Two results on words

by

Shuo Tan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The study of combinatorial patterns of words has raised great interest since the early 20th century. In this thesis we primarily study two combinatorial patterns. The first pattern is "abelian k-th power free" and the second one is "representability of a set of words of equal length".

In Chapter 1 we give a brief introduction to these two combinatorial patterns. In Section 2.1 we present a proof that the language of non-abelian squares is not context-free using Odgen's Lemma. In Section 2.2 we present a more elegant proof for this quadratic case by applying a characterization theorem for the bounded context-free languages. Strengthening the technique applied in Section 2.2, we prove in Section 2.3 that the language of non-abelian cubes is not context-free.

In Chapter 3 we study the representability of a set of words of a fixed length. A set S of words of length n is representable if there exists some word w such that the set of length-n factors of w equals S. In Section 3.1 we give a lower bound and an upper bound for the number of representable sets of words of fixed length. In Section 3.2 we give a lower bound $l(n)$ such that if a set S of words of length n is representable, then there exists a word w, with $|w| < l(n)$, such that the set of factors of w equals S. We study a variation of these problems in Section 3.3: we fix a length t, and try to evaluate the number of sets of words of length n such that there exists some word w of length t such that the set of length-n factors of w equals S. We give a closed-form formula in the case where $n \leq t < 2n$.

Finally in Chapter 4, we present some open problems which are related to these two combinatorial patterns.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Jeffrey Shallit, for his continuous support of my study and research, for his motivation and immense knowledge. I would also like to thank him for his patience in reviewing and fixing my thesis and paper.

Furthermore, I would also like to acknowledge Professor Bin Ma and Professor Ming Li for their comments on my thesis and presentation.

Dedication

I would like to dedicate this thesis to my family.

Table of Contents

List of Tables viii

List of Figures ix

1 Introduction 1

1.1 Combinatorial patterns . . . 1
1.2 Words . . . 3
1.3 Context-free languages . . . 4
1.4 Representable sets . . . 8

2 Non-abelian k-th Powers 9

2.1 Binary case: A first proof . . . 10
2.2 Binary case: A more elegant proof . . . 16
2.3 Cubic case: A generalization . . . 21

3 Representable sets of words of equal length 26

3.1 Bounds on the size of $\mathring{R}_n$ . . . 26
3.1.1 Lower bound . . . 26
3.1.2 Upper bound . . . 28
3.2 Shortest witness . . . 32
3.3 Fixed-length witnesses . . . 32
3.4 Numerical results . . . 37

4 Open problems 39

References 41

List of Tables

3.1 Numerical results on representable subsets 38
3.2 Numerical results on $T(t, n)$ 38
3.3 Numerical results on $T(t, n)$ 38

List of Figures

Chapter 1

Introduction

1.1 Combinatorial patterns

This thesis focuses primarily on combinatorial patterns. The study of combinatorial patterns has raised a lot of interest for quite some time. For instance, in the 20th century, much research has been done on the existence of infinite words over a given alphabet avoiding certain patterns. Thue [21] gave the first example of an infinite word avoiding square over an alphabet of size 3. Later, various results avoiding k-th powers were achieved. Most of the k-th power-free words are generated by iterating a particular morphism which preserves k-th power-freeness.

In 1961, Erdős [7] introduced the notion of abelian square (see Section 1.2). Erdős asked if there exists an infinite word over a given alphabet containing no abelian square as a factor. It is easy to check that there exists no such infinite word over an alphabet of size 3. Keränen [13] constructed a morphism on an alphabet of 4 symbols which preserves abelian square-freeness. Using this morphism, Keränen proved that abelian squares are avoidable with an alphabet size of 4.

Richmond and Shallit [17] gave an asymptotic estimate for the number of abelian squares of a fixed length over a given alphabet size. They showed that the number of abelian squares of length $2n$ over an alphabet of size k is asymptotically $k^{2n+\frac{k}{2}}(4\pi n)^{\frac{1-k}{2}}$. Blanchet-Sadri and Fox [1] considered abelian primitive partial words. They counted the number of abelian primitive words of a fixed length over a given alphabet size.

We study the context-freeness of the language of non-abelian squares, non-abelian

cubes, etc. We show in Section 2.1 that the language of non-abelian squares is not context-free. We present a much more elegant proof in Section 2.2 for the quadratic case by applying characterizations of bounded context-free languages (defined in Section 1.3). Finally, we consider the cubic case in Section 2.3, and prove that the language of non-abelian cubes is not context-free. We conjecture that the language of non-abelian k-th powers is not context-free.

Another combinatorial pattern we study is the representable sets of words of equal length. Representable sets have been studied in both algorithmic aspects and combinatorial aspects. Algorithmic aspects of related problems have been discussed under the name "shortest common superstring" and "representing words". A *common superstring* of a set of words S is a word containing each word in S as a factor. For example, the word 001100 is a common superstring of the set of words $\{0011, 110, 100\}$. An instance of the "shortest common superstring" problem is a given set S of words. The object of the problem is to find a shortest common superstring of S. Discussions on this topic include complexity class membership of variants of related problems (e.g., $\mathcal{P}$ and $\mathcal{NP}$) and approximation algorithms. For example, Gallant, Maier, and Storer [9] proved that the decision version of the "shortest common superstring" problem is $\mathcal{NP}$-complete; namely, given a set S of words and an integer K, deciding if there exists a common superstring w of S with $|w| < K$ is $\mathcal{NP}$-complete. After this complexity result had been achieved, researchers turned their attention to approximation algorithms. The first known approximation algorithm was given by Li [14], which achieved an approximation ratio of $\log n$. The first constant-bound algorithm, in terms of the approximation ratio, was found in 1991 [3]; while the best known algorithm was presented by Sweedyk [18], with an approximation ratio of 2.5.

Recently, Blanchet-Sadri and Simmons [2] studied the representability of sets by finite partial words, which are sequences containing holes that match all symbols. Formally, given an alphabet Σ, a *partial word* is a set of symbols in $\Sigma \cup \{\Box\}$, where $\Box \notin \Sigma$. A *full word*, in contrast, is an ordinary word which contains no $\Box$. A full word f is *compatible* with a partial word p if $f[i] = p[i]$ for every position i such that $p[i] \neq \Box$. For example, the full word 'elephant' is compatible with the partial word 'ele$\Box$ha$\Box$t'. The set of length-n factors of w is defined to be the set of full words of length n that are compatible with w. Blanchet-Sadri and Simmons showed that whether a given set of words of equal length equals the set of length-n factors of some partial word can be decided in polynomial time. In particular, they gave an upper bound for the length of representing (partial) word for a given set of words.

Our contribution on this topic basically lies in the combinatorial aspect of the problem. We consider the following problems respectively in Chapter 3: for how many different sets

S of words of length n can we find a word w_S such that the set of length-n factors of w_S equals S? For any set S of words of length n, if such a w_S exists, how long a word do we need to represent it? For how many sets S of words of length n can we find a word w of length t such that the set of length-n factors of w equals S? For the first problem, we give a lower and upper bound in the binary case. For the second problem, we give a weaker upper bound and some experimental data. For the third problem, we give a closed-form formula in the case where $n \leq t < 2n$; in particular, we give a characterization of those distinct words having the same subset of length-n factors.

Finally, in Chapter 4 we give some open problems that are related to the two combinatorial patterns we study.

1.2 Words

An alphabet Σ is a non-empty finite set of symbols. The elements of Σ are referred to as *symbols* or *letters*. We let Σ^* denote the set of all finite words over the alphabet Σ. A finite word over an alphabet Σ is a sequence of symbols in Σ. The *length* of a finite word w is the number of symbols in w. The *empty word*, of which the length equals 0, is denoted by ϵ.

A (right-)infinite word over an alphabet Σ is usually defined as a map from $\mathbb{N}^+$ to Σ. The set of all infinite words over the alphabet Σ is denoted by Σ^ω.

Let w, x, y, z be finite words (possibly empty). If $w = xyz$, then we say that y is a *factor* of w. Let $F_n(w)$ denote the set of length-n factors of an ordinary (non-circular) word w, and let $C_n(w)$ denote the set of length-n factors of w where w is interpreted circularly. For example, if $w = 0001$, then $F_2(w) = \{00, 01\}$; while if the word $w = 0001$ is interpreted circularly, then $C_2(w) = \{00, 01, 10\}$.

A *factor* of an infinite word $\mathbf{w}$ is a finite word y such that there exists a finite word x and an infinite word $\mathbf{z}$ such that $\mathbf{w} = xy\mathbf{z}$. We let $F_n(\mathbf{w})$ denote the set of length-n factors of $\mathbf{w}$.

For convenience, we let $w[i]$ denote the i'th letter of a finite word w and $w[i..j]$ denote the factor of w with length $j - i + 1$ that starts with the i'th letter of w. Thus $w = w[1..n]$ where $n = |w|$.

A finite word w is a *square* if it is of the form xx for some non-empty word x. A word is *square-free* if it contains no square as factor. For example, the word **abcdabcd** is a square; while the word **abcdcdba** is not a square, but it is not square-free since it contains the square **cdcd** as a factor. It is quite easy to check that any word contains at least 4 symbols constructed from a binary alphabet contains a square.

In general, we say that a finite word w is a *k-th* power if it is of the form x^k for some non-empty word x. A word is *k-th power-free* if it contains no k-th power as a factor. A word is *primitive* if it is not a k-th power for any $k > 1$. For example, the set of words $\{a^n b^n : n \geq 1\}$ are all primitive.

A word w is an *abelian square* if it is of the form $w_1 w_2$ where w_2 is a permutation of w_1. For example, the English word **reappear** is an abelian square as it can be factorized as two parts **reap** and **pear**, the second part being a permutation of the first part. It is quite obvious that a square word w is also an abelian square.

In general, a word w is an *abelian k-th power* if there exists a partition $w = w_1 w_2 \cdots w_k$ such that each w_i is a permutation of w_1 for $2 \leq i \leq k$. An instance of an abelian 4-th power is the word $a_4 = 010001100100$, where a_4 can be decomposed into four factors of length 3, each containing two 0's and one 1. If a word is not an abelian k-th power, we say that w is a *non-abelian k-th power*. A word w is an *abelian primitive word* if it is a non-abelian k-th power for all $k > 1$. For instance, the words $0^n 1^n$ for $n \geq 1$ are abelian primitive words. The word $0011(01)^{n-2}$ is an abelian primitive word if and only if n is prime; we will give a proof for this in Lemma 2.0.1. By applying this lemma, Domaratzki and Rampersad [6] proved that the language of abelian primitive words is not context-free. They also gave a linear-time algorithm deciding whether a word is abelian primitive.

1.3 Context-free languages

A *context-free language* is a language accepted by some pushdown automata or generated by some context-free grammar. The set of context-free languages is closed under union, reversal, concatenation, Kleene star, as well as morphism and inverse morphism. These closure properties are widely used to prove certain languages are context-free or noncontext-free. The closure properties we use intensively in this thesis are stated as follows:

Proposition 1.3.1. *If L is context-free and R is regular, then $L \cap R$ is context-free.*

Proposition 1.3.2. *If L is context-free and T is a finite-state transducer then $T(L)$ is context-free.*

These two closure properties are widely used to prove that certain languages are not context-free. For example, we consider the language

$$L = \{w : w \text{ contains the same number of 0's and 1's}\}.$$

We let $R = 0^*1^*$ be a regular language. The intersection $L \cap R = \{0^n1^n : n \geq 0\}$ is known to be noncontext-free. Thus we conclude that the language L is also noncontext-free by applying Lemma 1.3.1.

Ogden's lemma [16] is an extension of the pumping lemma. It states that

Theorem 1.3.3. *If a language L is context-free, then there exists a positive integer n, such that for all $z \in L$ with $|z| > n$, if n or more symbols of z are marked arbitrarily, there exists a decomposition $z = uvwxy$ such that*

1. *vx has at least one marked symbol;*
2. *vwx has at most n marked symbols;*
3. *$uv^iwx^iy \in L$ for all $i \geq 0$.*

Applying Ogden's lemma, we show in Chapter 2 that the language of non-abelian squares is not context-free. A few months after we found the first proof, we simplified the proof by using characterization theorems on the class of bounded context-free languages.

In general, a language L is *bounded* if there exists non-empty words $w_1, w_2, \ldots, w_m$ such that $L \subseteq w_1^*w_2^* \cdots w_m^*$. The words $w_1, w_2, \ldots, w_m$ are said to be the corresponding words of a bounded language L.

A nice characterization of the class of bounded context-free languages was given by Ginsburg [10]. Unlike the pumping lemma and Ogden's lemma, which give necessary conditions for a language to be regular or context-free, Ginsburg's characterization theorem gives a necessary and sufficient condition for the class of bounded context-free languages. Before presenting this beautiful theorem, we first introduce some definitions.

