. *如果非真实分布q是对真实分布p的最佳估计，则在码长分配*

$$l(x) = \left\lceil \log\frac{1}{q(x)} \right\rceil$$

下，关于$p(x)$的期望长度满足:

$$H(p) + D(p||q) \le \mathbb{E}_p l(X) < H(p) + D(p||q) + 1$$

Proof.

$$l(x) \in \left[\log\frac{1}{q(x)}, \log\frac{1}{q(x)} + 1\right)$$

类似thm4.1.4.带入即证。 □

以下定理表明，仅仅要求编码唯一可译，而不要求即时性，也无法找到更小码长：

Theorem 4.1.7 (McMillan定理)**.** *任一唯一可译的D元码也满足Kraft不等式$\sum D^{l_i} \le 1$；反之，给定一组码字长度$\{l_i\}$满足$\sum D^{l_i} \le 1$，必有对应的唯一可译码。*

Proof. $\forall k \in \mathbb{N}^*$，考虑C的k次扩展，有

$$l(x_!, \cdots, x_k) = \sum_{i=1}^{k} l(x_i)$$

注意到恒等式

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{x^k\in\mathcal{X}^k} D^{-l(x_1)} \cdots D^{-l(x_k)} = \sum_{x^k\in\mathcal{X}^k} D^{-l(\mathbf{x})}$$

按码字长度合并同类项：

$$\sum_{x^k\in\mathcal{X}^k} D^{-l(\mathbf{x})} = \sum_{m=1}^{kl_{\max}} a_m D^{-m}$$

唯一可译性要求$a_m \le D^m$，于是

$$\sum_{m=1}^{kl_{\max}} a_m D^{-m} \le kl_{\max}$$

即

$$\sum D^{-l_j} \le (kl_{\max})^{1/k}$$

令$k \to \infty$即得。 □

Corollary 4.1.8. 无限信源字母表$\mathcal{X}$的唯一可译码满足*Kraft*不等式。

Proof. 唯一可译码的有限子集仍是唯一可译的。 □

4.2 Huffman编码

Huffman编码是**最优编码的构造算法**。给定D元字母表Σ，X有概率分布$(p_1, \cdots, p_n)$，重复执行以下步骤：
①将概率按降序排列
②在这个排列方式中，选取末尾D个字符，依次赋码字$0 \sim D-1$
③赫夫曼合并(Huffman reduction)：将此D个字符视作一个新字符，概率为各字符概率之和
重复到只剩一个字符为止。

Codeword Length	Codeword	X	Probability				
2	01	1	0.25	0.3	0.45	0.55	1
2	10	2	0.25	0.25	0.3	0.45	
2	11	3	0.2	0.25	0.25		
3	000	4	0.15	0.2			
3	001	5	0.15				

图 4.1: 赫夫曼编码算法执行例

Remark 4.2.1. ①若$D \geq 3$，每一步减少字符数为D-1，字符总数必为1+k(D-1).如果$|\mathcal{X}|$不满足条件，需要添加概率为0的虚拟字符
②将p_j视为权重时可以不要求$\sum p_j = 1$，此时同样可以用赫夫曼码最小化码长加权和$\sum p_j l_j$
③赫夫曼码不唯一。例如

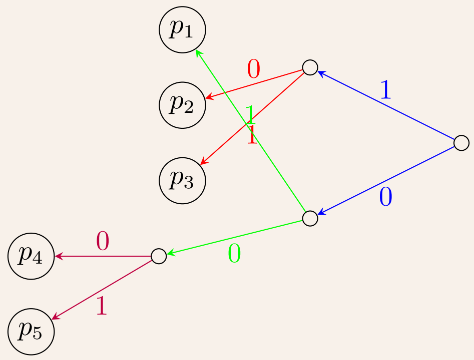
$$D = 2, p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

Codeword	X	Probability			
1	1	0.25	0.25	0.5	1.0
2	2	0.25	0.25	0.25	
01	3	0.2	0.2	0.25	
02	4	0.1	0.2		
000	5	0.1	0.1		
001	6	0.1			
002	Dummy	0.0			

图 4.2: D=3,虚拟字符使用例

上图中缩减无兄弟的节点后即得**编码树**（字符对应叶子节点）

Example 4.2.1. 图4.1.的编码树为



赫夫曼码是一种最优即时码。最优即时码是在期望长度最小意义下的即时码，它不是唯一的。

Lemma 4.2.1 (最优即时码的刻画)**.** *任给分布* $\mathbf{p}$*，* D*=2，最优即时码必然满足：*
(1)$p_i > p_j \implies l_i \le l_j$*，即码长与概率反序；*
(2)最长的两个码字拥有相同的长度，且对应两个最小可能字符。
此外，总能构造编码满足下述条件：
(3)最长的两个码字仅在最后一位有差别
称满足(3)的最优编码是典则的(canonical).

