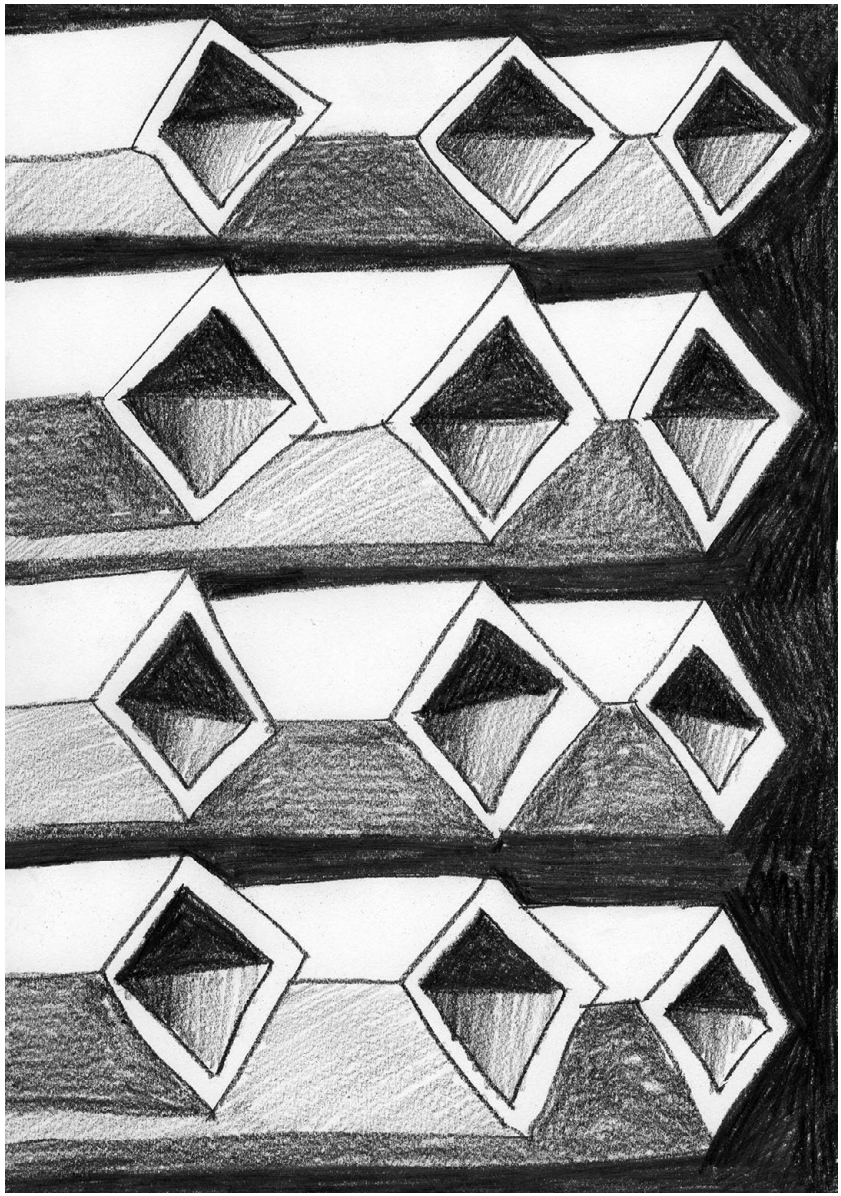
2 Combinatorial Puzzles



https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

1 Stringologic Proof of Fermat's Little Theorem

In 1640 the great French number theorist Pierre de Fermat proved the following property:

If p is a prime number and k is any natural number
then p divides $k^p - k$.

The statement is known as *Fermat's little theorem*. For example:

7 divides $2^7 - 2$ and 101 divides $10^{101} - 10$.

Question. Prove Fermat's little theorem using only stringologic arguments.

[**Hint:** Count conjugacy classes of words of length p.]

Solution

To prove the property we consider conjugacy classes of words of the same length. For example, the conjugacy class containing `aaaba` is the set $C(\texttt{aaaba}) = \{\texttt{aaaab}, \texttt{aaaba}, \texttt{aabaa}, \texttt{abaaa}, \texttt{baaaa}\}$. The next fact is a consequence of the Primitivity Lemma.

Observation. The conjugacy class of a primitive word w contains exactly $|w|$ distinct words.

Let us consider the set of words of length p, a prime number, over the alphabet $\{1, 2, \ldots, k\}$ and let $S_k(p)$ be its subset of primitive words. Among the k^p words only k of them are not primitive, namely words of the form a^p for a letter a. Thus we arrive at the following observation.

Observation. The number $|S_k(p)|$ of primitive words of length p, a prime number, on a k-letter alphabet is $k^p - k$.

Since words in $S_k(p)$ are primitive, the conjugacy class of each of them is of size p. Conjugacy classes partition $S_k(p)$ into sets of size p, which implies that p divides $k^p - k$ and that there are $(k^p - k)/p$ classes. This proves the theorem.

Notes

When a word $w = u^q$ of length n on a k-letter alphabet has a primitive root u of length d, we have $n = qd$ and the conjugacy class of w contains d elements. Running d over the divisors of n we get the equality $k^n = \Sigma\{d\psi_k(d) : d \text{ divisor of } n\}$, where $\psi_k(m)$ denotes the number of classes of primitive words of length m. It proves the theorem when n is prime. Further details are in the book by Lothaire [175, chapter 1].

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

2 Simple Case of Codicity Testing

A set $\{w_1, w_2, \ldots, w_n\}$ of words drawn from an alphabet A is a (uniquely decipherable) code if for every two sequences (noted as words) $i_1 i_2 \cdots i_k$ and $j_1 j_2 \cdots j_\ell$ of indices from $\{1, 2, \ldots, n\}$ we have

$$i_1 i_2 \cdots i_k \neq j_1 j_2 \cdots j_\ell \;\Rightarrow\; w_{i_1} w_{i_2} \cdots w_{i_k} \neq w_{j_1} w_{j_2} \cdots w_{j_\ell}.$$

In other words, if we define the morphism h from $\{1, 2, \ldots, n\}^*$ to A^* by $h(i) = w_i$, for $i \in \{1, 2, \ldots, n\}$, the condition means that the morphism is injective.

For an arbitrary integer n there is no known linear-time algorithm for testing the codicity property. However, the situation is extremely simple for $n = 2$: it is enough to check if the two codewords commute, that is, if $w_1 w_2 = w_2 w_1$.

Question. Show that $\{x, y\}$ is a code if and only if $xy \neq yx$.

Solution

A proof idea is given on page 5 as a consequence of the Periodicity Lemma. Below is a self-contained inductive proof.

If $\{x, y\}$ is a code, the conclusion follows by definition. Conversely, let us assume $\{x, y\}$ is not a code and prove the equality $xy = yx$. The equality holds if one of the words is empty, so we are left to consider the two words are not empty.

The proof is by induction on the length of $|xy|$. The induction base is the simple case $x = y$, for which the equality obviously holds.

Assume that $x \neq y$. Then one of the words is a proper prefix of the other and assume w.l.o.g. that x is a proper prefix of y: $y = xz$ for a non-empty word z. Then $\{x, z\}$ is not a code because the two distinct concatenations of x's and y's producing the same word translate into two distinct concatenations of x's and z's producing the word.

The inductive hypothesis applies because $|xz| < |xy|$ and yields $xz = zx$. Consequently $xy = xxz = xzx = yx$, which shows that the equality holds for x and y, and achieves the proof.

Notes

The same type of proof shows that $\{x, y\}$ is not a code if $x^k = y^\ell$ for two positive integers k and ℓ.

We do not know if there is a special codicity test for three words in terms of a fixed set of inequalities. For a finite number of words, an efficient polynomial-time algorithm using a graph-theoretical approach is given in Problem 52.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

3 Magic Squares and the Thue–Morse Word

The goal of the problem is to build magic squares with the help of the infinite Thue–Morse word $\mathbf{t}$ on the binary alphabet $\{0,1\}$ (instead of $\{a,b\}$). The word $\mathbf{t}$ is $\mu^\infty(0)$ obtained by iterating the morphism μ defined by $\mu(0) = 01$ and $\mu(1) = 10$:

$$\mathbf{t} = 01101001100101101001\cdots.$$

The $n \times n$ array S_n, where $n = 2^m$ for a positive natural number m is defined, for $0 \le i, j < n$, by

$$S_n[i,j] = \mathbf{t}[k](k+1) + (1 - \mathbf{t}[k])(n^2 - k),$$

where $k = i.n + j$. The generated array S_4 is

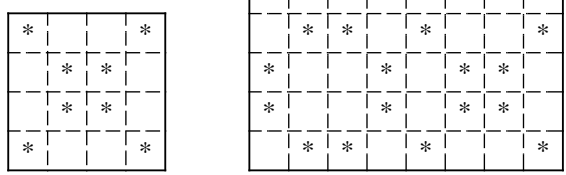
16	2	3	**13**
5	**11**	**10**	8
9	**7**	**6**	12
4	14	15	**1**

The array is a magic square because it contains all the integers from 1 to 16 and the sum of elements on each row is 34, as well as the sums on each column and on each diagonal.

Question. Show the $n \times n$ array S_n is a magic square for any natural number n power of 2.

Solution

To understand the structure of the array S_n let T_n be the Thue–Morse 2-dimensional word of shape $n \times n$, where $n = 2^m$, defined, for $0 \le i, j < n$, by $T_n[i,j] = \mathbf{t}[i.n + j]$. The picture displays T_4 and T_8, where $*$ substitutes for 0 and space substitutes for 1.



https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Notice that the table T_n satisfies two simple properties:

(i) Every row and every column is made up of blocks `0110` and `1001`.

(ii) Each of the two main diagonals is homogeneous, consisting only of `0`'s or only of `1`'s (respectively stars and spaces on the picture).

It is clear from the definition that the $n \times n$ matrix S_n is filled with all the integers from 1 to n^2. To prove it is a magic square we have to show that the sum of all entries in any row, in any column or in any of the two diagonals is the same, that is, $\frac{n}{2}(n^2+1)$.

