Math 5251 Kraft & McMillan inequalities (§3.2)

Can we arbitrarily specify the word lengths $\ell_1, \ldots, \ell_m$ for a code $\mathcal{C} = \{w_1, \ldots, w_m\}$ on alphabet Σ of size n ?

called an n-ary alphabet/code

e.g. $\Sigma = \{0,1\}$ binary = 2-ary

$\Sigma = \{0,1,2\}$ ternary = 3-ary

EXAMPLE $\mathcal{C} = \{0, 1, 20, 21, 22\}$ on $\Sigma = \{0,1,2\}$ $n=3$

has $(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5)$ $m=5$

$= (1, 1, 2, 2, 2)$

Certainly not arbitrarily, e.g. if $\Sigma = \{0,1\}$

then $(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5) = (2,2,2,2,2)$

is impossible since Σ^* has only 4 words of length 2 :

00
01
10
11

If we further insist on the code being uniquely decipherable, it imposes even more of a constraint on $(\ell_1, \dots, \ell_m)$; interestingly it's the same constraint for codes that are prefix.

THEOREM Let Σ be an alphabet with n letters, and $(\ell_1, \ell_2, \dots, \ell_m)$ positive integers.

(a) (Kraft) If $\displaystyle\sum_{i=1}^{m} \frac{1}{n^{\ell_i}} = \frac{1}{n^{\ell_1}} + \frac{1}{n^{\ell_2}} + \dots + \frac{1}{n^{\ell_m}} \leq 1$

then $\exists$ a prefix (instantaneous) code $\mathcal{C}$ on Σ with those lengths.

(b) (McMillan) If $\exists$ a uniquely decipherable code $\mathcal{C}$ on Σ with those lengths,

then $\displaystyle\sum_{i=1}^{m} \frac{1}{n^{\ell_i}} \leq 1$

SAME inequality for both! So one concludes

$\{$lengths of u.d. n-ary codes$\}$ $=$ $\{$lengths of prefix n-ary codes$\}$

$= \left\{(\ell_1, \dots, \ell_m) \text{ with } \displaystyle\sum_i \frac{1}{n^{\ell_i}} \leq 1\right\} =$

Kraft-McMillan inequality

EXAMPLES If $n = 3 = |\Sigma|$, say $\Sigma = \{0,1,2\}$

then $\nexists$ any u.d. code $\mathcal{C}$ with

word lengths $(1,1,2,2,2,3)$ because

$$\frac{1}{3^1} + \frac{1}{3^1} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^3} = \frac{9+9+3+3+3+1}{27} = \frac{28}{27} > 1$$

On the other hand, there does $\exists$ a prefix code $\mathcal{C}$ with lengths $(1,2,2,2,2,3,3)$ because

$$\frac{1}{3^1} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^3} = \frac{9+3+3+3+3+1+1}{27} = \frac{23}{27} \leq 1$$

In fact, let's prove Kraft first, via an algorithm to find $\mathcal{C}$. Assuming $(l_1, \dots, l_m)$ has t_i occurrences of length i, then the inequality assumes

$$\sum_{i=1}^{m} \frac{1}{n^{l_i}} = \frac{t_1}{n^1} + \frac{t_2}{n^2} + \frac{t_3}{n^3} + \dots \leq 1$$

and we try to pick the shorter words first.

EXAMPLE $(\ell_1,\dots,\ell_m) = (\underbrace{1}_{t_1=1},\underbrace{2,2,2,2}_{t_2=4},\underbrace{3,3}_{t_3=2})$

has $\dfrac{t_1}{3^1}+\dfrac{t_2}{3^2}+\dfrac{t_3}{3^3} = \dfrac{1}{3^1}+\dfrac{4}{3^2}+\dfrac{2}{3^3} \le 1$

$\Rightarrow \dfrac{t_1}{3^1} \le 1$

so $t_1 \le 3^1$

allowing us to pick t_1 words of length 1:

$$\left.\begin{array}{c} 0 \\ 1 \\ 2 \end{array}\right\} 3^1$$

1st pick $t_1=1$ of these

$\Rightarrow \dfrac{t_1}{3^1}+\dfrac{t_2}{3^2} \le 1$

so $3t_1 + t_2 \le 3^2$

$t_2 \le 3^2 - \underbrace{3t_1}$

this many length 2 words have a prefix from our length 1 choices

allowing us to pick t_2 words of length 2:

$$\left.\begin{array}{c} \cancel{00} \\ \cancel{01} \\ \cancel{02} \\ 10 \\ 11 \\ 12 \\ 20 \\ 21 \\ 22 \end{array}\right\} 3^2$$