Let $\Sigma = \{a_1, a_2, \ldots, a_m\}$ be an ordered alphabet. The *Parikh map* is a function $\psi : \Sigma^* \to \mathbb{N}^m$ such that $\psi(w) = (c_1, c_2, \ldots, c_m)$, where c_i is the number of occurrences of a_i in

w. For example, suppose $\Sigma = \{0, 1, 2\}$. We have $\psi(02121) = (1, 2, 2)$. For a language L, we let $\psi(L) = \{\psi(w) : w \in L\}$.

We say that two vectors $x = (x_1, x_2, \ldots, x_m)$ and $y = (y_1, y_2, \ldots, y_m) \in \mathbb{N}^m$ are *interleaved* if there exist indices $1 \leq i_x < j_x < i_y < j_y \leq k$ such that $x_{i_x}, x_{j_x}, y_{i_y}, y_{j_y}$ are all positive. For example, the two vectors $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$ are interleaved since we can take such $i_x = 1$, $i_y = 2$, $j_x = 3$ and $j_y = 4$.

Let $S \subseteq \mathbb{N}^m$ be a set of vectors. We say that S is *stratified* if every vector in S has at most two nonzero coordinates, and no pair of vectors in S is interleaved.

Let $X = \{v + Au : u \in \mathbb{N}^m\} \subseteq \mathbb{N}^k$ be a linear set with $m \geq 0, v \in \mathbb{N}^k$ and $A \in \mathbb{N}^{k \times m}$. We say that X has *stratified periods* or X is *stratified* if the column set of A is stratified.

For example, the linear set $X_1 = \{v_1 + A_1 u_1 : u_1 \in \mathbb{N}^2\}$ where $A_1 = \begin{pmatrix} 1 & 0 \\ 3 & 0 \\ 2 & 1 \end{pmatrix}$ is not stratified since the first column $(1, 3, 2)$ contains 3 nonzero entries. The linear set $X_2 = \{v_2 + A_2 u_2 : u_2 \in \mathbb{N}^2\}$ where $A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$ is not stratified as well, since the two columns of A_2 are interleaved, as explained in a previous example.

With these definitions, the characterization theorem is stated as follows [10]:

Theorem 1.3.4. *A bounded language L is context-free if and only if the set*

$$E(L) = \{(e_1, e_2, \ldots, e_m) \in \mathbb{N}^m : w_1^{e_1} w_2^{e_2} \cdots w_m^{e_m} \in L\},$$

where $w_1, w_2 \cdots w_m$ are the corresponding words of L, is a finite union of linear sets with stratified periods.

We use this theorem in our proof for the non-context-freeness of the language of non-abelian cubes. Another useful theorem [12] characterizes the class of **DLI** (**d**efined by **l**inear **i**nequality) context-free languages.

Let $\underline{m}$ denote the cyclical ordered set $\{1, 2, \ldots, m\}$. Then the set

$$E(\Theta, \delta, \epsilon) = \bigcap_{I \in \Theta} \{(e_1, e_2, \ldots, e_m) \in \mathbb{N}^m : \epsilon(I) \sum_{i \in I} \delta_i e_i \geq 0\}$$

is a ***DLI**-set* where

1. Θ is a set of subsets of $\underline{m}$. Θ is considered as a multi-set.
2. For any $1 \leq i \leq m$, $\delta_i \in \{-1, 0, 1\}$.
3. For any $I \in \Theta$, $\epsilon(I) \in \{-1, 1\}$.

A bounded language is a **DLI** language if the set

$$E(L) = \{(e_1, e_2, \ldots, e_l) \in \mathbb{N} : w_1^{e_1} w_2^{e_2} \cdots w_m^{e_m} \in L\}$$

is a **DLI**-set.

DLI languages are often used as examples or counterexamples for context-free languages. In such a case we have to decide whether a given **DLI** language is context-free. The following theorem [12] gives a necessary and sufficient condition for a **DLI**-set to be stratified semilinear.

Theorem 1.3.5. *The **DLI**-set*

$$E(\Theta, \delta, \epsilon) = \bigcap_{I \in \Theta} \{(e_1, e_2, \ldots, e_m) \in \mathbb{N}^m : \epsilon(I) \sum_{i \in I} \delta_i e_i \geq 0\}$$

is stratified semilinear if and only if for every $e \in E$ there is a hypergraph H, having the following properties:

1. *The vertices of H are the vertices of a convex m-polygon, indexed by the numbers $1, 2, \ldots, m$ according to their cyclical order.*
2. *The edges of H are one- or two-element subsets of the vertex set $V(H)$ of H.*
3. *If $\{i, j\}$ is a two-element edge of H, then $\delta_i = -\delta_j$.*
4. *The edge f is forbidden if there exists $I \in \Theta$ such that $f \cap I = \{i\}$ and $\epsilon(I) = -\delta_i$. Hypergraph H does not contain any forbidden edge.*
5. *The edges of H are non-crossing.*
6. *The degree of each vertex i is e_i.*

As an example, Kászonyi proved the language $\{0^a 1^b 2^c : a \leq b \leq c\}$ is not context-free by applying Theorem 1.3.5. Here we do not go into details explaining this theorem. The reader is referred to Kászonyi's paper [12] if details are needed. Theorem 10 in his paper is the original statement of this theorem. The example provided is from Example 1 in his paper.

1.4 Representable sets

Let Σ denote the alphabet. We say that a finite word w *witnesses* (resp., *circularly witnesses*) a subset S of Σ^n if $F_n(w) = S$ (resp., $C_n(w) = S$). A subset S of Σ^n is *representable* (resp., *circularly representable*) if there exists a non-empty finite word (resp., circular word) that witnesses S. For example, the set $\{0,1\}^2$ is a representable set of order 2 since the set of all length-2 factors of $w = 00110$ (interpreted non-circularly) equals $\{0,1\}^2$. The set $\{00, 11\}$ is not representable since any word w containing both 0 and 1 must contain either 01 or 10 as a length-2 factor. Let R_n denote the set of all non-empty representable subsets of Σ^n, and let $\mathring{R}_n$ denote the set of all non-empty circularly representable subsets of Σ^n.

Let $\mathrm{sw}(S)$ (resp., $\mathrm{scw}(S)$) denote the length of the shortest non-circular witness (resp., circular witness) for S. For example, we consider the set $S = \{010, 101\}$. We have $\mathrm{sw}(S) = 4$ since the shortest non-circular witness is 0101, of which the length equals 4; $\mathrm{scw}(S) = 3$ since 010 circularly witnesses S. Let μ_n (resp., ν_n) denote the maximum length of the shortest non-circular (resp., circular) witness over all representable subsets of Σ^n.

A *de Bruijn word* b_n of order n over the alphabet Σ is a shortest circular witness for the set Σ^n. It is known [5] that the length of a de Bruijn word of order n over Σ is 2^n. For example, one instance of b_2 is 0011 and two instances of b_3 are 00010111 and 11101000.

Chapter 2

Non-abelian k-th Powers

At the DLT conference in Milan, Italy, Maxime Crochemore asked if the language of non-abelian squares is context-free. We answered his question in an arxiv note [19].

We study the context-freeness of the language of non-abelian k-th power words in this chapter. Given an alphabet Σ, let L_{AP} denote the set of abelian primitive words over Σ; let L_{NAS} denote the set of non-abelian squares over Σ; and let L_{NAC} denote the set of non-abelian cubes over Σ.

By Lemma 1.3.1, we see that if $L_{AP} \cap \{0,1\}^*$ (respectively, $L_{NAS} \cap \{0,1\}^*$ and $L_{NAC} \cap \{0,1\}^*$) is not context-free, then L_{AP} (respectively, L_{NAS} and L_{NAC}) is not context-free. Thus, without loss of generality, we let the alphabet $\Sigma = \{0,1\}$ in this chapter.

We start with the simplest one, i.e., the language L_{AP}. In order to prove that L_{AP} is not context-free, Domaratzki and Rampersad [6] defined the regular language $R_{AP} = 0011(01)^*$.

Lemma 2.0.1. *Let $n > 1$ be an integer and $x_n = 0011(01)^{n-2}$. Then $x_n \in L_{AP}$ iff n is prime.*

Proof. The length of x is $2n$. The word x has exactly n 0's and n 1's. We first consider the case where $n < 5$. We have $x_2 = 0011 \in L_{AP}$, $x_3 = 001101 \in L_{AP}$, and $x_4 = 00110101$ is an abelian square. Now it suffices to consider the case where $n \geq 5$.

If $x_n \notin L_{AP}$, then there exists $k \geq 2$ such that x_n is an abelian k-th power. If $k = 2$, then n is even; otherwise the first half or the second half of w contain a different number

of 0's or 1's. It follows that p is not a prime number. If $k > 2$, then the length of x_n is divisible by k, which implies that n is not a prime number.

If $n \geq 5$ is not a prime number, then there exist $p, q \geq 2$ such that $n = pq$. Note that $x_n = 0011(01)^{q-2}((01)^q)^{p-1}$. Thus it is easy to see that x_n is an abelian q-th power. Thus $x_n \notin L_{AP}$. □

Theorem 2.0.2. *The set L_{AP} is not context-free.*

Proof. Let $M = L_{AP} \cap R_{AP}$. By Lemma 2.0.1, we obtain that $M = \{0011(01)^{p-2} : p \text{ is prime}\}$. The language M is easily seen to be not context-free by applying the pumping lemma. Hence L_{AP} is not context-free since the intersection of L_{AP} and R_{AP} is not context-free. □

2.1 Binary case: A first proof

The binary case, in which we consider the set of non-abelian squares, however, is not that simple. Our proof in the binary case basically uses the same idea as applied in the proof for the primitive case, i.e., construct a certain regular language R and prove that the intersection $L_{NAS} \cap R$ is not context-free. Much of the content of this section comes from my arxiv note [19].

Let $w_i = 10^{i-1}$ for all $i > 1$. We define $R_{NAS} = w_4^* w_3 w_2^* w_3 w_3^*$. We begin with a few lemmas.

Lemma 2.1.1. *The word $w_4^n w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)} \in L_{NAS} \cap (\Sigma^2)^* \cap R_{NAS}$.*

Proof. Suppose $z = w_4^n w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)}$. Clearly $z \in (\Sigma^2)^* \cap R_{NAS}$. Suppose m is the total number of 1's in z. Then $m = 3n! + 4n + 2$. However, the number of 1's in the second half of z is $\frac{4n!}{3} + 2n + 1$, which is not equal to $\frac{m}{2}$. Hence z is not an abelian square. Thus $z \in L_{NAS}$. □

Let $w = 0^{s_0}10^{s_1} \cdots 10^{s_k}$ be a word over Σ. We let $\mathrm{alt}(w)$ denote the number of occurrences of two consecutive blocks of 0's in w containing a different number of symbols. Formally, we define

$$\mathrm{alt}(w) = |\{1 \leq i < k : s_i \neq s_{i+1}\}|.$$

We define $\mathrm{alt}(\cdot)$ over a language K as follows: $\mathrm{alt}(K) = \max_{w \in K} \mathrm{alt}(w)$.

For example, we consider the regular language $K = 0^*1^*0^*1^*$. We see that $\text{alt}(K) \leq 4$ since there are at most 4 blocks of consecutive 0's and 1's. Also, we have $k = 0110001111 \in K$ and $\text{alt}(k) = 4$. It follows that $\text{alt}(K) = 4$.

We say that a sequence of non-negative integers $(a_k)_{k=1}^n$ is *uneven* if $n > 1$ and there exists $i \in [1, n]$ such that $a_i \neq a_{i+1}$. Here $a_{n+1} = a_1$. Otherwise, it is *even*.

Suppose w is a word that contains a 1. Then w is called *uneven* if the sequence $(s_1 + s_{k+1}, s_2, \ldots, s_k)$ is uneven, where $w = 0^{s_1}10^{s_2}1\cdots 0^{s_k}10^{s_{k+1}}$. Otherwise, it is called *even*.

Suppose w is an even word and of the form $0^{s_1}10^{s_2}1\cdots 0^{s_k}10^{s_{k+1}}$. Now we consider what w looks like. Since w is even, we get that $s_1+s_{k+1} = s_2 = \cdots = s_k$. Then $w = 0^s1(0^{s+t}1)^k0^t$ for some $s, t, k \geq 0$. We have the following lemmas:

Lemma 2.1.2. *If w is an uneven word, then* $\text{alt}(w^k) \geq k - 1$.

