Markus A. Whiteland

On the k-Abelian Equivalence Relation of Finite Words

Turku Centre *for* Computer Science

TUCS Dissertations
No 241, June 2019

On the k-Abelian Equivalence Relation of Finite Words

Äärellisten Sanojen k-Abelin Ekvivalenssirelaatiosta

Markus A. Whiteland

To be presented, with the permission of the Faculty of Science and Engineering of the University of Turku, for public criticism in Agora XXI on June 28, 2019 at 12 noon.

University of Turku
Department of Mathematics and Statistics
FI-20014 Turku, Finland

2019

Supervisors

Juhani Karhumäki
Department of Mathematics and Statistics
University of Turku
FI-20014 Turku
Finland

Svetlana Puzynina
St. Petersburg State University
29B Line 14th (Vasilyevsky Island)
199178 Saint Petersburg
Russia

Sobolev Institute of Mathematics
Prospekt Akademika Koptyuga, 4
630090, Novosibirsk
Russia

Reviewers

Robert Mercaş
Department of Computer Science
Loughborough University
Epinal Way
Loughborough LE11 3TU
United Kingdom

Michel Rigo
Department of Mathematics
University of Liège
Grande traverse 12 (B37)
B-4000 Liège
Belgium

Opponent

Dirk Nowotka
Department of Computer Science
Kiel University
24098 Kiel
Germany

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using Turnitin OriginalityCheck service.

ISBN 978-952-12-3837-6
ISSN 1239-1883

Abstract

This thesis is devoted to the so-called k-abelian equivalence relation of sequences of symbols, that is, words. This equivalence relation is a generalization of the abelian equivalence of words. Two words are abelian equivalent if one is a permutation of the other. For any positive integer k, two words are called k-abelian equivalent if each word of length at most k occurs equally many times as a factor in the two words. The k-abelian equivalence defines an equivalence relation, even a congruence, of finite words. A hierarchy of equivalence classes in between the equality relation and the abelian equivalence of words is thus obtained.

Most of the literature on the k-abelian equivalence deals with infinite words. In this thesis we consider several aspects of the equivalence relations, the main objective being to build a fairly comprehensive picture on the structure of the k-abelian equivalence classes themselves. The main part of the thesis deals with the structural aspects of k-abelian equivalence classes. We also consider aspects of k-abelian equivalence in infinite words.

We survey known characterizations of the k-abelian equivalence of finite words from the literature and also introduce novel characterizations. For the analysis of structural properties of the equivalence relation, the main tool is the characterization by the rewriting rule called the k-switching. Using this rule it is straightforward to show that the language comprised of the lexicographically least elements of the k-abelian equivalence classes is regular. Further word-combinatorial analysis of the lexicographically least elements leads us to describe the deterministic finite automata recognizing this language. Using tools from formal language theory combined with our analysis, we give an optimal expression for the asymptotic growth rate of the number of k-abelian equivalence classes of length n over an m-letter alphabet. Explicit formulae are computed for small values of k and m, and these sequences appear in Sloane's Online Encyclopedia of Integer Sequences.

Due to the fact that the k-abelian equivalence relation is a congruence of the free monoid, we study equations over the k-abelian equivalence classes. The main result in this setting is that any system of equations of k-abelian equivalence classes is equivalent to one of its finite subsystems, i.e., the monoid defined by the k-abelian equivalence relation possesses the compactness property.

Concerning infinite words, we mainly consider the (k-)abelian complexity function. We complete a classification of the asymptotic abelian complexities of pure morphic binary words. In other words, given a morphism which has an infinite binary fixed point, the limit superior asymptotic abelian complexity of the fixed point can be computed (in principle). We also give a new proof of the fact that the k-abelian complexity of a Sturmian word is $n + 1$ for length $n < 2k$, and $2k$ for $n \geqslant 2k$. In fact, we consider several aspects of the k-abelian equivalence relation in Sturmian words using a dynamical interpretation of these words. We reprove the fact that any Sturmian word contains arbitrarily large k-abelian repetitions. The methods used allow to analyze the situation in more detail, and this leads us to define the so-called k-abelian critical exponent which measures the ratio of the exponent and the length of the root of a k-abelian repetition. This notion is connected to a deep number theoretic object called the Lagrange spectrum.

Tiivistelmä

Väitöskirja käsittelee äärellisten symbolijonojen, eli sanojen, k-abelin ekvivalenssirelaatiota. Tämä käsite yleistää sanojen abelin ekvivalenssirelaation. Kaksi sanaa ovat abelin ekvivalentit, jos toinen sanoista on toisen permutaatio. Kaksi sanaa ovat k-abelin ekvivalentit, missä k on positiivinen kokonaisluku, jos jokainen korkeintaan pituutta k oleva sana esiintyy yhtä monta kertaa osasanana kummassakin sanassa. Näin määritelty relaatio on ekvivalenssirelaatio ja jopa kongruenssi. Edelleen, näin saadaan määriteltyä ekviavalenssiluokkien hierarkia abelin ekvivalenssin ja yhtäsuuruusekvivalenssin välillä.

Kirjallisuudessa k-abelin ekvivalenssia on tarkasteltu lähinnä äärettömien sanojen näkökulmasta. Tässä väitöskirjassa ekvivalenssirelaatiota tarkastellaan useista eri näkökulmista, ja päätarkoituksena on ollut antaa melko kattava kuva k-abelin ekvivalenssiluokkien rakenteesta. Niinpä suurin osa väitöskirjasta koskee k-abelin ekvaivalenssiluokkien rakenteen analysointia. Tarkastelemme myös k-abelin ekvivalenssia äärettömien sanojen yhteydessä.

Käymme läpi k-abelin ekvivalenssirelaation tunnetut luonnehdinnat kirjallisuudesta, ja esittelemme myös uusia luonnehdintoja. Esitetyistä luonnehdinnoista tärkein k-abelin ekvivalenssiluokkien rakenteen tarkastelemisen kannalta on uudelleenkirjoitussääntö k-switching. Tämän säännön avulla on suoraviivaista osoittaa, että ekvivalenssiluokkien leksikografisesti pienimpien alkioiden muodostama kieli on säännöllinen. Tarkastelemme syvällisemmin tämän edustajiston rakennetta kielenä, ja tämän analyysin johdosta saadaan kuvailtua kielen hyväksyvien äärellisten deterministen automaattien rakennetta. Käyttämällä formaalisten kielten teorian työkaluja, saadaan pituutta n olevien m-kirjaimisten k-abelin ekvivalenssiluokkien lukumäärälle asymptoottisesti tarkka lauseke. Esittelemme myös eksplisiittisiä lausekkeita kyseisten lukujen laskemiseksi pienillä lukujen k ja m arvoilla. Nämä lukujonot esiintyvät nyt verkkosivustolla Sloane's Online Encyclopedia of Integer Sequences.

Yhtälöitä yli k-abelin ekvivalenssiluokkien on mahdollista tarkastella, sillä k-abelin ekvivalenssi määrää kongruenssin. Väitöskirjassa osoitetaan, että jokainen tällainen yhtälöryhmä on ekvivalentti jonkin äärellisen osayhtälöryhmänsä kanssa. Toisin sanoen k-abelin ekvivalenssirelaation määräämällä monoidilla on kompaktisuusominaisuus.

