Combinatorics of Words

Article · April 1997

DOI: 10.1007/978-3-642-59136-5_6 · Source: OAI

2 authors, including:

Christian Choffrut
Paris Diderot University
138 PUBLICATIONS **1,452** CITATIONS

Combinatorics of words

Juhani Karhumäki

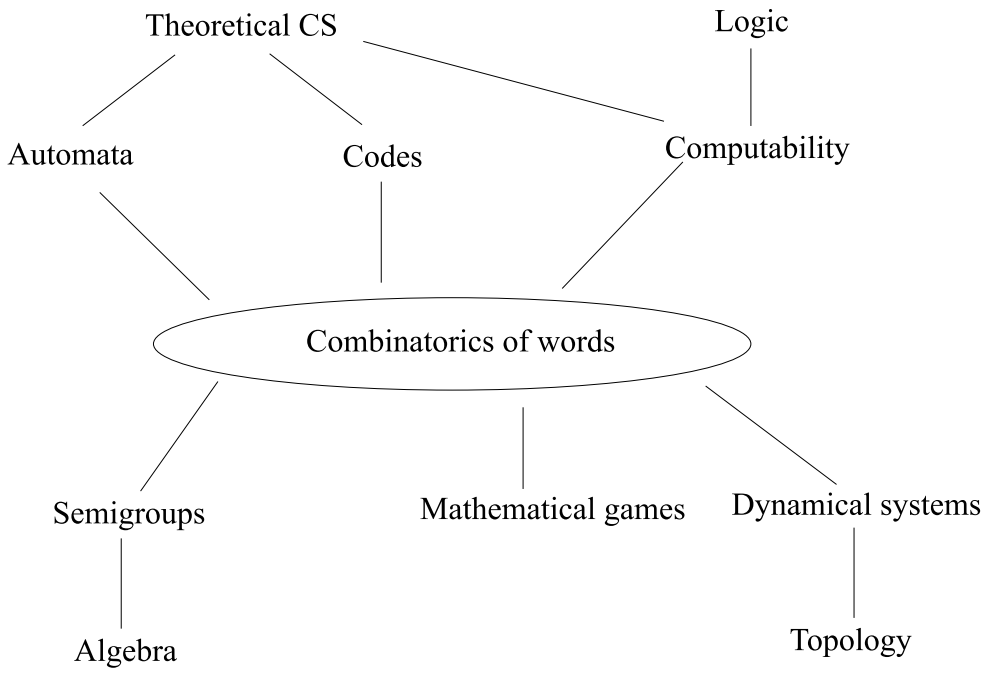
The central notion of this course is a *word,* i.e. a finite or infinite sequence of symbols taken from a finite set. It follows immediately that the mathematical research of words emphasizes two features, namely

- *discreteness,* and
- *noncommutativity.*

In addition an *algorithmic* point of view is often natural. It is probably mainly the noncommutativity which makes the field very challenging: many easily formulated problems are difficult to solve. This is connected to the general fact that there are much weaker mathematical tools to deal with noncommutative structures than commutative ones.

First important papers on words were written by A. Thue at the beginning of this century, cf.[Th1,Th2]; however they became noticed only much later and "classical" as late as 70's. A systematic study of words was initiated by M.P. Schützenberger in 60's. Two influencial papers were [LySch] and [LeSch]. The first, and still most comprehensive, book on words appeared in 1983, cf. [Lo]. A recent survey is [CK].

Combinatorics of words is connected to many modern, as well as classical, fields of mathematics. Connections to combinatorics - actually being part of it - are obvious, but also connections to algebra are deep. Indeed, a natural *environment of a word is a free semigroup.* More generally, the above connections can be illustrated as follows:



This course considers basic properties of words and (finite) sets of words, mainly from combinatorical, but also from algebraic, point of view. In more details topics covered will be:

- periodicity properties,
- equations on words,
- dimension properties such as
  - freeness and
  - defect effect,
- unavoidable regularities such as
  - repetitionfree words and
  - words with repetitions.

No particular prerequisities are required.

Books on the topic:

M.Lothaire, *Combinatorics on words,* Addison-Wesley, 1983.

C. Choffrut and J. Karhumäki, *Combinatorics of words,* in: G. Rozenberg and A. Salomaa (eds), *Handbook of Formal Languages,* Springer, 1997.

H.J. Shyr, *Free monoids and languages,* Hon Min Book Company Taiwan, 1991.

G. Lallement, *Semigroups and combinatorial applications,* John Wiley and Sons, 1979.

J. Berstel and D. Perrin, *Theory of codes,* Academic Press, 1985.

D.Lind and B. Marcus, *An introduction to symbolic dynamics and coding,* Cambridge University Press, 1996

M.Lothaire, *Algebraic combinatorics on words,* Cambridge University Press, 2002,
also in http://www-igm.univ-mlv.fr/~berstel/Lothaire/index.html

J. Berstel, J. Karhumäki, *Combinatorics on Words - A Tutorial,* Bull. EATCS 79, 178-228, 2003, also in
http://www.tucs.fi/Publications/insight.php?id=tBeKa03a&table=techreport

Historically important articles:

A. Thue, *Über unendliche Zeichenvihen,* Kra. Vidensk. Selsk. Skrifter. I Mat.-Nat.Kl., Christiana, Nr. 7, 1906.

A. Thue, *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen,* as above, Nr. 12, 1912.

R.C. Lyndon and M.P. Schützenberger, *The equation $a^m = b^n c^p$ in a free group,* Michigan Math. J. 9 289-298, 1962.

A. Lentin and M.P. Schützenberger, *A combinatorical problem in the theory of free monoids,* in: R.C. Bose and T.E. Dowling (eds), *Combinatorial Mathematics,* North Carolina Press, Chapel Hill, 112-144, 1967.

Contents

1 Notations and basic properties

We first fix some terminology of words.

Alphabet A **:** A nonempty finite set of symbols, like $A = \{a, b\}$.

Word w **:** A sequence of symbols from A, like $(a, b, a) = aba$. Can be finite or infinite (to the right); the latter are called ω*-words* . *Empty word* 1 is a sequence of zero symbols.

A^*, A^+ **and** A^ω **:** Sets of all finite, finite nonempty and infinite words over A, respectively.

Catenation or product of words : Operation defined as

$$a_1...a_n \cdot b_1...b_m = a_1...a_n b_1...b_m.$$

Clearly, this operation is associative and the empty word is the unit element with respect to this operation. Consequently, $A^* = (A^*, \cdot)$ and $A^+ = (A^+, \cdot)$ are a monoid and a semigroup. Moreover, they are *free* in the following sense - so-called *free-monoid* and *semigroup generated by A.*

Free semigroup or monoid : A semigroup (or monoid) S is called *free* if it has a subset B such that each element of S can be uniquely expressed as a product of elements of B. Such a B is referred to as a *free generating set* of S, or a *base* of S.

Language L **over** A **:** Any subset of A^*.

Let $w, u \in A^*$, $a \in A$ and $L, K \subseteq A^*$.

Length of $w = |w|$ **:** the total number of letters in w; $|1| = 0$.

$|w|_a$ **:** the number of a's in w.

Alphabet of w : $Alph(w) = \{a \mid |w|_a > 1\}$.

Factors : A word u is a *factor* of w (resp. *left factor* or a *prefix,* a *right factor* or a *suffix* if there exist words x and y such that

$$w = xuy \quad (\text{resp. } w = uy,\ w = xu).$$

All these are *proper* if they are different from w. We write $u \leq w$ (resp. $u < w$) denoting that u is a prefix (resp. a proper prefix) of w. The set of all prefixes of w is denoted by $\text{pref}(w)$, while $\text{pref}_k(w)$ means the prefix of lenght k of w (or w if $|w| < k$). Similarly, by $\text{suf}(w)$,for instance, we mean the set of suffixes of w.

Quotients : If $u \leq w$ there exists the unique y such that $w = uy$. Such a y is called a *left quotient of w by u,* and is denoted by $u^{-1}w$. In the case u is not a prefix of w $u^{-1}w$ is undefined, so that the function

$$(u, w) \mapsto u^{-1}w$$

becomes a well defined partial function. Similarly we define *right quotients* wu^{-1}.

Reverse of w : if $w = a_1...a_n$ with $a_i \in A$, then $w^R = a_n...a_1$.

Factorizations : A *factorization* of a word w is any sequence $u_1, ..., u_n$ of words such that

$$w = u_1...u_n. \tag{1}$$

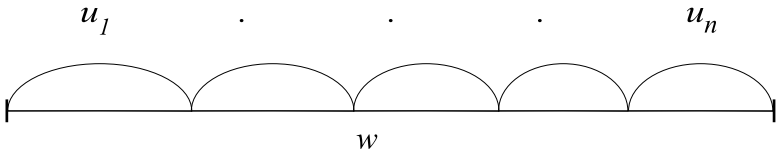
(1) is *L-factorization* if all u_i's are from L. It is natural to write

$$L^* = \{u_1...u_n \mid n \geq 0 \text{ and } u_i \in L\},$$

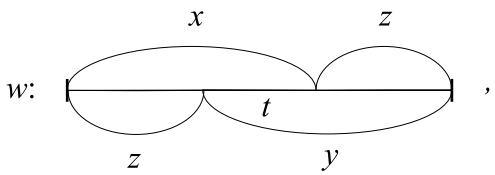
$$L^+ = \{u_1...u_n \mid n \geq 1 \text{ and } u_i \in L\}.$$

The sets L^* and L^+ are submonoids and subsemigroups of A^*, respectively, so-called sub-monoids and subsemigroups *generated* by L. Note that each w in L^* has at least one L-factorization, and if this is always unique, then L^* is free, and L is its base. Such an L is called a *code.*

Factorizations are illustrated by pictures :



or



,

the latter meaning that $w = xz = zy = ztz$.

Operations for languages : For languages we have

- Boolean operations :
  - *Union* $L \cup K$,
  - *Intersection* $L \cap K$,
  - *Complementation* $L^c = A^* \setminus L$.
- Operations connected to the product of words :
  - *Product* LK,
  - *Quotients* $L^{-1}K$ and KL^{-1},
  - *Iterations* L^* and L^+.

Here the product and quotients are defined componentwise, i.e. for example $L^{-1}K = \{l^{-1}k \mid l \in L,\ k \in K\}$. Further L^+ is usually called *1-free iteration* of L.

Morphism $h : A^* \to B^*$ (or $A^+ \to B^+$) : A mapping $h : A^* \to B^*$ which satisfies :

$$h(ww') = h(w)h(w') \quad \text{for all } w, w' \in A^*.$$

In particular, it follows that

- $h(1) = 1$ and
- h is completely specified by the words $h(a)$ with $a \in A$.

We call a morhphism h

- *1-free* if $h(a) \neq 1$ for all $a \in A$,
- *periodic* if $\exists z$ such that $h(a) \in z^*$ for all $a \in A$,
- *uniform* if $|h(a)| = |h(b)|$ for all $a, b \in A$,
- *prefix (resp. suffix)* if none of the words in $h(A)$ is a prefix (resp. suffix) of another,
- *code* if h is injective.

The notion of a morphism is very important in combinatorics of words!

Finally, we define a few more special notions of words.

Conjugates : Two words x and y are conjugates if there exist words u and v such that

$$x = uv \text{ and } y = vu,$$

or equivalently, that they are obtained from each other by a *cyclic permutation* $c : A^* \to A^*$ defined as

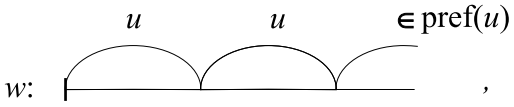
$$\begin{cases} c(1) & = 1 \\ c(w) & = \text{pref}_1^{-1}(w)w\text{pref}_1(w) \quad \text{for } w \in A^+, \end{cases}$$

i.e. $x = c^k(y)$ for some k. Note that in the second picture of page 2 x and y are conjugates. We denote the relation "being conjugates" by $\sim$; clearly this is an equivalence relation.

Periods of w : Let $w = a_1...a_n$ with $a_i \in A$. We say that number p is a *period* of w if

$$a_i = a_{i+p} \quad \text{for } i = 1, ..., n-p.$$

This can be illustrated as



where $u =\text{pref}_p(w)$. The smallest period of w is called *the period* of w, denoted as $p(w)$. The elements in the conjugacy class of $\text{pref}_{p(w)}(w)$ are called *cyclic roots* of w

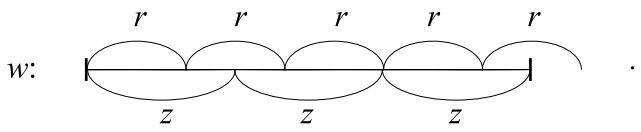
Example 1. A word can have several periods. For example words *abababa* and *aabaabbaabaa* have periods 2,4,6 and 7,10,11, respectively. Moreover, any number $\geq |w|$ is always a period of w.

Primitive words : We say that a word $w \neq 1$ is *primitive* if it is not a proper integer power of any of its cyclic roots.

Theorem 1. *A word $w \in A^+$ is primitive iff it satisfies*

$$\forall z \in A^* : [w = z^n \Rightarrow n = 1 \text{ and hence } w = z]. \tag{2}$$

Proof. Clearly, (2) implies the primitiveness. To prove the converse let w be primitive and $w = z^n$ with $n \geq 2$. Let $r = \text{pref}_{p(w)}(w)$. We can illustrate the situation as follows :



Since $|r|$ is the period of w, $|z| \geq |r|$. Moreover, by primitiveness of w we have $z \notin r^*$. Consequently, comparing the prefixes of length $|r|$ of the two first occurrences of z we can write

$$r = ps = sp \quad \text{with } p, s \neq 1.$$

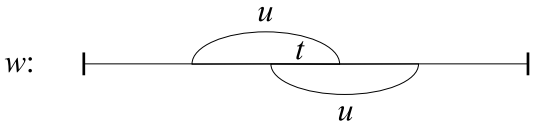
But now by Theorem 3 (be sure that we do not make a chain conclusion!) p and s are powers of a nonempty word, a contradiction since $|r|$ was the period of w. □

Note that often the primitiveness of a word is defined using the condition (2).

There exist two particularly important classes of primitive words, namely unbordered and Lyndon words.

Unbordered words : A word w is *unbordered* if its smallest period is $|w|$. In other words, w does not contain any nonempty word both as a proper prefix and as a suffix. Of course, a word is *bordered* if it is not unbordered.

Note that bordered words can *overlap* as factors of another word :

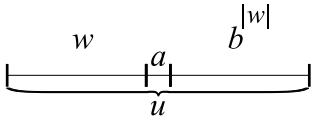


For unbordered words this is not possible.

Example 2. We give a simple construction how arbitrary word over at least two letter alphabet can be extended to an unbordered word. Consider a word w. Let

$$u = wab^{|w|}, \quad \text{where } a = \text{pref}_1(w) \text{ and } b \neq a.$$

The illustration is now as follows :



Now, by the choise of a and b,
- no nonempty suffix of u of length $\leq |w|$ is a prefix of u ;
- no prefix of u of length $\geq |w|$ is a suffix of u.

Consequently, u is unbordered.

Lyndon words : These are primitive words which are the smallest in their conjugacy class with respect to the lexicographic order, cf. page 17.

After fixing the terminology we now go to basic combinatorical results on words.

Theorem 2. *If words u,w,x and y over A satisfy uw=xy, then there exists a unique word t such that either*

i. u=xt and y=tw, or

ii. x=ut and w=ty.

Proof. By symmetry, we may assume that $|u| \geq |x|$. Then since uw and xy represent the same word over A, x is a prefix of u, i.e. there exists a t such that $u = xt$. Moreover such a t is unique. We can write

$$xy = uw = xtw.$$

Again this is a identity of words, so that $y = tw$. Hence (i) holds. □

Our next result (and it's corollary) characterizes when two words commute.

Theorem 3. *Let $u, v \in A^*$. The following conditions are equivalent :*

i. u and v commute, i.e. uv=vu,

ii. u and v satisfy a nontrivial relation,

iii. there exists a word t such that $u, v \in t^*$.

Proof. Since Theorem is obvious if u or v is empty we assume that this is not the case.

(i)⇒(ii) Clear, since uv=vu is a nontrivial relation on u and v.

(ii)⇒(iii) By induction on $|u| + |v|$. We assume that $|u| \geq |v|$; the other case being symmetric.

If $|u| = |v| = 1$, then clearly the implication holds.

Now let $\alpha = \beta$ be a nontrivial relation satisfied by u and v, i.e. $\alpha, \beta \in \{u, v\}^+$ and are not identical as words over $\{u, v\}$, but

$$\alpha = \beta \text{ in } A^*. \tag{3}$$

Since α and β are not identical over $\{u, v\}$ we may assume, possibly by removing common prefixes of α and β (in $\{u, v\}^*$), that $\alpha = u\alpha_1$ and $\beta = v\beta_1$ with $\alpha_1, \beta_1 \in \{u, v\}^*$.

By (3) and Theorem 2 there exists a word w such that

$$u = vw. \tag{4}$$

If $w = 1$ we are done as at the beginning. So let $w \neq 1$. Now let α_2 and β_2 be words over $\{v, w\}$ obtained from α and β by replacing each occurrance of u by the word vw. Then α_2 and β_2 are nonidentical over $\{v, w\}$ since the former starts with vw and the latter with vv. On the other hand, clearly $\alpha_2 = \beta_2$ in A^*, and since $|v| + |w| < |u| + |v|$ the induction hypothesis implies that there exists a word t such that $v, w \in t^*$. But, by (4), the same holds for u and v proving (iii).

(iii)⇒(i) Obvious. □

Theorem 3 has two interesting Corollaries. The first one is just a weaker form of Theorem 3

Corollary 1. *Two words $u, w \in A^*$ commute if and only if they are powers of a same word.*

The second one gives a representation result for words :

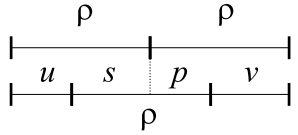
Corollary 2. *For each $w \in A^+$ there exists the unique primitive word $\rho(w)$ such that $w = \rho(w)^n$, for some $n \geq 1$.*

Proof. The existence of at least one required $\rho(w)$ is clear by Theorem 1: If w is not primitive write $w = z^n$ with $n \geq 2$. And if z is not primitive continue at the same way until a required $\rho(w)$ is found.

For the uniqueness assume that both ρ_1 and ρ_2 are primitive and $w = \rho_1^n = \rho_2^m$ with $n, m \geq 1$. Then ρ_1 and ρ_2 satisfy a nontrivial relation, and hence by Theorem 3, are powers of a same word. But by Theorem 1, a primitive word can be a power of a word only in a trivial way. Hence $\rho_1 = \rho_2$, and the Corollary 2 holds. □

The word $\rho(w)$ in Corollary 2 is called the *primitive root* of the word w, and the number n is the *exponent* of w.

Example 3. We claim that a primitive word ρ cannot be a factor of the square ρ^2 in a nontrivial way, i.e. if $\rho^2 = u\rho v$, then necessarily either $u = 1$ or $v = 1$. Assume the contrary : $\rho^2 = u\rho v$ with $u, v \neq 1$. Let p and s be words such that $us = \rho = pv$. This is illustrated in the following figure:



We have

$$u\rho v = \rho\rho = uspv, \text{ or equivalently } \rho = sp.$$

On The other hand, ρ has a prefix p and a suffix s, so that we also have $\rho = ps$. Consequently, $ps = \rho = sp$ and so by Theorem 3 p and s are powers of a same word, and therefore $\rho = z^n$, with $n \geq 2$; a contradiction since ρ is primitive. □

In Theorem 3 we characterized the commutation of words. Next we characterize when two words are conjugates.

Theorem 4. *Let $u, v \in A^+$. The following conditions are equivalent :*

i. u and v are conjugates,

ii. there exists a word z such that $uz = zv$,

iii. there exists words z,p and q such that

$$u = pq, \ v = qp \text{ and } z \in p(qp)^*.$$

Proof. The equivalence of (i) and (iii) is obvious. So it is enough to prove the equivalence of (ii) and (iii).

(iii)⇒(ii). If $u = pq, \ v = qp$ and $z = p(qp)^n$ with $n \geq 0$, then

$$uz = pqp(qp)^n = p(qp)^{n+1} = (pq)^{n+1}p = p(qp)^n qp = zv.$$

(ii)⇒(iii). Assume that $uz = zv$. Then for all n

$$u^n z = u^{n-1}uz = u^{n-1}zv \underset{\text{hyp}}{\overset{\text{ind.}}{=}} zv^{n-1}v = zv^n.$$

Now choose n such that

$$n|u| \geq |z| > (n-1)|u|,$$

and consider the equation

$$u^n z = zv^n. \tag{5}$$

Then, by Theorem 2,

$$z = u^{n-1}p \text{ and } zq = u^n \quad \text{for some words } p \text{ and } q.$$

Now

$$u^n = zq = u^{n-1}pq,$$

so that $u = pq$, and hence by (5) and above,

$$v^n = qz = q(pq)^{n-1}p = (qp)^n,$$

and so $v = qp$. This completes the proof. □

Note that in Theorem 4 the conjugacy relation is characterized in terms of

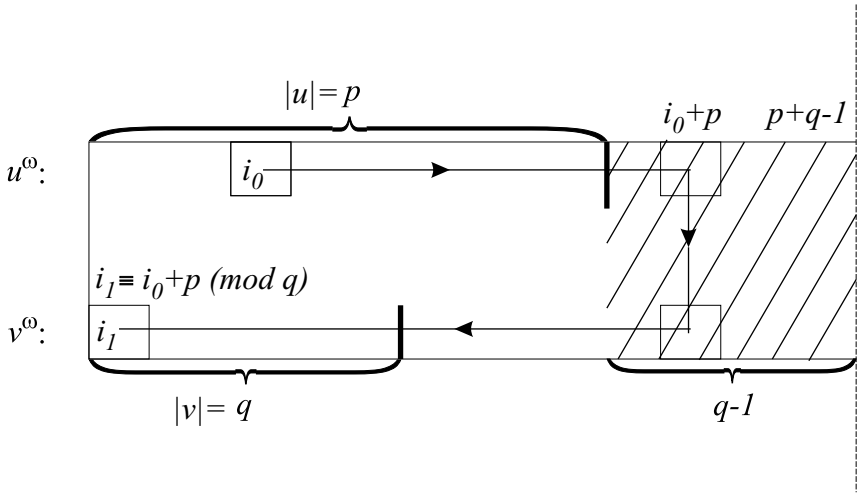
- equations in (ii), and
- solutions of an equation in (iii).

Next we prove a fundamental periodicity result of words often referred to as the Periodicity Lemma of Fine and Wilf.

Theorem 5 (Fine and Wilf, 1956). *Let $u, v \in A^+$. Then the words u and v are powers of a same word if and only if the words u^ω and v^ω have a common prefix of length $|u| + |v| - \gcd(|u|, |v|)$.*

Proof. We first note how to reduce the general case to the case where $\gcd(|u|, |v|) = 1$. If $\gcd(|u|, |v|) \neq 1$, say $|u| = dp$ and $|v| = dq$, with $\gcd(p, q) = 1$, we consider u and v as elements of $(A^d)^+$, i.e. over the alphabet A^d, where letters are words of length d in the original alphabet. In the larger alphabet $\gcd(|u|, |v|) = 1$, and if we can prove the theorem there it immediately gives the general proof.

So we assume that $|u| = p$ and $|v| = q$ with $\gcd(p, q) = 1$. In one direction the implication is trivial. To prove the converse we assume that u^ω and v^ω have a common prefix of length $p + q - 1$. We assume further, by symmetry, that $p > q$, and illustrate our proof in the following figure :



Here p and q denote the lengths of the words u and v, and positions of words u^ω and v^ω are numbered from $1, ..., p + q - 1$. The dashline tells how far the words u^ω and v^ω can be compared. Finally, the arrow describes the *procedure* defined as follows :

The purpose of this procedure is to fix the values of new positions to be the same as a given value of an initial position $i_0 \in [1, ..., q-1]$. Now, by our assumptions, the value of the position computed as follows:

$$i_0 \longmapsto i_0 + p \longmapsto i_1 \equiv i_0 + p \pmod{q}, \tag{6}$$

where i_1 is reduced to the interval $[1, ..., q]$, gets the same value as that of i_0. So the procedure computes from i_0 the number i_1. Since $\gcd(p, q) = 1$ the number i_1 is different from i_0. If it is different from q as well the procedure can be repeated, and the new position obtained is different from the previous ones. Indeed, if

$$i_0 + np \equiv i_0 + mp \pmod{q}, \text{ with } n, m \in [0, q-1],$$

then necessarily $n = m$, since $\gcd(p, q) = 1$.

The crucial observation here is that if the procedure can be repeated $q-1$ times, then all the positions in the shadowed area will be covered, and so those get the same value as the initial on i_0. But this means that v is over a unary alphabet, and consequently so is u. This would complete the proof.

But the procedure can be repeated $q-1$ times if we choose i_0 such that

$$i_0 + (q-1)p \equiv q \pmod{q}. \tag{7}$$

If this is the case, then all the values $i_0 + jp \pmod{q}$ for $j = 0, ..., q-2$ are different from q, which was the condition for an application of the procedure. Clearly, such an i_0 satisfying (7) can be found. □

There exists an obvious reformulation of Theorem 5.

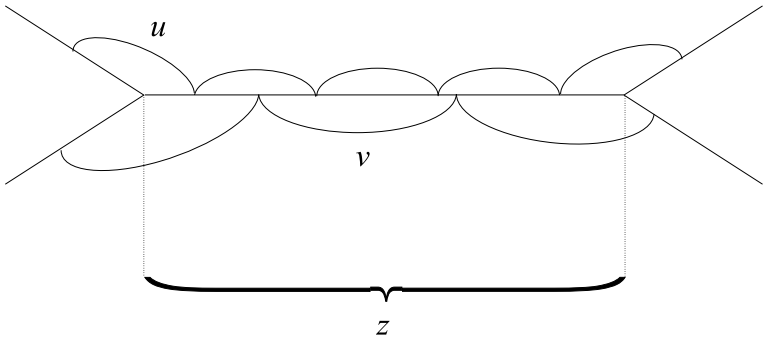
Corollary 1. *If a word has two periods p and q, and if it is of length at least $p + q - \gcd(p, q)$, then it has also a period $\gcd(p, q)$.*

The formulation of Theorem 5 emphasized the fact that the considered two words were periodic from the left end. Of course, the theorem can be formulated still differently: a word has a factor having two periodic presentations. To formulate this let us denote by $l(u, v)$ the length of a maximal common factor of words u and v. Then we can formulate Theorem 5 as follows:

Corollary 2. *For any two words $u, v \in A^+$ we have*

$$l(u^\omega, v^\omega) \geq |u| + |v| - \gcd(|u|, |v|) \Rightarrow \rho(u) \sim \rho(v).$$

Proof. The assumptions can be illustrated as



where $|z| \geq |u| + |v| - \gcd(|u|, |v|)$. Clearly, there are conjugates u_1 and v_1 of u and v, respectively, such that $z \leq u_1^\omega, v_1^\omega$. Hence, by Theorem 5, u_1 and v_1 are powers of a same word, say t, and thus also powers of the primitive word $\rho(t) = \rho$. So by Corollary 2, $\rho(u_1) = \rho(v_1) = \rho$.

We need the following simple

Claim: If x is primitive and $x' \sim x$, then also x' is primitive. The proof of the claim is easy: If $x' = s^n$, then clearly $c(x') = (c(s))^n$, where c is the cyclic permutation of the page 4, so that also $c(x')$ would be an nth power. Hence so would be x.

From the beginning we obtain

$$u_1 = \rho^n \text{ and } v_1 = \rho^m.$$

Hence, by the argument used to prove the claim, we conclude that

$$u = \bar{\rho}^n \text{ and } v = \bar{\bar{\rho}}^m,$$

where $\bar{\rho} \sim \rho \sim \bar{\bar{\rho}}$. By Claim, $\bar{\rho}$ and $\bar{\bar{\rho}}$ are primitive so that

$$\rho(u) = \bar{\rho} \sim \bar{\bar{\rho}} = \rho(v).$$

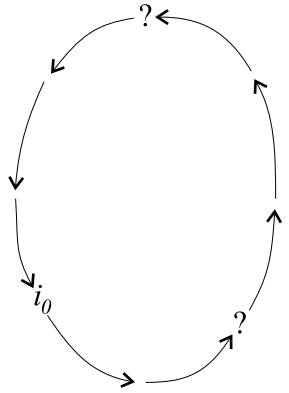
□

Example 4. Conjugates need not have the same smallest period, although by Claim they have equally long primitive roots, as shown by words

$$abaa \text{ and } aaab.$$

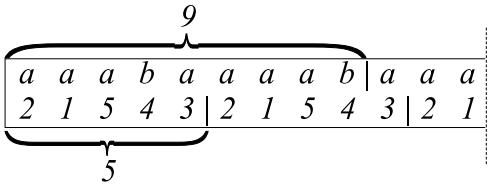
In particular, in Corollary 2 we cannot replace primitive roots by the periods. □

Example 5. Theorem 5 allows a word of length $p+q-2$, with $\gcd(p,q)=1$, to have periods p and q without being a power of a letter. We claim that such a word always exists, and moreover *it is binary and unique* up to renaming of letters. The proof of this claim follows directly from that of Theorem 5. In that proof we fixed the values of positions of the shorter word using the procedure. In that proof the procedure could be applied in all but one case, namely when the position was q. Now there are two such positions, namely q and $q-1$. In these situations the value of the new position could change. In the proof of Theorem 5 this could happen only once. But finally the new value would be the same as the value of the starting position i_0. So in fact all positions got the same value!