Proof. Suppose $w = 0^{s_1}10^{s_2}1\cdots 0^{s_k}10^{s_{k+1}}$. Then

$$w^k = 0^{s_1}(10^{s_2}\cdots 10^{s_k}10^{s_{k+1}+s_1})^{k-1}10^{s_2}\cdots 10^{s_k}10^{s_{k+1}}.$$

Since w is uneven, we get $\text{alt}(10^{s_2}\cdots 10^{s_k}10^{s_{k+1}+s_1}) \geq 1$. It follows that $\text{alt}(w^k) \geq k-1$. □

Theorem 2.1.3. *The intersection $L_{NAS} \cap (\Sigma^2)^* \cap R_{NAS}$ is not context-free.*

Proof. (By *Ogden's Lemma*) Let T denote the language $L_{NAS} \cap (\Sigma^2)^* \cap R_{NAS}$. For any $n > 4$, let $z = w_4^n w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)}$. By Lemma 2.1.1 we see that $z \in T$. Mark the first $4n$ bits of z, that is, the bits corresponding to w_4^n. Let $m(s)$ denote the number of bits marked in s. Now we show by contradiction that no decomposition $z = u_0v_0w_0x_0y_0$ satisfies all the following three conditions:

1. condition A: $m(v_0x_0) > 0$;
2. condition B: $m(v_0w_0x_0) \leq n$;
3. condition C: $\forall i \geq 0, u_0v_0^iw_0x_0^iy_0 \in T$.

We first mark z with different colors. Mark the bits corresponding to w_4^n red. Mark the next 3 bits corresponding to w_3 blue. Mark the bits corresponding to $w_2^{n!+n}$ green. Mark the bits corresponding to $w_3w_3^{2(n!+n)}$ black. Define a new function $m(\text{color}, x)$ as the number

of bits in x colored color. Note that $m(x)$ in our former definition is the same as $m(\text{red}, x)$. Here is a picture of how z is colored:

$$\underbrace{w_4 w_4 \cdots w_4}_{\text{red}} \underbrace{w_3}_{\text{blue}} \underbrace{w_2 \cdots w_2}_{\text{green}} \underbrace{w_3 w_3 \cdots w_3}_{\text{black}}.$$

Now we list all possible cases.

(i) Either v or x is the empty word. Without loss of generality, suppose x is empty.

(1) $v \in 0^+$.

(2) v contains a 1 and v is uneven.

(3) v contains a 1 and v is even.

(ii) Both v and x are non-empty words.

(1) $v \in 0^+$ or $x \in 0^+$.

(2) Both v and x contain a 1; v is uneven or x is uneven.

(3) Both v and x contain a 1 and are even.

(a) $m(\text{red}, v) = 0$, which means that w_4^n precedes v in z.

(b) $m(\text{red}, v) > 0$.

(i) $m(\text{red}, x) > 0$.

(ii) $m(\text{red}, x) = 0$ and $m(\text{green}, x) > 0$

- $m(\text{blue}, x) = 3$.
- $m(\text{blue}, x) = 2$.
- $m(\text{blue}, x) = 1$.
- $m(\text{blue}, x) = 0$ and $m(\text{green}, x) > 1$.
- $m(\text{blue}, x) = 0$ and $m(\text{green}, x) = 1$.

(iii) $m(\text{red}, x) = m(\text{green}, x) = 0$ and $m(\text{black}, x) > 0$

Suppose there exists a decomposition $z = uvwxy$ satisfying the above three conditions simultaneously.

Case i: First we consider the case when either v or x is empty. Without loss of generality, suppose x is empty. Then v cannot be empty, since vx is non-empty.

Case i.1: Suppose $v = 0^k$ for some $k \in \mathbb{N}^+$. Then we select $i = 4$. Since there are more than 3 successive 0's in v^4, this is also true for uv^4wx^4y. However, no word in T contains more than 3 successive 0's. Hence we get a contradiction.

Case i.2: Suppose v contains a 1 and is uneven. We pick $i = 6$. Then $\text{alt}(v^6) \geq 5 > \text{alt}(R) = 4$ by Lemma 2.1.2. So $uv^6wx^6y \notin T$, which violates condition C.

Case i.3: Now we consider when v is even. In this case v can be written in the form $0^k1(0^{k+s}1)^p0^s$ for some $k, s, p \in \mathbb{N}$. Then it follows that $m(\text{green}, v) = 0$ and $k + s = 3$ by the following argument. Suppose $m(\text{green}, v) > 0$. Then $m(\text{blue}, v) = 3$. That is to say, the w_3 between the occurrences of w_4's and w_2's lies in v. Then v must be of the form $r_101001r_2$ for some words r_1 and r_2. It follows that $k+s = 2$, since v is even. Now we select $i = 2$. Then $uv^2wx^2y = w_4^n w_3^l w_2^{n!+n} w_3 w_3^{2(n!+n)}$ for some $l > 1$, which violates condition C. Now suppose $k + s \neq 3$. Then we pick $i = 2$. It follows that uv^2wx^2y is of the form $w_4^l w_{k+s+1}^{2+2p} w_4^j w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)} \notin T$, which violates condition C again. Now let $i = \frac{n!}{1+p}$. It follows that $z_i = uv^iwx^iy = w_4^{n!+n} w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)} = (w_4^{n!+n} w_3 w_2^{n!+n})(w_3 w_3^{2(n!+n)})$ is an abelian square, a contradiction.

Case ii: Both v and x are non-empty. In this case, we first show that both v and x contain a 1. Then, we show v and x are even. Finally we rule out all subcases under the condition v and x are even.

Case ii.1: Suppose $v = 0^k$ or $x = 0^l$ for some $k, l \in \mathbb{N}^+$. By a similar analysis in Case i.i, we get that this case violates condition C.

Case ii.2: Suppose v is uneven. By a similar analysis in Case i.ii, we see that this case violates condition C. The same applies to the case when x is uneven.

Case ii.3: Now it remains to consider when v and x are even. Suppose $v = 0^k1(0^{k+s}1)^p0^s$ for some $k, s, p \in \mathbb{N}$, and $x = 0^c1(0^{c+d}1)^e0^d$ for some $c, d, e \in \mathbb{N}$.

Case ii.3.a: First of all we consider the case when $m(\text{red}, v) = 0$. Then $m(\text{red}, x) = 0$ since x precedes v in z. It follows that $m(\text{red}, vx) = 0$, which violates condition A.

Case ii.3.b: Now we turn to the case when $m(\text{red}, v) > 0$. By the same argument in Case i.iii, we get that $m(\text{green}, v) = 0$ and $k + s = 3$. Note that $p < n$, for otherwise the condition $m(\text{green}, v) = 0$ cannot be satisfied. Now we consider the following subcases:

Case ii.3.b.i: If $m(\text{red}, x) > 0$, then $m(\text{green}, x) = 0$ and $c + d = 3$. After selecting $i = \frac{n!}{2+p+e}$, we get that $z_i = uv^iwx^iy = w_4^{n!+n} w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)} = (w_4^{n!+n} w_3 w_2^{n!+n})(w_3 w_3^{2(n!+n)})$ is an abelian square, which violates condition C again.

Case ii.3.b.ii: If $m(\text{red}, x) = 0$ and $m(\text{green}, x) > 0$, then $m(\text{blue}, x) < 3$, for otherwise x cannot be even. There are again four subcases here.

1. The first subcase is $m(\text{blue}, x) = 2$. Then x is in the form $001r_1$, where r_1 is any word. We see that (a) $r_1 = \epsilon$ or (b) $r_1 = 0$, for otherwise x cannot be even. (a) If $r_1 = \epsilon$, then $x = 001$. We pick $i = 2$. It follows that $uv^2wx^2y = w_4^{n+2p+2}w_3^2w_2^{n!+n}w_3w_3^{2(n!+n)} \notin T$, which violates condition C. (b) If $r_1 = 0$, then $x = 0010$. We pick $i = 2$. It follows that $uv^2wx^2y = w_4^{n+2p+2}w_3w_4w_2^{n!+n}w_3w_3^{2(n!+n)} \notin T$, which also violates condition C.

2. The second case is $m(\text{blue}, x) = 1$. Here is a picture.

$$z = w_4w_4 \cdots w_410 \underbrace{0w_2 \cdots}_{x} \cdots w_2w_3 \cdots w_3$$

   Thus (a) $x = 010$ or (b) $x = (01)^l$ for some $l > 0$. (a) If $x = 010$, then we pick $i = 2$. It follows that $uv^2wx^2y = w_4^{n+2p+2}w_3^2w_2^{n!+n}w_3w_3^{2(n!+n)} \notin T$. (b) Otherwise $x = (01)^l$ for some $l > 0$. After picking $i = \frac{n!}{1+p}$, we get

$$\begin{aligned} z_i &= w_4^{n!+n}w_3w_2^{n!+n+\frac{ln!}{1+p}}w_3w_3^{2(n!+n)} \\ &= (w_4^{n!+n}w_3w_2^{n!+n})w_2^{\frac{ln!}{1+p}}(w_3w_3^{2(n!+n)}) \\ &= (w_4^{n!+n}w_3w_2^{n!+n}w_2^{\frac{ln!}{2(1+p)}})(w_2^{\frac{ln!}{2(1+p)}}w_3w_3^{2(n!+n)}). \end{aligned}$$

   Note that $\frac{ln!}{2(1+p)}$ is an integer since $n > 4$. Therefore z_i is an abelian square. Hence $z_i \notin T$, a contradiction.

3. The third case is $m(\text{blue}, x) = 0$ and $m(\text{green}, x) > 1$. Similarly x must be of the form $(01)^l$ or $(10)^l$ for some $l \in \mathbb{N}^+$, since x is even. We pick $i = \frac{n!}{1+p}$ and find the same result as in the second case.

4. The last case is exactly when $m(\text{blue}, x) = 0$ and $m(\text{green}, x) = 1$. Then x has to be the first or the last letter of the substring $w_2^{n!+n}$ of z, since x cannot be a single 0 or 1 (there are no successive 1's in elements of R). Moreover, we find that x cannot be the first letter of $w_2^{n!+n}$, since $m(\text{blue}, x) = 0$. It follows that x is the trailing 0 of $w_2^{n!+n}$. Under this circumstance, we find that $x = 0(100)^e10$, since x is even. Now

we select $i = \frac{n!}{1+p}$. It follows that

$$\begin{aligned} z_i &= uv^i wx^i y \\ &= w_4^{n!+n} w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)+\frac{(1+e)n!}{1+p}} \\ &= (w_4^{n!+n} w_3 w_2^{n!+n}) w_3^{\frac{(1+e)n!}{1+p}} (w_3 w_3^{2(n!+n)}) \\ &= (w_4^{n!+n} w_3 w_2^{n!+n} w_3^{\frac{(1+e)n!}{2(1+p)}})(w_3^{\frac{(1+e)n!}{2(1+p)}} w_3 w_3^{2(n!+n)}). \end{aligned}$$

Note that $\frac{(1+e)n!}{2(1+p)}$ is an integer since $n > 4$. Thus z_i is an abelian square. Therefore $z_i \notin T$, a contradiction.

Case ii.3.b.iii: The last possible subcase is when $m(\text{red}, x) = 0$ and $m(\text{green}, x) = 0$ and $m(\text{black}, x) > 0$. In this case, we get that $c + d = 2$. Now we pick $i = \frac{n!}{1+p}$. It follows that

$$\begin{aligned} z_i &= uv^i wx^i y \\ &= w_4^{n!+n} w_3 w_2^{n!+n} w_3 w_3^{2(n!+n)+\frac{(1+e)n!}{1+p}} \\ &= (w_4^{n!+n} w_3 w_2^{n!+n}) w_3^{\frac{(1+e)n!}{1+p}} (w_3 w_3^{2(n!+n)}) \\ &= (w_4^{n!+n} w_3 w_2^{n!+n} w_3^{\frac{(1+e)n!}{2(1+p)}})(w_3^{\frac{(1+e)n!}{2(1+p)}} w_3 w_3^{2(n!+n)}). \end{aligned}$$

Note that $\frac{(1+e)n!}{2(1+p)}$ is an integer since $n > 4$. Thus z_i is an abelian square. Therefore $z_i \notin T$, a contradiction again.

With the above discussion, we claim that no decomposition of z can satisfy all three conditions simultaneously. Thus T is not context-free. □

With Theorem 2.1.3, it follows that

Corollary. *The language L_{NAS} is not context-free.*

As the readers may notice, Theorem 2.1.3 contains quite a lot of case analysis. Also, the regular language R_{NAS} is not easily extended to a proof of the general case, i.e., there is no trivial way to construct a similar regular language to show that the language of non-abelian k-th powers is not context-free. In the next section, we construct another regular language R_2 and prove that $L_{NAS} \cap R_2$ in not context-free.