$3t_1$ have a bad prefix

2nd pick $t_2=4$ of these

$\Rightarrow \dfrac{t_1}{3^1}+\dfrac{t_2}{3^2}+\dfrac{t_3}{3^3} \le 1$

so $3^2 t_1 + 3^1 t_2 + t_3 \le 3^3$

$t_3 \le 3^3 - \underbrace{(3^2 t_1 + 3^1 t_2)}$

this many length 3 words have a bad prefix

allowing us to pick t_3 words of length 3

$$\left.\begin{array}{l} \left.\begin{array}{c} \cancel{000} \\ \cancel{001} \\ \vdots \\ \cancel{022} \end{array}\right\} 3^2 t_1 \text{ bad} \\ \left.\begin{array}{c} \cancel{100} \\ \cancel{101} \\ \cancel{102} \\ \cancel{110} \\ \cancel{111} \\ \cancel{112} \\ \cancel{120} \\ \cancel{121} \\ \cancel{122} \\ \cancel{200} \\ \cancel{201} \\ \cancel{202} \end{array}\right\} 3t_2 \text{ bad} \\ \quad 210 \\ \quad 211 \\ \quad 212 \\ \quad \vdots \\ \quad 222 \end{array}\right\} 3^3$$

3rd pick $t_3=2$ of these

proof of Kraft's inequality:

If $(l_1, \dots, l_m)$ has t_i occurrences of i and

$$\frac{t_1}{n^1} + \frac{t_2}{n^2} + \frac{t_3}{n^3} + \dots = \sum_{i=1}^{m} \frac{1}{n^{l_i}} \leq 1$$

we show how to pick a prefix code $\mathcal{C}$ with those lengths. Assuming one has already picked the words of length $\leq i-1$, and show they leave $\geq t_i$ words of length i that avoid them as prefixes.

Previously one has picked

t_{i-1} of length $i-1$ $\rightsquigarrow$ create $n t_{i-1}$ with bad prefix

t_{i-2} of length $i-2$ $\rightsquigarrow$ create $n^2 t_{i-2}$ with bad prefix

$\vdots \qquad \vdots$

t_2 of length 2 $\rightsquigarrow$ create $n^{i-2} t_2$ with bad prefix

t_1 of length 1 $\rightsquigarrow$ create $n^{i-1} t_1$ with bad prefix

Since there are n^i words of length i in total using alphabet Σ, ...

this leaves

$$n^i - \left(n^{i-1}t_1 + n^{i-2}t_2 + \dots + n^2 t_{i-2} + n t_{i-1}\right)$$

words of length i from which to choose t_i for $\mathcal{C}$.

We claim the above quantity is at least t_i,

since $\dfrac{t_1}{n^1} + \dfrac{t_2}{n^2} + \dots + \dfrac{t_{i-2}}{n^{i-2}} + \dfrac{t_{i-1}}{n^{i-1}} + \dfrac{t_i}{n^i} \leq 1$

multiply by n^i

$$n^{i-1}t_1 + n^{i-2}t_2 + \dots + n^2 t_{i-2} + n t_{i-1} + t_i \leq n^i$$

i.e. $t_i \leq n^i - \left(n^{i-1}t_1 + n^{i-2}t_2 + \dots + n^2 t_{i-2} + n t_{i-1}\right)$

$\blacksquare$

proof of McMillan inequality:

Assume C is a uniquely decipherable n-ary code having t_i codewords of length i for $i = 1, 2, \dots, \ell$.