Here the change of the values can happen twice - in positions denoted by ? in the figure. Now the latter change is back to the original one, so that actually only one change is possible. But this makes the considered word unique and binary. □

Example 6. The words of Example 5 are so-called *Sturmian* words. Let us compute such a word for values $q=5$ and $p=9$. The length of such a word is 12 and it is obtained by fixing values as follows :



Here the numbers tell the order in which the procedure fixes the values. The word looked for is

$$(aaaba)^2aa = (aaabaaaab)aaa.$$

And as we proved this is the only word of length 12, starting with a and having periods 5 and 9. □

A weaker version of the Theorem of Fine and Wilf can be formulated as follows:

Theorem 6. *For any two words $u, v \in A^+$ we have*

$$z \leq u^\omega, v^\omega \text{ and } |z| \geq |u| + |v| \Rightarrow \rho(u) = \rho(v).$$

Obviously this is just a special case of Theorem 5. On the other hand, it might be easier to remember, and actually strong enough for most of applications. For example, Theorem 6 immediately implies the unique representation of Corollary 2 on page 7: Indeed, if $w = \rho_1^n = \rho_2^m$, with ρ_1 and ρ_2 primitive, hold, then Theorem 6 forces ρ_1 and ρ_2 to be powers of a same word, and hence as primitive ones equal.

Our next goal is to compute the number $p_n(k)$ of all primitive words of length n over a k-letter alphabet. So let $|A| = k$. Define

$l_n(k) =$ the number of the conjugacy classes of primitive words of length n over A.

We note two simple facts.

Lemma 1. *If words x and y are conjugates so are their primitive roots. In particular, the exponents of x and y are equal.*

Proof. Lemma 1 follows directly from Claim of p. 11 and Corollary 2 of p. 7. □

Lemma 2. *Let the length and the exponent of x be n and e, respectively. Then the conjugacy class C_x of x contains exactly $\frac{n}{e}$ words.*

Proof. Let $x = \rho^e$ with ρ primitive. Then

$$c^{|\rho|}(x) = c^{|\rho|}(\rho^e) = \rho^e = x.$$

Hence C_x contains at most $|\rho| = \frac{n}{e}$ words, namely the words

$$x, c(x), ..., c^{|\rho|-1}(x).$$

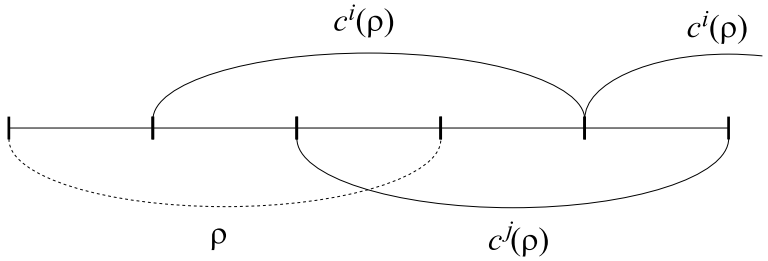
On the other hand, these are pairwise different, since if

$$c^i(x) = c^j(x) \quad \text{with } 0 \leq i < j \leq |\rho| - 1, \tag{8}$$

then also

$$c^i(\rho) = c^j(\rho),$$

which yields the situation:



By Lemma 1, $c^i(\rho)$ is primitive. Consequently, the assumption (8) would give a contradiction with Example 3. □

From Lemmas 1 and 2 (restricted to primitive words) we conclude:

$nl_n(k) =$ the number of primitive words of length n over A with $|A| = k$ $(= p_n(k))$.

Note also that $l_1(1) = 1$ and $l_n(1) = 0$ for $n \geq 2$. The above extends to

Lemma 3. $k^n = \sum_{d|n} dl_d(k)$

Proof. As usual the notation $\sum_{d|n}$ means the sum over all factors of n.

Now, by our representation result of words (Corollary 2 on p. 7) we have a *one-to-one* correspondence:

$$x \in A^n \longleftrightarrow (\rho, e) \text{ with } \rho \text{ primitive and } n = e \cdot |\rho|.$$

When x ranges here all possibilities we obtain the left hand side k^n of the required identity. On the other hand, the number of different possibilities obtained from the right hand side of the above correspondence is (by Lemma 2):

$$\sum_{\substack{e|n \\ n=de}} dl_d(n) = \sum_{d|n} dl_d(n).$$

□

In order to compute $l_n(k)$ we need so-called *Möbius-function* $\mu : \mathbb{N}\backslash\{0\} \to \mathbb{Z}$ defined as:

$$\mu(1) = 1,$$
$$\mu(n) = \begin{cases} (-1)^i & \text{if } n \text{ is a product of } i \text{ different primes,} \\ 0 & \text{otherwise.} \end{cases}$$

We still need one more general lemma, so-called *Möbius Inverse Formula.*

Lemma 4. *Let $\alpha, \beta : \mathbb{N} \setminus \{0\} \to \mathbb{Z}$ be functions. Then*

$$\alpha(n) = \sum_{d|n} \beta(d) \quad , \forall n \geq 1, \tag{9}$$

iff

$$\beta(n) = \sum_{d|n} \mu(d)\alpha(\frac{n}{d}) \quad , \forall n \geq 1. \tag{10}$$

Proof. Let S be the set of functions $\mathbb{N} \setminus \{0\} \to \mathbb{Z}$. Define a binary operation $*$ by the condition:

$$(f * g)(n) = \sum_{n=de} f(d)g(e).$$

Clearly, $*$ is well defined, and moreover comutative and associative (verify!). The function $\mathbf{1}$ defined by

$$\begin{aligned} &\mathbf{1}(1) = 1, \\ &\mathbf{1}(n) = 0 \quad \text{for } n > 1, \end{aligned}$$

is the unit element:

$$(\mathbf{1} * f)(n) = \sum_{n=de} \mathbf{1}(d) f(\frac{n}{d}) = f(n).$$

It follows that $(S, *)$ is a commutative monoid.

Let τ be the constant function:

$$\tau(n) = 1 \quad , \forall n \geq 1.$$

We claim that

$$\tau * \mu = \mathbf{1}. \tag{11}$$

If $n = 1$, then $(\tau * \mu)(n) = (\tau * \mu)(1) = \tau(1)\mu(1) = 1$. For the general case let $n = p_1^{k_1} \cdots p_m^{k_m}$ be the canonical representation of n. Then

$$\mu(d) \neq 0 \text{ iff } d = p_1^{l_1} \cdots p_m^{l_m} \text{ with each } l_j = 0 \text{ or } 1.$$

If this is the case, then $\mu(d) = (-1)^t$, where $t = \sum_{j=1}^{m} l_j$. Clearly, for a fixed t, there exist $\binom{m}{t}$ different such choices for d. So we can compute:

$$\begin{aligned} (\tau * \mu)(n) &= \sum_{n=ed} \tau(e)\mu(\frac{n}{e}) = \sum_{d|n} \mu(d) \\ &= \sum_{t=0}^{m} (-1)^t \binom{m}{t} = \sum_{t=0}^{m} \binom{m}{t} (-1)^t 1^{m-t} = (-1+1)^m = 0. \end{aligned}$$

So formula (11) has been proved.

Now, assuming (9) we conclude from the definition of τ that

$$\alpha(n) = \sum_{d|n} \beta(d) = \sum_{n=ed} \tau(e)\beta(d) = (\tau * \beta)(n),$$

in other words, that

$$\alpha = \tau * \beta.$$

But this implies that

$$\mu * \alpha = \mu * (\tau * \beta) = (\mu * \tau) * \beta = \mathbf{1} * \beta = \beta,$$

which is just a reformulation of (10). The reverse implication is exactly similar proving Lemma 4. □

Now, we are ready for our goal.

Theorem 7. *The number $p_n(k)$ of primitive words of length n over a k-letter alphabet is $\sum_{d|n} \mu(\frac{n}{d}) \cdot k^d$.*

Proof. As we noted on page 14

$$p_n(k) = n l_n(k),$$

where, by Lemma 3, $l_n(k)$ satisfies

$$\alpha(n) = k^n = \sum_{d|n} d l_d(k).$$

Finally, Lemma 4 allows to compute:

$$\beta(n) = n l_n(k) = \sum_{d|n} \mu(d) k^{\frac{n}{d}} = \sum_{d|n} \mu(\frac{n}{d}) \cdot k^d.$$

□

Example 7. Theorem 7 allows to show that asymptotically "almost all" words are primitive, namely that, cf. Exc.,

$$\lim_{n\to\infty} \frac{|\{w \in A^n |\ w \text{ is primitive}\}|}{|\{w \in A^* |\ |w| = n\}|} = 1.$$

□

As the last topic of this chapter we consider an important subclass of primitive words, namely so-called *Lyndon words.* In order to define these we need some terminology.

In many cases it is important to consider all words over A in some order, i.e. to define *total ordering* of A^*. Such an ordering can be defined in many different ways - two particularly important ones are lexicographic and alphabetic orderings. These are obtained by

- assuming a total ordering of A, i.e. of letters, and
- extending it to all words.

Assume that alphabet A is *totally ordered* by $\prec$, i.e. for each two letters $a \neq b$ either $a \prec b$ or $b \prec a$. Further for words u and v let $u \wedge v$ denote the *maximal common prefix* of u and v. Then we extend the total ordering $\prec$ of A to *lexicographic ordering* $\prec_l$ and *alphabetic ordering* $\prec_a$ of A^* by setting

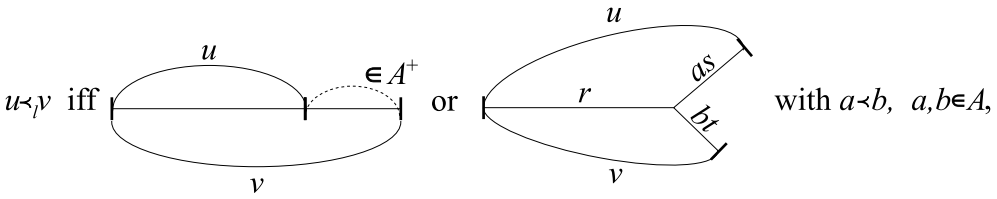
$$u \prec_l v \text{ iff } u^{-1}v \in A^+ \text{ or } \mathrm{pref}_1((u \wedge v)^{-1}u) \prec \mathrm{pref}_1((u \wedge v)^{-1}v)$$

and

$$u \prec_a v \text{ iff } |u| < |v| \text{ or } |u| = |v| \text{ and } u \prec_l v.$$

Clearly, both $\prec_l$ and $\prec_a$ are total orderings on A^*: For each pair (u, v) of words one of the relations $u = v$, $u \prec v$ or $v \prec u$ hold (where we have omitted the indices l and a). We write $u \preceq v$ iff $u = v$ or $u \prec v$. It is also worth noting that $\prec_l$ and $\prec_a$ *coincide on words of equal lengths.*

The definition of the lexicographic ordering can be illustrated as follows :



i.e. either u is a proper prefix of v or after the maximal common prefix of u and v u continues with a "smaller" letter. In particular, we have

$$u < v \quad \Rightarrow \quad u \prec_l v,$$

i.e. the lexicographic ordering is an extension of the relation of "being a proper prefix".

For *now on we concentrate on* lexicographic ordering and denote it simply by $\prec$. We have the following simple result.

Lemma 5. *For all words u,v,w and z we have:*

i. $u \prec v \quad \Leftrightarrow \quad wu \prec wv,$

ii. if $v \notin uA^*$, *then :* $u \prec v \quad \Rightarrow \quad uw \prec vz.$

Proof. (i) Obvious (from the illustration of the relation $u \prec_l v$).
(ii) As above. Indeed, if we have the second alternative in that illustration, the words can be extended arbitrarily preserving the relation. □

Now, a word $w \in A^+$ is *Lyndon word* if

i. w is primitive, and

ii. w is minimal in its conjugacy class with respect to the lexicographic ordering, i.e. satisfies:

$$\forall u, v \in A^+ : \ w = uv \ \Rightarrow \ w \prec vu.$$

Note that our earlier number $l_n(k)$ gives the number of Lyndon words of length n. Let L_y denote the set of all Lyndon words (over our fixed alphabet A).

Next we prove two characterizations of Lyndon words.

Theorem 8. *A word* $w \in A^+$ *is Lyndon word if and only if it is strictly smaller than any of its proper suffices, i.e.*

$$w \in L_y \ \Leftrightarrow \ [\forall v \in A^+, \ w \in A^+ v \ \Rightarrow \ w \prec v].$$

Proof. $\Rightarrow$. Let $w \in L_y$ and $w = uv$ with $u, v \in A^+$. We first show that v is not a prefix of w. Assume the contrary:

$$w = vt \quad \text{with } t \in A^+.$$

Then we have

$$uv = w = vt,$$

so that, by Theorem 4,

$$u = xy, \ t = yx \text{ and } v = (xy)^i x \text{ with } i \geq 0 \text{ and } x, y \in A^*.$$

Consequently,

$$w = (xy)^{i+1} x.$$

If here $x = 1$, then since $v \neq 1$, necessarily $i \geq 1$. But then w would not be primitive.

So we may assume that $x \neq 1$. Now since w is a Lyndon word we have

$$w = (xy)^{i+1}x \prec x(xy)^{i+1}.$$

This, by (i) of Lemma 5, yields

$$(yx)^{i+1} \prec (xy)^{i+1},$$

and further, by (ii) of Lemma 5, the relation

$$(yx)^{i+1}x \prec (xy)^{i+1}x = w.$$

This, however, is a contradiction with the fact $w \in L_y$.

So we have proved that $w \notin vA^*$. Now, if we would have

$$v \prec uv = w, \tag{12}$$

then by (ii) of Lemma 5, we would also have $vu \prec uv = w$, a contradiction since $w \in L_y$. Hence, (12) cannot hold as was to be proved.

$\Leftarrow$. So we assume that w is smaller than any of its proper sufficies. Write $w = uv$, where $u, v \in A^+$. Then we have

$$uv = w \prec v \prec vu,$$

showing that w is a Lyndon word, since by assumption it is also primitive. □

Our second characterization is as follows.

Theorem 9. *A word $w \in A^+$ is a Lyndon word if and only if $w \in A$ or there exist Lyndon words l and m such that $w = lm$ with $l \prec m$.*

Proof. $\Leftarrow$. Since $A \subseteq L_y$ we may assume that $w = lm$ with $l \prec m$. First we prove that $lm \prec m$.

If $m \notin lA^*$, then this follows from (ii) of Lemma 5 and the fact $l \prec m$. If, on the other hand, $m = lm'$, then by Theorem 8 $m \prec m'$, and so by (i) of Lemma 5, $lm \prec lm' = m$.

Next we show that lm is smaller than any of its proper sufficies, which by Theorem 8 means that lm is a Lyndon word. There are two alternatives:

First if $v \neq 1$ is a suffix of m, then since m is a Lyndon word we conclude from above that

$$lm \prec m \preceq v.$$

Second if $v = v'm$, where v' is a proper suffix of l, we obtain, since also l is a Lyndon word, that $l \prec v'$, and so by (ii) of Lemma 5, also $lm \prec v'm = v$.

So in all cases $lm \prec v$, proving the implication $\Leftarrow$.

$\Rightarrow$. Let w be a Lyndon word. Further we may assume that $w \notin A$. Let m be the longest proper suffix of w such that m is a Lyndon word. Clearly, such a suffix exits since always a letter is a Lyndon word. We set

$$w = lm.$$

If $l \in A$ we are done: $l \prec lm \prec m$, where the latter relation holds by Theorem 8. So we assume that $l \notin A$ and will prove that $l \in L_y$.

Consider an arbitrary proper suffix v of l. Then, by the choice of m, the word $vm \notin L_y$. Hence Theorem 8 guarantees the existence of a proper suffix t of vm such that $t \prec vm$. If now

$$v \prec t,$$

then

$$v \prec t \prec vm,$$

so that there exists a word $s \neq 1$ such that

$$t = vs \text{ and } s \prec m.$$

But this means, since t is a proper suffix of vm, that s is a proper suffix of m with $s \prec m$, a contradiction to the fact that $m \in L_y$ (Theorem 8).

It follows that necessarily

$$t \preceq v,$$

and so

$$l \prec lm \prec t \preceq v,$$

where the second inequality follows again from Theorem 8.

Now, since v was an arbitrary suffix of l, we conclude, again from Theorem 8, that l is a Lyndon word.

To complete the proof we have to prove that $l \prec m$. This, however, is a direct consequence of Theorem 8 and the fact that $lm \in L_y$. □

Theorem 10 (Lyndon, 1955). *Each word $w \in A^+$ can be expressed uniquely as a product of nonincreasing Lyndon words:*

$$w = l_1 ... l_n \text{ with } l_i \in L_y \text{ and } l_n \preceq l_{n-1} \preceq ... \preceq l_1.$$

Proof. Let $w \in A^+$. Since letters are Lyndon words, w has a representation as a product of Lyndon words. Consider now the representation

$$w = l_1 ... l_n \text{ with } l_i \in L_y \text{ and } n \textit{ is minimal.}$$

If now, for some i, $l_i \prec l_{i+1}$, then by Theorem 9 $l_i l_{i+1}$ is a Lyndon word, and so n above is not minimal. Hence, w has at least once required representation.

To prove the uniqueness let

$$l_1 ... l_n = w = l'_1 ... l'_m$$

where $l_i, l'_j \in L_y$ and $l_n \preceq l_{n-1} \preceq ... \preceq l_1$ and $l'_m \preceq l'_{m-1} \preceq ... \preceq l'_1$. Assume now that, for example, l'_1 is a proper prefix of l_1, i.e.

$$l_1 = l'_1 ... l'_i u \text{ with } u \preceq l'_{i+1}.$$

Then, by Theorem 8,

$$l_1 \prec u \preceq l'_{i+1} \preceq l'_1 \prec l_1.$$

This, however, is a contradiction, so that necessarily $l_1 = l'_1$, and hence inductively $n = m$ and $l_i = l'_i$ for $i = 1, ..., n$. □

We note that Lyndon words are used in several considerations in algebra.

2 Basics on equations

In this section we consider some basic properties of *word equations,* including methods of solving those. Implicitly some such methods have already been used in the previous section. It has to be emphasized that there does not exist any general method to solve a given word equation. The difficultiness of the probles is supported by the fact that probably the most important result on words is so-called *Makanin's algorithm* which gives (only!) an algorithm to *decide* whether a given equation has a solution.

Let us fix some terminology. Let A be a finite alphabet, and X a *finite* set of *unknowns* with $A \cap X = \emptyset$. An *equation* with X as the set of unknowns over A is a pair

$$(u, v) \in (A \cup X)^* \times (A \cup X)^*,$$

normally written as $u = v$. A *solution* of an equation $u = v$ is a morphism $h : (X \cup A)^* \to A^*$ satisfying

$$h(u) = h(v) \text{ and } h(a) = a \text{ for } a \in A,$$

i.e. *identifying* u and v. The set of all solutions of an equation $u = v$ is denoted by $\mathrm{Sol}(u = v)$. Obviously, these notions extend in a natural way to (not necessarily finite) systems of equations.

In above we allow equations to contain constants. An equation $u = v$ is *constant-free* if $u, v \in X^*$. Let E and E' be two systems of equations (with the same finite set of unknowns). We say that E and E' are

$$\textit{equivalent} \text{ if } \mathrm{Sol}(E) = \mathrm{Sol}(E').$$

Further we say that E is *independent* if for each $u = v \in E$ there exists a solution h in $\mathrm{Sol}(u = v) \backslash \mathrm{Sol}(E \setminus \{u = v\})$. Finally, we call an equation $u = v$ *reduced* if $\mathrm{pref}_1(u) \neq \mathrm{pref}_1(v)$ and $\mathrm{last}_1(u) \neq \mathrm{last}_1(v)$.

Note that a solution of an equation is an $|X|$-tuple of words over A. Hence we are solving equations in the *free monoid* A^*. Sometimes, however, it is interesting to solve these in a *free semigroup* A^+, i.e. by requiring that h is 1-free.

Example 1. Consider equations

$$e_1 : \ xy = yx \text{ and } e_2 : \ xxyyxx = yyxyxyy.$$

By Theorem 3, both of those have only *periodic* solutions, i.e. $x, y \in t^*$ for some t. However, they are not equivalent since

$$\mathrm{Sol}(e_1) = \{(t^n, t^m) | t \in A^*; \ n, m \geq 0\},$$
$$\mathrm{Sol}(e_2) = \{(t^{3n}, t^{2n}) | t \in A^*; \ n \geq 0\}.$$

□

Example 2. Equations

$$xz = zy \text{ and } xz^2 = z^2x$$

are equivalent. This was shown in the proof of Theorem 4 and Exc. 2/III. □

Example 3. As we have seen equations can be used to *characterize properties* of words:

$x \in (ab)^* \Leftrightarrow xab = abx$

x and y commute $\Leftrightarrow xy = yx$

$$x \text{ and } y \text{ are conjugates} \Leftrightarrow xz = zy \Leftrightarrow \begin{cases} x &= pq \\ y &= qp \end{cases}.$$

In the last characterization this is by *using additional unknowns.* Very little is known which properties of words are expressable in this way as components of solutions of an equation. In particular to show that something is *not expressable* seems to be very difficult. □

Next we introduce some methods to solve equations.

I. *Length argument.* Comparing the lengths of different sides of an equation (or a system of equations) we obtain a linear equation (or a linear system of equations) on numbers, which leads to *potential* solutions. This method gives for example the solution of e_2 in Example 1. Sometimes length argument can be used to preficies or sufficies of the equation:

$$xyyx = uvvu \Leftrightarrow \begin{cases} xy &= uv \\ yx &= vu \end{cases}.$$

Indeed, here the middle of both sides can be detected by the length argument, and the first halves of the equations must coincide.

II. *Splitting of an equation.* As above sometimes equation can be splitted into several equations. A criterium for splitting is often some form of the length argument, but might be also the fact some words are known to match. For example, if we know that

- x is unbordered, and
- $uxv = yxz$ with $|u|, |v|, |y|, |z| \leq |x|$,

then necessarily $u = y$ and $v = z$. Also Example 2 might be useful here.

III. *Elimination of the leftmost (or rightmost) unknown,* sometimes referred to as *Levi's Lemma* or *Neilsen transformation.* This method, which we have already used several times, is based on Theorem 1: If we have

$$x\alpha = y\beta \text{ with } \alpha, \beta \in \{X \cup A\}^*$$

we write

$$x = yt \text{ (or } y = xt),$$

and substitute this to the original equations to get

$$yt\alpha = y\beta \Leftrightarrow t\alpha = \beta.$$

The hope is that the new equation is "simpler", and thus leads to a solution. This, however, need not be the case, since when substituting x this has to be done in everywhere in α and β so that the total length of the equation may grow. Note also that the above method means a change of X:

$$X \longmapsto X \cup \{t\} \setminus \{x\}.$$

IV. *Use of already known equations.* Methods II and III lead to new equations, for which the solutions might be already known. Such basic equations are the commutation $xy = yx$, or the conjugacy $xz = zy$, or also the use of Fine and Wilf Theorem to conclude the periodicity.

Example 4. Consider the pair $\begin{cases} xy &= uv \\ yx &= vu \end{cases}$ from the previous page. The method III yields

$$x = ut \text{ (or } u = tx)$$

which implies that

$$v = ty \text{ and } yut = tyu.$$

Now, the latter condition means that t and yu commute, i.e. we can write

$$t = (\alpha\beta)^k,\ y = (\alpha\beta)^i\alpha \text{ and } u = \beta(\alpha\beta)^j,$$

where $\alpha, \beta \in A^*$ and $i, j, k \geq 0$. This leads to the general solution

$$\begin{cases} x &= \beta(\alpha\beta)^{j+k} \\ y &= (\alpha\beta)^i\alpha \\ u &= \beta(\alpha\beta)^j \\ v &= (\alpha\beta)^{k+i}\alpha \end{cases} \text{ or } \begin{cases} x &= \beta(\alpha\beta)^j \\ y &= (\alpha\beta)^{i+k}\alpha \\ u &= \beta(\alpha\beta)^{j+k} \\ v &= (\alpha\beta)^i\alpha \end{cases},$$

where $i, j, k \geq 0$ and $\alpha, \beta \in A^*$. Of course, here the second of solutions comes from the symmetric case. □

Example 5. Let us show that the equation

$$x^2y^2 = z^2 \tag{1}$$

has only periodic solutions, i.e. $\rho(x) = \rho(y) = \rho(z)$ (if $x, y, z \neq 1$). Assume that (x, y, z) is a solution. Now, taking a suitable conjugacy in (1) we conclude that there exists z' such that

$$z' \sim z \text{ and } z'^2 = xy^2x.$$

But then by the length argument $xy = z' = yx$ so that x and y are powers of a word. Therefore applying Fine and Wilf Theorem (or even its weaker form) to (1) we obtain the required claim.

The claim can be extended to the implication:

$$x^ny^m = z^k \text{ with } n, m, k \geq 2 \Rightarrow \exists t : x, y, z \in t^*.$$

Note how simple our proof of the speacial case was!

In Example 4 the general solution of an equation $xyyx = uvvu$ was expressed using *parametric words,* i.e. expression containing

- word parameters α and β, and
- integer parameters i, j and k.

Here, as in general in constant-free equations, the word parameters can be fixed to be any word over A. In the case of equations with constants, of course, values of unknowns in solutions might be fixed words (as in Example 1), or for example such that they start with a certain symbol.

Unfortunately, it can be proved that a general solution of a constant free equation need not have a finite expression using parametric words. In fact, even as simple equation as $xyz = ztx$ is such!

Our next goal is to show that any *Boolean combination of equations* can be transformed into a single equation, having often more unknowns, such that the solution sets restricted to the set of original unknowns coincide. So for equations $u = v$ and $u' = v'$ we have to consider "solutions" of

conjunction : $u = v$ and $u' = v'$,

disjunction : $u = v$ or $u' = v'$, and

inequation : $u \neq v$.

Obviously these solution sets can be formally defined as:

- $\mathrm{Sol}(u = v \wedge u' = v') = \mathrm{Sol}(u = v) \cap \mathrm{Sol}(u' = v')$
- $\mathrm{Sol}(u = v \vee u' = v') = \mathrm{Sol}(u = v) \cup \mathrm{Sol}(u' = v')$
- $\mathrm{Sol}(u \neq v) = A^* \times A^* \backslash \mathrm{Sol}(u = v)$.

In these considerations we assume that $|A| \geq 2$. We start with a simple result.

Theorem 11. *Each pair of equations is equivalent to a single equation.*

Proof. Consider equations $u = v$ and $u' = v'$. We claim that

$$\begin{cases} u & = v \\ u' & = v' \end{cases} \quad \Leftrightarrow \quad uau'ubu' = vav'vbv',$$

where a and b are two different constants of A. Obviously if the left hand side holds so does the right one. Conversily, assume that

$$uau'ubu' = vav'vbv'.$$

Now considering the lengths of both sides of this equality we can specify the middle point and hence split the equality to the pair

$$\begin{cases} uau' & = vav' \\ ubu' & = vbv' \end{cases}.$$

If here, for example, u would be longer than v we would have

$$u = va... = vb...$$

which is impossible. Therefore $u = v$ and $u' = v'$ as required. □

It is worth noting that here we do not need any new unknowns. So we have

Corollary 1. *Each finite system of equations is equivalent to a single equation.*

In the other two results we have to introduce new unknowns. Let us continue with the inequation.

For any two words $\alpha, \beta \in A^*$ we have

$$\begin{aligned} \alpha \neq \beta \quad \Leftrightarrow \quad & \exists a \in A, t \in A^* : \ \alpha = \beta a t, \\ & \exists a \in A, t \in A^* : \ \beta = \alpha a t, \text{ or} \\ & \exists a, b \in A, \ t, r, s \in A^* : \ \alpha = tar, \ \beta = tbs, \ a \neq b. \end{aligned}$$

This guides us to associate each inequation $u \neq v$, where $u, v \in \{X \cup A\}^*$, with the formula

$$\mathrm{F}(u,v) = (\bigvee_{a \in A} u = vaz) \vee (\bigvee_{a \in A} v = uaz) \vee (\bigvee_{\substack{a,b \in A \\ a \neq b}} u = zaz' \vee v = zbz'),$$

where z, z' and z'' are new unknowns. By Theorem 11, we can transform $\mathrm{F}(u,v)$ into the form

$$\bigvee_{i=1}^{n} (u_i = v_i) \quad \text{with each } u_i, v_i \in (X \cup \{z, z', z''\} \cup A)^*.$$

It follows from the construcion that for any values of the unknowns X we have

$$u \neq v \text{ iff } \exists i, \exists z, z', z'' \in A^* : \ u_i = v_i.$$

So we can formulate.

Theorem 12. *For each inequation $u \neq v$ with unknowns X there exists a finite set of equations $u_i = v_i$, where $i = 1, ..., N$, with unknowns $X \cup \{z, z', z''\}$ such that*

s is a solution of $u \neq v$ iff

$\exists z, z', z'' \in A^*$ such that (s, z, z', z'') is a solution of $u_i = v_i$ for some $i = 1, ..., N$.

Finally we reduce a disjunction of two equations into one equation.

