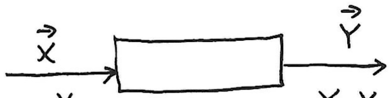
无失真信源编码



$x_1\ x_2 \cdots x_L$ $\quad$ $Y_1\ Y_2 \cdots Y_{K_L}$

$x_i \in \{a_1 \cdots a_n\}$ $\quad$ $Y_i \in \{b_1\ b_2 \cdots b_m\}$

一、 $\vec{X} = (X_1\ X_2 \cdots X_L)$，为独立同分布 (iid) 随机变量序列，具有渐近均分性质 (AEP)。
(Asymptotic equipartition property)

$$I(\vec{x}) = -\log p(x_1 x_2 \cdots x_L) = -\log p(x_1) - \log p(x_2) - \cdots - \log p(x_L)$$
$$= \sum_{i=1}^{L} I(x_i)$$

均值 $E\{I(\vec{X})\} = E\{\sum_{i=1}^{L} I(x_i)\} = \sum_{i=1}^{L} E\{I(x_i)\} = \sum_{i=1}^{L} H(x_i) = LH(x)$

方差 $D\{I(\vec{X})\} = \sum_{i=1}^{L} D\{I(x_i)\} = LD(I(x))$

$$D(I(x)) = E[(I(x) - H(x))^2]$$
$$= \sum_{i=1}^{n} p(x_i)(-\log p(x_i) - H(x))^2$$

切比雪夫不等式

$$\Pr\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} \qquad \varepsilon > 0,\ \sigma > 0$$

$$\Pr\{|I(\vec{X}) - LH(x)| \geq L\varepsilon\} \leq \frac{LD(I(x))}{(L\varepsilon)^2}$$

$$\Pr\{|\frac{I(\vec{X})}{L} - H(x)| \geq \varepsilon\} \leq \frac{D(I(x))}{L\varepsilon^2}$$

$$L \to \infty,\ \frac{D(I(x))}{L\varepsilon^2} \to 0,\quad \frac{-\log p(x_1 x_2 \cdots x_L)}{L} \to H(x)$$

$$p(x_1 x_2 \cdots x_L) \to 2^{-LH(x)}$$

$0 \sim p$
$1 \sim 1-p$

$$p(01101) = p^2 q^3$$

定义 A_ε（典型集）

$$A_\varepsilon = \{\vec{x} : |\frac{I(\vec{x})}{L} - H(x)| \leq \varepsilon\}$$

$$A_\varepsilon^c = \{\vec{x} : |\frac{I(\vec{x})}{L} - H(x)| > \varepsilon\}$$

$$A_\varepsilon + A_\varepsilon^c = X^L = \{\vec{x}_i\} \quad i = 1, 2, \cdots, n^L$$

A_ε 有以下性质：

(1) 若 $\vec{x} \in A_\varepsilon$，则

$$2^{-L(H(x)+\varepsilon)} \leq p(\vec{x}) \leq 2^{-L(H(x)-\varepsilon)}$$

（典型集中的元素几乎是等概率出现的）

(2) 当 L 充分大时，$\Pr\{A_\varepsilon\} > 1 - \varepsilon$ （典型集的概率近似为1）

$$\Pr\{A_\varepsilon\} = \sum_{\vec{x} \in A_\varepsilon} p(\vec{x}) > 1 - \varepsilon$$

