On the expected duration of a search for a fixed pattern in random data

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

example, for a length-3 channel the pulse response is $1 + 2D + D^2$. The constant K_2 for these channels and thus for all possible channels is overbounded by

$$K_2 \leq \sum_{n=1}^{\infty} 2n \left[\frac{l-1}{l}\right]^n = 2l(l-1). \tag{13}$$

IV. Conclusions

An estimate of an upper bound on performance has been developed for the VA which is useful for estimating performance over channels for which only the pulse response energy and duration are known.

References

[1] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory* vol. IT-13, pp. 260–269, Apr. 1967.

[2] G. D. Forney, Jr., "Lower bounds on error probability in the presence of large intersymbol interference," *IEEE Trans. Commun.* (Corresp.), vol. COM-20, pp. 76–77, Feb. 1972.

[3] ——, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 363–378, May 1972.

[4] F. R. Magee, Jr., and J. G. Proakis, "Adaptive maximum-likelihood sequence estimation for digital signaling in the presence of intersymbol interference," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-19, pp. 120–124, Jan. 1973.

[5] A. J. Viterbi, "Convolutional codes and their performance in communications systems," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 751–772, Oct. 1971.

[6] C. R. Wylie, Jr., *Advanced Engineering Mathematics.* New York: McGraw-Hill, 1966, ch. 11.

[7] M. E. Austin, "Decision feedback equalization for digital communication over dispersive channels," M.I.T. Lincoln Lab., Lexington, Mass., Tech. Rep. 437, Aug. 1967.

On the Expected Duration of a Search for a Fixed Pattern in Random Data

P. TOLSTRUP NIELSEN

***Abstract*—An expression is obtained for the expected duration of a search to find a given L-ary sequence in a semi-infinite stream of random L-ary data. The search time is found to be an increasing function of the lengths of the "bifices" of the pattern, where the term bifix denotes a sequence which is both a prefix and a suffix.**

I. Introduction

The most common method for obtaining or maintaining word or frame synchronization in a digital communication system utilizes the insertion of a fixed pattern at regular intervals into the transmitted sequence. Gilbert [1] (see also [2]) analyzed the constraints to be imposed on the data in applications where it is imperative to prevent the pattern from appearing between its intended locations. Usually, however, one does not care to constrain the data since the regularity with which the true pattern occurs makes it easily distinguishable from any spurious patterns; but in that event, Gilbert's results may be used to calculate the probability of not seeing the pattern between two successive insertions.

In this correspondence, we shall be concerned with the search for a specified pattern in a semi-infinite random sequence which has *not* been punctuated by any periodic insertions. This situation may arise during acquisition of frame synchronization in systems where the synchronizer has the additional task of discriminating between a number of signals (possibly received in time division multiplex), only one of which has been furnished with a recurrent synchronization pattern. The problem also relates to several other applications of unique words, for instance in digital burst communication, multiaccess systems, and various kinds of message switching.

It is entirely possible to adapt Gilbert's results to this somewhat different problem and hence obtain a fairly complete statistical description of the search. The results, however, are generally intractable and useful only for numerical calculations. In spite of this, we shall see in the next section that the *expected* duration of a search depends on the structure of the pattern in a very simple manner, and we shall derive a closed-form expression for this expectation.

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

II. The Search

Let A_L, $L \geq 2$, denote an alphabet of L letters, and consider a semi-infinite sequence

$$\boldsymbol{d} = [d_1, d_2, d_3, \cdots \tag{1}$$

of digits, where each digit d_i is selected independently and has probability L^{-1} of being any of the L letters in A_L. We shall call $\boldsymbol{d}$ a *random data stream.* (Later we shall occasionally assume for convenience that $\boldsymbol{d}$ is preceded by an initial digit d_0.) Finite data strings will be written

$$\boldsymbol{d}[i,j) = [d_i, d_{i+1}, \cdots, d_{j-1}] \tag{2}$$

for any $i \geq 0$, where of course $j \geq i + 1$. Furthermore, let

$$\boldsymbol{p} = [p_1, p_2, \cdots, p_n] \tag{3}$$

be a fixed pattern of n digits each chosen from A_L.

Since we are concerned with the statistics of the first occurrence of the pattern $\boldsymbol{p}$ in the data stream $\boldsymbol{d}$, we define the random variable x such that $x = i$ if and only if

$$\begin{aligned} &\boldsymbol{d}[i, i+n) = \boldsymbol{p} \\ &\boldsymbol{d}[j, j+n) \neq \boldsymbol{p}, \qquad 1 \leq j < i. \end{aligned} \tag{4}$$

Hence x is just the number of positions examined before the pattern is found. As we shall see, the expectation $E\{x\}$ is completely determined by the lengths of the "bifices" of $\boldsymbol{p}$ which are defined as follows.

