A note on bifix-free sequences

Nielsen, Peter Tolstrup

Published in:
I E E E Transactions on Information Theory

Publication date:
1973

Document Version
Publisher's PDF, also known as Version of record

III. Conclusion

We have derived an expression for the expected duration of a search for a fixed pattern in a semi-infinite random sequence. It was demonstrated that the presence of bifices in the pattern tends to delay its occurrence in random data. The analysis employed the use of indicator random variables with a somewhat novel twist, and it is anticipated that this approach may prove useful in other contexts as well.

Acknowledgment

Thanks are due to Prof. J. L. Massey, who assisted this work through many helpful discussions and who coined the word "bifix" to facilitate the statement of the results.

References

[1] E. N. Gilbert, "Synchronization of binary messages," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 470–477, Sept. 1960.
[2] J. J. Stiffler, *Theory of Synchronous Communications*. Englewood Cliffs, N.J.: Prentice-Hall, 1971, sect. 12.4.

A Note on Bifix-Free Sequences

P. TOLSTRUP NIELSEN

***Abstract*—A bifix of an L-ary n-tuple is a sequence which is both a prefix and a suffix of that n-tuple. The practical importance of bifix-free patterns is emphasized, and we devise a systematic way of generating all such sequences and determine their number.**

I. Unique Patterns

Much attention has been given to the problem of selecting from the set of all L-ary n-tuples a few patterns which are particularly suitable for use as synchronization patterns in digital communications or for other similar applications. Whenever such a unique pattern is to be inserted into random data in order to provide an easily recognizable reference point for the receiver, it is essential to avoid accidental imitation of the pattern by data-pattern overlaps. When noise is a major consideration, this requirement imposes certain bounds on the shape of the pattern correlation function. The Barker sequences [1] constitute a classical example of such patterns with "perfect" correlation properties. A more application-oriented treatment of the problem may be found in [2].

When noise is absent the shape of the correlation function is immaterial if only the pattern is "bifix"-free. By a *bifix* we shall understand a sequence which is both a prefix and a suffix of the pattern. A bifix-free pattern of arbitrary length greater than one may always be obtained in the form *aa—abb—b* where *a* and *b* denote two distinct letters from the alphabet. Such patterns, however, generally have rather poor correlation properties which leave them unsuitable for use in noisy systems. Fortunately, the set $S_L(n)$ of L-ary bifix-free n-tuples is always rich, and for large values of n contain far more patterns than those mentioned previously. In fact, the search for patterns with outstanding correlation characteristics may frequently be restricted to encompass only elements of $S_L(n)$. Since the search for long patterns may be very time consuming, even on large computers, significant savings may in some cases be achieved by generating $S_L(n)$ before the search is initiated. In this correspondence we count the number of elements in $S_L(n)$ and devise a systematic way of generating the elements for any values of $L \geq 2$ and $n \geq 2$.

Manuscript received July 5, 1972; revised February 20, 1973.
The author is with the Institute for Circuit Theory and Telecommunication, Technical University of Denmark, Lyngby, Denmark.

II. The Number of Bifix-Free Patterns

We shall write n-tuples in the form

$$\boldsymbol{p} = [p_1, p_2, \cdots, p_n], \qquad p_i \in A_L \tag{1}$$

where A_L ($L \geq 2$) is an alphabet of L letters. We define the *bifix indicators* h_i of $\boldsymbol{p}$, $1 \leq i < n$, in a manner such that $h_i = 1$ if $[p_1, p_2, \cdots, p_i]$ is a bifix of $\boldsymbol{p}$, i.e., if

$$[p_1, p_2, \cdots, p_i] = [p_{n-i+1}, p_{n-i+2}, \cdots, p_n] \tag{2}$$

and $h_i = 0$ otherwise. $\boldsymbol{p}$ is bifix-free if and only if $h_i = 0$, $1 \leq i < n$.

The following property was hinted at by Artom [3], who referred to bifix-free sequences as being "valid" for synchronization purposes. In this context, however, we shall need the strongest statement possible.