(3) $|A_\varepsilon| \leq 2^{L(H(x)+\varepsilon)}$ $\qquad$ $|A_\varepsilon|$ 为典型集中元素的个数

$$1 = \sum_{\vec{x} \in X^L} p(\vec{x}) \geq \sum_{\vec{x} \in A_\varepsilon} p(\vec{x}) \geq \sum_{\vec{x} \in A_\varepsilon} 2^{-L(H(x)+\varepsilon)} = |A_\varepsilon| 2^{-L(H(x)+\varepsilon)}$$

$$\Rightarrow |A_\varepsilon| \leq 2^{L(H(x)+\varepsilon)}$$

(4) $|A_\varepsilon| \geq (1-\varepsilon)2^{L(H(X)-\varepsilon)}$

L充分大时，$P_r\{A_\varepsilon\} > 1-\varepsilon$

$$1-\varepsilon < P_r\{A_\varepsilon\} = \sum_{\vec{x}\in A_\varepsilon} p(\vec{x}) \leq \sum_{\vec{x}\in A_\varepsilon} 2^{-L(H(X)-\varepsilon)} = |A_\varepsilon|\, 2^{-L(H(X)-\varepsilon)}$$

$$\Rightarrow \quad |A_\varepsilon| \geq (1-\varepsilon)2^{L(H(X)-\varepsilon)}$$

* AEP也可由弱大数定理直接得到：n很大时 $\frac{1}{n}\sum_{i=1}^{n} x_i \to E(X)$ （X i.i.d.）

$$-\frac{1}{L}\log p(x_1 x_2 \cdots x_L) = -\frac{1}{L}\sum_{i=1}^{L}\log p(x_i)$$

$$\to -E(\log p(x)) = H(X)$$

二、定长编码定理。

由L个符号组成的，每个符号的熵为 $H_L(\vec{X})$ 的无记忆平稳符号序列 $X_1, X_2, \cdots, X_L$，可用 K_L 个符号 $Y_1, Y_2, \cdots, Y_{K_L}$（每个符号有 m 种可能值）进行定长编码。对任意 $\varepsilon>0$，$\delta>0$，只要

$$\frac{K_L}{L}\log m \geq H_L(\vec{X}) + \varepsilon$$

则当 L 足够大时，必可使译码差错小于 δ；反之，当

$$\frac{K_L}{L}\log m \leq H_L(\vec{X}) - 2\varepsilon$$

时，译码差错一定是有限值，当 $L\to\infty$ 时，译码几乎必定出错。

证明：对 A_ε 中的元素进行定长编码，为了唯一可译，码字个数大于等于 $|A_\varepsilon|$，即

（$H_L(\vec{X})$ 平均符号熵写成 $H(X)$）

$m^{K_L} \geq |A_\varepsilon|$ 　此时 $|A_\varepsilon|$ 取上界

$$m^{K_L} \geq 2^{L(H(X)+\varepsilon)}$$

$$K_L \log m \geq L(H(X)+\varepsilon)$$

$$\frac{K_L}{L}\log m \geq H(X)+\varepsilon$$

此时，错误译码的概率 $P_e = P_r\{A_\varepsilon^c\} \leq \dfrac{D(I(x))}{L\varepsilon^2}$，

若需 $P_e \leq \delta$，$\dfrac{D(I(x))}{L\varepsilon^2} \leq \delta$，$L \geq \dfrac{D(I(x))}{\varepsilon^2\delta}$

反之，若 $\frac{K_L}{L}\log m \leq H(X) - 2\varepsilon$，

$$\log m^{K_L} \leq L(H(X)-2\varepsilon)$$

$$m^{K_L} \leq 2^{L(H(X)-2\varepsilon)} < \underbrace{(1-\varepsilon)2^{L(H(X)-\varepsilon)}}_{|A_\varepsilon|\text{下界}}$$

码字个数小于 A_ε 中元素个数

A_ε 中选取 m^{KL} 个元素进行一对一编码，

$$\Pr\{A_\varepsilon \text{中} m^{KL} \text{个元素}\} \le m^{KL} \cdot \max_{\vec{x}\in A_\varepsilon} p(\vec{x}) = m^{KL} 2^{-L(H(X)-\varepsilon)}$$

$$\le 2^{L(H(X)-2\varepsilon)-L(H(X)-\varepsilon)}$$

$$= 2^{-L\varepsilon}$$

译码错误概率为

$$P_e = 1-2^{-L\varepsilon}, \quad L\to\infty, \ P_e \to 1$$

三、单符号变长编码定理：

若离散无记忆信源的符号熵为 $H(X)$，每个信源符号用 m 进制码元进行变长编码，一定存在一种无失真编码方法，其（m 进制）码字平均长度 $\bar{k}$ 满足下列不等式：

$$\frac{H(X)}{\log m} \le \bar{k} < \frac{H(X)}{\log m} + 1$$

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ p(a_1) & p(a_2) & \cdots & p(a_n) \end{bmatrix}$$

