信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. Classes of Codes
2. Kraft Inequality
3. Optimal Codes: minimal average codeword length
   - Shannon Code
4. Relative Entropy

Consider a discrete memoryless source X with distribution $P_X(x), x \in \mathcal{X}$. A code of source $\mathcal{X}$ over the alphabet $\mathcal{D}$ consists of the following essentials.

$$\begin{aligned} \text{Encoding} \quad \phi: \quad & \mathcal{X} \to \mathcal{D}^* \\ & x \mapsto c(x) \\ \text{Average length} \quad & L = \sum_{x \in \mathcal{X}} P_X(x)\ell(c(x)) \end{aligned}$$

Definition 1

1. A source code is called non-singular if $c(x) \neq c(x')$ for $x \neq x'$.
2. A source code is called uniquely decodable if $c(\mathbf{x}) \neq c(\mathbf{x}')$ for $\mathbf{x} \neq \mathbf{x}'$, where $c(\mathbf{x}) = c(x_1)c(x_2)\ldots c(x_n)$.
3. A source code is called to be a prefix code or instantaneous code if no codeword is a prefix of any other codeword.

Table: source codes

X	Singular	non-singular	Uniquely decodable	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

Table: source codes

X	Singular	non-singular	Uniquely decodable	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

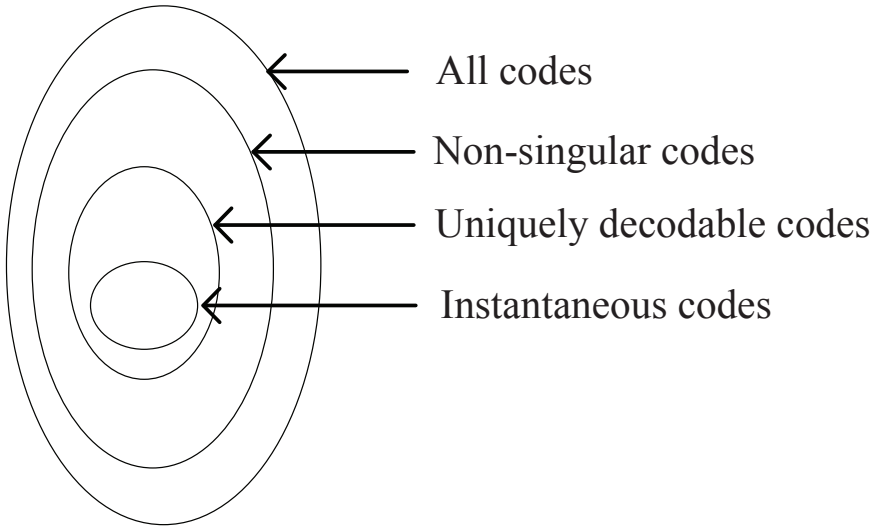

Figure: Classes of codes.

Theorem 2 (Kraft Inequality)

For any instantaneous code over the alphabet $\mathcal{D}$, the codeword lengths $\ell_1, \ell_2, \ldots, \ell_M$ must satisfy the inequality

$$\sum_{i=1}^{M} D^{-\ell_i} \leq 1. \tag{1}$$

Conversely, given a set of codeword lengths that satisfy this equality, there exists an instantaneous code with these word lengths.

Outline of proof:

(1) Consider a D-ary tree in which each node has D children. In the tree, the branches represent the symbols of the codeword; the path from the root traces out the symbols of the codeword; each codeword is represented by a leaf. The following figure depicts a 3-ary code tree.

Outline of proof:

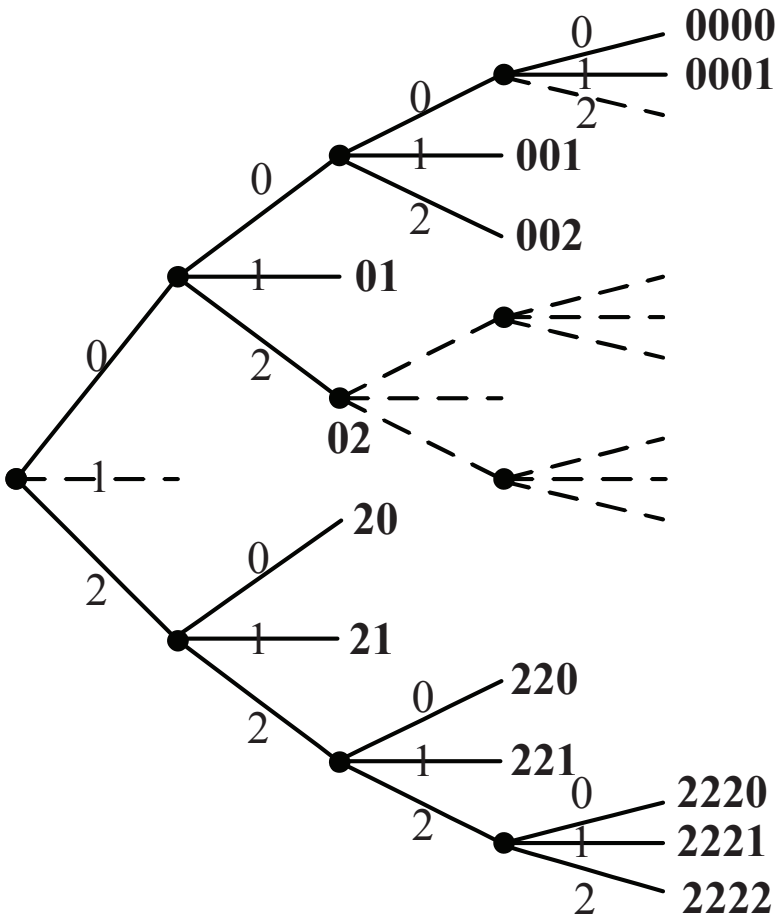

Figure: 3-ary code tree.

Outline of proof:

(2) For a prefix code, no codeword is an ancestor of any other codeword on the tree.

(3) Let ℓ_{max} be the length of the longest codeword. There are at most $D^{\ell_{max}}$ nodes at the ℓ_{max}th level of the code tree. A codeword at level ℓ_i has $D^{\ell_{max}-\ell_i}$ descendants at level ℓ_{max}. Hence

$$\sum_{i=1}^{M} D^{\ell_{max}-\ell_i} \leq D^{\ell_{max}} \tag{2}$$

or

$$\sum_{i=1}^{M} D^{-\ell_i} \leq 1 \tag{3}$$

(4) Conversely, we reorder the indexing such that $\ell_1 \leq \ell_2 \leq \ldots \leq \ell_M$ and then construct a tree by a full D-ary tree with level ℓ_M. Firstly, label the first node of depth ℓ_1 as codeword 1, and remove its descendants from the tree. Secondly, label the first remaining node of depth ℓ_2 as codeword 2, etc. Then a prefix code is constructed.

Outline of proof:

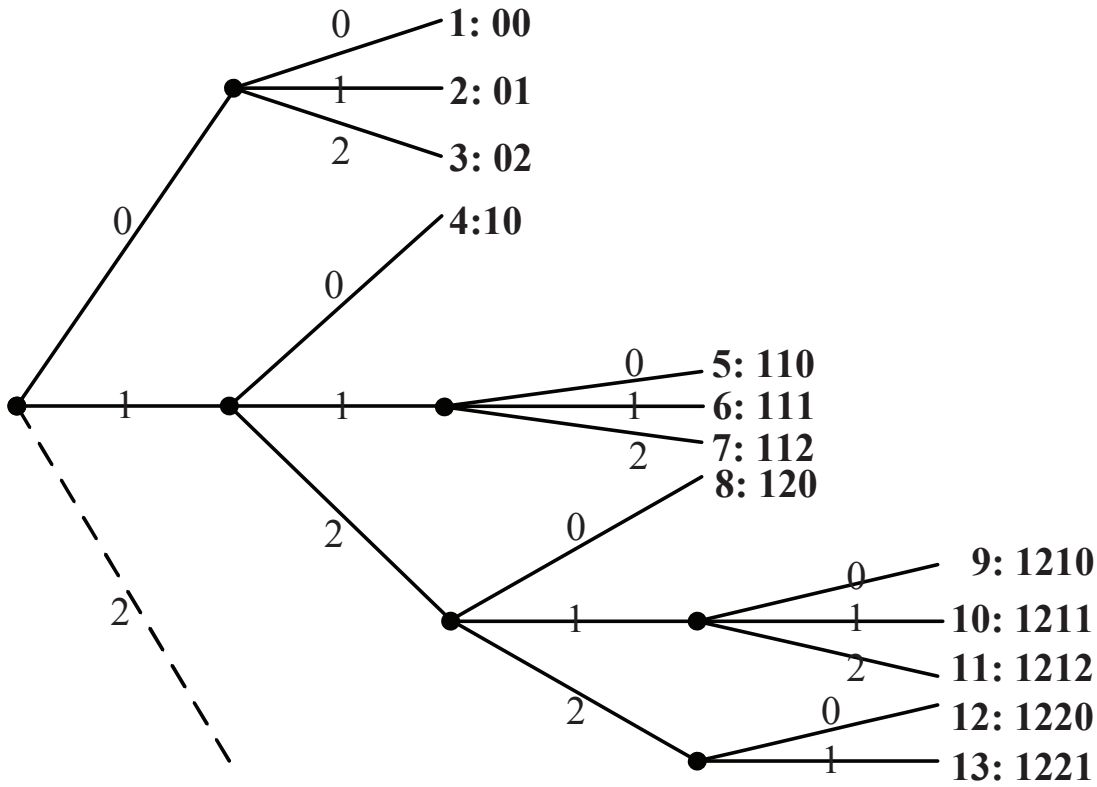

Figure: 3-ary prefix code tree.

