Reprinted from INFORMATION AND CONTROL, Volume 11, No. 4, October 1967

INFORMATION AND CONTROL **11**, 396–401 (1967)

On Synchronizing Prefix Codes

MARCEL PAUL SCHÜTZENBERGER

Faculté de Sciences, Paris

A study is made of the possible distribution of the word lengths in a synchronizing prefix code.

INTRODUCTION

Let X^* be the free monoid generated by a fixed finite alphabet X of $k > 1$ elements. A non-empty subset A of XX^* is a *prefix code* iff every word of X^* has at most one left factor in A, i.e. iff $A \cap AXX^* = \emptyset$. It is *synchronizing* iff $X^*\bar{a} \subset A^*$ for some $\bar{a} \in A^*$ where A^* denotes the submonoid generated by A. (See references).

Consider the enumerating sequence $\alpha = (\alpha_n = \textbf{\textit{Card}}\,(A \cap X^n))_{n\in\mathbf{N}}$ of a prefix code A and set $\sigma_n = k^n - \sum_{0<m\leqq n} \alpha_m k^{n-m} = \sigma_{n-1}k - \alpha_n$. We have $\alpha_0 = 0$ and $\sigma_0 = 1$ because of $A \subset XX^*$. Further, X^n is the disjoint union of the sets $S \cap X^n$ and $(A \cap X^m)X^{n-m} (m = 1, 2, \cdots, n)$ where $S = X^*\backslash AX^*$. It follows that $\textbf{\textit{Card}}\,(S \cap X^n) = \textbf{\textit{Card}}\,(X^n) - \sum_{0<m\leqq n} \textbf{\textit{Card}}\,(A \cap X^m).\ \textbf{\textit{Card}}\,(X^{n-m}) = \sigma_n$ showing

$$\sigma_n \geqq 0 \textit{ for all } n \in \mathbf{N}. \tag{1}$$

Assume that A is synchronizing and $\bar{a} \in X^p$. S contains the left factors of its members and no word having $\bar{a}$ as a right factor. Thus $S \cap X^{n+p}$ is a subset of $(S \cap X^n)(X^p\backslash\{a\})$ and

$$\sigma_{n+p} \leqq \sigma_n \,.\, (k^p - 1) \textit{ for some fixed positive } p \textit{ and all } n \in \mathbf{N}. \tag{2}$$

Finally, let d be the greatest common divisor of the elements of $N_\alpha = \{n \in \mathbf{N}: \alpha_n \neq 0\}$, i.e., let $A \subset (X^d)^*$. Since $A^* \subset (X^d)^*$, $f\bar{a} \subset A^* (f \in X^*)$ is possible only if $f \in (X^d)^*$. Thus

$$1 \textit{ is the } \text{g.c.d.} \textit{ of the elements of } N_\alpha\,. \tag{3}$$

We intend to verify the following converse property.

PROPERTY. *Let $\alpha = (\alpha_n)_{n\in\mathbf{N}}$ be a sequence of non-negative integers that*

satisfies $\alpha_0 = 0$, (1), (2) *and* (3). *It is the enumerating sequence of at least one synchronizing prefix code.*

It may be observed that when A is finite, i.e. when $\alpha_n = 0$ for all n larger than some finite value $\bar{n}$, Condition (2) is equivalent to $\sigma_{\bar{n}} = 0$. Indeed, from $\alpha_n = 0$, $(n > \bar{n})$, and $\sigma_{n+1} = \sigma_n k - \alpha_{n+1}$ we deduce $\sigma_{\bar{n}+p} = k^p \sigma_{\bar{n}}$. This implies $\sigma_{\bar{n}} = 0$ in view of $\sigma_{\bar{n}+p} \leqq \sigma_{\bar{n}} . (k^p - 1)$. Reciprocally, if $\sigma_{\bar{n}} = 0$, the hypothesis $0 \leqq \sigma_{n+1} = \sigma_n k - \alpha_{n+1}$ shows that $\alpha_n = \sigma_n = 0$ for all $n > \bar{n}$ and one has $\sigma_{n+\bar{n}} \leqq \sigma_n . (k^{\bar{n}} - 1)$ identically.