Theorem 13. *For any pair $u = v$ and $u' = v'$ of equations with unknowns X there exists an equation $x = y$ over unknowns $X \cup \{z, z'\}$ such that*

s is a solution of $u = v$ or $u' = v'$ iff

$\exists z, z' \in A^*$ such that (s, z, z') is a solution of $x = y$.

Proof. We start with two small reduction steps. Since clearly

$$u = v \vee u' = v' \quad \Leftrightarrow \quad uv' = vv' \vee vu' = vv'$$

we may assume that $v = v'$, i.e. the pair is of the form

$$u = v \vee v = u'.$$

We can also assume that $u \neq u'$ since otherwise the claim is trivial.

We associate to a word α a word

$$< \alpha >= \alpha a \alpha b, \text{ with } a \neq b.$$

First we note that for each α the period of $< \alpha >$ is longer, than half of it's length, in particular $< \alpha >$ is primitive, cf. Excercises. Now, the result is a consequence of the following equivalence:

$$u = v \text{ or } u' = v \quad \Leftrightarrow \quad \exists z, z' : w = zqz', \tag{2}$$

where

$$q =< uu' >^2 v < uu' >^2$$

and

$$w =< uu' >^2 u < uu' >^2 u' < uu' >^2 .$$

Proof of Equivalence (2). $\Rightarrow$. If $u = v$ then we can choose $z = 1$ and $z' = u' < uu' >^2$ and if $u' = v$ then we can choose $z =< uu' >^2 u$ and $z' = 1$.

$\Leftarrow$. First, since $< uu' >$ is primitive it occurs inside the word $< uu' >^2$ in exactly two places: as a prefix and as a suffix. Second, the word $< uu' >^2$ cannot occur in $< uu' > u < uu' >$ or in $< uu' > u' < uu' >$ since $< uu' >$ is longer than both u and u', and the period of $< uu' >$ is longer than half of its length. From these two facts we see that either $z = 1$ or $z' = 1$ and so $u = v$ or $u' = v$. □

We summarize results of Theorems 11-13 as follows.

Theorem 14. *For each Boolean combination B of equations with unknowns X we can construct a single equation E with unknowns $X \cup X'$ such that*

$$Sol(B) = Sol(E|X),$$

where Sol(B) denotes the set of all solutions of B and Sol($E|X$) the set of all solutions of E restricted to unknowns X.

□

The above results provide useful tools to show that some properties of words are *expressable as solutions of certain equations,* or more precisely, as values of some components of solutions of equations. We take a few examples:

Example 6. Following properties are easy to express:

"y **is a prefix of** x" : $x = yz$.

"y **is a proper prefix of** x" : $x = yz$ and $z \neq 1$, or $x = yz$ and $\bigvee_{a \in A} z = at$.

"x **contains a square as a factor**" : $x = yz^2w$.

"x **is imprimitive**" : $xy = yx$ and $x < y$, i.e. $xy = yx$ and $x = yz$ and $z \neq 1$.

We can also easily force two unknowns x and y

"to be powers of a word" : $xy = yx$

"to be unequal" : $x \neq y$

by adding the above formulas

On the other hand, it seems to be not known whether, for example, properties "x is primitive" or "$x \in \{a, b\}^*$, with $\{a, b\} \subset A$, are expressable.

We conclude this section by considering an algorithmic problem of deciding whether a given equation has any solution. As we already noted this problem is decidable, as shown by Makanin 1976. However, the algorithm is very complicated. Note that for constant-free equations the problem is trivial since such an equation has always a solution, where all unknowns are equal to 1. Note also that a given equation need not have any solution as shown by encoding the properties "x starts with a" and "x starts with b" into one equation:

$$x = az \text{ and } x = by \;\Leftrightarrow\; xaxxbx = azabyazbby.$$

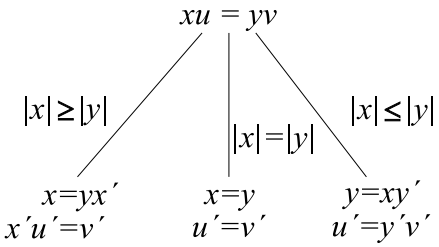
We present a solution for the above algorithmic problem in a non-trivial case, namely in the case *the equation contains each unknown at most twice.*

Let

$$xu = yv \text{ with } x, y \in X \cup A,\ u, v \in (X \cup A)^* \tag{3}$$

be an equation. We intend to solve it by *applying exhaustively Levi's Lemma,*

i.e method III: We have three possibilities illustrated as:



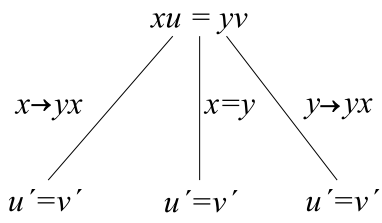
(1)

where x' (resp. y') is a new unknown and new equations containing u' and v' are obtained from the original one by substituting $x = yx'$ (or $x = y$ or $y = xy'$). So what was done here is that the set X of unknowns is changed to

$$(X \setminus \{x\}) \cup \{x'\},\ X \setminus \{x\} \text{ or } (X \setminus \{y\}) \cup \{y'\}.$$

Note that the middle case is actually included in the others, but "needed for termination".

Since in (1) we only replace an unknown by another (or identify it with another) we prefer not to rename it. Accordingly we write (1) in the form:



(2)

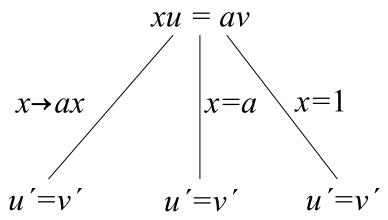
Next we note that

$xu = yv$ has a solution with $|x| \geq |y|$ iff

$u' = v'$ has a solution.

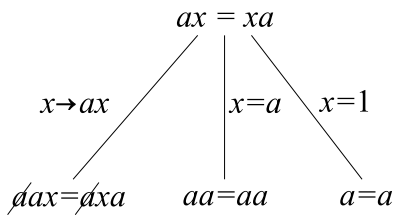
Moreover, from any solution of $u' = v'$ a solution of $u = v$ with $|x| \geq |y|$ is obtained by changing the x-component to yx. Consequently, all solutions of the equation $u = v$ are obtained by finding all solutions on leaves of graph (2).

In the case $y \in A$, say $y = a$, the graph (2) is as follows:



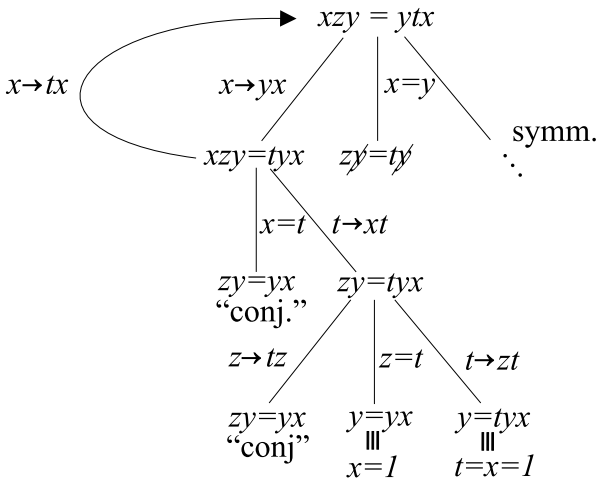
Let us take two examples.

Example 7. Consider the equation $ax = xa$. Now the above method leads to the following finite graph



On the two rightmost branches we end up with an identity, so solutions are found. On the leftmost branch we have returned to the original equation. Hence all solutions are found by repeating the leftmost branch a finite number of times, and then terminating on the other branches. Hence, indeed the general solution is $x \in a^*$.

Example 8. The equation $xzy = ytx$ (which was the one having no finite expression for the general solution using parametric words) leads to the graph:



This graph, which is now finite, describes all solutions of the equation $xzy = ytx$. Indeed, they are obtained by solving the conjugacy equation, fixing some unknowns to be empty and applying Levi's Lemma in the reverse order. for example, starting from the leftmost leaf of the bottom line we found the following solutions (underlined below):

$$\begin{aligned}(x,y,z,t) =& (pq, qpq, qp, t) \to (pq, qpq, tqp, t) \to \\ & (pq, qpq, tqp, pqt) \to \underline{(qpqpq, qpq, tqp, pqt)} \to^2 \\ & \underline{(qpqpqtqpqpq, qpq, tqp, pqt)} \to^2 .\end{aligned}$$

□

Above considerations yield easily:

Theorem 15. *It is algorithmically decidable whether a given equation containing each unknown at most twice has a solution.*

Proof. Let $xu = yv$ be an equation of the above form. Let us consider the operations of (1). We claim that the total length of the equation cannot increase:

(i) If we identify $x = y$, then the length reduces by two.

(ii) If we set $x = yx$, then

- the x at the beginning of u remains as x;
- the y at the beginning of v is cancelled;
- the potential other occurrance of x is replaced by xy; while
- all other occurrances of symbols in u and v remain as they were.

In all cases the equation remains in the required form. This proves the claim.

But the claim means that the whole graph is finite! Hence, the existence of a solution is reduced to check whether on some leaf of the graph we get a solution (and not a contradiction). □

It is interesting to note that if we replace the phrase "at most twice" in Theorem 15 by "at most three times" we are in the general problem for system of equations: Let $u = v$ be an equation containing n occurrances of x. Then let $u' = v'$ be the equation obtained from $u = v$ by replacing the ith occurrance of x by a new unknown x_i. Then we have

$$u = v \;\Leftrightarrow\; u' = v',\; x_1 = x_2,\; x_2 = x_3, ...,\; x_{n-1} = x_n,\; x_n = x_1.$$

Doing the same for all unknowns we get the result. Finally, note that Theorem 15 and its proof hold for systems of equations as well.

3 Repetition-free words

The study of repetition-free words has been one of the central areas in combinatorics of words. It was initiated by A. Thue at the beginning of this century, when he proved several basic results of the field. Repetition-free words has a lot applications e.g. in algebra.

We start by fixing some terminology. A *square* is a word of the form uu, with $u \neq 1$, and a *cube* is a word of the form uuu with $u \neq 1$. More generally, a word w is a *kth power* if

$$w = u^k = \mathrm{pref}_{k \cdot |u|}(u^\omega), \text{ with } u \neq 1.$$

Here it is not required that $k \in \mathbb{N}$, only that $|u| \cdot k = |w|$ (which requires k to be a rational). For example, $(aaab)^{\frac{5}{4}} = aaaba$. We say that a word w contains a *repetition of order k* if it contains as a factor a *kth or higher* power of a word.

Next we define the central notion of the *repetition freeness,* or actually there are three related notions. Let $k \in \mathbb{R}_+$ and $w \in A^* \cup A^\omega$. We say that w is

k-free, if it does not contain a repetition of order k;

k^+-free, if for any $k' > k$, it is $k' - free$;

k^--free, if it is k-free, but not k'-free for any $k' < k$.

It follows immediately that

$$w \text{ is } k^-\text{-free } \Rightarrow \ w \text{ is } k\text{-free } \Rightarrow \ w \text{ is } k^+\text{-free},$$

while the reverse implications are not true, in general. The special cases 2-free, 2^+-free and 3-free are called as *square-free, overlap-free* and *cube-free.* As an illustration let us note that

- the 2-freeness means that w does not contain a square,
- the 2^+-freeness means that w can contain a square, but no factor of the form $uvuvu$, with $u \neq 1$, and
- the 2^--freeness means that w does not contain a square, but does contain, for any $\epsilon > 0$, a repetition higher than $2 - \epsilon$.

Now we formulate:

Thue's Problem. Find as long as possible, preferably infinite, word over an n-letter alphabet such that it is k-free (or k^+-free or k^--free).

Two remarkable results of Thue were:

i. There exists an infinite 2^+-free word over a binary alphabet, and

ii. There exists an infinite 2-free word over a 3-letter alphabet.

Example 1. It is very easy to see that in (i) the 2^+-freeness can not be replaced by the 2-freeness. Indeed, any 2-free binary word is of length at most 3, the words *aba* and *bab* being the maximal. □

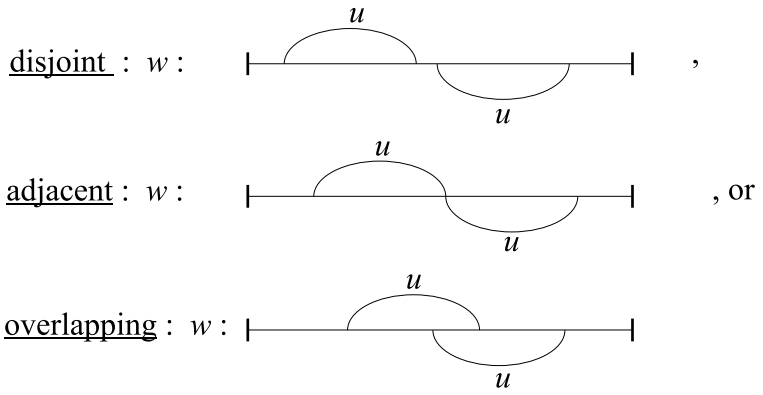
Thue's Problem has natural *commutative* variants *when* $k \in \mathbb{N}$. We say that a word w contains an *Abelian repetition of order k,* if it contains a factor $u_1...u_k$ such that $\Pi(u_1) = \Pi(u_2) = ... = \Pi(u_k)$, where the mapping Π gives the *commutative image* of a word, i.e.

$$\Pi(u) = (|u|_{a_1}, ..., |u|_{a_n}), \text{ when } A = \{a_1, ..., a_n\}.$$

Now, a word is *Abelian k-free* if it does not contain an Abelian repetition of order k, and *commutative variants of Thue's problem* asks to find as long as possible Abelian k-free word over an n-letter alphabet.

Example 2. As in Example 1, there exist only rather short words which are Abelian 3-free in a binary alphabet, or Abelian 2-free in a 3-letter alphabet. To find exact bounds are left as an excercise. □

Now we go to solutions of Thue's Problems. First we characterize 2^+-free words in terms of how different occurrances of factors can situate inside a word. We note that two occurrances of a factor u of w can be



In the last case we say that w contains an *overlapping factor.*

Theorem 16. *Let $w \in A^+ \cup A^\omega$. Then the following conditions are equivalent:*

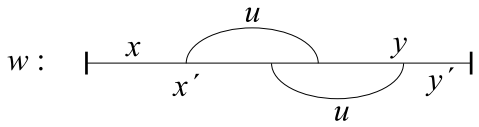
i. w is 2^+-free;

ii. w does not contain a factor of the form $avava$ with $a \in A,\ v \in A^*$;

iii. w does not contain an overlapping factor.

Proof. (i)⇒(ii). Obvious, since if w contains $avava$ as a factor it contains a repetition of order at least $2 + \dfrac{1}{|v|+1}$.

(ii)⇒(iii). We assume that w contains an overlapping factor u and show that then it also contains a factor of the form $avava$ with $a \in A$ and $v \in A^*$. But this is quite obvious, since if $w = xuy = x'uy'$ with $|x| < |x'| < |xu| < |x'u|$, i.e. we have the illustration



then

$$w \text{ has a factor } avava,$$

where $a = \mathrm{pref}_1(u)$ and $v = (\mathrm{pref}_1(u))^{-1}\mathrm{pref}_{|x'|-|x|}(u)$.

(iii)⇒(i). Now, if w is not 2^+-free it contains a repetition of order k, with $k > 2$, and hence a factor $u^2\mathrm{pref}_1(u)$. This means that w contains an overlapping factor $avava$ where $a = \mathrm{pref}_1(u)$ and $v = \mathrm{pref}_1(u)^{-1}u$. Hence also the third implication is proved. □

Now we are ready for Thue's construction. Consider a morphism $h : \{a,b\}^* \to \{a,b\}^*$ defined by

$$h : \begin{cases} a \mapsto ab \\ b \mapsto ba \end{cases} .$$

Then clearly

$$a = h^0(a) \le h(a) = ab, \tag{1}$$

so that (by induction)

$$h^i(a) \le h^{i+1}(a) \quad \text{for all } i \ge 0. \tag{2}$$

Consequently, there exists the unique infinite word $\alpha \in \{a,b\}^\omega$ such that

$$\alpha = \lim_{i\to\infty} h^i(a) = abbabaabbaababba... \tag{3}$$

Note that for any morphism h (1), i.e. condition $a \leq h(a)$, implies (2), and also (3), if only $\lim_{i\to\infty} |h^i(a)| = \infty$. Clearly, this is the case if

- $h(a) \in aA^+$ and
- h is 1-free.

Such morphism are called *prolongable.* So it follows that for each prolongable morphism h there exists the unique infinite word α_h obtained by *iterating h at point a.* Moreover, α_h is a fixed point of h, i.e.

$$h(\alpha_h) = \alpha_h.$$

Now let us go back to the word α of (3) which is usually referred as *Thue-Morse word.*We prove

Theorem 17 (Thue, 1906). *There exists an infinite 2^+-free word over a binary alphabet. In particular Thue-Morse word α is such.*

Proof. Based on two lemmas. □

Lemma 6. *Let $X = \{ab, ba\}$. If $x \in X^*$, then $axa,\ bxb \notin X^*$.*

Proof. By induction on $|x|$.

(i) $|x| = 0$. Clear since $aa, bb \notin X^*$.

(ii) Assume that $x \in X^*$ with $x \neq 1$. Assume further that $u = axa \in X^*$ (the case $bxb \in X^*$ being symmetric). We can write

$$u = abx_1...x_{r-1}ba \text{ with } r \geq 1,\ x_i \in X.$$

Set $y = x_1...x_{r-1}$ so that $x = byb$ with $y \in X^*$. Hence, by induction hypothesis $x \notin X^*$, a contradiction. So necessarily $axa \notin X^*$ proving Lemma 6. □

Lemma 7. *Let $w \in \{a, b\}^+$. If $h(w)$ has an overlapping factor so does w.*

Proof. Assume that $h(w)$ has an overlapping factor. By Theorem 16 it can be assumed to be of the form $cvcvc$ with $c \in \{a, b\}$ and $v \in \{a, b\}^*$, that is we can write

$$h(w) = xcvcvcy.$$

Now $|cvcvc|$ is odd, and since $h(w) \in \{ab, ba\}^*$ necessarily $|xy|$ is odd as well.

This yields two possibilieties:

i. $|x|$ is even, and $x,\ cvcv,\ cy \in X^*$, or

ii. $|x|$ is odd, and $xc,\ vcvc,\ y \in X^*$.

We claim that $|v|$ is odd. If this would not be the case, then both in (i) and (ii) $v,\ cvc \in X^*$, a contradiction with Lemma 6.

Consider first case (i). Since $|v|$ is odd, necessarily $cv \in X^*$, so that $w = rsst$, where $h(r) = x$, $h(s) = cv$ and $h(t) = cy$. Hence, by the form of h, both t and s start with letter c. But then ssc is an overlapping factor of w.

In case (ii) accordingly $vc \in X^*$, so that $w = rsst$, where $h(r) = xc,\ h(s) = vc$ and $h(t) = y$. Now, again by the form of h, r and s end with c, showing that css is an overlapping factor of w.

So we have proved Lemma 7. □

Proof of Theorem 17 is now easy. Assume that α is not 2^+-free. Hence, by Theorem 16, it has an overlapping factor. This means that for some $i \geq 0,\ h^i(a)$ has an overlapping factor as well. Then, by Lemma 7, also $h^{i-1}(a)$, and hence inductively a has an overlapping factor, a contradiction. It follows that α is 2^+-free. □

Corollary 1. *There exists a cube-free infinite word in the binary alphabet.*

By Example 1, Theorem 17 is optimal: Squares *cannot be avoided* in infinite binary words. On the other hand, as we show next, they can be avoided in infinite ternary words.

Let $A = \{a, b\}$ and $B = \{a, b, c\}$. Define a morphism $\delta : B^* \to A^*$ by

$$\delta : \begin{cases} a \mapsto abb \\ b \mapsto ab \\ c \mapsto a \end{cases} .$$

Then δ viewed as a mapping of B^ω is a bijection $B^\omega \to (a^+\{b, bb\})^\omega$: Clearly, since δ is injective. It is also surjective since any ω-word starting with a and not containing three consecutive b's is an image of some word over B under δ. In particular, any cube-free binary word starting with a is an image of the unique infinite word under δ. So there exists $\overline{\alpha} \in \{a, b, c\}^\omega$ such that

$$\delta(\overline{\alpha}) = \alpha,$$

where α is the Thue-Morse word.

Theorem 18 (Thue, 1906). *There exists an infinite 2-free word over a ternary alphabet.*

Proof. Let $\overline{\alpha}$ be the word defined above. if it would contain a square, say uu, as a factor, then $\alpha = \delta(\overline{\alpha})$ would contain $\delta(u)\delta(u)a$ as a factor. This, however, is overlapping, by the form of δ, proving the theorem. □

Example 3. The 2-free word $\overline{\alpha}$ constructed above starts as

$$\overline{\alpha} = abcacbabcbacabcacb... \quad .$$

As in the case of α, also $\overline{\alpha}$ can be defined as a fixed point of a morphism:

$$\overline{\alpha} = \lim_{i \to \infty} \varphi^i(a),$$

where

$$\varphi : \begin{cases} a \mapsto abc \\ b \mapsto ac \\ c \mapsto b \end{cases} .$$

□

The method of the proof of Theorem 17 has been used to prove many other results similar to Theorems 17 and 18. We give two such examples.

Example 4. We claim that the infinite word defined by iterating the morphism

$$h : \begin{cases} a \mapsto aba \\ b \mapsto abb \end{cases}$$

at point a is 3^--free. Let $\beta = \lim_{i \to \infty} h^i(a)$.

First, β is not k-free for any $k < 3$, since

i. aab is a factor of β; and

ii. any word of the form $uuu(\mathrm{suf}_1(u))^{-1}$ is mapped under h to a word of the same form.

Second to prove that β is 3-free, a crucial observation is that words of length two can be covered by aba and abb only in the following ways:

$$\overbrace{a\,\overbrace{a}} \quad , \quad \overbrace{a\,b} \quad , \quad \overbrace{\underbrace{b\,a}} \quad \text{and} \quad \overbrace{b\,b} \quad .$$

Using this to a prefix of length two of an assumed cube in β, it is straightforward to conclude that β contains smaller and smaller cubes, which is a contradiction. The details are left as an excercise. □

Example 5. Using the above ideas, but more complicated considerations, one can show that so-called *Fibonacci word* defined as

$$\gamma = \lim_{i\to\infty} F^i(a) = abaababaabaababaababaabaab...,$$

where F is so-called *Fibonacci morphism*

$$F: \begin{cases} a \mapsto ab \\ b \mapsto a \end{cases},$$

is $(2+\varphi)^-$-free, where φ is the golden number $\frac{1}{2}(\sqrt{5}+1)$. Note that here k is irrational. □

By Theorem 17 and Example 1, we know exactly which repetitions can be avoided in infinite binary words. These are repetitions of order greater than 2. This motivates the following definition. For each $n \geq 2$, the *repetitiveness treshold* in an n-letter alphabet is a number $T(n)$ satisfying:

i. There exists a $T(n)^+$-free infinite word over an n-letter alphabet;

ii. Each $T(n)$-free word over an n-letter alphabet is finite.

It follows that if $T(n)$ exists then it is rational, since for a irrational number r, the notions of the r-freeness and r^+-freeness coincide.

Example 6. By Theorem 18, $T(3) < 2$ (if exists). It indeed exists and is equal to $\frac{7}{4}$, since:

i. Any $\frac{7}{4}$-free word over a 3-letter alphabet is shorter than 39, and

ii. the infinite word defined by iterating the morphism

$$h: \begin{cases} a \mapsto abcacbcabcbacbcacba \\ b \mapsto bcabacabcacbacabacb \\ c \mapsto cabcbabcabacbabcbac \end{cases}$$

at point a is $\frac{7}{4}^+$-free! □

The other known values of $T(n)$ are as follows:

$\lvert A\rvert$	2	3	4	5	6	7	8	9	10	11
$T(n)$	2	$\frac{7}{4}$	$\frac{7}{5}$	$\frac{5}{4}$	$\frac{6}{5}$	$\frac{7}{6}$	$\frac{8}{7}$	$\frac{9}{10}$	$\frac{10}{11}$	$\frac{11}{12}$
max(n)	3	38	122	6	7	8	9	10	11	12

Here max(n) tells the length of the maximal $T(n)$-free word in an n- letter alphabet. It is conjectured that for $n \geq 12$ $T(n) = \frac{n}{n+1}$.

As we saw repetition-free infinite words can be constructed by iterating a morphism, and this is by far the most commonly used method to define those. However, not all repetition-free infinite words can be obtained this way - as we shall see in a moment.

Clearly, a *sufficient* condition for a prolongable morphism h to define a k-free (resp. k^+- or k^--free) infinite word, when iterated at a point a, is that it itself is *k-free* (resp. *k^+- or k^--free*) in the following sense:

whenever $w \in A^+$ is k-free (resp. k^+- or k^--free) so is $h(w)$. This leads to the floowing natural problem:

Given a morphism decide whether it is repetition-free of a certain kind.

We shall give a solution for a special case of this problem in our next result. But before that we state a few central results known on this problem.

I. It *is decidable* whether a given morphism is 2-free. This follows, for example, from the next characterization of 2-free morphisms (Crochemore TCS, vol. 18, No.2, 1982):

II. A morphism h is 2-free iff it is so on all words of length

$$t(h) = (M(h) - 3)/m(h),$$

where $M(h) = \max\{|h(a)||a \in A\}$ and $m(h) = \min\{|h(a)||a \in A\}$. Moreover, if $|A| = 3$, then the above number $t(h)$ can be replaced by a constant 5.

III. On the other hand, *it is not known* any algorithm to decide whether a given morphism h

- is 3-free, or
- more generally, k-free for a given $k \in \mathbb{N}$.

IV. Finally, 2^+-free morphisms over a binary alphabet are completely characterized: They are of the forms T^n or $T^n \circ \mu$, where T is a Thue-Morse morphism and μ is the permutation $a \mapsto b$ and $b \mapsto a$.

Next we prove.

Theorem 19. *A uniform morphism $h : A^* \to A^*$ is 2-free iff it is so on 2-free words of length at most 3.*

Proof. "$\Rightarrow$". Trivial.

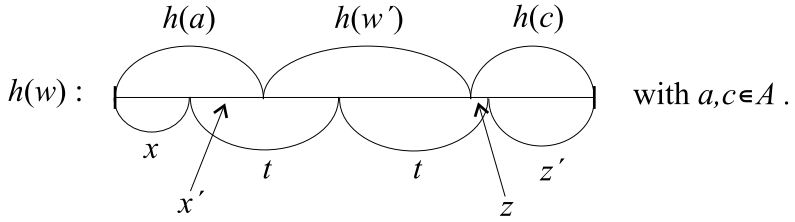
"$\Leftarrow$". Assume that h is uniform, i.e. $|h(a)| = |h(b)|$ for all $a, b \in A$, and that $h(w)$ is 2-free whenever $|w| \leq 3$ and w is 2-free. We first note that

- h is 1-free, and
- h is a prefix and a suffix (since otherwise we could have $h(ab) = h(a)h(b) = h(a)h(a)u$, for example).

Assume that w is 2-free and of the minimal length such that

$$h(w) = xttz'.$$

This can be illustrated as:



By our assumption $w' \neq 1$, and hence, by the uniformity, $|h(a)| < |xt|$ and $|h(c)| < |tz'|$ so that we can write

$$x'h(u)y = t = y'h(v)z \text{ where } h(b) = yy' \text{ with } b \in A.$$

There are two cases.

I. $|x'| = |y'|$. Now necessarily $x' = y'$, $h(u) = h(v)$ and $y = z$. Therefore

$$h(abc) = xx'yy'zz' = x(x'y)^2z'$$

meaning that abc is not 2-free, i.e. either $a = b$ or $b = c$. Further since $h(u) = h(v)$ and h is injective (since a prefix) necessarily $u = v$. All in all this means that

$$w = aubvc \text{ is either } auauc \text{ or } aubub,$$

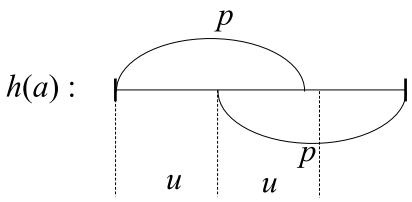
i.e. contains a square.

II. $x' = y'p$ with $p \neq 1$. Let $v = dv'$ with $d \in A$. Now we have

$$h(a) = xx' = xy'p, \text{ with } h(d) = pp'.$$

Consequently, $h(ad)$ contains a square, meaning that we must have $a = d$. This, in turn, means that the word $h(a)$ has p as a border, that is as a prefix and as a suffix.

Now, if $2|p| \geq |h(a)|$, $h(a)$ contains a square:



So necessarily $h(a) = pqp$ with $q \neq 1$. But now $q \leq h(ub)$ so that

$$h(aub) = pqpq... \ ,$$

which contradicts with the minimality of w.

So Theorem 19 has been proved. □

Note that Theorem 19 does not guarantee the existence of 2-free uniform morphisms. It only gives an easy criterium to test whether a given uniform morphism is 2-free.

