CDMTCS
Research
Report
Series

On Maximal Prefix Codes

Ludwig Staiger
Martin-Luther-Universität
Halle-Wittenberg

CDMTCS-280
May 2006

Centre for Discrete Mathematics and
Theoretical Computer Science

On Maximal Prefix Codes

Ludwig Staiger
Martin-Luther-Universität Halle-Wittenberg
Institut für Informatik
von-Seckendorff-Platz 1
D–06099 Halle (Saale), Germany
staiger@informatik.uni-halle.de

Abstract

Kraft's inequality is a classical theorem in Information Theory which establishes the existence of prefix codes for certain (admissible) length distributions. We prove the following generalisation of Kraft's theorem: For every admissible infinite length distribution one can construct a maximal prefix codes whose codewords satisfy this length distribution.

Prefix codes are widely used in data transmission or in (algorithmic) information theory (see [3, 4]). A set of nonempty words $C \subseteq X^*$ over an alphabet X is called a *prefix code* provided $w \in C$ is *not* a prefix of $v \in C$, for every pair of distinct words $w, v \in C$.

A classical theorem about the existence prefix codes is called Kraft's inequality [2].

Theorem 1 (Kraft's inequality). *Let X be a finite alphabet, $I \subseteq \mathbb{N}$ and let $f : I \to \mathbb{N}$ be a non-decreasing function such that $\sum_{n \in I} |X|^{-f(n)} \leq 1$. Then there is a prefix code $C = \{v_n : n \in I\} \subseteq X^*$ such that $|v_n| = f(n)$.*

Here $|X|$ denotes the cardinality of the set X, and $|v|$ denotes the length of the word v and $\sum_{n \in I} |X|^{-f(n)} \leq 1$ means that the length distribution $f : I \to \mathbb{N}$ is admissible.

The aim of this note is to show that a simple modification of Kraft's construction (see e.g. [4]) is suitable for the construction of infinite maximal prefix codes $C \subseteq X^*$ whenever $\sum_{v \in C} |X|^{-|v|} \leq 1$.

Here a code $C \subseteq X^*$ is referred to as *maximal prefix* if C is a prefix code and for every prefix code $C' \supseteq C$ implies $C' = C$. It is known that a maximal prefix code need not be maximal as a code (see e.g. [1, II. Example 3.1]). For finite codes $C \subseteq X^*$, however, a maximal prefix code satisfies $\sum_{v \in C} |X|^{-|v|} = 1$ and is also maximal as a code.

Theorem 2. *Let $f : \mathbb{N} \to \mathbb{N}$ be a non-decreasing function such that $\sum_{n\in\mathbb{N}} |X|^{-f(n)} \leq 1$. Then there is a maximal prefix code $C = \{v_n : n \in \mathbb{N}\} \subseteq X^*$ such that $|v_n| = f(n)$.*

We use the following characterisation of maximal prefix codes whose proof is given here for the sake of completeness.

Lemma 3. *Let M be an infinite subset of $\mathbb{N}$. A code $C \subseteq X^*$ is maximal prefix if and only if for all $w \in \{v : v \in X^* \wedge |v| \in M\}$ there is a $v \in C$ such that $w \sqsubseteq v$ or $v \sqsubseteq w$.*

Proof. If C is not maximal prefix then there is a $w \notin C$ such that $C \cup \{w\}$ is a prefix code. Consider $wu \in X^*$ where $|wu| \in M$. Since $w \not\sqsubseteq v$ and $v \not\sqsubseteq w$ for every $v \in C$, the same holds true for the word wu.

Conversely, if for some $w \in \{v : v \in X^* \wedge |v| \in M\}$ there is no $v \in C$ such that $w \sqsubseteq v$ or $v \sqsubseteq w$ then $C \cup \{w\}$ is a prefix code properly containing C. □

Now, using this lemma we construct a prefix code which satisfies the condition of Lemma 3 for some infinite set $M \subseteq \{f(n) : n \in \mathbb{N}\}$. This is done by the following algorithm MaxKraft.

Algorithm MaxKraft

```
0  n := 0;  l := 0;  C := ∅;  M := ∅
1  For  i = 1  to  ∞  do
2  l := f(n);  W := X^l \ C · X^*;  M := M ∪ {l}
3          Let W = {w_1, ..., w_|W|}
4          For  j = 0  to  |W| − 1  do
5                  C := C ∪ {w_{j+1} · 0^{f(n+j)−l}}
6          Endfor
7          n := n + |W|
8  Endfor
```

Here the set M is included just to have a reference to Lemma 3.

At stage $i + 1$ our parameters before constructing the new approximation C_{i+1} are C_i, n_i and $l_{i+1} = f(n_i)$ where $f(n_i - 1) = \sup\{|w| : w \in C_i\}$.

Then the set $W_{i+1} = X^{l_{i+1}} \setminus C_i \cdot X^*$ is the set of words which have no prefix in C_i. For each of the words $\{w_1, \ldots, w_{|W_{i+1}|}\}$, the body of the **For**-loop (lines 4 to 6) adds the word $w_{j+1} \cdot 0^{f(n_{i+1}+j)-l_{i+1}}$ of length $f(n_{i+1} + j)$ to C_i. Thus $f(j)$ is the length of the jth word in C_{i+1} if $j \leq |C_{i+1}|$, in particular $f(n_{i+1} - 1) = \sup\{|w| : w \in C_{i+1}\}$.

As in the proof of Kraft's inequality, we obtain that

$$|W_{i+1}| = \sum_{v\in C_i} |X|^{l_{i+1}-|v|} = |X|^{l_{i+1}} \cdot \sum_{j=1}^{|C_i|} |X|^{-f(j)} < |X|^{l_{i+1}} .$$

Consequently, the algorithm does not stop, that is, $C_i \subset C_{i+1}$, and returns an infinite set $C = \bigcup_{i=1}^{\infty} C_i$ in which the word constructed in step j has length $f(j)$.

Clearly, the resulting C_{i+1} is a prefix-code, if C_i is a prefix-code, and by the steps in lines 4 and 5 every word of length l_{i+1} has a prefix in $C_i \subseteq C_{i+1}$ or is a prefix of some word in C_{i+1}.

At the next stage this process is repeated for the new (greater) length $l_{i+2} := f(n_{i+1} + |W_{i+1}|)$. So, by induction, it is seen that $C = \bigcup_{i=1}^{\infty} C_i$ is a prefix code for which the infinite set $M = \{l_i : i = 1, \ldots\}$ is a witness for its prefix maximality.

The algorithm depends on the monotonicity of the function $f : \mathbb{N} \to \mathbb{N}$. The monotonicity guarantees that, when, at some stage i, the finite approximation C_i of the code C is constructed, all words $w \in C \setminus C_i$ will have length $|w| \geq f(n_i - 1)$.

References

[1] J. Berstel and D. Perrin. *Theory of Codes*. Academic Press, 1985.

[2] L.G. Kraft. *A Device for Quantizing Grouping and Coding Amplitude Modulated Pulses*, MS Thesis, Electrical Eng. Dept., MIT, Cambridge, Ma., 1949.

[3] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, New York, 1993.

[4] R. Johannesson, *Informationstheorie*, Addison-Wesley, 1992.