Codes II

Prof. Dan A. Simovici

UMB

The lengths of the words of a finite prefix code are characterized in the next theorem.

Theorem

(Kraft-McMillan Inequality for Prefix Codes) *Let* $A = \{a_0, \ldots, a_{k-1}\}$ *be an alphabet. A prefix code L on the alphabet A with word lengths* $\ell_0, \cdots, \ell_{m-1}$ *exists if and only if*

$$\sum_{i=0}^{m-1} \frac{1}{k^{\ell_i}} \leq 1.$$

Proof

If $m = 1$, all words of L have the same length, so L is a prefix code; therefore, assume that $m > 1$.
Suppose that $\ell_0 \leq \cdots \leq \ell_{m-1}$, and consider the complete labeled ordered tree T of height ℓ_{m-1}. T has $k^{\ell_{m-1}}$ leaves. If a word of length ℓ_i is included in the code, none of its descendants in T can be labeled by a word in the code. Thus, a word of length ℓ_i eliminates $k^{\ell_{m-1}-\ell_i}$ leaves. The total number of excluded leaves is $\sum_{i=0}^{m-1} k^{\ell_{m-1}-\ell_i}$ and it cannot exceed $k^{\ell_{m-1}}$. Thus, we have

$$\sum_{i=0}^{m-1} k^{\ell_{m-1}-\ell_i} \leq k^{\ell_{m-1}}$$

or, equivalently,

$$\sum_{i=0}^{m-1} \frac{1}{k^{\ell_i}} \leq 1.$$

Proof (cont'd)

Conversely, suppose that we have the nonnegative integers $\ell_0, \ldots, \ell_{m-1}$ such that

$$\sum_{i=0}^{m-1} \frac{1}{k^{\ell_i}} \leq 1.$$

Equivalently, we have

$$\sum_{i=0}^{m-1} k^{\ell_{m-1}-\ell_i} \leq k^{\ell_{m-1}}. \tag{1}$$

Select a word x_0 of length ℓ_0 that is the label of a vertex v_0, to include in the code. This corresponds to $k^{\ell_{m-1}-\ell_0}$ leaves in the labeled ordered tree $\mathtt{T}$. Since $m > 1$, the inequality (1) gives $k^{\ell_{m-1}-\ell_0} < k^{\ell_{m-1}}$, so $\mathsf{LEAVES}(\mathtt{T}) - \mathsf{LEAVES}(\mathtt{T}_{[v_0]}) \neq \emptyset$.

Proof (cont'd)

Let w_1 be a leaf in $\text{LEAVES}(\mathtt{T}) - \text{LEAVES}(\mathtt{T}_{[v_0]})$, and let v_1 be a vertex on the path that joins the root to w_1 whose label x_1 is of length ℓ_1. It is clear that neither of the words x_0, x_1 is a prefix of the other. The tree $\mathtt{T}_{[v_1]}$ has $k^{\ell_{m-1}-\ell_1}$ leaves and, if $m \geq 2$, then $k^{\ell_{m-1}-\ell_0} + k^{\ell_{m-1}-\ell_1} < k^{\ell_{m-1}}$, there are leaves of $\mathtt{T}$ that are not included in $\text{LEAVES}(\mathtt{T}_{[v_0]}) \cup \text{LEAVES}(\mathtt{T}_{[v_1]})$, etc. By a repeated application of this technique we construct a prefix code with word lengths $\ell_0, \ell_1, \ldots, \ell_{m-1}$.

Observe that the proof of the theorem contains an algorithm for generating prefix codes that have a prescribed list of word lengths. Since, in general, there are several choices for the words of a given length, the algorithm is nondeterministic.

Example

Let $S = \{s_0, s_1, \ldots, s_7\}$ be the set of symbols of a source and let $A = \{a, b, c\}$ be an alphabet. Suppose that we intend to design a code $h : S^* \longrightarrow A^*$ such that the lengths of the words of the code set $h(S)$ are

s	s_0	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$\lvert h(s) \rvert$	1	2	2	2	3	3	3	3

Since

$$\frac{1}{3} + 3 \cdot \frac{1}{3^2} + 4 \cdot \frac{1}{3^3} = \frac{22}{27} \leq 1,$$

we know that there is a prefix code such that the corresponding code set consists of words having the prescribed length.

Applying the above construction we can obtain the following prefix code:

s	s_0	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$h(s)$	a	ba	bb	cc	bcb	bcc	cab	cbb

On the other hand, there is no prefix code (and, as we shall see, no code) having as its list of lengths of code words $(1, 1, 2, 2, 2, 3, 3, 3)$ because

$$2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3^2} + 3 \cdot \frac{1}{3^3} = \frac{10}{9} > 1.$$

The Kraft-McMillan inequality can be extended to arbitrary codes.

