Discrete Mathematics 310 (2010) 1152–1159

Contents lists available at ScienceDirect

Discrete Mathematics

journal homepage: www.elsevier.com/locate/disc

The discrepancy of the lex-least de Bruijn sequence

Joshua Cooper [a,*], Christine Heitsch [b]

[a] Department of Mathematics, University of South Carolina, 1523 Greene Street, Columbia, SC 29208, USA
[b] School of Mathematics, 686 Cherry Street, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 22 March 2009
Received in revised form 7 October 2009
Accepted 10 November 2009
Available online 24 November 2009

Keywords:
De Bruijn sequence
Ford sequence
Discrepancy
Greedy algorithm

ABSTRACT

We answer the following question: What is the discrepancy of the lexicographically least binary de Bruijn sequence? Here, "discrepancy" refers to the maximum (absolute) difference between the number of ones and the number of zeros in any initial segment of the sequence. We show that the answer is $\Theta(2^n \log n/n)$.

1. Introduction

A *binary de Bruijn sequence of order k* is a word $a_1 \cdots a_{2^k}$ over the alphabet $\{0, 1\}$ that contains every k-word exactly once as a subword when the indices are interpreted cyclically. It is well known (see, e.g., [12]) that the number of de Bruijn cycles of order k is given by

$$2^{2^{k-1}-k}.$$

Among these is the "Ford sequence",[1] the remarkable cyclic binary word which is

1. the lexicographic least de Bruijn sequence,
2. the result of applying the least-first greedy algorithm to constructing a de Bruijn sequence (starting with 1^k),
3. the result of concatenating all "Lyndon" words (lexicographically minimal representatives of free conjugacy classes) of each length dividing k in lexicographic order, and
4. the de Bruijn sequence generated by a shift register whose truth table has minimum weight.

Every de Bruijn sequence has a number of random-like properties: each word of the appropriate length appears as a subsequence, the number of runs of various lengths is "right", the number of 0's equals the number of 1's, etc. (See [7] for a classical discussion.) However, the Ford sequence is very non-random-like in another sense. Since the greedy algorithm uses 0's before 1's whenever possible, it is natural to suspect that this special sequence has an excess of 0's early on, i.e, the difference between the number of 0's and 1's in initial segments is large. Indeed, Huang comments in [8] that

* Corresponding author. Tel.: +803 777 3180; fax: +803 777 3783.
E-mail addresses: cooper@math.sc.edu (J. Cooper), heitsch@math.gatech.edu (C. Heitsch).

[1] See the excellent survey [6] for a history of this and related sequences. The eponym, due to Fredricksen, refers to a 1957 unpublished manuscript of Ford [5]. However, subsequent research has revealed earlier references. In [6], the author proposes that a 1934 paper of Martin [13] is the earliest appearance. Knuth [11] agrees, and refers to this sequence as the "granddaddy of all de Bruijn cycle constructions."

