上周作业解答

Exercise 1.

a) 举例说明非奇异码不一定是唯一可译码；

b) 说明等长的非奇异码是唯一可译码；

c) 说明前缀码是唯一可译码，但唯一可译码可以不是前缀码。

d) 说明前缀码反序之后得到的码是唯一可译码。

上周作业解答

a) 例：如 $1 \to 0, 2 \to 10, 3 \to 010$，则对于序列 010，它满足非奇异码的条件，但有可能译为 1, 2，也有可能译为 3。唯一可译码是指同一段序列只能划分成唯一的码字序列，且每个码字有唯一的信源符号对应。而非奇异码只要求每个码字有唯一的信源符号对应。

b) 由于码字等长，因此不会分割出错误的码字。又因为是非奇异码，同一个码字只能对应于一个符号，所以不会错译，是唯一可译码。

c) 前缀码中码字之间不互为前缀，因此码字互不相同，也满足非奇异码的条件。又因为前缀码中，任何码字都不是其它码字的前缀，因此对编码后的码字序列进行的码字划分是唯一的，所以为唯一可译码。

非前缀码也有可能满足唯一可译码的条件，只需要令编码方案中所有存在前缀关系的不同码字对中，任意一一对中较长的码字在除去前缀段之后的剩余部分无法独自或和其它码字配合译出消息序列，便满足唯一可译的条件。

上周作业解答

d) 前缀码反序后为后缀码，即任意码字之间不互为后缀。这样的码字也互不相同，因此满足非奇异码条件。又因为各码字不互为后缀，因此对码字序列的划分也是唯一的，因此满足唯一可译码的条件。

上周作业作业解答

Exercise 2.[王育民(2013)]

令离散无记忆信源

$$U=\left\{\begin{array}{cccc} a_1 & a_2 & \cdots & a_M \\ P(a_1) & P(a_2) & \cdots & P(a_M) \end{array}\right\}$$

其中 $a_1, a_2, \cdots, a_M$ 依概率从大到小排序。定义

$$F_i=\sum_{k=1}^{i-1} P(a_k), i>1$$

而 $F_1=0$，今按下述方法进行二元编码。消息 a_k 的码字为实数 F_k的二元数字表示序列的截短（例如 1/2 的二元数字表示序列为 $1/2 \rightarrow 1000\cdots, 1/4 \rightarrow 0100\cdots$），保留的截短序列长度 ℓ_k 是大于或等于 $I(a_k)$的最小整数。

上周作业解答

(a)对信源 $\left\{\begin{array}{cccccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 1/4 & 1/4 & 1/8 & 1/8 & 1/16 & 1/16 & 1/16 & 1/16 \end{array}\right\}$ 构造码。

(b)证明上述编码法得到的码满足异字头条件，且平均码长 $\bar{\ell}$ 满足 $H(U) \leq \bar{\ell} < H(U) + 1$。

上周作业解答

(a) $F_1 = 0$，$\ell_1 = \lceil \log \frac{1}{P(a_1)} \rceil = 2$ 则符号a_1的码字为00

$F_2 = \frac{1}{4}$ ，$\ell_2 = \lceil \log \frac{1}{P(a_2)} \rceil = 2$ 则符号a_2的码字为$01000\cdots$截取前2位为01

$F_3 = \frac{1}{2}$ ，$\ell_3 = \lceil \log \frac{1}{P(a_3)} \rceil = 3$ 则符号a_3的码字为$10000\cdots$截取前3位为100

$F_4 = \frac{1}{2} + \frac{1}{8}$ ，$\ell_4 = \lceil \log \frac{1}{P(a_4)} \rceil = 3$ 则符号a_4的码字为$10100\cdots$截取前3位为101

$F_5 = \frac{1}{2} + \frac{1}{4}$ ，$\ell_5 = \lceil \log \frac{1}{P(a_5)} \rceil = 4$ 则符号a_5的码字为$11000\cdots$截取前3位为1100