Correctness for rows. According to property (i) each block in a row is of type `0110` or type `1001`. Consider a block `0110` whose first element is the kth element in the array. Then

$$S[k,k+1,k+2,k+3] = [n^2-k,\, k+2,\, k+3,\, n^2-k-3],$$

which sums to $2n^2+2$. For a block whose type is different from `0110` we get $[k+1, n^2-k-1, n^2-k-2, k+4]$, whose sum is the same value. Since we have $n/4$ such blocks in a row, the sum of all their contributions is

$$\frac{n}{4} \cdot (2n^2+2) = \frac{n}{2}(n^2+1),$$

as required.

The correctness for columns can be shown similarly.

Correctness for diagonals. Let us consider only the diagonal from $(0,0)$ to $(n-1, n-1)$ since the other diagonal can be treated similarly. Entries on the diagonal are $1, 1+(n+1), 1+2(n+1), \ldots, 1+(n-1)(n+1)$, listed bottom-up. Their sum is

$$n + (n+1)\sum_{i=0}^{n-1} i = n + (n+1)\frac{n}{2}(n-1) = \frac{n}{2}(n^2+1),$$

as required.

This achieves the proof that S_n is a magic square.

Notes

More on magic squares and their long history may be found on Wikipedia: https://en.wikipedia.org/wiki/Magic_square.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

4 Oldenburger–Kolakoski Sequence

The Oldenburger–Kolakoski sequence is an autodescriptive and self-generating infinite sequence of symbols $\{1,2\}$. More technically, it is its own run-length encoding. The sequence, denoted here by $\mathbf{K}$, is one of the strangest sequences. Despite the simplicity of its generation it appears to have a random behaviour.

By a block of letters in a word we mean a run of letters, that is, a maximal factor consisting of occurrences of the same letter. The operation $blocks(S)$ replaces each block of a word S by its length. For example,

$$blocks(2\,1\,1\,1\,2\,2\,1\,2\,2\,2) = 1\,3\,2\,1\,3.$$

The sequence $\mathbf{K}$ is the unique infinite sequence over the alphabet $\{1,2\}$ that starts with 2 and satisfies $blocks(\mathbf{K}) = \mathbf{K}$.

Remark. Usually the sequence is defined to start with 1, but it is more convenient here that it starts with 2. In fact, these are the same sequences after removing the first occurrence of 1.

Question. Show that we can generate online the first n symbols of the sequence $\mathbf{K}$ in $O(n)$ time and $O(\log n)$ space.

[**Hint:** Produce $\mathbf{K}$ by iterating $h = blocks^{-1}$ from 2.]

The very small space used for the generation of $\mathbf{K}$ is the most interesting element of the question.

Solution

As h is defined, $h(x) = y$ if and only if y starts with 2 and $blocks(y) = x$.

How to generate $h^{k+1}(2)$ from $h^k(2)$. Let $x = h^k(2)$. Then $y = h^{k+1}(2) = h(x)$ results by replacing the letter $x[i]$ of x either by $x[i]$ occurrences of letter 2 if i is even or by $x[i]$ occurrences of letter 1 if i is odd. The word $\mathbf{K}$ is the limit of $\mathbf{K}_k = h^k(2)$ when k goes to infinity. The first iterations of h give

$$\begin{cases} h(2) = 22 \\ h^2(2) = 22\,11 \\ h^3(2) = 22\,11\,2\,1 \\ h^4(2) = 22\,11\,2\,1\,22\,1 \end{cases}$$

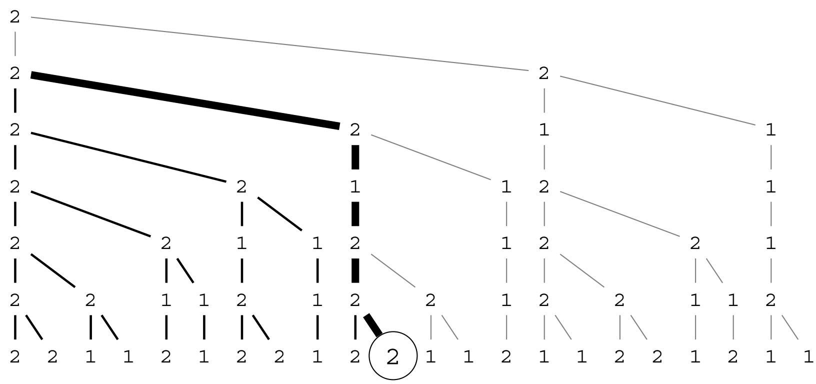
We leave for the reader the following technical fact.

Observation. $n = O(\log |\mathbf{K}_n|)$ and $\sum_{k=0}^{n} |\mathbf{K}_k| = O(|\mathbf{K}_n|)$.

Let T be the parsing tree associated with $\mathbf{K}_n$. Its leaves correspond to positions on $\mathbf{K}_n$. For a position i, $0 \le i < |\mathbf{K}_n|$, $RightBranch(i)$ denotes

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

the path from the ith leaf upwards to the first node on the leftmost branch of the tree (see picture).



The figure illustrates the parsing tree of $\mathbf{K}_6 = h^6(2)$. Each level represents $h^k(2)$ for $k = 0, 1, \ldots, 6$. The RightBranch of position 10 (circled leaf) consists of the thick edges and their endpoints. It starts from the leaf and goes up to finish at the first node on the leftmost branch.
To every node on the RightBranch is attached one bit of information: the parity of the numbers of nodes to the left on its level.

If for each node we know its label and whether it is a left child, then from *RightBranch*(i) the symbol at position $(i+1)$ as well as the whole *RightBranch*$(i+1)$ are computed in logarithmic space and amortised constant time due to the observation (since lengths of paths are logarithmic and the size of the whole tree is linear). The process works as follows on a suffix of the RightBranch. It goes up the tree to find the first left child, then goes down to the right from its parent and continues until it reaches the next leaf. Basically it goes up to the lowest common ancestor of leaves i and $i+1$ and in a certain sense each iteration can be seen as an in-order traversal of the parsing tree using small memory.

The RightBranch may grow upwards, as happens when changing *RightBranch*(13) to *RightBranch*(14) in the example. This is a top-level description of the algorithm and technical details are omitted.

Notes

The Oldenburger–Kolakoski sequence, often referred to as just the Kolakoski sequence, was designed by Oldenburger [197] and later popularised by Kolakoski [166]. The sequence is an example of a smooth word, see [46]. Our sketch of the algorithm is a version of the algorithm by Nilsson [195]; see also https://en.wikipedia.org/wiki/Kolakoski_sequence.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

5 Square-Free Game

A non-trivial square is a word over an alphabet A of the form uu, where $|u| > 1$, and it is an odd-square if in addition $|u|$ is an odd number.

The *square-free game* of length n over A is played between two players, Ann and Ben. The players extend an initially empty word w by alternately appending letters to the word. The game ends when the length of the emerging word is n or a non-trivial square has been created earlier. We assume that Ben makes the first move and that n is even. Ann wins if there are no non-trivial squares in the final word. Otherwise, Ben is the winner.

Odd square-free game. In this limited game Ann wins if no odd-square occurs. On the alphabet $A = \{0,1,2\}$ we describe Ann's winning strategy as follows. Ann never makes the same move as Ben's last move, and if Ben repeats Ann's last move then she does not repeat his previous move.

To do so, Ann remembers the pair (b,a), where a is the letter appended during her previous move and b is that from Ben's previous move. In other terms, the word w is of even length and after the first move is of the form $w = vba$. Then Ben adds c and Ann responds by adding d to get $w = vbacd$, where

$$d = \begin{cases} a & \text{if } c \neq a, \\ 3 - b - a & \text{otherwise.} \end{cases}$$

Ann behaves like a finite deterministic automaton whose output has six states. A possible sequence of moves starting with 1 2, potentially winning for Ann, is

1 2 1 1 2 2 2 0 1 0 0 2 1 2 2 0.

Question. **(A)** Show that Ann always wins against Ben in the odd square-free game of any even length n.
(B) Describe a winning strategy for Ann in the square-free game over an alphabet of size 9.

[**Hint:** To prove (A) show w contains no odd-square. For point (B) mix a simple even-square strategy with the former strategy.]

Solution

Point (A). We show point (A) by contradiction that Ann's strategy is winning and assume the word w (history of the game) contains an odd-square uu ($|u| > 1$).

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Case 1. The first letter of uu is from a move by Ben.
The square is of the form

$$uu = b_0a_1b_1a_2b_2\cdots a_kb_k\ a'_0b'_1a'_1b'_2a'_2\cdots b'_ka'_k,$$

where the letters b_i and b'_j correspond to Ben's moves and the others to Ann's moves.

Since uu is a square we get $b_0 = a'_0, a_1 = b'_1, \ldots, b_k = a'_k$. Due to Ann's strategy we have $a_1 \neq b_0, a_2 \neq b_1$, etc.; that is, each two adjacent letters in uu are distinct. In particular, this implies that Ben never repeats the last move of Ann in uu.

Consequently all moves of Ann are the same; that is, all letters a_i, a'_j are the same. Hence $a_k = a'_k$ but at the same time $a'_k = b_k$ since uu is a square. This implies $b_k = a_k$ and that Ben repeats the last move of Ann, a contradiction. This completes the proof for this case.

Case 2. The first letter of uu is from a move by Ann.
The square is of the form

$$uu = a_0b_1a_1b_2a_2\cdots b_ka_k\ b'_0a'_1b'_1a'_2b'_2\cdots a'_kb'_k,$$

where as before the letters b_i, b'_j correspond to Ben's moves and the others to Ann's moves.

Similarly to the previous case we can prove that Ben always makes a move different from the last move of Ann, except that it can happen that $a_k = b'_0$. If so, $a'_1 \neq b_k$, since $a'_1 = 3 - a_k - b_k$, and later $a'_1 = a'_2 = \cdots = a'_k$. Consequently $a'_k \neq b_k$ but at the same time $a'_k = b_k$, since uu is a square, a contradiction.

If $a_k \neq b'_0$ all moves of Ben are different from those of Ann, who consequently always does the same move in uu. This leads to a contradiction in the same way as in case 1.