2.2 Binary case: A more elegant proof

In our effort to generalize the proof in Section 2.1, we tried many regular languages. One of these regular languages which seems promising is

$$R_2 = (00)^*(11)^*(00)^*(11)^*.$$

We have the following lemma:

Lemma 2.2.1. *The word* $w = 0^{2a}1^{2b}0^{2c}1^{2d}$ *is an abelian square if and only if* $((a = c) \wedge (b \geq d)) \vee ((a \leq c) \wedge (b = d))$.

Proof. If w is an abelian square, two possible factorizations for w are $w = w_1w_2$ and $w = w_3w_4$ where $w_1 = 0^{2a}1^{b_1}$ and $w_2 = 1^{b_2}0^{2c}1^{2d}$; $w_3 = 0^{2a}1^{2b}0^{c_1}$ and $w_4 = 0^{c_2}1^{2d}$. For the first case we have $a = c$ and $b_1 = b_2 + 2d$. Thus $2b \geq 2b_2 + 2d \geq 2d$; it follows that $b \geq d$ and $a = c$. By symmetry, we have $b = d$ and $a \leq c$ for the second case. The result follows immediately. □

Luke Schaeffer (Personal communication, July 2012) showed that the language

$$P = \{0^a1^b2^c3^d : a > c \text{ or } b > d \text{ or } (a < c \text{ and } b < d)\}$$

is not context-free. He applied the fundamental characterization theorem on the class of bounded context-free languages. Here we use Schaeffer's technique to prove that $L_{NAS} \cap R_2$ is not context-free. We begin with a definition of *dimension* of discrete sets.

Given a set $X \subseteq \mathbb{N}^k$, we define a point counting function $\phi_X(R) : \mathbb{N} \to \mathbb{N}$ where

$$\phi_X(R) = |\{x \in X : ||x||_1 \leq R\}|.$$

Here $||.||_1$ is the L_1 norm.

We say that X has *dimension* d ($\dim X = d$) if $\phi_X(R) \in \Theta(R^d)$. Note that this dimension is not well defined for all subsets of $\mathbb{N}^k$.

Proposition 2.2.2. *The dimension of* $\mathbb{N}^k$ *is* k.

Proof. Let

$$A = \{(x_1, x_2, \ldots, x_k) : \sum_{i=1}^{k} x_i \leq R\},$$
$$B = \{(x_1, x_2, \ldots, x_k) : x_i \leq R \text{ for } 1 \leq i \leq k\},$$
$$C = \{(x_1, x_2, \ldots, x_k) : x_i \leq \frac{R}{k} \text{ for } 1 \leq i \leq k\}.$$

Clearly $C \subseteq A \subseteq B$. Thus we have

$$\phi_{\mathbb{N}^k}(R) = |A| \leq |B| = R^k$$

and

$$\phi_{\mathbb{N}^k}(R) = |A| \geq |C| = R^k \frac{1}{k^k}.$$

It follows that $\dim(\mathbb{N}^k) = k$. □

Proposition 2.2.3. *Let X, Y be subsets of $\mathbb{N}^k$ such that $\dim X$ and $\dim Y$ exist. Then*

1. *If $X \subseteq Y$, then $\dim X \leq \dim Y$;*
2. *$\dim X \cup Y = \max\{\dim X, \dim Y\}$.*
3. *$\dim \bigcup_i X_i = \max_i\{\dim X_i\}$.*

Proof. If $X \subseteq Y$, then for any $R > 0$, we have $\phi_X(R) \leq \phi_Y(R)$. Thus $\dim X \leq \dim Y$.

For the union, we have $\phi_{X \cup Y}(R) \leq \phi_X(R) + \phi_Y(R) \leq 2\max\{\phi_X(R), \phi_Y(R)\}$. Thus $\dim X \cup Y \leq \max\{\dim X, \dim Y\}$. Also, we have $\dim X \leq \dim X \cup Y$ since $X \subseteq X \cup Y$. Similarly, $\dim Y \leq \dim X \cup Y$. Thus $\dim X \cup Y = \max\{\dim X, \dim Y\}$.

The third statement follows immediately from the second one. □

Lemma 2.2.4. *Let $X = \{v + Au : u \in \mathbb{N}^m\}$ be a linear set with $m \geq 0$, $v \in \mathbb{N}^k$ and $A \in \mathbb{N}^{k \times m}$. If A has rank m, then $\dim X = m$.*

Proof. Since A has full rank, then the map $u \to Au$ is injective. It follows that

$$\begin{aligned}\phi_X(R) &= |\{v + Au : u \in \mathbb{N}^m \text{ such that } ||v + Au||_1 \leq R\}| \\ &= |\{u \in \mathbb{N}^m : ||v + Au||_1 \leq R\}|.\end{aligned}$$

The triangle inequality gives us

$$||Au||_1 - ||v||_1 \leq ||v + Au||_1 \leq ||Au||_1 + ||v||_1.$$

Let X' denote the set $\{Au : u \in \mathbb{N}^m\}$. We have

$$\phi_{X'}(R - ||v||_1) \leq \phi_X(R) \leq \phi_{X'}(R + ||v||_1).$$

The matrix norm inequality gives us

$$||Au||_1 \leq ||A||_1 ||u||_1.$$

On the other hand, let a_{ij} denote the element of A located in the i-th row and j-th column. Every column of A is nonzero; thus

$$||Au||_1 = \sum_{i=1}^{m} u_i \sum_{j=1}^{k} a_{ji} \geq \sum_{i=1}^{m} u_i = ||u||_1.$$

This gives us

$$\begin{aligned} \phi_{\mathbb{N}^m}(\frac{R}{||A||_1}) &= |\{u : ||A||_1 |u|_1 \leq R\}| \\ &\leq |\{u : ||Au||_1 \leq R\}| = \phi_{X'}(R) \\ &\leq |\{u : |u|_1 \leq R\}| = \phi_{\mathbb{N}^m}(R). \end{aligned}$$

Thus we have

$$\frac{R^m}{{||A||_1}^m} \leq \phi_{X'}(R) \leq R^m.$$

Thus

$$\frac{(R - ||v||_1)^m}{{||A||_1}^m} \leq \phi_X(R) \leq (R + ||v||_1)^m.$$

It follows immediately that $\dim X = m$. □

Lemma 2.2.5. *Let $A \in \mathbb{N}^{n \times n}$ be a square, nonsingular matrix. There exists a permutation $\sigma \in S_n$ such that $A_{1\sigma(1)}, A_{2\sigma(2)}, \ldots, A_{n\sigma(n)}$ are all nonzero.*

Proof. The square matrix A is nonsingular; thus $\det(A) = \Sigma_{\sigma \in S_n} \prod_{i=1}^{n} A_{i\sigma_i} \neq 0$. Thus $\prod_{i=1}^{n} A_{i\sigma_i}$ is nonzero for some $\sigma \in S_n$. It follows that $A_{i\sigma_i}$ are all nonzero for $1 \leq i \leq n$. □

Lemma 2.2.6. *Let l be an even number. Let Y denote the set*

$$\{(2a, 2b, 2c, 2d) \in \mathbb{N}^4 : a < c \text{ and } d < b\}.$$

Then $Y \cap (l\mathbb{N})^4$ has dimension 4.

Proof. We can show that $Y \cap (l\mathbb{N})^4 = \{v + Au : u \in \mathbb{N}^4\}$, where

$$A = l \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

and

$$v = \begin{pmatrix} 0 \\ 2 \\ 2 \\ 0 \end{pmatrix}.$$

Since A is non-singular, $Y \cap (l\mathbb{N})^4$ has dimension 4. □

Theorem 2.2.7. *The intersection $L_{NAS} \cap R_2$ is not context-free.*

Proof. We let M denote the intersection $L_{NAS} \cap R_2$. Lemma 2.2.1 says that the word $0^{2a}1^{2b}0^{2c}1^{2d}$ is an abelian square if and only if the condition $C = ((a = c) \wedge (b \geq d)) \vee ((a \leq c) \wedge (b = d))$ holds. Thus, we have $M = \{0^{2a}1^{2b}0^{2c}1^{2d} : \overline{C} \text{ holds}\}$, where $\overline{C}$ is the negation of C.

Suppose M is context-free. By Theorem 1.3.4, $\varphi(M) = \{(2a, 2b, 2c, 2d) : \overline{C} \text{ holds}\}$ is a finite union of linear sets with stratified periods, say $\bigcup_i X_i$, where $X_i = \{v_i + A_i u : u \in \mathbb{N}^{m_i}\}$. Let Y denote the set $\{(2a, 2b, 2c, 2d) \in \mathbb{N}^4 : a < c \text{ and } d < b\}$. Let $T : \mathbb{N}^4 \to \mathbb{N}^2$ be a map such that $T((a, b, c, d)) = (a - c, d - b)$. Choose l a natural number such that l is divisible by both entries of $T(w)$ for each column w of A_i, for all i.

Clearly $Y \subseteq \varphi(M)$. Thus $Y \cap (l\mathbb{N})^4 \subseteq \varphi(M)$. By Lemma 2.2.6, $Y \cap (l\mathbb{N})^4$ has dimension 4. We have

$$\begin{aligned} Y \cap (l\mathbb{N})^4 &= Y \cap (l\mathbb{N})^4 \cap \varphi(M) \\ &= Y \cap (l\mathbb{N})^4 \cap \bigcup_i X_i \\ &= \bigcup_i ((l\mathbb{N})^4 \cap X_i). \end{aligned}$$

Following Proposition 2.2.3, we see that there exists some X_i such that $Y \cap (l\mathbb{N})^4 \cap X_i$ has dimension 4. For convenience, we denote this X_i by X, and the corresponding A_i (resp., v_i) by A (resp., v).

By Lemma 2.2.5, there are four columns w_1, w_2, w_3, w_4 of A such that the ith entry of w_i is positive, for all i. For convenience, suppose $w_i = (a_i, b_i, c_i, d_i)$ and $T(w_i) = (x_i, y_i)$.

We claim that $x_1 > 0$ or $y_4 > 0$; otherwise, we have $x_1 = a_1 - c_1 \leq 0$ and $y_4 = d_4 - b_4 \leq 0$. It follows that $0 < a_1 \leq c$ and $0 < b_4 \leq d_4$. Thus, w_1 and w_4 are interleaved, contradicting the fact that the periods of A form a stratified set. Thus, $x_1 > 0$ or $y_4 > 0$. Since the two cases are very similar, we assume $x_1 > 0$ and omit the other case.

Let $z = (a, b, c, d)$ be any element of $Y \cap (l\mathbb{N})^4 \cap \varphi(M)$ and let $T(z) = (x, y)$. We have two cases, depending on whether $y_1 > 0$.

1. If $y_1 \leq 0$ then let $z_1 = z + \lambda_1 w_1$, where $\lambda_1 = -\frac{x}{x_1}$. We claim that $\lambda_1 > 0$ since $x = a - c < 0$ and $x_1 > 0$. Also, $\lambda_1 \in Z$ since $x = a - c$ is divisible by l and l is divisible by x_1. Thus we have $z_1 \in X$ since $z \in X$ and we add $\lambda_1 \in N$ copies of w_1, a period of X. On the other hand, we have

   $$T(z_1) = (x + \lambda_1 x_1, y + \lambda_1 y_1) = (0, \hat{y}),$$

   where $\hat{y} = y + \lambda_1 y_1 < 0$. Hence, $z_1 \notin \varphi(M)$, and thus $z_1 \notin X$, which leads to a contradiction.

2. If $y_1 > 0$, we define z_1 as in the previous case. We also define $z_2 = z + \lambda_2 w_1$, where $\lambda_2 = -\frac{y}{y_1}$. Similarly, we can prove that $\lambda_2 \in \mathbb{N}^+$. Thus we have

   $$T(z_1) = (x + \lambda_1 x_1, y + \lambda_1 y_1) = (0, \hat{y}),$$

   $$T(z_2) = (x + \lambda_2 x_2, y + \lambda_2 y_1) = (\hat{x}, 0),$$

   where $\hat{x} = x + \lambda_2 x_1 = \frac{xy_1 - yx_1}{y_1}$ and $\hat{y} = y + \lambda_1 y_1 = \frac{yx_1 - xy_1}{x_1}$. It follows that $\hat{x}\hat{y} = -\frac{(yx_1 - xy_1)^2}{x_1 y_1} \leq 0$, and thus $\hat{x} \leq 0$ or $\hat{y} \leq 0$; thus $z_1 \notin X$ or $z_2 \notin X$. This is a contradiction.

In both cases, we reach a contradiction. So M is not context-free. □

It follows immediately that

Corollary. *The language L_{NAS} is not context-free.*

We achieved this elegant proof for the noncontext-freeness of L_{NAS} by applying Schaeffer's idea. Unfortunately, we are still not able to generalize this proof to show that the language of non-abelian k-th powers is not context-free. However, by strengthening Schaeffer's idea, we obtain some lemmas and prove that the language of non-abelian cubes is not context-free.

