Combinatorics on Words

Gabriele Fici

CWI, Amsterdam — April 2024

Part 1: Basics

Words

An alphabet Σ is a set of symbols, whose elements are called letters.

Definition 1

Given an alphabet Σ, a word w over Σ is a finite sequence of letters from Σ. A word over an alphabet of size 2 is called a binary word.

The length $|w|$ of a word w is the number of its letters. The unique word of length 0 is called the empty word and is denoted by ε.

We let Σ^* denote the set of all words of any length over Σ and Σ^+ the set of all words of positive length over Σ, that is, $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$.

Finally, for a given $n \geq 0$, Σ^n denotes the set of all words over Σ of length n.

Words

With wz we denote the concatenation of words w and z.

With w^k we denote the concatenation of k copies of the word w.

Notice that $w^0 = \varepsilon$.

Indeed, equipped with the operation of concatenation, the set Σ^+ is a free semigroup, while the set Σ^* is a free monoid.

Remark 2

Over an alphabet of cardinality k, there are k^n possible words of length n, for every $n \geq 0$.

Orders

With Σ_k we denote the ordered alphabet $\{0, 1, \ldots, k-1\}$.

The order on Σ_k induces the lexicographic order $\leq$ on the set of words Σ_k^*, defined by $x \leq y$ if and only if x is a prefix of y or in the first position in which x and y disagree, the letter occurring in x is smaller than the letter occurring in y.

Notice that, although $\leq$ is a total order, it is not a well-order, in the sense that there exist infinite sets of words without a least element.

For example, if we start from the word 0, the next word in lexicographic order will be 00, then 000, etc. So, listing all words in lexicographic order, there are infinitely many words before we encounter the word 1.

Orders

For this reason, we will also use another order, $\leq_s$, called genealogical (or shortlex, or radix, or military) order, defined by: $x \leq_s y$ if the length of x is smaller than the length of y[1] or, if $|x| = |y|$, $x < y$.

The first few words over Σ_2 in genealogical order are:

$$\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100 \dots$$

Notice that if we take the subset of words that start with 1, we get the sequence of binary representations of positive integers:

$$1, 10, 11, 100, 101, 110, 111, 1000, \dots$$

[1] Notice that without any other conditions this would not be a total order.

Definition 3

Given a finite or infinite word x, we say that a word v is a factor of x if $x = uvz$ for some words u and z.

We say that v is a prefix (resp., a suffix) of x if $u = \varepsilon$ (resp., $z = \varepsilon$).

We let *Fact*(x), *Pref*(x), *Suff*(x) denote, respectively, the set of factors, prefixes, suffixes of the word x.

We implicitly assume that the empty word ε is a prefix, a suffix and a factor of any word. So, a word w of length n has exactly $n+1$ prefixes and $n+1$ suffixes. It has $O(n^2)$ distinct factors.

A word of length n in which each letter occurs exactly once has $\Theta(n^2)$ distinct factors.

A binary word of length n has at most $2^{k+1} - 1 + \binom{n-k+1}{2}$ distinct factors, where k is the unique integer such that $2^k + k - 1 \leq n \leq 2^{k+1} + k$.

Equivalently, the maximum number of distinct factors of a binary word of length n is

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

Factors, Borders and Powers

Definition 4

We say that a word v is a border of a finite word x if v is both a prefix and suffix of x. A word v is unbordered if it has only trivial borders (ε and v).

Example 5

Let $x = 0101010$. The borders of x are ε, 0, 010, 01010 and x. The word $x = 00100101$ is unbordered.

Notice that if v is a border of a word x and u is a border of v, then u is a border of x.

It is easy to see that a bordered word x has at least one nonempty border of length smaller than or equal to $|x|/2$.

That is, if x is bordered, then there exist v, y, with v nonempty, such that $x = vyv$.

In general, if x is a border of a word w, we can write $w = yx = xz$, for some words y, z of the same length. If $z = y$, we are in a very special situation: the word w can be written both as xy or yx.

The following theorem is fundamental.