This completes the proof of this case and shows that Ann's strategy is winning.

Point (B). If the game concerns non-trivial even squares on the alphabet $\{0,1,2\}$ a winning strategy for Ann is extremely simple: in her kth move she adds the kth letter of any (initially fixed) square-free word over the same alphabet.

Combining in a simple way strategies (using them simultaneously) for non-trivial odd and even square-free games, Ann gets a winning strategy avoiding general non-trivial squares on a 9-letter alphabet. The alphabet now consists of pairs (e,e') of letters in $\{0,1,2\}$. The history of the game is a word of the form

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

$w = (e_1, e'_1)(e_2, e'_2)\cdots(e_k, e'_k)$ for which $e_1e_2\cdots e_k$ contains no odd-square and $e'_1e'_2\cdots e'_k$ contains no non-trivial even square.

Notes

The solution of the game presented in the problem is described in [132], where the number of letters was additionally decreased to 7 using more complicated arguments. However, a flaw was discovered by Kosinski et al.; see [169], where the number of letters is reduced just to 8.

6 Fibonacci Words and Fibonacci Numeration System

Let $r(m)$ denote the Fibonacci representation of a non-negative integer m. It is a word x of length ℓ on the alphabet $\{0,1\}$ ending with 1 except for $m = 0$, containing no two consecutive occurrences of 1 and that satisfies $m = \sum_{i=0}^{\ell-1} x[i] \cdot F_{i+2}$, where F_{i+2} is the $(i+2)$th Fibonacci number (recall that $F_0 = 0$, $F_1 = 1$, $F_2 = 1$, $F_3 = 2$, etc.).

For example: $r(0) = 0$, $r(1) = 1$, $r(2) = 01$, $r(3) = 001$, $r(4) = 101$, $r(5) = 0001$, $r(6) = 1001$, $r(7) = 0101$.

Note that the usual positional Fibonacci representation of an integer m is $r(m)^{\mathrm{R}}$, the reverse of $r(m)$. Also note that Fibonacci coding used to encode an integer m in a data stream is $r(m)1$, terminating with 11 to allow its decoding.

Question. Show that the sequence of first digits of Fibonacci representations of natural numbers in increasing order is the infinite Fibonacci word when letters are identified to digits: a to 0, b to 1.

Let $pos(k, c)$, $k > 0$, denote the position of the kth occurrence of letter c in the infinite Fibonacci word **f**.

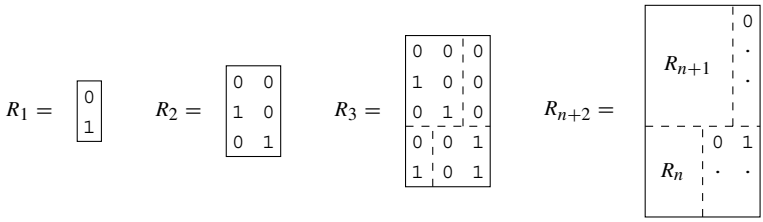
Question. Show how to compute the position of the kth occurrence of letter a in the Fibonacci word **f** in time $O(\log k)$. The same applies for the letter b.

[**Hint:** Show the following formulas: $r(pos(k, \mathtt{a})) = 0 \cdot r(k-1)$ and $r(pos(k, \mathtt{b})) = 10 \cdot r(k-1)$.]

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Solution

To understand the structure of Fibonacci representations, let us consider the rectangle R_n whose rows are representations of the first $|fib_n| = F_{n+2}$ natural numbers. Representations are possibly right padded with `0`'s to get n digits. The rectangles are given by the recurrence shown in the picture below.



Answer to the first question. Rows of rectangles R_1 and R_2 are representations of first $|fib_1|$ and $|fib_2|$ integers in increasing order respectively. Let us show by recurrence it holds for R_{n+2}, $n > 0$. Indeed, the first $|fib_{n+1}|$ rows of R_{n+2} are representations padded with `0` of the first $|fib_{n+1}|$ integers by the recurrence hypothesis. The next $|fib_n|$ rows are representations of the form $x \cdot \mathtt{01}$ (they cannot end with `11`). Since x is a row of R_n and using again the recurrence hypothesis, the next rows represent the next $|fib_n|$ integers, which shows that R_{n+2} satisfies the property and ends the recurrence.

It is clear from the recurrence that the sequence of first digits (the first column at the limit) corresponds to the infinite Fibonacci word. This answers the first question.

Answer to the second question. The limit of tables R_n is the infinite table R_∞ of Fibonacci representations of all consecutive natural numbers in increasing order. In each row, letters to the right of the rightmost occurrence of `1` are non-significant digits equal to zero.

Zeros in the first column of R_∞ correspond to `a`'s in the Fibonacci word. Rows starting with `0`'s are of the form

$$0 \cdot x_0,\ 0 \cdot x_1,\ 0 \cdot x_2,\ \ldots,$$

where

$$x_0,\ x_1,\ x_2,\ \ldots$$

is the sequence of representations of consecutive natural numbers.

Hence the kth zero corresponds to x_{k-1} and occurs at position $0 \cdot x_{k-1}$, which gives $r(pos(k, \mathtt{a})) = \mathtt{0} \cdot r(k-1)$.

Similarly we get $r(pos(k, \mathtt{b})) = \mathtt{10} \cdot r(k-1)$, since all rows containing `1` in the first column of R_∞ start in fact with `10`.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Therefore, computing the kth occurrence of a letter in the Fibonacci word amounts to computing the Fibonacci representation of an integer and doing the inverse operation, both taking $O(\log k)$ time as expected.

positions of the

5th occurrence of a:
$(0 \cdot 101)_{\mathrm{F}} = 7$

4th occurrence of b:
$(10 \cdot 001)_{\mathrm{F}} = 9$

0	a	0	0	0	0	0	·
1	b	1	0	0	0	0	·
2	a	0	1	0	0	0	·
3	a	0	0	1	0	0	·
4	b	1	0	1	0	0	·
5	a	0	0	0	1	0	·
6	b	1	0	0	1	0	·
7	a	0	1	0	1	0	·
8	a	0	0	0	0	1	·
9	b	1	0	0	0	1	·
10	a	0	1	0	0	1	·
11	a	0	0	1	0	1	·
12	b	1	0	1	0	1	·
·	·	·	·	·	·	·	·

Notes

The problem material is by Rytter [216].

7 Wythoff's Game and Fibonacci Word

Wythoff's game, a variant of the game of Nim, is a two-player game of strategy. It is played with two piles of tokens, one being initially non-empty. Players take turns removing either a positive number of tokens from one pile or the same number of tokens from both piles. When there are no tokens left, the game ends and the last player is the winner.

A configuration of the game is described by a pair of natural numbers (m,n), $m \leq n$, where m and n are the number of tokens on the two piles. Note that $(0,n)$ as well as (n,n), $n > 0$ are winning configurations. The smallest losing configuration is $(1,2)$ and all configurations of the form $(m+1,m+2)$, $(1,m)$ and $(2,m)$ for $m > 0$ are winning configurations.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

It is known that losing configurations follow a regular pattern determined by the golden ratio. Thus we pose the following question.

Question. Is there any close relation between Wythoff's game and the infinite Fibonacci word?

Solution

Losing configurations in Wythoff's game are closely related to the Fibonacci word. Let *WytLost* denote the set of losing configurations. It contains pairs of the form (m,n), $0 < m < n$:

$$WytLost = \{(1,2),(3,5),(4,7),(6,10),(8,13),\ldots\}.$$

Denoting by (m_k,n_k) the kth lexicographically smallest pair of the set we get

$$WytLost = \{(m_1,n_1),(m_2,n_2),(m_3,n_3),\ldots\},$$

with $m_1 < m_2 < m_3 < \cdots$ and $n_1 < n_2 < n_3 < \cdots$

Let $pos(k,c)$, $k > 0$, denote the position of the kth occurrence of the letter c in the infinite Fibonacci word **f**. The following property relating **f** to Wythoff's game is stated as follows.

Fact 1. $m_k = pos(k,\mathtt{a}) + 1$ and $n_k = pos(k,\mathtt{b}) + 1$.

Let $M = \{m_1,m_2,m_3,\ldots\}$ and $N = \{n_1,n_2,n_3,\ldots\}$. The following fact is well known and not proved here.

Fact 2.

(i) $M \cap N = \emptyset$ and $M \cup N = \{1,2,3,\ldots\}$.

(ii) $n_k = m_k + k$ for every $k > 0$.

Fact 2 is used to derive Fact 1. It is enough to prove that both properties (i) and (ii) hold for the sets $M' = \{pos(k,\mathtt{a}) + 1 : k > 0\}$ and $N' = \{pos(k,\mathtt{b}) + 1 : k > 0\}$.

Property (i) obviously holds and property (ii) follows from the hint presented and proved in Problem 6:

$$r(pos(k,\mathtt{a})) = \mathtt{0} \cdot r(k-1) \text{ and } r(pos(k,\mathtt{b})) = \mathtt{10} \cdot r(k-1),$$

where $r(i)$ stands for the Fibonacci representation of the natural number i. To show that $pos(k,\mathtt{b}) + 1 - pos(k,\mathtt{a}) + 1 = k$ it is sufficient to prove that for any Fibonacci representation x of a positive integer we have $(\mathtt{10}x)_{\mathrm{F}} - (\mathtt{0}x)_{\mathrm{F}} = (x)_{\mathrm{F}} + 1$, where $(y)_{\mathrm{F}}$ denotes the number i for which $r(i) = y$. But this follows directly from the definition of the Fibonacci representation and achieves the proof.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Notes

The game was introduced by Wythoff [240] as a modification of the game of Nim. He discovered the relation between losing configurations and the golden ratio; see https://en.wikipedia.org/wiki/Wythoff's_game. Specifically, the kth losing configuration (m_k, n_k), $k > 0$, is given by $m_k = \lfloor k\Phi \rfloor$ and $n_k = \lfloor k\Phi^2 \rfloor = m_k + k$. He also showed that sequences of m_k's and of n_k's are complementary; that is, each positive integer appears exactly once in either sequence.