Example 7. As an application of Theorem 19 it is not too complicated to verify that the morphism

$$h_6 : \begin{cases} a \mapsto abacabcacbabcbacabacbc \\ b \mapsto abacabcacbcabcbabcacbc \\ c \mapsto abacabcbabcacbabcbacbc \\ d \mapsto abacabcbabcacbcabcbabc \\ e \mapsto abacabcbacabacbabcacbc \\ f \mapsto abacabcbacbcacbabcacbc \end{cases}$$

is 2-free. Here $|h_6(a)| = 22$, and it can be shown that this is the smallest possible for uniform 2-free morphisms over a 6-letter alphabet! □

From the above considerations we obtain two interesting results.

Theorem 20. *For any alphabet A there exists a uniform 2-free morphism $h : \ A^* \to \{a, b, c\}^*$.*

Proof. For $n \leq 6$, the morphism h_6 restricted to an n-letter alphabet works although is not always minimal.

For $n = 12$ (and hence also for $n \leq 12$) the required h can be constructed as follows. Let

$$h_6 : \ A^* \to \{a, b, c\}^*$$
$$h'_6 : \ A'^* \to \{a', b', c'\}^*$$

be two copies of h_6 defined on disjoint alphabets. Define

$$g_{12} : \ (A \cup A')^* \to \{a, b, c, a', b', c'\}^*$$

by setting

$$g_{12}(a) = \begin{cases} h_6(a) \text{ if } a \in A \\ h'_6(a) \text{ if } a \in A'. \end{cases}$$

Now, since h_6 and h'_6 are square-free, so is g_{12}, and so is also $\overline{h}_6 \circ g_{12}$, where $\overline{h}_6$ is a copy of h_6 defined on $\{a, b, c, a', b', c'\}^*$. So we have constructed the required h for $n = 12 = 3 \cdot 2^2$.

Using the above procedure iterating we find the required h for any n of the form $3 \cdot 2^k$. □

To state the second result we need one notation. Let $\mathrm{SF}_k(n)$ denote the set of all 2-free words of length n over a k-letter alphabet. Here we allow n to be infinite. We have

Theorem 21. *i. There exists a constant $\alpha > 1$ such that*

$$|SF_3(n)| \geq (1/2)\alpha^n \quad \text{for all } n \geq 1.$$

ii. $SF_3(\infty)$ is nondenumerable.

Proof. (i). By Theorem 18, there exists a 2-free word w of a given length l. Let $\tau : \ \{a, b, c\}^* \to \{a, b, c, \overline{a}, \overline{b}, \overline{c}\}^*$ be a *finite substitution* defined as

$$\tau(x) = \overline{x} \text{ for } x \in \{a, b, c\}.$$

Then $\tau(w)$ contains 2^l different words, namely those obtained from w by adding "bars" into it in all possible ways. It is also clear that

$$\tau(w) \subseteq \mathrm{SF}_6(l).$$

Hence, also

$$h_6(\tau(w)) \subseteq \mathrm{SF}_3(22l),$$

where h_6 is the morphism of Example 7. Next we recall that since h_3 is 2-free it is injective even on $\{a, b, c, \overline{a}, \overline{b}, \overline{c}\}$. Consequently the number of square-free words over $\{a, b, c\}$ of length $n = 22l$ is at least

$$2^l = (2^{\frac{1}{22}})^n.$$

More generally, let $n \in [22(l-1), 22l]$. Then, by above

$$|\mathrm{SF}_3(n)| \geq 2^{l-1} = \frac{1}{2} \cdot (2^{\frac{1}{22}})^{22l} \geq \frac{1}{2} \cdot (2^{\frac{1}{22}})^{n}.$$

From this (i) follows directly.

(ii). This follows directly from the above proof after the following observations: After choosing a fixed infinite 2-free word w we have:

- $\tau(w)$ is nondenumerable,
- $\tau(w) \subseteq \mathrm{SF}_3(\infty)$,
- h_6 is 3-free also on infinite words,
- h_6 (as a prefix) is injective also on infinite words

Clearly, each of these points is true. □

Several remarks connected to the above considerations are in order.

Remark 1. By Theorem 21 (ii) "most" of 2-free infinite words cannot be obtained by the standard method of iterating a morphisms, since "the number of morphisms is denumerable"!!

Remark 2. The number α in Theorem 21 is just a bit more than 1, and actually $|\mathrm{SF}_3(n)|$ does not grow very rapidly. The smallest exact values are: 3,6,12,18,30,42,60,78,108,144,...

Remark 3. Theorem 21 can be modified for 3-free words over a binary alphabet. The proof is in principle exactly the same.

Remark 4. For 2^+-free words over a binary alphabet the situation is different:

- the number of such words of length n is $\mathcal{O}(n^2)$, but still
- the cardinality of such infinite words is nondenumerable.

Now we move to consider Abelian repetition-free words.

Theorem 22 (Dekking, 1979). *There exists an infinite Abelian 4-free word over a binary alphabet.*

Proof. The basic idea of the proof is as in Theorem 17. We shall prove that the infinite word ω defined by iterating the morphism

$$h: \begin{cases} a \mapsto abb \\ b \mapsto aaab \end{cases} \tag{4}$$

at point a is Abelian 4-free simply by showing that from a factor of ω which is Abelian 4th power we can construct a shorter similar factor. The fact that the consecutive blocks in the 4th power are only commutatively equal, and not equal, makes the proof much more complicated. As we shall see there are four, from (i) to (iv), special properties of h which are used in the following considerations.

First we associate with a word u its *value* in the group $\mathbb{Z}_5$ by a morphism $\mu: \{a,b\}^* \to \mathbb{Z}_5$ defined as

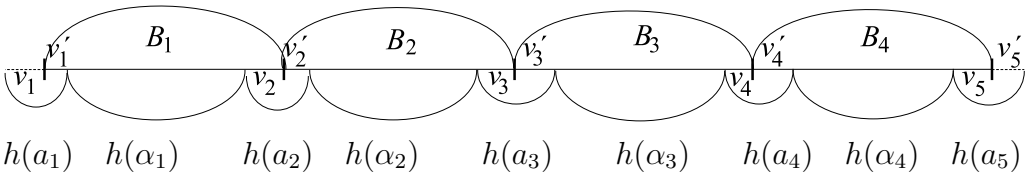
$$\mu(a) = 1 \text{ and } \mu(b) = 2.$$

This implies our first requirement for h namely that

(i) $\mu(h(w)) = 0$ for all $w \in \{a,b\}^*$,

which indeed is satisfied by our morphism (4).

Now, assume that $B_1B_2B_3B_4$ is an Abelian 4-repetition in ω. This together with the fact that these B_i's are covered by h-images is illustrated as follows:



Formally the above means that

$$h(a_1\alpha_1...\alpha_4a_5) = v_1B_1B_2B_3B_4v_5' \text{ with } a_i \in A \text{ and } \alpha_j \in A^*,$$

where for $i = 1,...,5$ and $j = 1,...,4$

$$h(a_i) = v_iv_i' \text{ and } B_j = v_j'h(\alpha_j)v_{j+1} \text{ with } v_i \in A^* \text{ and } v_i' \in A^+.$$

Recall now that μ is a morphism so that, by (i), we obtain

$$\begin{aligned} \mu(v_{j+1}) &= \mu(B_j) - \mu(h(\alpha_j)) - \mu(v_j') \\ &= \mu(B_j) + \mu(v_j) = g + \mu(v_j), \end{aligned}$$

where g is a constant since B_j's are commutatively equal. It follows that the sequence

$$\mu(v_1), \mu(v_2), \mu(v_3), \mu(v_4), \mu(v_5) \tag{5}$$

is an arithmetic progression of length 5. We want to allow only trivial such progressions. This guides us to require that

(ii) $S = \{a \in \mathbb{Z}_5 | \exists z \in \text{pref}\{h(a), h(b)\} : a = \mu(z)\}$

is *5-progression free,* i.e. does not contain an arithmetic progression of length 5, with $g \neq 0$. That our morphism satisfies this is easy to check:

$$\{\mu(a), \mu(ab)\} = \{1, 3\} \text{ and } \{\mu(a), \mu(aa)\mu(aaa)\} = \{1, 2, 3\} \tag{6}$$

so that $S = \{0, 1, 2, 3\}$, while in $\mathbb{Z}_5$ any arithmetic progression of length 5, with $g \neq 0$, equals to the whole $\mathbb{Z}_5$, as is easy to see.

Since v_i's in (5) are prefixes of $h(a)$ and $h(b)$ we can write (5) in the form

$$\mu(v_1) = \mu(v_2) = \mu(v_3) = \mu(v_4) = \mu(v_5).$$

We now return to Figure 1. We want that B_i's, possibly after a shift, would match with the $\{h(a), h(b)\}$-factorization of ω. This is achieved if either the words v_i or the words v_i' coincide. This motivates our next condition required for h and μ. We say that μ *is h-injective,* if for all factorizations $v_i v_i' \in \{h(a), h(b)\}$ with $i = 1, ..., 5$, we have

(iii) $\mu(v_1) = \mu(v_2) = ... = \mu(v_5) \Rightarrow v_1 = v_2 = ... = v_5$ or $v_1' = v_2' = ... = v_5'$.

From our computations in (6) we see that the only case to be checked here is the case when $v_1 = ab$ and $v_2 = aaa$, and then indeed $v_1' = b = v_2'$. So for our choice of μ and h μ is h-injective.

We are almost done. We know now that the words v_i or v_i' coincide. Consequently, the four Abelian repetitions B_i can be shifted to match with the morphism h: instead of B_i's we now consider the commutatively equal blocks

$$D_i = v_i B_i v_i^{-1} \text{ (or } D_i = {v'}_i^{-1} B_i v_i') \text{ for } i = 1, ...4.$$

Then there are words C_i such that

$$h(C_i) = D_i \text{ with } \Pi(D_i) = \Pi(D_j) \text{ for } i, j = 1, ...4, \tag{7}$$

where Π gives the commutative image of a word. If we would know that C_i's were commutatively equal the proof would be complete. Indeed, then ω would contain a shorter Abelian 4-repetition, and hence inductively also either *aaaa* or *bbbb* as a factor. This, however, is not the case.

So to complete the proof we impose one more requirement for h, namely that

(iv) $M(h) = \begin{pmatrix} |h(a)|_a & |h(a)|_b \\ |h(b)|_a & |h(b)|_b \end{pmatrix}$ *is invertible.*

Then, by (7), we have

$$\Pi(C_i) \cdot M(h) = \Pi(D_i) \text{ for } i = 1, ...4,$$

or equivalently that

$$\Pi(C_i) = \Pi(D_i) \cdot M(h)^{-1} \text{ for } i = 1, ..., 4.$$

This means that C_i's are commutatively equal. Finally, we note that our h indeed is invertible since $M(h) = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$.

So our proof is complete. □

We purposely pointed out the requirements (i)-(iv) in the above proof, since those can be used to prove - using $\mathbb{Z}_7$ instead of $\mathbb{Z}_5$ and the morphism

$$h : \begin{cases} a \mapsto aabc \\ b \mapsto bbc \\ c \mapsto acc \end{cases}$$

that there also exists an infinite Abelian 3-free word over a 3-letter alphabet.

The problem of whether there exists an infinite Abelian 2-free word over a 4-letter alphabet was for a long time a challenging open problem, until it was solved by Veikko Keränen: the answer is "yes" obtained by iterating a uniform morphism with $|h(a)| = 85$!!

Finally, we are ready to summarize the knowledge on the existence of repetition-free words of different lengths:

size of alphabet \ order of repetition	2	3	
2	3	∞	...
3	∞	...	...
	...	...	...

Abelian case	2	3	4	
2	3	9	∞	...
3	7	∞	...	...
4	∞	...	...	...
	...	...	...	...

Table 1. Maximal lengths of repetition-free words in different alphabets.

Next we consider briefly an extension of the repetition-freeness introduced by Bean, Ehrenfeucht and McNulty, namely *avoidability.* Let X be a set of unknowns and A our usual alphabet. A *pattern* is any nonempty word p in X^+. We say that a pattern p is

i. *avoidable in* A if there exists an infinite word $w \in A^\omega$ such that, for any morphism $h: X^+ \to A^+$, the word $h(p)$ is not a factor in w, and

ii. *unavoidable in* A otherwise.

Infinite (or finite) words w in (i) are said to *avoid the pattern* p .

Our next result, which is based on a simple general combinatorial trick, states the unavoidability in terms of finite words.

Lemma 8. *A pattern p is unavoidable iff for any morphism $h: X^+ \to A^+$ there exist only finitely many finite words avoiding the word $h(p)$.*

Proof. $\Leftarrow$. Trivial, since if p is avoidable there exists an infinite word w avoiding $h(p)$, for all h, and so do all finite prefixes of w.

$\Rightarrow$. Assume the contrary: There exist infinitely many finite words, say $w_1, w_2, w_3, ...$, avoiding $h(p)$ for some fixed morphism h. Now since A *is finite,* of the words w_i infinitely many start with a common letter, say $a = \alpha_1$. Next consider only those words w_i which starts with α_1. Repeating the argument we conclude that, for each $k \geq 1$, there exists a word α_k of length k such that

- infinitely many of words w_i start with α_k, and
- α_j is a prefix of α_k for $j \leq k$.

By the second fact

$$\alpha = \lim_{k \to \infty} \alpha_k$$

is well-defined, and by the first fact α avoids $h(p)$, showing that p is avoidable, a contradiction. □

Example 8. With the above notions we can reformulate our earlier results (cf. Example 1, Theorem 17, its corollary and Theorem 18) as follows:

- pattern xx is avoidable in a ternary alphabet, but not in a binary alphabet.
- pattern $xyxyx$ is avoidable in a binary alphabet.

Note that, as we formulated the above statements, the avoidability depends only on the cardinality of A, and not on A itself.

Example 9. By Example 8, the pattern xx separates the binary and ternary alphabets. Similarly - with quite a long considerations - one can show that the pattern

$$ABuDCvCAwBAzAC$$

separates the 3- and 4-letter alphabets, i.e. is avoidable in the latter, but not in the former.

Interestingly, it is not known any pattern separating larger alphabets!

Example 10. Assume that both X and A are binary, i.e. we are interested in which *binary patterns* are avoidable in a binary alphabet. This question is completely solved: In the following tree all inside nodes are unavoidable, while all leaves and all their right extensions are avoidable. Note, that in this sense the tree covers all patterns starting with x. The numbers in Table 2 tell the length of the longest binary word avoiding the considered pattern.

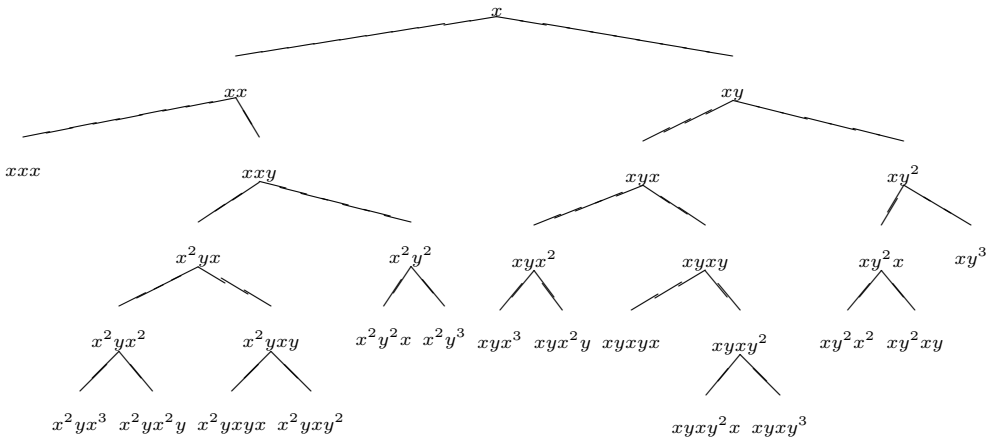


Figure 2. Tree characterizing avoidable binary patterns over a binary alphabet, cf. text.

p	x	xy	x^2	x^2y	xyx	x^2yx	xy^2x	x^2y^2	$xyxy$	x^2yx^2	x^2yxy
$\max(p)$	0	1	3	4	4	9	10	11	18	18	38

Table 2. The maximal lengths of word avoiding p.

The method of proving above is a standard one: In each avoidable case the required infinite word is constructed either by iterating a suitable morphism, or by mapping the infinite word defined by iterating a morphism by another morphism.

4 Applications of repetition-free words

In this section we consider a few applications of repetition-free words, as well as some related problems.

Application I. (Unending chess; Morse, Hedlund, 1943).
Let us consider the following problem (which was a motivation to rediscover the Thue-Morse word): Consider a game, like the chess, which *allows only finitely many different configurations;* in each move the game goes from on configuration to another. A game terminates

- when one of the players wins, or
- it is judged to be a draw.

If the draw is declared whenever a configuration repeats, then clearly each game terminates, i.e. is finite. Hence, under this rule of draw no infinite game is possible. Now, a natural question is: Under which "meaningfull" rules of draw (if any) a game can be infinite?

We use Thue-Morse word to describe such a situation. We set

Rule of draw: If after two identical sequences of configurations the game continues with the first move of these sequences, then the game is judged to be a draw.

We first emphasize that this rule is quite acceptable from practical points of view: repetitions of moves are allowed once, but "not a bit more". And in this case the game indeed can be infinite:

Theorem 23. *Under the above rule of draw there exist infinite games.*

Proof. Let α and β be two sequences of moves such that

i. both of those start and end at a same configuration;

ii. neither of those repeats any other configuration;

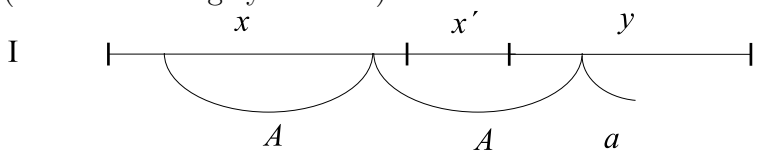
iii. α and β start with a different move.

We consider α and β as letters and denote by $w(\alpha, \beta)$ the infinite Thue-Morse word over $\{\alpha, \beta\}$. By (i), $w(\alpha, \beta)$ corresponds to a sequence of moves, i.e. an instance of the game.

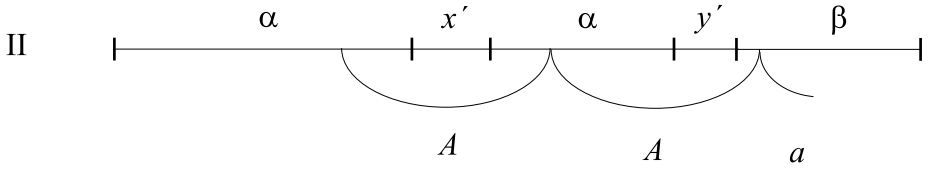
Claim. The moves corresponding to $w(\alpha, \beta)$ define an infinite game.

Proof of Claim. Of course we assume silently that neither of the players wins during α or β. Then the contra assumption is: $w(\alpha, \beta)$ contains a sequence of moves of the $AA\mathrm{pref}_1(A) = AAa$.

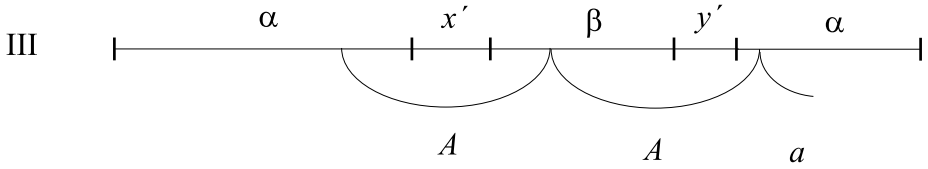
We first note that if any of the three configurations at the beginning or at the end of either of A's mathces with the end configuration of α or β, then, by (ii), so do all these three configurations. But then, by (iii), $w\{\alpha, \beta\}$ would not be 2^+-free. So we have to analyze the following three situations (the other being symmetric):



with $x, y \in \{\alpha, \beta\},\ x' \in \{\alpha, \beta\}^*$,



with $x', y' \in \{\alpha, \beta\}^*$,



with $x', y' \in \{\alpha, \beta\}^*$.

Case I. Impossible by (ii).

Case II. Again by (ii) the first occurrance of A both starts and ends at the same place inside α. Therefore if $|x'| > |y'|$ (resp. $|y'| > |x'|$), then inside β (resp. α) there would be a repetition of a configuration. So, by (i) and (ii), necessarily $|x'| = |y'|$, and hence $x' = y'$, implying that $\text{pref}_1(\alpha) = \text{pref}_1(\beta)$, a contradiction with (iii), or if $\beta = \alpha$ with 2^+-freeness of $w(\alpha, \beta)$.

Case III. We first note, as the consequence of (ii), that
"the prefix of the first occurrance of α up to the beginning of the first A"
coincides with
"the prefix of the second occurrance of α up to the end of the second A."
Consequently, we can shift the occurrance of the form AAa step by step to left starting from the beginning of α. During this process the end point of the first A inside β stays properly inside β, until at the very end when it matches with the beginning of β. Indeed, by (i) and (ii), it cannot hit the beginning of β before, and, by the same reason, it has to do so at the end of the process. So it follows that

$$\text{pref}_1(\alpha) = \text{pref}_1(A') = \text{pref}_1(\beta).$$

where A' is the new A obtained at the end of the process. This contradicts

with (iii). □

Note that in the above considerations we didn't specify the initial configuration of the game. Hence, some details remain to be fixed if we want to formalize the above in details for the chess, for example.

Application II (Burnside Problem). The *Burnside Problem for semigroups* asks:

Is a finitely generated semigroup, all elements of which generate finite semigroups, itself finite?

Actually, Burnside formulated this problem in 1900 for groups. We formulated it for semigroups, since repetition-free words give a very simple and clear solution in the semigroup case. The answer is "no" both for semigroups and groups.

We recall that a *semigroup* S is any set provided with an associate operation product $\cdot$. For any subset $F \subseteq S$ we can define a subsemigroup of S, so-called subsemigroup of S *generated* by F, by taking all finite products of F:

$$\langle F \rangle = F^* = \{f_1 \cdots f_n \mid n_{\geq} 1,\ f_i \in F\}.$$

The subsemigroup of S generated by a single element a is

$$\langle a \rangle = \{a^i \mid i \geq 1\}.$$

Theorem 24. *The Burnside Problem for semigroups has a negative answer.*

Proof. Let $A = \{a, b, c\}$. Denote by SF the set of all square-free words of A so that

$$A^+ \setminus SF = \bigcup_{x \in A^+} A^* x x A^*.$$

We introduce a new element, denoted by 0, and define a semigroup

$$S = (SF \cup \{0\}\ ;\ \cdot),$$

where the product is defined as follows: for all $\alpha, \beta \in SF$ we set

$$\alpha \cdot \beta = \begin{cases} \alpha\beta & \text{if } \alpha\beta \in SF \\ 0 & \text{otherwise,} \end{cases}$$
$$\alpha \cdot 0 = 0 \cdot \alpha = 0 \text{ and } 0 \cdot 0 = 0.$$

Clearly, the procudt is well-defined and associative - as soon as a product of words contains a square it becomes 0 in S. Note also that, as indicated, 0 is the zeroelement of S.

To complete the proof we note that

- S is finitely generated, since $\langle a, b, c\rangle = S$;
- each element $s \in S$ generates a finite subsemigroup, since $s^2 = 0$ always;
- S is infinite by Theorem 18.

□

Remark. The semigroup S in the proof of Theorem 24 was constructed from $A^+ = \{a, b, c\}^+$ by

- adjoining the zero into it to obtain $A_0^+ = A^+ \cup \{0\}$, and
- identifying all words containing a square into 0 to obtain S.

In more algebraic terms this means that

$$S = A^+ / \approx ,$$

where $\approx$ is the congruence generated by the relation $\sim$

$$xx \sim 0 \text{ for all } x \in A^+ \tag{1}$$

Intuitively, (1) identifies all squares to 0, and $\approx$ all words containing a square to 0.

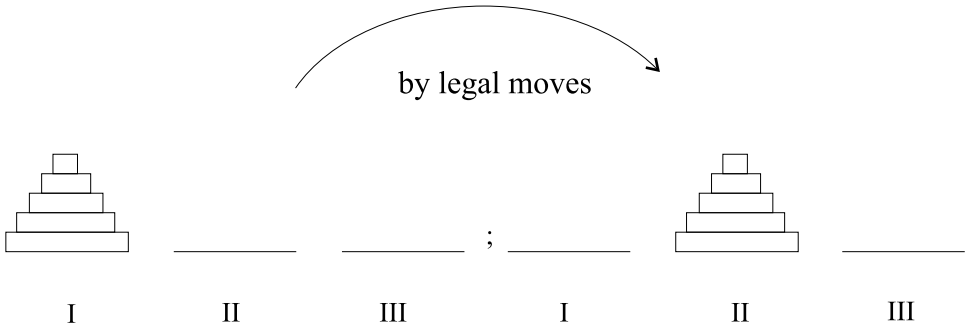
Application III (Tower of Hanoi Puzzle and square-free words).
Our goal is to show that the optimal solution of Tower of Hanoi Problem can be described as a word obtained by iterating a morphism, and moreover this word is square-free!

Recall that an instance of Tower of Hanoi Problem, THP for short, consist of

- N disks of different sizes, say $1, ..., N$,
- three sites 1, 2 and 3.

Initially, disks are in site 1 in decreasing order. A *move* consists of taking a topmost disk from one site and putting it to the top of another site. A move is *legal,* if it does not put a larger disk on the top of a smaller one. The goal of the puzzle is to describe a sequence of legal moves which moves the disks from the initial configuration to either site 2 or 3, and again in decreasing order.

The puzzle with $N = 5$ is illustrated as follows:



Tower of Hanoi Problem.

Clearly, at any moment there exist six different moves (all of which are not legal at any moment):

$$\begin{array}{llll} a: & \mathrm{I} \to \mathrm{II} & \overline{a}: & \mathrm{II} \to \mathrm{I} \\ b: & \mathrm{II} \to \mathrm{III} & \overline{b}: & \mathrm{III} \to \mathrm{II} \\ c: & \mathrm{III} \to \mathrm{I} & \overline{c}: & \mathrm{I} \to \mathrm{III} \end{array} \tag{2}$$

where the barred moves are *inverses* of the nobarred ones.

As is well known THP has a recursive solution

i. Move $N-1$ topmost disks to III using II;

ii. Move the disk N from I to II;

iii. Move $N-1$ disks from III to II using I.

Moreover, its complexity $\mathrm{T}(N)$, i.e. the number of moves, satisfies

$$\mathrm{T}(N) = 2\mathrm{T}(N-1) + 1,\ \mathrm{T}(1) = 1$$

so that

$$\mathrm{T}(N) = 2^N - 1.$$

Here we described a solution moving N disks from I to II using III as an auxilary site. Similarly, a solution moving the disks from I to III using II can be described.

The above solution, although very simple, is *optimal.* Indeed, any solution has to move the disk N, and in order to be able to do that all other disks must be in one site. Consequently, no solution can do better than (i) moving $N-1$ top disks to III using the optimal strategy, (ii) move N from I to

II, and (iii) again by the optimal way move $N-1$ disks from III to II. In particular, this holds for the optimal algorithm, so that if its complexity is $\mathrm{T}'(N)$ we have

$$\mathrm{T}'(N) \geq 2\mathrm{T}'(N-1)+1,\ \mathrm{T}'(1)=1,$$

yielding

$$\mathrm{T}'(N) \geq 2^N - 1,$$

as required.

In terms of our encoding (2) our optimal solution for $N=3$ is the word

$$a\overline{c}bac\overline{b}a.$$

To continue we define two morphisms $\sigma,\ \overline{\sigma}$ on $\{a,b,c,\overline{a},\overline{b},\overline{c}\}$ by

$$\begin{array}{llll} \sigma(a)=b & \sigma(\overline{a})=\overline{b} & & \overline{\sigma}(a)=c \quad \overline{\sigma}(\overline{a})=\overline{c} \\ \sigma(b)=c & \sigma(\overline{b})=\overline{c} & \text{and} & \overline{\sigma}(b)=a \quad \overline{\sigma}(\overline{b})=\overline{a} \\ \sigma(c)=a & \sigma(\overline{c})=\overline{a} & & \overline{\sigma}(c)=b \quad \overline{\sigma}(\overline{c})=\overline{b} \end{array}$$

Hence the morphism σ maps a symbol (either barred or unbarred) to the next one, and $\overline{\sigma}$ to the previous one, respectively.

From now on we fix our optimal solution to the unique one by considering only the solutions:

From I to II, if N is odd;

From I to III, if N is even.

With this convention let us denote by H_N, for $N \geq 1$, the sequence of moves in the optimal solution. Then we have

$$\begin{cases} H_{2N+1} = H_{2N}a\overline{\sigma}(H_{2N}) & \text{for } N \geq 0, \\ H_{2N} = H_{2N-1}\overline{c}\sigma(H_{2N-1}) & \text{for } N \geq 1. \end{cases} \tag{3}$$

Therefore we compute

$$H_1 = a,\ H_2 = a\overline{c}b,\ H_3 = a\overline{c}bac\overline{b}a, ...$$

and moreover,

$$H_j \leq H_{j+1} \quad \text{for all } j,$$

so that

$$H = \lim_{j\to\infty} H_j = a\overline{c}bac\overline{b}a\overline{c}b\overline{a}cba\overline{c}b...$$

exists. This was the reason why we chose the optimal solutions such as we did.

