Source coding

Which horse won in the horse racing?

X	Pr	Code I	Code II
0	1/2	000	0
1	1/4	001	10
2	1/8	010	110
3	1/16	011	1110
4	1/64	100	111100
5	1/64	101	111101
6	1/64	110	111110
7	1/64	111	111111

$$H(X) = -\sum p_i \log p_i = 2\text{bits}$$

Which code is better?

Data compression

- We interpret that $H(X)$ is the best achievable data compression.
- We want to develop practical lossless coding algorithms that approach, or achieve the entropy limit $H(X)$.

Terminology

X	Pr	Code I	Code II
0	1/2	000	0
1	1/4	001	10
2	1/8	010	110
3	1/16	011	1110
4	1/64	100	111100
5	1/64	101	111101
6	1/64	110	111110
7	1/64	111	111111

- Source alphabet $\mathcal{X} = \{0, 1, 2, 3, 4, 5, 6, 7\}$.
- Code alphabet $\mathcal{D} = \{0, 1\}$.
- Codeword, e.g., 010 for $X = 2$ in Code 1.
- Codeword length, e.g., codeword length for Code 1 is 3.
- Codebook: all the codewords.

Source Coding

Notation (Alphabet Extension)

The set of all possible sequences based on a finite alphabet $\mathcal{D}$ is denoted by $\mathcal{D}^*$. E.g.,
$\mathcal{D} = \{0, 1\} \rightarrowtail \mathcal{D}^* = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$.

Definition (Source Code)

Let $\mathcal{X}$ be the alphabet of a random variable X, and $\mathcal{D}$ be the alphabet of code. A *source code* C for the random variable X is a map

$$\begin{aligned} C: \quad & \mathcal{X} \to \mathcal{D}^* \\ & x \mapsto C(x) \end{aligned}$$

where $C(x)$ is the codeword associated with x. Let $\ell(x)$ denote the length of C(x).

Definition

The *expected length* $L(X)$ of a source code C for a random variable X with probability mass function $p(x)$ is

$$L(X) = E\ell(X) = \sum_{x \in \mathcal{X}} p(x)\ell(x).$$

X	Pr	Code I	Code II
0	1/2	000	0
1	1/4	001	10
2	1/8	010	110
3	1/16	011	1110
4	1/64	100	111100
5	1/64	101	111101
6	1/64	110	111110
7	1/64	111	111111

$L_1(X) = 3$
$L_2(X) = 2$

Set of codes

For $\mathcal{X} = \{1, 2, 3, 4\}$ and $\mathcal{D} = \{0, 1\}$, consider

x	$p(x)$	C_I	C_{II}	C_{III}	C_{IV}
1	1/2	0	0	10	0
2	1/4	0	1	00	10
3	1/8	1	00	11	110
4	1/8	10	11	110	111
$H(X)$	1.75	–	–	–	–
$E\ell(X)$	–	1.125	1.25	2.125	1.75

- Code efficiency $= H(X)/E[\ell(X)]$
- Which code is best? Would we prefer C_I or C_{II}?

  Consider C_I and decode string: 00001. It would come from 1, 2, 1, 2, 3 or 2, 1, 2, 1, 3 or 1, 1, 1, 1, 3, or etc.

C_I与C_{II}会产生歧义

- Consider C_{III}. Can we decode 1100000000?

  Yes. But if we only see a prefix, such as 11, we don't know until we see more bits to the end.

  $1100000000 = 3, 2, 2, 2, 2$
  $11000000000 = 4, 2, 2, 2, 2$

C_{III}要读完最后一个bit才能解码

- Consider C_{IV}. This code seems at least feasible (since $E[\ell] \geq H$). Decoding seems easy: (e.g., $111110100 = 111, 110, 10, 0 = 4, 3, 2, 1$).

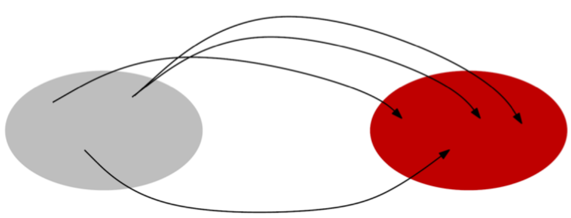
Code types

Definition (Nonsingular Code)

A code C is called *nonsingular* if every realization of $\mathcal{X}$ maps onto a difference codeword in $\mathcal{D}^*$, i.e.,

$$x \neq x' \Rightarrow C(x) \neq C(x').$$

不同的realization有不同的编码.