Another consequence of the above properties is a surprising algorithm that generates the infinite Fibonacci word (or prefixes of it as long as required). To do so, assume we start with the infinite word $Fib = \sqcup^\infty$ and apply the following instruction.

```
1  for k ← 1 to ∞ do
2      i ← smallest position on Fib of ⊔
3      Fib[i] ← a
4      Fib[i + k] ← b
```

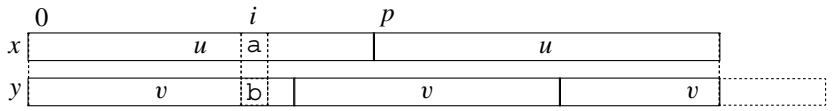
Then properties (i) and (ii) imply *Fib* becomes the Fibonacci word.

8 Distinct Periodic Words

In this problem we examine how much different two periodic words of the same length can be. The difference is measured with the Hamming distance. The Hamming distance between x and y of the same length is $\mathrm{HAM}(x, y) = |\{j : x[j] \neq y[j]\}|$.

We consider a word x whose period is p, a word y of length $|x|$ whose period q satisfies $q \leq p$ and we assume there is at least a mismatch between them. Let i be the position on x and on y of a mismatch, say, $x[i] = \mathtt{a}$ and $y[i] = \mathtt{b}$. On the picture $x = u^2$, $|u| = p$, and $|v| = q$.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press



Example. Let $x = (\texttt{abaababa})^2$ of period 8 and $y = (\texttt{abaaa})^3\texttt{a}$ of period 5. The words are distinct and have more than one mismatch. They are at positions 4, 9, 11, 12 and 14.

	0	1	2	3	**4**	5	6	7	8	**9**	10	**11**	**12**	13	**14**	15
x	a	b	a	a	b	a	b	a	a	b	a	a	b	a	b	a
y	a	b	a	a	a	a	b	a	a	a	a	b	a	a	a	a

Question. What is the minimal Hamming distance between two distinct periodic words of the same length?

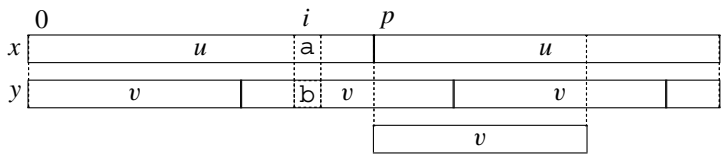
[**Hint:** Consider different cases of position i according to periods p and q.]

Solution

Since x is periodic its length is at least $2p$. W.l.o.g. it can be assumed that $x = x[0\,..\,2p-1] = u^2$. By symmetry we can also consider the mismatch position i satisfies $0 \le i < p$. Let $v = y[0\,..\,q-1]$ be the prefix period of y. Note that u and v are primitive words.

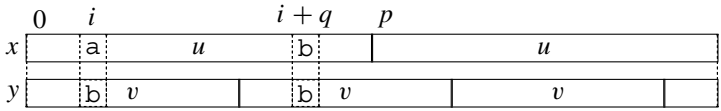
For example, aa and bb of period 1 have exactly two mismatches, as well as bbcabcbbcabc and abcabcabcabc of respective periods 6 and 3. In fact, if p is a multiple of q, that is, $p = hq$ for a positive integer h, it is clear that there is another mismatch at position $i + p$. Then $\mathrm{HAM}(x, y) \ge 2$.

If p is not a multiple of q, we prove the same inequality by contradiction, then assume the two words x and y match except at position i on them. Let us consider three cases illustrated by the three pictures that follow.

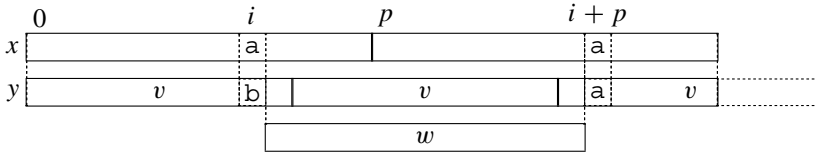


Case $i \ge q$. The word v as a prefix of u occurs at position p on both x and y. It is then an internal factor of v^2, which contradicts its primitivity by the Primitivity Lemma.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press



Case $i < q$ **and** $i + q < p$**.** Since $y[i] = y[i + q] = x[i + q]$, we get $x[i] \neq x[i + q]$. Then q is not a period of u though its occurrence at position p has period q, a contradiction.



Case $i < q$ **and** $i + q \geq p$**.** Let us first show that $w = y[i + 1 \,..\, i + p - 1]$ has period $p - q$. Indeed, for a position j, if $i < j < p$ we have

$$y[j] = x[j] = x[j + p] = y[j + p] = y[j + p - q]$$

and if $p \leq j < i + q$, we get

$$y[j] = y[j - q] = x[j - q] = x[j + p - q].$$

Then, w of length $p - 1$ has period $p - q$ in addition to period q as a factor of y longer than v. The Periodicity Lemma implies that $\gcd(q, p - q)$ is also a period, which contradicts the primitivity of v because $p - q < q$.
To conclude, when p is not a multiple of q, we have $\mathrm{HAM}(x, y) \geq 2$ as before, which achieves the whole proof.

Notes

A different proof of the result is by Amir et al. [12], and more developments can be found in [9].

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

9 A Relative of the Thue–Morse Word

Let $\mathbf{c} = (c_0, c_1, c_2, \ldots)$ be the least increasing sequence of positive integers starting with 1 and satisfying the condition

$$(*) \qquad n \in \mathbf{C} \Leftrightarrow n/2 \notin \mathbf{C},$$

where $\mathbf{C}$ is the set of elements in the sequence $\mathbf{c}$. The first elements of the sequence $\mathbf{c}$ are

$$1, 3, 4, 5, 7, 9, 11, 12, 13, 15, 16, 17, 19, 20, 21, 23, 25, 27, 28, 29, \ldots$$

Observe both that all odd integers are in the sequence and that gaps between two consecutive elements are either 1 or 2.

Question. What is the relation between the sequence $\mathbf{c}$ and the infinite Thue–Morse word $\mathbf{t}$?

Solution

Recall the Thue–Morse word $\mathbf{t}$ is $\mu^\infty(\mathtt{a})$, where the morphism μ is defined from $\{\mathtt{a}, \mathtt{b}\}$ to itself by $\mu(\mathtt{a}) = \mathtt{ab}$ and $\mu(\mathtt{b}) = \mathtt{ba}$. Let *end-pos*$(x, y)$ denote the set of ending positions of occurrences of a word x inside a word y.

Key property. For a positive integer n,

$$(**) \qquad n \notin \mathbf{C} \Leftrightarrow n \in \textit{end-pos}(\mathtt{aa}, \tau) \cup \textit{end-pos}(\mathtt{bb}, \tau).$$

The table below shows a prefix of $\mathbf{t}$ and a few first elements of $\mathbf{C}$ associated with it (even values in bold) to illustrate the property.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
t	a	b	b	a	b	a	a	b	b	a	a	b	a	b	b	a	b	a
c		1		3	**4**	5		7		9		11	**12**	13		15	**16**	17

From its definition, the word $\mathbf{t}$ satisfies, for $k > 0$:

(i) $\mathbf{t}[n] = \overline{\mathbf{t}[k]}$ and $\mathbf{t}[n-1] = \mathbf{t}[k]$ if $n = 2k + 1$.

(ii) $\mathbf{t}[n] = \mathbf{t}[k]$ and $\mathbf{t}[n-1] = \overline{\mathbf{t}[k-1]}$ if $n = 2k$.

Then property (i) rules out equivalence $(**)$ for odd integers and property (ii) does it by induction for even integers, which shows the relation between $\mathbf{c}$ and $\mathbf{t}$.

Notes

Referring to the equivalent definition of the Thue–Morse word using the parity of the number of 1's in the binary representation of integers (see page 8)

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

the property '$n \in \mathbf{C} \Leftrightarrow \nu(n)$ is even,' where $\nu(n)$ denotes the length of the end-block of 0's in the binary representation of n and also characterises the sequence **c**. (Note $\nu(n) = 0$ if and only if n is odd.)

10 Thue–Morse Words and Sums of Powers

For a finite set of natural numbers I let $Sum_k(I) = \sum_{i \in I} i^k$. Given two finite sets I and J of natural numbers we consider the property $\mathbf{P}(n, I, J)$:

$$\text{for any } k, 0 < k < n,\ Sum_k(I) = Sum_k(J),$$

which we examine with regards to sets of positions on the nth Thue–Morse word τ_n of length 2^n. Namely, the sets are

$$T_{\mathtt{a}}(n) = \{i : \tau_n[i] = \mathtt{a}\} \text{ and } T_{\mathtt{b}}(n) = \{j : \tau_n[j] = \mathtt{b}\}.$$

For example, the Thue–Morse word $\tau_3 = \mathtt{abbabaab}$ provides

$$T_{\mathtt{a}}(3) = \{0, 3, 5, 6\} \text{ and } T_{\mathtt{b}}(3) = \{1, 2, 4, 7\}.$$

The property $\mathbf{P}(3, T_{\mathtt{a}}(3), T_{\mathtt{b}}(3))$ holds due to the equalities:
$0 + 3 + 5 + 6 = 1 + 2 + 4 + 7 = 14$,
$0^2 + 3^2 + 5^2 + 6^2 = 1^2 + 2^2 + 4^2 + 7^2 = 70$.

Question. Show that the property $\mathbf{P}(n, T_{\mathtt{a}}(n), T_{\mathtt{b}}(n))$ holds for any integer $n > 1$.

Solution

For a natural number d let $I + \{d\} = \{a + d : a \in I\}$. Note the following fact, whose proof is a matter of simple calculation, for any number d and sets I, J.

Observation. Assume $\mathbf{P}(n, I, J)$ holds. Then the two other properties hold as well:

$$\mathbf{P}(n, I + \{d\}, J + \{d\}) \text{ and } \mathbf{P}(n + 1, I \cup (J + \{d\}), J \cup (I + \{d\})).$$

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

The solution of the problem, that is, the proof of the statement in the question, reduces then to a simple induction on n, using the observation above and the following recurrence on $n > 1$:

$T_{\mathtt{a}}(n+1) = T_{\mathtt{a}}(n) \cup (T_{\mathtt{b}}(n) + 2^n)$ and $T_{\mathtt{b}}(n+1) = T_{\mathtt{b}}(n) \cup (T_{\mathtt{a}}(n) + 2^n)$.