Äärettömien sanojen yhteydessä tarkastelemme lähinnä (k-)abelin kompleksisuuksia. Täydennämme puhtaasti morfisten binääristen sanojen asymptoottisten abelin kompleksisuuksien ylärajakasvun luokittelun. Toisin sanoen, annettuna morfismi, joka määrää binäärisen kiintopisteen, kyseisen kiintopisteen abelin kompleksisuuden ylärajakasvu voidaan laskea ainakin periaatteessa. Todistamme uudelleen, että Sturmin sanan k-abelin kompleksisuus on $n + 1$, kun $n < 2k$ ja $2k$, kun $n \geqslant 2k$. Tarkastelemme myös muita k-abelin ekvivalenssiin liittyviä käsitteitä Sturmin sanoissa käyttäen Sturmin sanojen luonnehdintaa dynaamisena systeeminä. Annamme vaihtoehtoisen todistuksen sille seikalle, että Sturmin sanoissa esiintyy mielivaltaisen suuria k-abelin toistoja. Käyttämämme menetelmät sallivat aiempaa tarkemman analyysin tilanteesta, ja tämä seikka johdattelee määrittelemään kriittisen k-abelin eksponentin, joka vertaa sanassa esiintyvän k-abelin toiston eksponenttia sen juuren pituuteen. Tämä taas johdattaa syvälliseen lukuteoreettiseen rakenteeseen, niin kutsuttuun Lagrangen spektriin.

Acknowledgments

Finally, the project is almost completed. I could never have done this on my own. It has not been an easy project, and a lot of doubts arose during the work on this project. For the support during these years, I am grateful for the people and organizations who helped me achieve this feat.

I would like to express my deepest gratitude to my supervisors Professor Juhani Karhumäki and Docent Svetlana Puzynina for their guidance and support throughout this project. It has been an enlightening opportunity to work with you and to explore our research field. I would also like to thank you for your patience in me. It has been a long ride, but it is coming to an end. Thank you Juhani, for the experiences in banding ravens. This was something I did not see coming when I applied for doctoral studies in mathematics. Thank you also for the course on combinatorics on words, which woke my interest into this topic. Thank you Svetlana for your scientific advising. I have learned tons from you.

I am grateful for the financial support from the University of Turku Graduate School and the Yrjö, Ville, Kalle Väisälä foundation of the Finnish Academy of Science and Letters. I would also like to thank the doctoral program MATTI for the financial support for traveling to conferences.

I am grateful to Doctor Robert Mercaş and Professor Michel Rigo for agreeing to review this dissertation, and for their suggestions which greatly improved the readability of the text. I am equally grateful to Professor Dirk Nowotka for agreeing to act as the opponent.

I would like to thank Professor Jarkko Kari for acting as my research director. Thank you for the many inspiring discussions on mathematics and the extremely interesting courses in automata theory and cellular automata. They are among the most interesting courses I have taken.

I would like to thank the director of the Department of Mathematics and Statistics, Professor Iiro Honkala, for presenting several interesting teaching opportunities, as well as time for research and finalizing the thesis.

To Doctor Jarkko Peltomäki I express my deepest gratitude for the numerous invigorating discussions on mathematics, life, and all kinds of other stuff, as well as your friendship. It has been a real joy to collaborate with you. Thank you for these years. I hope you feel good about yourself for pushing me over the edge into the world of fountain pens.

I would also like to thank the members of the symbolic dynamics seminar group for interesting and brilliant talks on the subject. Merci Beaucoup Doctor Thibault Godin, Joonatan Jalonen, Johan Kopra, Etienne Moutot, and Doctor Ville Salo. I would like to thank the several colleagues who have shared an office with me. Thank you Doctor Mikhail Barash, Doctor Reino Niskanen, Tuomo Lehtilä, Juho Salmensuu, Doctor Michal Szabados, Timo Vesalainen, and Doctor Jetro Vesti for the time spent during these years, for the numerous lunch discussions, and for the help in fighting occasional computer issues.

The atmosphere in the department was always warm and welcoming, and the staff of the Department of Mathematics and Statistics is to be thanked. Thank you Doctors Kaisa Joki, Ville Junnila, Eija Jurvanen, Jyrki Lahtonen, Tommi Meskanen, and Eila Seppänen and Outi Montonen and Eila Seppänen for the numerous interesting and fun discussions during coffee breaks. I'd especially like

to thank Ville Junnila and Tommi Meskanen for their humour without which life would be quite sunny. I would like to thank Doctor Petteri Harjulehto for several discussions on the pedagogical aspects of teaching mathematics. Your views helped me a lot in the teaching duties. Doctor Arto Lepistö, thank you for helping out with the all things related to computers and thank you for sharing your stories in programming. I would like to thank Sonja Vanto and Tuire Huuskonen for help in bureaucratic matters during the beginning of my studies. Thank you to Professor Marko Mäkelä and Doctors Petteri Harjulehto, Jyrki Lahtonen, Yuri Nikulin, Reino Niskanen, Mikko Pelto, and Tuomas Nurmi for the tough but fair games of floorball. Thank you Doctor Thibault Godin, Joonatan Jalonen, Ville Laitinen, and Etienne Moutot for lots of enjoyable moments during bouldering and climbing.

Lopuksi haluan kiittää ystäviäni ja läheisiäni. Haluan kiittää kaikkia henkilöitä, jotka ovat elämääni kuuluneet. Olette opettaneet minulle elämästä paljon. Haluan kiittää Henryä lukemattomista kokemuksista, jotka olen saanut jakaa kanssasi. Kiitos siitä, että olen saanut kokea kanssasi aitoa iloa, kummastusta ja ihmettelyä, naurua ja rehellisyyttä. Kiitos myös, että autoit ymmärtämään kaiken edellä olleen tyhjyyden. Ystäväni, joiden kanssa olen kasvanut lapsesta aikuiseksi, kiitos, että olette muistaneet minua, vaikka tiemme ovat välillä eriytyneet pidemmäksikin aikaa. Erityisesti haluan kiittää Juri and Leena Viitaniemiä lämmöstä ja ystävyydestä, sekä kaikesta avusta matkan varrella. Kiitos myös Atelle, Juholle, Santerille, Konstalle ja kaikille Vanhoille Kamuille, että olette olemassa. Haluan myös kiittää Samia ja Akua tuesta ja ystävyydestä elämän lähitaistelutilanteissa. Haluan kiittää myös isovanhempiani Leeviä, jonka rakkaalle muistolle omistan tämän väitöskirjan, ja Terttua tuesta ja rakkaudesta. Kiitos Heli kaikista keskusteluista elämän realiteeteista. Kiitän siskoani Annia mukavista ja ei niin mukavista hetkistä lapsuudessamme ja nuoruudessamme. Kiitos myös Eetille ja Fiiralle, joiden ansiosta voin kokea uutta ihmetystä maailmaa kohtaan.

I would like to thank my parents Alan and Tuula for their love and support. Thank you for making this a possibility for me.

Turku, June 6, 2019 Markus A. Whiteland

Contents

List of Original Publications and Manuscripts

[1] J. Cassaigne, J. Karhumäki, S. Puzynina, and M.A. Whiteland. "k-Abelian Equivalence and Rationality". *Fundam. Inform.* 154.1-4 (2017), pp. 65–94. DOI: 10.3233/FI-2017-1553.

[2] J. Karhumäki, S. Puzynina, M. Rao, and M.A. Whiteland. "On cardinalities of k-abelian equivalence classes". *Theor. Comput. Sci.* 658 (2017), pp. 190–204. DOI: 10.1016/j.tcs.2016.06.010.

[3] J. Karhumäki and M.A. Whiteland. "Regularity of k-Abelian Equivalence Classes of Fixed Cardinality". In: *Adventures Between Lower Bounds and Higher Altitudes - Essays Dedicated to Juraj Hromkovič on the Occasion of His 60th Birthday*. 2018, pp. 49–62. DOI: 10.1007/978-3-319-98355-4_4.

[4] J. Peltomäki and M.A. Whiteland. *On k-Abelian Equivalence and Generalized Lagrange Spectra*. (Submitted). 2018. URL: https://arxiv.org/abs/1809.09047v1.

[5] M.A. Whiteland. "Asymptotic Abelian Complexities of Certain Morphic Binary Words". *Journal of Automata, Languages and Combinatorics* 24.1 (2019), pp. 89–114. DOI: 10.25596/jalc-2019-089.