Theorem 6 (Lyndon, Schützenberger, 1962)

Let $x, y \in \Sigma^+$. Then the following conditions are equivalent:

1. $xy = yx$ *(x and y commute);*
2. *There exists $z \in \Sigma^+$ such that $x = z^r$ and $y = z^s$ for some integers r and s;*

Notice that the word z in the previous theorem must have a length that divides both $|x|$ and $|y|$. Actually, z can be chosen of minimal length, that is, of length $\gcd(|x|, |y|)$.

The following theorem is a generalization of the Lyndon and Schützenberger theorem.

Theorem 7

The equation

$$w^i = x^j y^k$$

$w, x, y \in \Sigma^+$, $i, j, k \geq 2$, *holds if and only if there exists* $z \in \Sigma^+$ *such that* $w = z^l$, $x = z^m$, $y = z^n$, $li = mj + nk$.

Definition 8

Two words x and y are conjugates if there exists v such that $xv = vy$.

This definition of conjugacy comes from the fact that Σ^* is a monoid with respect the operation of concatenation, but not a group. Therefore, one cannot define the conjugacy in the classical way $y = v^{-1}xv$, but the definition is still possible by "multiplying" both members to the left by v, thus obtaining $vy = xv$.

Conjugacy is an equivalence relation. An equivalence class of words with respect to the conjugacy relation is sometimes called a necklace, or a circular word.

For example, the conjugacy class of 0101 is $\{0101, 1010\}$, while the conjugacy class of 010 is $\{001, 010, 100\}$.

Definition 9

A nonempty word x is called primitive if the cardinality of its conjugacy class is equal to its length $|x|$; that is, if the words in the conjugacy class of x are all distinct.

By definition, if a word is primitive, then every its conjugate is primitive.

Remark 10

An unbordered word is primitive. Conversely, a primitive word may have a nontrivial border, e.g., $x = 01001$.

The following result expresses the conjugacy of two words *à la* Lyndon and Schützenberger.

Lemma 11

Let x, y be nonempty words such that $x \neq y$ and $xv = vy$ for some word v (that is, x and y are conjugates). Then, there exists a unique pair of words (p, q) and a unique integer $m > 0$ such that pq is primitive and

$$x = (pq)^m, \ y = (qp)^m, \ v \in (pq)^* p.$$

For example, let $0100 \cdot 010 = 010 \cdot 0010$. We have that $v = 010$ is a border of $x = 0100$ and $y = 0010$. Let $p = 010$ and $q = 0$. Then $x = pq$, $y = qp$ and $v = (pq)^0 p$.

So, we have:

primitive $\Leftrightarrow$ the equation $xy = yx$ has only trivial solutions

unbordered $\Leftrightarrow$ the equation $xv = vy$ has only trivial solutions

Proposition 12

A word is primitive if and only if it is conjugate to an unbordered word.

Proof.

If a word w is primitive, then its least conjugate in lexicographic order, w', is unbordered. Indeed, if w' had a border then we could write $w' = xyx$, for some x, y, both nonempty (if y were empty w' would not be primitive). Now, x must be lexicographically smaller than y, for otherwise the conjugate $w'' = yxx$ would be lexicographically smaller than w'. But then the conjugate $w''' = xxy$ is smaller than w', against the assumption that w' is the least conjugate in its class.

Conversely, if a word w is not primitive, then there are two conjugates xy and yx of w that coincide. Therefore, by Lyndon and Schützenberger, $w = z^n$, for some word z and $n > 1$. Hence, all conjugates of w have a border. □

As a corollary, any primitive word has at least one unbordered conjugate.

Proposition 13

A word w is primitive if and only if it does not occur internally in ww (that is, it appears only as a prefix and as a suffix in ww).

Thus, an efficient algorithm to check if a word w is primitive is to locate the occurrences of w in ww.

Exercise 14

Prove Proposition 13.

Definition 15

A word w of the form $w = z^n$ for a nonempty z and $n > 0$ (that is, a word that is not primitive) is called an integer power, or simply a power.