2.3 Cubic case: A generalization

In this section we strengthen Schaeffer's technique and prove that the language of non-abelian cubes is not context-free. We begin with the definition of plane.

Let $p \in \mathbb{N}^3$. Let $n \in \mathbb{Z}^3$ be a nonzero vector. A *plane* $P(p, n)$ is a set of points $s \in \mathbb{N}^3$, such that every point $s \in P(p, n)$ satisfies the condition $(s - p) \cdot n = 0$. For example, $P(\{0, 0, 0\}, \{0, 0, 1\})$ gives the set of points satisfying the equation $z = 0$. Let $P^+(p, n) = \{s \in \mathbb{N}^k : (s - p) \cdot n > 0\}$. That is, P^+ is the set of points above the plane P.

Proposition 2.3.1. *Let $n \in \mathbb{Z}^3$ such that exactly one entry of n is negative. Let $p \in \mathbb{N}^3$. Then* $\dim(P(p, n)) = 2$.

Proof. For one direction, we prove that $\dim(P) \geq 2$. Let $p = (p_1, p_2, \ldots, p_k)$ and $n = (n_1, n_2, \ldots, n_k)$. Without loss of generality, suppose $n_1 < 0$ and $n_2, n_3 \geq 0$. Fix any $R > 2(p_1 + p_2 + p_3)$. Let $L = 2n_2n_3 - n_1n_2 - n_1n_3$. Clearly $L > 0$. Let B denote the set $\{x \in \mathbb{N}^3 : (x - p) \cdot n = 0 \text{ and } ||x||_1 \leq R\}$.

We claim that for any $x_2 \in \{p_2 - an_1n_3 : a \in [1, \frac{R}{2L}]\}, x_3 \in \{p_3 - an_1n_2 : a \in [1, \frac{R}{2L}]\}$, there exists a $x_1 \in \mathbb{N}$, such that $(x_1, x_2, x_3) \in B$. For any such x_2, x_3, i.e., $x_2 = p_2 - a_2n_1n_3, x_3 = p_3 - a_3n_1n_2$, where $a_2, a_3 \in [1, \frac{R}{2L}]$, we let $x_1 = p_1 + (a_2 + a_3)n_2n_3$. It is easy to verify that $(x_1, x_2, x_3) \in B$. Thus, we have $\phi_P(R) = |B| \geq \frac{R^2}{4L^2}$. Thus $\dim(P) \geq 2$.

For the other direction we prove $\dim(P) \leq 2$. We fix any $x_1, x_2 \in \mathbb{N}^3$ such that $x_1 + x_2 \leq R$. There exists at most one x_3 such that $(x_1, x_2, x_3) \in B$. Thus, we have $\phi_P(R) = |B| \leq \phi_{\mathbb{N}^2}(R)$. Thus $\dim(P) \leq 2$.

Finally we conclude from these two directions that $\dim(P) = 2$. □

Lemma 2.3.2. *Let $P_1 = P(v_1, n_1)$ and $P_2 = P(v_2, n_2)$ be two planes. Let $X = \bigcup_i X_i$ where $X_1, X_2, \ldots, X_k$ are linear sets such that*

1. *The set $P_1^+ \cap P_2^+ \subseteq X$.*
2. *The plane $P_1 \cap X = \emptyset$.*
3. *The plane $P_2 \cap X = \emptyset$.*

Let T_1 be a set of points p such that $(p - v_1) \cdot n_1$ is divisible by $w \cdot n_1$ for every column of X_i, for all i. Similarly, let T_2 be a set of points p such that $(p - v_2) \cdot n_2$ is divisible by $w \cdot n_2$ for every column of X_i, for all i.

Then for any $X_i = \{v_i + A_i u : u \in \mathbb{N}^{m_i}\}$ with $X_i \cap P_1^+ \cap P_2^+ \cap T_1 \cap T_2 \neq \emptyset$, every column w of A_i satisfies the conditions $w \cdot n_1 \geq 0$ and $w \cdot n_2 \geq 0$.

Proof. We give a proof for the first condition, i.e., for any $X_i = \{v_i + A_i u : u \in \mathbb{N}^{m_i}\}$ with $X_i \cap S \cap T \neq \emptyset$, every column w of A_i satisfies the condition $w \cdot n_1 \geq 0$. The proof for the second condition is omitted since it is quite similar to the first one. For convenience, we let $T = T_1 \cap T_2$ and $S = P_1^+ \cap P_2^+$.

Suppose there exists some X_i with $X_i \cap S \cap T \neq \emptyset$ such that some column w of A_i satisfies the condition $w \cdot n_1 < 0$. We pick any $z \in X_i \cap S \cap T$. Let $z' = z + \lambda w$, where $\lambda = -\frac{(z - v_1) \cdot n_1}{w \cdot n_1}$. We first claim that $\lambda > 0$. This is because $w \cdot n_1 < 0$ and $(z - v_1) \cdot n_1 > 0$ since $z \in P_1^+$. We also observe that $(z - v_1) \cdot n_1$ is divisible by $w \cdot n_1$ since $z \in T_1$. Thus $z \in \mathbb{N}^+$. It follows that $z' \in X_i$. However, we can show that $(z' - v_1) \cdot n_1 = 0$, which implies that $z' \in P_1$. This leads to a contradiction since $P_1 \cap X_i = \emptyset$. Thus, we conclude that for every X_i with $X_i \cap S \cap T \neq \emptyset$, every column w of A_i satisfies the condition $w \cdot n_1 \geq 0$. □

Let $R_3 = 0^*1^*0^*1^*0^*$ be a regular language. Again, we try to prove that $L_{NAC} \cap R_3$ is not context-free. Since R_3 is a bounded language, we will show that the Parikh set of $L_{NAC} \cap R_3$ is not a finite union of linear sets with stratified periods. For convenience, let $M = L_{NAC} \cap R_3$.

Lemma 2.3.3. *The word $w = 0^a1^b0^c1^d0^e$ is an abelian cube if and only if one of the conditions C_1, C_2, C_3 is satisfied, where*

$$C_1 : (a = c = e) \wedge (b \leq 2d) \wedge (d \leq 2b)$$
$$C_2 : (a \leq c) \wedge (2a = c + e) \wedge (b = 2d)$$
$$C_3 : (e \leq c) \wedge (a + c = 2e) \wedge (d = 2b)$$

Here we omit the proof and just list the three possible factorizations of w as an abelian cube. The three possible factorizations (which correspond to the above three conditions respectively) are:

$$0^a 1^{b_1} \cdot 1^{b_2} 0^c 1^{d_1} \cdot 1^{d_2} 0^e$$

$$0^a 1^{b_1} \cdot 1^{b_2} 0^{c_1} \cdot 1^{c_2} 1^d 0^e$$

$$0^a 1^b 0^{c_1} \cdot 0^{c_2} 1^{d_1} \cdot 1^{d_2} 0^e$$

.

Let M' denote the language $\{0^a 1^c 2^e : \overline{C'} \text{ holds.}\}$, where C' is the condition $(a = c = e) \vee ((a \leq c) \wedge (2a = c + e)) \vee ((e \leq c) \wedge (a + c = 2e))$ and $\overline{C'}$ is the negation of C'.

Lemma 2.3.4. *If L_{NAC} is context-free, then M' is context-free.*

Proof. If L_{NAC} is context-free, so is M by Proposition 1.3.1. Then, we construct a finite-state transducer converting M to M'. This transducer removes all 1's in the input, and converts each 0 to 1 (resp., 2) in the second (resp., third) block of 0. By Proposition 1.3.2, M' is context-free. □

With Lemma 2.3.4, it suffices to show that M' is not context-free. We tried to apply Schaeffer's idea to prove M' is not context-free; however, it doesn't thoroughly solve the problem. Schaeffer's technique finds certain linear set (X as in Theorem 2.2.7) and on the other hand shows this set is not linear, which leads to a contradiction. We alter Schaeffer's technique by showing directly that M' cannot be a finite union of linear sets with stratified periods.

Theorem 2.3.5. *The language M' is not context-free.*

Proof. Suppose M' is context-free. By Theorem 1.3.4, $\varphi(M')$ is a finite union of linear sets with stratified periods, say $\bigcup_i X_i$, where $X_i = \{v_i + A_i u : u \in \mathbb{N}^{m_i}\}$. Let $P_1 = P(\mathbf{0}, n_1)$ and $P_2 = P(\mathbf{0}, n_2)$ be two planes, where $\mathbf{0}$ denotes $\{0, 0, 0\}$, $n_1 = \{1, 1, -2\}$ and $n_2 = \{-2, 1, 1\}$. Let $S^+ = P_1^+ \cap P_2^+$. It is easy to verify that the following three conditions hold:

1. $S^+ \subseteq \varphi(M')$.
2. $P_1 \cap \varphi(M') = \emptyset$.
3. $P_2 \cap \varphi(M') = \emptyset$.

Choose l a natural number such that if w is a column of A_i (for any i) then both $w \cdot n_1$ and $w \cdot n_2$ divides l. Let $T = (l\mathbb{N})^3$. By Lemma 2.3.2, for any X_i with $X_i \cap S^+ \cap T \neq \emptyset$, every column w of X_i satisfies the condition $w \cdot n_1 \geq 0$ and $w \cdot n_2 \geq 0$.

Choose any $v \in S^+$ which minimizes $v \cdot n_1$. We name this v by $\hat{v}$. Note that $\hat{v}$ exists since every $v \in P_1$ satisfies that $v \cdot n_1 > 0$.

Let $P' = P(\hat{v}, n_1)$ be a plane. Let $W = P' \cap P_2^+ \cap T$. Clearly $W \subseteq S^+$. By slightly altering Proposition 2.3.1, we can see that $\dim(W) = 2$. We also observe that

$$\begin{aligned}\dim(W \cap \varphi(M')) &= \dim(W \cap \bigcup_i X_i) \\ &= \dim(\bigcup_i (W \cap X_i))\end{aligned}$$

By Proposition 2.2.3, there exists some X_i such that $\dim(W \cap X_i) = \dim(W) = 2$. Then $\dim(P' \cap X_i) = 2$. Let $x \in P' \cap X_i$. We have $(x - \hat{v}) \cdot n_1 = 0$ and there exists $u \in \mathbb{N}^3$ such that $x = A_i u + v_i$. Hence, $(A_i u + v_i - \hat{v}) \cdot n_1 = 0$, which implies $(A_i u) \cdot n_1 = \hat{v} \cdot n - v_1 \cdot n$. By the definition of $\hat{v}$ we immediately get that $\hat{v} \cdot n - v_1 \cdot n \leq 0$; thus

$$(A_i u) \cdot n_1 \leq 0.$$

On the other hand, let $a_1, a_2, \cdots, a_k$ be the column vectors of A_i. Then $(A_i u) \cdot n_1 = \sum_{j=1}^{k} u_j a_j \cdot n_1$. Since every column a of X_i satisfies the condition $a \cdot n_1 \geq 0$, we obtain the inequality

$$(A_i u) \cdot n_1 \geq 0.$$

Thus we have $\sum_{j=1}^{k} u_j a_j \cdot n_1 = 0$. It follows that for any $1 \leq j \leq k$ if $a_j \cdot n_1 \neq 0$, then $u_j = 0$.

Let I denote the set of columns vectors a_j of A_i such that $a_j \cdot n_1 = 0$. Let $a = (b, c, d)$ be any element in I. Then $b + c = 2d$. Since the column vectors of A_i form a stratified linear set, at least one of b, c, d is zero. If $b = 0$, then $c = 2d$ and $a = s(0, 2, 1)$ for some $s > 0$. If $c = 0$, then $b = 2d$; thus $a \cdot n_2 = -2b + c + d < 0$, which leads to a contradiction. The element d cannot vanish since otherwise a vanishes. Thus, for any $j \in I$ we have $a_j = s_j(0, 2, 1)$ for some $s_j > 0$.

Now we consider $\dim(P' \cap X_i)$. It is easy to see that

$$P' \cap X_i = \{Au + v_i : u_j = 0 \text{ for } j \notin I\}.$$

Thus, we can show that any $x \in P' \cap X_i$ can be written in the form $v_i + s(0, 2, 1)$ for some $s > 0$. It follows immediately that $\dim(P' \cap X_i) \leq 1$ which contradicts the fact that $\dim(P' \cap X_i) = 2$.