[6] M.A. Whiteland. *On Equations Over Monoids Defined by Generalizations of Abelian Equivalence*. (in preparation). Parts presented at the Fifth Russian-Finnish Symposium on Discrete Mathematics (RuFiDiM), Veliky Novgorod, Russia, May 19–22. 2019.

The thesis is based on the publications and manuscripts listed above. The notations and results of these works have have been modified and unified for the presentation of this thesis.

Chapter 1

Introduction

1.1 Background

Sequences of symbols, i.e. words, appear all around in the research of mathematics. Especially in number theory, modern algebra, and probability theory, the combinatorial analysis of these kinds of sequences is of major significance. In the area of theoretical computer science and formal language theory, the usage of words is paramount, and is even a central notion of the research areas. Other areas of science having particular use of word combinatorics are for example physics in crystallography and biology in gene assembly. The increasing use of words and related notions led to the birth of a research area of words called combinatorics on words, a branch of combinatorics. The topic of this thesis falls well into the areas of combinatorics on words and theoretical computer science.

The study of finite and infinite words as an object of independent interest can be traced back to the work [103, 104] (for a modern interpretation of these results see [9]) of A. Thue in the early 1900's. In these papers, Thue exhibited the existence of an infinite word over three letters which avoids squares, that is, two consecutive identical contiguous subwords, or factors. No such infinite words over two letters exists (each binary word of length four contains a square). On the other hand, he constructed an infinite binary word avoiding cubes, that is, no three consecutive identical factors occur. This infinite word constructed by Thue is now known as the *Prouhet–Thue–Morse word*. The first implicit use of this word can be found in the work of Prouhet [83] as early as 1851 [5], and was also used by M. Morse in [71]. The word has been rediscovered several times in the literature. Thue considered this problem on infinite words of independent interest and without any particular application in mind. See [5] for a comprehensive treatment of the history of this word.

The first broad and unifying treatment of combinatorics on words was performed in the 1983 book [63] titled *Combinatorics on Words* by the pseudonym M. Lothaire. At that point the topic had ripened to a level of having mature theory and clearly articulated major topics of research. The research progressed quite quickly, and the area attracted a lot of interest. The subsequent book [64] titled *Algebraic Combinatorics on Words* (2002) deepened the theories presented in the previous book as well as presented new topics that had not been ripe enough for

the first exposition. Later on, in 2005, a third title [65] from the pseudonym was published: *Applied Combinatorics on Words*. This book presents the wide use of combinatorics on words in other areas such as "... core algorithms for text processing, natural language processing, speech processing, bioinformatics, and several areas of applied mathematics such as combinatorial enumeration and fractal analysis" [65]. These three books on combinatorics on words testify the solidification of this topic as a research area in its own right.

As an algebraic structure, the set of finite words is extremely rigid. Two finite words are distinct if either they have different lengths, or there exists an index at which the two words have distinct letters. When applying combinatorics on words to problems arising in other areas of mathematics, it is natural to identify distinct words by using equivalence relations. For example, one such equivalence relation is the so-called *abelian equivalence*: two words are called abelian equivalent if each symbol of the alphabet occurs equally many times in the two words. Thus two words are abelian equivalent if we may permute either of the words to obtain the other. The topic of this thesis concerns a generalization of the abelian equivalence, the so-called *k-abelian equivalence* introduced by J. Karhumäki in [53]. In fact, the notion of k-abelian equivalence contains a hierarchy of equivalence relations. For each $k \geqslant 1$, two words are k-abelian equivalent if each word of length at most k occurs equally many times in u and v. The 1-abelian equivalence and the abelian equivalence coincide. For any $k \geqslant 1$, two words being k-abelian equivalent implies that they are k'-abelian equivalent for each $k' < k$, but not necessarily conversely. Thus the k-abelian equivalence may be seen as a refinement of the abelian equivalence as an equivalence relation on words. Now two words are equal if and only if they are k-abelian equivalent for each $k \geqslant 1$. By interpreting the equality relation as the ∞-abelian equivalence, we obtain an infinite hierarchy of refinements between the abelian equivalence and the equality relation.

Several aspects of k-abelian equivalence in infinite words have been studied in the recent literature. In [53] the notion was introduced as a generalization of the so-called *Parikh mapping* and considered in decidability problems, such as the equivalence of HD0L-systems (in a k-abelian sense) and problems related to Post Correspondence Problem. It is not inaccurate to say that such considerations have been inspired by the central questions of combinatorics on infinite words applied in the k-abelian setting. The main aspects considered have been the avoidance of k-abelian repetitions and the k-abelian complexity functions, though several other notions are considered as well [17, 20, 50, 56, 57, 58, 86, 88]. Indeed, these problems have been strongly motivated by the results concerning the central topics of avoiding powers or abelian powers, and the notions of factor complexity and abelian complexities, respectively. Especially interesting questions regarding the k-abelian setting are when the corresponding behaviours of the equality relation and abelian equivalence differ. For example, cubes (three consecutive identical blocks) are avoidable in binary words, but abelian cubes are not. (Abelian cubes are avoidable over ternary words [29] and so are thus k-abelian cubes for any $k \geqslant 1$.) What about k-abelian cubes in binary words? Similarly, squares can be avoided in ternary words, but not abelian squares (abelian squares are avoidable with four letters [60]). What about k-abelian squares? Questions of these sorts were among the first considered in the literature regarding k-abelian aspects in infinite words. The search for the minimal k for which k-abelian cubes can be

avoided over the binary alphabet was initiated in [48], where it was shown that $k \leqslant 8$. The search of the optimal k was continued in [69] (showing $k \leqslant 5$) and [68] (showing $k \leqslant 3$), and was finally settled to $k = 2$ in [86], by constructing a binary word avoiding 2-abelian cubes. Similarly, in [49] it was shown that 2-abelian squares over the ternary alphabet cannot be avoided,[1] but, on the other hand, in [46] it was shown that 64-abelian cubes can be avoided. In [86] a ternary word avoiding 3-abelian squares was constructed, settling this question.

The study of somehow quantifying the complexity of an infinite word is a well-motivated and actively studied research area. Especially fruitful is the notion of relating the complexity of an infinite word to the complexity of the word's finite factors. The *factor complexity function* $\mathcal{C}_{\mathbf{w}} : \mathbb{N} \to \mathbb{N}$ of an infinite word $\mathbf{w}$ counts, for each $n \in \mathbb{N}$, the number of distinct blocks of length n occurring in $\mathbf{w}$. The notion is a fundamental one in combinatorics of infinite words. Indeed, the theorem of M. Morse and G.A. Hedlund [72], which characterizes *ultimately periodic* words as exactly the words admitting $\mathcal{C}(n_0) \leqslant n_0$ for some $n_0 \in \mathbb{N}$, already showcases the usefulness of this notion. For surveys on factor complexity we refer the reader to [21, 23]. Sturmian words comprise a large class of extensively studied words with strong connections to number theory, particularly to continued fractions (see, e.g., [10], [64, Chapter 2], [84, Chapter 6] and references therein). An infinite word $\mathbf{w}$ is Sturmian if and only if $\mathcal{C}_{\mathbf{w}}(n) = n + 1$ for each $n \geqslant 0$. Sturmian words may be characterized in several different ways (see [64, Chapter 2]).

The notion of factor complexity has prompted other quantifications of complexity. One such is the *abelian complexity* of infinite words. For other related complexity measures, see for instance [57, 79, 93] and the survey [92]. The *abelian complexity function* $\mathcal{C}^{\mathrm{ab}}_{\mathbf{w}} : \mathbb{N} \to \mathbb{N}$ of an infinite word $\mathbf{w}$ counts, for each n, the number of distinct abelian equivalence classes of length n occurring in the word $\mathbf{w}$. (The subscript is omitted when $\mathbf{w}$ is clear from context.) E.M. Coven and G.A. Hedlund [24] characterize *purely periodic words* to be exactly the words w for which $\mathcal{C}^{\mathrm{ab}}_{\mathbf{w}}(n_0) = 1$ for some $n_0 \geqslant 1$. This creates the starting point of the study of the abelian complexity function. Even though the notion has been around for some time, the study was formally initiated only recently by G. Richomme, K. Saari, and L.Q. Zamboni in [91].