Moreover, z can always be chosen to be primitive, and with this assumption n is called the order of the power w and z is called the primitive root of w.

Hence, sometimes, a primitive word is defined as a nonempty word w such that if $w = z^n$ for some z, then $n = 1$.

Remark 16

If p is a prime number, then a word of length p is either a power of a single letter or it must be primitive, since $w = z^n$ implies that $|z|$ must divide $|w|$.

The previous remark can be used to give a very simple proof of the famous Fermat's Little Theorem.

Theorem 17 (Fermat's Little Theorem)

Let p be a prime and k a positive integer. Then $k^p - k$ is a multiple of p.

Proof.

Since p is a prime, the $k^p - k$ words of length p over Σ_k that are not powers of a single letter are grouped in conjugacy classes consisting of primitive words, hence they all have cardinality equal to p. $\square$

Theorem 18 (Shyr and Yu)

Let x, y be primitive words. Then there exists at most one non-primitive word of the form $x^n y^m$, $n, m \geq 1$. If x and y are also unbordered, then every word of the form $x^n y^m$, $n, m \geq 1$, is primitive.

As a consequence of the previous theorem, if a word (of length at least 2) can be written as the concatenation of two (distinct) nonempty unbordered words, then it is primitive — this statement is easy to prove by contraposition. The converse is not true in general; for example 00100 is primitive but cannot be written as the concatenation of two distinct unbordered words.

On the other hand, every unbordered word (of length at least 2) can be written as the concatenation of two (distinct) nonempty unbordered words, as shown in the next proposition — however, observe that $0110 = 011 \cdot 0$ can be written as the concatenation of two distinct nonempty unbordered words, yet it is not unbordered.

Proposition 19

Let x be an unbordered word of length > 1. Let u be the longest proper unbordered prefix (resp., suffix) of x and write $x = uv$ (resp., $x = vu$). Then v is unbordered.

Proof.

Let x be an unbordered word of length > 1, and let v be the longest proper suffix of x that is unbordered. We prove that the prefix u of x such that $uv = x$ is also unbordered. Suppose by contradiction that u has a nonempty border u'. We can write $u = u'zu'$ for some word z (recall that any bordered word has a border whose length is no more than half its length). Write $x = u'z\hat{v}$. The word $\hat{v}$ is a proper suffix of x and is longer than v, so $\hat{v}$ has a nonempty border v'. If $|v'| \leq |u'|$, then v' is a prefix of u, hence of x, and is a suffix of $\hat{v}$, hence of x, against the hypothesis that x is unbordered. If $|v'| > |u'|$, then $v' = u'u''$ for some nonempty u''. But in this case u'' is a prefix and a suffix of v, against the hypothesis that v is unbordered. The other case is symmetric. □

Proposition 20 (Duval factorization)

Every word w can be written uniquely as a concatenation of unbordered prefixes of w.

For example, if $w = 011001110011$, then the Duval factorization of w is $w = 01100111 \cdot 0 \cdot 011$.

The Duval factorization can be computed from right to left by recursively removing the shortest nonempty border of w.

Factors, Borders and Powers

Recall that if f and g are arithmetic functions, the Dirichlet convolution of f and g is defined as

$$f * g = \sum_{d|n} f(d) g\left(\frac{n}{d}\right) = \sum_{ab=n} f(a)g(b)$$

and $*$ is associative and commutative. The Möbius inversion formula says that if $g(n) = \sum_{d|n} f(d)$ then $f = \mu * g$, that is,

$$f(n) = \sum_{d|n} \mu(d) g\left(\frac{n}{d}\right)$$

where $\mu(n)$ is the Möbius function: $\mu(1) = 1$, $\mu(n) = (-1)^j$ if n is the product of j distinct primes or 0 otherwise, i.e., if n is divisible by the square of a prime number.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\mu(n)$	1	-1	-1	0	-1	1	-1	0	0	1	-1	0	-1	1

Table: The first few values of the Möbius function.

Proof.