Definition (Code Extension)

The *extension* of a code $C\colon \mathcal{X} \to \mathcal{D}^*$ is defined by

$$C(x_1x_2 \cdots x_n) = C(x_1)C(x_2) \cdots C(x_n).$$

不同的sequence有不同的编码

Definition (Unique Decodable Code)

A code is called *uniquely decodable* if its extension is nonsingular.

u.d.比singular强.
区分可不可用，看其是否u.d.

$$x_1x_2 \ldots x_m \neq x_1'x_2' \ldots x_n' \Rightarrow C(x_1x_2 \ldots x_m) \neq C(x_1'x_2' \ldots x_n')$$

编码不互为前缀

区分好不好用，看其是否 prefix

Definition (Prefix Code)

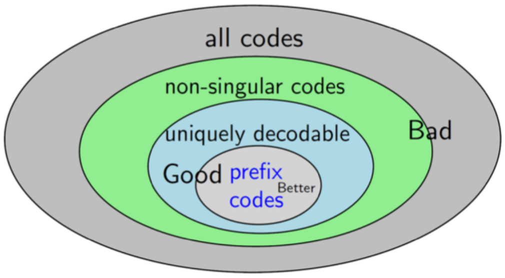
A code C is called a *prefix code* (a.k.a. *instantaneous*) iff no codeword of C is a prefix of any other codeword of C.

For $\mathcal{X} = \{1, 2, 3, 4\}$ and binary code, consider

x	$p(x)$	C_I	C_{II}	C_{III}	C_{IV}
1	1/2	0	0	10	0
2	1/4	0	1	00	10
3	1/8	1	00	11	110
4	1/8	10	11	110	111
$H(X)$	1.75	–	–	–	–
$E\ell(X)$	–	1.125	1.25	2.125	1.75

- C_I is singular.
- C_{II} is non-singular, but not uniquely decodable.
- C_{III} is non-singular, uniquely decodable, but NOT prefix.
- C_{IV} is non-singular, uniquely decodable, and prefix.

Classes of codes



- Goal: to find a prefix code with minimum expected length.

Kraft Inequality

Theorem 5.2.1 (Kraft Inequality)

For any prefix code over an alphabet of size D, the codeword lengths $\ell_1, \ell_2, \ldots, \ell_m$ must satisfy the inequality

$$\sum_i D^{-\ell_i} \le 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists a prefix code with these codeword lengths.

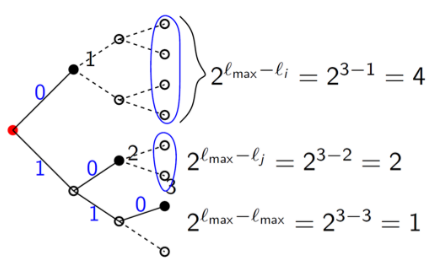
Proof Idea. (A small example) To prove: A prefix code with lengths $\ell_1, \ell_2, \ldots, \ell_m$, the inequality

$$\sum_i D^{-\ell_i} \le 1 \quad \text{holds.}$$

Depth: 0 1 2 3

x	$c(x)$
1	0
2	10
3	110

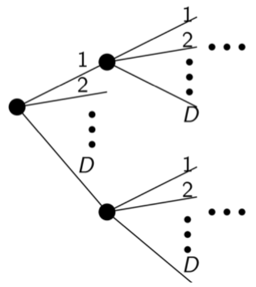
$\ell_{\max} = 3$



$$\sum_i 2^{-\ell_i} \le 1 \Leftarrow \sum_i 2^{\ell_{\max}-\ell_i} \le 2^{\ell_{\max}}$$

Proof. (in general)

- Represent the set of prefix codes on a D-ary tree:



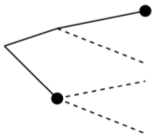
- Codewords correspond to leaves
- Path from root to each leaf determines a codeword
- Prefix condition: won't get to a codeword until we get to a leaf (no descendants of codewords are codewords)

到叶子节点的路径=编码

- $\ell_{\max} = \max_i(\ell_i)$ is the length of the longest codeword.