Again, natural questions arise for the k-abelian equivalence in relation to complexity functions. For each $k \geqslant 1$, the *k-abelian complexity function* $\mathcal{C}^{(k)}_{\mathbf{w}}(n) : \mathbb{N} \to \mathbb{N}$ of the infinite word $\mathbf{w}$ is defined by setting $\mathcal{C}^{(k)}_{\mathbf{w}}(n)$ to equal the number of distinct k-abelian equivalence classes represented by factors of $\mathbf{w}$ of length n. This notion was introduced in [57], where the usefulness of the notion was immediately showcased: if, for some $k \geqslant 1$ and $n_0 \geqslant 1$, the value $\mathcal{C}^{(k)}_{\mathbf{w}}(n)$ is too small (depending on k), $\mathbf{w}$ is ultimately periodic (we discuss the precise result in the beginning of Chapter 9). Ever since the introduction of this notion there has been quite a bit of interest towards the k-abelian complexity function [20, 22, 78].

The critical exponent of an infinite word $\mathbf{w}$ is the supremum of exponents of fractional powers occurring in $\mathbf{w}$. The Prouhet–Thue–Morse word mentioned above has critical exponent 2, as shown by Thue in 1906 [103]. In other words, it does not avoid squares, but it avoids factors of the form $axaxa$, where a is a letter and x is a word. Thus the notion of a *critical exponent* is a central one

[1] The longest ternary words avoiding 2-abelian squares have length 537 [47, 48].

in combinatorics on words. The powers occurring in Sturmian words are well-understood, and a formula for the critical exponent of a Sturmian word has been determined by Damanik and Lenz [27], and Justin and Pirillo [52]. For example, the critical exponent of the Fibonacci word, the fixed point of the morphism $a \mapsto ab$, $b \mapsto a$ (see Subsection 2.2.3 for a definition), is $(5+\sqrt{5})/2$ [70]. The critical exponent of the Fibonacci word is minimal among all Sturmian words. In the recent years, there has been a substantial amount of research in generalizations of the concept of a power. A popular generalization is that of an abelian power; other generalizations are k-abelian powers (studied here) and those based on k-binomial equivalence [93]. Also abelian repetitions in Sturmian words are well-investigated, see, e.g. [36] and references therein.

As mentioned previously, most of the literature on k-abelian equivalence is on aspects in infinite words. There are of course results on finite words and the equivalence classes themselves. Several characterizations of k-abelian equivalence are known (see Chapter 3 for a review). Also, for example, the number $\mathcal{P}_m^{(k)}(n)$ of k-abelian equivalence classes of words of length n over an m-letter alphabet was considered in [57], and an asymptotic formula has been obtained (see Chapter 2 for notation of asymptotic analysis):

Theorem 1.1. *Let $k \geqslant 1, m \geqslant 1$. We have $\mathcal{P}_m^{(k)}(n) = \Theta(n^{m^{k-1}(m-1)})$, where the constants implied by Θ depend on k and m.*

This theorem gives a polynomial order of growth for the number of k-abelian equivalence classes of words of length n over an m-letter alphabet. We emphasize this result since the question of sharpening this bound was the main motivation behind the research on the structure of k-abelian equivalence classes.

Algorithmic aspects of the k-abelian equivalence relation have also been considered. A linear time algorithm for finding the largest k for which two given words are k-abelian equivalent was obtained in [33], together with text processing algorithms in the k-abelian sense.

In this thesis, we mainly take the viewpoint of studying the k-abelian equivalence relation, the corresponding equivalence classes and the properties of representatives of the classes. We elaborate on the topics covered in this thesis in the following section.

1.2 Structure of the Thesis

We structure the thesis as follows. In Chapter 2 we recall basic notions and terminology from the literature on combinatorics on words and related areas. We also prove some basic results needed in the thesis. The rest of the thesis can be seen to be comprised of two parts, one on aspects of the k-abelian equivalence on finite words and the other one on infinite words. In Chapters 3–7 we consider different aspects of the k-abelian equivalence in finite words: characterizations of k-abelian equivalence, language theoretic aspects, numbers of k-abelian equivalence classes, and k-abelian equations. Chapters 8 and 9 then focus on the k-abelian aspects of infinite words. Let us briefly summarize the contents of each of these chapters.

In Chapter 3 we review several different equivalent characterizations of the k-abelian equivalence relation from the literature. We also introduce novel characterizations, a rewriting rule called the k-switching (introduced in [55]) and a matrix

representation ([110]) (see Sections 3.2 and 3.4, respectively). These characterizations open up different aspects of the k-abelian equivalence relation. When considering structural properties of the k-abelian equivalence classes, we employ two of these characterizations. One of them involves the so-called de Bruijn graphs and the other is the k-switching. The matrix representation of the k-abelian equivalence is relevant when studying algebraic properties of the k-abelian equivalence classes.

In Chapter 4 we focus on certain representatives of the k-abelian equivalence classes, namely, the lexicographically least elements of each class. We see how these lexicographically least representatives may be seen in a couple of different viewpoints, which we then exploit in consequent chapters. The structure of these representatives as words are studied in a word combinatorial sense together with an interpretation of them in certain graphs. We also consider the structure of the so-called *k-abelian singletons*, that is, words who are k-abelian equivalent to only themselves. This chapter lays down the basic notions which we use in the rest of the thesis. Most of this chapter is based on parts of the article [19].

In Chapter 5 we turn to automata theoretic aspects of the k-abelian equivalence classes. We first give two proofs of the regularity of the language of lexicographically least representatives. Similar arguments show that the language of k-abelian singletons is also regular. Both the proofs are constructive, but have different consequences. We construct automata for these languages for some small values of k and of the cardinality of the alphabet. We also show that the language of all words representing classes of a fixed cardinality is regular, as well as other language theoretic aspects related to the k-abelian equivalence classes. This chapter is based on the works [19] and [59].

In Chapter 6 we turn to quantitative aspects of the k-abelian equivalence classes. The main result is the sharp asymptotic expression for the growth of the number of k-abelian equivalence classes of length n; We show that the limit

$$\lim_{n\to\infty} \frac{\mathcal{P}_m^{(k)}(n)}{n^{m^{k-1}(m-1)}}$$

exists and equals to a rational number. This can be considered as a sharpening of Theorem 1.1. We also give an upper bound on the asymptotic growth of the number $\mathcal{S}_m^{(k)}(n)$ of k-abelian singletons of length n over an m-letter alphabet:

$$\mathcal{S}_m^{(k)}(n) = \mathcal{O}(n^{N_m(k-1)-1}),$$

where

$$N_m(\ell) = \tfrac{1}{\ell} \sum_{d|\ell} \varphi(d) m^{\ell/d}$$

is the number of conjugacy classes (or necklaces) of words of length ℓ over an m-letter alphabet (see Chapter 2 for definitions). The main tool used for these results is the general treatment of the asymptotic complexities of regular languages having polynomial growth. More precisely, we give a sufficient condition for a regular language having polynomial growth to have growth asymptotic to a polynomial. Explicit formulae for computing the number of k-abelian equivalence classes and k-abelian singletons of length n over an m letter alphabet for small values of k and

m are also given. Finally, we discuss the relation of the problem of lower bounding the asymptotic growth of k-abelian singletons to other problems in the literature. This chapter is based on the articles [19, 55].