Next, for $x \in \{a, b, c\}$, we denote by X either x or $\overline{x}$. Then we can express H as a kind of "pseudoperiodic" word modulo 6, that is we claim that

$$H = (aCbAcB)^{\omega}. \tag{4}$$

This indeed follows from the formulas

$$H_{2N+1} = (aCbAcB)^{(2^{2N+1}-2)/6}a \quad \text{for } N \geq 0$$

and

$$H_{2N} = (aCbAcB)^{(2^{2N}-4)/6}aCb \quad \text{for } N \geq 1,$$

which, in turn, follows from (3) by induction (Exc.).

Now we turn to show how H is obtained as a *fixed point* by iterating a morphism. We define a uniform morphism φ on the alphabet $\{a, b, c, \overline{a}, \overline{b}, \overline{c}\}$ by

$$\varphi \begin{cases} a \mapsto a\overline{c} \\ b \mapsto c\overline{b} \\ c \mapsto b\overline{a} \end{cases} \quad \text{and} \quad \begin{cases} \overline{a} \mapsto ac \\ \overline{b} \mapsto cb \\ \overline{c} \mapsto ba \end{cases} .$$

Then we compute

$$\varphi(H_0) = 1,\ \varphi(H_1) = a\overline{c},\ \varphi(H_2) = a\overline{c}bac\overline{b},\ \varphi(H_3) = a\overline{c}bac\overline{b}a\overline{c}b\overline{a}cba\overline{c}, ...$$

implying that

$$\begin{aligned} \varphi(H_0)a = H_1, &\quad \varphi(H_1)b = H_2 \\ \varphi(H_2)a = H_3, &\quad \varphi(H_3)b = H_4. \end{aligned}$$

The above guides us to guess:

Lemma 9. *For each $i \geq 0$, we have*

$$\varphi(H_{2i})a = H_{2i+1} \quad \textit{and} \quad \varphi(H_{2i+1})b = H_{2i+2}.$$

Proof. We first note, using our earlier notations, that for all $w \in \{a, b, c, \overline{a}, \overline{b}, \overline{c}\}^*$

$$\varphi\sigma(w) = \overline{\sigma}\varphi(w)$$

and

$$\varphi\overline{\sigma}(w) = \sigma\varphi(w).$$

Since each of the mappings involved is a morphism it is enough to show these formulas for letters, for example

$$\varphi\sigma(b) = \varphi(c) = b\overline{a} = \overline{\sigma}(c\overline{b}) = \overline{\sigma}\varphi(b).$$

Now, the lemma is proved by induction on i. For $i = 0$ we already verified it. For $i > 0$ we compute

$$\begin{aligned}\varphi(H_{2i})a &\overset{(3)}{=} \varphi(H_{2i-1}\overline{c}\sigma(H_{2i-1}))a \\ &= \varphi(H_{2i-1})ba\varphi(\sigma(H_{2i-1}))a \\ &\underset{\text{above}}{\overset{\text{by}}{=}} \varphi(H_{2i-1})ba\overline{\sigma}(\varphi(H_{2i-1})b \\ &\overset{\text{i.h.}}{=} H_{2i}a\overline{\sigma}(H_{2i}) \overset{(3)}{=} H_{2i+1}.\end{aligned}$$

A similar computation shows the other formula. □

Now we are ready for

Theorem 25. *The solution H is obtained by iterating φ at a, that is $H = \lim_{i\to\infty}\varphi^i(a)$.*

Proof. By Lemma 9, for each $i \geq 0$, φ maps any prefix of H_i to a prefix of H_{i+1}. Consequently, since $a \leq H_1$ we have $\varphi^i(a) \leq H_{i+1}$ for all i. So the theorem follows when i tends to ∞. □

We turn to show that our optimal solution for THP is square-free, that is does not contain any repetition (of any length) of legal moves. This result becomes "provable" by Theorem 25, that is by the fact that the solution can be obtained by iterating a morphism.

Theorem 26. *Our optimal solution for THP is square-free.*

Proof. First we recall some terminology fixed above:

$$H = h_0h_1h_2... = \lim_{i\to\infty}\varphi^i(a) \in \{a, b, c, \overline{a}, \overline{b}, \overline{c}\}^\omega, \tag{5}$$

where φ is defined on page 57. We start with three simple claims:

Claim I. Ignoring the bars H is periodic modulo 3, namely $(abc)^\omega$

Claim II. For each $i \geq 0$,

$$h_i = \begin{cases} a & \text{if } i \equiv 0 \pmod 6 \\ b & \text{if } i \equiv 2 \pmod 6 \\ c & \text{if } i \equiv 4 \pmod 6 \end{cases}.$$

In particular, symbols in even positions of H are unbarred.

Claim III. H does not contain 4 consecutive unbarred symbols.

Proofs of claims. Claims I and II follow directly from (4). To prove Claim III assume the contrary that H contains 4 consecutive unbarred symbols, say $h_i h_{i+1} h_{i+2} h_{i+3}$. We have two cases:

"i is even": Set $i = 2k$. Then, by (5) and the fact that h is uniform, we have

$$\varphi(h_k h_{k+1}) = h_i h_{i+1} h_{i+2} h_{i+3}.$$

Since only images of barred letters under φ are unbarred it follows that h_k and h_{k+1} are barred, a contradiction with Claim II.

"i is odd": Now, by Claim II, h_{i+1} is unbarred, so that we can repeat the argument of first case for the word $h_{i-1} h_i h_{i+1} h_{i+2}$.

Proof of Theorem 26 (continued). As usual in such proofs assume that H contains a square xx. Without loss of generality we may assume that $|x|$ is of minimal length, and that there exists no squares in H of length $|xx|$ starting earlier than xx.

Let $x = x_1 ... x_n$ with each x_i being a letter. We have three cases:

Case 1. n is odd. Then, by Claim II, the whole x is unbarred. This contradicts with Claim III if $n \geq 3$, and with Claim I and (3) if $n = 1$.

Case 2. n is even, and xx starts at an even position in H, i.e. $n = 2m$ and

$$xx = h_{2j} h_{2j+1} \dots h_{2j+2m-1} h_{2j+2m} \dots h_{2j+4m-1}.$$

Then, as in the proof of Claim III, we have

$$\varphi(h_j \dots h_{j+m-1} h_{j+m} \dots h_{j+2m-1}) = xx.$$

Now, by the uniformity of φ, $\varphi(h_j \dots h_{j+m-1}) = \varphi(h_{j+m} \dots h_{j+2m-1}) = x$, and so by the injectivity of φ, we have $h_j \dots h_{j+m-1} = h_{j+m} \dots h_{j+2m-1}$, a contradiction with the minimality of x.

Case 3. n is even, and xx starts at an odd position in H. Now we can write

$$qxx = q h_{2j+1} \dots h_{2j+n} h_{2j+n+1} \dots h_{2j+2n},$$

where $q = h_{2j}$. From Claim I we conclude that $n \equiv 0 \pmod 3$, so that, by Claim II, $q = h_{2j+2n}$, and hence also $q = h_{2j+n}$. Therefore $(qx(q^{-1}))^2$ is a square of length $|xx|$ and occurring in H earlier than xx, a contradiction.

So our proof is complete. □

Application IV. (Free idempotent semigroup). Actually, this is not an application of the repetition-freeness, but rather related consideration to Application II. There, as a solution to the Burnside Problem, we defined (algebraicly, cf. Remark on p. 54) the semigroup

$$S = A^+ \cup \{0\} / \approx_s$$

where $\approx_s$ was the congruence generated by the relation

$$xx \sim_s 0 \quad \forall x \in A^+.$$

Now we start from the *idempotency relation*

$$xx \sim_i x \quad \forall x \in A^*$$

and denote by $\approx_i$ the congruence it generates. Then we define the semigroup

$$\mathfrak{I} = A^*/\approx_i$$

referred to as the *free idempotent* semigroup generated by A. This was the algebraic definition. However, as in the case of Burnside Problem we define $\mathfrak{I}$ in a more intuitive way as follows: Let A be a finite alphabet. We say that two words u and v are equivalent if we can derive v from u, and vice versa, by a finite (possibly 0) number of applications of the rules

- replace a factor x by its square xx, or
- replace a square factor xx by the word x.

Clearly, this relation, say $\equiv_i$, is an equivalence relation. Moreover, the operation defined on these equivalence classes by

$$[u][v] = [uv]$$

is well-defined, so that the set of equivalence classes of this relation forms a semigroup, which is isomorphic to the above $\mathfrak{I}$ (as is clear for those who knows more about the semigroup theory).

We keep the notation $\mathfrak{I}$ for this semigroup. According to our intuitive definition it consists of all words over A with the operation of catenation, but words which can be transformed to each other by applying *rewrite rules* $xx \to x$ and $x \to xx$ for factors are identified. Recall that in the semigroup S we identified all nonsquare-free words with the zero element.

Example 1. We claim that $bacbcabc = x \equiv_i y = bacabc$. We first compute

$$\begin{gathered} uy = abcaca.bacabc \equiv_i abcacabc \equiv_i abcabc \equiv_i abc, \\ x = bacbcabc \equiv_i bacbcuy = bacbcabcacay = vy, \end{gathered}$$

and

$$\begin{aligned} xr &= bacbcabc.bcabacbcacbcbac \\ &\equiv_i bacbcabacbcacbcbac \\ &\equiv_i bacbcacbcbac \\ &\equiv_i bacbcbac \equiv_i bacbac \equiv_i bac. \end{aligned}$$

So that

$$y = bacabc \equiv_i xrabc = xs,$$

and we can conclude

$$x \equiv_i vy \equiv_i vyy \equiv_i xy \equiv_i xxs \equiv_i xs \equiv_i y.$$

Note that the numbers of rewriting steps above are 3,1,3,5,1 and 5, so altogether 18, and the longest word encountered in these steps is xxs, which is of length $34 \geq 4\max\{|x|, |y|\}$. □

As a contrast to the Burnside semigroup S we prove

Theorem 27. *For any finite alphabet A the free idempotent semigroup generated by A, say $\mathfrak{I}$, is finite.*

Proof. Recall that $Alph(w)$ denotes the set of all letters appearing in w. We also note immediately

$$x \equiv_i y \quad \Rightarrow \quad Alph(x) = Alph(y).$$

Claim I. If $Alph(y) \subseteq Alph(x)$, then $\exists u : \ x \equiv_i xyu$.

Proof is by induction on $|y|$. If $|y| = 0$, then $x \equiv_i x$ (or xx). Consider $y = y'a$ with $a \in A$. By induction hypothesis there exists u' such that

$$x \equiv_i xy'u'.$$

Now, by the assumption, a is in $Alph(x)$, i.e. we can write $x = zaz'$. Then we can choose $u = z'y'u'$:

$$xyu = zaz'y'az'y'u' \equiv_i zaz'y'u' = xy'u' \overset{\text{i.h.}}{\equiv}_i x,$$

proving the claim.

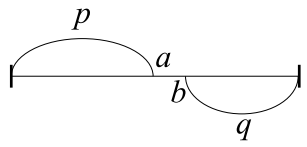
Next with each $x \in A^*$ we associate a quadruple

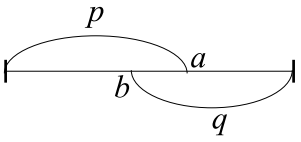
$$x\widehat{=}(p, a, b, q) \tag{6}$$

by requiring that

$$Alph(p)\dot{\cup}\{a\} = A = Alph(q)\dot{\cup}\{b\}, \tag{7}$$

where $\dot{\cup}$ denotes the disjoint union, and p is a prefix of x and q is a suffix of x:

X: 

or 

with (7).

Clearly, the words pa and bq are the *shortest* prefix and suffix of x, respectively, such that they contain all letters of x.

Claim II. if $x \widehat{=} (p, a, b, q)$, then $x \equiv_i pabq = \hat{x}$.
To prove Claim II let $x = pay = zbq$. Then we have $Alph(y) \subseteq Alph(pa) = Alph(x)$, so that, by Claim I,

$$pa \equiv_i payu \quad \text{for some } u.$$

Similarly using the inclusion $Alph(pa) \subseteq Alph(bq)$ and the dual form of Claim I we obtain

$$bq \equiv_i vpabq \quad \text{for some } v.$$

These identities allow to compute:

$$\hat{x} = pabq \equiv_i payubq = xubq = xw$$

and

$$x = zbq \equiv_i zvpabq = zv\hat{x} = t\hat{x}.$$

Now we can complete the proof of Claim II:

$$x \equiv_i t\hat{x} \equiv_i t\hat{x}\hat{x} \equiv_i x\hat{x} \equiv_i xxw \equiv_i xw \equiv_i \hat{x}.$$

Now, we are ready to finish the proof of Theorem 27. This is done by induction on $|A|$.

If $|A| = 1$, then any two nonempty words are equivalent, implying that $\mathfrak{I}$ consists of two elements: $\{1\}$ and $A^* \setminus \{1\}$. The induction step follows from the representation (6) and Claim II: Each word x is equivalent to $\hat{x} = (p, a, b, q)$, where p and q are over a proper subalphabet of x. Hence there exist only finitely many nonequivalent p's and q's, respectively, and so also only finitely many nonequivalent x's. □

Theorem 27 deserves two remarks.

Remark 1. Based on the fact (which is not difficult to prove) that for equivalent words x and x' their representations (6) $\hat{x} = (p, a, b, q)$ and $\hat{x}' = (p', a', b', q')$ satisfy $p \equiv_i p',\ a = a',\ b = b'$ and $q \equiv_i q'$ it is possible to prove that

$$|\mathfrak{I}| = \sum_{k=0}^{n} \binom{n}{k} c_k, \text{ with } c_k = \prod_{i=1}^{k} (k - i + 1)^{2^i},$$

where $n = |A|$, and in particular, for $|A| = 0$, we define $|\mathcal{I}| = 1$. These numbers grow very fast: 1,2,7,160,332381,...

Remark 2. Our considerations in Application IV are connected to the notion of "rewriting a word under certain rules", cf. Example 1. Here the rules were that x can be replaced by xx, and vice versa. Example 1 indicated that is not easy to check whether two words are equivalent under certain rewriting rules. As an illustration let us mention so-called *word problem* for finitely generated semigroups. Assume that S is a semigroup generated by a finite set F and satisfying relations $u_i = v_i$, for $i \in I$, where $u_i, v_i \in F^*$. Now, the word problem for S asks to decide whether two elements of S, say $x = s_1 \dots s_n$ and $y = t_1 \dots t_m$ with $s_i, t_j \in F^*$, are the same element in S, that is whether x can be transformed to y by using the rewriting rules $u_i \to v_i$ and $v_i \to u_i$, for $i \in I$, a finite number of times. One of the important results is that the word problem (with a finite I) is algorithmically *undecidable.*

In our considerations of free idempotent semigroups we started from the relation $x \sim_i xx$, or from the rewriting rules $x \to xx$ and $xx \to x$. An equally natural starting relation would be

$$x^2 \sim x^3 \quad \text{for all } x \in A^*,$$

or in terms of rewriting rules

$$x^2 \to x^3 \text{ and } x^3 \to x^2 \quad \text{for all } x \in A^*.$$

As earlier, we can define a semigroup, where equivalent words under these rules are identified. Is the semigroup finite or not?

5 Free monoids and semigroups

We defined free monoids and semigroups already on page 1. In what follows we concentrate with the monoid case. The results are easy to modify for semigroups.

We recall that a monoid M is *free* if it has a subset $B \subseteq M$ such that

i. $M = B^*$, and

ii. For all $n, m \geq 0$ and $x_1, \ldots, x_n, y_1, \ldots, y_m \in B$ we have:

$$x_1 \ldots x_n = y_1 \ldots y_m \;\Rightarrow\; n = m \text{ and } x_i = y_i \text{ for } i = 1, \ldots n.$$

Condition (i) means that B *generates* M, i.e. is a *generating set* of M, and condition (ii) requires that each element of M has the *unique* representation as a product of elements of B. The subset B of M satisfying (i) and (ii) is called a *base* of S. Note that the identity 1 of M is never in the base.

The goal of this section is to consider first some general properties of free monoids (in the spirit of semigroup theory), and then concentrate on free monoids of words, i.e. free submonoids of A^* (in the spirit of combinatorics of words).

Let us start with a few examples.

Example 1. For any alphabet A (not only for finite ones) the monoid A^* (or the semigroup A^+) is free with the base A. These are called *the free monoid and semigroup generated by A.*

Example 2. Let $X = \{a, ab, ba\}$, $Y = \{a, ab, bb\}$ and $Z = \{a, ab, b\}$. Then X^*, Y^* and Z^* are clearly monoids, and also submonoids of A^*, and moreover

- X^* is not free since $a.ba = aba = ab.a$, *and* any subset of X^* generating it contains X;
- Y^* is free, since no word in Y^+ has two different Y-factorizations (which is easy to conclude by reading w from right to left);
- Z^* is free, since $Z^* = \{a, b\}^*$, but not freely generated by Z.

Example 3. For any words $x, y \in A^+$, with $\rho(x) \neq \rho(y)$, the monoid $\{x, y\}^*$ is free, since, by Theorem 3, words x and y cannot satisfy a nontrivial relation.

Let M be a free monoid with a finite or countable base B, and A an alphabet of the same cardinality than B. Then, by the definition of free monoids, any bijection $h : A \to B$ extends in a natural way to an isomorphism $A^* \to M$. So we have

Lemma 10. *Each free monoid with a finite or infinite base is isomorphic to a word monoid A^*.*

□

Lemma 10 does not mean that free monoids are met only in connection with words:

Example 4. We claim that the multiplicative monoid generated by nonnegative integer matrices

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

is free. Let M denote this monoid. Assume that

$$X_1 \dots X_n = Y_1 \dots Y_m \text{ with } X_i, Y_j \in \{A, B\}.$$

If $X_1 = Y_1$, then since both A and B are invertible (over $\mathbb{Q}$) we can cancel the first members of the product and obtain

$$X_2 \dots X_n = Y_2 \dots Y_m,$$

for which we can apply induction (on $n + m$) to conclude that $n = m$ and $X_i = Y_i$ for $i = 1, \dots, n$, as was to be proved.

So it remains the case $X_1 = A$ and $Y_1 = B$ (or symmetric one). Set

$$\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = X_2 \dots X_n \text{ and } \beta = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = Y_2 \dots Y_m.$$

Then

$$\begin{aligned} A\alpha = B\beta \quad &\Leftrightarrow \quad \begin{cases} a + c & = a' \\ c & = a' + c' \end{cases} \text{ and } \begin{cases} b + d & = b' \\ d & = b' + d' \end{cases} \\ &\Leftrightarrow \quad \begin{cases} a & = c' = 0 \\ c & = a' \end{cases} \text{ and } \begin{cases} b & = d' = 0 \\ d & = b' \end{cases}, \end{aligned}$$

where the latter equivalence is since α and β are with non-negative entries. Hence we concluded that $a = 0$, a contradiction since, by the form of A and B, no product of those contain the zero on the left upper corner. Hence the claim is proved. □

We need a few notions. Let M be a monoid and F its generating set. We say that F is a *minimal* generating set if no proper subset of F is a generating

set of M. Further an element x of M is called *indecomposable* or *atomic* if it cannot be expressed in the form $x = yz$ with $y, z \neq 1$. Finally, we say that M is a *monoid with length,* if there exists a mapping $lg : M \to \mathbb{N}_0$ such that

$$lg(xy) = lg(x) + lg(y) \text{ for all } x, y \in M,$$

and

$$lg(x) = 0 \text{ iff } x = 1,$$

where $\mathbb{N}_0$ is the additive monoid of nonnegative integers. We prove

Theorem 28. *Each monoid M with length has the unique minimal generating set consisting of indecomposable elements of M, that is the set $(M \setminus \{1\}) \setminus (M \setminus \{1\})^2$.*

Proof. Clearly, the set of indecomposable elements of M coincides with $F = (M \setminus \{1\}) \setminus (M \setminus \{1\})^2$. Further, since, by definitions, each indecomposable element is in any generating set of M, F is the unique minimal generating set, if it is a generating set, that is if

$$F^* = M.$$

To prove this we first note that the inclusion $F^* \subseteq M$ is trivial. The reverse inclusion follows from the fact that all elements of M with the smallest length are - by the definitions - in F.

So consider $y \in M$ with $lg(y) = n + 1$. If $y \neq uv$ for all $u, v \in M \setminus \{1\}$, then y is indecomposable and so in F. Otherwise we write $y = uv$ with $u, v \in M \setminus \{1\}$. Then $1 \leq lg(u), lg(v) \leq n$, so that induction hypothesis apply: $u, v \in F^*$, so that also $y \in F^*$.

This completes the proof. □

Example 5. Consider the free monoid A^* generated by A. Obviously A^* is a monoid with length: we can set $lg(w) = |w|$. Moreover, any submonoid of A^*, say X^* with $X \subseteq A$, is a monoid with length. The length function of X^* can be chosen to be that of A^* restricted to X^*. It is worth noting that lg defined in this way is a morphism from A^* onto $\mathbb{N}_0$, and also from X^* into $\mathbb{N}_0$.

Example 5 implies a corollary to Theorem 28.

Corollary 1. *For each subset $X \subseteq A^*$ the monoid X^* has the unique minimal generating set $(X^+ \setminus \{1\}) \setminus (X^+ \setminus \{1\})^2$.*

As another general result we prove the following characterizatrion of free monoids.

Theorem 29 (Levi, 1940). *A monoid M is free if and only if it satisfies the following two conditions:*

i. There exists a morphism $h: M \to \mathbb{N}_0$ such that $h^{-1}(0) = \{1\}$;

ii. Whenever $u_1u_2 = u_3u_4$ with $u_1, u_2, u_3, u_4 \in M$, then one of the following conditions holds:

- *there exists $u_5 \in M$ such that $u_1 = u_3u_5$ and $u_5u_2 = u_4$, or*
- *there exists $u_6 \in M$ such that $u_1u_6 = u_3$ and $u_2 = u_6u_4$.*

Proof. $\Rightarrow$: Assume that M is free with the base B. We define the morphism $lg: M \to \mathbb{N}_0$ by the condition

$lg(x)$=the number of elements of B in the unique representation of x as the product of elements of B.

Obviously lg is a well-defined morphism, and further $lg^{-1}(0) = \{1\}$. Now assume that $u_1u_2 = u_3u_4$ with $u_i \in M$. We write u_i's as products of elements of B: $u_1 = \alpha_1 \dots \alpha_p$, $u_2 = \beta_1 \dots \beta_q$, $u_3 = \gamma_1 \dots \gamma_s$ and $u_4 = \delta_1 \dots \delta_r$. Then we have

$$\alpha_1 \dots \alpha_p\beta_1 \dots \beta_q = \gamma_1 \dots \gamma_s\delta_1 \dots \delta_r.$$

Since B is the base of M the requirements (ii) follow: indeed $p + q = s + r$ and corresponding elements of both sides are equal, so that, for example, if $p \geq s$ we can choose $u_5 = \alpha_{s+1} \dots \alpha_p$.

$\Leftarrow$: Now we assume that M satisfies the conditions (i) and (ii). Then, by (i), M is a monoid with length. Set

$$B = (M \setminus \{1\}) \setminus (M \setminus \{1\})^2,$$

so that B is exactly the set of indecomposable elements of M, and moreover $B^* = M$. We have to show that B is a free generating set of M, i.e. the implication

$$x_1 \dots x_n = y_1 \dots y_m \text{ with } x_i, y_j \in B \;\Rightarrow\; n = m \text{ and } x_i = y_i \text{ for } i = 1, \dots, n.$$

We may assume that $n \leq m$, and will prove the implication by induction on n. The case $n = 1$ follows from the definition of B. Assuming that the implication for $n < k$ holds we consider the relation

$$x_1 \dots x_k = y_1 \dots y_m \text{ with } x_i, y_j \in B \text{ and } k \leq m.$$

Writing this in the from $x_1(x_2 \dots x_k) = y_1(y_2 \dots y_m)$ we can conclude from (ii) the existence of u such that either $x_1 = y_1u$ and $ux_2 \dots x_k = y_2 \dots y_m$ or $x_1u = y_1$ and $x_2 \dots x_k = uy_2 \dots y_m$. But since x_1 and y_1 are indecomposable

necessarily $u = 1$, so that $x_1 = y_1$ and $x_2 \dots x_k = y_2 \dots y_m$. Now induction hypothesis applies and we conclude that $n = m$ and $x_i = y_i$ for $i = 1, \dots, n$, as was to be proved. □

Note that our basic result on words, namely Theorem 2, is a weak variant of Theorem 29!

Now we turn to consider submonoids of a word monoid A^*, which are called *F-semigroups,* and in particular free submonoids of A^*. We adjust our terminology more traditional to combinatorics of words.

We say that a subset $X \subseteq A^*$ is a *code* if it satisfies the following condition: For all $n, m \geq 1$ and $x_1, \dots, x_n, y_1, \dots, y_m \in X$

$$x_1 \dots x_n = y_1 \dots y_m \;\Rightarrow\; n = m \text{ and } x_i = y_i \text{ for } i = 1, \dots, n. \qquad (1)$$

Note that (1) is just a reformulation of condition (ii) in the definition of free monoids. Usually (1) is referred as the *decoding condition.* Not also that the empty word 1 is never in a code.

Example 6. There is no need to require that a code is finite. Indeed, the set

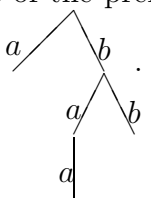
$$C_\infty = \{a^i b \mid i \geq 1\} \subseteq \{a, b\}^*$$

clearly satisfies condition (1), and is thus a code. □

Example 7. There are three important special cases of codes. A set X is a *prefix, suffix* or *bifix* if

- none of the words of X is a prefix is of another,
- none of the words of X is a suffix of another, and
- none of the words of X is either a prefix or a suffix

of another word in X, respectively.

Clearly, each of these special sets satisfies the decoding condition and is thus a code. Note also that prefix codes can be illustrated as trees: paths of the tree from the root to leaves correspond the elements of the prefix code; for example the code $X = \{a, baa, bb\}$ is illustrated as 

.

Next simple result gives connections of codes and free monoids.

Theorem 30. *Let $X \subseteq A^*$. Then the following conditions are equivalent:*

i. X is a code,

ii. X is a free generating set, or a base, of the monoid X^*,

iii. X^* is free and X is its minimal generating set.

Proof. (i)⇒(ii). So let X be a code. By definition of X^*, the set X generates X^*, and, by (1), it generates X^* freely.

(ii)⇒(iii). Now we assume that X generates X^* freely, i.e. X generates X^*, and each element of X^* can be expressed as a product of elements of X in the unique way. Then, by the definition of the freeness, X^* is free. So by Corollary 1 to Theorem 28 we have to show that

$$X = (X^+ \setminus \{1\}) \setminus (X^+ \setminus \{1\})^2. \tag{2}$$

Recall here that the right hand side coincides with the set of indecomposable elements of X^*. Remember also that since X is a free generating set it does not contain the empty word 1.

Now assume first that $w \in X$. Then $w \in X^+ \setminus \{1\}$, by above, and $w \notin (X^+ \setminus \{1\})^2$, since X is a free generating set. Second, if $w \in (X^+ \setminus \{1\}) \setminus (X^+ \setminus \{1\})^2$, that is, is indecomposable, then $w \in X$. Hence (2) holds.

(iii)⇒(i). Now we assume that X^* is free, and X is its minimal generating set, i.e., by Corollary 1 to Theorem 28,

$$X = (X^+ \setminus \{1\}) \setminus (X^+ \setminus \{1\})^2.$$

Since X^* is free it has a free generating set, say Y. Since Y generates X^*, $1 \notin Y$ and all elements of X are indecomposable, necessarily each w in X is also in Y, i.e. $X \subseteq Y$. So since Y satisfies decoding condition (1), as a free generating set, so does X. Hence X is a code, and our proof is complete. □

Remark. Assume that $X \subseteq A^*$ is a code. Then X^* is free with X as a base, and by Lemma 10, we conclude that there exists an alphabet B, with $|B| = |X|$, and an *injective* morphism $h_X : B^* \to A^*$ such that $h_X(B) = X$. Morphism h_X is called a *coding morphism* of X. The converse holds also: if $h : B^* \to A^*$ is an injective morphism then $X = h(B)$ is a code. We can say that h_X above *encodes* B^*, and that h_X^{-1} *decodes* X^*.

It follows that the *theory of codes is nothing but the theory of injective morphisms* $B^* \to A^*$. This theory is sometimes called the *theory of variable length codes* as a distinction to the *theory of error correcting codes.* The theory of codes was initiated by M.P.Schützenberger in 50's.

The use of coding morphisms yields easily.

Theorem 31. *Let $f: A^* \to C^*$ be an injective morphism. Then*

i. if $X \subseteq A^*$ is a code so is $f(X)$; and

ii. if $Y \subseteq C^*$ is a code so is $f^{-1}(Y)$.

Proof. (i). Let $h_X: B^* \to A^*$ be a coding morphism of X. Then the composition $f \circ h_X: B^* \to C^*$ is an injective morphism, and so

$$f \circ h_X(B) = f(X)$$

is a code.

(ii). Let $Y \subseteq C^*$ be a code and $X = f^{-1}(Y)$. Now if

$$x_1 \dots x_n = y_1 \dots y_m \text{ with } x_i, y_j \in X,$$

then

$$\alpha(x_1) \dots \alpha(x_n) = \alpha(y_1) \dots \alpha(y_m) \text{ with } \alpha(x_i), \alpha(y_j) \in Y,$$

so that, since Y is a code, necessarily $n = m$ and $\alpha(x_i) = \alpha(y_i)$ for $i = 1, \dots, n$. Hence X satisfies the decoding condition. □

Theorem 30 characterizes free submonoids of A^* in terms of codes, i.e. bases. The next theorem gives another characterization but without using bases. We need some terminology.

