Counting Special Sets of Binary Lyndon Words

Joseph L. Yucas

Southern Illinois University
Carbondale, Illinois 62901-4408
U.S.A.

1. Introduction.

A *binary string of length* n will be an n-tuple written $w = w_1 w_2 \dots w_n$ with $w_i \in \{0, 1\}$ for $i = 1, 2, \dots, n$. We let Z_2^n denote the set of all binary string of length n and set $Z_2^* = \cup_n Z_2^n$, the union over all non-negative integers n. The action of the permutation $\pi = (12 \dots n)$ on Z_2^n given by $w^\pi = w_{\pi(1)} w_{\pi(2)} \dots w_{\pi(n)}$ yields an equivalence relation on Z_2^n; $v \sim w$ if $v = w^{\pi^m}$ for some positive integer m. The resulting equivalence classes are referred to as *circular binary strings.*

A binary string $w \in Z_2^n$ is *aperiodic* if $w \neq v^m$ for any substring v and positive integer m, where v^m denotes the concatenation of m copies of v. A circular string is *aperiodic* if every word in the equivalence class is aperiodic or equivalently, if the equivalence class contains n distinct binary strings. By an elementary Möbius inversion, (see [2]), the number of aperiodic binary circular strings of length n is given by $\frac{1}{n} \sum_{d|n} \mu(n/d) 2^d$ where μ is the Möbius function of elementary number theory.

For two strings, u, v in Z_2^* we say that u is *lexicographically less* than v, written $u < v$, if

(i) $v = uw$ for some non-empty string $w \in Z_2^*$, or
(ii) $u = ras$, $v = rbt$ for some $r, s, t \in Z_2^*$ and some non-empty $a, b \in Z_2^*$ with $a < b$.

We let L_n be the set of binary strings of length $n \geq 1$ in Z_2^n which are lexicographically least in the aperiodic equivalence classes determined by $\sim$. The strings in L_n are called *Lyndon words of length* n. As above $|L_n| = \frac{1}{n} \sum_{d|n} \mu(n/d) 2^d$.

Recent interest in L_n stems from [1] and [3] where L_n is used for a code with bounded synchronization delay and where Hamilton paths are built in the n-cube from words in L_n. A well-known classification of Lyndon words is given in the following proposition. A proof of Proposition 1.1 may be found in [4].

Proposition 1.1. *For a string* w*, the following statements are equivalent:*

(1) w *is a Lyndon word;*
(2) $w = uv$ *where* u *and* v *are Lyndon words with* $u < v$*;*
(3) w *is strictly less than each of its proper right factors.*

The equivalence of 1 and 2 above yields a recursive algorithm for generating all the words in L_n but unfortunately many repetitions are generated.

ARS COMBINATORIA 31(1991), pp. 21-29

Some strings are obviously Lyndon words. Namely, the words that end with 1 and begin with a string of 0 's longer than any other string of 0 's appearing in the word. We let Z_n denote the collection of these Lyndon words that have length n with one exception. We do not include $0\ \overbrace{1\ 1 \ldots 1}^{n-1}$ in Z_n for reasons which will become apparent later. Let $B_n = L_n \backslash Z_n$.

Example: The 18 Lyndon words in L_7 are

Z_n	B_n
0000001	0010011
0000011	0101111
0000101	0101011
0000111	0110111
0001001	0111111
0001011	
0001101	
0001111	
0010101	
0010111	
0011011	
0011101	
0011111	

Fortunately, words of the type in Z_n account for most of the Lyndon words and they can all be constructed by preceeding words that end in 1 with enough zeros. In the following section we give a count of $|Z_n|$ and $|B_n|$ in terms of the $|F_n^k|$'s of Table 1.

Section 2.

Let F_n^k be the strings of length n having at least one substring of k 0 's but no substring of (k + 1) 0 's.