0012-365X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.disc.2009.11.010

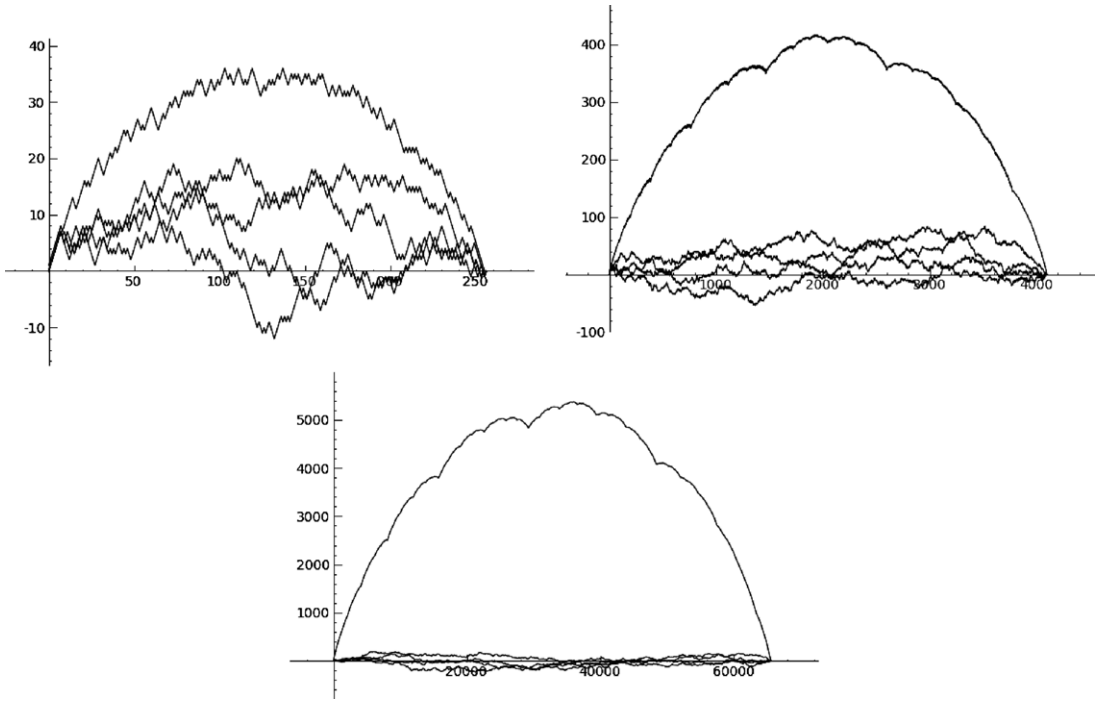

Fig. 1. The discrepancy of the Ford sequence versus four uniformly random choices of linear de Bruijn sequences for $n = 8, 12, 16$.

> The "prefer one" algorithm proposed by Fredricksen joins the pure cycles of [a] circulating register (CR) in order according to the weights of the n-tuples... so some part of the sequence may contain many heavily weighted n-tuples and it leads to a bad local 0–1 balance.

Therefore, for some applications of pseudorandom bit strings, the Ford sequence should be avoided. For example, a sequence with a large "discrepancy" used as a carrier signal in direct-sequence spread spectrum communications could generate an unwanted spectral peak at the symbol rate, risking interference with devices operating near this frequency [9]. Alternatively, the biochemical properties of DNA probes constructed with the use of de Bruijn sequences (see, for example, [2,15]) may display unacceptable variation if the distribution of symbols is too biased.

In the present note, we show that the maximum discrepancy has order $2^n \log n/n$. To compare, a random sequence of 2^n bits has maximum discrepancy of order $2^{n/2}\sqrt{\log n}$, by the law of the iterated logarithm. It is also known that every de Bruijn sequence has discrepancy at most $O(2^n/\sqrt{n})$. (See [3,4], where discrepancy arises as a character sum over non-linear recurrence sequences.) Fig. 1 shows a comparison between some random de Bruijn cycles and the Ford sequence.

Define the equivalence relation $\sim$ ("conjugacy") on binary words by setting $xy \sim yx$ for any $x, y \in \{0, 1\}^*$. For a word $w \in \{0, 1\}^*$, define w° to be the lexicographic least element of the $\sim$-equivalence class $[\![w]\!]$ of w. If w is aperiodic (i.e., if $w = xy$ with $x, y \neq \epsilon$, then $w \neq yx$), then w° is called a "Lyndon word". Then the lexicographically least binary order-n de Bruijn sequence $\mathcal{L}_n$ consists of the concatenation of all Lyndon words of length dividing n, *in lexicographic order*.

For a word $w \in \{0, 1\}^*$, write w_k for its kth symbol from left to right, starting with zero. Then we define the *discrepancy* of w to be

$$disc(w) = \max_M \left| \sum_{k=0}^{M} (-1)^{w_k} \right|.$$

Theorem 1. $disc(\mathcal{L}_n) = \Theta(2^n \log n/n)$.