For every $n > 1$ one has
$$\sum_{d|n} \mu(d) = 0 \tag{1}$$
that is, μ sums up to 0 on every set that is the set of divisors of an integer. Indeed, $\mu(n)$ depends only on the set of primes dividing n and every set has an equal number of odd- and even-cardinality subsets. Now,

$$\begin{aligned}
(\mu * g)(n) &= \sum_{d|n} \mu(d) g\left(\frac{n}{d}\right) && \text{(by definition)}\\
&= \sum_{d|n} \mu(d) \sum_{d'|(n/d)} f(d') && \text{(by hypothesis)}\\
&= \sum_{d'|n} \mu(d') \sum_{d|(n/d')} f(d) && \text{(d and d' both range over all div.s of n)}\\
&= \sum_{d'|n} f(d') \sum_{d|(n/d')} \mu(d) && \text{(by commutativity of Dirichlet convol.)}\\
&= f(n) && \text{(by 1 since the sum is $\neq 0$ only when $d' = n$)}
\end{aligned}$$

$\square$

Factors, Borders and Powers

Theorem 21

The number of primitive words of length n over Σ_k is

$$P_k(n) = \sum_{d|n} \mu(d) k^{n/d}$$

Proof.

The number of words of length n is $S_k(n) = k^n$ and it is equal to $\sum_{d|n} P_k(d)$, where $P_k(d)$ is the number of primitive words of length d. By Möbius inversion formula, $P_k(n) = \sum_{d|n} \mu(d) S_k(n/d)$. □

For example, let $k = 2$ and $n = 4$. The primitive binary words of length 4 are: 0001, 0010, 0011, 0100, 0110, 0111 and their binary complements obtained exchanging 0s and 1s. So we have 12 words in total.

Applying the theorem, we have
$\sum_{d|n} \mu(n) k^{n/d} = \mu(1)2^4 + \mu(2)2^{4/2} + \mu(4)2^{4/4} = 1 \cdot 16 + (-1) \cdot 4 + 0 \cdot 2 = 12$.

Remark 22

From the previous theorem, it follows that the number of conjugacy classes of primitive words of length n *over* Σ_k *is*

$$L_k(n) = \frac{P_k(n)}{n} = \frac{1}{n} \sum_{d|n} \mu(d) k^{n/d},$$

because each class contains n *primitive words.*

The number of conjugacy classes of words (i.e., the number of necklaces) of length n over Σ_k, instead, is

$$N_k(n) = \frac{1}{n} \sum_{d|n} \varphi(d) k^{n/d}$$

where $\varphi(n)$ is the Euler totient function, i.e., the function counting the positive integers smaller than n and coprime with n. Indeed, it follows from the definition of φ that $\sum_{l|d} \varphi(l) = d$ (a number divides d if and only if it divides a divisor of d), and, by Möbius inversion,

$$\varphi(d) = \sum_{l|d} \mu(l) \frac{d}{l}$$

(continues)

Now, since we can associate in a bijective way every word of length n with its primitive root, we have $N_k(n) = \sum_{d|n} L_k(d)$. Thus,

$$\begin{aligned} nN_k(n) &= \sum_{d|n} nL_k(d) \\ &= \sum_{d|n} \frac{n}{d} \sum_{l|d} \mu(l) k^{d/l} \\ &= \sum_{d|n} k^{n/d} \sum_{l|d} \mu(l) \frac{d}{l} \\ &= \sum_{d|n} k^{n/d} \varphi(d) \end{aligned}$$

where we used the commutativity of the Dirichlet convolution.

Periods

An important tool to study and classify finite and infinite words is the notion of period.

Definition 23

A word p is a word-period of a word w if $p = \varepsilon$ or w is a prefix of a power of p. Notice that any word p such that w is a prefix of p is a word-period of w, so every word has at least one word-period.

The shortest nonempty word-period of w is called the fractional root of w and is denoted by ρ_w.

For example, the word $w = 0010010$ has word periods $\varepsilon, 001, 001001, 0010010$, etc.; its fractional root is $\rho_w = 001$.

Exercise 24

Show that a word p is a word-period of a word w if and only if w is a prefix of pw.