$F_6 = \frac{1}{2} + \frac{1}{4} + \frac{1}{16}$，$\ell_6 = \lceil \log \frac{1}{P(a_6)} \rceil = 4$ 则符号a_6的码字为$11010\cdots$截取前3位为1101

$F_7 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$ ，$\ell_7 = \lceil \log \frac{1}{P(a_7)} \rceil = 4$ 则符号a_7的码字为$11100\cdots$截取前3位为1110

$F_8 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}$ ，$\ell_8 = \lceil \log \frac{1}{P(a_8)} \rceil = 4$ 则符号a_8的码字为$11110\cdots$截取前3位为1111

上周作业解答

(b) 题干中的编码实际为二进制下的 Shannon 码，而满足异字头条件即满足前缀码条件，因此题目等价于证明 Shannon 码为前缀码。证明如下：

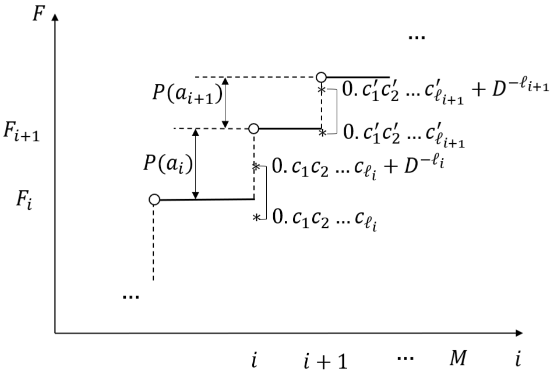

Figure: Shannon Code 图解

上周作业解答

在本题中， $D=2$。

方法1. 一个码字对应的区间长度为

$$2^{-\ell_i}=2^{-\lceil\log\frac{1}{P(a_i)}\rceil}\leq P(a_i)\leq P(a_{i-1})$$

而该区间的下端点小于等于 F_i，所以，该区间的上端点开区间不会包含 F_{i+1}。又因为该区间的上端点严格大于 F_i，所以该区间的下端点一定严格大于 F_{i-1}。

所以，可以知道，第 i 个码字的区间 $[0.c_1c_2\cdots c_{\ell_i},\ 0.c_1c_2\cdots c_{\ell_i}+D^{-\ell_i})$ 只包含累积分布函数 F 的一个阶跃点，这意味着一个码字对应一个符号。在这个结论基础上，若能证明相邻区间没有重叠的部分，便意味着各码字之间不互为前缀，证明如下:

上周作业解答

假设相邻区间有重叠的部分，则点 $0.c_1^{'}c_2^{'}\cdots c_{\ell_{i+1}}^{'} < 0.c_1c_2\cdots c_{\ell_i} + D^{-\ell_i}$。可以知道，一个区间内的所有点对应的序列都以下端点 $c_1c_2\cdots c_{\ell_{i+1}}$ 为前缀，因此，以下端点为前缀的所有序列都在对应的区间中。若点 $0.c_1^{'}c_2^{'}\cdots c_{\ell_{i+1}}^{'}$ 落在第 i 个码字的区间内，则意味着 $c_1^{'}c_2^{'}\cdots c_{\ell_{i+1}}^{'}$ 以 $c_1c_2\cdots c_{\ell_i}$为前缀。

因此，所有以 $c_1^{'}c_2^{'}\cdots c_{\ell_{i+1}}^{'}$ 为前缀的序列都以 $c_1c_2\cdots c_{\ell_i}$ 为前缀，因而都在第 i 个码字的区间。所以第 $i+1$ 个码字的小区间在 i 个码字的小区间内，第 i 个区间包含两个阶跃点，与刚才的结论矛盾。

所以各码字对应区间无重叠部分，Shannon 码是前缀码。

上周作业解答

方法2. 用 $c|_\ell$ 表示数 c 在(二进制)小数位上长度为 ℓ 的截断，即 $0.c_1c_2\cdots|_\ell = 0.c_1c_2\cdots c_\ell$ 。