Finally, we obtain that M' is not context-free. □

By Theorem 2.3.5, M' is not context-free. So $L_{NAC} \cap R_3$ is not context-free. Thus,

Corollary. *L_{NAC} is not context-free.*

Chapter 3

Representable sets of words of equal length

This chapter primarily discusses representability of set of words of equal length. We focus on the binary alphabet in this chapter. Let $\Sigma = \{0, 1\}$. Recall that $\mathring{R}_n$ is the number of circularly representable subsets of Σ^n. Much of the content of this chapter is taken verbatim from our paper [20]

3.1 Bounds on the size of $\mathring{R}_n$

In this section, we give lower and upper bounds on the size of $\mathring{R}_n$, both of which are of the form α^{2^n}. Our lower bound has $\alpha = \sqrt{2}$ while our upper bound has $\alpha = \sqrt[4]{10}$. Note that our lower bound also works for the size of R_n, since every circularly representable subset is also representable.

3.1.1 Lower bound

Our argument for the lower bound derives from constructing a set of circularly representable subsets.

Proposition 3.1.1. *Let b_n be any de Bruijn word of order n. Then $|C_{n+1}(b_n)| = 2^n$.*

Proof. Every de Bruijn word of order n is of length 2^n; thus there are 2^n length-$(n+1)$ factors of b_n (considered circularly). These length-$(n+1)$ factors are pairwise distinct, for

if $w \in \Sigma^{n+1}$ appears more than once as a factor of b_n, then $w[1..n]$ appears more than once as a factor of b_n. However, every length-n factor appears only once in b_n, a contradiction. Hence $|C_{n+1}(b_n)| = 2^n$. □ □

Lemma 3.1.2. *Given a de Bruijn word b_n, let Y denote the set $\Sigma^{n+1} \backslash C_{n+1}(b_n)$. For any $y \in Y$, the set $\{y\} \cup C_{n+1}(b_n)$ is circularly witnessed by a word w for which both the length-2^n prefix and the length-2^n suffix equal b_n.*

Proof. We construct such a witness for $\{y\} \cup C_{n+1}(b_n)$.

Let $w = b_n b_n b_n b_n$. Let $y_1 = y[1..n]$ and $y_2 = y[2..n+1]$. Let i_1 denote the index of the first occurrence of y_1 in w; namely, the index i_1 is the minimal integer such that $y_1 = w[i_1..i_1 + n - 1]$. Let i_2 denote the index of the last occurrence of y_2 in w; namely, the index i_2 is the maximal integer such that $y_2 = w[i_2..i_2 + n - 1]$.

We argue that the first occurrence of y_1 does not overlap the last occurrence of y_2. We have $i_1 \leq 2^n$, since every possible factor of length n appears in the circular word b_n. Similarly, we obtain $i_2 > 3 \cdot 2^n - n$. Thus we have

$$i_1 + n - 1 - i_2 < -2 \cdot 2^n + 2n - 1 < 0,$$

and hence the first occurrence of y_1 does not overlap the last occurrence of y_2.

Now consider the circular word

$$w_y = b_n b_n w[1..i_1 - 1] w[i_1..i_1 + n - 1] w[i_2 + n - 1] w[i_2 + n..2^{n+2}] b_n b_n.$$

We argue that w_y is a witness for $\{y\} \cup C_{n+1}(b_n)$. For one direction, every element of $\{y\} \cup C_{n+1}(b_n)$ appears as a length-$(n+1)$ factor of w_y. This is a consequence of the following two facts:

1. $b_n b_n$ witnesses $C_{n+1}(b_n)$.
2. $w[i_1..i_1 + n - 1] w[i_2 + n - 1] = y[1..n] y[n+1] = y$.

For the other direction, we can see that all factors of length $n+1$ in w_y are elements of $\{y\} \cup C_{n+1}(b_n)$ by inspection. Note that the length-2^n prefix and the length-2^n suffix of w_y both equal b_n. Hence we conclude that there exists a word for which the prefix and the suffix equal b_n and this circular word circularly witnesses $\{y\} \cup C_{n+1}(b_n)$. □

As an example, we let $n = 2$. One of the de Bruijn words of order 2 is $b_2 = 0011$. We have $C_3(b_2) = \{001, 011, 110, 100\}$. Thus $Y = \{000, 010, 101, 111\}$. Let $y = 010$. The following circular word demonstrates that the set $\{y\} \cup C_{n+1}(b_n)$ is representable:

$$w_{010} = (\underbrace{00110011}_{b_2b_2})(\underbrace{0}_{w[1..i_1-1]})(\underbrace{01}_{w[i_1..i_1+n-1]=y_1})(\underbrace{0}_{w[i_2+n-1]})(\underbrace{011}_{w[i_2+n..2^{n+2}]})(\underbrace{00110011}_{b_2b_2}).$$

Proposition 3.1.3. *Given a de Bruijn word b_n, let Y denote the set $\Sigma^{n+1} \backslash C_{n+1}(b_n)$. For any subset $S \subseteq Y$, the set $S \cup C_{n+1}(b_n)$ is a circularly representable subset of Σ^{n+1}.*

Proof. We have proved this proposition for the case where $|S| = 1$ by Lemma 3.1.2. Now we turn to the general case. Let $S = \{s_1, s_2, \ldots, s_m\}$. By Lemma 3.1.2, for each $1 \leq i \leq m$, there exists a circular word w_i that witnesses $\{s_i\} \cup C_{n+1}(b_n)$ and both the prefix and the suffix of w_i equal b_n. We argue that the circular word $w_S = w_1 w_2 \cdots w_m$ witnesses $S \cup C_{n+1}(b_n)$.

First, for any $1 \leq i \leq m$, s_i appears in w_i and thus in w_S. Moreover, every element of $C_{n+1}(b_n)$ appears in the prefix of w_S: $b_n b_n$. Thus, it suffices to show that every length-$(n+1)$ factor of w_S is a member of $S \cup C_{n+1}(b_n)$. This is shown by the fact that for any $1 \leq i < m$, both the suffix of w_i and the prefix of w_{i+1} equal b_n, which implies that the concatenation of t_i and t_{i+1} does not produce any new factor of length $n+1$ in w_S.

Thus, we conclude that for any subset S of Y, there exists a witness for the set $S \cup C_{n+1}(b_n)$. □

Corollary. *A lower bound for the size of $\mathring{R}_{n+1}$ is $2^{2^n} = \sqrt{2}^{2^{n+1}}$.*

3.1.2 Upper bound

An obvious upper bound for $|\mathring{R}_n|$ is 2^{2^n}, since $\mathring{R}_n \subseteq 2^{\Sigma^n}$, where $|2^{\Sigma^n}| = 2^{2^n}$. In this section, we will show that a tighter upper bound is α^{2^n}, where $\alpha = \sqrt[4]{10}$.

Let $S \subseteq \Sigma^{n+1}$ and $T \subseteq \Sigma^n$. We say that S *is incident on* T if there exists a circular word w such that w witnesses both S and T. For example, we fix $n = 3$. Let $w = 0110$. Then w is a witness for the set $S = \{0110, 1100, 1001, 0011\} \in \mathring{R}_4$ and $T = \{011, 110, 100, 001\} \in \mathring{R}_3$. It follows that S is incident on T. Note that $w' = 01100110$ is also a witness for S, and a witness for T as well.

In fact we can argue that if S is incident on T, then every word that witnesses S also witnesses T.

Proposition 3.1.4. *Every set $S \in \mathring{R}_{n+1}$ is incident on exactly one set in $\mathring{R}_n$.*

Proof. Let

$$T = \{t \in \Sigma^n : \exists w \in S \text{ such that } t \text{ is a length-}n \text{ prefix or suffix of } w\}.$$

Then a word w which witnesses S also witnesses T. Thus S is incident on T. Moreover, if S is incident on T and T', then every witness of S must also witness T and T'. Thus we have $T = T'$. So we conclude that every set $S \in \mathring{R}_{n+1}$ is incident on exactly one set in $\mathring{R}_n$. □

Now we give a partition of $\mathring{R}_{n+1}$. Let

$$\mathring{R}_{n+1}[T] = \{S \in \mathring{R}_{n+1} : S \text{ is incident on } T\}.$$

Proposition 3.1.4 implies that $\{\mathring{R}_{n+1}[T]\}_{T \subseteq \Sigma^n}$ is a pairwise disjoint partition of the set $\mathring{R}_{n+1}$. Namely, (1) for every $T_1 \neq T_2$, we have $\mathring{R}_{n+1}[T_1] \cap \mathring{R}_{n+1}[T_2] = \emptyset$ and (2) $\bigcup_{T \in \mathring{R}_n} \mathring{R}_{n+1}[T] = \mathring{R}_{n+1}$.

Thus we have $|\mathring{R}_{n+1}| = \sum_{T \subseteq \Sigma^n} |\mathring{R}_{n+1}[T]|$. So to give an upper bound for $|\mathring{R}_{n+1}|$, it suffices to give a upper bound for the size of $\mathring{R}_{n+1}[T]$.

Let x be a word of length n. We say that $P_x = \{0x, 1x\}$ is a *pair* of order n w.r.t. x, that $S_x = \{0x, 1x, x0, x1\}$ is a *skeleton* of order n w.r.t. x, and $N_x = \{0x0, 0x1, 1x0, 1x1\}$ is a *net* of order n w.r.t. x. We also say that a set S contains P_x (resp., S_x and N_x) if $P_x \subseteq S$ (resp., $S_x \subseteq S$ and $N_x \subseteq S$).

For any $T \subseteq \Sigma^n$, let $\sigma(T)$ denote the number of skeletons of order $n-1$ in T and let $\rho(T)$ denote the number of pairs of order $n-1$ in T. We have the following proposition:

Proposition 3.1.5. *For any $T \subseteq \Sigma^n$, we have $|\mathring{R}_{n+1}[T]| \leq 7^{\sigma(T)}$.*

Before giving the proof for Proposition 3.1.5, we introduce another definition.

We say that a set R is *feasible* for a set $T \subseteq \Sigma^n$ if there exists $S \in \mathring{R}_{n+1}[T]$ such that $R \subseteq S$.

We observe that $\Sigma^{n+1} = \bigcup_{x \in \Sigma^{n-1}} N_x$ and thus any subset $S \subseteq \Sigma^{n+1}$ is a disjoint union of subsets of nets of order $n-1$. Formally, for any subset $S \subseteq \Sigma^{n+1}$, we have $S = \bigcup_{x \in \Sigma^{n-1}} R_x$, where $R_x \subseteq N_x$.

Proof of Proposition 3.1.5. Let F_x denote the set of feasible subsets (for T) of the net N_x. If $S \in R_{n+1}[T]$, then S is a disjoint union of feasible subsets (for T) of nets. Thus we have $|R_{n+1}[T]| \leq \prod_{x\in\Sigma^n} |F_x|$. In order to prove this proposition, it now suffices to show that for any $x \in \Sigma^{n-1}$, the following condition holds.

- if $S_x \subseteq T$, then $|F_x| \leq 7$;
- otherwise $|F_x| \leq 1$.

For any $x \in \Sigma^{n-1}$, we consider all the possible feasible subsets of N_x. Let F denote any feasible subset of N_x.

- For the first case where $S_x \subseteq T$, we have the following properties:

  1. Either $0x0 \in F$ or $0x1 \in F$ since $0x \in T$;
  2. Either $1x0 \in F$ or $1x1 \in F$ since $1x \in T$;
  3. Either $0x0 \in F$ or $1x0 \in F$ since $x0 \in T$;
  4. Either $0x1 \in F$ or $1x1 \in F$ since $x1 \in T$.

  Hence we have at most 7 possible feasible subsets of N_x which are listed as follows: $\{0x0, 1x1\}$, $\{0x0, 0x1,1x1\}$, $\{0x0, 1x0, 1x1\}$, $\{0x0, 0x1, 1x0, 1x1\}$, $\{0x0, 0x1, 1x0\}$, $\{0x1, 1x0\}$, $\{0x1, 1x0, 1x1\}$. Thus $|F_x| \leq 7$.

- For the second case where $S_x \not\subseteq T$, we argue that $|F_x| \leq 1$. Without loss of generality, suppose $0x \notin T$. It follows that:

  1. $0x0$ and $0x1$ cannot occur in F since $0x \notin T$;
  2. $1x0 \in F$ if and only if $x0 \in T$;
  3. $1x1 \in F$ if and only if $x1 \in T$;

  Hence, F is fixed. It follows that $|F_x| \leq 1$.

By finishing the argument on the above two cases, we conclude that $|\mathring{R}_{n+1}[T]| \leq 7^{\sigma(T)}$. □

Now, we are close to the core part. Instead of computing the number of skeletons, which is quite complex, we consider the number of pairs.

Proposition 3.1.6. *The size of the set* $|\mathring{R}_{n+1}|$ *is bounded by* $10^{2^{n-1}}$.