Very often, one is interested only in the length of a word-period. So, we give the following

Definition 25

An integer $|p| \geq 0$ is a period of a word w if the letters occurring in w at positions i and j coincide whenever $i = j \mod |p|$. Notice that any integer $|p| \geq |w|$ is a period of w, so every word has at least one positive period.

The minimum (or smallest) positive period of w is denoted by π_w.

For example, the word $w = 0010010$ has periods $0, 3, 6, 7$, etc.; its minimum positive period is $\pi_w = 3$.

Notice that if $|p|$ is a period of w, then any multiple of $|p|$ also is.

Remark 26

If w has period $|p|$, then every factor of w has period $|p|$ as well.

Another remark, due to de Luca and De Luca, is the following:

Lemma 27

Let w be a word. An integer $|p| \leq |w|$ is a period of w if and only if all the factors of w of length $|p|$ are conjugates.

Periods

The notions of word-period and border are intimately related. Indeed, a nonempty word w has a word-period p (with $|p| < |w|$) if and only if w has a border of length $|w| - |p|$.

Equivalently, w has a period $|p|$ shorter than its length if and only if w has a border of length $|w| - |p|$.

We therefore have the following

Remark 28

A word w is unbordered if and only if its smallest positive period is $|w|$.

A word w is primitive if and only if its smallest period dividing $|w|$ is $|w|$.

If w has a nonempty word-period p, then we can write $w = p^n p'$, where $n \geq 1$ and p' is a (possibly empty) prefix of p.

Therefore, a word w is unbordered if and only if $w = p^n p'$ implies $p' = \varepsilon$ and $n = 1$.

Definition 29

A word w is called periodic if $\pi_w \leq |w|/2$. Equivalently, a word is periodic if it has a border that overlaps with itself in w.

Notice that any non-primitive word is periodic. However, there are periodic primitive words, e.g., 01010.

Periods

Theorem 30 (Fine and Wilf)

Let w be a word having positive periods $|p|$ and $|q|$ such that $|p| + |q| - \gcd(|p|, |q|) \leq |w|$. Then w has also period $d = \gcd(|p|, |q|)$.

Proof.

For a fixed d, by induction on $|p| + |q|$. The base case ($|p| = |q| = d$) is trivial. Suppose the statement holds for all integers smaller than $|p| + |q|$. Assume $|p| > |q|$ and let $w = uv$, where $|u| = |p| - d$. Now, for any $1 \leq i \leq |q| - d$, we have $u_i = w_i = w_{i+|p|} = w_{i+|p|-|q|} = u_{i+|p|-|q|}$, and so u has period $|p| - |q|$. Since u has also period $|q|$ and $\gcd(|p| - |q|, |q|) = d$, the inductive hypothesis shows that u has period d. Now, $|u| \geq |q|$ implies that the prefix q of length $|q|$ of w has period d. Since w has period $|q|$, and d divides $|q|$, it follows that w has period d, too. $\square$

Periods

The value $|p| + |q| - \gcd(|p|, |q|)$ is the smallest one that makes the Fine and Wilf's theorem true. As an example showing that the condition on $|w|$ is necessary, the word 0001000 has periods 4 and 6, but not $2 = \gcd(4, 6)$. This can happen because its length is $7 < 4 + 6 - 2 = 8$.

A word w with two coprime periods $|p|$ and $|q|$ and length equal to $|w| = |p| + |q| - 2$ is called a central word. For example, 010 is a central word with coprime periods 2 and 3. Central words are binary palindromic words.

Fine and Wilf theorem has an immediate corollary in the case of a word with two coprime periods.

Corollary 31

Let w be a word having coprime periods $|p|$ and $|q|$ and length $|w| > |p| + |q| - 2$. Then w is a power of a single letter.

Periods

In some applications, one often needs only a weaker version of the Fine and Wilf's theorem:

Theorem 32 (Periodicity Lemma, or Weak Fine and Wilf)

If a word w has positive periods $|p|$ and $|q|$ such that $|p| + |q| \leq |w|$, then $\gcd(|p|, |q|)$ is also a period of w.