We call a submonoid M of A^* *stable, right unitary* or *left unitary* if for all words u, v and w:

$$\text{whenever} \quad u, v, uw, wv \in M, \text{ then } w \in M, \tag{3}$$

$$\text{whenever} \quad u, uv \in M, \text{ then } v \in M, \tag{4}$$

$$\text{whenever} \quad u, vu \in M, \text{ then } v \in M, \tag{5}$$

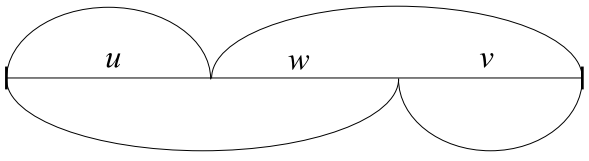
respectively. The implication (3) can be stated in the form

$$M^{-1}M \cap MM^{-1} \subseteq M,$$

or equivalently, since $1 \in M$, in the form

$$M^{-1}M \cap MM^{-1} = M.$$

Further the assumption of (3) can be illustrated as



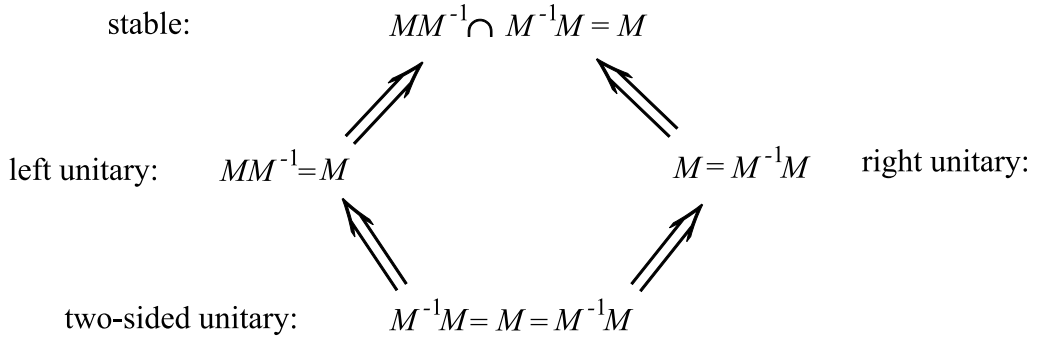
,

where an arc means that "belongs to M". And the conclusion is that w is in M.

Similarly, the conditions for the right and left unitary can be rewritten as

$$M^{-1}M = M \text{ and } MM^{-1} = M.$$

We note the following immediate connections of the above notions:



Now we are ready for the characterizations:

Theorem 32 (Schützenberger's criterium). *A submonoid of A^* is free if and only if it is stable.*

Proof. Let M be a stable submonoid of A^*. By Theorem 30 it is enough to show that the minimal generating set

$$X = (M^+ \setminus \{1\}) \setminus (M^+ \setminus \{1\})^2$$

of M is a code. Assume the contrary: a word w has two X-factorizations, i.e.

$$w = x_1 \dots x_n = y_1 \dots y_m \text{ with } x_i, y_j \in X,$$

where moreover, assuming w as short as possible $|x_1| < |y_1|$. Then we can write $y_1 = x_1 z$ with $z \in A^+$ so that

$$x_1, y_2 \dots y_m, x_1 z, z y_2 \dots y_m \in M.$$

Consecuently, by the stability of M, we have: $z \in M$. But then $y_1 = x_1 z \in X \cap (M \setminus \{1\})^2$, a contradiction.

To prove the other direction let M be free and X its base. Further assume that words u, v and w satisfy

$$u, v, uw, wv \in M,$$

as illustrated in Figure on page 70. We write these words as products of elements of the base X:

$$u = x_1 \dots x_k, \quad wv = x_{k+1} \dots x_r,$$
$$uw = y_1 \dots y_l, \quad v = y_{l+1} \dots y_s,$$

where $x_i, y_j \in X$. Since $u(wv) = (uw)v$ we can write

$$x_1 \dots x_k x_{k+1} \dots x_r = y_1 \dots y_l y_{l+1} \dots y_s.$$

Therefore, since X is a code, necessarily $r = s$ and $x_i = y_i$ for $i = 1, \dots, r$. Since $|uw| \geq |u|$, necessarily $l \geq k$ and we can write

$$uw = x_1 \dots x_k x_{k+1} \dots x_l = u x_{k+1} \dots x_l,$$

which implies that $w = x_{k+1} \dots x_l$, and hence in M.

So M is stable. □

The following result shows the power of Theorem 32.

Corollary 1. *Any intersection of free submonoids of A^* is free.*

Proof. Let M_i be a free submonoid of A^* for each $i \in I$. Consider

$$M = \bigcap_{i \in I} M_i.$$

Clearly, M is a submonoid of A^*: Indeed

- $1 \in M_i$ for all i, and hence $1 \in M$,
- if $m, m' \in M$, then $m, m' \in M_i$ for all i, and hence $mm' \in M$.

It is also free by Theorem 32:

$$\begin{aligned} u, v, uw, wv \in M \;&\Rightarrow\; u, v, uw, wv \in M_i \text{ for all } i \overset{\text{Thm } 32}{\Rightarrow} \\ w \in M_i \text{ for all } i \;&\Rightarrow\; w \in M. \end{aligned}$$

□

Example 8. Let $M_1 = \{aab, aba\}^*$ and $M_2 = \{a, baaba\}^*$. Then they are free since sets $\{aab, aba\}$ and $\{a, baaba\}$ are codes, in fact even prefixies. Consequently, $M_1 \cap M_2$ is free. However, it not finitely generated (cf. Exc), while M_1 and M_2 are so. □

Example 9. The monoid $M = \{w \in A^* \mid |w|_a \equiv 0 \pmod 2\}$ is both right and left unitary. Indeed:

$$u, uv \in M \;\Rightarrow\; |u|_a \text{ and } |uv|_a \text{ are even } \Rightarrow$$
$$|v|_a \text{ is even } \Rightarrow\; v \in M.$$

Further its minimal generating set is, assuming that $A = \{a, b\}$,

$$B = \{b\} \cup ab^*a.$$

This is a code as it must be by Theorems 30 and 32. In fact it is even a bifix as it must be by our next result. □

Theorem 32 has variants for right and left unitary monoids.

Theorem 33. *A submonoid of A^* is right (resp. left) unitary if and only if its minimal generating set is a prefix (resp. suffix). In particular, such monoids are free.*

Proof. $\Rightarrow$. Let M be a right unitary submonoid of A^* and B its minimal generating set. By Theorem 28

$$B = (M \setminus \{1\}) \setminus (M \setminus \{1\})^2.$$

Consider words u and v such that $u, uv \in B$. Hence $u, uv \in M$ and since M is right unitary necessarily $v \in M$. If $v \neq 1$, then $vu \in B \cap (M \setminus \{1\})^2$, a contradiction. Consequently $v = 1$, implying that B is a prefix.

$\Leftarrow$. Now we assume that, with the above notations, the minimal generating set B of M is a prefix, and consider two words $u, uv \in M = B^*$. We write

$$u = x_1 \dots x_n \text{ and } uv = y_1 \dots y_m \text{ with } x_i, y_j \in B,$$

and conclude the identity

$$x_1 \dots x_n v = y_1 \dots y_m.$$

Since B is a prefix necessarily $x_1 = y_1$, and so inductively $x_i = y_i$ for $i = 1, \dots, n$, and $v = y_{n+1} \dots y_m$. The last equality shows that $v \in B^* = M$, and therefore we have proved that M is right unitary.

The proof for left unitary submonoids is completely symmetric. □

As in the case of Theorem 32 we have also now an immediate consequence:

Corollary 1. *Any intersection of right (resp. left) unitary submonoids of A^* is right unitary (resp. left unitary).*

Example 8 (Continued). Actually, by Theorem 33, the monoids M_1 and M_2 in Example 8 are right unitary, i.e. their minimal generating sets are prefixies. Hence, the minimal generating set of the monoid $M_1 \cap M_2$ is a prefix, but as we claimed in Example 8, not finite. □

Corollaries of Theorems 32 and 33 show their algebraic usefulness. Even a better evidence of that is given in the following so-called *Defect Theorem.*

In order to formulate it we need some terminology. Let $X \subseteq A^*$ be a finite set. Define

$$FM(X) = \bigcap_{\substack{X \subseteq M \subseteq A^* \\ M \text{ is free}}} M.$$

Consequently, $FM(X)$ is the intersection of all free submonoids of A^* containing X. By Corollary of Theorem 32, $FM(X)$ is a free submonoid. Since it *contains* X, and *is contained* in any free submonoid of A^* containing X it is *the smallest free submonoid of* A^* *which contains* X. The minimal generating set of $FM(X)$ is called the *free hull* of X, and is denoted by $\widehat{X}$.

Similarly, using Corollary to Theorem 33, we can define the smallest right unitary monoid containing X, say $RUM(X)$, by the formula

$$RUM(X) = \bigcap_{\substack{X \subseteq M \subseteq A^* \\ M \text{ is right unitary}}} M.$$

Let us denote the minimal generating set of $RUM(X)$ by $\widehat{X}(p)$.

Now, we are ready to state Defect Theorem:

Theorem 34 (Defect Theorem). *For each finite set $X \subseteq A^*$ its free hull $\widehat{X}$ satisfies*

$$|\widehat{X}| \leq |X|,$$

and moreover, the equality holds if and only if X is a code.

Proof. If X is a code, then clearly the smallest free submonoid of A^* containing X is X^*, so that $\widehat{X} = X$. Therefore $|X| = |\widehat{X}|$.

Consequently, to prove Theorem 34 it is enough to show:

$$X \text{ is not a code } \Rightarrow \ |\widehat{X}| \leq |X| - 1. \tag{6}$$

This is the essential message of Theorem 34.

The proof of this is based on the following claim:

Claim. $\widehat{X} \subseteq X(\widehat{X}^*)^{-1} \cap (\widehat{X}^*)^{-1}X$, i.e. each word of the free hull $\widehat{X}$ occurs as the first (and as the last) factor of some word of X in its $\widehat{X}$-factorization.

Proof of Claim: Assume the contrary: there exists an $\widehat{x}$ such that $\widehat{x} \in \widehat{X} \setminus (\widehat{X}^*)^{-1}X$. The case $x \notin X(\widehat{X}^*)^{-1}$ is symmetric. It follows, as we already noted, that no word of X contains $\widehat{x}$ as the last factor in its $\widehat{X}$- factorization, in other words, that

$$X \subseteq \{1\} \cup \widehat{X}^*(\widehat{X} \setminus \{\widehat{x}\}). \tag{7}$$

We set

$$Z = \widehat{x}^*(\widehat{X} \setminus \{\widehat{x}\}).$$

Then

$$Z^+ = \widehat{X}^*(\widehat{X} \setminus \{\widehat{x}\}) \tag{8}$$

since:

$$\begin{aligned}\widehat{X}^*(\widehat{X} \setminus \{\widehat{x}\}) &= \{w \in \widehat{X}^* \mid w \text{ has an } \widehat{X}\text{-factorization such that the last} \\ &\qquad \text{factor is in } \widehat{X} \setminus \{\widehat{x}\}\} \\ &= (\widehat{x}^*(\widehat{X} \setminus \{\widehat{x}\}))^+ \\ &= Z^+.\end{aligned}$$

From (7) and (8) we obtain

$$X \subseteq Z^*, \tag{9}$$

i.e. the monoid Z^* contains X.

Next we show that

$$Z^* \text{ is free with the base } Z. \tag{10}$$

Now recall that $\widehat{X}$ is a base of a free monoid. Therefore each word $z \in Z^* \subseteq \widehat{X}^*$ can be written uniquely in the form

$$z = x_1 \dots x_n \text{ with } x_i \in \widehat{X} \text{ and } x_n \neq \widehat{x}. \tag{11}$$

Consequently, z has also the unique representation in the form:

$$z = \underbrace{\widehat{x}^{p_1} z_1}_{\in Z} \underbrace{\widehat{x}^{p_2} z_2}_{\in Z} \dots \underbrace{\widehat{x}^{p_r} z_r}_{\in Z} \text{ with } z_i \in \widehat{X} \setminus \{\widehat{x}\} \text{ and } p_i \geq 0.$$

It follows that Z is a free generating set, i.e. (10) holds true.

Now the proof of Claim is easy: Since $\widehat{X}$ is a code, $\widehat{x}$ cannot be written in form (11), and so $\widehat{x} \notin Z^*$. Therefore, by (9), we have:

$$X \subseteq Z^* \subset \widehat{X}^*, \text{ with } Z^* \text{ free.}$$

This contradicts with the fact that $\widehat{X}$ is the free hull.

Proof of Theorem 34 (continued).

Case 1. $1 \in X$. Then the monoids X^* and $(X \setminus \{1\})^*$ have the same free hull, so that this case is reduced to the other one.

Case 2. $1 \notin X$. We define the mapping $\alpha : X \to \widehat{X}$ by

$$\alpha(x) = \widehat{x} \text{ if } x \in \widehat{x}\widehat{X}^*.$$

Now

- the value of α is *always defined* since $X \subseteq \widehat{X}^*$;
- α is *well-defined* since $\widehat{X}$ is a code; and
- α is *surjective* by Claim.

Now to prove the implication (6) we finally use the fact that X is not a code. This means that some word has two X-factorizations:

$$x_1 \dots x_n = x'_1 \dots x'_m \text{ with } x_i, x'_j \in X \text{ and } x_1 \neq x'_1.$$

Hence from the definition of α we obtain:

$$\alpha(x_1)\widehat{X}^* \cap \alpha(x'_1)\widehat{X}^* \neq \emptyset.$$

Since $\widehat{X}$ is a code, this is possbile only if $\alpha(x_1) = \alpha(x'_1)$. This means that α is not injective. So from the surjectivity we conclude that $|\widehat{X}| = |\alpha(X)| \leq |X| - 1$, as was to be proved. □

Defect Theorem deserves a few remarks.

Remark 1. Our basic theorem of words, Theorem 3 is a consequence of Defect Theorem: If two words satisfy a nontrivial relation, i.e. the set is not a code, then they are powers of a same word, i.e. the free hull contains only one word.

Remark 2. Defect Theorem can be viewed as a weak *dimension property* of words: If a finite set X of words satisfies a nontrivial relation, i.e. "is dependent", then these words can be expressed as products of fewer than $|X|$ words, i.e. "belong to a smaller subspace". We consider these matters more in Chapter 6.

Remark 3. Our defect theorem was based on Corollary to Theorem 32. A similar result based on Corollary to Theorem 33 can be proved, where the "free hull" would be not only a code, but even a prefix. Also this is more considered in Chapter 6.

Next we turn to consider the important *problem of deciding whether a given set $X \subseteq A^*$ is a code,* i.e. a free generating set of X^*. Obviously X is a code if and only if

$$xX^* \cap yX^* = \emptyset \text{ for all } x, y \in X \text{ with } x \neq y. \tag{12}$$

Indeed, this condition is equivalent to the fact that no word in X^+ has two X-factorizations.

The theory of finite automata gives a simple (but unefficient) method to solve whether (12) holds for a finite set X: The problem is reduced to the emptiness problem of so-called regular languages.

We do not assume any knowledge of finite automata, but instead give a more combinatorial solution to this problem (which is actually essentially the same given by the automata theory). The idea of the proof is very simple: We search systematically for a word having a double X-factorization, if such a word is found X is not a code. Consequently, the problem is when can we stop the search if such a word does not exist.

Let $X \subseteq A^+$ be arbitrary. We define recursively

$$\begin{aligned} U_1 &= X^{-1}X \setminus \{1\}, \\ U_{n+1} &= X^{-1}U_n \cup U_n^{-1}X \quad \text{for } n \geq 1, \end{aligned}$$

and prove:

Theorem 35. *A set $X \subseteq A^+$ is a code if and only if none of the sets U_n contains the empty word* 1.

The proof of Theorem 35 is based on the following technical lemma.

Lemma 11. *Let X and U_n for $n \geq 1$ be as above. Then for all n and $k \leq n$ we have: $1 \in U_n$ if and only if there exist $u \in U_k$ and $i, j \leq n$ such that*

$$uX^i \cap X^j \neq \emptyset \text{ with } i + j + k = n. \tag{13}$$

Proof. The proof is by decreasing induction on k. Let $k = n$. Now if $1 \in U_n$ we can choose in (13) $u = 1$ and $i = j = 0$. On the other hand, if (13) holds, necessarily $i = j = 0$, and so $1 \in U_n$.

Ind. step. We consider a fixed value k, and assume that Lemma 11 holds for all *larger* values. We have to show that the equivalence of Lemma 11 holds for our fixed value of k.

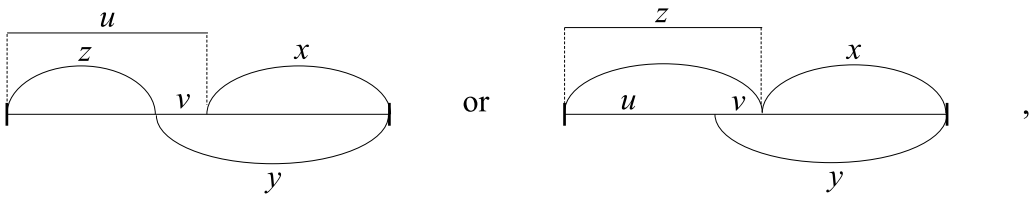
$\Rightarrow$: Assume that $1 \in U_k$. Then by induction hypothesis there exist $v \in U_{k+1}$ and $i, j \geq 0$ such that

$$vX^i \cap X^j \neq \emptyset \text{ with } i + j + k + 1 = n.$$

So we can write $vx = y$ for some $x \in X^i$ and $y \in X^j$. On the other hand, since $v \in U_{k+1}$ one of the following two cases holds:

$$zv = u \text{ with } z \in X \text{ and } u \in U_k, \text{ or}$$
$$z = uv \text{ with } z \in X \text{ and } u \in U_k.$$

These alternatives can be illustrated as:



,

where the arc means that "belongs to X^+".

In the first case $ux = zy$, meaning that

$$uX^i \cap X^{j+1} \neq \emptyset \text{ with } u \in U_k,$$

and in the second case $zx = uy$, implying that

$$uX^j \cap X^{i+1} \neq \emptyset \text{ with } u \in U_k.$$

Consequently (13) holds in both cases.

$\Leftarrow$: Now we assume that there exist $u \in U_k$ and $i, j \geq 0$ such that

$$uX^i \cap X^j \neq \emptyset \text{ with } i + j + k = n.$$

Now we write

$$ux_1 \dots x_i = y_1 \dots y_j \quad \text{for some } x_i, y_j \in X.$$

If $j = 0$ so is i. Therefore $k = n$, i.e. we are in the case considered at the beginning of the proof. Assuming $j \geq 1$ we have again two cases:

i. $u = y_1 v$ with $v \in A^*$. Then $v \in X^{-1}U_k \subseteq U_{k+1}$ and moreover

$$vx_1 \dots x_i = y_2 \dots y_j.$$

Consequently, $vX^i \cap X^{j-1} \neq \emptyset$, with $v \in U_{k+1}$, so that by induction hypothesis $1 \in U_k$.

ii. $y_1 = uv$ with $v \in A^*$. Now $v \in U_k^{-1}X \subseteq U_{k+1}$ and moreover

$$x_1 \dots x_i = vy_2 \dots y_j.$$

Hence in this case $vX^{j-1} \cap X^i \neq \emptyset$ with $v \in U_{k+1}$, so that again induction hypothesis yields $1 \in U_k$.

This completes the proof of Lemma 11. □

Proof of Theorem 35. Assume first that X *is not* a code. Then we can write

$$x_1 \dots x_p = y_1 \dots y_q \quad \text{for some } x_i, y_j \in X \text{ with } x_1 \neq y_1.$$

Assuming, by symmetry, that $|x_1| < |y_1|$ we can write $y_1 = x_1 u$ with $u \in A^+$. Consequently,

$$u \in U_1 \text{ and } uX^{p-1} \cap X^{q-1} \neq \emptyset.$$

Hence, by Lemma 11, $1 \in U_{p+q-1}$.

Second assume that $1 \in U_n$ for some n. We apply Lemma 11 for $k = 1$: there exist $u \in U_1$ and $i, j \leq n$ such that $uX^i \cap X^j \neq \emptyset$. Since $u \in U_1$ we can write $xu = y$ for some $x, y \in X$, and moreover $x \neq y$; otherwise U_1 would contain the empty word. It follows that $xuX^i \cap xX^j \neq \emptyset$, or equivalently that $yX^i \cap xX^j \neq \emptyset$. But then X is not a code. □

Note that the detailed proof of Theorem 35 is more complicated than the intuition behind it!

It is also worth noticing that for a *finite* X we have:

i. if $u \in U_n$, for some $n \geq 1$, then $|u| \leq \max\{|x| \mid x \in X\}$;

ii. if $U_i = U_j$, then, for any $t \geq 0$, $U_{i+t} = U_{j+t}$.

Condition (i) implies that there exist only finitely many different U_n sets. Condition (ii), in turn, guarantees that once a repetition in the sequence $U_1, U_2, \dots$ is found all U_i sets are found as well. Hence we conclude

Corollary 1 (Sardinas-Patterson's Algorithm). *Let $X \subseteq A^+$ be finite and $i \geq 2$ such that $U_i = U_{i-t}$ for some $t > 0$. Then*

$$X \text{ is a code } \Leftrightarrow\ 1 \notin \bigcup_{j=1}^{i-1} U_j.$$

In particular, the problem whether X is a code is decidable.

Actually the above corollary holds also for some other sets, such as for *regular sets.* We give an example of this.

Example 11. Let $X = \{abaa, baa, baab\} \cup (aba)^+b^+$. Then the U_i sets are as follows

$$\begin{aligned}
U_1 &= \{b\} \cup ba(aba)^*b^+, \\
U_2 &= \{aa, aab\} \cup (ba(aba)^*b^+ \cup a(aba)^*b^+), \\
U_3 &= \emptyset \cup (ba(aba)^*b^+ \cup a(aba)^*b^+), \\
U_4 &= U_3
\end{aligned}$$

Therefore $U_j = U_3$ for $j \geq 3$, and since $1 \notin U_1 \cup U_2 \cup U_3$ we conclude that X is a code. □

The rest of this chapter is devoted to study two special types of codes, so-called maximal and complete codes. Moreover we concentrate on such finite codes.

A code $X \subseteq A^*$ is *maximal* if it is not properly included in any code over A, i.e. $X \cup \{x\}$ with $x \in A^* \setminus X$ is never a code. Complete codes are defined as follows. We say that a set $X \subseteq A^*$ is *dense* if any word of A^* is a factor of X, i.e.

$$F(X) = A^*, \text{ or equivalently } A^* = (A^*)^{-1}X(A^*)^{-1}.$$

Now a set $X \subseteq A^*$ is *complete* if X^* is dense, i.e. any word of A^* is a factor of a word in X^*. Finally by a *complete code* we mean a complete set which is a code.

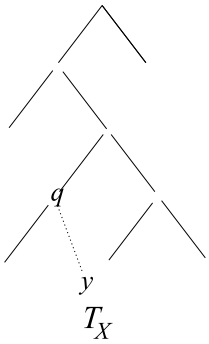
We can also talk about *maximal prefixies,* i.e. prefixies which cannot be extended in the considered alphabet.

Bot of the above notions have a practical motivation: maximal codes use the coding capacity in a maximal way, and complete codes allow any word to be a part of an encoded message.

We start by considering finite prefixies. We recall that finite prefixies can be viewed as finite trees, cf. 68.

Theorem 36. *A finite prefix $X \subseteq \{a, b\}^*$ is maximal if and only if each of its nodes has 0 or 2 descendants.*

Proof. Let T_X denote the tree associated to X.



Then if T_X contains a node q having only one descendant then $X \cup \{y\}$ would be a prefix, where y is obtained by extending the path going through the above node q using the other symbol. Hence X is not a maximal prefix.

Conversely, if T_X has always only two or zero descendants, then, due to the fact that $|A| = 2$, any word z is comparable to a word of X, i.e. one of those is a prefix of another. Hence $X \cup \{z\}$ is not a prefix. □

Despite the simplicity of Theorem 36 it has two interesting consequences.

Corollary 1. *Each finite prefix $X \subseteq \{a, b\}^*$ can be completed, i.e. extended, to a finite maximal prefix.*

Proof. Obvious by Theorem 36 □

Corollary 2. *Each maximal finite prefix $X \subseteq \{a, b\}^*$ is a maximal code.*

Proof. Follows easily from Theorem 36 and a necessary condition for maximal codes proved in Corollary 2 of Theorem 37. □

In order to characterize maximal finite codes we associate with a set $X \subseteq A^*$ of words its numerical value, so-called *measure,* as follows. Let

$$\Pi : \ A^* \to \mathbb{R}_+,$$

where $\mathbb{R}_+$ is a multiplicative monoid of nonnegative real numbers, be a morphism satisfying

$$\sum_{a \in A} \Pi(a) = 1.$$

We call such a morphism a *bernoulli distribution,* or simply a distribution. A distribution is *positive* if $\Pi(a) > 0$ for all $a \in A$ and *uniform* if $\Pi(a) = |A|^{-1}$ for all $a \in A$. Since Π is a morphism

$$\Pi(1) = 1,$$

and moreover

$$\sum_{u \in A^n} \Pi(u) = 1 \quad \text{for all } n \geq 1,$$

as is easily seen by induction:

$$\sum_{u \in A^{n+1}} \Pi(u) = \sum_{\substack{v \in A^n \\ a \in A}} \Pi(va) = \sum_{\substack{v \in A^n \\ a \in A}} \Pi(v)\Pi(a) = (\sum_{v \in A^n} \Pi(v)) \cdot \sum_{a \in A} \Pi(a) \stackrel{\text{i.h.}}{=} 1.$$

Next we extend Π to a function

$$\Pi : \ 2^{A^*} \to \mathbb{R}_+ \cup \{\infty\}$$

by setting

$$\Pi(X) = \sum_{x \in X} \Pi(x) \quad \text{for all } X \subseteq A^*.$$

Note that here the value ∞ is needed if X is allowed to be infinite. Note also that since $\Pi(x) \geq 0$ for all x the value of the above sum is well-defined, that is independent of the order of the summation (cf. Anal II). The value $\Pi(X)$ is called the *measure* of X with respect to Π.

We observe the following inequalities:

$$\Pi(\bigcup_{i\in I} X_i) \leq \sum_{i\in I} \Pi(X_i), \tag{14}$$

where the equality holds at least when the sets X_i are pairwise disjoint,

$$\Pi(XY) \leq \sum_{x\in X}\sum_{y\in Y} \Pi(x)\Pi(y) = \sum_{x\in X} \Pi(x)(\sum_{y\in Y} \Pi(y)) = \Pi(X)\cdot\Pi(Y), \tag{15}$$

and

$$\Pi(X^*) \overset{(14)}{\leq} \sum_{n\geq 0} \Pi(X^n) \overset{(15)}{\leq} \sum_{n\geq 0} (\Pi(X))^n, \tag{16}$$

and moreover in the last formula $\Pi(X^*) < \infty$, if $\Pi(X) < 1$.

When X is a code we can say more.

Lemma 12. *Let $X \subseteq A^+$ and Π be a distribution of A^*.*

i. If X is a code, then for all $n \geq 1$:

$$\Pi(X^n) = (\Pi(X))^n, \tag{17}$$

and

$$\Pi(X^*) = \sum_{n\geq 0} (\Pi(X))^n.$$

In particular, $\Pi(X^*) < \infty$ if and only if $\Pi(X) < 1$.

ii. Conversely, if Π is positive, $\Pi(X) < \infty$ and X satisfies (17), then X is a code.

Proof. Let $S_n = X \times \cdots \times X_n$ be the n-folded Cartesian product of X.
(i). Since X is a code the mapping

$$(x_1, \ldots, x_n) \mapsto x_1 \ldots x_n$$

is a bijection $S_n \to X^n$. Therefore

$$\Pi(X^n) = \sum_{x\in X^n} \Pi(x) = \sum_{(x_1,\ldots,x_n)\in S_n} \Pi(x_1)\ldots\Pi(x_n) = (\Pi(X))^n,$$

where the second equality is due to the above bijection and the fact that Π is a morphism. Since X is a code the sets X^n are pairwise disjoint so that (14) and above yield

$$\Pi(X^*) = \sum_{n\geq 0} (\Pi(X))^n.$$

Finally the last sentence of (i) follows from properties of geometric series.

(ii). Assume the contrary: X is not a code. Then there exists a word u such that

$$u = x_1 \dots x_n = x'_1 \dots x'_m \quad \text{with } x_i, x'_j \in X \text{ and } x_1 \neq x'_1.$$

Then the word uu has two different X-factorizations of length $k = n + m$. Therefore

$$(\Pi(X))^k = \sum_{(y_1,\dots,y_k)\in S_k} \Pi(y_1) \dots \Pi(y_k) \geq \Pi(X^k) + \Pi(uu).$$

But since X satisfies (17), necessarily $\Pi(uu) \leq 0$, a contradiction. □

Now we are ready for a necessary condition for codes.