We conjecture a slightly stronger statement:

Conjecture 1. *There is some C so that* $\lim_{n\to\infty} \frac{n\, disc(\mathcal{L}_n)}{2^n \log n} = C$.

Our argument will estimate the discrepancy of $\mathcal{L}_n$ by considering substrings consisting of Lyndon words w° grouped by the length k of their 0^k1 prefix. For $0 < k < n$, let S_k be the set of binary words of length n which, when the indices are read cyclically, contain the subword 0^k but not the subword 0^{k+1}. For instance, $w = 00100 \in S_4$. Thus, the elements of S_k are

precisely those w so that w° begins with 0^k. Define $S_k^\circ = \{w^\circ : w \in S_k\}$, and let ℓ_k be the concatenation of the elements of S_k° in lexicographic order. Since the elements of S_k° precede those of S_{k-1}° in the lexicographic order, this means that

$$\mathcal{L}_n = 0 \cdot \left(\prod_{k=1}^{n-1} \ell_{n-k} \right) \cdot 1,$$

as long as n is prime.

For a binary string w of length n, we define the *skew* of w to be

$$\mathrm{sk}(w) = \sum_{i=0}^{n-1} (-1)^{w_i}$$

so that

$$disc(\mathcal{L}_n) = \max_{1 \le t \le n-2} \left(1 + \sum_{k=1}^{t} \mathrm{sk}(\ell_{n-k}) + disc(\ell_{n-t-1}) \right)$$

when n is prime. This will allow us to bound the discrepancy of $\mathcal{L}_n$.

2. Preliminaries

Define $\alpha_k(n)$ to be the number of elements of $\{0, 1\}^n$ containing no subword 0^k, and let $\beta_k(n)$ be defined by

$$\beta_k(n) = \sum_{\substack{w \in \{0,1\}^n \\ 0^k \notin w}} \mathrm{sk}(w).$$

For the remainder of this section, we fix a $k \ge 2$.

Lemma 2. *The sequences $a_n = \alpha_k(n)$ and $b_n = \beta_k(n)$ satisfy:*

1. $a_n = \sum_{j=1}^{k} a_{n-j}$ *for* $n \ge k$, *and*
2. $b_n = \sum_{j=1}^{k} [(j-2)a_{n-j} + b_{n-j}]$ *for* $n \ge k$.

Furthermore, $a_j = 2^j$ for $0 \le j < k$ and $b_j = 0$ for $0 \le j < k$.

Proof. Both recurrences follow from the following consideration: any string of length at least k not containing a subword 0^k has a left-most 1. Therefore, we may partition the 0^k-free sequences into those which begin with a string of the form $0^j 1$ for $0 \le j < k$. The "base case" formulas trivially follow from the fact that *every* string of length less than k is 0^k-free. □

Lemma 3. *For $n - 1 \ge k \ge 3$,*

$$a_{n-1} = k + \sum_{j=3}^{k} (j-2)a_{n-j} + (k-1) \sum_{j=0}^{n-k-1} a_j.$$

Proof. We proceed by induction. First, we verify that $a_k = k + \sum_{j=3}^{k} (j-2)a_{k+1-j} + (k-1)a_0$. Note that, by the "base case" part of Lemma 2, $a_j = 2^j$ in the relevant range, except that $a_k = 2^k - 1$. Therefore,

$$\begin{aligned}
k + \sum_{j=3}^{k} (j-2)a_{k+1-j} + (k-1)a_0 &= k + \sum_{j=3}^{k} (j-2)2^{k+1-j} + k - 1 \\
&= \sum_{j=1}^{k-2} j 2^{k-j-1} + 2k - 1 \\
&= 2^{k-2} \sum_{j=1}^{k-2} j 2^{-(j-1)} + 2k - 1 \\
&= 2^{k-2}(4 - k2^{-k+3}) + 2k - 1 \\
&= 2^k - 2k + 2k - 1 \\
&= 2^k - 1 = a_k.
\end{aligned}$$