Let us show an example of application of the Fine and Wilf's theorem:

Lemma 33

Let w be a word over Σ, with $|\Sigma| > 1$. Then there exists a letter $a \in \Sigma$ such that wa is primitive.

Proof.

We give the proof for $\Sigma = \{0, 1\}$. If $w = \varepsilon$, then $w0$ and $w1$ are both primitive. Suppose then $|w| > 0$, and assume that $w0 = v^k$ and $w1 = u^\ell$ for some primitive words u, v and integers $k, \ell \geq 2$. Both $|u|$ and $|v|$ are periods of w, and since $k, \ell \geq 2$, we have

$$|w| = k|v| - 1 = \ell|u| - 1 \geq 2\max\{|u|, |v|\} - 1 \geq |u| + |v| - 1.$$

By Fine and Wilf, also $d = \gcd(|u|, |v|)$ is a period of w. Since d divides both $|u|$ and $|v|$, and u and v are primitive, we conclude that $|u| = |v| = d$. Since u and v are prefixes of w, we have $u = v$, contradicting the fact that u and v end with different letters. $\square$

As another application, we have the following

Proposition 34

Suppose that w has two distinct primitive word-periods p and q, and let $w = p^n p' = q^m q'$, for some p' prefix of p and q' prefix of q. Then, $n = 1$ or $m = 1$.

Proof.

By contradiction, if $n > 1$ and $m > 1$, then $|p| \leq |w|/2$ and $|q| \leq |w|/2$, so that $|p| + |q| \leq |w|$. By the Periodicity Lemma, $\gcd(|u|, |v|)$ is a period of w, and thus also a period of p and of q, so that at least one between p and q has a period smaller than its length and dividing its length, against the hypothesis that p and q are both primitive. $\square$

Hence, a primitive word can have at most one period that is smaller or equal than half its length.

Periods

Actually, the Fine and Wilf's theorem can be seen as a particularization of the Lyndon–Schützenberger theorem. Indeed, we can state the following general theorem:

Theorem 35

Let $x, y \in \Sigma^+$. Then the following conditions are equivalent:

1. $xy = yx$;
2. *There exist integers $i, j > 0$ such that $x^i = y^j$;*
3. *There exist integers $i, j > 0$ such that $x^i y^j = y^j x^i$;*
4. *There exists $z \in \Sigma^+$ such that $x = z^i$ and $y = z^j$, for some $i, j > 0$ (i.e., $\{x, y\}$ is not a code, i.e., the submonoid $\{x, y\}^*$ has rank 1, i.e., $\{x, y\}^* = z^*$);*
5. $x^* = y^*$;
6. *xy and yx have a longest common prefix of length $|x| + |y| - \gcd(|x|, |y|)$ (Fine and Wilf property);*
7. *xy and yx have the same minimum period p and a common prefix of length p (and $p = \gcd(|x|, |y|)$);*

Exercise 36

Write a proof of the previous theorem.

Corollary 37

If $x^i = y^j$, with x, y primitive, $i, j > 0$, then $x = y$ and $i = j$.

In other words, if x and y are distinct primitive words, then $x^* \cap y^* = \{\varepsilon\}$.

Periods

A beautiful result on the set of periods of a finite word is the following theorem, due to Guibas and Odlyzko, which states that the set of periods of a finite word is independent of the alphabet size (provided that the alphabet has more than one letter).

Theorem 38

For every nonempty word w over any alphabet Σ such that $|\Sigma| > 2$, there exists a word over Σ_2 having the same set of periods as w.

Halava, Harju and Ilie gave a constructive proof of this theorem from which it is possible to construct the binary image of any word in linear time.