Proposition 2.1. *If* $0 \leq n < k$ *then* $|F_n^k| = 0$. $|F_n^n| = 1$ *and if* $n > k$

$$|F_n^k| = \sum_{i=1}^{k} |F_{n-i}^k| + \sum_{i=0}^{k} |F_{n-k-1}^i|.$$

Proof: If $n > k$, every string in F_n^k can be written in the form $\overbrace{0\ 0 \ldots 0}^{i-1}\ 1\ w$ for some string w of length $n - i$ and some $1 \leq i \leq k + 1$. Now there are $|F_{n-i}^k|$ strings in F_n^k of the form $\overbrace{0\ 0 \ldots 0}^{i-1}\ 1\ w$ for $1 \leq i \leq k$ and there are

$$|F_{n-k-1}^0| + |F_{n-k-1}^1| + \ldots + |F_{n-k-1}^k|$$

strings in F_n^k of the form $\overbrace{0\ 0\ \ldots 0}^{k}\ 1\ w$. ∎

Table 1 gives the values of $|F_n^k|$ for $0 \leq n \leq 20, 0 \leq k \leq 10$. Notice that $|F_n^k|$ is just the number of compositions with no part greater than k and at least one part equal to k, see, for example, [5, Example 12, p. 155]. The limiting sequence $1, 2, 5, 12, 28, 64, 144, \ldots$ has $(n+1)$th term given by $(n+3)2^{n-2}$ for $n \geq 2$. By Proposition 2.1 the (n, k)th entry in Table 1 is obtained by adding along row $n-k-1$ until column k then adding down column k from row $n-k$ to row $n-1$. For example,

$$\begin{aligned} 1+7+&\ 5 \\ &+11 \\ &+23 \\ &=47 \end{aligned}$$

At times, in the formulae that follow, $|F_i^k|$ with $i < 0$ will appear. By convention, in this case, take $|F_{-1}^0| = 1$ and $|F_i^k| = 0$ otherwise.

Theorem 2.2. $|Z_1| = |Z_2| = 0$. *For* $n \geq 3$,

$$|Z_n| = \sum_{k=0}^{[\frac{n-3}{2}]} \sum_{i=k}^{n-k-3} |F_i^k|.$$

Proof: Every word in Z_n can be written $\overbrace{0\ 0\ \ldots 0}^{i}\ 1\ w\ 1$, where $w \in \bigcup_{k=0}^{i-1} F_{n-i-2}^k$. Thus we count, for n odd

type	**number of words of this type**
$\overbrace{0\ 0\ \ldots 0}^{n-2}\ 1\ 1$	$\lvert F_0^0\rvert$
$\overbrace{0\ 0\ \ldots 0}^{n-3}\ 1\ w\ 1$	$\lvert F_1^0\rvert + \lvert F_1^1\rvert$
$\overbrace{0\ 0\ \ldots 0}^{n-4}\ 1\ w\ 1$	$\lvert F_2^0\rvert + \lvert F_2^1\rvert + \lvert F_2^2\rvert$
$\vdots$	$\vdots$
$\overbrace{0\ \ldots 0}^{\frac{n-1}{2}}\ 1\ w\ 1$	$\lvert F_{\frac{n-3}{2}}^0\rvert + \ldots + \lvert F_{\frac{n-3}{2}}^{\frac{n-3}{2}}\rvert$
$\vdots$	$\vdots$
$0\ 0\ 1\ w\ 1$	$\lvert F_{n-4}^0\rvert + \lvert F_{n-4}^1\rvert$
$\overbrace{0\ 0\ \ldots 0}^{n-1}\ 1$	$\lvert F_{n-3}^0\rvert$

and for n even

type	**number of words of this type**
$\overbrace{0\,0\ldots0}^{n-2}\,1\,1$	$\lvert F_0^0\rvert$
$\overbrace{0\,0\ldots0}^{n-3}\,1\,w\,1$	$\lvert F_1^0\rvert+\lvert F_1^1\rvert$
$\vdots$	$\vdots$
$\overbrace{0\,0\ldots0}^{n/2}\,1\,w\,1$	$\lvert F^0_{\frac{n-4}{2}}\rvert+\ldots+\left\lvert F^{\frac{n-4}{2}}_{\frac{n-4}{2}}\right\rvert$
$\overbrace{0\,0\ldots0}^{\frac{n-2}{2}}\,1\,w\,1$	$\lvert F^0_{\frac{n-2}{2}}\rvert+\ldots+\left\lvert F^{\frac{n-4}{2}}_{\frac{n-2}{2}}\right\rvert$
$\vdots$	$\vdots$
$0\,0\,1\,w\,1$	$\lvert F^0_{n-4}\rvert+\lvert F^1_{n-4}\rvert$
$\overbrace{0\,0\ldots0}^{n-1}\,1$	$\lvert F^0_{n-3}\rvert$

where $\overbrace{0\,0\ldots0}^{n-1}\,1$ is replacing $0\,\overbrace{1\,1\ldots1}^{n-1} \notin Z_n$ in the natural order of enumeration. In either case, adding down the columns yields the desired result. ∎