Now, suppose the statement holds for n. Applying the first recurrence in Lemma 2,

$$\begin{aligned}
a_n &= \sum_{j=1}^{k} a_{n-j} \\
&= a_{n-1} + \sum_{j=2}^{k} a_{n-j} \\
&= k + \sum_{j=3}^{k} (j-2)a_{n-j} + (k-1)\sum_{j=0}^{n-k-1} a_j + \sum_{j=2}^{k} a_{n-j} \\
&= k + \sum_{j=2}^{k} (j-1)a_{n-j} + (k-1)\sum_{j=0}^{n-k-1} a_j \\
&= k + \sum_{j=3}^{k} (j-2)a_{n+1-j} + (k-1)a_{n-k} + (k-1)\sum_{j=0}^{n-k-1} a_j \\
&= k + \sum_{j=3}^{k} (j-2)a_{n+1-j} + (k-1)\sum_{j=0}^{n-k} a_j. \quad \square
\end{aligned}$$

Corollary 4. *$b_n < 0$ for all $n - 1 \geq k \geq 3$.*

Proof. If we combine the recurrence for b_n from Lemma 2 with the above Lemma 3,

$$\begin{aligned}
b_n &= \sum_{j=1}^{k} [(j-2)a_{n-j} + b_{n-j}] \\
&= -a_{n-1} + \sum_{j=3}^{k} (j-2)a_{n-j} + \sum_{j=1}^{k} b_{n-j} \\
&= -k - (k-1)\sum_{j=0}^{n-k-1} a_j + \sum_{j=1}^{k} b_{n-j} < 0, \qquad (1)
\end{aligned}$$

by induction. $\square$

Let ρ_k be the largest (in absolute value) root of the polynomial $g(z) = z^{k+1} - 2z^k + 1$. It is proven in [14] that ρ_k is real, lies between 5/3 and 2, and is unique in these respects. It is also shown in [14] that $\rho_k \to 2$ as $k \to \infty$. Note that

$$z^k - \sum_{j=0}^{k-1} z^j = \frac{z^{k+1} - 2z^k + 1}{z - 1},$$

so that ρ_k is a root of the left-hand polynomial $f(z)$ here as well. Since $f(z)$ is the characteristic polynomial for the recurrence that the a_n satisfy, ρ_k is the growth rate of the a_n, i.e., $\lim_{n\to\infty} \log a_n/n = \rho_k$.

Lemma 5. *For all $n \geq 1$, $a_n \geq \rho_k a_{n-1}$.*

Proof. Since $\rho_k < 2$, and $a_n = 2^n$ for $0 \leq n < k$, the claimed bound holds for n in this range. Suppose it holds for all $n < N$. Then by Lemma 2,

$$\begin{aligned}
a_n &= \sum_{j=1}^{k} a_{n-j} \\
&\geq \sum_{j=1}^{k} \rho_k a_{n-j-1} \\
&= \rho_k a_{n-1}. \quad \square
\end{aligned}$$

Lemma 6. *For $k \geq 4$ and all $n \geq k$, $b_n \geq -2ka_n/3$.*

Proof. By (1),

$$b_n = -k - (k-1)\sum_{j=0}^{n-k-1} a_j + \sum_{j=1}^{k} b_{n-j}.$$

If we suppose that $b_j \geq -\gamma k a_j$ for all $j < n$, then

$$\begin{aligned} b_n &\geq -k - k\sum_{j=0}^{n-k-1} a_j - \gamma k \sum_{j=1}^{k} a_{n-j} \\ &= -k - k\sum_{j=0}^{n-k-1} a_j - \gamma k a_n. \end{aligned}$$