We also recall the known fact that $\alpha_0 = 0$ and Condition (1) suffice to insure that a sequence $\boldsymbol{\alpha}$ of non-negative integers is the enumerating sequence of at least one prefix code A. Indeed, let $A_0 = \emptyset$ and, inductively, let A_n be the union of A_{n-1} with an arbitrary set of $\alpha_n' = \boldsymbol{Min}\{\alpha_n, \boldsymbol{Card}\,(X^n \backslash A_{n-1}X^*)\}$ words from $X^n \backslash A_{n-1}X^*$. Each A_n and $A = \bigcup_{0 \leqq n} A_n$ is a prefix code. Suppose $\alpha_m' = \alpha_m$ and $\sigma_m = \boldsymbol{Card}\,(X^m \backslash A_m X^*)$ is already verified for $m < n$. We have $X^n \backslash A_{n-1}X^* = (X^{n-1} \backslash A_{n-1}X^*)X$ since $A_{n-1} \cap X^n = \emptyset$ and $\alpha_n \leqq \sigma_{n-1}k = \boldsymbol{Card}\,((X^{n-1} \backslash A_{n-1}X)X)$ since $0 \leqq \sigma_n = \sigma_{n-1}k - \alpha_n$. Thus $\alpha_n' = \alpha_n$ and $\sigma_n = \boldsymbol{Card}\,(X^n \backslash A_n X^*)$ and the observation is proved.

Finally, to simplify our later discussion, we establish the following remark in which x is any fixed letter of X.

Remark 1. Let $\boldsymbol{\alpha}$ satisfy $\alpha_0 = 0$, (1) and (2). It is the enumerating sequence of a prefix code A such that $S x^{2p} \cap S = \emptyset$. Thus the Property is true when, further, $\alpha_1 \geqq 1$.

Proof. Let $\bar{m}$ be any fixed integer and, for $n \geqq \bar{m}$, let the α_n words of $A_n \backslash A_{n-1}$ be successively chosen so that $fx^r \in A_n \backslash A_{n-1}$, $(f \in X^{n-r})$ only if every word of $X^n \backslash A_{n-1}X^*$ of the form $f'x^{r'}$, $(f' \in X^{n-r'}, r' > r)$ is already taken in A_n. Then to prove $sx^{2p} \notin S$ for each $s \in S \cap X^n$ it suffices to verify that Condition (2) implies $\sigma_n < \sum_{0<m \leqq 2p} \alpha_{n+m}$. Now, by definition

$$0 \leqq \sigma_{n+p} = \sigma_n k^p - \sum_{0 < m \leqq p} \alpha_{n+m} k^{p-m} \tag{4}$$

Hence, by (2),

$$\sigma_n \leqq \sum_{0<m \leqq p} \alpha_{n+m} k^{p-m}. \tag{5}$$

Replacing n by $n + p$ in (5) and using (4) we get

$$\sigma_n k^p - \sum_{0<m \leqq p} \alpha_{n+m} k^{p-m} = \sigma_{n+p} \leqq \sum_{0<m \leqq p} \alpha_{n+p+m} k^{p-m}$$

that is,

$$\sigma_n \leqq \sum_{0<m\leqq p} \alpha_{n+m}k^{-m} + \sum_{0<m\leqq p} \alpha_{n+p+m}k^{-m} \leqq \sum_{0<m\leqq 2p} \alpha_{n+m}\,,$$

and the first part of the Remark is proved. When $\alpha_1 = 1$, we can take $\bar{m} = 0$ and then, $x \in A$. Letting $\bar{a} = x^{2p}$ we have $S\bar{a} \subset A^*$. Hence $X^*\bar{a} \subset A^*$ because A is a prefix code and therefore every word of X^* has one and only one factorisation in the form as ($a \in A^*$, $s \in S$), i.e. because $X^* = A^*S$.