Theorem

(Kraft-McMillan Inequality) *Let* $A = \{a_0, \ldots, a_{k-1}\}$ *be an alphabet. A code on the alphabet A with word lengths* $\ell_0, \cdots, \ell_{m-1}$ *exists if and only if*

$$\sum_{i=0}^{m-1} \frac{1}{k^{\ell_i}} \leq 1.$$

Proof

It is clear that the inequality of the theorem is sufficient since every prefix code is a code.
Conversely, let L be a code that has word lengths $\ell_0, \ell_1, \ldots, \ell_{m-1}$ and let $r = \max\{\ell_i \mid 0 \leq i \leq m-1\}$. Define

$$K = \sum_{i=0}^{m-1} \frac{1}{k^{\ell_i}}.$$

Using the generalized binomial formula We can write

$$K^n = \sum \left\{ \frac{n!}{n_0! \cdots n_{p-1}!} k^{-(n_0\ell_0 + \cdots + n_{p-1}\ell_{p-1})} \;\middle|\; n_0 + \cdots + n_{p-1} = n \right\},$$

where $n \leq n_0\ell_0 + \cdots + n_{p-1}\ell_{p-1} \leq nr$. If μ_m is the sum of the coefficients c of the terms of the form ck^{-m}, then $K^n = \sum\{\mu_m k^{-m} \mid n \leq m \leq mr\}$. The last equality is obtained by regrouping the sum by collecting coefficients of like powers. Note that μ_m equals the number of solutions in the natural numbers of the equation $x_0\ell_0 + \cdots + x_{p-1}\ell_{p-1} = m$, since each of these solutions corresponds to a word in L^*.

Proof (cont'd)

On the other hand, the number of words of length m in L^+ is at least equal to the number of the solutions in the natural numbers of the equation $x_0\ell_0 + \cdots + x_{p-1}\ell_{p-1} = m$ since each each such solution corresponds to a word $x \in L^+$. Since there are at most k^m words of length m in L^+, we have $\mu_m \leq k^m$. This implies

$$\begin{aligned} K^n &= \sum\{\mu_m k^{-m} \mid n \leq m \leq mr\} \\ &\leq \sum\{k^m k^{-m} \mid n \leq m \leq mr\} = mr - m + 1. \end{aligned} \tag{2}$$

for $n \in \mathbb{N}$. If $K > 1$, then $\lim_{n\to\infty} K^n = +\infty$, and this would contradict the existence of the upper bound given in (2). This implies the Kraft-McMillan inequality.

Corollary

Let $A = \{a_0, \dots, a_{k-1}\}$ *be an alphabet. A code on the alphabet* A *with word lengths* $\ell_0, \cdots, \ell_{m-1}$ *exists if and only if there exists a prefix code that has the same list of word lengths.*

Proof

Suppose that there exists a code on A. By Theorem 3, its list of lengths $\ell_0, \cdots, \ell_{m-1}$ satisfies the Kraft-McMillan Inequality. Then, there exists a prefix code that has the same list of word lengths.
The reverse implication is obvious.

The Kraft-McMillan inequality does not guarantee that a language L is a code as shown next.

Example

Let $A = \{a, b\}$ and let $L = \{aa, aab, baa\}$. We have $\ell_0 = 2$, and $\ell_1 = \ell_2 = 3$, so

$$\sum_{i=0}^{2} \frac{1}{k^{\ell_i}} = \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}.$$

However, L is not a code since $(aab)(aa) = (aa)(baa)$.

The Kraft-McMillan inequality suggests the introduction of a numerical characteristic of languages.

Definition

Let A be an alphabet and let L be a language on A. The *code indicator* of a language L is

$$\mathrm{ci}(L) = \sum_{x \in L} \frac{1}{|A|^{|x|}}.$$

Now we extend the Kraft-McMillan to arbitrary (not necessarily finite) codes.

Theorem

If a language $L = \{x_0, x_1, \ldots\}$ *on the alphabet* A *is a code, then* $\mathrm{ci}(L) \leq 1$.

Proof

Suppose that L is a code and $\mathrm{ci}(L) > 1$. There is a finite language K such that $K \subseteq L$ and $\mathrm{ci}(K) > 1$. Since every subset of a code is also a code, this contradicts Theorem 3.

Definition

A code L on an alphabet A is maximal if $L \cup \{x\}$ is not a code for every $x \in A^+ - L$.

Corollary

If L is a finite code on the alphabet L and $\mathrm{ci}(L) = 1$*, then L is a maximal code.*

This statement follows immediately from Theorem 3.

Example

Let A be an alphabet and let L_k be the block code that consists of all words of length k in A^*. Since $\mathrm{ci}(L_k) = 1$, it follows that L_k is a maximal code.