Theorem 3 (Extended Kraft Inequality)

For any instantaneous code with countably infinite codewords over the alphabet $\mathcal{D}$, the codeword lengths $\ell_1, \ell_2, \ldots$ must satisfy the inequality

$$\sum_{i=1}^{\infty} D^{-\ell_i} \leq 1. \tag{4}$$

Conversely, given a set of codeword lengths that satisfy this equality, there exists an instantaneous code with these word lengths.

Outline of proof:

(1) Denote the ith codeword by $c_1 c_2 \ldots c_{\ell_i}$. Let $0.c_1 c_2 \ldots c_{\ell_i}$ be the real number given by the D-ary expansion $0.c_1 c_2 \ldots c_{\ell_i} = \sum_{j=1}^{\ell_i} c_j D^{-j}$. This codeword corresponds to the interval

$$[0.c_1 c_2 \ldots c_{\ell_i}, 0.c_1 c_2 \ldots c_{\ell_i} + D^{-\ell_i}), \tag{5}$$

the set of all real numbers whose D-ary expansion begins with $0.c_1 c_2 \ldots c_{\ell_i}$.

Outline of proof:

(2) It is a subinterval of the unit interval $[0, 1]$ and of length $D^{-\ell_i}$. By the prefix condition, these intervals are disjoint. Hence $\sum_{i=1}^{\infty} D^{-\ell_i} \leq 1$.

(3) Conversely, we reorder the indexing such that $\ell_1 \leq \ell_2 \leq \ldots$. Then disjoint subintervals of the unit interval are constructed, and a prefix code is constructed by assign the low ends of these subintervals as codewords.
Take a 3-ary code with codeword lengths $\ell_1 = 1$, $\ell_2 = 3$, $\ell_3 = 4$, ... as an example. The subintervals are
$[0, 3^{-1})$, $[3^{-1}, 3^{-1} + 3^{-3})$, $[3^{-1} + 3^{-3}, 3^{-1} + 3^{-3} + 3^{-4})$, ...
and the codewords are 0, 10, 110, ...

Kraft inequality for the class of uniquely decodable codes

Remark: The uniquely decodable codes also satisfy Kraft inequality (see Theorem 5.5.1 and its corollary in pp.117-118 in [Cover2006]). This implies that the class of uniquely decodable codes does not offer any further choices for the set of codeword lengths than the class of prefix codes. Hence the bounds derived on the optimal codeword lengths continue to hold for uniquely decodable codes.

$$\left[\sum_{i=1}^{m} D^{-\ell(c(x_i))}\right]^k = \sum_{i=1}^{k\ell_m} D_i D^{-i}$$

其中 D_i 是码长之和为 i 的 k 长信源符号序列的个数，所以 $D_i \leq D^i$，因此 $\left[\sum_{i=1}^{m} D^{-\ell(c(x_i))}\right]^k \leq k\ell_m$。

An interesting problem is to find a prefix code with the minimal average codeword length. That is,

$$\min_{\{\ell_i\}} \sum_i p_i \ell_i, \text{ subject to } \sum_i D^{-\ell_i} \leq 1. \tag{6}$$

By the method of Lagrange multipliers to optimize this objective, we obtain optimal length set as

$$\ell_i^* = -\log_D p_i, \tag{7}$$

and the optimal length is

$$L^* = -\sum_{x_i} p(x_i) \log_D p(x_i) = H_D(X). \tag{8}$$

However, ℓ_i^* may not be an integer!

Theorem 4 (Minimal Length)

The average length L^* of the optimal prefix code over a D-ary alphabet for the source $\mathcal{X}$ must satisfy

$$H_D(X) \leq L^* \leq H_D(X) + 1. \tag{9}$$

Outline of proof:

Firstly, we prove that, for any prefix code, $H_D(X) \leq L$ with a method different from the preceding discussion. Let $\mathbf{r} = \{r_i\}$ with $r_i = D^{-\ell_i} / \sum_j D^{-\ell_j}$. Then $\mathbf{r}$ is a probability vector and we have

$$\begin{aligned} L - H_D(X) &= \sum_{x_i} p(x_i)\ell(c(x_i)) + \sum_{x_i} p(x_i) \log_D p(x_i) \\ &= \sum_i p_i \log_D D^{\ell_i} + \sum_i p_i \log p_i \\ &= \sum_i p_i \log_D(\frac{p_i}{r_i}) - \sum_i p_i \log_D(\sum_j D^{-\ell_j}) \\ &= D(p_{\mathbf{x}} \| \mathbf{r}) - \log_D(\sum_j D^{-\ell_j}) \geq 0 \end{aligned}$$

since the prefix code satisfies the Kraft inequality $\sum_j D^{-\ell_j} \leq 1$.