Lemma 1: A necessary and sufficient condition for $\boldsymbol{p}$ to be bifix-free is that

$$h_i = 0, \qquad i = 1, 2, \cdots, \left\lfloor \frac{n}{2} \right\rfloor. \tag{3}$$

(Here and hereafter $\lfloor \cdot \rfloor$ denotes the integer part of the argument.)

Proof: Necessity is obvious. To prove sufficiency, we assume that (3) holds and that for some l, $\lfloor n/2 \rfloor < l < n$, we have $h_l = 1$. Then $\boldsymbol{p}$ has the bifix $\boldsymbol{b} = [p_1, p_2, \cdots, p_l]$; but since $\boldsymbol{b}$ is both a prefix and a suffix of $\boldsymbol{p}$ and its length exceeds half the length of $\boldsymbol{p}$ it follows that the sequence $\boldsymbol{b}' = [p_{n-l+1}, \cdots, p_l]$ of length $2l - n < l$ must be a bifix of $\boldsymbol{b}$ and, therefore, also of $\boldsymbol{p}$. By repeating this argument we eventually conclude that $\boldsymbol{p}$ must have a bifix of length less than or equal to $\lfloor n/2 \rfloor$, contradicting (3), and the proof is complete.

Given a bifix-free pattern $\boldsymbol{p}$ of *even* length, we now proceed to demonstrate how to construct longer bifix-free patterns by inserting extra digits in the "middle" of $\boldsymbol{p}$. Letting $\boldsymbol{p}_1 = [p_1, p_2, \cdots, p_{n/2}]$ and $\boldsymbol{p}_2 = [p_{n/2+1}, \cdots, p_n]$ we may write $\boldsymbol{p} = \boldsymbol{p}_1 \boldsymbol{p}_2$. Next, we define the $(n+1)$-digit pattern $\boldsymbol{p}'$ and the $(n+2)$-digit pattern $\boldsymbol{p}''$ as

$$\boldsymbol{p}' = \boldsymbol{p}_1 \pi_1 \boldsymbol{p}_2 \tag{4}$$

$$\boldsymbol{p}'' = \boldsymbol{p}_1 \pi_1 \pi_2 \boldsymbol{p}_2 \tag{5}$$

where π_1 and π_2 are single letters from A_L.

Lemma 2: For any choice of $\pi_1 \in A_L$ and $\pi_2 \in A_L$ the following implications are true:

a) $\boldsymbol{p}'$ is bifix-free $\Leftrightarrow$ $\boldsymbol{p}$ is bifix-free;
b) $\boldsymbol{p}''$ is bifix-free $\Rightarrow$ $\boldsymbol{p}$ is bifix-free.

(Note that the second implication is undirectional.) If $\boldsymbol{p}$ is bifix-free, $\boldsymbol{p}''$ will have a bifix if and only if both of the following conditions are satisfied:

i) $[\pi_1, \pi_2] = [p_n, p_1]$
ii) $[p_2, \cdots, p_{n/2}] = [p_{n/2+1}, \cdots, p_{n-1}]$, $\quad n \geq 4$ only.

Proof: Let the bifix indicators of $\boldsymbol{p}'$ and $\boldsymbol{p}''$ be h_i', $i = 1, 2, \cdots, n$, and h_i'', $i = 1, 2, \cdots, n+1$, respectively. From the constructions (4) and (5) we clearly have

$$h_i' = h_i'' = h_i = 0, \qquad 1 \leq i \leq \frac{n}{2} = \left\lfloor \frac{n+1}{2} \right\rfloor < \left\lfloor \frac{n+2}{2} \right\rfloor \tag{6}$$

TABLE I
SYSTEMATIC LISTING OF ALL BIFIX-FREE BINARY n-TUPLES ($2 \le n \le 6$)

$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
10	100	1000	10000	100000
	110		10100	101000
				101100
		1100	11000	110000
			11100	110100
				111000
				111100
		1110	11010	110010
			11110	111010
				111110

Only half of the complete list is shown; the other half results from an interchange of 0's and 1's.

TABLE II

L	V_∞
2	0.267786
3	0.556979
4	0.687748
5	0.760064
6	0.805577
7	0.836743
8	0.859378

and the first part of the lemma then follows immediately from Lemma 1. Furthermore, we note that because of (6) p'' will have a bifix if and only if $h''_{n/2+1} = 1$. The lemma states the necessary and sufficient condition for this to happen.