Notes

The present problem is a particular case of the so-called Tarry–Escott problem; see [6].

11 Conjugates and Rotations of Words

Two words x and y are conjugate if there exist two words u and v for which $x = uv$ and $y = vu$. They are also called rotations or cyclic shifts of one another. For instance, the word $\mathtt{abaab} = \mathtt{aba}\cdot\mathtt{ab}$ is a conjugate of $\mathtt{ababa} = \mathtt{ab}\cdot\mathtt{aba}$. It is clear that conjugacy is an equivalence relation between words but it is not compatible with the product of words.

Below are the seven conjugates of aabaaba (left) and the three conjugates of aabaabaab (right).

```
a a b a a b a
  a b a a b a a
    b a a b a a a
      a a b a a a b
        a b a a a b a
          b a a a b a a
            a a a b a a b
```

```
a a b a a b a a b
  a b a a b a a b a
    b a a b a a b a a
```

Question. Show that two non-empty words of the same length are conjugate if and only if their (primitive) roots are conjugate.

On the above example, $\mathtt{aabaabaab} = (\mathtt{aab})^3$ and $\mathtt{baabaabaa} = (\mathtt{baa})^3$ are conjugate, like their respective roots aab and baa.

A more surprising property of conjugate words is stated in the next question.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

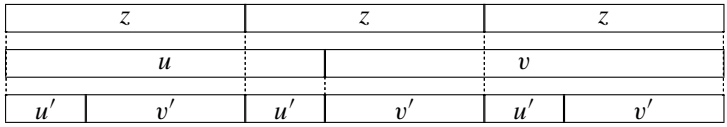
Question. Show that two non-empty words x and y are conjugate if and only if $xz = zy$ for some word z.

On the above example (left), `aabaaba` and `baabaaa` are conjugate and `aabaaba` $\cdot$ `aa` $=$ `aa` $\cdot$ `baabaaa`.

Solution

Assume words x and y of the same length have conjugate roots. Let uv be the root of x and vu be the root of y. Then $x = (uv)^k$ and $y = (vu)^k$ with $k > 0$, since they have the same length. Thus $x = u \cdot v(uv)^{k-1}$ and $y = v(uv)^{k-1} \cdot u$, which shows they are conjugate.

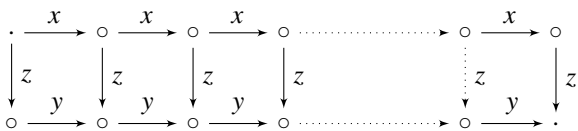
Conversely, assume x and y are conjugate and let u and v be such that $x = uv$ and $y = vu$. Let z be the root of x and $k > 0$ with $x = z^k$. Let also u' and v' be defined by $z = u'v'$, u' is a suffix of u and v' is a prefix of v.



Then, $y = vu = (v'u')^{k'}v'(u'v')^{k''}u'$, where $k' + k'' = k - 1$. This gives $y = (v'u')^k$ and shows that the root t of y satisfies $|t| \leq |u'v'| = |z|$ using Lemma 2. But since the roles of x and y are symmetric, this also proves $|z| \leq |t|$ and thus $|z| = |t|$ and $t = v'u'$. Therefore, the respective roots z and t of x and y are conjugate.

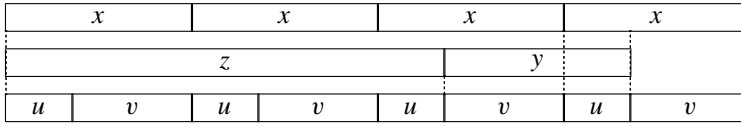
To answer the second question, let us first assume x and y are conjugate, that is $x = uv$ and $y = vu$. Then $xu = (uv)u = u(vu) = uy$, which proves the conclusion with $z = u$.

Conversely, assume $xz = zy$ for some word z. For any positive integer ℓ we get $x^\ell z = x^{\ell-1}zy = x^{\ell-2}zy^2 = \cdots = zy^\ell$. This is illustrated by the next diagram, expansion of the initial left square diagram associated with $xz = zy$, in which $\circ$ denotes the concatenation.



Considering the integer k that satisfies $(k-1)|x| \leq |z| < k|x|$, z is a proper prefix of x^k at least as long as x^{k-1} ($k = 3$ in the picture below).

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press



Then, there exist two words u and v for which $x = uv$ and $z = x^{k-1}u$. It follows that $xz = (uv)^k u = zvu$, which implies $y = vu$ from the condition $xz = zy$. Therefore x and y are conjugate.

Notes

Conjugacy of words is intimately related to their periodicity as seen on page 3. More on conjugate words may be found in Lothaire [175].

12 Conjugate Palindromes

The problem is related to the two operations on words consisting of reversing a word and taking one of its conjugate. The operations are essentially incompatible in the sense that only a few conjugates of a word are also its reverse.

To examine the situation, we consider palindromes that are conjugates of each other. For example, the words `abba` and `baab` are both palindromes and conjugate of each other. On the contrary, the word `aabaa` has no other conjugate palindrome, that is to say, its conjugacy class contains only one palindrome.

Question. What is the maximal number of palindromes in the conjugacy class of a word?

[**Hint:** Consider the primitive root of two conjugate palindromes.]

The conjugacy class of `abba`, set {`abba`, `bbaa`, `baab`, `aabb`}, contains only two palindromes. This is also the case for the word $(\texttt{abba})^3$ whose conjugacy class contains `abbaab` `baabba` and `baabba` `abbaab`, two palindromes. But the conjugacy class of $(\texttt{abba})^2$ has only one palindrome among its four conjugates.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Solution

The preceding examples suggest a conjugacy class contains no more than two palindromes. Before showing it we prove an intermediate result.

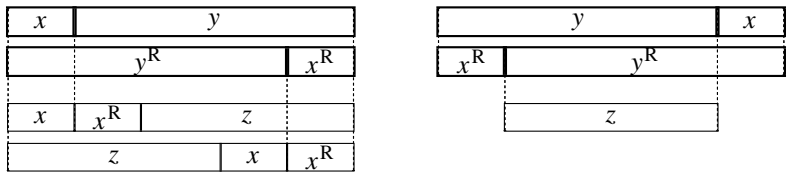
***Lemma* 4** *If $x \neq x^{\mathrm{R}}$ and $xx^{\mathrm{R}} = w^k$ for a primitive word w and a positive integer k, then k is odd and $w = uu^{\mathrm{R}}$ for some word u.*

Proof If k is even, $xx^{\mathrm{R}} = (w^{k/2})^2$ and then $x = x^{\mathrm{R}}$, a contradiction. So k is odd and then $|w|$ is even. Let $w = uv$ with $|u| = |v|$. Since u is a prefix of x and v is a suffix of x^{R}, we get $v = u^{\mathrm{R}}$, as expected. ∎

For two non-empty words x and y, assume the conjugate words xy and yx are distinct palindromes. We have both $xy = (xy)^{\mathrm{R}} = y^{\mathrm{R}}x^{\mathrm{R}}$ and $yx = (yx)^{\mathrm{R}} = x^{\mathrm{R}}y^{\mathrm{R}}$.

To prove that no more than two palindromes can be conjugate we first show that $xy = (uu^{\mathrm{R}})^k$ and $yx = (u^{\mathrm{R}}u)^k$, where k is a positive integer and u is a word for which uu^{R} is primitive. There are two cases according to x and y having the same length or not.

If $|x| = |y|$, we have $y = x^{\mathrm{R}}$, which implies $xy = xx^{\mathrm{R}}$ and $yx = x^{\mathrm{R}}x$. In addition, $x \neq x^{\mathrm{R}}$ because of the hypothesis $xy \neq yx$. Using the result of Lemma 4, the primitive root of xy is of the form uu^{R} and $xy = (uu^{\mathrm{R}})^k$ for some odd integer.



If $|x| \neq |y|$, w.l.o.g. we assume $|x| < |y|$ (see picture). Then, x is a proper border of y^{R} and x^{R} is a proper border of y, which implies that xx^{R} is a proper border of xy. The word $z = (x^{\mathrm{R}})^{-1}y$ is also a border of xy. Then the word xy has two periods $|xx^{\mathrm{R}}|$ and $|z|$ that satisfy the Periodicity Lemma condition. Thus $q = \gcd(|xx^{\mathrm{R}}|, |z|)$ is also a period of xy and divides its length. Considering the primitive root w of xy, the latter word is of the form w^k, $k > 1$, where $p = |w|$ is a divisor of q. Using Lemma 4 again, the primitive root is of the form uu^{R}, with $u \neq u^{\mathrm{R}}$ because it is primitive. Then $xy = (uu^{\mathrm{R}})^k$, where k is an odd integer.

Whether x and y have the same length or not, we get the same conclusion. To achieve the proof we just have to consider the conjugacy class of a

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

palindrome $(uu^{\mathrm{R}})^k$ where uu^{R} is primitive. Such a class contains another palindrome, namely $(u^{\mathrm{R}}u)^k$.

Since conjugates of $(u^{\mathrm{R}}u)^k$ are of the form $(st)^k$ where st is a conjugate of uu^{R}, applying the above argument again, the inequalities $u \neq u^{\mathrm{R}}$ and $s \neq u^{\mathrm{R}}$ would lead to a contradiction with the primitivity of uu^{R}. This achieves the proof that a conjugacy class contains no more than two palindromes.

Notes

The result of this problem is by Guo et al. [133] and the present proof is adapted from their article.

13 Many Words with Many Palindromes

The problem deals with the number of words containing as many palindrome factors as possible. A word w is called *palindrome rich* if it contains $|w|$ distinct non-empty palindromes as factors, including single-letter palindromes.

Example. The words `poor`, `rich` and `abac` are rich, while the words `maximal` and `abca` are not. Indeed, the set of palindromes occurring in `abac` is {`a`, `aba`, `b`, `c`}, while it is {`a`, `b`, `c`} for `abca`.