$l_{max} = \max_x l(x) =$ 树的深度

- We can expand the full-tree down to depth $\ell_{\max}$:



The nodes at the level $\ell_{\max}$ are either

1. codewords
2. descendants of codewords
3. neither

- Consider a codeword i at depth ℓ_i in tree
- There are $D^{\ell_{\max}-\ell_i}$ descendants in the tree at depth $\ell_{\max}$

对于在深度为l_i的编码节点$c(x_i)$，其在full expand的时候会生出$D^{l_{max}-l_i}$个后代

- Descendants of code i are disjoint from decedents of code j (prefix free condition)

$c(x_i)$与$c(x_j)$的full expand节点不会重合.

- All the above implies:

$$\sum_i D^{\ell_{\max}-\ell_i} \le D^{\ell_{\max}} \quad \Rightarrow \sum_i D^{-\ell_i} \le 1$$

因此，所有编码节点新长出的节点数之和，必然少于总共的节点数，也即$\sum_i D^{l_{max}-l_i} \le D^{l_{max}}$

Proof. (in general)

- Conversely: given codewords lengths $\ell_1, \ell_2, \ldots, \ell_m$ satisfying Kraft inequality, try to construct a prefix code.

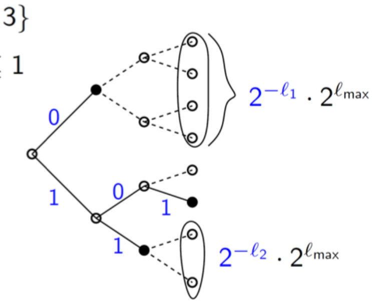
分配节点时从浅到深

$\{\ell_1, \ell_2, \ell_3\} = \{1, 2, 3\}$

$2^{-1} + 2^{-2} + 2^{-3} \le 1$

x	$c(x)$
1	0
2	11
3	101

C is prefix.



Outline

- **Extended Kraft inequality for prefix code**
- **Kraft inequality for uniquely decodable code**

  Uniquely decodable code does NOT provide more choices than prefix code

- **Bounds on optimal expected length**

  Entropy length is achievable when jointly encoding a random sequence.

Extended Kraft Inequality

Theorem 5.5.1 (Extended Kraft Inequality)

Kraft inequality holds also for all countably infinite set of codewords, i.e., the codeword lengths satisfy the extended Kraft inequality,

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

Conversely, given any $\ell_1, \ell_2, \ldots$ satisfying the extended Kraft inequality, we can construct a prefix code with these codeword lengths.

Theorem 5.2.2 (Extended Kraft Inequality)

Kraft inequality holds also for all countably infinite set of codewords.

Proof.

Consider the ith codeword $y_1 y_2 \cdots y_{\ell_i}$. Let $0.y_1 y_2 \cdots y_{\ell_i}$ be the real number given by the D-ary expansion

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

which corresponds to the interval

$$[0.y_1 y_2 \cdots y_{\ell_i}, 0.y_1 y_2 \cdots y_{\ell_i} + \frac{1}{D^{\ell_i}}).$$

□

不同codeword映射到的区间是互不相交的.

如 $[0.y_1, 0.y_1 + \frac{1}{D})$ 与 $[0.y_1'y_2, 0.y_1'y_2 + \frac{1}{D^2})$ 是不相交的 (因为 $y_1 \neq y_1'$)

Proof. (cont.)

By the prefix condition, these intervals are disjoint in the unit interval $[0, 1]$. Thus, the sum of their lengths is ≤ 1. This proves that

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

For converse, reorder indices in increasing order and assign intervals as we walk along the unit interval. □

Conversely, 有长度 $l_1, l_2, \ldots$ 满足 $\sum_i D^{-l_i} \leq 1$.