Equipped with Lemma 2, we can now show how to generate in a systematic manner all bifix-free sequences of arbitrary length n. First, we list all $L(L-1)$ bifix-free patterns of length 2. This set forms the basis of a recursion in which each bifix-free pattern of even length n gives birth to L distinct bifix-free $(n+1)$-digit patterns as shown in (4), and to either L^2 or $L^2 - 1$ (depending on whether condition ii) of Lemma 2 is satisfied or not) distinct bifix-free patterns of length $n + 2$ in accordance with (5). The procedure is illustrated in Table I for binary patterns.

If u_n denotes the number of bifix-free n-tuples for a fixed alphabet size L, we immediately see that

$$u_{n+1} = Lu_n, \qquad n \text{ even}, \quad n \ge 2 \tag{7}$$

but in order to get a similar expression for u_{n+2} we must determine how many of the u_n n-tuples branch into only $L^2 - 1$ rather than L^2 $(n+2)$-tuples. To this end, consider the $(n/2+1)$-digit pattern $p^* = [p_1, p_2, \cdots p_{n/2-1}, p_{n/2}, p_n]$ $(n \ge 4)$ with bifix indicators h_i^*, $i = 1, 2, \cdots, n/2$. Supposing that p satisfies condition ii) of Lemma 2, we see that $h_i^* = h_i$, $i = 1, 2, \cdots, n/2$, and we conclude from Lemma 1 that p is bifix-free if and only if p^* is bifix-free. Hence the number of bifix-free patterns p which branch into only $L^2 - 1$ $(n+2)$-tuples must equal $u_{n/2+1}$, so that

$$u_{n+2} = L^2 u_n - u_{n/2+1}, \qquad n \text{ even}, \quad n \ge 4. \tag{8}$$

Theorem 1: The number, u_n, of n-digit bifix-free patterns satisfies

$$u_n = \begin{cases} Lu_{n-1} - u_{n/2}, & n \text{ even} \\ Lu_{n-1}, & n \text{ odd} \end{cases} \tag{9}$$

with $u_0 = 1$.

The sequence $v_0, v_1, v_2, \cdots$, where $v_n \triangleq L^{-n} u_n$, is monotonically nonincreasing and it converges to a constant, v_∞, which satisfies

$$v_\infty \ge 1 - L^{-1} - L^{-2} > 0. \tag{10}$$

Proof: The first part of the theorem follows immediately from a rearrangement of (7) and (8) along with a direct verification for the smallest values of n. As to the second part, we rewrite (9) in terms of the v_n's,

$$v_n = \begin{cases} v_{n-1} - L^{-n/2} v_{n/2}, & n \text{ even} \\ v_{n-1}, & n \text{ odd} \end{cases} \tag{11}$$

where $v_0 = 1$. Assigning to n a fixed even value, say $n = 2m$, we obtain for $m \ge 1$

$$\begin{aligned} v_{2m} &= v_{2m-1} - L^{-m} v_m \\ v_{2m+1} &= v_{2m} \\ v_{2m+2} &= v_{2m+1} - L^{-(m+1)} v_{m+1} \\ &= v_{2m-1} - L^{-m} v_m - L^{-(m+1)} v_{m+1} \\ &\ge v_{2m-1} - v_m (L^{-m} + L^{-(m+1)}). \end{aligned} \tag{12}$$

Generally, for any $m \ge 1$ and $q \ge 0$

$$\begin{aligned} v_{2m+2q} &= v_{2m-1} - \sum_{j=0}^{q} L^{-(m+j)} v_{m+j} \\ &\ge v_{2m-1} - L^{-m} v_m \sum_{j=0}^{q} L^{-j} \\ &= v_{2m-1} - L^{-m} v_m \frac{L - L^{-q}}{L - 1} \\ &\ge v_{2m-1} - v_m \frac{L^{-m+1}}{L - 1}. \end{aligned} \tag{13}$$

Setting $m = 2$, we have $v_{2m-1} = v_m = 1 - L^{-1}$ so that for any $n \ge 4$ we have

$$v_n \ge (1 - L^{-1}) \left(1 - \frac{L^{-1}}{L - 1}\right) = 1 - L^{-1} - L^{-2}. \tag{14}$$

The sequence $v_0, v_1, v_2, \cdots$ must converge since it has been shown to be nonincreasing and to satisfy the aforementioned inequality (14). Equation (10) then follows immediately.