Thus, $|Z_n|$ is given by adding the entries in an upper triangular block of Table 1. For example,

$$|Z_7| = \begin{array}{l} \quad 1 \\ +\ 1+1 \\ +\ 1+2+1 \\ +\ 1+4 \\ +\ 1 \end{array} = 13 \quad \text{and} \quad |Z_8| = \begin{array}{l} \quad 1 \\ +\ 1+1 \\ +\ 1+2+1 \\ +\ 1+4+2 \\ +\ 1+7 \\ +\ 1 \end{array} = 23$$

As an easy corollary of Theorem 2.2 we get the following recursion describing $|Z_{n+1}|$. Corollary 2.3 may be expressed as adding the bottom of the triangle onto $|Z_n|$ to get $|Z_{n+1}|$.

Corollary 2.3.

$$|Z_{n+1}| = |Z_n| + \sum_{i=0}^{[\frac{n-2}{2}]} |F^i_{n-i-2}|.$$

Proof:

$$|Z_{n+1}| - |Z_n|$$
$$= \sum_{k=0}^{[\frac{n-2}{2}]} \sum_{i=k}^{n-k-2} |F_i^k| - \sum_{k=0}^{[\frac{n-3}{2}]} \sum_{i=k}^{n-k-3} |F_i^k| = \sum_{k=0}^{[\frac{n-2}{2}]} |F_{n-k-2}^k|.$$

∎

We now turn our attention to $|B_n|$. Let

$$G_n^k = \left\{ w \in F_n^k \mid w^{\pi^m} \in F_n^k, \ \forall\, m \geq 0 \right\}$$

where $\pi = (1\ 2 \ldots n)$. Notice that $B_n = L_n \cap \cup_k G_n^k$. This is why we included $0\ \overbrace{1\ 1 \ldots 1}^{n-1}$ in B_n instead of Z_n. Let

$$D_n^k(m) = \left\{ w \in G_n^k \mid w = v^m \text{ for some string } v \right\}.$$

Proposition 2.4.

$$|B_n| = \frac{1}{n} \sum_{k=1}^{[\frac{n}{2}]-1} \sum_{d|n} \mu(n/d) \mid D_n^k(d)|.$$

Proof: Let $S_n^k(m) = \{w \in D_n^k(m) \mid w \neq v^r \text{ for } r < m\}$. Then $|D_n^k(n)| = \sum_{d|n} S_n^k(d)$ so by Möbius inversion,

$$|S_n^k(n)| = \sum_{d|n} \mu(n/d)\, |D_n^k(d)|$$

and thus the result follows. ∎

Proposition 2.5. $|D_n^k(1)| = 0$ *for* $k < n$, $|D_n^k(n)| = |G_n^k|$, *and for* $d|n$, $d \neq 1, n$

$$|D_n^k(d)| = \sum_{i=0}^{k} (i|F_{d-i-1}^k| + (k+1)|F_{d-k-2}^i|).$$

Proof: For $d|n$, $d \neq 1, n$ notice that a string in $D_n^k(d)$ must be a repeated string of length d. There are $|F_{d-i-j-2}^k|$ of them that have the form of repeating $\overbrace{0\ 0 \ldots 0}^{i}$ $1\ w\ 1\ \overbrace{0\ 0 \ldots 0}^{j}$ where $0 \leq i, j \leq k-1$ and $i + j < k$. Hence a total of

$$\sum_{\substack{0 \leq i,j \leq k-1 \\ i+j<k}} |F_{d-i-j-2}^k| = \sum_{i=1}^{k} i|F_{d-i-1}^k|.$$

All other strings in $D_n^k(d)$ must have the form of repeating $\overbrace{0\ 0 \ldots 0}^{i}\ 1\ w\ 1\ \overbrace{0\ 0 \ldots 0}^{j}$ where $0 \leq i, j \leq k$ and $i + j = k$. There are $\sum_{m=0}^{k} F_{d-i-j-2}^{m}$ of these. Summing over all i, j with $i + j = k$ yields a total of $\sum_{i=0}^{k} (k+1)|F_{d-k-2}^{i}|$. Notice above that the string $\overbrace{0\ 0 \ldots 0}^{i}\ 1\ \overbrace{0\ 0 \ldots 0}^{j}$ is counted by F_{-1}^{0}. ∎

By Proposition 2.4 and Proposition 2.5 we see that in order to write $|B_n|$ in terms of the $|F_n^k|$'s it suffices to concentrate on the $|G_n^k|$'s. Let H_n^k denote the collection of strings in F_n^k having at least two substrings of k 0's.