Let $Rich_k(n)$ denote the number of rich words of length n over an alphabet of size k.

Note that each position on a word is the (starting) position of the rightmost occurrence of a palindrome that it contains at most once. This is due to the fact that a second shorter palindrome sharing the position would be a proper suffix of the longer palindrome and then would occur later, a contradiction. This implies the following fact.

Observation. There are at most $|w|$ palindromes, factors of a word w.

The most interesting case to discuss is that of binary words, that is, $k = 2$, because we have

$$\begin{cases} Rich_2(n) = 2^n & \text{for } n < 8, \\ Rich_2(n) < 2^n & \text{for } n \geq 8. \end{cases}$$

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Question. Show that $Rich_2(2n)$ grows exponentially; that is, there is a positive constant c for which $Rich_2(2n) \geq 2^{cn}$.

[**Hint:** Use the fact that the number of partitions of integers grows exponentially.]

Solution

Consider all partitions of the number n into different positive integers:

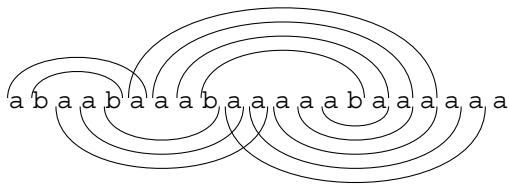
$$n = n_1 + n_2 + \cdots + n_k \text{ and } n_1 < n_2 < \cdots < n_k.$$

For each such partition $\pi = (n_1, n_2, \ldots, n_k)$ let us consider the word w_π of length $n + k - 1$ defined as follows:

$$w_\pi = \mathtt{a}^{n_1}\mathtt{b}\mathtt{a}^{n_2}\mathtt{b}\ldots\mathtt{b}\mathtt{a}^{n_k}.$$

It is fairly easy to see that the word w_π is palindrome rich.

The figure below displays non-unary palindromes occurring in the word $\mathtt{aba}^2\mathtt{ba}^3\mathtt{ba}^5\mathtt{ba}^6$ of length 21 associated with the partition of 17 $(1, 2, 3, 5, 6)$. In addition to the 14 palindromes shown in the picture, the word contains the unary palindromes a, aa, aaa, aaaa, aaaaa, aaaaaa and b for a total of 21 palindromes.



Appending $\mathtt{b}^{n-k+1}$ to w_π produces the word $v_\pi = w_\pi\mathtt{b}^{n-k+1}$ of length $2n$ that contains the additional palindromes $\mathtt{ba}^{n_k}\mathtt{b}$, $\mathtt{b}^2$, $\mathtt{b}^3$, $\ldots$, $\mathtt{b}^{n-k+1}$. Then v_π is also rich. It is known that the number of partitions of an integer n into pairwise distinct positive integers grows exponentially with n. Hence $Rich_2(2n)$ also grows exponentially with n.

Notes

The problem is based on the survey by Glen et al. [130] on the palindromic richness of words.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

14 Short Superword of Permutations

The goal of the problem is to show that a certain set of patterns can be packed into a single word in a space-economic way. This can be viewed as a compression technique for the specific set.

The present patterns called n-permutations are drawn from the alphabet of natural numbers. They are words on the alphabet $\{1, 2, \ldots, n\}$ in which every number from $\{1, 2, \ldots, n\}$ appears exactly once. The aim is to build words, called n-superwords, which contain all the n-permutations as factors.

For $n = 2$ the word `121` is a shortest 2-superword, since it contains the two 2-permutations `12` and `21`. For $n = 3$, `123121321` is a shortest 3-superword. The six 3-permutations appear in it in the order

$$\pi_1 = 123, \pi_2 = 231, \pi_3 = 312, \pi_4 = 213, \pi_5 = 132, \pi_6 = 321.$$

Note how is the structure of `123121321`: each occurrence of letter `3` is flanked by two occurrences of a 2-permutation.

The two examples of superwords are of respective lengths $\alpha_2 = 3$ and $\alpha_3 = 9$, where $\alpha_n = \sum_{i=1}^{n} i!$. But it is not clear whether a shortest n-superword is of length α_n for $n \geq 4$.

The problem consists in constructing a short n-superword, which may not be of minimal length.

Question. Show how to construct an n-superword of length α_n for each natural number n.

[**Hint:** Use this above remark on the structure of `123121321` to build an n-superword from an $(n-1)$-superword.]

Solution

The construction is done iteratively, starting with the base case $n = 2$ (or $n = 3$), as follows.

Let w_{n-1} be an $(n-1)$-superword of length α_{n-1}. The $(n-1)$-permutations are considered in their order of appearance along w_{n-1}. Let i_k be the ending position on w_{n-1} of the first occurrence of the kth $(n-1)$-permutation in w_{n-1}. This means that there are exactly $k-1$ distinct $(n-1)$-permutations with an ending position $i < i_k$ (some $(n-1)$-permutations can repeat).

The n-superword w_n is built by inserting some n-permutations in w_{n-1}. The selected n-permutations are all the words $n \cdot \pi_k$ where π_k, $1 \leq k \leq (n-1)!$, is the kth $(n-1)$-permutation occurring in w_{n-1}. All these words are inserted simultaneously immediately after their respective position i_k in w_{n-1}.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

From the definition of i_k, insertions generate factors of the form $\pi_k \cdot n \cdot \pi_k$ in w_n for each $(n-1)$-permutation π_k.

Example. Building w_4 from $w_3 = 123121321$. The ending positions on w_3 of the six 3-permutations π_i above are

$$i_1 = 2,\ i_2 = 3,\ i_3 = 4,\ i_4 = 6,\ i_5 = 7,\ i_6 = 8.$$

The insertion of the six 4-permutations of the form $4 \cdot \pi_i$ produces the following 4-superword of length $\alpha_4 = 33$:

$$123\mathbf{4}1231\mathbf{4}2312\mathbf{4}3121\mathbf{4}2132\mathbf{4}1321\mathbf{4}321,$$

in which occurrences of 4 are emphasised.

The length of the word w_n is α_n. Since there are $(n-1)!$ insertions of words of length n, the length of the resulting word w_n is $|w_{n-1}| + (n-1)!\,n = \sum_{i=1}^{n} i! = \alpha_n$, as required.

All n-permutations are factors of the word w_n. A n-permutations occurring in w_n is of the form $u \cdot n \cdot v$, where uv is a word of length $n-1$ that does not contain the letter n. This permutation occurs inside the factor $vu \cdot n \cdot vu$ of w_n, where $vu = \pi_k$ for some $(n-1)$-permutation π_k. Since by construction all words of the form $\pi_k \cdot n \cdot \pi_k$ appear in w_n all n-permutations appear in w_n. This answers the question.

Notes

It was conjectured that α_n is the minimal length of the shortest n-superwords. The conjecture was confirmed for $n = 4$ and $n = 5$ by Johnston [152] but was disproved for $n = 6$ by Houston [143].

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

15 Short Supersequence of Permutations

The problem deals with the idea of storing efficiently a set of patterns into a word. Contrary to the definition of a superword, in this problem patterns are stored as subsequences of a word called a supersequence.

The present patterns called n-permutations are drawn from the alphabet $\{1, 2, \ldots, n\}$. They are words in which every number from $\{1, 2, \ldots, n\}$ appears exactly once. The aim is to build words, called n-supersequences, that contain all n-permutations as subsequences.

For $n = 3$ the word `1213212` of length 7 is a shortest 3-supersequence. For $n = 4$ the word `123412314321` of length 12 is a shortest 4-supersequence. These two supersequences are of lengths $n^2 - 2n + 4$ (for $n = 3, 4$). Observe that for $n = 4$ our 4-supersequence has length 12 while a shortest 4-superword, obviously longer, is of length 33 (see Problem 14).

A simple way to produce an n-supersequence is to consider a word of the form π^n for any n-permutation π, or of the form $\pi_1 \pi_2 \pi_3 \ldots \pi_n$ where π_is are any n-permutations. It is clear they contain all the $n!$ n-permutations as subsequences but their length is n^2, far from optimal.

The aim of the problem is to show how to construct a moderately short n-supersequence, which may not be of minimal length.

Question. Show how to construct an n-supersequence of length $n^2 - 2n + 4$ for each natural number n.

[**Hint:** Starting from a straight n-supersequence of length n^2, show how to tune it to get the required length.]

Solution

To get the result, the n-supersequence $x = \pi_1 \pi_2 \pi_3 \ldots \pi_n$ as above is shortened in two steps.

Length $n^2 - n + 1$. The length of x is reduced by selecting permutations of the form $n \cdot \rho_i$, for $i = 1, \ldots, n-1$, where ρ_i is an $(n-1)$-permutation, and by considering the word

$$y = n \cdot \rho_1 \cdot n \cdot \rho_2 \cdot n \cdots n \cdot \rho_{n-1} \cdot n.$$

This obviously shortens the n-supersequence by $n - 1$ letters and gives the expected length.

Length $n^2 - 2n + 4$. Now the construction technique becomes slightly more tricky. The main idea is to choose more carefully the $(n-1)$-permutations ρ_i of y.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Having solutions for $n \leq 4$ we develop a solution for $n \geq 5$. To do so, let γ_1, γ_2 and γ_3 be three $(n-1)$-permutations of respective forms $3 \cdot \gamma_1' \cdot 2$, $1 \cdot \gamma_2' \cdot 3$ and $2 \cdot \gamma_3' \cdot 1$, where γ_1' is a permutation of $\{1, 2, \ldots, n-1\} \setminus \{2, 3\}$ and similarly for the other γ's.

We first concatenate in an alternative way $n-1$ blocks of type γ_i and then insert n between them, which gives successively

$$\gamma_1 \cdot \gamma_2 \cdot \gamma_3 \cdot \gamma_1 \cdot \gamma_2 \cdot \gamma_3 \cdots,$$
$$w = n \cdot \gamma_1 \cdot n \cdot \gamma_2 \cdot n \cdot \gamma_3 \cdot n \cdots n,$$
$$w = n \cdot 3 \cdot \gamma_1' \cdot 2 \cdot n \cdot 1 \cdot \gamma_2' \cdot 3 \cdot n \cdot 2 \cdot \gamma_3' \cdot 1 \cdot n \ldots n.$$

It follows from the previous case that this is an n-supersequence and that its length is $n^2 - n + 1$.