Theorem 37 (Kraft-MacMillan inequality). *For any code $X \subseteq A^+$ and any distribution Π of A^* we have $\Pi(X) \leq 1$.*

Proof. Assume first that the words of X are of length at most k (which still would allow X to be infinite if A is infinite!). So we have

$$X \subseteq A^1 \cup A^2 \cup \dots \cup A^k,$$

and therefore, for all $n \geq 1$,

$$X^n \subseteq A^1 \cup A^2 \cup \dots \cup A^{nk},$$

implying that

$$\Pi(X^n) \leq nk.$$

If we would have $\Pi(X) = 1 + \epsilon$, with $\epsilon > 0$, then by Lemma 12

$$\Pi(X^n) = (1 + \epsilon)^n,$$

and so we would have

$$(1 + \epsilon)^n \leq nk \quad \text{for all } n \geq 1.$$

Since this is impossible necessarily $\Pi(X) \leq 1$. □

Theorem 37 has several useful consequences.

Corollary 1. *If X is a code over a k-letter alphabet, then $\sum_{x\in X} k^{-|x|} \leq 1$.*

Proof. Apply Theorem (37) for the uniform distribution. □

Corollary 2. *If X is a code and there exists a positive distribution Π such that $\Pi(X) = 1$, then X is maximal.*

Proof. Immediate from Theorem 37 and the positiveness of Π. □

Example 12. Consider the sets

$$\begin{aligned} X &= \{a, ba, bb\}, \\ Y &= \{b, ab, ba\}, \\ Z &= \{ab, aba, aab\}. \end{aligned}$$

Clearly, X is a code while Y and Z are not. Note that Z is obtained from Y by replacing b by ab Now if Π is a distribution such that $\Pi(a) = p$, $\Pi(b) = 1 - p = q$, we compute:

$$\begin{aligned} \Pi(X) &= p + pq + qq = p + pq + (1-p)q = p + q = 1, \\ \Pi(Y) &= q + pq + qp = q(1 + 2p) = q(3 - 2q) = \begin{cases} 1 & \text{if } q = \dfrac{1}{2} \\ \dfrac{10}{9} & \text{if } q = \dfrac{2}{3} \end{cases}, \\ \Pi(Z) &= pq + pqp + ppq = pq + 2p^2q = (p - p^2) + 2(p^2 - p^3) \\ &\leq \frac{1}{4} + \frac{8}{27} < \frac{1}{4} + \frac{1}{2} < 1, \end{aligned}$$

where the first inequality follows by searching the maximal values of the functions $p - p^2$ and $p^2 - p^3$ on interval $[0, 1]$.

So we conclude

- X can be proved maximal by Corollary 2 using *any* distribution, cf Exc. 3/I.
- Y can be shown to be a noncode by using a *suitable* distribution,
- Z cannot be shown to be a noncode by *any* distribution.

This means that sometimes we can settle the question whether $X \subseteq A^+$ is a code or a maximal code simply by computing one number! Intuitively Theorem 37 says that a code cannot contain "too many short words". Corollary 2, in turn, tells that once the measure of a code X reaches 1, then it is maximal. □

Next our goal is to show that the property "$\Pi(X) = 1$ " for finite codes characterizes the maximal ones, i.e. that also the converse of Corollary 2 holds for maximal codes.

In doing so we have to consider complete codes, and more precisely the problem whether a given code can be extended (by adding some words) to a complete one. If this is the case we say that a code *can be completed.* Next result shows how this is always possible.

Theorem 38. *Let $X \in A^+$ be a code and $y \in A^+$ an unbordered word such that $A^*yA^* \cap X^* = \emptyset$. Then*

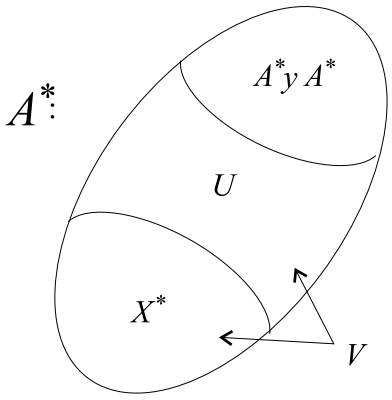
$$Y = X \cup y(Uy)^*, \tag{18}$$

where

$$U = (A^* \setminus X^*) \setminus A^*yA^*, \tag{19}$$

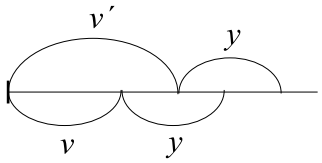
is a complete code.

Proof. Set $V = A^* \setminus A^*yA^*$. Then, by the choice of y, $X^* \subseteq V$ and further $U = V \setminus X^*$, as illustrated in Figure.



Claim I. $Z = Vy$ is a prefix.

To prove this assume that $vy < v'y$ with $v, v' \in V$. There are two possibilities: either $v' < vy$ or $vy \leq v'$. In the first case y is bordered:



And in the second case $v' \in A^*yA^*$. Since both of these are impossible Claim I is proved.

Claim II. Y is a code.

Proof of Claim II. Assume the contrary:

$$y_1 \dots y_n = y'_1 \dots y'_m \quad \text{with } y_i, y'_j \in Y \text{ and } y_1 \neq y'_1 \tag{20}$$

Since X is a code not all words in (20) are from X, i.e. at least one is from $Y \setminus X \subseteq yUy$. Say y_p is such a word. We can further assume that p is minimal, in other words, that words $y_1, \ldots, y_{p-1}$ are in X. Now, by our assumptions $y \notin F(X^*)$ so that also $y_p \notin F(X^*)$. This means that in (20) at least one of the words y'_j is from yU^*y. Again denote by y'_q the one where q is minimal.

Now, by (20), the words

$$y_1 \ldots y_{p-1}y, y'_1 \ldots y'_{q-1}y \in Vy = Z$$

are comparable. But, by Claim I, Z is a prefix, so that they are equal. Consequently, we have

$$y_1 \ldots y_{p-1} = y'_1 \ldots y'_{q-1} \text{ with } y_i, y'_j \in X,\ y_1 \neq y'_1.$$

This, in turn, is possible only if $p = q = 1$.

So we can write

$$\begin{cases} y_1 = yu_1yu_2 \ldots yu_ky & \text{with } k \geq 0 \text{ and } u_1 \in U \\ y'_1 = yu'_1yu'_2 \ldots yu'_ly & \text{with } l \geq 0 \text{ and } u'_i \in U \end{cases} \tag{21}$$

By symmetry, we may assume that $y_1 < y'_1$. Since $U \subseteq V$, the words u_iy and u'_jy are in Z. Hence, Claim I implies that

$$u_1 = u'_1, \ldots, u_k = u'_k,$$

and further that

$$y_1^{-1}y'_1 = t = u'_{k+1}y \ldots yu_ly.$$

Consequently (20) comes into the form

$$y_2 \ldots y_n = ty'_2 \ldots y'_m.$$

Since $y_1 < y'_1$ the word y is a factor of t, and hence also a factor of $y_2 \ldots y_m$. So again, since $y \notin F(X^*)$, some y_j is in yU^*y, and let r be the corresponding minimal value of j. Then the words $y_2 \ldots y_{r-1}y$ and $u'_{k+1}y$ are comparable and in Z, and so, by Claim I, we have:

$$u'_{k+1} = y_2 \ldots y_{r-1} \quad \text{with } y_2, \ldots y_{r-1} \in X,\ u'_{k+1} \in U = V \setminus X^*.$$

This contradiction proves Claim II.

Proof of Theorem 38 (continued). It remains to be proved that Y is complete. Let $w \in A^+$ be an arbitrary word. We write it in the form

$$w = v_1yv_2y \ldots v_{n-1}yv_n \quad \text{with } n \geq 1 \text{ and } v_i \in A^* \setminus A^*yA^*. \tag{22}$$

We show that

$$ywy \in Y^*,$$

which completes the proof. Let $v_{i_1}, \ldots, v_{i_s}$ be exactly those v_i's in (22) which are in X^*. Then we can write

$$ywy = (yv_1y \ldots v_{i_1-1}y)v_{i_1}(yv_{i_1+1}y \ldots v_{i_2-1}y)v_{i_2} \ldots v_{i_s}(yv_{i_s+1}y \ldots v_ny),$$

which shows that $ywy \in Y^*$. □

A reformulation of Theorem 38 is as follows:

Corollary 1. *Each code $X \subseteq A^+$ can be completed.*

Proof. If X is not complete, there exists a word z which is not a factor in X^*. Then Example 2 in Chapter 1 shows that z can be extended to an unbordered word, say y. Of course, y is neither a factor in X^*, so that we can apply the construction of Theorem 38. □

Another consequence is the following.

Theorem 39. *Each maximal code is complete.*

Proof. If $|A| = 1$ the only codes over A are singletons and hence both maximal and complete. If $|A| \geq 2$ and X is not complete, then, by the proof of Corollary 1, it is not maximal either. □

We still need one more lemma.

Lemma 13. *If $X \subseteq A^+$ is finite and complete, and Π is a positive distribution, then $\Pi(X) \geq 1$.*

Proof. Since X is complete we have $A^* = (A^*)^{-1}X^*(A^*)^{-1}$. But by the finiteness of X we can rewrite this as

$$A^* = P^{-1}X^*S^{-1},$$

where P and S are finite sets of words; in fact $P = \mathrm{pref}(X)$ and $S = \mathrm{suf}(X)$. We need only the finiteness of P and S.

Now, we recall that

$$\Pi(A^*) = \sum_{n\geq 0} \Pi(A^n) = \sum_{n\geq 0} (\Pi(A))^n = \infty.$$

It follows that for some $p \in P$ and $s \in S$ necessarily

$$\Pi(p^{-1}X^*s^{-1}) = \infty. \tag{23}$$

Clearly

$$p(p^{-1}X^*s^{-1})s \subseteq X^*,$$

and therefore

$$\Pi(s) \cdot \Pi(p^{-1}X^*s^{-1}) \cdot \Pi(p) \leq \Pi(X^*).$$

So (23), together with the positiveness of Π, implies that $\Pi(X^*) = \infty$.

Hence the lemma follows from the estimates

$$\Pi(X^*) \leq \sum_{n \geq 0} \Pi(X^n) \leq \sum_{n \geq 0} (\Pi(X))^n$$

and properties of geometric series. □

Let us compare Lemma 13 to Theorem 37: The results are in some sense dual. By Theorem 37 the measure of any code is "small", in fact at most 1, while Lemma 13 guarantees that the measure of complete (and finite) sets is "large", namely at least 1.

Now we are ready for a characterization of maximal finite codes.

Theorem 40. *Let $X \subseteq A^+$ be a finite code. The following conditions are equivalent:*

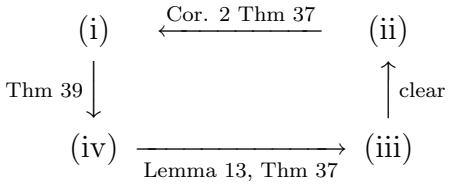
i. X is maximal,

ii. there exists a positive distribution Π such that $\Pi(X) = 1$,

iii. for all positive distributions Π, we have $\Pi(X) = 1$,

iv. X is complete.

Proof. Follows from our earlier considerations:



□

We complete this chapter with two examples. The first one shows that, in general, the classes of complete codes and maximal codes are not the same.

Example 13. Let $A = \{a, b\}$ and define $r(w) = |w|_a : |w|_b$ with the convention that $r(a^n) = \infty$. We consider

$$D = \{w \in A^+ \mid r(w) = 1 \text{ and } \forall u \in A^+,\ u < w :\ r(u) \neq 1\}.$$

Clearly D is a prefix and

$$D^+ = \{w \in A^+ \mid r(w) = 1\}.$$

Claim I. D is dense, and hence complete
Indeed, for any $w \in A^+$ the word

$$v = a^{2|w|_b} w b^{|w|}$$

is in D as can be straightforwardly seen. Hence the Claim I holds.

Claim II. $D' = D \setminus \{x\}$ with $x \in D$ is also dense.
Indeed, any factor of x is a factor of xx which, in turn, is a factor of a word in D (since D is dence), and hence also in $D \setminus \{x\}$. On the other hand, any word which is not a factor of x is also a factor of a word in D (since D is dense), and hence also of a word in $D \setminus \{x\}$.

So D' is a complete code which is not maximal. However, D itself is also maximal (cf. Exc.). □

Our second example is connected to the problem of extending a given code into a maximal one. By Corollary 1 to Theorem 38 any finite code can be completed to a complete, and hence by Theorem 39, to a maximal code Y. The Y given in Theorem 39 is not finite, and by the next example, cannot be so in general.

Example 14. The set

$$X = \{a^5, ba^2, ab, b\}$$

is clearly a code, and we show that it is not a subset of any maximal *finite* code.

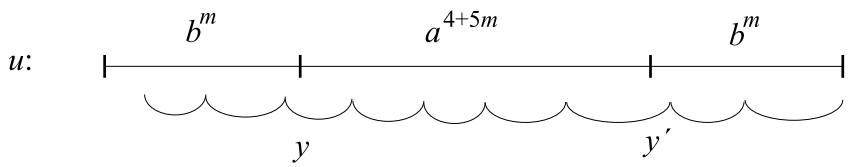
Assume the contrary: $X \subseteq Y$ with Y maximal. Set $m = \max\{|y| \mid y \in Y\}$ and consider the word

$$u = b^m a^{4+5m} b^m.$$

Since Y is maximal it is also complete (Theorem 40) so that u is a factor of a word in Y^*. By the choice of m, the words b^m and a^{4+5m} are not factors of Y. Hence we can write u in the form

$$u = b^p y a^q y' b^r \quad \text{with } p, q, r \geq 0,\ b^r \in Y^* \text{ and } y, y' \in Y \cup \{1\},$$

illustrated as follows:



The word a^5 is the only word in Y not containing b. Hence $q \equiv 0 \pmod 5$, and so $|y|_a + |y'|_a \equiv 4 \pmod 5$. We write

$$y = b^h a^{5s+i} \text{ and } y' = a^{5t+j} b^k \text{ with } 0 \leq i, j \leq 4.$$

Further since $i + j \equiv 4 \pmod 5$, necessarily $i + j = 4$.

There remain 5 cases, and we exclude each of those one by one:
I $i = 0$ and $j = 4$. Then $k \geq 1$, since $\{a^5, a^{5t+4}\}$ is not a code. But then there exists a word with two Y-factorizations:

$$ba^2.a^{5t+4}b^k = b.a^{5(t+1)}.ab.b^{k-1}.$$

II $i = 1$ and $j = 3$. Then we have

$$b^h a^{5s+1}.b = b^h.a^{5s}.ab.$$

III $i = 2$ and $j = 2$. Then we have

$$b.a^{5t+2}b^k = ba^2.a^{5t}.b^k.$$

IV $i = 3$ and $j = 1$. Then, as in I, $h \geq 1$, and so

$$b^h a^{5s+3}.b = b^{h-1}.ba^2.a^{5s}.ab.$$

V $i = 4$ and $j = 0$. Then we have

$$b^h a^{5s+4}.ab = b^h.a^{5(s+1)}.b.$$

Therefore each possible case leads to a contradiction, proving that X indeed cannot be completed to a finite maximal code. □

Our final remark is the following. It is always possible to extend a code X to a maximal one. Indeed, this can be shown by using Zorn's Lemma - and hence the proof is very nonconstructive!

6 Dimension properties

As a starting point we recall our two earlier results, namely Theorems 3 and 34. The first one stated that if two words satisfy a nontrivial relation, then they are powers of a word, and the second one that if words from a finite set X satisfy a nontrivial relation, then the free hull of X contains less than $|X|$ words.

Both of these results are examples of so-called *defect effect:* "If n words satisfy a nontrivial relation they can be expressed (simultaneously) as products of at most $n-1$ words." Clearly, the defect effect can be viewed as a dimension property of words. However, as we shall see

- dimension properties of words are rather weak, and
- there exist, not only one, but several results, which formalize the above defect effect.

Let $X \subseteq A^+$ be a finite set. We define its

combinatorial rank (or c-rank) as the smallest number of words needed to express all words of X, i.e.

$$r_c(X) = \min\{|F| \mid X \subseteq F^*\};$$

prefix rank (or p-rank) as the size of the base of the smallest right unitary submonoid of A^* containing X, i.e.

$$r_p = |B|, \text{ where } B \text{ is the base of } RUM(X) = \bigcap_{\substack{X \subseteq M \subseteq A^* \\ M \text{ is right unit.}}} M;$$

free rank (or f-rank) as the size of the base of the smallest free submonoid of A^* containing X, i.e.

$$r_f(X) = |\widehat{X}|, \text{ where } \widehat{X} \text{ is the base of } FM(X) = \bigcap_{\substack{X \subseteq M \subseteq A^* \\ M \text{ is free}}} M.$$

The above definitions deserve several comments.

Remark 1. Clearly, the combinatorial rank of X is well defined. For the other two ranks this is not obvious, but follows from Theorem 32 (Schützenberger criterium) and the definition of right unitary monoids.

Remark 2. By considerations of the previous chapter the bases of $FM(X)$ and $RUM(X)$ *are unique.* These are called *free hull* and *prefix hull of* X*,* and are denoted by

$$\widehat{X}(f) \text{ and } \widehat{X}(p),$$

so that, in our earlier notations, $\widehat{X}(f) = \widehat{X}$, and

$$r_f(X) = |\widehat{X}(f)| \text{ and } r_p(X) = |\widehat{X}(p)|.$$

Moreover,

$$\widehat{X}(f) \textit{ is a code, and}$$
$$\widehat{X}(p) \textit{ is a prefix.}$$

Concerning the combinatorial rank there need not be the unique Y such that $c_r(X) = |Y|$. However, by the definition of $c_r(X)$, and defect theorem, any Y satisfying $c_r(X) = |Y|$ is necessarily a code.

Remark 3. Note that the combinatorial rank $r_c(X)$ emphasizes *combinatorial* aspects of the notion of a rank, while the other two emphasize more *algebraic aspects* of a rank.

Remark 4. Finally, we note a few obvious connections of the above notions

$$r_c(X) \leq \min\{|X|, |A|\},$$
$$r_c(X) \leq r_f(X) \text{ and } r_c(X) \leq r_p(X).$$

Example 1. Let $X = \{a, ab, cc, bccdd, dda\}$. We compute the prefix hull $\widehat{X}(p)$ of X. Now the words of X satisfy just one minimal relation:

$$\overbrace{a\ \underbrace{b\ c\ c}\ d\ d\ a}$$

(1)

Since $a, ab \in X \subseteq \widehat{X}(p)^+$ and $\widehat{X}(p)^+$ is right unitary, necessarily $b \in \widehat{X}(p)^+$. Similarly, from $bccdd, bcc \in \widehat{X}(p)^+$ we obtain that $dd \in \widehat{X}(p)^+$. Therefore

$$\{a, b, cc, dd\} \subseteq \widehat{X}(p)^+,$$

and hence also

$$\{a, b, cc, dd\}^+ \subseteq \widehat{X}(p)^+.$$

But the set $\{a, b, cc, dd\}$ is a prefix, so that by the minimality of $RUM(X) = \widehat{X}(p)^+$, we obtain

$$\widehat{X}(p) = \{a, b, cc, dd\},$$

and so

$$r_p(X) = 4.$$

Consequently, also $r_c(X) \leq 4$. That it indeed equals to 4 has to be checked by an exhaustive search: Clearly, any Y such that $X \subseteq Y^*$ must satisfy

- Y contains a,
- Y contains c or cc,
- Y contains b or ab,
- Y contains a word containing d.

Therefore $|Y| \geq 4$, implying that $r_c(X) = 4$. Note also that

$$X \subseteq \{a, b, c, d\}^+ \text{ and } X \subseteq \{a, b, cc, dd\}^+$$

showing that the Y defining the value of $r_c(X)$ is not unique.

Finally, let us compute $\widehat{X}(f)$ from (1). The reasoning is the same what we have already had (and is very similar to that used to compute $\widehat{X}(p)$): From (1) we see that

$a, ab, bccdda, ccdda \in \widehat{X}(f)^+$ and $abcc, abccdd, dda, a \in \widehat{X}(f)^+$, and hence, by SC (Schützenberger's Criterium),

$$b, dd \in \widehat{X}(f)^+.$$

Therefore $\{a, b, cc, dd\} \subseteq \widehat{X}(f)^+$, and the minimality of $\widehat{X}(f)^+$, and the fact that $\{a, b, cc, dd\}$ is a code, imply that

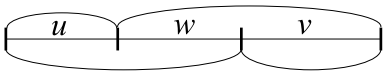
$$\widehat{X}(f) = \{a, b, cc, dd\}.$$

It follows that

$$\widehat{X}(p) = \widehat{X}(f) \text{ and } r_c(X) = r_p(X) = r_f(X) = 4 < |X|.$$

□

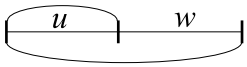
Let us look at a bit more closely the methods used in Example 1 to compute $\widehat{X}(f)$ and $\widehat{X}(p)$. The free hull $\widehat{X}(f)$ is computed by using SC to a double factorization like



(2)

to conclude that $w \in \widehat{X}(f)^+$, and hence to obtain a candidate X_1 for $\widehat{X}(f)$. In Example 1 X_1 would be $\{a, b, cc, ccdd, dda\}$. Since this is not a code, indeed $ccdd.a$ and $cc.dda$, we can continue to obtain $X_2 = \{a, b, cc, dd\}$. But this is a code so that $\widehat{X}(f) = X_2$.

In computing $\widehat{X}(p)$ we work under *weaker* assumptions than (2) namely under the assumption



to conclude that $w \in \widehat{X}(p)^+$. Hence, if the procedure to compute $\widehat{X}(f)$ leads to a prefix code (and not only a code), then we obtained $\widehat{X}(p)$, i.e. $\widehat{X}(f) = \widehat{X}(p)$, as was the case in Example 1.

It is worth emphasizing that although the above method to compute the free hull $\widehat{X}(f)$ works nicely in concrete examples, it is not so clear how to formalize it, i.e. how to define the sequence $X_0, X_1, \ldots, X_n$ of finite sets such that $X_0 = X, X_n = \widehat{X}(f)$ and X_{i+1} is obtained from X_i by application of SC. This, indeed is one reason why our proof for the Defect Theorem was completely different!

On the other hand to compute the prefix hull the situation is much nicer as we now show. We define the following

Procedure: Given a finite set $X \subseteq A^+$, considered as an unambiguous multiset.

1. Find two words $x, y \in X$ such that $x < y$. If there exist no such words go to 4:
2. Set $X \leftarrow X \cup \{x^{-1}y\} \setminus \{y\}$ as a multiset;
3. If X is ambiguous identify equal elements, and go to 1;
4. Output $\widehat{X}(p) \leftarrow X$.

Using above we obtain a variant to the Defect Theorem.

Theorem 41. *Let $X \subseteq A^+$ be finite. The prefix hull $\widehat{X}(p)$ satisfies*

$$|\widehat{X}(p)| \leq |X|, \tag{3}$$

and moreover,

$$|\widehat{X}(p)| < |X|, \tag{4}$$

if X is not a code.

Proof. Recall first that the smallest right unitary submonoid of A^+ containing X exists, by Corollary 1 to Theorem 33 and hence the prefix hull as the base of this monoid exists, too; in our earlier notations

$$\widehat{X}(p)^+ = RUM(X) = \bigcap_{\substack{X \subseteq M \subseteq A^+ \\ M \text{ is right unit.}}} M.$$

If X is a prefix, clearly $\widehat{X}(p) = X$. If on the other hand, $x, y \in X$, where $y = xt$ with $t \in A^+$, then by the definition of the right unitary, $x, t \in \widehat{X}(p)^+$, and hence $X_1 = (X \setminus \{y\}) \cup \{t\} \subseteq \widehat{X}(p)^+$. Since also $X \subseteq X_1^+$, if X_1 is a prefix the minimality of $\widehat{X}(p)$ yields that $X_1 = \widehat{X}(p)$. Otherwise we repeat the construction. But this is exactly what is done in the above Procedure. Hence it computes the prefix hull correctly, and clearly implies (3).

It remains to prove (4). So assume that X satisfies a nontrivial relation, say $x\alpha = y\beta$, with $\alpha, \beta \in X^*$. Further let $s(X)$ denote the *size* of X, i.e. $s(X) = \sum_{x \in X} |x|$. In each round of the procedure the new X, say $X' = (X \setminus \{y\}) \cup \{t\}$ with $y = xt,\ t \in A^+$ satisfies

$$s(X') < s(X), \tag{5}$$

and moreover,

i. X' satisfies a nontrivial relation, namely $\alpha = t\beta$, if $y = xt$ with $t \in A^+$, or

ii. $|X'| < |X|$, when an identification in step 3 is done.

By (5), (i) cannot take place forever, so that (ii) must be encountered, proving (4). □

It is worth noting that the proof of Theorem 41 is really much simpler than that of the Defect Theorem (Theorem 34). Theorem 41 and its proof yield several consequences.

Corollary 1. *The prefix hull can be computed in time* $\mathcal{O}(s(X)^4)$.

Proof. Indeed, each step of the Procedure can be done by a naive algorithm in time $\mathcal{O}(s(X)^3)$, and, by (5), there are at most $\mathcal{O}(s(X))$ rounds. □

Corollary 2. *For each finite set* $X \subseteq A^+$ *we have* $r_p(X) \leq r_f(X)$.

Proof. Since $\widehat{X}(p)^+$ is free, by the minimality of $\widehat{X}(f)^+$, we have

$$\widehat{X}(f) \subseteq \widehat{X}(p)^+,$$

and moreover $X \subseteq \widehat{X}(f)^+$. From the first inclusion we obtain $(\widehat{\widehat{X}(f)})(p) \subseteq \widehat{X}(p)^+$, and hence, by the latter and the minimality of $\widehat{X}(p)^+$, this inclusion cannot be proper. That means that

$$(\widehat{\widehat{X}(f)})(p) = \widehat{X}(p).$$

Consequently, by Theorem 41, $|\widehat{X}(p)| \leq |\widehat{X}(f)|$ as was to be proved. □

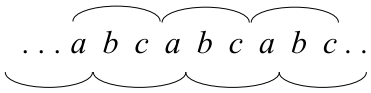
Actually the essential message of Theorem 41, namely (4), can be sharpened. We say that a subset $X \subseteq A^+$ is an *ω-code,* if each infinite word $\omega \in A^\omega$ has at most one X-factorization. Obviously each ω-code is an ordinary code.

Corollary 3. *Each finite set $X \subseteq A^+$ which is not an ω-code satisfies $|\widehat{X}(p)| < |X|$.*

Proof. Exactly as that of Theorem 41, except that instead of "nontrivial relations" in Theorem 41 we now consider "nontrivial ω-relations". □

Corollary 3 states that any finite set $X \subseteq A^+$ such that some infinite word w has two X-factorizations possesses a defect effect. This does not hold if an infinite word is replaced by a 2-way infinite word:

Example 2. Let $X = \{abc, bca, c\}$. Then, as is straightforward to see, $r_c(X) = 3$. However, we have



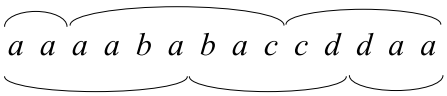
i.e. there exists a 2-way infinite word having two X-factorizations.

We saw in Example 1 that all the ranks we have defined may coincide. Our next example shows that they can also be different.

Example 3. Let

$$X = \{aa, aaaaba, aababac, baccd, cddaa, daa, baa\}.$$

The only minimal relation is:



Hence, by SC, *aaba* and *bac* are in $\widehat{X}(f)$. Therefore

$$X \subseteq X_1^+ \text{ with } X_1 = \{aa, aaba, bac, cd, daa, baa\} \subseteq \widehat{X}(f)^+.$$

Note that each word of X^+ factorizes uniquely in X_1^+. However X_1^+ is not free since it satisfies one minimal relation, namely:

This implies that

$$X_2 = \{aa, ba, c, d, baa\} \subseteq \widehat{X}(f)^+.$$

But now X_2 is a code, so that $X_2 = \widehat{X}(f)^+$. To continue we apply the Procedure of p. 94 to compute the prefix hull:

$$\underline{\overline{ba}a} \quad \Rightarrow \quad a \in \widehat{X}(p),$$

implying that

$$X_3 = \{a, ba, c, d\} \subseteq \widehat{X}(p)^+.$$

And since X_3 is a prefix we conclude that $\widehat{X}(p) = X_3$.

As the conclusion, we have

$$X \subset X_1^+ \subset \widehat{X}(f)^+ \subset \widehat{X}(p)^+,$$

or in terms of ranks

$$4 = r_p(X) < r_f(X) < |X_1| < |X| = 7.$$

In this particular case $r_c(X)$ is also 4. However, if we replace X by $h(X)$, where $h: \{a, b, c, d\}^+ \to \{a, b, c\}^+$ is a morphism defined by $h(a) = a,\ h(b) = b,\ h(c) = c$ and $h(d) = bb$, then the situation changes. Clearly $r_c(h(X)) = 3$, while all the other ranks remain unchanged. Indeed, the above considerations would not change at all. □

To summarize the above considerations we conclude: For any finite set $X \subseteq A^+$, we have (by Corollary 2 of Theorem 41)

$$r_c(X) \leq r_p(X) \leq r_f(X) \leq |X|. \tag{6}$$

Moreover, if X is not a code, i.e. satisfies a nontrivial relation, then

$$r_f(X) < |X|,$$

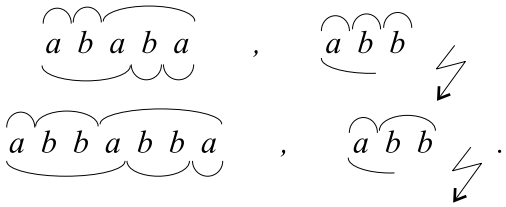
i.e. X possesses a defect effect. In general, the inequalities in (6) can be proper or not. For instance, by Examples 1 and 3, there exists noncodes for which the two first inequalities are both strict or both equalities.