Proof. (1)对$p_i > p_j$，设编码C满足$l_i \le l_j$，C'为C交换i,j编码所得，则有

$$\begin{aligned} L(C') - L(C) &= p_j l_i + p_i l_j - p_i l_i - p_j l_j \\ &= (p_j - p_i)(l_i - l_j) \ge 0 \end{aligned}$$

(2)(3)两个最长码字在编码树中对应兄弟节点，否则必然可以修剪编码树（删除码字最后一位）。 □

Remark 4.2.2. (1)赫夫曼码满足

$$H(X) \le L(C) < H(X) + 1$$

(2)赫夫曼码是贪心算法，即局部最优→全局最优

4.3 其他编码

1.香农码

前面已经介绍。对于单个字符，赫夫曼码与香农码都可能有更短的码字长度，但平均意义下赫夫曼码才是最优编码。

香农码具有唯一竞争最优性：在单局游戏中，以甲、乙对同一随机序列的编码长短决定胜负，则香农码是一个很好的策略。

以下设D=2.

Theorem 4.3.1. *(1)设C为香农码，C'为其他唯一可译码，则*

$$P\{l(X) \geq l'(X)+c\} \leq \frac{1}{2^{c-1}}$$

(2)（香农码的唯一竞争最优性）设$p(x)$是二进制的，$l(x)=\log\frac{1}{p(x)}$为香农码的码字长度，则

$$\forall C', P\{l(X) < l'(X)\} \geq P\{l(X) > l'(X)\}$$

取等 $\Longleftrightarrow$ $\forall x, l(x)=l'(x)$。

Corollary 4.3.2. *对于非二进制的$p(x)$，有*

$$\mathbb{E}sgn(l(X)-l'(X)-1) \leq 0$$

其中l为香农码C的码字长度。

2.**费纳(Fano)编码**

将概率值递减排列，然后选取k使得

$$\left|\sum_{i=1}^{k} p_i - \sum_{i=k+1}^{m} p_i\right|$$

最小（**二分概率**），概率高者赋0，概率低者赋1，再对子字符集重复此过程，直到每一字符分得唯一码字。

一般来说，这种编码不是最优编码，但是可以达到$L(C) \leq H(X)+2$.

3.**Shannon-Fano-Elias编码**

考虑累积分布函数$F(x)=\sum_{a \leq x} p(a)$及其修正$\overline{F}(x)$，$\overline{F}(x)$在间断点处取左右平均值。

以$\overline{F}(x)$的二进制表示的小数部分作为字符x的码字。在码字无限时，取前$\left\lceil \log\frac{1}{p(x)} \right\rceil + 1$位即可。

编码的唯一性：以$\lfloor \overline{F}(x) \rfloor_{l(x)}$记$\overline{F}(x)$保留$l(x)$位。取$l(x)=\left\lceil \log\frac{1}{p(x)} \right\rceil + 1$，则

$$F(x) - \lfloor \overline{F}(x) \rfloor_{l(x)} < \frac{1}{2^{l(x)}} < \frac{1}{2} 2^{\log p(x)} = \frac{p(x)}{2}$$

即$\lfloor \overline{F}(x) \rfloor_{l(x)}$在F(x)下方跳变幅度以内，对应了唯一的区间。

这种编码的特点是：应用广泛；$L(C) < H(X)+2$；必为前缀码（考虑区间树可证）

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

(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 (离散无记忆信道)**.** 离散无记忆信道(DMC)的转移概率模型可以用转移概率分布$p(y|x)$描述。

设输入字母表为$\mathcal{X}$，输出字母表为$\mathcal{Y}$，该DMC可以用

$$(\mathcal{X}, p(x|y), \mathcal{Y})$$

表示

离散无记忆信道的n次扩展可以用

$$(\mathcal{X}^n, p(x^n|y^n), \mathcal{Y}^n)$$

表示，其中

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \forall k \in \{1, 2, \cdots, n\}$$

进一步，如果信道不带反馈（如不明确指出，则默认不带反馈），即输入字符不依赖于过去的输出字符，即

$$p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1}), \forall k$$

则n次扩展的信道转移函数简化为

$$p(y^k|x^k) = \prod_{i=1}^{k} p(y_i|x_i), \forall k$$

Definition 5.1.4 (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=任意小误差传输比特数/信道使用次数。

$I(X;Y)$关于输入分布$p(x)$是紧集上的连续的凹函数，因此最大值存在且唯一，使用max是合理的。

Proposition 5.1.1 (信息信道容量的性质).