CONSTRUCTION OF THE PREFIX CODE A

We consider a fixed sequence $\boldsymbol{\alpha}$ satisfying $\alpha_0 = \alpha_1 = 0$, (1), (2), (3) and we let x and y be two distinct letters of X. We set the following:

$\bar{q}$ = the least member of N_α such that 1 is the g.c.d. of the elements of N_α, less or equal to $\bar{q}(2 < \bar{q} < \infty)$.

$\bar{n}$ = the supremum of the elements of $N_\alpha(\bar{q} \leqq \bar{n}$ and $\bar{n} = \infty$ iff $\sigma_{\bar{n}} > 0$ for all $n \in \mathbf{N})$.

Define inductively a sequence $(\tau_n)_{n\in\mathbf{N}}$ by $\tau_0 = 0$ and $\tau_n = \boldsymbol{Max}\{0, 1 + \tau_{n-1} - \alpha_n\}$. We have $\sigma_n > \tau_n$ for all $n < \bar{n}$. Indeed, this is true for $n = 0$; for $n < \bar{n}$ it follows from $\sigma_n > 0$ and the induction hypothesis since $\sigma_n - \tau_n$ is equal to σ_n or to $\sigma_{n-1}k - 1 - \tau_{n-1} = (\sigma_{n-1} - \tau_{n-1}) + \sigma_{n-1}(k-1) - 1$, depending upon $\tau_n = 0$ or $= 1 + \tau_{n-1} - \alpha_n$.

Remark 2. Let $1 + \tau_{n-1} - \alpha_n \geqq 0$ for all $n \geqq \bar{q}$. The Property is true for $\boldsymbol{\alpha}$.

Proof. We first show that the hypothesis entails $k = 2$ and $\sigma_n = \alpha_n = 1$ for all $n \geqq \bar{q}$, hence $\bar{n} = \infty$. Indeed for $n \geqq \bar{q}$ we have

$$\begin{aligned}\sigma_n - \tau_n - 1 &= \sigma_{n-1}k - \alpha_n - 1 - \tau_{n-1} + \alpha_n - 1 \\ &= (\sigma_{n-1} - \tau_{n-1} - 1)k + \tau_{n-1}(k-1) + k - 2.\end{aligned} \tag{6}$$

It follows that $\sigma_n - \tau_n - 1$ is identically zero because, otherwise, we would have for every $m \geqq 0$ the inequality $\sigma_{n+mp} - \tau_{n+mp} - 1 \geqq k^{mp}$ in contradiction with Condition (2) which imposes $\sigma_{n+mq} \leqq \sigma_n(k^q - 1)^m$. Thus, for $n \geqq \bar{q}$, we must have successively $k = 2$, $\tau_{n-1} = 0$, $\sigma_{n-1} = 1$ and $\alpha_n = 1$ which follow from $1 = \sigma_n = \sigma_{n-1}k - \alpha_n = 2 - \alpha_n$.

Using the construction described in Remark 1, we choose the words of the sets $A_n \backslash A_{n-1}$ in such a way that for $n \geqq \bar{q}$, $A_n \backslash A_{n-1} = \{x^{n-1}y\}$. Then S consists of words of lengths $<\bar{q}$ and of x^*. Let $\bar{a} = x^r y$ where $r \geqq 2\bar{q}$. If $s \in S$ has the form $x^{r'}$), we have $s\bar{a} = x^{r'+r}y \in A$; if not, s

has length $<\bar{q}$, sa does not belong to S and, accordingly, $sa = a'x^{r'}y$ where $a' \in A$ and $x^{r'}y \in A$. The Remark is verified.