Definition: The sequence $\boldsymbol{a} = [a_1, a_2, \cdots, a_m]$ $(1 \leq m < n)$ is a bifix of $\boldsymbol{p} = [p_1, p_2, \cdots, p_n]$ if

$$[p_1, p_2, \cdots, p_m] = [p_{n-m+1}, p_{n-m+2}, \cdots, p_n] = \boldsymbol{a} \tag{5}$$

i.e., if $\boldsymbol{a}$ is both the m-digit prefix and the m-digit suffix of $\boldsymbol{p}$.

We further 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 $\boldsymbol{p}$ and $h_i = 0$ otherwise. If $h_i = 0$, $1 \leq i < n$, we say that $\boldsymbol{p}$ is *bifix-free.* By way of convention we define $h_0 = h_n = 1$.

Example 1: Consider $A_2 = \{0,1\}$ and the six-digit pattern $\boldsymbol{p} = [1,0,1,1,0,1]$. [1] and [1,0,1] are the only bifices of $\boldsymbol{p}$, so that $h_0 = h_1 = h_3 = h_6 = 1$ and $h_2 = h_4 = h_5 = 0$.

The following conditional expectation will play a crucial role in what follows.

Lemma 1: For any n-digit pattern $\boldsymbol{p}$, $1 \leq n < \infty$,

$$E\{x \mid \boldsymbol{d}[0,n) = \boldsymbol{p}\} = L^n. \tag{6}$$

Proof: We define the indicator random variables z_i, $1 \leq i < \infty$, as

$$z_i = \begin{cases} 1, & \text{if } \boldsymbol{d}[i, i+n) = \boldsymbol{p} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

and for arbitrary N consider the sample mean

$$s_N = \frac{1}{N}\sum_{i=1}^{N} z_i. \tag{8}$$

With no constraints on the data stream, we have

$$\begin{aligned} E\{z_i\} &= \Pr\{z_i = 1\} \\ &= \Pr\{\boldsymbol{d}[i, i+n) = \boldsymbol{p}\} \\ &= L^{-n} \end{aligned} \tag{9}$$

so that from (8)

$$E\{s_N\} = L^{-n}. \tag{10}$$

But from (7), it follows that z_i and z_j are statistically independent if $|i-j| \geq n$. Moreover, we have always $|z_i z_j| \leq 1$ so that it follows from (8) that

$$\begin{aligned} E\{s_N{}^2\} &= E\left\{\frac{1}{N^2}\sum_{|i-j|<n} z_i z_j + \frac{1}{N^2}\sum_{|i-j|\geq n} z_i z_j\right\} \\ &\leq \frac{1}{N}(2n-1) + (L^{-n})^2 \end{aligned} \tag{11}$$

and hence that

$$\operatorname{var}\{s_N\} \leq \frac{1}{N}(2n-1). \tag{12}$$

Since the variance of s_N vanishes as N approaches infinity, it follows that for any $\varepsilon > 0$ and any $\delta > 0$

$$\Pr\left\{L^n - \varepsilon < \frac{1}{s_N} < L^n + \varepsilon\right\} \geq 1 - \delta \tag{13}$$

for all suitably large N. But s_N is just the fraction of the N positions $1,2,\cdots,N$ which contain the initial digit of length-n subsequences where $\boldsymbol{p}$ has been observed to occur. Thus $s_N{}^{-1}$ is just the observed average distance between occurrences of $\boldsymbol{p}$ and (13) is merely the law of large numbers for this distance. The lemma then follows immediately.

Although, as we have seen in the previous proof of Lemma 1, the expected distance *between* occurrences of the pattern $\boldsymbol{p}$ depends only on the length of the pattern; the expected distance x to the *first* occurrence of the pattern depends also on the structure of the pattern. It is somewhat surprising that, as we now proceed to show, the conditional expectation of x that was easily found by indicator random variable arguments can be utilized to obtain the desired unconditional expectation of x. We shall obtain $E\{x\}$ by an inductive argument, the basis of which will be as follows.

Lemma 2: If the n-digit pattern $\boldsymbol{p}$ is bifix-free, then

$$E\{x\} = 1 + L^n - n. \tag{14}$$

Proof: Suppose that $\boldsymbol{d}[0,n) = \boldsymbol{p}$. Then $\boldsymbol{d}[i,n) \neq [p_1,\cdots, p_{n-i}]$ for $1 \leq i < n$ since $\boldsymbol{d}[i,n)$ is the $(n-i)$-digit suffix of $\boldsymbol{p}$ and by hypothesis $\boldsymbol{p}$ has no bifices. Thus

$$E\{x \mid \boldsymbol{d}[0,n) = \boldsymbol{p}\} = n - 1 + E\{x\} \tag{15}$$

since the search for the next occurrence of $\boldsymbol{p}$ when $\boldsymbol{d}[0,n) = \boldsymbol{p}$ is equivalent to the search for the first occurrence of $\boldsymbol{p}$ in the random data stream $[d_n, d_{n+1}, d_{n+2}, \cdots$. The lemma now follows from (15) and Lemma 1.