码长 $l_1 \quad l_2 \quad \cdots \quad l_n$

$$\bar{k} = \sum_{i=1}^{n} p(a_i) l_i$$

证明：(1)

必存在无失真编码，使得 $H(X)-(\log m)\bar{k} \le 0$

$$H(X)-(\log m)\bar{k} = -\sum_{i=1}^{n} p(a_i)\log p(a_i) - \sum_{i=1}^{n} p(a_i) l_i \log m$$

$$= \sum_{i=1}^{n} p(a_i) \log \frac{m^{-l_i}}{p(a_i)}$$

Jensen 不等式 $E[f(x)] \le f(E(x))$

$$H(X)-(\log m)\bar{k} \le \log \sum_{i=1}^{n} p(a_i) \frac{m^{-l_i}}{p(a_i)} = \log \sum_{i=1}^{n} m^{-l_i}$$

唯一可译码存在（无失真），因此 $\sum_{i=1}^{n} m^{-l_i} \le 1$

$$H(X)-(\log m)\bar{k} \le \log 1 = 0$$

因此有 $\bar{k} \ge \dfrac{H(X)}{\log m}$

若要等号成立，$m^{-l_i} = p(a_i)$，$l_i = \dfrac{-\log p(a_i)}{\log m}$

(2) 证明 $\bar{k} < \dfrac{H(X)}{\log m} + 1$

符号 a_i 的码字长度 l_i 需要是整数，因此取

$$\frac{-\log p(a_i)}{\log m} \le l_i < \frac{-\log p(a_i)}{\log m} + 1$$

由 $\frac{-\log p(a_i)}{\log m} \le l_i \Rightarrow -\log p(a_i) \le \log m^{l_i}$

$$p(a_i) \ge m^{-l_i}$$

$$\sum_{i=1}^{n} m^{-l_i} \le \sum_{i=1}^{n} p(a_i) = 1$$ 表明 l_i 满足唯一可译码存在的条件

由 $l_i < \frac{-\log p(a_i)}{\log m} + 1$

$$\sum_{i=1}^{n} l_i\, p(a_i) < \sum_{i=1}^{n} p(a_i)\left(\frac{-\log p(a_i)}{\log m} + 1\right)$$

$$\bar{k} < \frac{H(X)}{\log m} + 1$$

四、离散平稳无记忆序列变长编码定理（香农第一定理）

对于平均符号熵为 $H_L(\vec{X})$ 的离散平稳无记忆信源，必存在一种无失真编码方法，使平均信息率 $\bar{K}$ 满足不等式

$$H_L(\vec{X}) \le \bar{K} < H_L(\vec{X}) + \varepsilon$$

其中 ε 为任意小正数。

证明：将 $\vec{X} = (X_1 X_2 \cdots X_L)$ 作为一个整体，构成一个新的单符号信源，因此有

$$\frac{L H_L(\vec{X})}{\log m} \le \bar{k}_L < \frac{L H_L(\vec{X})}{\log m} + 1$$

$$H_L(\vec{X}) \le \frac{\bar{k}_L}{L}\log m < H_L(\vec{X}) + \frac{\log m}{L}$$ ，令 $\bar{K} = \frac{\bar{k}_L}{L}\log m$

$$H_L(\vec{X}) \le \bar{K} < H_L(\vec{X}) + \varepsilon. \quad (L \to \infty)$$