Thus we can now define

$$q = \text{the least integer} \geqq \bar{q} \text{ such that } 1 + \tau_{q-1} - \alpha_q < 0.$$

Remark 3. Let α satisfy $\alpha_0 = \alpha_1 = 0$, (1), (2), (3) and $q < \infty$. It is the enumerating sequence of a prefix code A such that:

$$x^q \in A, \tag{7}$$

$$B = A \cap x^*yx^* \text{ consists of } q \text{ words } b_i = x^iyx^{\lambda_i - i - 1} \tag{8}$$

$(i = 0, 1, \cdots, q-1)$ where $\lambda_i \geqq i - 1$;

$$\lambda_i \leqq q \text{ identically with equality for } i \geqq q - \tau_{q-1} - 1; \tag{8.1}$$

$$1 \text{ is the g.c.d. of } \{\lambda_0, \lambda_1, \cdots, \lambda_{q-1}\}; \tag{8.2}$$

$$Sx^{rq} \cap S = \emptyset \text{ where } rq \geqq 2p. \tag{9}$$

Proof. We again use the construction of Remark 1. We can take $x^q \in A_q \backslash A_{q-1}$ since $\alpha_q \neq 0$ and $\sigma_n > 0$ for $n < q$ because $q \in N_\alpha$.

Let (for $n \leqq q$)

$Y_n = \{x^iyx^{n-1-i}, \quad i = 0, 1, \cdots, n-1\}$,

$B_0 = B_1 = \emptyset$ and, inductively,

B_n = the union of B_{n-1} with the $\beta_n = \textbf{\textit{Min}}\{\alpha_n, \textbf{\textit{Card}}(Y_n \backslash B_{n-1}x^*)\}$ words $b_i = x^iyx^{n-1-i}$ of $Y_n \backslash B_{n-1}x^*$ for which i has its least values.

Setting $T_n = Y_n \backslash B_n x^* = (\{x^{n-1}y\} \cup T_{n-1}x) \backslash (B_n \backslash B_{n-1})$ and $\tau_n' = \textbf{\textit{Card}}\, T_n$, we have $\beta_n = \textbf{\textit{Min}}\{\alpha_n, 1 + \tau'_{n-1}\}$ and $\tau_n' = \textbf{\textit{Max}}\{0, 1 + \tau'_{n-1} - \beta_n\} = \textbf{\textit{Max}}\{0, 1 + \tau'_{n-1} - \alpha_n\}$. Thus, by induction, $\tau_n' = \tau_n$ and $\tau_q = 0$ follows from our choice of q. This proves (8.1), and (8.2) follows from $q \geqq \bar{q}$ and the fact that by construction $\beta_n > 0$ for each $n \in N_\alpha$ less or equal to q.

Finally, (9) results from the same reasons as in Remark 1.

It only remains to verify that A is synchronizing. In view of (7) and (9) and of $X^* = A^*S$ it suffices to show that the submonoid of A^* generated by $C = \{x^q\} \cup B$ contains at least one word $\bar{c}$ such that $x^m\bar{c} \in A^*$ for every m in the interval $I = \{0, 1, \cdots, q-1\}$. To do this we first verify the following Remark in which, for any $z \in \mathbf{Z}$, $[z]_{q'}$ denotes the least non-negative integer congruent of z modulo q' and for $d \in \mathbf{N}$, $I_d = \{q-1-d, q-d, q+1-d, \cdots, q-1\}$:

Remark 4. Let t be a positive integer less than q and define two maps u and v of I into itself by letting for each $i \in I$

$$\text{i.u} = [i + 1]_q \text{ and i.v} = [i - \lambda_i + 1]_q$$

where, identically, $1 + i \leqq \lambda_i \leqq q$ and $\lambda_i = q$ iff $i \in I_t$. If the g.c.d. of the numbers $\{\lambda_0, \lambda_1, \cdots, \lambda_{q-1}\}$ is 1, the monoid M generated by u and v contains an element sending I onto 0.

