

# 信息论

Leo Yan

2026 年 2 月 26 日

# 目录

<b>第一章 熵, 相对熵, 互信息</b>	<b>3</b>
1.1 基本概念 . . . . .	3
1.2 多维随机变量 . . . . .	4
1.3 不等式 . . . . .	6
1.4 其他背景下的不等式 . . . . .	7
<b>第二章 演进均分性</b>	<b>10</b>
2.1 演进均分性 . . . . .	10
<b>第三章 熵率与Markov链</b>	<b>12</b>
3.1 熵率 . . . . .	12
3.2 Markov链 . . . . .	12
<b>第四章 数据压缩</b>	<b>15</b>
4.1 编码理论 . . . . .	15
4.2 Huffman编码 . . . . .	15
<b>第五章 信道容量</b>	<b>16</b>
5.1 通信系统模型 . . . . .	16

# 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。” $\parallel$ ”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).

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

**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).

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

**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  $\iff X \sim U_{\mathcal{X}}$

**Corollary 1.3.6** (熵的独立界).

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

取等  $\iff X_1, X_2, \dots, X_n$  相互独立

**Proof.** 由链式法则,

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

□

## 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 $X \rightarrow Y \rightarrow Z$   
条件独立的意思是

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

X,Z对称, 即 $X \rightarrow Y \rightarrow Z \iff Z \rightarrow Y \rightarrow X$ , 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 $X \rightarrow Y \rightarrow Z$ , 则有

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

取等  $\iff X \rightarrow Z \rightarrow Y$

**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.**

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

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

*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

$$\theta \rightarrow X \rightarrow T(X)$$

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 $\theta$ 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 $\theta$ 条件独立, 即 $\theta \rightarrow T(X) \rightarrow X$

**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 $\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X$

*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 $\theta$ 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 $X \rightarrow Y \rightarrow \hat{X}$ 。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 $E = \mathbf{1}_{\{\hat{X} \neq X\}}$ , 则有马尔可夫链 $X \rightarrow Y \rightarrow \hat{X} \rightarrow E$ 。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 演进均分性

### 2.1 演进均分性

**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

$$\forall \epsilon > 0, P\{|X_n - X| \geq \epsilon\} \rightarrow 0 (n \rightarrow \infty)$$

则称 $X_n$ 依概率收敛于 $X$ , 记为 $X_n \xrightarrow{P} X$

②如果

$$P\left\{\lim_{n \rightarrow \infty} X_n = X\right\} = 1$$

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

$$\mathbb{E}(X_n - X)^2 \rightarrow 0 (n \rightarrow \infty)$$

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 $X_1, X_2, \dots, X_n i.i.d. \sim p(x)$ , 则有

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

**Proof.**

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

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

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

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 2.1.1.* 直观:

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

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

因而理论上用 $nH(X)$ 比特即可表示序列 $x_1, x_2, \dots, x_n$  (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。

$$A_\epsilon^{(n)} \leq 2^{n(H(X)+\epsilon)} \Rightarrow \lceil n(H(X) + \epsilon) \rceil \leq n(H(X) + \epsilon) + 1 \text{ (bit)}$$

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

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

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

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

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

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

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

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

**Theorem 3.1.1.** 定义

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

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

**Proof.**

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

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

由链式法则，

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

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 $\mu$ ，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.**

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

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

□

**Theorem 3.2.2.** 设  $\mu_n, \mu'_n$  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  $\mu_n, \mu'_n$  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  $\mu$  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  $\iff P$  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

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

**Proof.**

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

令 $k \rightarrow \infty$ ，则

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

□

**Lemma 3.2.6.**

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

**Proof.**

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

□

又由3.1.1，

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

**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

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

且

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

# 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  
(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

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

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,  $C = 13bit/\text{传输}$ :

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  $\iff p(x) \sim U$

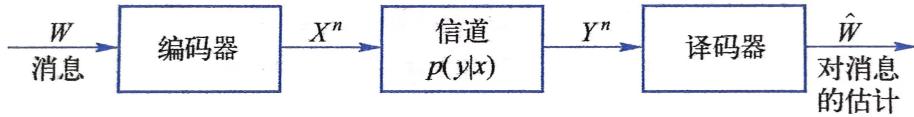


图 5.1: 离散通信信道模型

**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 $\{1, 2, \dots, M\}$
- 编码函数(encoding function)

$$x^n : \{1, 2, \dots, M\} \rightarrow \mathcal{X}^n$$

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

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

- 译码函数

$$g : \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

**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

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

(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,  $P_e^{(n)} \leq \lambda^{(n)}$ 。通常希望这二者差异大, 但将证明: 在相同码率下, 平均误差概率小  $\Rightarrow$  最大误差概率小

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

$$R = \frac{\log M}{n} (\text{bit/传输})$$

如果存在一个( $\lceil 2^{nR} \rceil, n$ )码, 满足  $\lambda^{(n)} \rightarrow 0, n \rightarrow 0$ , 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  $\epsilon$  为界靠近的  $n$  长序列构成的集合, 即:

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

其中

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