By iterating Lemma 5, we have

$$\begin{aligned} b_n &\geq -k - k\sum_{j=0}^{n-k-1} \rho_k^{j-n} a_n - \gamma k a_n \\ &\geq -a_n k\left(\frac{1}{a_n} + \sum_{j=0}^{\infty} \rho_k^{j-n} + \gamma\right) \\ &= -a_n k\left(\frac{1}{a_n} + \frac{\rho_k^{-n}}{1-\rho_k^{-1}} + \gamma\right) \\ &\geq -a_n k\left(\frac{1}{a_n} + \frac{5}{2}\rho_k^{-n} + \gamma\right). \end{aligned}$$

We may begin by taking $\gamma = \frac{2k-1}{k(2^{k+1}-3)} \leq \frac{7}{116}$ by considering $a_{k+1} = 2^{k+1} - 3$ and $b_{k+1} = 1 - 2k$. Then, γ increases by at most

$$\begin{aligned} \sum_{n=k+1}^{\infty}\left(\frac{1}{a_n} + \frac{5}{2}\rho_k^{-n}\right) &\leq \sum_{n=k+1}^{\infty} \frac{1}{\rho_k^{n-k} a_k} + \frac{5}{2}\sum_{n=k+1}^{\infty} \rho_k^{-n} \\ &= \frac{\rho_k^k}{2^k - 1}\sum_{n=k+1}^{\infty} \rho_k^{-n} + \frac{5}{2}\sum_{n=k+1}^{\infty} \rho_k^{-n} \\ &= \left(\frac{\rho_k^k}{2^k-1} + \frac{5}{2}\right)\rho_k^{-k-1}\sum_{n=0}^{\infty} \rho_k^{-n} \\ &= \left(\frac{\rho_k^{-1}}{2^k-1} + \frac{5}{2\rho_k^{k+1}}\right)\cdot\frac{1}{1-\rho_k^{-1}} \\ &\leq \left(\frac{3}{5\cdot 15} + \frac{5}{2(5/3)^5}\right)\cdot\frac{5}{2} = \frac{293}{500}. \end{aligned}$$

The conclusion follows for all $n \geq k+1$, since $\frac{293}{500} + \frac{7}{116} = \frac{2343}{3625} \leq \frac{2}{3}$. It is also easy to verify that $b_k \geq -2ka_k/3$. □

3. Main result

Here we prove Theorem 1 stated in the introduction.

Proposition 7. *For $4 \leq k < n$ and n prime,*

$$\frac{k}{3} - 2 \leq \frac{\mathrm{sk}(\ell_k)}{\alpha_{k+1}(n-k-2)} \leq 2k - 3.$$

Proof. The set S_k contains each sequence of the form $0^k 1w$ where w is a 0^k-free word of length $n-k-1$. However, the quantity $\mathrm{sk}(S_k)$ is not quite the sum of the skews of all 0^k-free sequences of length $n-k-1$ prefixed by 0^k1: it must include all elements of S_k°, not just those that have prefix 0^k and contain no other runs 0^k. For each word w of length n which contains more than one run of the form 0^k, but no runs of the form 0^{k+1}, only one of its conjugates (namely, w°) appears in

S_k°. Define $\mathrm{run}(w)$ to be the maximum k so that $0^k \in w$, and let $\rho_k(w)$ be the number of subwords of the form 0^k in w, where $\mathrm{run}(w) = k$. (Set $\rho_k(w) = 0$ otherwise.) Since we may assume that each w is aperiodic, this means that

$$\begin{aligned}
\mathrm{sk}(\ell_k) &= \sum_{w\in S_k^\circ} \mathrm{sk}(w)\\
&= \sum_{\substack{w\in\{0,1\}^n\\ \mathrm{run}(w)=k}} \mathbb{1}(w = w^\circ)\,\mathrm{sk}(w)\\
&= \sum_{t\geq 0} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} \frac{\mathrm{sk}(0^k1w1)}{t+1}\\
&= \sum_{t\geq 0} \frac{1}{t+1} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2+\mathrm{sk}(w)).
\end{aligned}$$