Proof. Let $L_{k,i}$ denote the number of subsets $T \in \mathring{R}_n$, such that $|T| = k$ and $\rho(T) = i$. There are in total 2^{n-1} pairs in Σ^n, and we first choose i pairs from them. Then, we choose the other $k - 2i$ elements which do not form any pair from the remaining $2^n - 2i$ elements (which forms $2^{n-1} - i$ pairs); it is equivalent to pick $k-2i$ pairs from the remaining $2^{n-1} - i$ pairs and randomly choose one element from each selected pair. Thus, we have

$$L_{k,i} = \binom{2^{n-1}}{i}\binom{2^{n-1}-i}{k-2i}2^{k-2i}.$$

Note that $k \geq 2i$ since a set of k elements can contain at most $\lfloor \frac{k}{2} \rfloor$ pairs and the term $L_{k,i}$ vanishes when $k - 2i > 2^{n-1} - i$. Thus we have

$$|\mathring{R}_{n+1}| = \sum_{T \subseteq \Sigma^n} |\mathring{R}_{n+1}[T]| \leq \sum_{k=0}^{2^n} \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} L_{k,i} 7^i.$$

The inequality holds since we count the number of pairs instead of the number of skeletons and the number of pairs is always greater than or equal to the number of skeletons. Then we can see that

$$|\mathring{R}_{n+1}| \leq \sum_{k=0}^{2^n} \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} \binom{2^{n-1}}{i}\binom{2^{n-1}-i}{k-2i}2^{k-2i}7^i \quad \leq \sum_{i=0}^{2^{n-1}} \binom{2^{n-1}}{i} 7^i \sum_{k=2i}^{2^n} \binom{2^{n-1}-i}{k-2i}2^{k-2i}$$

by writing $L_{k,i}$ in closed form. Note that

$$\sum_{k=2i}^{2^n} \binom{2^{n-1}-i}{k-2i}2^{k-2i} = \sum_{k=0}^{2^n-2i} \binom{2^{n-1}-i}{k}2^k = \sum_{k=0}^{2^{n-1}-i} \binom{2^{n-1}-i}{k}2^k = 3^{2^{n-1}-i}.$$

So we have

$$|\mathring{R}_{n+1}| \leq \sum_{i=0}^{2^{n-1}} \binom{2^{n-1}}{i} 7^i 3^{2^{n-1}-i} = 10^{2^{n-1}}.$$

□

Proposition 3.1.6 directly implies the upper bound we claimed in the beginning of this section.

3.2 Shortest witness

Recall that μ_n (resp., ν_n) is the maximum length of the shortest non-circular witness (resp., circular witness) over all subsets of Σ^n. The quantities of μ_n and ν_n are of interest since we can enumerate all sequences of length less than or equal to μ_n (resp., ν_n) in order to list all the representable (resp., circularly representable) subsets of Σ^n. In this section we obtain an upper bound on μ_n and ν_n.

We need the following result of Hamidoune [11, Prop. 2.1]. Since the result is little-known and has apparently not appeared in English, we give the proof here. By a *Hamiltonian walk* we mean a closed walk, possibly repeating vertices and edges, that visits every vertex.

Proposition 3.2.1. *Let $G = (V, E)$ be a directed graph on n vertices. If G is strongly connected (that is, if there is a directed path from every vertex to every vertex), then there is a Hamiltonian walk of length at most $\lfloor (n+1)^2/4 \rfloor$. Furthermore, this bound is best possible.*

Proof. Let L be a longest simple path in G. (A simple path does not repeat edges or vertices.) Let $V - L = \{v_i : 1 \leq i \leq k\}$. Let v_0 be the last vertex in L and v_{k+1} be the first vertex in L. Let L_i be a simple path from v_i to v_{i+1}. Then a Hamiltonian walk W is obtained by following the edges in $L_0, L_1, \ldots, L_k$, and then those in L. So the number of edges in W is at most $(k+2)|L| = |L|(n+1-|L|)$. But it is easy to see that $r(n+1-r)$ is maximized when $r = \lceil n/2 \rceil$, so $r(n+1-r) = \lfloor (n+1)^2/4 \rfloor$, as claimed.

To see that this bound is best possible, consider a graph where there is a directed chain of $\lfloor n/2 \rfloor$ vertices, where the last vertex has a directed edge to $\lceil n/2 \rceil$ other vertices, and each of those vertices have a single directed edge back to the start of the chain. The shortest walk covering all the vertices traverses the chain, then an edge to one of the other vertices, then a single edge back, and repeats this $\lceil n/2 \rceil$ times. The total length is then $(\lfloor n/2 \rfloor + 1)\lceil n/2 \rceil = \lfloor (n+1)^2/4 \rfloor$. So the bound is tight. □

From this we immediately get

Proposition 3.2.2. *An upper bound for μ_n and ν_n is $2^{2n-2} + 2^{n-1}$.*

3.3 Fixed-length witnesses

Restriction on the length of a witness leads us to another interesting problem. Let $T(t, n)$ denote the number of subsets of $\{0,1\}^n$ witnessed by some word of length $t \geq n$. Is there

any characterization of $T(t,n)$? We focus on ordinary (non-circular) words for this question and derive a closed-form formula for $T(t,n)$ in the case where $n \leq t < 2n$.

In order to compute $T(t,n)$, we consider the number of words that witness the same subset of Σ^n. Suppose $S \subseteq \Sigma^n$. Let $C_t(S)$ denote the number of words of length t that witness S. Then we have

$$T(t,n) = 2^t - \sum_{\substack{S \subseteq \Sigma^n \\ C_t(S) > 1}} (C_t(S) - 1).$$

It suffices to characterize what subsets S satisfy $C_t(S) > 1$ and to determine $C_t(S)$.

For $t < 2n$, we have such a characterization by Theorem 3.3.1 below. Before stating the theorem, we first introduce some notation.

Let w be a word. Let $\mathrm{Pref}(w)$ denote the set of prefixes of w. A *period* p of w is a positive integer such that w can be factorized as

$$w = s^k s', \text{ with } |s| = p, \ s' \in \mathrm{Pref}(s), \text{ and } k \geq 1.$$

Let $\pi(w)$ denote the minimal period of w.

The *root* of a word w is the prefix of w with length $\pi(w)$. Let $r(w)$ denote the root of w. Two words w and w' are *conjugate* if there exist $u, v \in \Sigma^*$ such that $w = uv$ and $w' = vu$; w and w' are *root-conjugate* if their roots $r(w)$ and $r(w')$ are conjugate.

The following theorem is crucial for our work and of independent interest.

Theorem 3.3.1. *Let t, n, k be such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. Let w and w' be distinct words of length t over an arbitrary alphabet. Then $F_n(w) = F_n(w')$ iff $\pi(w) = \pi(w') \leq k + 1$ and w, w' are root-conjugate.*

One direction is easy: if w and w' are root-conjugate with period $p \leq k+1$, then there are p places to begin, and considering consecutive factors of length $n+p-1$ gives exactly p distinct length-n factors.

For the other direction, we need three lemmas.

Lemma 3.3.2. *(Fine-Wilf theorem [8, Theorem 1]) Let w_1, w_2 be two words. If w_1 and w_2 have a common prefix of length $\pi(w_1) + \pi(w_2) - 1$, then $r(w_1) = r(w_2)$.*

Lemma 3.3.3. *For any $w \in \Sigma^+$, if there exists a factorization $w = xyz$ such that $xy = yz$ and $x, y, z \in \Sigma^+$, then w is periodic with $\pi(w) \leq |x|$.*

Proof. By the Lyndon-Schützenberger theorem [15, Lemma 2], there exist $u \in \Sigma^+, v \in \Sigma^*$ and an integer $e \geq 0$ such that $x = uv, y = (uv)^e u, z = vu$. Thus $w = (uv)^{e+2}u$. Thus w is periodic with $\pi(w) \leq |x|$. □

Lemma 3.3.4. *Let t, n, k be integers such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. Let w be a word of length t with $\pi(w) \leq k + 1$. If w' is any word such that $F_n(w) = F_n(w')$, then w and w' are root-conjugate.*

Carpi and de Luca proved a stronger proposition [4, Proposition 6.2] which directly implies this lemma. We first introduce some relevant notation from that paper.

A factor s of a word w is said to be *right-special* in w if there exist two distinct symbols a and b such that sa and sb are factors of w. Let R_w denote the minimal length m such that there exists no factor of length m that is right-special.

A factor s of a word w is said to be *right-extendable* (resp., *left-extendable*) in w if there exists a symbol a such that sa is a factor of w (resp., as is a factor of w). Let K_w and H_w denote the length of the shortest factor which is not right-extendable (resp., left-extendable).

A word is *semiperiodic* if $R_w < H_w$.

proof of Lemma 3.3.4. Carpi and de Luca proved [4, Lemma 3.2] that $\pi(w) > R_w$. Also, we have $H_w \geq \pi(w)$ since the length-$(\pi(w) - 1)$ prefix of w is left-extendable. Thus w is semiperiodic. Moreover we have $F_n(w) = F_n(w')$ where $n \geq k + 1 \geq \pi(w) \geq 1 + R_w$. Then we can apply [4, Proposition 6.2] to prove this lemma. □

proof of Theorem 3.3.1. We give a proof for Theorem 3.3.1 by induction on k.

The base case is when $k = 0$. In this case $t = n$ and thus $F_n(w) = \{w\}$ and $F_n(w') = \{w'\}$. Thus $w = w'$, contradicting the fact that w and w' are distinct.

Now we deal with the induction step. We assume the result holds for $k - 1$ and we prove it for k. For convenience, we let $p_i(w)$ denote the length-i prefix of the word w; let $s_i(w)$ denote the length-i suffix of the word w.

We first consider the case where $H_w < n$. We have $p_n(w) \in F_n(w) = F_n(w')$. If $p_n(w) \neq p_n(w')$, then there exists $a \in \Sigma$ such that $ap_{n-1}(w) \in F_n(w')$. Thus we have $ap_{n-1}(w) \in F_n(w)$ which leads to the contradiction that $H_w \geq |ap_{n-1}(w)| = n$. Hence $p_n(w) = p_n(w')$.

Now let $s = w[2..t]$ and $s' = w'[2..t]$. Clearly $|s| = |s'| = t - 1$. The prefix $p_n(w)$ appears only once as a factor of w, otherwise $p_{n-1}(w)$ is left-extendable in w which contradicts the

fact that $H_w < n$. Thus we have $F_n(s) = F_n(w)\backslash\{p_n(w)\}$. Similarly we have $F_n(s') = F_n(w')\backslash\{p_n(w)\}$. Thus $F_n(s) = F_n(s')$. Let $k' = k - 1$. We have $t - 1 = n + k - 1 = n + k'$ and $n \geq k + 1 > k' + 1$. By induction, we have either

Case 1: $s = s'$; or

Case 2: s and s' are root-conjugate and $\pi(s) = \pi(s') = \rho$, where $\rho \leq k' + 1 = k$.

In Case 1, it follows that $w = w'$, contradicting the fact that w, w' are distinct. In Case 2, we prove that $s = s'$ by showing that their roots are identical. Suppose s and s' have a common prefix of length d. We have $d \geq n - 1$, since w and w' have a common prefix of length at least n. If $d \geq \rho$, then the root of s is identical to the root of s'. Otherwise, we have the chain of inequalities $k \geq \rho \geq d + 1 \geq n \geq k + 1$, which is trivially a contradiction. Thus neither Case 1 nor Case 2 can occur and we are done with the case where $H_w < n$.

Similarly we can prove the induction step when $K_w < n$. Thus it suffices to consider the case where $H_w \geq n$ and $K_w \geq n$. We first claim $\pi(w) \leq k + 1$. There are several cases to settle:

- The first case is when $p_{n-1}(w) = s_{n-1}(w)$ and the occurrence of $p_{n-1}(w)$ and $s_{n-1}(w)$ do not overlap; namely we have $w = p_{n-1}(w)Lp_{n-1}(w)$, where $L \in \Sigma^*$. We have the inequality $n + k = t = |w| = 2|p_{n-1}(w)| + |L| = 2(n - 1) + |L|$. Thus $|L| = k + 2 - n$. Hence $\pi(w) \leq |p_{n-1}(w)L| = n - 1 + k + 2 - n = k + 1$.

- The second case is when $p_{n-1}(w) = s_{n-1}(w)$ and these occurrences overlap. Formally we put it as follows: there exist $x, y, z \in \Sigma^+$, such that $p_{n-1}(w) = xy = yz$ and $w = xyz$. It follows that $\pi(w) \leq |x| \leq k + 1$ by Lemma 3.3.3.