Proposition 2.6. $|G_n^1| = |F_{n-1}^1| + |F_{n-3}^1| + 1$ *and for* $k > 1$

$$|G_n^k| = \sum_{i=1}^{k} i\,|H_{n-i-1}^k| + (k+1)|F_{n-k-2}^k|.$$

Proof: For $k = 1$, there are $|F_{n-1}^1|$ strings in G_n^1 of the form $1\ w$ and $|F_{n-3}^1| + |F_{n-3}^0|$ strings in G_n^1 of the form $0\ 1\ w\ 1$. For $k > 1$, there are $|H_{n-i-j-2}^k|$ strings in G_n^k of the form $\overbrace{0\ 0 \ldots 0}^{i}\ 1\ w\ 1\ \overbrace{0\ 0 \ldots 0}^{j}$ where $0 \leq i, j \leq k-1$ with $i + j < k$ yielding a total of $\sum_{i=1}^{k} i|H_{n-i-1}^k|$. All other strings in $|G_n^k|$ have the form $\overbrace{0\ 0 \ldots 0}^{i}\ 1\ w\ 1\ \overbrace{0\ 0 \ldots 0}^{j}$ where $0 \leq i, j \leq k$ and $i + j = k$. There are $|F_{n-i-j-2}^k|$ of these. Summing over all i, j with $i + j = k$ yields $(k+1)|F_{n-k-2}|$. ∎

Proposition 2.7.

$$|H_n^k| = \sum_{i=0}^{k} |H_{n-i}^k| + |F_{n-k-1}^k|.$$

Proof: Every string in H_n^k can be written $\overbrace{0\ 0 \ldots 0}^{i}\ 1\ w$ with $i < k$ and $w \in H_{n-i-1}^k$ or $\overbrace{0\ 0 \ldots 0}^{k}\ 1w$ with $w \in F_{n-k-1}^k$. Now, there are $|H_{n-i-1}^k|$ of the first type and $|F_{n-k-1}^k|$ of the second. Hence, $|H_n^k| = \sum_{i=0}^{k-1} |H_{n-i-1}^k| + |F_{n-k-1}^k|$. ∎

We let f_n^k be the kth-generalized Fibonacci sequence, that is,

$$f_n^k = \begin{cases} 1 & \text{if } n = 0 \\ 2^{n-1} & \text{if } 1 \leq n \leq k \\ \sum_{i=n-k}^{n-1} f_i^k & \text{if } n > k. \end{cases}$$

Proposition 2.8.

$$|H_n^k| = \sum_{i=1}^{n-2k} f_{i-1}^k |F_{n-k-i}^k|.$$

Proof: We induct on n. By Proposition 2.7,

$$|H_n^k| = \sum_{i=0}^{k-1} |H_{n-i-1}^k| + |F_{n-k-1}^k|.$$

By induction we have

$$\begin{array}{rlll} |H_n^k| = & f_0^k|F_{n-k-2}| + f_1^k|F_{n-k-3}^k| + \cdots & & + f_{n-2k-2}^k|F_k^k| \\ + & \quad f_0^k|F_{n-k-3}^k| + \cdots & & + f_{n-2k-3}^k|F_k^k| \\ + & \quad \ddots & & \\ \vdots & \qquad \ddots & & \\ + & \qquad f_0^k|F_{n-2k-1}^k| + \cdots + f_{n-3k-1}^k|F_k^k| + |F_{n-k-1}^k| & & \end{array}$$

Adding columns yields

$$|H_n^k| = f_1^k|F_{n-k-2}^k| + f_2^k|F_{n-k-3}^k| + \cdots + f_{n-2k-1}^k|F_k^k| + f_0^k|F_{n-k-1}^k|$$

since $1 = f_0^k = f_1^k$, $\sum_{i=1}^{s} f_i^k = f_{s+1}^k$ for $s < k$ and $\sum_{i=r-k}^{r-1} f_i^k = f_r^k$. ∎

We note in closing that even the generalized Fibonacci coefficients of Proposition 2.8 can be written in terms of the $|F_n^k|$'s as partial row sums.