Outline of proof:

Secondly, we prove that there exists a prefix code with $L \leq H_D(X) + 1$. From the optimization problem, we know that the optimal lengths should be $\ell_i^* = \log_D \frac{1}{p_i}$. Since ℓ_i^* may not be an integer, we set

$$\ell_i = \lceil \log_D \frac{1}{p_i} \rceil$$

where $\lceil x \rceil$ is the smallest integer not less than x. Then

$$\log_D \frac{1}{p_i} \leq \ell_i \leq \log_D \frac{1}{p_i} + 1.$$

These lengths satisfy the Kraft inequality since

$$\sum D^{-\lceil \log_D \frac{1}{p_i} \rceil} \leq \sum D^{-\log_D \frac{1}{p_i}} = \sum p_i = 1.$$

Then we can construct a prefix code with these lengths and the average length is

$$L = \sum_i p_i \lceil \log_D \frac{1}{p_i} \rceil < \sum_i p_i (\log_D \frac{1}{p_i} + 1) = H_D(X) + 1.$$

Hence the average codeword length L^* of the optimal prefix code satisfies $L^* \leq L \leq H_D(X) + 1$.

Shannon Code

Consider the following method for generating a code for a random variable X that takes on m values $\{1, 2, ..., m\}$ with probabilities $p_1, p_2, \cdots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to ℓ_i bits, where $\ell_i = \lceil \log \frac{1}{p_i} \rceil$.
(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

Suppose that a probability mass function **q** is used in practice instead of the true probability mass function **p**, then the Shannon code is of lengths $\ell_i = \lceil \log \frac{1}{q_i} \rceil$.

Theorem 5

The average length under $\mathbf{p}$ *of the Shannon code assignment* $\ell_i = \lceil \log \frac{1}{q_i} \rceil$ *satisfies*

$$H(\mathbf{p}) + D(\mathbf{p}\|\mathbf{q}) \leq \mathbf{E}_{\mathbf{p}}[\ell(X)] < H(\mathbf{p}) + D(\mathbf{p}\|\mathbf{q}) + 1.$$

Remark: Using the wrong distribution incurs a penalty of $D(\mathbf{p}\|\mathbf{q})$ in the average description length.

Outline of proof: We have

$$\begin{aligned}
\mathbf{E_p}[\ell(X)] &= \sum p_i \lceil \log \frac{1}{q_i} \rceil < \sum p_i (\log \frac{1}{q_i} + 1) \\
&= \sum p_i \log(\frac{p_i}{q_i}\frac{1}{p_i}) + 1 = \sum p_i \log \frac{p_i}{q_i} + \sum p_i \log \frac{1}{p_i} + 1 \\
&= D(\mathbf{p}\|\mathbf{q}) + H(\mathbf{p}) + 1,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E_p}[\ell(X)] &= \sum p_i \lceil \log \frac{1}{q_i} \rceil \geq \sum p_i \log \frac{1}{q_i} \\
&= \sum p_i \log(\frac{p_i}{q_i}\frac{1}{p_i}) = \sum p_i \log \frac{p_i}{q_i} + \sum p_i \log \frac{1}{p_i} \\
&= D(\mathbf{p}\|\mathbf{q}) + H(\mathbf{p}).
\end{aligned}$$

Relative Entropy

The *relative entropy* or *Kullback Leibler distance* between two pmfs P and Q over the finite set $\mathcal{X}$ is defined as

$$D(P\|Q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

- Usually, $D(P\|Q) \neq D(Q\|P)$.
- $D(P\|Q) \geqslant 0$.
- $D(P\|Q) = 0$ iff $p(x) = q(x)$ for all $x \in \mathcal{X}$. Equivalently, $D(P\|Q) > 0$ iff $p(x) \neq q(x)$ for some x.
- Let Q be an estimation of the true distribution P on the set $\mathcal{X}$. Then $D(P\|Q)$ can be considered as the penalty of source coding.

作业

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。

作业

Exercise 2.[王育明(2013)]

令离散无记忆信源

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

定义

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

而 $Q_1 = 0$，今按下述方法进行二元编码。消息 a_k 的码字为实数 Q_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 $1/2 \to 1000\cdots, 1/4 \to 0100\cdots$），保留的截短序列长度 n_k是大于或等于 $I(a_k)$的最小整数。

作业

(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 $\bar{n}$ 满足 $H(U) \leq \bar{n} < H(U) + 1$。

作业

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 $H(X) \leq \log |\mathcal{X}|$。

谢谢！