Define the "run-print" $\mathrm{rp}(w)$ of a word $w \in \{0,1\}$ with $\mathrm{run}(w) = k$ to be the set of indices $j \in [n]$ so that w has a run 0^k starting at index j. Then we may write

$$\begin{aligned}
\mathrm{sk}(\ell_k) &= \sum_{t\geq 0} \frac{1}{t+1} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2+\mathrm{sk}(w))\\
&= \sum_{t\geq 0} \frac{1}{t+1} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2) + \sum_{t\geq 0} \frac{1}{t+1} \sum_{S\in\binom{[n-k-2]}{t}} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \mathrm{rp}(w)=S}} \mathrm{sk}(w).
\end{aligned}$$

Now, for a given S of cardinality t and w with $\mathrm{rp}(w) = S$, there is a 0^k run starting at location s for each $s \in S$. Each such run is bounded on both sides by a 1. In between the runs are intervals, the sum over whose skews is nonpositive, by Corollary 4. Therefore,

$$\sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \mathrm{rp}(w)=S}} \mathrm{sk}(w) \leq \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \mathrm{rp}(w)=S}} t(k-1),$$

so we have

$$\begin{aligned}
\mathrm{sk}(\ell_k) &\leq \sum_{t\geq 0} \frac{1}{t+1} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2) + \sum_{t\geq 0} \frac{1}{t+1} \sum_{S\in\binom{[n-k-2]}{t}} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \mathrm{rp}(w)=S}} t(k-1)\\
&< \sum_{t\geq 0} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2) + (k-1) \sum_{t\geq 0} \sum_{S\in\binom{[n-k-2]}{t}} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \mathrm{rp}(w)=S}} 1\\
&= (k-2)\alpha_{k+1}(n-k-2) + (k-1)\alpha_{k+1}(n-k-2)\\
&= (2k-3)\alpha_{k+1}(n-k-2).
\end{aligned}$$

On the other hand, by Lemma 6,

$$\begin{aligned}
\mathrm{sk}(\ell_k) &= \sum_{t\geq 0} \frac{1}{t+1} \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} (k-2+\mathrm{sk}(w))\\
&\geq \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=0}} (k-2) + \sum_{\substack{w\in\{0,1\}^{n-k-2}\\ \rho_k(w)=t}} \mathrm{sk}(w)\\
&= (k-2)\alpha_{k+1}(n-k-2) + \beta_{k+1}(n-k-2)\\
&\geq (k/3-2)\alpha_{k+1}(n-k-2). \quad \square
\end{aligned}$$

In the proof of Theorem 1 below, we use the following useful inequality of Janson (see, for example, [10]). The lower bound is standard; the upper bound is an easy modification of the one presented in [1]. Let X be a finite set and let P be a random subset of X, with elements $x \in X$ chosen independently with probability p_x. Let $\{Z_i : i \in \mathcal{I}\}$ be a system of subsets of X, and let A_i denote the event that $Z_i \subset P$. If $Z_i \cap Z_j = \emptyset$, then A_i and A_j are independent. Let

$$\Delta = \sum P(A_i \wedge A_j),$$

where the sum is taken over all ordered pairs $i \neq j$ with $Z_i \cap Z_j \neq \emptyset$. Finally, define $\mu = \sum_i P(A_i)$.