The main step of the technique eventually consists in removing $n-3$ letters in w, which gives the required length $n^2 - n + 1 - (n-3) = n^2 - 2n + 4$. This is done by removing the letter i from each γ_i' occurring in w, except from their first and last occurrences, to produce the word z.

The word z is an n-supersequence. Observe that the removal of letter i from the block γ_i', for $i = 1, 2, 3$, is compensated by the presence of i to the left and to the right of γ_i beyond the letter n. Then an argument similar to the one applied to the above word y proves that z is an n-supersequence.

It achieves the construction of an n-supersequence of length $n^2 - 2n + 4$.

Example. We illustrate the construction by the case $n = 6$. Let $\gamma_1 = 31452$, $\gamma_2 = 12453$ and $\gamma_3 = 23451$ be the selected 5-permutations. Let also γ_i^{Rem} be the word γ_i after removal of letter i.

Considering the sequence

$$w = 6 \cdot \gamma_1 \cdot 6 \cdot \gamma_2 \cdot 6 \cdot \gamma_3 \cdot 6 \cdot \gamma_1 \cdot 6 \cdot \gamma_2 \cdot 6,$$

the required 6-supersequence is obtained by removing the letter i from each block γ_i, except from the first and last blocks, which produces

$$z = 6 \cdot \gamma_1 \cdot 6 \cdot \gamma_2^{\mathrm{Rem}} \cdot 6 \cdot \gamma_3^{\mathrm{Rem}} \cdot 6 \cdot \gamma_1^{\mathrm{Rem}} \cdot 6 \cdot \gamma_2 \cdot 6;$$

that is

$$z = 6\ 31452\ 6\ 1453\ 6\ 2451\ 6\ 3452\ 6\ 12453\ 6.$$

Notes

The above method is a version of the construction by Mohanty [191]. It is known that the present construction gives a shortest supersequence of length

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

$n^2 - 2n + 4$ for $2 < n \leq 7$. However, for $n \geq 10$ the construction by Zalinescu [242] gives supersequences of length $n^2 - 2n + 3$. The exact general formula for the smallest length of n-supersequences is still unknown; it is only known so far that it is $n^2 - o(n^2)$.

16 Skolem Words

A Skolem word of order n, for a positive integer n, is a word over the alphabet $A_n = \{1, 2, \ldots, n\}$ satisfying, for each $i \in A_n$, the properties:

(i) The letter i appears exactly twice in the word,

(ii) Consecutive occurrences of i are at distance i.

Skolem words have a definition very similar to that of Langford words (Problem 17) but the small change in the distance makes a big difference.

If igi is a factor of a Skolem word, the gap word g does not contain the letter i and $|g| = i - 1$. For example, `11` is an obvious Skolem word of order 1, `23243114` a Skolem word of order 4 and `4511435232` is a Skolem word of order 5. But a mere checking shows there is no Skolem word of order 2 or of order 3.

Question. Discuss for which positive integer n there exists a Skolem word of order n and design an algorithm to build it when possible.

[**Hint:** Discuss according to n modulo 4.]

Solution

We examine different cases depending on n modulo 4.

Case $n = 4k$. The word `23243114` is an example of a Skolem word of order 4. Let $n = 4k$ for $k > 1$. The following procedure builds a Skolem word of order n.

The basic bricks of the construction are the two words w_{even} and w_{odd}. The first is made of the increasing sequence of even numbers in A_n and the second of the increasing sequence of odd numbers in $A_n \setminus \{n - 1\}$ (the largest odd number is discarded).

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

Algorithm SKOLEM produces the expected word.

SKOLEM(n multiple of 4 larger than 4)

1 $(c,d) \leftarrow (n/2-1, n-1)$
2 $w_{\text{odd}} \leftarrow 1\ 3 \cdots n-3$ ▷ no letter $n-1$ in w_{odd}
3 $\alpha \cdot c \cdot \beta \cdot 1\ 1 \cdot \beta^{\text{R}} \cdot c \cdot \alpha^{\text{R}} \leftarrow$ decomposition of $w_{\text{odd}}^{\text{R}} w_{\text{odd}}$
4 $v \leftarrow \alpha \cdot 1\ 1 \cdot \beta \cdot c \cdot d \cdot \beta^{\text{R}} \cdot \alpha^{\text{R}} \cdot c$
5 $w_{\text{even}} \leftarrow 2\ 4 \cdots n$
6 **return** $v \cdot w_{\text{even}}^{\text{R}} \cdot d \cdot w_{\text{even}}$

The instruction at line 3 factorises the two ends of the word $w_{\text{odd}}^{\text{R}} w_{\text{odd}}$ around letter c.

Example. For $n = 12$ the algorithm computes successively, from the words $w_{\text{odd}} = 1\ 3\ 5\ 7\ 9$ and $w_{\text{even}} = 2\ 4\ 6\ 8\ 10\ 12$, the decomposition of $w_{\text{odd}}^{\text{R}} w_{\text{odd}}$ with $c = 5$

$$9\ 7 \cdot 5 \cdot 3 \cdot 1\ 1 \cdot 3 \cdot 5 \cdot 7\ 9,$$

where $\alpha = 9\ 7$ and $\beta = 3$; then

$$v = 9\ 7 \cdot 1\ 1 \cdot 3 \cdot 5 \cdot 11 \cdot 3 \cdot 7\ 9 \cdot 5$$

and eventually produces the Skolem word of order 12:

$$9\ 7\ 1\ 1\ 3\ 5\ \mathbf{11}\ 3\ 7\ 9\ 5\ 12\ 10\ 8\ 6\ 4\ 2\ \mathbf{11}\ 2\ 4\ 6\ 8\ 10\ 12,$$

in which $d = 11$ is emphasised.

Why does it work? First note that property (i) is satisfied. Then it is clear that occurrences of each letter in $u = w_{\text{odd}}^{\text{R}} w_{\text{odd}}$, in v and in the suffix $w_{\text{even}}^{\text{R}} \cdot d \cdot w_{\text{even}}$ of the output are at correct distances.

So it remains to show property (ii) holds for letters c and d. Inside v the distance between the occurrences of c is $|\alpha| + |\beta| + 1$, the number of odd numbers different from 1 and c; that is, $n/2 - 2$, as required.

The distance between the two occurrences of letter d in the output is $|\alpha| + |\beta| + 1 + |w_{\text{even}}|$, that is, $|A_n \setminus \{1, d\}| = n - 2$, as required as well.

Therefore SKOLEM(n) is a Skolem word of order n.

Case $n = 4k + 1$. This case works essentially in the same way as the previous case, except that d is set to n and c is set to $\lfloor n/2 \rfloor - 1$. Let w_{even} be, as before, the increasing sequence of even numbers in A_n and let w_{odd} be the increasing sequence of odd numbers in $A_n \setminus \{n\}$ (the largest odd number is discarded).

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

With this instance of length n, Algorithm SKOLEM produces the expected word. Observe that in the first case v and the output contain the factor $c \cdot d$ while in the present case they contain the factor $d \cdot c$.

SKOLEM(n in the form $4k+1$ larger than 4)

1 $(c,d) \leftarrow (\lfloor n/2 \rfloor - 1, n)$
2 $w_{\text{odd}} \leftarrow \texttt{1 3} \cdots n-2$ ▷ no letter n in w_{odd}
3 $\alpha \cdot c \cdot \beta \cdot \texttt{1 1} \cdot \beta^{\text{R}} \cdot c \cdot \alpha^{\text{R}} \leftarrow$ decomposition of $w_{\text{odd}}^{\text{R}} w_{\text{odd}}$
4 $v \leftarrow \alpha \cdot \texttt{1 1} \cdot \beta \cdot d \cdot c \cdot \beta^{\text{R}} \cdot \alpha^{\text{R}} \cdot c$
5 $w_{\text{even}} \leftarrow \texttt{2 4} \cdots n-1$
6 **return** $v \cdot w_{\text{even}}^{\text{R}} \cdot d \cdot w_{\text{even}}$

Example. For $n = 13$ the algorithm computes successively, from the words $w_{\text{odd}} = \texttt{1 3 5 7 9 11}$ and $w_{\text{even}} = \texttt{2 4 6 8 10 12}$, the decomposition of $w_{\text{odd}}^{\text{R}} w_{\text{odd}}$ with $\lfloor n/2 \rfloor - 1 = c = \texttt{5}$:

$$\texttt{11 9 7} \cdot \mathbf{5} \cdot \texttt{ 3} \cdot \texttt{1 1} \cdot \texttt{3} \cdot \mathbf{5} \cdot \texttt{7 9 11},$$

where $\alpha = \texttt{11 9 7}$ and $\beta = \texttt{3}$; then

$$v = \texttt{11 9 7 } \mathbf{1\ 1} \texttt{ 3 } \mathbf{13\ 5} \texttt{ 3 7 9 11 } \mathbf{5}$$

and eventually produces the Skolem word of order 13:

$$v = \texttt{11 9 7 1 1 3 } \mathbf{13\ 5} \texttt{ 3 7 9 11 } \mathbf{5} \texttt{ 12 10 8 6 4 2 } \mathbf{13} \texttt{ 2 4 6 8 10 12},$$

in which $c = \texttt{5}$ and $d = \texttt{13}$ are emphasised.

Impossibility of other cases. Let $odd(n)$ be the number of odd natural numbers not exceeding n.

Observation. If there is a Skolem word of order n, we have the equality $odd(n) \bmod 2 = n \bmod 2$.

To prove the observation we consider sums modulo 2, called 2-sums, of positions on a Skolem word w of order n. First, the 2-sum of all positions on w is n mod 2. Second, let us pair positions of the same letter i to compute the sum. If i is even the (two) positions of its occurrences have the same parity, so their contribution to the 2-sum is null. But if i is odd the corresponding positions have different parities. Hence the 2-sum of positions is $odd(n) \bmod 2$. Consequently we have $odd(n) \bmod 2 = n \bmod 2$, as stated.