In Chapter 7 we turn to equations in the k-abelian setting as well as equations in terms of a related equivalence relation called the *k-binomial equivalence*. This related equivalence relation has been of interest in the recent literature, and we consider the k-binomial setting here, since it is quite interesting to see how these two equivalence relations differ. We consider two concrete examples of equations, namely commutation and conjugacy. We also consider the general properties of systems of equations. As the main result we show that any system of equations in the k-abelian setting has an equivalent finite subsystem. The same property is shown to hold in the k-binomial setting. We conclude the chapter by considering the number of equations in the so-called independent systems of equations: We give a uniform upper bound on the number of equations in such a system (in both the k-abelian and the k-binomial settings). This chapter is based on the unpublished manuscript [110].

All the previous chapters consider properties of the k-abelian equivalence classes. The last two chapters of the thesis concern k-abelian aspects in infinite words. In Chapter 8 we consider abelian complexities of morphic binary words (for definitions see Subsection 2.2.3). When considering the existence of infinite words having certain properties, the question is quite often answered, when positive, by constructing such a morphic word. One example is the Prouhet–Thue–Morse word introduced previously; it answers positively the question whether there exist infinite binary words avoiding cubes. The morphic words are thus an interesting class of words by themselves. The celebrated theorem of Pansiot [76] classifies the factor complexities of pure morphic words; the factor complexity is asymptotically one of five possibilities: $\Theta(1)$, $\Theta(n)$, $\Theta(n \log\log n)$, $\Theta(n \log n)$, or $\Theta(n^2)$. The study of abelian complexities of pure morphic words is inspired by this result. The research on abelian complexities was initiated in [12]. We study certain pure morphic binary words and give limit superior and inferior asymptotics for the abelian complexity. The words studied here are the only binary pure morphic words whose abelian complexities had not yet been considered. Thus we complete the classification of the asymptotic limit superior abelian complexities of pure morphic binary words. We then consider morphic binary words, and give several such words having asymptotically the same abelian complexity, but having distinct factor complexities. The considerations in this chapter appear in the article [109].

In Chapter 9 we consider several aspects of the k-abelian equivalence in the Sturmian words. We give new proofs of the k-abelian complexity (see Chapter 9) of Sturmian words and the fact that there are arbitrarily large k-abelian repetitions in Sturmian words—results already proved in [57]. We thus give alternative proofs through our starting point of Sturmian words as encodings of irrational rotations. This viewpoint gives us tools to analyze the situation more precisely, and this can be seen in the statements of our results. Indeed, we sharpen some of the results of [57]. We also consider other notions from the literature related to Sturmian words concerning abelian equivalence, and generalize such notions to k-abelian equivalence. This chapter is based on the manuscript [81].

Chapter 2

Preliminaries

2.1 Basic Notation and Terminology

The symbols $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ denote the set of natural numbers (including zero), the integers, the rational numbers, the real numbers, and the complex numbers respectively. The cardinality of a set S is denoted by $\#S$ and sometimes by $|S|$. We only consider cardinalities of finite sets in this thesis.

We recall the Bachmann–Landau notation for asymptotic comparison of functions. Let f and g be functions $\mathbb{N} \to \mathbb{R}$, with f having non-negative values and g having positive values. We write

- $f(n) = \mathcal{O}(g(n))$ if there exist $n_0 \in \mathbb{N}$ and $C > 0$ such that $f(n) \leqslant Cg(n)$ for all $n \geqslant n_0$;
- $f(n) = \Omega(g(n))$ if there exist $n_0 \in \mathbb{N}$ and $C > 0$ such that $f(n) \geqslant Cg(n)$ for all $n \geqslant n_0$;
- $f(n) = \Theta(g(n))$ if both $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$;
- $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$;
- $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.

We make use of linear algebraic notation in several occasions. When explicitly stated, a vector $\vec{\mathbf{u}}$ is a finite ordered tuple of elements from a ring R, and the components are indexed with natural numbers starting from 1. The ith element of the vector $\vec{\mathbf{x}} = (a_1, a_2, \ldots, a_n) \in R^n$ is expressed as $\vec{\mathbf{x}}[i]$. The dot product $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$, or $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}$, of vectors $\vec{\mathbf{x}}, \vec{\mathbf{y}} \in R^n$, is thus the element $\sum_{i=1}^{n} \vec{\mathbf{x}}[i] \cdot \vec{\mathbf{y}}[i]$ of R. We often do not require an explicit expression of the indexing, in which cases we express the elements of a vector in some other way. For example, let A be a finite set and let R be a ring. Let R^A denote the set of functions from A to R. Then R^A is seen as a finitely generated left module by identifying the elements of R^A with vectors of dimension $\#A$ over the ring R. We may index the elements of such a vector $\vec{\mathbf{x}}$ by elements of A, whence the notation $\vec{\mathbf{x}}[a]$ is understood as the element at "position" a, i.e., the element at the position corresponding to the element a.

Here some ordering of the coordinates is implicitly assumed. Thus, for $\vec{\mathbf{x}}, \vec{\mathbf{y}} \in R^A$, where A is a finite set and R is a ring, we may compute the dot product as

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = \sum_{a \in A} \vec{\mathbf{x}}[a] \cdot \vec{\mathbf{y}}[a].$$

Note that this dot product does not depend on the order of which the components are defined, since the sum operation is commutative (in R). Thus no explicit ordering is required for such an operation.

For a matrix M defining a linear mapping $R^A \to R^B$, where A and B are finite non-empty sets, we index the rows of M by the elements of B and the columns of M by the elements of A. Thus, for $a \in A$ and $b \in B$, the element $M[b, a]$ means the "ath" element of the "bth" row of M. For $\vec{\mathbf{x}} \in R^A$ we thus have $(M\vec{\mathbf{x}})[b] = \sum_{a \in A} M[b, a] \cdot \vec{\mathbf{x}}[a]$ for each $b \in B$.

Next we set some terminology on graphs. Throughout this thesis, when talking about graphs, we usually mean directed multigraphs with loops allowed. Sometimes we consider labeled multigraphs, so that each edge is distinguishable, and at other times it is not important. If the choice of interpretation is not clear from context, we shall explicitly mention if the edges need to be distinguishable. For a graph $G = (V, E)$, where V is some set and E consists of edges (labeled or not) between elements of V, we let $V(G)$ denote the set V of vertices and $E(G)$ the set E of edges. For an edge $e \in E(G)$ from vertex x to vertex y, we call the vertex x the *tail* of e, denoted by $\mathrm{tail}(e)$, and y the *head* of e, denoted by $\mathrm{head}(e)$. We denote by $d^+_G(u)$ (resp., $d^-_G(u)$) the number of outgoing (resp., incoming) edges of u. For $u, v \in V$, the number of edges from u to v in G is denoted by $m_G(u, v)$. In all of these notations we omit the subscript G when the graph is clear from context.

A sequence $W = (e_i)_{i=1}^t$ of edges satisfying $\mathrm{head}(e_i) = \mathrm{tail}(e_{i+1})$ for each $i \in [1, t-1]$, is called a *walk* (in G). We set $\mathrm{tail}(W) = \mathrm{tail}(e_1)$ and $\mathrm{head}(W) = \mathrm{head}(e_t)$. Further, we let $|W|$ denote the length t of W. For a given walk $W = (e_i)_{i=1}^t$ we let $V(W)$ denote the set

$$\{\mathrm{tail}(e_i), \mathrm{head}(e_i) \colon i \in [1, t]\}$$

of vertices and $E(W)$ the set

$$\{e_i \colon i \in [1, t]\}$$

of edges along W. We call W a *path* if $\mathrm{tail}(e_i) \neq \mathrm{tail}(e_j)$ when $i \neq j$, and $\mathrm{head}(e_t) \neq \mathrm{tail}(e_i)$ for all $i \in [1, t]$. In other words, a path P does not visit any vertex twice. If the walk $(e_i)_{i=1}^{t-1}$ is a path and e_0 is an edge such that $\mathrm{tail}(e_0) = \mathrm{head}(e_{t-1})$ and $\mathrm{head}(e_0) = \mathrm{tail}(e_1)$, then we call the walk $(e_i)_{i=0}^{t-1}$ a *cycle*. (We index the edges of a cycle starting from 0 for notational reasons.) In the literature, our notion of a cycle is often called a *simple cycle*, while a cycle is defined as a walk W for which $\mathrm{tail}(W) = \mathrm{head}(W)$. A graph G is called *Eulerian* if there exists an Eulerian *cycle*, that is, a walk W in G which contains each edge exactly once, and $\mathrm{tail}(W) = \mathrm{head}(W)$.