Theorem 2.9.

$$f_n^k = \sum_{i=0}^{k-1} |F_{n-1}^i|.$$

Proof: For $n = 0$ notice that $f_n^k = 1$ and

$$\sum_{i=0}^{k-1} |F_{n-1}^i| = |F_{-1}^0| + |F_{-1}^1| + \cdots + |F_{-1}^{k-1}| = 1.$$

For $2 \le n \le k$, $f_n^k = 2^{n-1}$ and

$$\sum_{i=0}^{k-1} |F_{n-1}^i| = |F_{n-1}^0| + \cdots + |F_{n-1}^{n-1}|$$

which is a count of all binary strings of length $n-1$ and hence equal to 2^{n-1}. For $n > k$ we induct on n.

$$f_n^k = \sum_{i=n-k}^{n-1} f_i^k = \sum_{i=0}^{k-1} |F_{n-k-1}^i| + \sum_{i=0}^{k-1} |F_{n-k}^i| + \ldots + \sum_{i=0}^{k-1} |F_{n-2}^i|$$

$$\begin{aligned}
= \; & |F_{n-k-1}^0| + |F_{n-k-1}^1| + \cdots\cdots\cdots + |F_{n-k-1}^{k-1}| \\
+ \; & |F_{n-k}^0| + |F_{n-k}^1| + \cdots\cdots\cdots + |F_{n-k}^{k-1}| \\
+ \; & \vdots \\
+ \; & |F_{n-3}^0| + |F_{n-3}^1| + \cdots\cdots\cdots + |F_{n-3}^{k-1}| \\
+ \; & |F_{n-2}^0| + |F_{n-2}^1| + \cdots\cdots\cdots + |F_{n-2}^{k-1}| \\
& \quad \| \qquad\quad \| \qquad\qquad\qquad \| \qquad\quad \| \\
= \; & |F_{n-1}^0| + |F_{n-1}^1| + \cdots + |F_{n-1}^{k-2}| + |F_{n-1}^{k-1}|
\end{aligned}$$

∎

TABLE 1 ($|F_n^k|$)

n \ k	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0
2	1	2	1	0	0	0	0	0	0	0	0
3	1	4	2	1	0	0	0	0	0	0	0
4	1	7	5	2	1	0	0	0	0	0	0
5	1	12	11	5	2	1	0	0	0	0	0
6	1	20	23	12	5	2	1	0	0	0	0
7	1	33	47	27	12	5	2	1	0	0	0
8	1	54	94	59	28	12	5	2	1	0	0
9	1	88	185	127	63	28	12	5	2	1	0
10	1	143	360	269	139	64	28	12	5	2	1
11	1	232	694	563	303	143	64	28	12	5	2
12	1	376	1328	1167	653	315	144	64	28	12	5
13	1	609	2526	2400	1394	687	319	144	64	28	12
14	1	986	4781	4903	2953	1485	699	320	144	64	28
15	1	1596	9012	9960	6215	3186	1519	703	320	144	64
16	1	2583	16929	20135	13008	6792	3277	1531	704	320	144
17	1	4180	31709	40534	27095	14401	7026	3311	1535	704	320
18	1	6764	59247	81300	56201	30391	14984	7117	3323	1536	704
19	1	10945	110469	162538	116143	63872	31808	15218	7151	3327	1536
20	1	17710	205606	324020	239231	133751	67249	32392	15309	7163	3328

References

1. L. Cummings, *Connectivity of synchronizable codes in the n-cube*, JCMCC **3** (to appear).
2. S. Golomb, *A Mathematical Theory of Discrete Classification*, in "Proceedings of the Fourth London Symposium on Information Theory", Butterworths, London, 1961.
3. S. Golomb and B. Gordon, *Codes with bounded synchronization delay*, Info. and Control **8** (1965), 355-372.
4. M. Lothaire, "Combinatorics on Words", Addison-Wesley, Reading, Mass., 1983.
5. J. Riordan, "An Introduction to Combinatorial Analysis", Wiley, New York, 1958.