We are now in a position to prove the main result of this correspondence.

Theorem 1: For an n-digit pattern $\boldsymbol{p}$ ($n \geq 1$), with bifix indicators $h_1, h_2, \cdots, h_{n-1}$,

$$E\{x\} = \sum_{i=0}^{n} h_i L^i - n \tag{16}$$

where, by convention, $h_0 = h_n = 1$.

Proof: We proceed by induction on the number k of nonzero bifix indicators h_i, $0 < i < n$. Note that (16) reduces to (14) for $k = 0$, i.e., for a bifix-free pattern, so a basis has been established for the induction.

Suppose that $k > 0$ and that (16) holds for all patterns with $k-1$ or fewer nonzero bifix indicators. Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k$ be the bifices of $\boldsymbol{p}$ in order of increasing length. Since $\boldsymbol{p}_k$ of length $n' < n$ is both a prefix and a suffix of $\boldsymbol{p}$, as are also the shorter sequences $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{k-1}$, it follows that the latter sequences are all bifices of $\boldsymbol{p}_k$ and the only bifices of $\boldsymbol{p}_k$. Letting x' be the expected search duration to find $\boldsymbol{p}_k$ in random data, we have then by the induction hypothesis

$$E\{x'\} = \sum_{i=0}^{n'} h_i L^i - n'. \tag{17}$$

Since we may consider the search for $\boldsymbol{p}$ as a search for $\boldsymbol{p}_k$ followed by a search for the entire pattern $\boldsymbol{p}$ (which latter search may terminate immediately since $\boldsymbol{p}_k$ is a prefix of $\boldsymbol{p}$), we may write

$$E\{x\} = E\{x'\} + (E\{x \mid \boldsymbol{d}[1, n'+1) = \boldsymbol{p}_k\} - 1). \tag{18}$$

We next observe that

$$E\{x \mid \boldsymbol{d}[0,n) = \boldsymbol{p}\} = E\{x \mid \boldsymbol{d}[1, n'+1) = \boldsymbol{p}_k\} + n - n' - 1 \tag{19}$$

since with the conditioning $\boldsymbol{d}[0,n) = \boldsymbol{p}$ it is impossible for the pattern to begin in any position i, $1 \leq i < n - n'$. This impossibility follows from the fact that otherwise $\boldsymbol{d}[i,n)$ would be both a prefix and a suffix of $\boldsymbol{p}$ of length $n'' = n - i$ which, since $n' < n'' < n$, would contradict the fact that n' is the length of the longest bifix of $\boldsymbol{p}$. Equation (19) simply states that the search for $\boldsymbol{p}$ in the data stream $\boldsymbol{d}$ with the condition $\boldsymbol{d}[0,n) = \boldsymbol{p}$ consists of $n - n' - 1$ steps in which the search cannot possibly terminate, followed by a search for $\boldsymbol{p}$ in a random data stream in which the first n' digits are known to equal $\boldsymbol{p}_k$. Invoking Lemma 1, we obtain from (19)

$$E\{x \mid \boldsymbol{d}[1, n'+1) = \boldsymbol{p}_k\} = 1 + L^n - n + n' \tag{20}$$

and substitution of (17) and (20) into (18) then gives

$$E\{x\} = \sum_{i=0}^{n'} h_i L^i + L^n - n = \sum_{i=0}^{n} h_i L^i - n \tag{21}$$

which establishes the theorem.

Corollary: For any n-digit pattern $\boldsymbol{p}$, the expected search to find $\boldsymbol{p}$ in random data satisfies

$$L^n - n + 1 \leq E\{x\} \leq \frac{L^{n+1} - 1}{L - 1} - n \tag{22}$$

with equality on the left if and only if $\boldsymbol{p}$ is bifix-free and equality on the right if and only if $\boldsymbol{p}$ consists of n repetitions of the same digit of A_L.

Inequality (22) follows immediately from (16) upon noting that $h_i = 0$ for $0 < i < n$ if and only if $\boldsymbol{p}$ is bifix-free and $h_i = 1$ for $0 < i < n$ if and only if $\boldsymbol{p}$ is a constant sequence. We note also that the upper and lower bounds in (22) differ by a factor of only (slightly less than) two for the important case of a binary alphabet ($L = 2$) and that their ratio approaches 1 with increasing alphabet size.

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

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