- $C \geq 0$，取等 $\iff$ 信道输出与输入独立
- $C \leq \log|\mathcal{X}|$，取等 $\iff$ X, Y独立且$X \sim U_{\mathcal{X}}$
- $C \leq \log|\mathcal{Y}|$，取等 $\iff$ 存在输入分布使得输出分布为均匀分布

Definition 5.1.5 (对称信道). 如果离散无记忆信道的概率转移矩阵$p(x|y)$是数独矩阵，则称该信道为对称信道(symmetric channel)。数独矩阵必然是双随机的。
数独矩阵可以描述为：任何两行互为置换、任何两列也互为置换。
如果每一行$p(\cdot|x)$都是同一分布的置换，且所有列的元素和$\sum_x p(y|x)$相同，则该信道是弱对称信道(weakly symmetric channel)

Theorem 5.1.2. 对于弱对称信道，信息信道容量为

$$C = \log|\mathcal{Y}| - H(p(\cdot|x))$$

其中$p(\cdot|x)$为任一行的分布。取等 $\iff$ 输入分布为均匀分布

Example 5.1.1 (有噪打字机信道). 该种信道的概率转移矩阵矩阵的基本特征是，对角线元素（不出错的概率）接近于1，除了对角线，每一行都有个别元素数值较大，而其他元素极小或为0。描述了考虑失误的打字机。
输入、输出字母表相同，输入以0.5的概率被输出端无改变地接收，以0.5的概率变为下一个字母。字母表大小为26时，$C = 13bit/$传输：
一方面，每次传输可以无误差地传输其中13个字符（仅传输奇数位置或偶数位置的字符）；
另一方面，

$$C = \max I(X;Y) = \max[H(Y) - H(Y|X)] = \max H(Y) - 1 = \log 26 - 1 = \log 13bit/\text{传输}$$

取等 $\iff$ $p(x) \sim U$

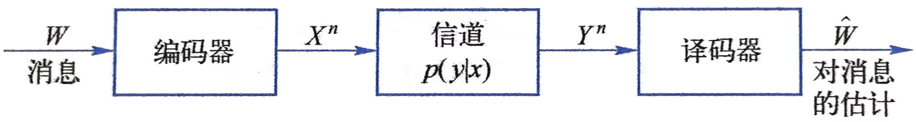


图 5.1: 离散通信信道模型

Definition 5.1.6 ((M,n)码). 信道$(\mathcal{X}, p(y|x), \mathcal{Y})$的$(M,n)$码由以下部分构成：

- 指标集$\{1,2,\cdots,M\}$
- 编码函数(encoding function)

$$x^n : \{1,2,\cdots,M\} \to \mathcal{X}^n$$

生成码字(codewords)$x^n(1), x^n(2), \cdots, x^n(M)$，所有码字的集合称为码本(code-book)

$$\mathcal{C} = \{x^n(1), x^n(2), \cdots, x^n(M)\} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M\}$$

- 译码函数

$$g : \mathcal{Y}^n \to \{1,2,\cdots,M\}$$

Definition 5.1.7 (条件误差概率). 已知指标i被发送，（给定发送符号时的）条件误差概率(conditional probability of error)为

$$\lambda_i = P\{g(Y^n) \neq i | X^n = x^n(i)\}$$

它衡量了特定符号在传输过程中的脆弱性或抗干扰能力。
已知接收到符号y，（给定接收符号时的）条件误差概率为

$$1 - \max_x P\{X = x | Y = y\}$$

反映了收到特定y后，接收端对这次判决的不确定性。条件误差概率通常指前者。

Definition 5.1.8. (M,n)码的最大误差概率(maximun probability of error)为

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

（算数）平均误差概率(average probability of error)为

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

Remark 5.1.1. 如果指标W的选取服从均匀分布，则$P_e^{(n)} = P\{W \neq g(Y^n)\}$。

Remark 5.1.2. 显然，$P_e^{(n)} \leq \lambda^{(n)}$。理应希望这二者差异大，但将证明：在相同码率下，平均误差概率小 $\Longrightarrow$ 最大误差概率小