The impossibility of having Skolem words for $n = 4k+2$ and for $n = 4k+3$ follows directly from the observation, since $odd(n) \bmod 2 \neq n \bmod 2$ in these cases.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

To conclude, Skolem words exist only if n is of the form $4k$ or $4k+1$.

Notes

Words considered in the problem have been introduced by Skolem [227].

17 Langford Words

A Langford word of order n, for a positive integer n, is a word over the alphabet $A_n = \{1, 2, \ldots, n\}$ satisfying, for each $i \in A_n$, the properties:

(i) Letter i appears exactly twice in the word,

(ii) Consecutive occurrences of i are at distance $i+1$.

Langford words have a definition very similar to that of Skolem words (Problem 16) but the small change in the distance makes a big difference.

If igi is a factor of a Langford word the gap word g does not contain the letter i and $|g| = i$. For example, `312132` is a Langford word of order 3 and `41312432` a Langford word of order 4.

Question. Discuss for which positive integer n there exists a Langford word of order n and show how to build it when possible.

[**Hint:** Discuss according to n modulo 4.]

Solution

We examine different cases depending on n modulo 4.

Case $n = 4k+3$. The above example shows a Langford word of order 3. For $n \geq 7$, that is, $k > 0$, let $X_n = \{2k+1, n-1, n\}$ and $A'_n = A_n \setminus X_n$. Let both w_{even} be the increasing sequence of even numbers in A'_n and w_{odd} be the increasing sequence of odd numbers in A'_n.

Note that A'_n has $4k$ elements, then exactly $2k$ even letters and $2k$ odd letters. Both w_{even} and w_{odd} can be split into halves: $w_{\text{even}} = p_1 \cdot p_2$, where $|p_1| = |p_2| = k$, and $w_{\text{odd}} = p_3 \cdot p_4$, where $|p_3| = |p_4| = k$.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

To get a solution let start with the following word that is almost a Langford word:

$$u = p_2^{\mathrm{R}}\ p_3^{\mathrm{R}} * p_3 * p_2 * p_4^{\mathrm{R}}\ p_1^{\mathrm{R}} * * p_1 * p_4,$$

where $*$ stands for a missing letter to be inserted. It is clear that the distance between the two occurrences of each $i \in A'_n$ equals $i + 1$.

Now it is enough to substitute twice the remaining elements of A_n to $*$, which is done in the order

$$4k + 2,\ 4k + 3,\ 2k + 1,\ 4k + 2,\ 2k + 1,\ 4k + 3.$$

Since each p_j has length k, it is straightforward to compute the distances between copies of inserted elements from $\{2k+1, n-1, n\}$ and see they comply with property (ii), producing a Langford word of order n.

Example. Let $n = 11 = 4 \times 2 + 3$, $k = 2$. We have $X_{11} = \{5, 10, 11\}$ and $A'_{11} = \{1, 2, 3, 4, 6, 7, 8, 9\}$; then $p_1 = 2\ 4$, $p_2 = 6\ 8$, $p_3 = 1\ 3$ and $p_4 = 7\ 9$. The first step produces

$$u = 8\ 6\ 3\ 1\ *\ 1\ 3\ *\ 6\ 8\ *\ 9\ 7\ 4\ 2\ *\ *\ 2\ 4\ *\ 7\ 9,$$

which leads to the Langford word of order 11

$$8\ 6\ 3\ 1\ \mathbf{10}\ 1\ 3\ \mathbf{11}\ 6\ 8\ \mathbf{5}\ 9\ 7\ 4\ 2\ \mathbf{10}\ \mathbf{5}\ 2\ 4\ \mathbf{11}\ 7\ 9,$$

in which $2k + 1 = 5$, $n - 1 = 10$ and $n = 11$ are emphasised.

Case $n = 4k$. An example of Langford word of order 4 is shown above. Then let $n = 4k+4$, with $k > 0$. This case is similar to the previous case. A solution u associated with $4k + 3$ is first built. Then a few changes are made to insert in it the largest element n. It is done by substituting it for the first copy of $2k+1$, an element that is moved to the end of the word and becomes its second occurrence. The second copy of n is placed after it.

To say it differently, insertions inside the word u associated with $4k + 3$ are done in the order

$$4k + 2,\ 4k + 3,\ n,\ 4k + 2,\ 2k + 1,\ 4k + 3,\ 2k + 1,\ n,$$

where the last two letters are appended at the end of the word.

Distances between elements smaller than n are not changed and the distance between occurrences of the largest element $n = 4k+4$ is as required, producing a Langford word of order n.

Impossibility of other cases. Any Langford word w over A_{n-1} can be transformed into a Skolem word over A_n by adding 1 to all elements in w

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

and by inserting at the beginning the two copies of letter 1. For example, the Langford word 312132 is so transformed into the Skolem word 11423243.

It is known that Skolem words do not exist for n of the forms $4k + 2$ nor $4k + 3$ (see Problem 16). The same observation works for Langford words and proves none exist when n is of the forms $4k + 1$ or $4k + 2$.

To conclude, a Langford word exists only when n is of the form $4k + 4$ or $4k + 3$ ($k \geq 0$).

Notes

There are various notions of Langford words. For example, property (i) can be dropped. In this case Berstel [32] showed that the associated words are square free.

18 From Lyndon Words to de Bruijn Words

The combinatorial result of the problem provides the basis for an efficient online construction of de Bruijn words.

A binary word (on the alphabet {0, 1}) is a de Bruijn word of order (rank or span) k if it contains cyclically each binary word of length k exactly once as a factor. The word is of length 2^k. There is a surprising relation between these words and the lexicographic ordering, which shows once more that ordering words is a powerful tool in text algorithms.

A Lyndon word is a primitive word that is the (lexicographically) smallest word in its conjugacy equivalence class.

Let p be a prime number and $\mathcal{L}_p = (L_0, L_1, \ldots, L_m)$ the sorted sequence of binary Lyndon words of length p or 1. Let also

$$\mathbf{b}_p = L_0 \cdot L_1 \cdots L_m$$

be the concatenation of words in $\mathcal{L}_p$.

For example, the sorted list of Lyndon words of length 5 or 1 is

$$\mathcal{L}_5 = (0, 00001, 00011, 00101, 00111, 01011, 01111, 1)$$

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

and the concatenation of its words is

$$\mathbf{b}_5 = \texttt{0 00001 00011 00101 00111 01011 01111 1}.$$

It is the lexicographically smallest de Bruijn word of order 5 and has length $32 = 2^5$.

Question. For a prime number p, show that the word $\mathbf{b}_p$ is a de Bruijn word of order p.

Solution

The number of binary Lyndon words of length p in $\mathcal{L}_p$ is $(2^p - 2)/p$ (see Problem 1). Therefore the length of $\mathbf{b}_p$ is $p(2^p - 2)/p + 2 = 2^p$. Then to show it is a de Bruijn word we just have to prove that each word w of length p appears cyclically in $\mathbf{b}_p$.

Let us start with a preliminary observation. For a word x in $\mathcal{L}_p$, $x \neq \texttt{1}$, let $next(x)$ denote the word following x in the sequence.

Observation. If $|x| = |next(x)| = p$, $x = uv$ and v contains an occurrence of $\texttt{0}$, then u is a prefix of $next(x)$.

Proof Assume, to the contrary, that $next(x) = u'v'$ with $|u| = |u'| = t$ and $u' \neq u$. Then $u < u'$ due to the order of elements in $\mathcal{L}_p$. However, the word $u \cdot \texttt{1}^{n-t}$ is a Lyndon word that is lexicographically between uv and $u'v'$, which contradicts $next(x) = u'v'$. Thus u is a prefix of $next(x)$. ∎

Lyndon words of length p are all factors of $\mathbf{b}_p$ by construction. Words $\texttt{0}^p$ and $\texttt{1}^p$ are respectively prefix and suffix of $\mathbf{b}_p$. Words of length p in $\texttt{1}^+\texttt{0}^+$ occur cyclically at the last $p - 1$ positions of $\mathbf{b}_p$. Thus it remains to prove that words of length p which are not Lyndon words and do not belong to $\texttt{1}^*\texttt{0}^*$ appear (non-cyclically) in $\mathbf{b}_p$. Let w be such a word and L_i be its Lyndon conjugate; then

$$w = vu \text{ and } L_i = uv$$

for v and u non-empty words because $w \neq L_i$.

There are two cases to consider whether v contains an occurrence of $\texttt{0}$ or not.

Case v contains $\texttt{0}$. Then u is a prefix of $L_{i+1} = next(L_i)$ from the observation. Hence $w = vu$ is a factor of L_iL_{i+1}.

Case v does not contain $\texttt{0}$. Then $v = \texttt{1}^t$, for some $t > 0$. Let L_j be the first word in $\mathcal{L}_p$ prefixed by u and let $L_{j-1} = u'v'$ with $|v'| = t$. Then v' cannot

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press

contain the letter 0, because otherwise $u' = u$ and L_j would not be the first word in $\mathcal{L}_p$ prefixed by u. Consequently $v' = 1^t = v$ and the concatenation $L_{j-1}L_j = u' \cdot v \cdot u \cdots$ contains vu as a factor.

In both cases w has an occurrence in $\mathbf{b}_p$. This concludes the proof that $\mathbf{b}_p$ is a de Bruijn word.

Notes

The list $\mathcal{L}_p$ can be generated online using only $O(p)$ memory space. The above construction then leads to an online generation of a de Bruijn word, using only a window of size $O(p)$ for storing the last computed letters of the word.

When the order k of de Bruijn words is not a prime number, a similar construction applies. In that case, the sorted list $\mathcal{L}_k$ is composed of Lyndon words whose length divides k. The concatenation of these sorted words gives in fact the lexicographically smallest de Bruijn word of order k over the given alphabet. The algorithm was initially developed by Fredricksen and Maiorana [120]. See also [192] for a simplified complete proof of the general case.

https://doi.org/10.1017/9781108835831.003 Published online by Cambridge University Press