While (10) provides only a relatively coarse bound on v_∞, (13) may be utilized to evaluate v_∞ with any desired degree of accuracy. Observe from (13) that

$$v_{2m} - v_m \frac{L^{-m}}{L - 1} \le v_\infty \le v_{2m} \tag{15}$$

where the size of the interval within which v_∞ is bounded vanishes exponentially with m. Hence only a few applications of the recursion (11) suffice to determine v_∞ with high accuracy. Some numerical results are listed in Table II for alphabet sizes $L \le 8$. Table II shows the tightest six-significant-digit lower bounds on the fraction of bifix-free L-ary sequences of arbitrarily large length n.

III. CONCLUSION

In situations where a computerized search is needed to find a long unique word satisfying certain requirements, it will often be allowable to restrict the search to include only bifix-free patterns. From the results in Table II we note that the saving in computer time thus obtainable may be significant, particularly in the binary case where the set to be searched is then reduced by

almost 75 percent. The algorithm presented in this correspondence for generating bifix-free patterns is well suited for computer programming.

References

[1] R. H. Barker, "Group synchronization of binary digital systems," in *Communication Theory*, W. Jackson, Ed. New York: Academic Press, 1953, p. 273.

[2] M. W. Williard, "Optimum code patterns for PCM synchronization," in *Proc. Nat. Telem. Conf.*, session 5–5, May 1962.

[3] A. Artom, "Choice of prefix in self-synchronizing codes," *IEEE Trans. Commun.* (Corresp.), vol. COM-20, pp. 253–254, Apr. 1972.

Bounds on Rate-Distortion Functions for Stationary Sources and Context-Dependent Fidelity Criteria

BARRY M. LEINER AND ROBERT M. GRAY

***Abstract*—A class of lower bounds to rate-distortion functions of stationary sources with context-dependent fidelity criteria is derived by mapping the source and distortion measure into an equivalent restricted-transition stationary source with a single-letter fidelity criterion, and then applying the composite bound. This approach is seen to yield bounds which, although sometimes quite loose, apply to general stationary sources and context-dependent fidelity criteria. Two examples are presented.**

Rate-distortion theory is usually restricted to single-letter fidelity criteria due to the complexity of more general cases. In many cases, however, the fidelity of the reproduction depends on the context of the message. To take context into account, Shannon [1] introduced local distortion measures as follows. Let the information source produce a sequence of letters $\{X_i\}$ from an alphbeat A_X which is reproduced as the sequence $\{\hat{X}_i\}$ from an alphabet $A_{\hat{X}}$. A local distortion measure of span g is then any function $\rho_g(x_1,x_2,\cdots,x_g;\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_g) \triangleq \rho_g(x_g;\hat{x}_g)$ of source and reproduction sequences of length g such that $\rho_g(\cdot) \geq 0$. The distortion between n-tuples is then given by

$$\rho_n(x;\hat{x}) = (n-g+1)^{-1} \sum_{k=1}^{n-g+1} \rho_g(x_k,x_{k+1},\cdots,x_{k+g-1};\hat{x}_k,\cdots,\hat{x}_{k+g-1}).$$

When $g \geq 2$, the family of distortion measures $\{\rho_n: n \geq g\}$ is said to be context dependent.

For simplicity assume the source is discrete; let $Q_{X_n}(x)$ be the source probability mass function and define $R_{X_n}(D)$ by

$$R_{X_n}(D) = \inf_{p_n \in \mathscr{P}_n} I(X;\hat{X})$$

where

$$\mathscr{P}_n = \left\{ p_n(\hat{x} \mid x): \sum_{x,\hat{x}} p_n(\hat{x} \mid x) Q_{X_n}(x) \rho_n(x;\hat{x}) \leq D \right\}$$

Manuscript received October 31, 1972; revised March 9, 1973. This work was supported in part by the National Science Foundation under Grant GK-31630 and by the Joint Services Program at Stanford Electronics Laboratories, Stanford, Calif., under U.S. Navy Contract N00014-67-A-0112-0044.