Let $W = (e_i)_{i=1}^s$ and $W' = (e'_i)_{i=1}^t$ be non-empty walks such that $\mathrm{tail}(W') = \mathrm{head}(W)$. We define the *concatenation* of W and W', denoted by $W \cdot W'$, as the walk $(d_i)_{i=1}^{s+t}$, where $d_i = e_i$ if $i \leqslant s$, and $d_i = e'_{i-s}$ if $i > s$. For an empty walk W,

we define $W \cdot W' = W' \cdot W = W'$. Note here that a cycle C can be concatenated with itself arbitrarily many times. We say that a walk W is a *repetition of a cycle* if we may write $W = C^r$ for some $r \geqslant 1$. A vertex x is said to be an *internal* vertex of a walk W, if we may write $W = W_1 W_2$ for some non-empty walks W_1, W_2 with $\mathrm{head}(W_1) = x$. A vertex x is an *extremal* vertex of W if $\mathrm{tail}(W)$ or $\mathrm{head}(W) = x$.

We say that a cycle $C = (d_j)_{j=0}^{s-1}$ *occurs along the walk* W if we may write

$$W = W_1 \cdot (d_{r+j \bmod s})_{j=0}^{s-1} \cdot W_2$$

for some $0 \leqslant r < s$ and some (possibly empty) walks W_1, W_2. We say that W *enters* C *via the vertex* $\mathrm{tail}(d_r)$ if W_1 is either empty or $d_{r-1 \bmod s}$ is not the last edge of W_1. In this case we say that W enters C at position $|W_1|+1$. We say that W leaves C via the vertex $\mathrm{head}(d_{r+s-1 \bmod s})$ if W_2 is empty or d_r is not the first edge of W_2. In this case we say that W leaves C at position $|W_1| + s$.

Let G be a graph and let the set $V(G)$ of vertices be ordered as $V(G) = \{v_1, \ldots, v_{\#V}\}$. The *adjacency matrix of* G is the matrix $(m(v_i, v_j))_{i,j}$.

2.2 Notions and Terminology of Combinatorics on Words

We recall some basic notions of combinatorics of words used in this thesis. We also mention some basic but relevant results used along this thesis.

2.2.1 Finite Words

An *alphabet* Σ is a non-empty set of symbols which are called *letters*. All alphabets in this thesis should be considered finite unless explicitly otherwise stated. A finite or infinite sequence of letters over the alphabet Σ is called a *word*. The *empty word* is denoted by ε. The set of finite words over Σ is denoted by Σ^*, the set of non-empty finite words by Σ^+, and the set of infinite words by $\Sigma^{\mathbb{N}}$. For a word $u \in \Sigma^*$ and a natural number $n \geqslant 0$, we let u^n denote the concatenation of n copies of u; $u^n = uu \cdots u$ (n times). A word u is called *primitive*, if $u = v^n$ for some word v implies that $n = 1$. A set $L \subseteq \Sigma^*$ of words is called a *language*. For a (usually finite) language $S \subseteq \Sigma^*$, S^* denotes the language of finite concatenations of elements of S interpreted as words over Σ. One should always see such languages as languages over the alphabet Σ, and not interpret S as an alphabet. The sets S^+ and $S^{\mathbb{N}}$ are defined analogously. For $u \in \Sigma^+$ we let u^* and u^+ denote the sets $\{u\}^*$ and $\{u\}^+$, respectively. The infinite word u^ω denotes the singleton element of $\{u\}^{\mathbb{N}}$. When talking about *the binary alphabet*, we mean the alphabet $\{a, b\}$. Throughout the thesis we let $\mathbb{B}$ denote the binary alphabet. For a finite word $w \in \Sigma^*$, the *length* $|w|$ of w is the number of letters occurring in w. The set of words of length n over Σ is denoted by Σ^n. For a language $L \subseteq \Sigma^*$, we define the mapping $\mathcal{C}_L : \mathbb{N} \to \mathbb{N}$, called the *complexity* or the *growth function* of L, defined by $\mathcal{C}_L(n) = \#(L \cap \Sigma^n)$ for each $n \geqslant 0$.

A word $u \in \Sigma^*$ is a *factor* of $w \in \Sigma^*$ if there exist $p, q \in \Sigma^*$ such that $w = puq$. For a non-empty word u we let $|w|_u$ denote the number of occurrences of u in w as a factor. The set of factors of w is denoted by $F(w)$. We let $F_n(w)$ denote the set $F(w) \cap \Sigma^n$. For w as above, if $p = \varepsilon$ (resp., $q = \varepsilon$) then u is called a *prefix* (resp.,

suffix) of w. Further, if $q \neq \varepsilon$ (resp., $p \neq \varepsilon$) then u is called a *proper prefix* (resp., *proper suffix*). For $w = pq$ we define $p^{-1}w = q$. Similarly we define $wq^{-1} = p$. The set of prefixes (resp., suffixes) of w is denoted by $\mathrm{pref}(w)$ (resp., $\mathrm{suff}(w)$) and the length k prefix (resp., suffix) of w, with $k \leqslant |w|$, is denoted by $\mathrm{pref}_k(w)$ (resp., $\mathrm{suff}_k(w)$). For a word $w = a_0a_1 \cdots a_{n-1} \in \Sigma^*$ and indices $0 \leqslant i \leqslant j < n$, the factor $a_i \cdots a_j$, is denoted by $w[i, j]$. For $i > j$ we set $w[i, j] = \varepsilon$. Similarly, for $i < j$ we set $w[i, j) = w[i, j - 1]$ and we set $w[i, j) = \varepsilon$ when $i \geqslant j$. For a finite word $w \in \Sigma^*$, we let $w[i..]$ (resp., $w[..i]$) denote the suffix $w[i, |w|)$ (resp., the prefix $w[0, i]$) for brevity. We say that a word $x \in \Sigma^*$ *has position i in w* if the word $w[i..]$ has x as a prefix.

Let $w = a_0a_1 \cdots a_{n-1} \in \Sigma^*$. A word $u \in \Sigma^+$ is called a *subword* (also known as a *scattered subword* or *scattered factor* in the literature) of w if there exist an increasing sequence of indices $0 \leqslant i_1 < i_2 < \ldots < i_m < n$ such that $u = a_{i_1}a_{i_2} \cdots a_{i_m}$. For $u \in \Sigma^+$ we let $\binom{w}{u}$ denote the number of occurrences of u in w as a subword, that is, the number of distinct increasing sequences of indices $(i_j)_{j=1,\ldots,m}$ such that $a_{i_1}a_{i_2} \cdots a_{i_m} = u$. We stress the distinction between factors and subwords.

For words $x, y \in \Sigma^*$ we say that x is conjugate to y if there exists $z \in \Sigma^*$ such that $xz = zy$. This is equivalent to the existence of $p, q \in \Sigma^*$ such that $x = pq$ and $y = qp$. Note that conjugacy is an equivalence relation on Σ^*. We call a conjugacy class a *necklace*. In our future considerations we recall the number of necklaces of a certain length. Let $N_m(\ell)$ be the number of necklaces of length ℓ over an m-letter alphabet. We have

$$N_m(\ell) = \frac{1}{\ell} \sum_{d|\ell} \varphi(d) m^{\ell/d}, \tag{2.1}$$

where φ is *Euler's totient function*, that is, $\varphi(n)$ counts the number of natural numbers less than n which are coprime with n (see, e.g., [95]). The sequence $(N_2(\ell))_{\ell=0}^{\infty}$ is sequence A000031 in N. Sloane's On-line Encyclopedia of Integer Sequences (http:/oeis.org). The first few values of the sequence are

$$1, 2, 3, 4, 6, 8, 14, 20, 36, 60, 108, 188, 352, 632, 1182, 2192, 4116, 7712, \ldots$$

See also the sequences A001867–A001869.