- The last case is when $p_{n-1}(w) \neq s_{n-1}(w)$. Let i_p denote the index of the last occurrence of $p_{n-1}(w)$; namely $i_p = \sup\{i \geq 1 : p_{n-1}(w) = w[i..i + n - 2]\}$. Note that i_p exists since $p_{n-1}(w)$ is left-extendable and $i_p \leq t - n + 2$ since $p_{n-1}(w) \neq s_{n-1}(w)$. We argue that $w_1 = w[1..i_p + n - 2]$ is periodic with $\pi(w_1) \leq i_p - 1 \leq k$. If the first occurrence of $p_{n-1}(w)$ (the prefix of w) overlaps the last occurrence of $p_{n-1}(w)$, then by Lemma 3.3.3, we see that w_1 is periodic with $\pi(w_1) \leq i_p - 1 \leq k$. Otherwise, we have $2(n - 1) \leq |w_1| \leq t - 1$; thus $k = n - 1$ and $|w_1| = 2(n - 1)$. Then we have $w_1 = p_{n-1}(w)p_{n-1}(w)$, where w_1 is periodic with $\pi(w_1) \leq n - 1 = i_p - 1 = k$. For both cases, we have w_1 is periodic with $\pi(w_1) \leq i_p - 1 \leq k$.

  Similarly we let i_q denote the index of the first occurrence of $s_{n-1}(w)$ and $w_2 = w[i_q..t]$. We have $1 < i_q \leq t - n + 2$ and $\pi(w_2) \leq t - n + 2 - i_q$. The factors w_1 and w_2 overlap

for at least $|w_1| + |w_2| - t \geq \pi(w_1) + \pi(w_2) - 1$ symbols. Let D denote the overlap of w_1 and w_2. We have $|D| \geq \pi(w_1) + \pi(w_2) - 1$. Also $\pi(w_1)$ is a period of D since $|D| \geq \pi(w_1)$ and D can be factorized as

$$D = d^l d', \text{ where } d \text{ is conjugate to the root of } w_1, \ d' \in \mathrm{Pref}(d), \text{ and } l \geq 1.$$

By Lemma 3.3.2, the overlap D has the same root as w_2. Since root-conjugacy is an equivalence relation, we have w_1 and w_2 are root-conjugate. Let l_1 denote the length of the root of w_1. We argue that w is periodic with $\pi(w) \leq l_1 \leq k+1$ by the fact that l_1 is also a period of w. It suffices to show that $w[l_1 + i] = w[i]$ for $1 \leq i \leq t - l_1$. For the case where $1 \leq i \leq |w_1| - l_1$, we have $w[i + l_1] = w_1[i + l_1] = w_1[i] = w[i]$; for the other case where $|w_1| - l_1 < i \leq t - l_1$, we have $w[i + l_1] = w_2[i + l_1 - i_q + 1] = w_2[i - i_q + 1] = w[i]$. Thus, we see that w is periodic with $\pi(w) \leq k + 1$.

Finally by Lemma 3.3.4, we get that w and w' are root-conjugate and their periods $\pi(w) = \pi(w') \leq k + 1$. By all cases, we finish the induction and complete the proof of Theorem 3.3.1. □

The following corollary gives $T(t, n)$ when $t < 2n$.

Corollary. *For $n \leq t < 2n$, we have $T(t, n) = 2^t - \sum_{k=1}^{t-n+1} \frac{k-1}{k} \sum_{d|k} \mu(\frac{k}{d}) 2^d$, where $\mu(\cdot)$ is the Möbius function.*

Proof. Let $k = t - n$. We have $n \geq t - n + 1 = k + 1$. By Theorem 3.3.1, we know that for any set $S \subseteq \Sigma^n$, $C_t(S) > 1$ if and only if there exists a word w that witnesses S with $\pi(w) \leq k + 1$. In this case we have $C_t(S) = \pi(w)$; that is, the set of words that witness S is the same as the set of the words that are root-conjugate to w. Thus each S such that $C_t(S) > 1$ corresponds to a set of root-conjugate words, which can be represented by their lexicographically least roots (the Lyndon words).

Thus we have

$$\begin{aligned} T(t, n) &= 2^t - \sum_{\substack{S \subseteq \Sigma^n \\ C_t(S) > 1}} (C_t(S) - 1) = 2^t - \sum_{\substack{w \text{ is a Lyndon word} \\ \pi(w) \leq k+1}} (\pi(w) - 1) \\ &= 2^t - \sum_{i=1}^{k+1} (i - 1) \cdot L(i), \end{aligned}$$

where $k = t - n$ and $L(i) = \frac{1}{i} \sum_{d|i} \mu(\frac{i}{d}) 2^d$ is the number of Lyndon words of length i. □

3.4 Numerical results

To finish this chapter, we give some tables listing several numerical results.

It is not feasible to enumerate every single word to verify whether a subset is circularly representable (or non-circularly representable). For this reason, we exploit ideas from graph theory.

Formally, we define $G_n = (V_n, E_n)$, where

$$V_n = \{(S, u, v) : S \subseteq \Sigma^n \text{ and } u, v \in \Sigma^n\} \text{ and}$$
$$E_n = \{((S, u, v), (S \cup \{x\}, u, x)) : S \subseteq \Sigma^n,\ u, v, x \in \Sigma^n,\ \text{and } v[2..n] = x[1..n-1]\}.$$

We say that a node (S, u, v) is *valid* if S is witnessed by a non-circular word w for which the length-n prefix is u and the length-n suffix is v.

We use a breadth-first search strategy to compute all the possible valid nodes in G_n. Let I denote a subset of nodes $\{(\{u\}, u, u) : u \in \Sigma^n\}$ in G_n. Nodes in G_n that are connected to any node in I can be proven valid by induction. Thus, a breadth-first search begins with the subset I and enumerates all nodes that are connected to nodes in I.

The relation between valid nodes in G_n and non-empty representable subsets of Σ^n is that any subset $S \subseteq \Sigma^n$ is representable if and only if there exist $u, v \in \Sigma^n$ such that (S, u, v) is valid. This relation can be proved by induction. Similarly, any subset $S \subseteq \Sigma^n$ is circularly representable if and only if there exists $u \in \Sigma^n$ such that (S, u, u) is valid and the minimum distance d between (S, u, u) and nodes in I satisfies the inequality $d \geq n - 1$.

With the above properties, we can enumerate all the possible non-empty representable (or circularly representable) subsets of order n. Our results are shown in the Table 3.1. The last two columns give words w of length ν_n (resp., μ_n) for which no shorter word witnesses $C_n(w)$ (resp., $F_n(w)$).

We present some numerical results for $T(t, n)$ in Table 3.2 and Table 3.3. The numbers in bold follow from Corollary 3.3.

n	$\lvert\mathring{R}_n\rvert$	$\lvert R_n\rvert$	ν_n	μ_n	longest circ. witness	longest witness
1	3	3	2	2	01	01
2	6	14	4	5	0011	00110
3	27	121	9	10	000100111	0001011100
4	973	5921	24	24	000010001011100011101111	000010010101100101101111
5	2466131	20020315	82	77	—	—

Table 3.1: Numerical results on representable subsets

n \ t	1	2	3	4	5	6	7	8
1	**2**	3	3	3	3	3	3	3
2		**4**	**7**	11	12	12	12	12
3			**8**	**15**	**27**	48	72	94
4				**16**	**31**	**59**	**114**	216
5					**32**	**63**	**123**	**242**
6						**64**	**127**	**251**
7							**128**	**255**
8								**256**

Table 3.2: Numerical results on $T(t, n)$

n \ t	9	10	11	12	13	14	15	16
1	3	3	3	3	3	3	3	3
2	12	12	12	12	12	12	12	12
3	100	103	101	103	101	103	101	103
4	391	677	1087	1621	2246	2928	3595	4235
5	**474**	933	1795	3421	6399	11682	20704	35914
6	**498**	**986**	**1965**	3899	7709	15171	29710	57726
7	**507**	**1010**	**2010**	**4013**	**8001**	15969	31789	63256
8	**511**	**1019**	**2034**	**4058**	**8109**	**16193**	**32367**	64671

Table 3.3: Numerical results on $T(t, n)$

Chapter 4

Open problems

We present some open problems that are related to these two combinatorial patterns in this chapter.

1. Is the language of non-abelian k-th powers context-free?

   In this thesis, we prove that the language of non-abelian squares and non-abelian cubes are not context-free. It remains to consider the case where $k \geq 4$.

   Theorems about bounded context-free languages may help solve this problem. To be more precise, a promising method is to construct a bounded regular language R, and prove that the intersection of R with the target language is not bounded context-free. Some really interesting constructions of R are: $w_2^* w_4^* w_1^* w_3^*$ and $w_2^* w_1^* w_4^* w_3^*$, where $w_i = 10^{i-1}$.

2. Does the limit $\lim_{n\to\infty} |\mathring{R}_n|^{\frac{1}{2^n}}$ exist?

3. Find better bounds for μ_n and ν_n.

4. Derive a formula for $T(t, n)$ where $t = 2n$.

   It is easy to see that Theorem 3.3.1 fails for $n < k + 1$. Indeed, it is possible to have $F_n(x) = F_n(y)$ in this case, and yet $\pi(x) \neq \pi(y)$. For example, take $n = k - 1$ so that $t = 2k - 1$, and consider $x = 0^k 10^{k-2}$ and $y = 0^{k-1}10^{k-1}$. Then $F_n(x) = F_n(y)$ but $\pi(x) = k + 1$ and $\pi(y) = k$.

   The remaining case is $n = k$, so that $t = 2k$. We conjecture that if x and y are distinct binary words of length $2n$ with $F_n(x) = F_n(y)$ then $\pi(x) = \pi(y)$ and furthermore

x and y are root-conjugate. However, it is possible in this case that $\pi(x) > n + 1$. Furthermore it seems that if $\pi(x) > n + 1$, then $x = uv01vu$ and $y = uv10v^R u$ (or vice versa) for some nonempty words u, v where u is the longest palindrome prefix of uv and $\pi(x) = t - |v|$.

As an example, consider $x = 010110$, $y = 011010$. Then

$$F_3(x) = F_3(y) = \{010, 011, 101, 110\}$$

but $\pi(x) = \pi(y) = 5$. Here $u = 0$, $v = 1$.

References

[1] F. Blanchet-Sadri and Nathan Fox. Abelian-primitive partial words. *Theoret. Comput. Sci.* **485** (2013), 16–37.

[2] F. Blanchet-Sadri and Sean Simmons. Deciding representability of sets of words of equal length. *Theoret. Comput. Sci.* **475** (2013), 34–46.

[3] Avrim Blum, Tao Jiang, Ming Li, John Tromp, and Mihalis Yannakakis. Linear approximation of shortest superstrings. *J. Assoc. Comput. Mach.* **41** (1994), 630–647.

[4] A. Carpi and A. de Luca. Semiperiodic words and root-conjugacy. *Theoret. Comput. Sci.* **292** (2003), 111–130.

[5] N. G. de Bruijn. A combinatorial problem. *Nederl. Akad. Wetensch., Proc.* **49** (1946), 758–764. (= *Indagationes Math.* **8** (1946), 461–467.)

[6] Michael Domaratzki and Narad Rampersad. Abelian primitive words. *Internat. J. Found. Comput. Sci.* **23** (2012), 1021–1033.

[7] P. Erdős. Some unsolved problems. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221–264.

[8] N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965), 109–114.

[9] J. Gallant, D. Maier, and J. A. Storer. On finding minimal length superstrings. *J. Comput. System Sci.* **20** (1980), 50–58.

[10] S. Ginsburg. The Mathematical Theory of Context-Free Languages. McGraw-Hill, New York, 1966.

[11] Y. O. Hamidoune. Sur les sommets de demi-degré h d'un graphe fortement h-connexe minimal. *C. R. Acad. Sci. Paris Sér. A-B* **286** (1978), A863–A865.

[12] L. Kászonyi. A pumping lemma for DLI-languages. *Discrete Math.* **258** (2002), 105–122.

[13] Veikko Keränen. Abelian squares are avoidable on 4 letters. *Lecture Notes in Comput. Sci.* **623** (1992), 41–52.

[14] Ming Li. Towards a DNA sequencing theory (learning a string)(preliminary version). *31st Annual Symposium on Foundations of Computer Science* **Vol. I, II** (1990), 125–134.

[15] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.

[16] William Ogden. A helpful result for proving inherent ambiguity. *Math. Systems Theory* **2** (1968), 191–194.

[17] L. B. Richmond and Jeffrey Shallit. Counting abelian squares. *Electron. J. Combin.* **16** (2009), Research Paper 72, 9.

[18] Z. Sweedyk. A $2\frac{1}{2}$-approximation algorithm for shortest superstring. *SIAM J. Comput.* **29** (2000), 954–986.

[19] Shuo Tan. The non-abelian squares are not context-free. *Preprint,* http://arxiv.org/pdf/1110.4136v1.pdf (Oct. 2011).

[20] Shuo Tan and Jeffrey Shallit. Sets Represented as the Length-n Factors of a Word. *Preprint,* http://arxiv.org/pdf/1304.3666v1.pdf (Apr. 2013).

[21] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22.