由Shannon Code图解可得，对于第 $i+1$ 个区间和第 i 个区间的端点而言，有：

$$
\begin{aligned}
0.c_1'c_2'\cdots c_{\ell_{i+1}}' &= (0.c_1c_2\cdots + P(a_i))|_{\ell_{i+1}} \\
&\geq 0.c_1c_2\cdots c_{\ell_{i+1}} + P(a_i)|_{\ell_{i+1}} \\
&\geq 0.c_1c_2\cdots c_{\ell_{i+1}} + 2^{-\ell_i} \\
&\geq 0.c_1c_2\cdots c_{\ell_i} + 2^{-\ell_i}
\end{aligned}
$$

所以，第 $i+1$ 个区间的下端点大于等于第 i 个区间的上端点，相邻区间不重合，因此Shannon 码是前缀码。

上周作业解答

方法3. 假设上述方法求得的码不满足异字头条件，设a_i的码字为a_{i+k}的前缀，则有

$$F_{i+k} = F_i + P(a_i) + \cdots + P(a_{i+k-1})$$

那么

$$P(a_i) \leq F_{i+k} - F_i < 2^{-\ell_i}$$

又由题意得

$$\ell_i \geq \log \frac{1}{P(a_i)}$$

则有

$$P(a_i) \geq 2^{-\ell_i}$$

与题目条件矛盾。则用shannon码编码方法得到的码满足异字头条件。

上周作业解答

证明完上述编码法得到的码满足异字头条件后，现在证明平均码长 $\bar{n}$ 满足的条件。

前面所求得的平均码长

$$\bar{\ell} = \sum_{i=1}^{8} P(a_i) \lceil \log \frac{1}{P(a_i)} \rceil$$

易得

$$\bar{\ell} = \sum_{i=1}^{8} P(a_i) \lceil \log \frac{1}{P(a_i)} \rceil \geq H(U) = \sum_{i=1}^{8} P(a_i) \log \frac{1}{P(a_i)}$$

上周作业解答

又

$$\begin{aligned}
\bar{\ell} &= \sum_{i=1}^{8} P(a_i)\lceil \log \frac{1}{P(a_i)} \rceil \\
&< \sum_{i=1}^{8} P(a_i)(\log \frac{1}{P(a_i)} + 1) \\
&= \sum_{i=1}^{8} P(a_i) \log \frac{1}{P(a_i)} + \sum_{i=1}^{8} P(a_i) \\
&= H(U) + 1
\end{aligned}$$

则有

$$H(U) \leq \bar{\ell} < H(U) + 1$$

上周作业解答

Exercise 3.

利用拉格朗日乘子法，证明离散随机变量 X (取有限 $|\mathcal{X}|$ 个值)的熵 $H(X) \leq \log |\mathcal{X}|$。

上周作业解答

Answer:

令$n=|\mathcal{X}|, f(p_1,p_2,\cdots,p_n)=H(X)=-\sum_{i=0}^{n}p_i\log p_i$
且由题意，有约束条件为$g(p_1,p_2,\cdots,p_n)=\sum_{i=0}^{n}p_i=1$

设$F(p_1,p_2,\cdots,p_n,\lambda)=f(p_1,p_2,\cdots,p_n)+\lambda\left[g(p_1,p_2,\cdots,p_n)-1\right]$

对$F(p_1,p_2,\cdots,p_n,\lambda)$求偏导得到$\frac{\partial F(p_1,p_2,\cdots,p_n,\lambda)}{\partial p_i}=-\log p_i-\frac{1}{\ln 2}+\lambda$

$\frac{\partial F(p_1,p_2,\cdots,p_n,\lambda)}{\partial \lambda}=\sum_{i=0}^{n}p_i-1$

令偏导数为0，得到$p_1=p_2=\cdots=p_n=2^{\frac{1}{\ln 2}+\lambda}$，$\sum_{i=0}^{n}p_i-1=0$

此时离散随机变量X的所有取值概率相等，为$\frac{1}{n}$时，易知熵达到极大值，同时极大值为最大值，最大值为$\log n=\log|\mathcal{X}|$，
即$H(X)\le\log|\mathcal{X}|$。