We now turn to some more recent notions in the literature of combinatorics of words used in this thesis.

Let us first of all articulate the k-abelian equivalence in the form of a definition.

Definition 2.1. Let $k \geqslant 1$ and let $u, v \in \Sigma^*$. Then u and v are called *k-abelian equivalent* if, for each $x \in \Sigma^*$ of length at most k, we have $|u|_x = |v|_x$.

Example 2.2. Let $u = aaba$ and $v = abaa$. One sees that $|u|_a = |v|_a = 3$, $|u|_b = |v|_b = 1$, and $|u|_{aa} = |u|_{ab} = |u|_{ba} = |v|_{aa} = |v|_{ab} = |v|_{ba} = 1$. Consequently $u \sim_2 v$. The fact that $|u|_{aab} = 1 \neq 0 = |v|_{aab}$ implies $u \not\sim_3 v$.

Let $x = aaba$ and $y = baab$. Even though $|x|_t = |y|_t$ for each $t \in \Sigma^2$, we have $|x|_a \neq |y|_a$ which implies $x \not\sim_2 y$.

Since the number of occurrences of letters in a word $u \in \Sigma^*$ sum up to the length of u, we immediately have that $u \sim_k v$, for $k \geqslant 1$, implies that $|u| = |v|$. Further,

if $u \sim_k v$, then $u \sim_t v$ for all $t \leqslant k$. The relation $\sim_k$ is clearly an equivalence relation. We let $[u]_k$ denote the k-abelian equivalence class defined by u. We may consider also the equivalence relation with $k = \infty$, which we understand as the equality relation. We obtain a hierarchy of equivalence relations:

$$u \sim_1 v \Leftarrow u \sim_2 v \Leftarrow \cdots u \sim_k v \Leftarrow u \sim_{k+1} v \Leftarrow \cdots \Leftarrow u = v.$$

For each $k \geqslant 1$ and $u \in \Sigma^*$ we define its *generalized Parikh vector* $\Psi_k(u)$ *of order* k as $\Psi_k(u) = (|u|_x)_{x \in \Sigma^k}$. The vector $\Psi_1(u) = (|u|_a)_{a \in \Sigma}$ is known in the literature as the *Parikh vector*. Thus two words $u, v \in \Sigma^*$ are k-abelian equivalent if and only if $\Psi_\ell(u) = \Psi_\ell(v)$ for each $\ell = 1, \ldots, k$.

We shall briefly consider also another equivalence relation on words recently introduced by M. Rigo and P. Salimov in [93].

Definition 2.3. Two words $u, v \in \Sigma^*$ are *k-binomial equivalent*, $u \equiv_k v$ in symbols, if $\binom{u}{e} = \binom{v}{e}$ for all words e of length at most k.

Observe also that the k-abelian equivalence and the k-binomial equivalence are incomparable as equivalence relations. Indeed, there exist k-abelian equivalent words that are not k-binomial equivalent and vice versa [93]. The notion has also gathered quite a bit of interest in recent years [39, 62, 87]

Basic properties of binomial coefficients $\binom{u}{v}$ are presented in [63, Chapter 6]. We repeat the main properties here. We define, for $a, b \in \Sigma$, $\delta_{a,b} = 1$ if $a = b$, otherwise $\delta_{a,b} = 0$. For all $p, q \in \mathbb{N}$, $u, v \in \Sigma^*$, and $a, b \in \Sigma$ we have

$$\binom{a^p}{a^q} = \binom{p}{q}; \quad \binom{u}{\varepsilon} = 1; \quad |u| < |v| \text{ implies } \binom{u}{v} = 0; \quad \binom{ua}{vb} = \binom{u}{vb} + \delta_{a,b}\binom{u}{v}.$$

The last three relations completely determine the binomial coefficient $\binom{u}{v}$ for all $u, v \in \Sigma^*$. We repeat a couple of basic results of k-binomial equivalence from [93].

Proposition 2.4. *Let $u, v, e \in \Sigma^*$ and $a \in \Sigma$.*

- *We have $\binom{uv}{e} = \sum_{e_1 e_2 = e} \binom{u}{e_1}\binom{v}{e_2}$.*
- *Let $\ell \geqslant 0$. We have $\binom{u}{a^\ell} = \binom{|u|_a}{\ell}$ and $\sum_{|v|=\ell} \binom{u}{v} = \binom{|u|}{\ell}$.*

We refine the second point of the above proposition:

Lemma 2.5. *Let $u, v \in \Sigma^*$. Then $\sum_{v' \equiv_1 v} \binom{u}{v'} = \prod_{a \in \Sigma} \binom{|u|_a}{|v|_a}$.*

Proof. We count the number of choices of subwords v' of u having $|v'|_a = |v|_a$ for each $a \in \Sigma$. For each $a \in \Sigma$, we may choose the occurrences of a in $\binom{|u|_a}{|v|_a}$ ways. Since the choices of distinct letters are independent, the total number of choices equals $\prod_{a \in \Sigma} \binom{|u|_a}{|v|_a}$. Each of these choices corresponds to an occurrence of a subword $v' \equiv_1 v$ of u. □

Example 2.6. Given two words $x, y \in \Sigma^*$, we have that $x \equiv_2 y$ if and only if $x \equiv_1 y$ and $\binom{x}{ab} = \binom{y}{ab}$ for all pairs of letters $a, b \in \Sigma$ with $a \lhd b$. Indeed, $x \equiv_1 y$ implies that $\binom{x}{aa} = \binom{|x|_a}{2} = \binom{|y|_a}{2} = \binom{y}{aa}$, and Lemma 2.5 implies that, for $a \lhd b$,

$$\binom{x}{ba} = |x|_a|x|_b - \binom{x}{ab} = |y|_a|y|_b - \binom{y}{ab} = \binom{y}{ba}.$$

2.2.2 Infinite Words

For an infinite word $\mathbf{x} \in \Sigma^{\mathbb{N}}$, we define factors, prefixes, and left quotients analogously and we use the same notation as for finite words. To distinguish infinite words from finite words, we write infinite words in boldface. An infinite word $\mathbf{y} \in \Sigma^{\mathbb{N}}$ such that $\mathbf{x} = u\mathbf{y}$ for some $u \in \Sigma^*$, is called a *tail* of $\mathbf{x}$. We call $\mathbf{x}$ *ultimately periodic* if there exist $u \in \Sigma^*, v \in \Sigma^+$ such that $\mathbf{x} = uv^\omega$. If, in the above, $u = \varepsilon$, then $\mathbf{x}$ is called *purely periodic*. If no such u and v exist, then $\mathbf{x}$ is called *aperiodic*.

An infinite word $\mathbf{x}$ is called *recurrent* if every non-empty factor $u \in F(\mathbf{x})$ occurs infinitely many times in $\mathbf{x}$. Moreover, $\mathbf{x}$ is called *uniformly recurrent* if, for each factor $u \in F(\mathbf{x})$, there exists an $N \in \mathbb{N}$ depending on u such that u occurs in each factor of x of length N. Further, $\mathbf{x}$ is called *linearly recurrent* if, for each $u \in F(\mathbf{x})$, there exists $K \in \mathbb{N}$ such that u occurs in each factor of $\mathbf{x}$ of length $K|u|$.