重排使得 $l_1 \leq l_2 \leq \cdots$

第一个区间: $[0, 0+2^{-l_1}) = [0, 0.\overbrace{00\ldots1}^{l_1})$

第二个区间: $[2^{-l_1}, 2^{-l_1}+2^{-l_2}) = [0.\underbrace{00\ldots1}_{l_1}, 0.\overbrace{\underbrace{00\ldots1}_{l_1}\ldots1}^{l_2})$

⋮

第n个区间: $[\sum_{i=1}^{n-1} 2^{-l_i}, \sum_{i=1}^{n} 2^{-l_i})$

⋮

Kraft Inequality for Uniquely Decodable Codes

Theorem 5.2.3 (McMillan)

The codeword lengths of any uniquely decodable D-ary code must satisfy the Kraft inequality

$$\sum D^{-\ell_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.

Proof.

Consider C^k, the k-th extension of the code by k repetitions. Let the codeword lengths of the symbols $x \in \mathcal{X}$ be $\ell(x)$. For the k-th extension code, we have

$$\ell(x_1, x_2, \dots, x_k) = \sum_i^k \ell(x_i).$$

Proof. (cont.)

Consider

$$\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^k = \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} D^{-\ell(x_1)}D^{-\ell(x_2)}\dots D^{-\ell(x_k)}$$
$$= \sum_{x_1,x_2,\cdots x_k\in\mathcal{X}^k} D^{-\ell(x_1)}D^{-\ell(x_2)}\dots D^{-\ell(x_k)}$$
$$= \sum_{x^k\in\mathcal{X}^k} D^{-\ell(x^k)}$$

Proof. (cont.)

Let $\ell_{\max}$ be the maximum codeword length and $a(m)$ is the number of source sequences x^k mapping into codewords of length m. Unique decodability implies that $a(m) \leq D^m$. We have

$$\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^k = \sum_{x^k\in\mathcal{X}^k} D^{-\ell(x^k)} = \sum_{m=1}^{k\ell_{\max}} a(m)D^{-m}$$
$$\leq \sum_{m=1}^{k\ell_{\max}} D^m D^{-m}$$
$$= k\ell_{\max}$$

Proof. (cont.)

$$\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^k \leq k\ell_{\max}.$$

Hence,

$$\sum_j D^{-\ell_j} \leq (k\ell_{\max})^{1/k}$$

holds for all k. Since the RHS$\to 1$ as $k\to\infty$, we prove the Kraft inequality. For the converse part, we can construct a prefix code as in **Theorem 5.2.1**, which is also uniquely decodable.

$$x^k = x_1x_2\dots x_k \Rightarrow l(x^k) = l(x_1)+l(x_2)+\dots+l(x_k)$$

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{x_1} D^{-l(x_1)} \cdot \sum_{x_2} D^{-l(x_2)} \cdots \sum_{x_k} D^{-l(x_k)}$$
$$= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_k} D^{-[l(x_1)+l(x_2)+\dots+l(x_k)]}$$

令codeword的最大长度为l_{max}，则
$\sum_{i=1}^k l(x_i) \in [k, kl_{max}] \subset [1, kl_{max}]$

令$a(m)$为codeword长度是m的x^k的数量，也即满足$\sum_{i=1}^k l(x_i)=m$的$(x_1,x_2,\dots,x_k)$的数量。

那么 $$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{x^k\in\mathcal{X}^k} D^{-l(x^k)}$$
$$= \sum_{m=1}^{kl_{max}} a(m) D^{-m}$$

因为长度为m的D进制编码数最多为D^m，即$a(m)\leq D^m$，所以 (这时编码方案为uniquely decodable)

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k \leq \sum_{m=1}^{kl_{max}} D^m D^{-m}$$
$$= kl_{max}, \forall k=1,2,\dots,\infty$$

$$\Rightarrow \sum_{x\in\mathcal{X}} D^{-l(x)} \leq (kl_{max})^{\frac{1}{k}} = k^{\frac{1}{k}} l_{max}^{\frac{1}{k}} \xrightarrow{k\to\infty} 1$$

从该定理可知u.d编码和prefix编码是一一对应的.

Optimal Codes

Problem To find the set of lengths $\ell_1, \ell_2, \ldots, \ell_m$ satisfying the Kraft inequality and whose expected length $L = \sum p_i \ell_i$ is minimized.

Optimization:

minimize $L = \sum p_i \ell_i$

subject to $\sum D^{-\ell_i} \le 1$ and ℓ_i's are integers.