The structure of the sets that are period sets of some word (not exceeding the length of the word) is described in the following

Theorem 39 (Breslauer)

Let $S = \{0 = p_0 < p_1 < \ldots < p_s = n\}$ be a set of integers and let $d_h = p_h - p_{h-1}$, $1 \leq h \leq s$. Then S is the set of periods of a word of length n if and only if for each h such that $d_h + p_h \leq n$, one has:

1. *$p_h + d_h \in S$ and*
2. *if $d_h = kd_{h+1}$ for some integer k, then $k = 1$.*

For example, for $S = \{0, 5, 7, 10\}$ the first-differences are $\{5, 2, 3\}$. The set S is not a valid period set since for $h = 2$ we have $9 = 7 + 2 \leq 10$ but 9 is not in S, so condition 1 is violated.

For $S = \{0, 2, 4, 6, 8, 9, 10\}$ condition 1 is not violated, yet it is not a valid period set since the first-differences are $\{2, 2, 2, 2, 1, 1\}$ and we have, for $h = 4$, $8 - 6 = 2(9 - 8)$ and $8 + 2 \leq 10$, so condition 2 is violated.

Using the previous theorem, one can construct the set Γ_n of all valid period sets of words of length n from the set Γ_{n-1} of valid period sets of words of length $n-1$ by the following algorithm:

For each $S \in \Gamma_{n-1}$, if $S \cup \{n\}$ does not violate any of the two conditions of the theorem, add it to Γ_n; if $S \setminus \{n-1\} \cup \{n\}$ does not violate any of the two conditions of the theorem, add it to Γ_n.

Periods

There are interesting connections between the minimum positive period π_w of a word and the maximal length ℓ_w of an unbordered factor of w.

For example, since for every factor u of w, one clearly has $\pi_u \leq \pi_w$ (every period of w is a period of u), it follows that $\ell_w \leq \pi_w$ (since $\ell_u = \pi_u = |u|$ for an unbordered word u).

So, a natural question is if the equality holds. This is not true for any word. For example, in $w = 00110010$, one has $\ell_w = |110010| = 6 < \pi_w = 7$.

Since any primitive word has at least one unbordered conjugate (Proposition 12), every periodic word w must contain all the conjugates of its fractional root (which is primitive). Hence, $|w| \geq 2\pi_w$ implies $\ell_w = \pi_w$.

Periods

Holub and Nowotka solved a problem raised by Ehrenfeucht and Silberger and proved that if $|w| \geq \frac{7}{3}\ell_w$, then $\ell_w = \pi_w$.

Note that the following example, provided by Assous and Pouzet,

$$w = a^n b a^{n+1} b a^n b a^{n+2} b a^n b a^{n+1} b a^n$$

where $n \geq 0$, verifies $\ell_w = 3n + 6$, $\pi_w = 4n + 7$ and $|w| = 7n + 10$, that is, $\ell_w < \pi_w$ and $|w| = \frac{7}{3}\ell_w - 4$.

Palindromes

Palindromes appear frequently in mathematics, theoretical computer science and also in theoretical physics. In fact, palindromes can be used to give interesting descriptions of some properties of sequences.

Definition 40

The reversal of a word $w = w_1 w_2 \cdots w_n$ is the word $\tilde{w} = w_n w_{n-1} \cdots w_1$ obtained by reversing the order of the letters. That is $\tilde{w}_i = w_{n-i+1}$ for every $i = 1, \ldots, n$. The reversal of the empty word is the empty word.

A word that coincides with its reversal is called a palindrome. For example, a, 010010, and $radar$ are all palindromes.

Sometimes one distinguishes between even and odd palindromes. An even palindrome is of the form $x\tilde{x}$ for some word x, while an odd palindrome is of the form $xa\tilde{x}$, for some word x and letter a.

Palindromes

Remark 41

Every border of a palindrome is a palindrome.

Some structural results about palindromes:

Lemma 42

Let w be a word and $n \geq 0$. Then w is a palindrome if and only if so is w^n.

Proposition 43

If $w = pq$, with p and q palindromes, then w has a conjugate $w' = p'q'$, with p' and q' palindromes whose length difference is at most 2.

Palindromes

Proposition 44

For all nonempty palindromes u, v, the word uv is a palindrome if and only if both u and v are powers of some palindrome z.

Theorem 45

Every conjugacy class of words contains at most two palindromes. A conjugacy class contains two palindromes if and only if it contains a word of the form $(x\tilde{x})^i$, where $x\tilde{x}$ is a primitive word and $i \geq 1$.