Next we turn to consider the *defect effect of several relations.* For Example, if a set $X \subseteq A^+$ satisfies two "different" nontrivial relations, can the words of X be expressed as products of at most $|X|-2$ words? Unfortunately, the answer to this questions is negative, showing that *dimension properties of words are actually very weak.*

Example 4. Consider the following pair of equations

$$\begin{cases} xyz &= zyx \\ xyyz &= zyyx \end{cases}.$$

Clearly this pair has a solution $x = z = a$ and $y = b$ of (any) rank two. However, the equations are even independent: The triple (a, b, aba) is a solution of the former, but not of the latter, and $(a, bb, abba)$ is a solution of the latter, but not of the former:



□

Before continuing we have to fix some terminology. A finite set $X \subseteq A^+$ of words (considered as an ordered set) can be identified with a solution of a constant-free equation it satisfies. In this view it is more natural to interpret the defect effect we have been considering by saying that

- a nontrivial equation causes a defect effect (i.e. any of its solutions is of rank smaller than the number of unknowns), rather than
- any noncode X possesses a defect effect.

This view becomes even more natural when we now consider defect effect of several relations.

Recall that a *constant-free equation* over A^* with Ξ as the set of unknowns is a pair $(u, v) \in \Xi^+ \times \Xi^+$, usually denoted as $u = v$. Its *solution* is a morphism $h : \Xi^* \to A^*$ satisfying $h(u) = h(v)$. Systems of equations and their solutions are defined in a natural way. Further two systems are *equivalent,* if they have exactly the same solutions. We recall that

- an equation $u = v$ is *reduced* if $\mathrm{pref}_1(u) \neq \mathrm{pref}_1(v)$ and $\mathrm{suf}_1(u) \neq \mathrm{suf}_1(v)$,
- a system S of equations is *independent,* if it is not equivalent to any of its proper subsystems.

Let $X \subseteq A^+$ be finite, and Ξ a finite set of unknowns such that $|X| = |\Xi|$. Relations in X^+ are viewed - under a renaming $X \to \Xi$ - as equations with Ξ as the set of unknowns and X as a solution. Of course, this requires to

consider X as an ordered set. It also allows to express the set of all relations of X^+, in symbols R_X, as a system of equations having Ξ as the set of unknowns and X as a solution. We denote by $E(X)$ *the set of all reduced equations of* X.

The intuitive notion that X satisfies two "different" relations is now formalized that the corresponding set of equations forms an independent pair of equations.

In what follows, unlike earlier, it is important that $1 \notin X$, i.e. *the equations are over a free semigroup* A^+, and not over the free monoid A^*.

Now let $X = \{u_1, \ldots, u_n\} \subseteq A^+$ and $E(X) \subseteq \Xi^+ \times \Xi^+$ be the set of reduced equations determined by X. This means that $X = h(\Xi)$ for some morphism $h : \Xi^+ \to A^+$ satisfying $h(\alpha) = h(\beta)$ for all $(\alpha, \beta) \in \Xi^+ \times \Xi^+$. With each equation e in $E(X)$, say

$$e : \ x\alpha = y\beta \text{ with } x \neq y, \ x, y \in \Xi \text{ and } \alpha, \beta \in \Xi^*$$

we associate the set

$$\Pi(e) = \{h(x), h(y)\},$$

and with the system $E(X)$ we associate the graph $\mathcal{G}_{E(X)}$ such that

- the set of nodes of $\mathcal{G}_{E(X)}$ is X, and
- the set of edges of $\mathcal{G}_{E(X)}$ are defined by:

$$(u, v) \text{ is an edge in } \mathcal{G}_{E(X)} \ \Leftrightarrow \ \Pi(e) = \{u, v\}.$$

We use $\mathcal{G}_{E(X)}$ to define an equivalence relation on X:

$u, v \in X$ are equivalent $\Leftrightarrow$ u, v are in the same component of $\mathcal{G}_{E(X)}$.

Let us denote by $c(\mathcal{G}_{E(X)})$ the number of connected components of $\mathcal{G}_{E(X)}$. Using this quantity we now generalize Theorem 41 as follows. Since we consider elements of X as unknowns it is natural to allow X to be a multiset. Indeed, from the technical point of view this is important in the next proof.

Theorem 42. *For any finite set* $X \subseteq A^+$ *we have*

$$r_c(X) \leq r_p(X) \leq c(\mathcal{G}_{E(X)}).$$

Proof. As we have seen the first inequality holds. To prove the second we have to recall the Procedure of p. 94 to compute the prefix hull of X.

Let u—v be an edge in $\mathcal{G}_{E(X)}$. Assuming, by symmetry, that $u \leq v$ we have two possibilities:

1. u=v, and then we identify u and v,

2. $v = ut$ with $t \in A^+$, and then we replace X by $X \cup \{t\} \setminus \{v\}$.

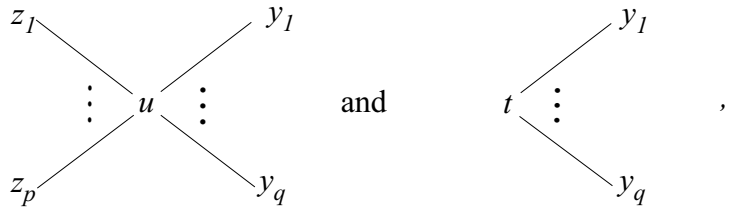
Let $X' \subseteq A^+$ be a multiset obtained from X by performing (1) or (2) once. Note that, due to (2), X' can be a multiset even if X would be unambiguous. Our claim is:

$$c(\mathcal{G}_{E(X')}) \leq c(\mathcal{G}_{E(X)}). \tag{7}$$

Now, if the operation performed is (1) there is nothing to be proved. So it remains to analyze what happens to the graph $\mathcal{G}_{E(X)}$ when 2 is performed. In particular, we have to consider what happens to a subgraph of it of the form:

(8)

Clearly, the connections z_i—u remain, and connection v—y_j are replaced by u—y_j. Moreover, the node v will disappear and a new node t will be connected in $\mathcal{G}_{E(X')}$ to all y_k's in X such that $uy_k\alpha = v\beta$, with $\alpha, \beta \in X^*$, is in $E(X)$. In addition, a new node t may create some completely new edges to $\mathcal{G}_{E(X')}$. But what is important is that, if $\mathcal{G}_{E(X)}$ contains the subgraph of (8), then $\mathcal{G}_{E(X')}$ contains the following subgraphs:



where, moreover, the nodes y_k are nodes of $\mathcal{G}_{E(X)}$, i.e. belong to some of the components of $\mathcal{G}_{E(X)}$. Therefore, the replacement of v by t does not increase the number of components, proving (7).

To complete the proof we first not that $s(X') < s(X)$. Consequently, an iterative application of rules (1) and (2) leads finally to a discrete graph, the edges of which are labeled by a set $\overline{X} \subseteq A^+$.

Now, it is important to note that when performing (1) and (2) we are actually following the Procedure of p. 94. Hence, by the proof of Theorem

41, $\overline{X}$ is included in $\widehat{X}(p)^+$, and therefore, by the minimalities of $\widehat{X}(p)^+$ and $\widehat{\overline{X}}(p)^+$, necessarily $\widehat{\overline{X}}(p) = \widehat{X}(p)$, implying

$$|\widehat{\overline{X}}(p)| = |\widehat{X}(p)| \leq |\overline{X}|.$$

On the other hand, by (7) and the discreteness of $\mathcal{G}_{E(\widehat{X})}$, we also have

$$|\overline{X}| = c(\mathcal{G}_{E(\widehat{X})}) \leq c(\mathcal{G}_{E(X)}).$$

These two inequalities completes the proof. □

Theorem 42, like Theorem 41, has a number of interesting consequences.

Corollary 1. *Let $X \subseteq A^+$ be finite. If the graph $\mathcal{G}_{E(X)}$ is connected then X is periodic, i.e. there exists $z \in A^+$ such that $X \subseteq z^+$.*

□

As a special case of the above we obtain

Corollary 2. *If a three-element set $X = \{u, v, w\} \subseteq A^+$ satisfies the relations*

$$\begin{cases} u\alpha &= v\beta \\ u\gamma &= w\delta \end{cases} \quad \text{with } \alpha, \beta, \gamma, \delta \in X^*. \tag{9}$$

then u, v and w are powers of a same word.

Proof. Indeed the graph associated to (9) is connected:

□

Corollary 2 can be viewed as a generalization of Theorem 3 (cf. also Exc. 7/III). As shown by Example 4, the pair (9) cannot be replaced by a pair of independent equations. Connected to Corollary 2 there exist the following interesting

Open Problem : Does there exist an independent system of three equations with three unknowns such that it has a nonperiodic solution in A^+?

The third Corollary to Theorem 42 is like that of Theorem 41. Instead of considering finite relations of X^+ we can consider one-way infinite relations. Using those we can associate with each $X \subseteq A^+$ a graph, say $\mathcal{G}_{E_\omega(X)}$, exactly as $\mathcal{G}_{E(X)}$ was associated to X. And the proof of Theorem 42 immediately extends to

Corollary 3. *For each finite $X \subseteq A^+$ we have*

$$r_c(X) \leq r_p(X) \leq c(\mathcal{G}_{E_\omega(X)}).$$

□

We continue with a few examples. The first one points out clearly that it is important to assume that $1 \notin X$.

Example 5. The graph of the pair

$$\begin{cases} x &= zx \\ y &= zy \end{cases}$$

is clearly connected. However, *in* A^* it has a solution of rank 2, namely $x = a, y = b$ and $z = 1$. □

Although we do not know an answer to the open problem of the p. 101, we shall show in next examples -as a further evidence of the weakness of dimension properties of words - that there exist "large" independent systems of equations forcing only a "small" defect effect.

Example 6. Let

$$\Xi = \{x, y\} \cup \{u_i, v_i, w_i \mid i = 1, \ldots, n\}$$

be a set of unknowns and

$$S: \ xu_jw_kv_jy = yu_jw_kv_jx \quad \text{for } j, k = 1, \ldots, n,$$

a system of equations over A^+ with Ξ as the set of unknowns. Then clearly

$$|S| = n^2 \quad \text{and} \quad |\Xi| = 3n + 2.$$

We claim that

1. S has a solution of combinatorial rank $3n + 1$,
2. S is independent.

The condition (1) is easy to satisfy: choose $x = y$, so that all equations become trivial, and hence a required solution can be found when $|A| \geq 3n+1$.

The fact that S is independent is more difficult to see. We have to show that, for each pair (j, k), there exists a solution of

$$S(j, k) = S \setminus \{xu_jw_kv_jy = yu_jw_kv_jx\},$$

which is not a solution of the whole S. Here is such a solution:

$$\begin{cases} x & = b^2ab \\ y & = b \\ u_t & = \begin{cases} ba & \text{if } t = j \\ bab & \text{otherwise} \end{cases} \\ w_{t'} & = \begin{cases} bab^2 & \text{if } t' = k \\ b & \text{otherwise} \end{cases} \\ v_t & = \begin{cases} ba & \text{if } t = j \\ a & \text{otherwise} \end{cases} \end{cases}. \tag{10}$$

Then if $t = j$ and $t' = k$ we compute

$$xu_jw_kv_jy = b^2ab.ba\cdots \neq b.ba.bab^2\cdots = yu_jw_kv_jx,$$

and conclude that (10) is not a solution of S. That it is a solution of $S(j,k)$ is a matter of simple computations:

$$\begin{aligned} t \neq j \wedge t' \neq k &\ : \ b^2ab.bab.b.a.b = b.bab.b.a.b^2ab, \\ t \neq j \wedge t' = k &\ : \ b^2ab.bab.bab^2.a.b = b.bab.bab^2.a.b^2ab, \\ t = j \wedge t' \neq k &\ : \ b^2ab.ba.b.ba.b = b.ba.b.ba.b^2ab. \end{aligned}$$

As a modification of Example 6 we present

Example 7. Now let

$$\Xi = \{x_i, y_i, u_i, v_i, w_i \mid i = 1, \ldots, n\}$$

and

$$S' : \ x_iy_jw_kv_jy_i = y_iu_jw_kv_jx_i \quad \text{for } i,j,k = 1,\ldots,n.$$

Consequently, S' contains n^3 equations having only $5n$ unknowns. Note that S' is obtained form S of Example 6 by introducing both for x and y n copies. Next we extend the solution (10) as follows:

$$\begin{cases} x_{t''} & = \begin{cases} b^2ab & \text{if } t'' = i \\ a & \text{otherwise} \end{cases} \\ y_{t''} & = \begin{cases} b & \text{if } t'' = i \\ a & \text{otherwise} \end{cases} \end{cases}. \tag{11}$$

It follows directly from computations in Example 6 that (10) and (11) define a solution which satisfies all equations of S' except the one $x_iu_jw_kv_jy_i = y_iu_jw_kv_jx_i$. Therefore also S' is independent, and still has a nonperiodic solution. □

We make the following remarks connected to above considerations, and in particular to Examples 6 and 7

Remark 1. Example 6 can be interpreted as follows: There exists an independent system of equations of size $\Omega(n^2)$ containing n unknowns such that it forces only the minimal defect effect, i.e. the defect effect of order 1. Similarly, Example 7 shows that there exists an independent system of equations of size $\Omega(n^3)$ containing n unknowns which does not force the maximal defect effect, i.e. force solutions to be periodic.

Remark 2. The above bounds were in the semigroup A^+. In the free monoid A^* these lower bounds can be made $\Omega(n^3)$ and $\Omega(n^4)$, respectively.

Remark 3. In Example 6 we needed that $|A| \geq 3n-1$ in order to find a solution of combinatorial rank of at least $3n+1$. If instead of the combinatorial rank we would consider the prefix rank, then A can be binary: Indeed, the solution of c-rank $3n+1$ in Example 6 can be transformed to a solution having the prefix rank equal to $3n+1$ by encoding letters, say $a_1, \ldots, a_{3n+1}$, to binary words as follows: $a_i \rightarrow a^i b$.

Remark 4. Finally, as we saw in Corollary 1 of Theorem 41, the prefix rank is not difficult to compute. Similarly, one could show that the free rank of a finite set X can be computed in polynomial time. The same does not hold for the combinatorial rank: the problem of deciding whether the combinatorial rank of a given finite set is less than a given number k is so-called NP-complete problem. Consequently, it is not likely that there exists a polynomial time algorithm to compute the combinatorial rank of a given finite set X!

So far we have considered (different types of) ranks of finite sets $X \subseteq A^+$, or those of a solution of an equation with a finite number of unknowns. Now we turn to consider ranks of equations defined as follows: The *rank of an equation* $u = v$ is the maximal of the ranks of its solutions.

So it looks that different notions of the rank of a finite set seem to lead different notions of the rank of an equation. *Fortunately, this is not the case,* as we shall next show. More precisely, we show that the combinatorial and the prefix rank of an equation - defined as above - coincide. A similar argumentation could be used to show that the same holds for the combinatorial and the free rank.

Theorem 43. *Let $e: \ u = v$ be a constant-free equation having the unknowns Ξ. Then the combinatorial and the prefix ranks of e coincide, i.e.*

$$\max\{r_c(h(\Xi)) \mid h \in Sol(e)\} = \max\{r_p(h(\Xi)) \mid h \in Sol(e)\}. \qquad (12)$$

Proof. For each solution h of e we clearly have

$$r_c(h(\Xi)) \leq r_p(h(\Xi)),$$

showing that the left hand side of (12) is at most as large as the right hand side.

To prove the converse we construct, for each solution $h : \Xi^+ \to A^+$, a new solution $h' : \Xi^+ \to A^+$ such that

$$r_c(h'(\Xi)) = r_p(h(\Xi)). \tag{13}$$

Let the prefix hull of $h(\Xi)$ be $U = \{u_1, \ldots, u_d\}$. Consequently, for each $x \in \Xi$, $h(x)$ has the unique U-factorization

$$h(x) = u_{i_1} \ldots u_{i_t}. \tag{14}$$

Next let A' be a new alphabet of size d, and $\Theta : (A')^* \to U^*$ be an isomorphism. For each $i = 1, \ldots, d$, let c_i be an element in A' such that $\Theta(c_i) = u_i$.

Now the morphism $h' : \Xi^+ \to (A')^+$ is defined by the condition:

$$h'(x) = c_{i_1} \ldots c_{i_t} \;\Leftrightarrow\; h(x) = u_{i_1} \ldots u_{i_t} \text{ with } u_{i_j} \in U.$$

Clearly, h' is well-defined, and moreover $\Theta(h'(x)) = h(x)$ for all x in Ξ. Since Θ is an isomorphism, A' is the prefix hull of $h'(\Xi)$. Indeed otherwise $h'(\Xi)$ would be included in a smaller right unitary submonoid of $(A')^*$, and hence its image under Θ would be a smaller right unitary submonoid of A^+ than U^+.

From above it follows that $r_c(h'(\Xi)) \leq d = r_p(h(\Xi))$. If $r_c(h'(\Xi)) < d$, then there would be at most $d-1$ words of $(A')^*$ such that each word $h'(x)$ could be expressed as products of these words. Therefore also words in (14) could be expressed as products of at most $d-1$ words of U^+. This, however, contradicts with the fact that, for any $X \subseteq A^+$, the prefix hull $\widehat{X}(p)$ satisfies: Each element of $\widehat{X}(p)$ occurs as the last factor in the $\widehat{X}(p)$-factorization of some word of X. This fact follows directly from the Procedure to construct the prefix hull.

Hence it follows that $r_c(h'(\Xi)) = d$, which proves (13), and hence the whole theorem. □

Theorem 43 deserves a few comments.

Remark 1. In the formulation of Theorem 43 we purposely did not specify the alphabet, where the equation is solved. If we would like to do it we could, by the proof, fix it to be of the size $|\Xi| - 1$.

Remark 2. Essentially the same proof, using Lemma of Defect Theorem, shows that the prefix rank can be replaced by the free rank.

Remark 3. Finally, we emphasize, that due to Theorem 43, the rank of an equation can be defined, in an equivalent way, by using the notion of the rank we have defined for finite sets.

We have seen that words, i.e. free semigroups, have actually rather weak dimension properties. However, the following fundamental result - Ehrenfeucht's Compactness Theorem - shows that they are not extremely weak.

Theorem 44 (Ehrenfeucht's Compactness Theorem). *Each system of equations over A^+ and with a finite number of unknowns is equivalent to some of its finite subsystems and hence any independent system of equations over A^* with a finite number of unknowns is finite.*

Proof. Let Ξ be a finite set of unknowns in the equations

$$S: \ u_i = v_i \quad \text{for } i \in I, \tag{15}$$

and A^* the free monoid, where the system is solved. We assume that equations are constant-free.

If $|A| = 1$ the result follows directly from linear algebra. All the other cases (including the case $|A| = \infty$) are equivalent due to embeddings, cf. Example 6 in Chapter 5,

$$A_i^* \longrightarrow A_2^*,$$

where A_i denotes the alphabet of size i. For convenience we denote

$$\Xi = \{a_0, a_1, \ldots, a_{n-1}\}.$$

The basic idea in the proof is to *convert* a word equation into a pair of polynomial equation over integers. This, in turn, becomes possible by the fact that a word w over A can be interpreted as a number, namely the n-ary number it presents.

Let

$$u = v \quad \text{with } u, v \in \Xi^+ \tag{16}$$

be a word equation. We choose two copies of Ξ, say Ξ_1 and Ξ_2, and associate to (16) the following pair of polynomial equations over $\mathbb{Z}$:

$$\begin{cases} l(u) - l(v) & = 0 \\ n(u) - n(v) & = 0 \end{cases}, \tag{17}$$

where l and n are mappings from Ξ^+ into the set of integer polynomials over $\Xi_1 \cup \Xi_2$, i.e. into $\mathbb{Z}\langle \Xi_1 \cup \Xi_2 \rangle$, defined recursively as follows:

$$\begin{cases} l(a) = a_1 & \text{for } a \in \Xi \\ l(wa) = l(w)a_1 & \text{for } a \in \Xi, w \in \Xi^+ \\ n(a) = a_2 & \text{for } a \in \Xi \\ n(wa) = n(w)l(a) + n(a) & \text{for } a \in \Xi, w \in \Xi^+ \end{cases} . \tag{18}$$

Clearly, the values $l(w)$ and $n(w)$ are well-defined polynomials over *commuting unknowns* $\Xi_1 \cup \Xi_2$. Note that the coefficients of the monomials of these polynomial are in the set $\{-1, 0, 1\}$, which however is not important. Note also that, by induction, the function n satisfies

$$n(ww') = n(w)l(w') + n(w') \quad \text{for } w, w' \in \Xi^+. \tag{19}$$

Next we associate to a word

$$w = a_{i_{k-1}} \dots a_{i_0} \quad \text{with } a_{i_j} \in A,$$

two numbers

$$\delta(w) = a_{i_0} + a_{i_1} n + \cdots + a_{i_{k-1}} n^{k-1}$$

and

$$\delta_0(w) = n^k.$$

Consequently, $\delta(w)$ is the value of w as an n-ary number, and $\delta_0(w)$ is the value $n^{|w|}$. This guides us to set $\delta(1) = 0$ and $\delta_0(1) = n^0 = 1$. Obviously the correspondence

$$w \longleftrightarrow (\delta(w), \delta_0(w))$$

is one-to-one, and we use this to show that

$$h : \ \Xi^* \to A^* \text{ is a solution of (17)}$$

if and only if

the $2n$-tuple $(\delta_0(h(a_0)), \dots, \delta_0(h(a_{n-1})), \delta(h(a_0)), \dots, \delta(h(a_{n-1})))$ is a solution of (16).

To prove this equivalence let us denote

$$s = (h(a_0), \dots, h(a_{n-1}))$$

and

$$s_1 = \delta_0(s) \quad \text{and} \quad s_2 = \delta(s),$$

where δ_0 and δ are applied componentwise.

First assume that s is a solution of (16), i.e. $h(u) = h(v)$. Then

$$l(u)\big|_{s_1} = n^{|h(u)|} = n^{|h(v)|} = l(v)\big|_{s_1},$$

showing that s_1 is a solution of the equation $l(u) - l(v) = 0$. Similarly, factorizing $u = u'u''$, with $h(u'), h(u'') \neq 1$, we conclude

$$\begin{aligned} n(u)\big|_{s_1,s_2} &\overset{(19)}{=} n(u')\big|_{s_1,s_2} \cdot l(u'')\big|_{s_1,s_2} + n(u'')\big|_{s_1,s_2} \\ &\overset{\text{i.h.}}{=} \delta(h(u'))n^{|h(u'')|} + \delta(h(u'')) \overset{\text{def.}}{=} \delta(h(u'u'')) = \delta(h(u)). \end{aligned}$$

The above holds also, due to the definitions of $\delta(1)$ and $\delta_0(1)$ as the basis of induction, when there does not exist the above factorization. Similarly, we conclude that

$$n(v)\big|_{s_1,s_2} = \delta(h(v)).$$

Consequently, the pair (s_1, s_2) is a solution of the equation $n(u) - n(v) = 0$.

Conversely, if a pair (s_1, s_2) is a solution of (17) then the above calculations show that $\delta(h(u)) = \delta(h(v))$ and $\delta_0(h(u)) = \delta_0(h(v))$, which implies that $h(u) = h(v)$, i.e. h is a solution of (16). Let

$$P: \ p_j = p_j(\Xi_1, \Xi_2), \ \text{for } j \in J,$$

be the set of polynomial equations obtained from word equations of S by the formula (17). Next we use Hilbert's Bases Theorem, which we prove in a moment, and which says that P is *finitely based,* i.e. there exists a finite subset $P_0 = \{p_j \mid j \in J_0\}$ of P such that each polynomial p in P can be expressed as a linear combination of polynomials in P_0:

$$p = \sum_{j \in J_0} q_j p_j \quad \text{with } q_j \in \mathbb{Z}\langle \Xi_1 \cup \Xi_2 \rangle.$$

Consequently, the systems

$$\text{"}p_j = 0 \text{ for } j \in J\text{"} \quad \text{and} \quad \text{"}p_j = 0 \text{ for } j \in J_0\text{"}$$

have exactly the same solutions. Therefore, by the beginning of the proof, our original system S is equivalent to its subsystem consisting of those word equations needed to determine (a super set of) P_0. □

To complete the proof of Theorem 44 we present also a proof of Hilbert's Bases Theorem. In what follows we consider polynomials with integer coefficients and over a fixed (but arbitrary) finite set of commuting unknowns. With these we formulate:

Theorem 45 (Hilbert's Bases Theorem). *For each at most denumerable set P of polynomials there exists a finite subset P_0 of P such that each polynomial p can be expressed in the form*

$$p = \sum_{i=1}^{r} q_i p_i$$

with each p_i in P_0 and q_i being a polynomial.

Proof. Let X be the set of unknowns of polynomials of P, and $r = |X|$. The proof is by induction on r.

$r = 0$. Now P is an arbitrary set of integers. Let p_0 be a number in P having the minimal absolute value, and for each $j = 1, \ldots, |p_0| - 1$, let p_j be any element of P such that

$$p_j \equiv j \pmod{p_0},$$

if such an element exists. Then, clearly, we can choose

$$P_0 = \{p_0\} \cup \{p_j \mid j = 1, \ldots, |p_0| - 1,\ p_j \text{ is defined}\}$$

as a required set.

Induction step: We consider a set $X' = X \cup \{x\}$ of variables, where $|X| = r$, and assume that the theorem holds for polynomials over X.

As an auxilary result we shall use the following fact. For a set Q of polynomials let us define its linear closure $\overline{Q}$ as the set of all linear combinations of polynomials in Q, i.e.

$$\overline{Q} = \{\sum_{i=1}^{n} s_i q_i \mid n \geq 1, q_i \in Q \text{ and } s_i \text{ is a polynomial}\}.$$

Further let us call Q *linearly closed* if $Q = \overline{Q}$. Then we have

Claim. If Theorem 45 holds for linearly closed sets of polynomials it holds for all set of polynomials.

Proof. Let Q be a set of polynomials. Clearly, $\overline{Q}$ is linearly closed and hence finitely based; let $\overline{Q}_0$ be such a finite subset. But elements of $\overline{Q}_0$ are finite linear combinations of polynomials in Q. Hence, also Q is finitely based.

Now, we are ready to prove the induction step. By claim we can assume that P is linearly closed. Consider an arbitrary polynomial p in P, and write it in the form

$$p = c_0 x^m + c_1 x^{m-1} + \cdots + c_m, \tag{20}$$

where c_j's are polynomials over X. Let C be the set of polynomials c_0 in (20). By induction hypothesis, C is finitely based, i.e. each polynomial in C is a linear combination of polynomials from a finite subset C_0 of C. (Of course, in these combinations coefficients are polynomials and not constants!). Next, for each polynomial c in C_0 choose a polynomial p_c from P satisfying

$$p_c = cx^m + c'q(x),$$

where $\deg q \leq m-1$ and c' is a polynomial over X. Let

$$P_0' = \{p_c \mid c \in C_0\} = \{p_i \mid i = 1, \ldots, t\}.$$

and

$$M = \max\{m \mid p_c \in P_0'\}.$$

Now let $p \in P$, and assume that $m > M$ in its representation (20). By the construction of P_0', we can write

$$p = \sum_{i=1}^{t} \gamma_i p_i + p', \tag{21}$$

where γ_i's and p' are polynomials over $X \cup \{x\}$, and moreover the highest exponent of x in p' is $\leq m-1$.

We assumed that our original set P is linearly closed, and hence, by (21), p' is in P. Therefore the above procedure can be repeated, so that finally we can assume in (21) that the highest exponent of x in $p' \leq M$, which is a fixed constant.

Secondly, we consider polynomials p_m of P, for which the number m in their representations (20) are fixed and $\leq M$. Repeating the argument of the beginning of the proof, we conclude that there exists a finite set of polynomials, say $\{p_1', \ldots, p_s'\}$ such that any of the considered polynomial p_m can be expressed in the form

$$p_m = \sum_{i_1}^{s} \gamma_i' p_i' + p'',$$

where the highest exponent of x in p'' is $\leq m-1$.

Obviously, we can repeat the second procedure for polynomials of lower and lower, but fixed, degrees, so that finally we find the required finite subset P_0 as the union of sets obtained in each of the above steps, including that for constructing P_0'. □

We conclude with a few remarks

Remark 1. The proof of Theorem 44 has one very peculiar looking feature: It connects a problem on word equation, i.e. on *noncommuting* unknowns, to a problem of polynomial equations, i.e. on commuting unknowns!

Remark 2. Although it is not very visible in the above presentation of the proof of Theorem 44, it is, or can be, based on the embedding of A^* into the multiplicative monoid of integer matrices, cf. Example 4 in Chapter 5.

Remark 3. The other unavoidable tool in the proof is Hilbert's Bases Theorem.