We want to show $\underbrace{\frac{t_1}{n^1} + \frac{t_2}{n^2} + \dots + \frac{t_\ell}{n^\ell}}_{\text{call this sum } A\text{; want } A \leq 1.} \leq 1$

IDEA: Instead, for each $p = 1, 2, 3, \dots$ we will show

$$A^p = \sum_{s=1}^{p\ell} \frac{c_s}{n^s}$$ for some coefficients $c_s \leq n^s$

$$\Rightarrow A^p \leq \sum_{s=1}^{p\ell} 1 = p\ell$$

take p^{th} root of both sides

$$\Rightarrow A \leq (p\ell)^{\frac{1}{p}}$$

$$\Rightarrow A \leq \lim_{p \to \infty} (p\ell)^{\frac{1}{p}} = 1 \text{, as desired}$$

Calculus interlude!

$$\lim_{p\to\infty} (p\ell)^{1/p} = \lim_{p\to\infty} e^{\ln(p\ell)^{1/p}} = e^{\lim_{p\to\infty} \frac{\ln(p)}{p} + \frac{\ln(\ell)}{p}}$$

$$\overset{\text{L'Hôpital}}{=} e^{\lim_{p\to\infty} \frac{1/p}{1} + 0}$$

$$= e^0 = 1$$

So for $\mathcal{C}$ u.d., we need to show

$$A := \frac{t_1}{n^1} + \frac{t_2}{n^2} + \dots + \frac{t_\ell}{n^\ell} \text{ has } A^p = \sum_{s=1}^{p\ell} \frac{c_s}{n^s} \text{ with } c_s \le n^s$$

In fact, we can interpret c_s as counting the number of messages $(w_1, w_2, \dots, w_p)$ of p words from $\mathcal{C}$ with a total length of s letters from Σ.

Since there are n^s strings in Σ^* with s letters, and $\mathcal{C}$ is uniquely decipherable, this shows $c_s \le n^s$; each string comes from at most one message.

ACTIVE LEARNING

If $W \xrightarrow{f} \{0,1,2\}^*$ has codewords $\mathcal{C} = \mathrm{im}(f) = \{\underbrace{0,1}_{\substack{t_1=1 \\ \text{of length } 1}} \mid \underbrace{20,21,22}_{\substack{t_2=3 \\ \text{of length } 2}}\}$,

(a) how many messages (w_1, w_2) with 2 source words end up with encodings $f^*(w_1,w_2) = f^*(w_1)f^*(w_2)$ that are

- 2 letters long (like $1|0$ or $0|1$) ?
- 3 letters long (like $1|21$ or $22|0$) ?
- 4 letters long ?
- 5 letters long ?

(b) How would those answers change if (Σ had more letters and)

$t_1 = 10$

$t_2 = 1000$?

(proof by) EXAMPLE that $A := \frac{t_1}{n^1} + \frac{t_2}{n^2} + \dots + \frac{t_\ell}{n^\ell}$

has $A^p = \sum_{s=1}^{p\ell} \frac{c_s}{n^s}$ with $c_s \le n^s$

where

c_s = number of messages $(w_1, w_2, \dots, w_p)$ of p words

from $\mathcal{C}$ with a total length of s letters from Σ.

EXAMPLE: $\mathcal{C} = \{\underbrace{0, 1}_{t_1=2}, \underbrace{20, 21, 22}_{t_2=3}\}$, $\Sigma = \{0,1,2\}$

$n=3$

$p=2$

$$\left(\frac{t_1}{3^1} + \frac{t_2}{3^2}\right)^2 = \frac{\overbrace{t_1 \cdot t_1}^{c_2}}{3^2} + \frac{\overbrace{(t_1 t_2 + t_2 t_1)}^{c_3}}{3^3} + \frac{\overbrace{t_2 \cdot t_2}^{c_4}}{3^4}$$

$$= \frac{2 \cdot 2}{3^2} + \frac{2 \cdot 3 + 3 \cdot 2}{3^3} + \frac{3 \cdot 3}{3^4}$$

0|0	0|20	20|0	20|20
0|1	0|21	20|1	20|21
1|0	0|22	21|0	20|22
1|1	1|20	21|1	21|20
	1|21	22|0	21|21
	1|22	22|1	21|22
			22|20
			22|21
			22|22

∎

RECAP: We showed

$\exists$ a u.d. n-ary code C with lengths $(l_1, \dots, l_m)$

(u.d. $\Leftarrow$ prefix)

$\exists$ a prefix n-ary code C with lengths $(l_1, \dots, l_m)$

McMillan

Kraft

$$\sum_{i=1}^{m} \frac{1}{n^{l_i}} \leq 1$$

so all 3 statements are equivalent.