Proof. For $d \in \mathbf{N}$, let M_d denote the subsemigroup of M consisting of all $m \in M$ such that i.m $\in I_d$ for $i \in I$ and $i.m = i$ for $i \in I_d$. By definition $i.vu^{q-1} = q - (\lambda_i - i) =$ an element strictly greater than itself if $i \in I \backslash I_t$ and $=$ an element of I_t if $i \in I_t$. Thus some large enough power of vu^{q-1} belongs to M_t.

Let $d = \textbf{\textit{Min}} \{d' : M_{d'} \neq \emptyset\}$. If $d = 1$ and $m \in M_1$ the proof is complete because mu sends I onto 0. Thus from now on we can assume $d > 1$ and we verify the following statement: M_d consists of a single map m sending each $q - i \in I$ onto $q - [i]_d$.

Indeed, by the definition of M_d we have $(q - i) \cdot m = q - [i]_d$ for every $m \in M_d$ and $i \in I_d$. For the sake of contradiction, let $j > d$ be the least value for which $(q - j) \cdot m \neq q - [j]_d$, $(m \in M_d)$. Consider the map $u^{j-d-1}m$. It sends I onto I_d and its restriction to I_d is a permutation. Thus one of its positive powers, say m', belongs to M_d and $(q - d - 1) \cdot m' = q - d' \neq q - 1$ because of $(q - j) \cdot m \neq q - [j]_d$. It follows that we can assume $m = m'$, that is $j = d + 1$.

Under this supplementary hypothesis mu^{q-1} has the following properties:

(a) sends I onto the interval $\{q - 1 - d, q - d, q + 1 - d, \cdots, q - 2\}$,

(b) reduces to a permutation on $J' = \{q - 1 - d, q - d, \cdots, q - 1 - d'\}$,

(c) sends every $i \in I_d$, onto $i - 1$.

Thus one of its positive powers, say m'', sends I onto J' and reduces to the identity on J'. Multiplying m'' on the left and on the right by suitable powers of u, we obtain a map which belongs to $M_{d+1-d'}$ where $d + 1 - d' < d$ by construction. Since this contradicts the minimal character of d, the statement is verified.

Consider now $vu^{q-1}m$ where $\{m\} \in M_d$. Since vu^{q-1} reduces to the identity on I_t and $t \geqq d$, this map also belongs to M_d. However, $m = vu^{q-1}m$ is possible only if for each $i \in I$ one has $i \cdot vu^{q-1} = [i]_d$, that is, according to the definition of u and of v, only if all the numbers λ_i are

congruent to 0 modulo d. This shows $d = 1$ when the g.c.d. of the λ_i is 1 and it establishes the Remark.

To conclude, the proof to each $m \in M$ of the form $m = z_1 z_2 \cdots z_r$, where $z_i = u$ or v, we associate the word μm obtained by replacing in m every $z_i = u$ by x and every $z_i = v$ by yx^q, $(i = 1, 2, \cdots, r)$. By $x^i y x^{\lambda_i - i - 1} \in A$, $x^q \in A$ and our definition of the maps u and v, we have identically $i \cdot m = i'$ iff $x^i \mu m = a x^{i'}$ where $a \in A^*$. Thus $x^* \mu(mu) \subset A^*$ where $\{m\} = M_1 \neq \emptyset$ and the Property is entirely proved.

RECEIVED: JULY 25, 1967

REFERENCES

GILBERT, E. N. AND MOORE, E. F. (1959), Variable length binary encodings. *Bell System Tech. J.* **38,** 933–968.

GOLOMB, S. W. AND GORDON, B. (1965), Codes with bounded synchronization delay. *Inform. Control.* **8,** 355–372.

MANDELBROT, B. (1954), On recurrent noise limiting coding. *Proc. Symp. Inform. Networks.* pp. 146–148, New York, N. Y.

NEUMANN, P. G. (1962), Efficient error limiting variable length codes. *IRE Trans.* **IT-8,** 292–304.

WINOGRAD, S. (1964), Input error limiting automata. *Commun. Comp. Mach.* **11,** 338–351.