Lemma 8. *With μ, Δ as above, if $\Delta \geq \mu/2$, then*

$$e^{-\mu} \leq \bigwedge_{i\in I} \overline{A_i} \leq e^{-\mu^2/3\Delta}.$$

Proof of Theorem 1. Suppose for the moment that n is prime and $k \geq 4$. We know that

$$disc(\mathcal{L}_n) = \max_k \left(1 + \sum_{j=1}^{k-1} \mathrm{sk}(\ell_{n-j}) + disc(\ell_{n-k})\right).$$

From Proposition 7, we have that

$$\begin{aligned}\sum_{k=\log n+1}^{n} \mathrm{sk}(\ell_k) &\geq \sum_{k=\log n+1}^{n} (k/3-2)\cdot\alpha_{k+1}(n-k-2)\\ &\geq \sum_{k=\log n+1}^{n} (k/3-2)\cdot 2^{n-k-1}(1-n2^{-k})\\ &= \Omega\left(\frac{2^n \log n}{n}\right).\end{aligned}$$

On the other hand, for any t,

$$\begin{aligned}\sum_{k=t}^{n-1} \mathrm{sk}(\ell_k) &\leq \sum_{k=0}^{n-2} (2k-3)\cdot\alpha_{k+1}(n-k-2)\\ &\leq \sum_{k=1}^{n-1} 2k\cdot\alpha_k(n-k-1).\end{aligned}$$

We estimate this quantity using the inequality of Janson stated above. In this case, we take $X = [n]$, P is the set of indices where a 0 appears, $p_x = 1/2$ for every x, $\mathcal{I} = [n-k+1]$, $Z_i = [i, i+k-1]$ (i.e., the ith length k interval of $[n]$), and A_i is the event that a length n word has a subsequence of the form 0^k on some Z_i. Then

$$\mu = (n-k+1)2^{-k}$$

and

$$\begin{aligned}\Delta &= \sum_{\substack{1\leq i,j\leq n-k+1\\ 0<|i-j|<k}} 2^{-k-|i-j|}\\ &< 2^{-k+1}(n-k+1)\sum_{s=1}^{\infty} 2^{-s} = 2^{-k+1}(n-k+1) = 2\mu.\end{aligned}$$

Furthermore,

$$\Delta \geq 2^{-k}(n-k+1)\sum_{s=1}^{k-1} 2^{-s} > 2^{-k-1}(n-k+1) = \mu/2,$$

so the hypotheses hold. Therefore, for a uniform random choice of $w \in \{0,1\}^n$,

$$P(0^k \notin w) \leq e^{-\mu/12} = e^{-(n-k+1)/(12\cdot 2^k)}.$$

Applying this bound to the above computations,

$$k\alpha_k(n-k-1) \leq k\cdot 2^{n-k} e^{-(n-2k)/(12\cdot 2^k)}.$$

Let $T = \lfloor \log n \rfloor$. Then

$$\begin{aligned}\sum_{k=1}^{n-1} k\alpha_k(n-k-1) &\leq \sum_{k=1}^{n-1} k\cdot 2^{n-k} e^{-(n-2k)/(12\cdot 2^k)}\\ &= 2^n \sum_{k=1}^{2\log n} k\cdot 2^{-k} e^{-(n-2k)/(12\cdot 2^k)} + 2^n \sum_{k=2\log n+1}^{n-1} k\cdot 2^{-k} e^{-(n-2k)/(12\cdot 2^k)}\end{aligned}$$

$$\begin{aligned}
&\leq 2^n \sum_{k=-\infty}^{\infty} k \cdot 2^{-k} \mathrm{e}^{-n/(24\cdot 2^k)} + o\left(\frac{2^n \log n}{n}\right) \\
&\leq 2^n \sum_{k=-\infty}^{\infty} (T-k) \cdot 2^{k-T} \mathrm{e}^{-n/(24\cdot 2^{T-k})} + o\left(\frac{2^n \log n}{n}\right) \\
&\leq 2^n \sum_{k=-\infty}^{\infty} \frac{2\log n}{n} \cdot 2^k \mathrm{e}^{-2^k/48} + o\left(\frac{2^n \log n}{n}\right) \\
&= O\left(\frac{2^n \log n}{n}\right) \cdot \sum_{k=-\infty}^{\infty} 2^k \mathrm{e}^{-2^k/48} = O\left(\frac{2^n \log n}{n}\right).
\end{aligned}$$

Therefore, the total discrepancy is $\Theta(2^n \log n/n)$.