B. M. Leiner is with the Department of Electrical Engineering, Stanford University, Stanford, Calif., and GTE-Sylvania, Inc., Mountain View, Calif.

R. M. Gray is with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.

and $I(X;\hat{X})$ is the average mutual information between the source n-tuple and the output of the "test channel" p_n

$$I(X;\hat{X}) = E \log \left[\frac{p_n(\hat{X} \mid X)}{\omega(\hat{X})} \right]$$

$$\omega(\hat{x}) = \sum_x p_n(\hat{x} \mid x) Q_{X_n}(x)$$

and the sums are taken over $A_{X_n} = \{x: Q_{X_n}(x) > 0\}$ and $A_{\hat{X}_n}$, the available reproducing alphabet. The rate-distortion function for the source X and context-dependent distortion measure generated by ρ_g is defined by

$$R_X(D) = \lim_{n\to\infty} n^{-1} R_{X_n}(D).$$

The appropriate source coding theorem and converse have been proved when the source is stationary ergodic [1], [2], [5].

When the fidelity criterion is context dependent, the rate-distortion function $R_X(D)$ is extremely difficult to calculate in general. Thus far the only calculation of a rate-distortion function for a context-dependent fidelity criterion has been a bound for a class of distortion measures called modular distortion measures. This bound is tight in the case of an equiprobable memoryless source [3]. Berger [6] and Goblick [7] have suggested using a map of the original source to a source with an expanded alphabet and single-letter distortion measure ($g = 1$) to evaluate the rate-distortion function for context-dependent fidelity criteria. In what follows, the equivalence of these two rate-distortion functions is established and a lower bound is obtained using a recent result for single-letter distortion measures. This bound is quite general, since it applies to any finite local distortion measure and discrete-time stationary source. Unfortunately, however, the bound is often quite loose, particularly when the context distortion matrix is sparse.

The equivalent source is described in terms of successive g-tuples of the source X. Define the source $\{U_k\}$ by $U_k = (X_k,X_{k+1},\cdots,X_{k+g-1})$. Since each letter of the source U is a g-tuple, define $U_k = (U_{k,1},U_{k,2},\cdots,U_{k,g})$. Clearly $U_{k,i} = U_{k-1,i+1}$; and U_k and U_{k+1} share $g-1$ components. A block U_n from the source U will have probability

$$\begin{aligned} Q_{U_n}(u_1,u_2,\cdots,u_n) &= Q_{U_n}((u_{1,1},\cdots,u_{1,g}),(u_{2,1},\cdots,u_{2,g})),\cdots,(u_{n,1},\cdots,u_{n,g}) \\ &= \begin{cases} Q_{X_{n+g-1}}(u_{1,1},u_{2,1},u_{3,1},\cdots,u_{n,1},u_{n,2},\cdots,u_{n,g-1}), \\ \qquad \text{if } u_{k,i} = u_{k+1,i-1},\ k = 1,\cdots,n-1, \\ \qquad\qquad 1 < i \leq g; \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned}$$

The equivalent source U has restricted transitions since it is not true that $Q_{U_2|U_1}(u_2 \mid u_1) > 0$ for all $u_1, u_2 \in A_{U_1} = A_{X_g}$. Defining the distortion measure on U in terms of the original X sequence mapped into U results in a single-letter distortion measure, since

$$\begin{aligned} d_n(u_n,\hat{u}_n) &= \rho_{n+g-1}(x_{n+g-1},\hat{x}_{n+g-1}) \\ &= n^{-1} \sum_{k=1}^{n} \rho_g(x_k,x_{k+1},\cdots,x_{k+g-1};\hat{x}_k,\cdots,\hat{x}_{k+g-1}) \\ &= n^{-1} \sum_{k=1}^{n} d(u_k,\hat{u}_k) \end{aligned}$$

where $d(u_k,\hat{u}_k)$ is the per-letter distortion measure on U induced by the span-g distortion measure defined on X. To calculate a rate-distortion function, it is necessary to specify the available