Theorem 5.3.1

The expected length L of any prefix D-ary code for a random variable X is no less than $H_D(X)$, i.e.,

$$L \ge H_D(X),$$

with equality iff $D^{-\ell_i} = p_i$.

→对数的底为D

Proof.

$$L - H_D(X) = \sum p_i \ell_i - \sum p_i \log_D \frac{1}{p_i}$$

$$= -\sum p_i \log_D D^{-\ell_i} + \sum p_i \log_D p_i$$

$$= \sum p_i \log_D \frac{p_i}{r_i} - \log_D c$$

$$= D(\mathbf{p} \| \mathbf{r}) + \log_D \frac{1}{c} \ge 0$$

"=" holds if $c = 1$ and $r_i = p_i$.

where $r_i = D^{-\ell_i} / \sum_j D^{\ell_j}$ and $c = \sum D^{-\ell_i} \le 1$. □

$L - H_D(X) = \Sigma p_i l_i - \Sigma p_i \log \frac{1}{p_i} = -\Sigma p_i \log D^{-l_i} + \Sigma p_i \log p_i$

$= \Sigma p_i \log_D \frac{p_i}{D^{-l_i}}$

令 $c = \Sigma D^{-l_i}$, $r_i = \frac{D^{-l_i}}{c}$, 也即 r_i 是 D^{-l_i} 的归一化.

$= \Sigma p_i \log_D \frac{p_i}{r_i} + \Sigma p_i \log_D \frac{1}{c}$

$= D(p \| r) - \log_D c$

由于 $c = \Sigma D^{-l_i} \le 1$, 则 $-\log_D c \ge 0$. 且有 $D(p\|r) \ge 0$, 故 ≥ 0.

取等时 $\begin{cases} D(p\|r) = 0 \Rightarrow p_i = r_i, \forall i \\ c = 1 \Rightarrow \sum_i D^{-l_i} = 1 \end{cases}$

$\Downarrow$

$r_i = D^{-l_i} \longrightarrow p_i = D^{-l_i}, \forall i$

$\Rightarrow L \ge H_D(X)$

Definition

A probability distribution is called D-adic if each of the probabilities is equal to D^{-n} for some n. Thus, we have equality in the theorem iff the distribution of X is D-adic.

Remark

$H_D(X)$ is a lower bound on the optimal code length. The equality holds iff p is D-adic.

Bound on the Optimal Code Length

Theorem 5.4.1 (Shannon Codes)

Let $\ell_1^*, \ell_2^*, \ldots, \ell_m^*$ be optimal codeword lengths for a source distribution $\mathbf{p}$ and a D-ary alphabet, and let L^* be the associated expected length of an optimal code ($L^* = \sum p_i \ell_i^*$). Then

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

Proof.

Take $\ell_i = \lceil -\log_D p_i \rceil$. Since

$$\sum_{i \in \mathcal{X}} D^{-\ell_i} \le \sum p_i = 1,$$

these lengths satisfy Kraft inequality and we can create a prefix code. Thus,

$$L^* \le \sum p_i \lceil -\log_D p_i \rceil$$

$$< \sum p_i (-\log_D p_i + 1)$$

$$= H_D(X) + 1.$$

□

Theorem 5.4.2

Consider a system in which we send a sequence of n symbols from X. The symbols are assumed to be i.i.d. according to p(x). The minimum expected codeword length per symbol satisfies

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq L_n^* < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}.$$

Proof.

First,

$$L_n = \frac{1}{n}\sum p(x_1, x_2, \ldots, x_n)\ell(x_1, x_2, \ldots, x_n) = \frac{1}{n}E[\ell(X_1, X_2, \ldots, X_n)]$$

We also have

$$H(X_1, X_2, \ldots, X_n) \leq E[\ell(X_1, X_2, \ldots, X_n)] < H(X_1, X_2, \ldots, X_n) + 1.$$

Since $X_1, X_2, \ldots, X_n$ are i.i.d., $H(X_1, X_2, \ldots, X_n) = nH(X)$. □

$\Rightarrow H(X) \leq L_n^* < H(X) + \frac{1}{n}$. 当 $n \to \infty$ 时, $L_n^* \to H(X)$

将sequence看作新的随机变量, 应用上一个定理也能证出.