There are two more terms to consider: $\mathrm{sk}(\ell_k)$ with $k \leq 3$, and $\max_k disc(\ell_{n-k})$. The former terms are bounded by $O(\rho_4^k) = O(1.93^k)$, and therefore make an insignificant contribution. As for the latter, the length of ℓ_{n-k} is bounded above by $\alpha_{k+1}(n-k-2)$, and the above analysis shows that this quantity is $o(2^n \log n/n)$. Since the length of ℓ_{n-k} is an upper bound for $disc(\ell_{n-k})$, this term also does not affect the order of $disc(\mathcal{L}_n)$.

Finally, we may drop the assumption that n is prime. If not, then the above analysis is wrong: some words of length n, which would be part of the concatenation that gives rise to an ℓ_k, are in fact periodic, and therefore only appear as their minimal roots in $\mathcal{L}_n$. (All Lyndon words of length dividing n arise in this way.) However, the total number of symbols they contribute is at most

$$\sum_{d|n, d<n} d2^d < n^2 2^{n/2} = o\left(\frac{2^n \log n}{n}\right).$$

Hence, the asymptotic bound holds. □

Acknowledgements

Thanks to Aaron Dutle for his careful reading of an earlier draft of this paper, and to Ron Graham for suggesting the problem.

References

[1] N. Alon, J. Spencer, The Probabilistic Method, in: Wiley-Interscience Series in Discrete Math. and Optimization, John Wiley & Sons, Inc, Hoboken, NJ, 2008.
[2] A. Ben-Dor, R. Karp, B. Schwikowski, Z. Yakhini, Universal DNA tag systems: A combinatorial design scheme, J. Comput. Biol. 7 (3–4) (2000) 503–519.
[3] S.R. Blackburn, I.E. Shparlinski, Character sums and nonlinear recurrence sequences, Discrete Math. 306 (12) (2006) 1126–1131.
[4] G. Everest, A.J. van der Poorten, I.E. Shparlinski, T.B. Ward, Recurrence Sequences, Amer. Math. Soc., 2003.
[5] L.R. Ford, A cyclic arrangement of m-tuples, Report P-1071, Rand Corp., Santa Monica, CA, 1957.
[6] H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, SIAM Rev. 24 (2) (1982) 195–221.
[7] S.W. Golomb, Shift Register Sequences, Holden-Day, Inc, San Francisco, 1967.
[8] Y.J. Huang, A new algorithm for the generation of binary de Bruijn sequences, J. Algorithms 11 (1) (1990) 44–51.
[9] V.P. Ipatov, Spread Spectrum and CDMA: Principles and Applications, John Wiley & Sons Ltd, West Sussex, 2005.
[10] S. Janson, Poisson approximation for large deviations, Random Structures Algorithms 1 (1990) 221–229.
[11] D. Knuth, The Art of Programming, Section 7.2.1.1, Pre-Fascicle 2A, 2001.
[12] J.H. van Lint, R.M. Wilson, A Course in Combinatorics, Second ed., Cambridge University Press, Cambridge, 2001.
[13] M.H. Martin, A problem in arrangements, Bull. Amer. Math. Soc. 40 (1934) 859–864.
[14] A.M. Odlyzko, Asymptotic enumeration methods, in: Handbook of Combinatorics, vol. 2, Elsevier, Amsterdam, 1995, pp. 1063–1229.
[15] A.A. Philippakis, A.M. Qureshi, M.F. Berger, M.L. Bulyk, Design of compact, universal DNA microarrays for protein binding microarray experiments, J. Comput. Biol. 15 (7) (2008) 655–665.