Definition 5.1.9 (码率). (M,n)码的码率(rate)为

$$R=\frac{\log M}{n}(bit/\text{传输})$$

如果存在一个$(\lceil 2^{nR}\rceil,n)$码，满足$\lambda^{(n)}\to 0,n\to 0$，则称码率R是可达的(achievable)。一般将此编码简记为$(2^{nR},n)$。
信道容量定义为可达码率的上确界。

Definition 5.1.10 (联合典型序列). 服从分布$p(x,y)$的联合典型序列(jointly typical sequence)$\{(x^n,y^n)\}$所构成的集合$A_\epsilon^{(n)}$是指其经验熵与真实熵以ϵ为界靠近的n长序列构成的集合，即：

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

其中

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

Theorem 5.1.3 (联合AEP). 设(X^n,Y^n)为服从$p(x^n,y^n)=\prod_{i=1}^n p(x_i,y_i)$的$i.i.d.$的$n$长序列，则
①$P\{(X^n,Y^n)\in A_\epsilon^{(n)}\}\to 1,n\to\infty$
②$|A_\epsilon^{(n)}|\le 2^{n(H(X,Y)+\epsilon)}$
③如果$(\tilde{X}^n,\tilde{Y}^n)\sim p(x^n)p(y^n)$（$i.e.\tilde{X}^n$与$\tilde{Y}^n$独立且具有$p(x^n,y^n)$的边界分布，与$X^n,Y^n$分别同分布），则

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

且对于充分大的n，有

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

Proof. ①由弱大数定律，

$$-\frac{1}{n}\log p(X^n)\to-\mathbb{E}\log p(X)=H(X)\qquad\text{依概率}$$

故$\forall\epsilon>0,\exists n_1,n>n_1$时，有

$$P\left\{\left|-\frac{1}{n}\log p(X^n)-H(X)\right|\ge\epsilon\right\}<\frac{\epsilon}{3}$$

同理，

$$-\frac{1}{n}\log p(Y^n)\to-\mathbb{E}\log p(Y)=H(Y)\qquad\text{依概率}$$

$$-\frac{1}{n}\log p(X^n,Y^n)\to -\mathbb{E}\log p(X,Y)=H(X,Y) \qquad \text{依概率}$$

从而有n_2,n_3，

$$n>n_2 \implies P\left\{\left|-\frac{1}{n}\log p(Y^n)-H(Y)\right|\geq\epsilon\right\}<\frac{\epsilon}{3}$$

$$n>n_3 \implies P\left\{\left|-\frac{1}{n}\log p(X^n,Y^n)-H(X,Y)\right|\geq\epsilon\right\}<\frac{\epsilon}{3}$$

取$n>\max\{n_1,n_2,n_3\}$，则

$$P\{(X^n,Y^n)\in A_\epsilon^{(n)}\}>1-\epsilon$$

②注意到

$$\begin{aligned}1&=\sum p(x^n,y^n)\\&\geq\sum_{A_\epsilon^{(n)}}p(x^n,y^n)\\&\geq|A_\epsilon^{(n)}|\cdot 2^{-n(H(X<Y)+\epsilon)}\end{aligned}$$

这与要证的式子是等价的。最后一行成立是因为，根据$A_\epsilon^{(n)}$的定义（最后一条），每个$A_\epsilon^{(n)}$的元素(x^n,y^n)都满足

$$-n(H(X,Y)+\epsilon)<\log p(x^n,y^n)<-n(H(X,Y)-\epsilon)$$

③在这种假设下，有

$$\begin{aligned}P\left\{\left(\tilde{X}^n,\tilde{Y}^n\right)\in A_\epsilon^{(n)}\right\}&=\sum_{(x^n,y^n)\in A_\epsilon^{(n)}}p(x^n)p(y^n)\\&\leq 2^{n(H(X,Y)+\epsilon)}\cdot 2^{-n(H(X)-\epsilon)}\cdot 2^{-n(H(Y)-\epsilon)}\\&=2^{-n(I(X;Y)-3\epsilon)}\end{aligned}$$

在n充分大时，$P\left\{\left(\tilde{X}^n,\tilde{Y}^n\right)\in A_\epsilon^{(n)}\right\}\geq 1-\epsilon$，因此

$$\begin{aligned}1-\epsilon&\leq\sum_{(x^n,y^n)\in A_\epsilon^{(n)}}p(x^n,y^n)\\&\leq|A_\epsilon^{(n)}|\cdot 2^{-n(H(X,Y)-\epsilon)}\end{aligned}$$

即证后一不等式。 $\square$

第六章　最后一页