Palindromes

Proposition 46

A word is a conjugate of its reversal if and only if it is the concatenation of two palindromes.

Exercise 47

Prove Proposition 46.

Palindromes

Every primitive binary word of length greater than 1 has at least two unbordered conjugates (this will be proved easily in the chapter dedicated to Lyndon words). The following theorem is due to Holub and Muller:

Theorem 48

Let w be a primitive binary word of length greater than 1. If w has only two unbordered conjugates, then w is the concatenation of two palindromes.

The converse of the previous statement does not hold true in general. For example, $w = 00101101 = 00 \cdot 101101$ has 4 unbordered conjugates, namely 01011010, 01101001, 10100101 and 10010110.

Palindromes

The following result has been shown by de Luca and Mignosi.

Proposition 49

Every primitive word has at most one factorization in two palindromes.

Notice that there are primitive words that cannot be factored in two palindromes, e.g. 0110.

Palindromes

It is easy to see that a word of length n contains at most n nonempty factors that are palindromes. Indeed, any position between 1 and n cannot be the ending position of the first occurrence of more than one new palindromic factor.

Definition 50

A word is called rich if it contains the maximum number of nonempty palindromic factors, that is therefore equal to its length.

For example, the word 01001 has length 5 and contains 5 nonempty palindromic factors, 0, 1, 00, 010 and 1001, so it is rich; whereas the words 00101100 and 0120 are not rich.

Remark 51

The shortest binary palindrome that is not rich has length 14. *An example is* 00110100101100.

Palindromes

Proposition 52

A word w is rich if and only if every prefix (resp., suffixes) v of w has one palindromic suffix (resp., prefix) unrepeated in v.

For example, let $w = 00101100$. The prefix 0 has the palindromic suffix 0 that is unrepeated in it; the prefix 00 has the palindromic suffix 00 that is unrepeated in it; the prefix 001 has the palindromic suffix 1 that is unrepeated in it; and so on up to the prefix 0010110, which has the palindromic suffix 0110 that is unrepeated in it; but w itself does not have this property, hence it not rich, as all its palindromic suffixes (0 and 00) are repeated.

Corollary 53

If w is rich, then:

1. *it has exactly one unrepeated palindromic suffix;*
2. *all of its factors are rich;*
3. *its reversal is also rich.*

Remark 54

If w is rich, it may have a conjugate that is not rich. For example, 00001011 is rich but its conjugate 00101100 is not.

A word such that all its conjugates are rich is called circularly rich.

Palindromes

Proposition 55

A word w over Σ_2 is not rich if and only if there exists a non-palindromic word v such that $0v0$, $1v1$, $0\tilde{v}1$ and $1\tilde{v}0$ are factors of w.

Rich words are also characterized by a property involving complete returns. We say that a word w is a complete return to v if v appears in w exactly twice, once as a prefix and once as a suffix, i.e., with no internal occurrences.

Proposition 56

A word w is rich if and only if all its factors that are complete returns to palindromes are palindromes.

For example, 00101100 is not rich since it is a complete return to the palindrome 00 but itself is not a palindrome.

In particular, then, consecutive occurrences of a letter in a rich word are separated by palindromes. For example, 0120 is not rich since the factor separating the two occurrences of 0 is not a palindrome.

Palindromes

It is possible to count the number of palindromic factors of a word, hence to decide if a word is rich, in time linear in the length of the word.

The number of rich words of length n over an alphabet of cardinality k is denoted $R_k(n)$. For the binary alphabet, Rubinchik and Shur proved that $R_2(n) \leq c1.605n$ for some constant c.

In addition, Guo, Shallit and Shur proved that the number of rich words grows superpolynomially and conjectured that it grows slightly slower than $n^{\sqrt{n}}$.

Rukavicka proved that $\lim_{n \to \infty} \sqrt[n]{R_k(n)} = 1$ for every k, i.e., $R_k(n)$ has a subexponential growth for every alphabet size.