Let $\mathbf{x} \in \Sigma^{\mathbb{N}}$ and let u be a non-empty factor of $\mathbf{x}$. The set of *complete first returns to u in $\mathbf{x}$*, denoted by $\Re_{\mathbf{x}}(u)$, is defined as

$$\Re_{\mathbf{x}}(u) = \{v \in F(\mathbf{x}) \colon u \in \operatorname{pref}(v), u \in \operatorname{suff}(v), \text{ and } |v|_u = 2\}.$$

An element of $\Re_{\mathbf{x}}(u)u^{-1}$ is called a *return to u in $\mathbf{x}$*. We recall a result from [31].

Proposition 2.7 ([31, part of Proposition 2.6.])**.** *Let $p_1, \ldots, p_n \in \Re_{\mathbf{x}}(u)u^{-1}$. Then $|p_1 \cdots p_n u|_u = n + 1$ and $u \in \operatorname{pref}(p_i \cdots p_n u)$ for all $i = 1, \ldots, n$.*

The following result is used implicitly in [31]. We give a proof for the sake of completeness.

Corollary 2.8. *The set $\Re_{\mathbf{x}}(u)u^{-1}$ is an ω-*code*, and thus a* code*. That is, if $p_1 p_2 \cdots = q_1 q_2 \cdots$ for some $p_i, q_i \in \Re_{\mathbf{x}}(u)u^{-1}$, $i = 1, \ldots, \infty$, then $p_i = q_i$ for all $i = 1, \ldots, \infty$.*

Proof. Assume that $\mathbf{y} = p_1 p_2 \cdots = q_1 q_2 \cdots$ for some $p_i, q_i \in \Re_{\mathbf{x}}(u)u^{-1}$. Now, for all $n \geqslant 1$, we have $|p_1 p_2 \cdots p_n u|_u = n+1$. Let $n \geqslant |u|+1$. It follows that $p_2 \cdots p_n u$ begins with u, and thus $p_2 \cdots p_n$ begins with u, since $p_2 \cdots p_n$ has length at least $|u|$. By repeating the above to $q_1 q_2 \cdots q_n u$ for n large enough, we see that $q_2 \cdots q_n$ begins with u. Now $\mathbf{y}$ begins with both $p_1 u$ and $q_1 u$. Since $q_1 u$ and $p_1 u$ are complete first returns to u in $\mathbf{x}$, it follows that $q_1 = p_1$. We may argue inductively, using similar arguments, to show that $p_i = q_i$ for each $i = 1, \ldots, \infty$. The claim follows. □

We define complete first returns in finite words as well. We shall not make any distinction between these notions. We refer the reader to [31, 106] for more on the notion of first return words.

Definition 2.9. Let $\mathbf{x}$ be a finite or infinite word. The *complexity function* $\mathcal{C}_{\mathbf{x}}(n) : \mathbb{N} \to \mathbb{N}$ of $\mathbf{x}$ is defined as $\mathcal{C}_{\mathbf{x}}(n) = \#F_n(\mathbf{x})$.

As was remarked in the introduction, the notion of factor complexity turns out to be very fruitful: If $\mathcal{C}_{\mathbf{x}}(n) \leqslant n$ for some $n \geqslant 1$, then $\mathbf{x}$ is ultimately periodic. A family of words has been very tightly related to this notion, namely, the so-called *Sturmian words*.

Definition 2.10. An infinite word $\mathbf{s}$ is called *Sturmian*, if $\mathcal{C}_{\mathbf{s}}(n) = n + 1$ for each $n \geqslant 0$.

In particular, Sturmian words are binary words. An example of a Sturmian word is the famous Fibonacci word, which may be defined as the fixed point of the morphism $\varphi : \mathbb{B} \to \mathbb{B}$, $a \mapsto ab$, $b \mapsto a$.

The idea of identifying the complexity of an infinite word with the complexity of its finite factors may be extended to abelian equivalence and k-abelian equivalence.

Definition 2.11. Let $\mathbf{x} \in \Sigma^*$. For $k \geqslant 1$, we define the k-abelian complexity $\mathcal{C}^{(k)}_{\mathbf{x}}(n)$ of $\mathbf{x}$ is defined by $\mathcal{C}^{(k)}_{\mathbf{x}}(n) = F_n(\mathbf{x})/\!\sim_k$, that is, $\mathcal{C}^{(k)}_{\mathbf{x}}(n)$ counts the number of distinct k-abelian equivalence classes among words of length n represented by factors of $\mathbf{x}$.

In the special case of $k = 1$, we use the notation $\mathcal{C}^{\mathrm{ab}}_{\mathbf{x}}(n) = \mathcal{C}^{(1)}_{\mathbf{x}}(n)$.

In general, the abelian complexity function $\mathcal{C}^{\mathrm{ab}}$ can be strongly fluctuating (see, e.g., [66, 58]), so, in a general setting, it is more meaningful to study the asymptotic behavior of the abelian complexity function. To this end, we define the *upper* (resp., *lower*) *abelian complexity functions*, $\mathcal{U}^{\mathrm{ab}}_{\mathbf{x}}$ (resp., $\mathcal{L}^{\mathrm{ab}}_{\mathbf{x}}$), of a word $\mathbf{x} \in \Sigma^{\mathbb{N}}$ as

$$\mathcal{U}^{\mathrm{ab}}_{\mathbf{x}}(n) = \max\{\mathcal{C}^{\mathrm{ab}}_{\mathbf{x}}(m) \colon 0 \leqslant m \leqslant n\} \text{ (resp., } \mathcal{L}^{\mathrm{ab}}_{\mathbf{x}}(n) = \min\{\mathcal{C}^{\mathrm{ab}}_{\mathbf{x}}(m) \colon m \geqslant n\}.)$$

The asymptotic growth rates of these functions indicate how large the fluctuation of the abelian complexity of x can be.

2.2.3 Other Basic Notions

A mapping $\varphi : \Delta^* \to \Sigma^*$ from the language Δ^* to the language Σ^* is called a *morphism* if $\varphi(uv) = \varphi(u)\varphi(v)$ for all $u, v \in \Delta^*$. The notion of a morphism extends naturally to infinite words, and we will not make a distinction between the two. We say that φ is *uniform* if the lengths of the images of letters are all equal. Throughout the text, when speaking of *binary morphisms*, we specifically mean morphisms $\mathbb{B}^* \to \mathbb{B}^*$.

For an ordering of $\Sigma = \{a_1, a_2, \ldots, a_{|\Sigma|}\}$ and a morphism $\varphi : \Sigma^* \to \Sigma^*$, the *incidence matrix* A_φ of φ is defined as $A_\varphi[i, j] = |\varphi(a_j)|_{a_i}$. In other words, the jth entry of the ith row equals the number of occurrences of a_i in $\varphi(a_j)$. It is straightforward to conclude that when $\varphi \colon \Sigma^* \to \Sigma^*$, we have $A_{\varphi^n} = A^n_\varphi$ for all $n \in \mathbb{N}$. The morphism φ is called *primitive* if there exists $n_0 \in \mathbb{N}$ such that $A^{n_0}_\varphi$ contains only positive entries. In the case of the binary alphabet $\mathbb{B}$, we fix $a_1 = a$, $a_2 = b$ so that, given a binary morphism φ, A_φ is of the form

$$A_\varphi = \begin{pmatrix} |\varphi(a)|_a & |\varphi(b)|_a \\ |\varphi(a)|_b & |\varphi(b)|_b \end{pmatrix}.$$

Let $\varphi : \Sigma^* \to \Sigma^*$ be a morphism satisfying $\varphi(a) = ah$ for some $a \in \Sigma$ and a word $h \in \Sigma^+$ such that $\lim_{n\to\infty} |\varphi^n(h)| = \infty$. Then the word $\varphi^\omega(a) = \lim_{n\to\infty} \varphi^n(a)$ exists and is a fixed point of φ. A word $\mathbf{x} \in \Sigma^{\mathbb{N}}$ is called *pure morphic* if there exist a letter $a \in \Sigma$ and a morphism φ such that $\mathbf{x} = \varphi^\omega(a)$. Further, $\mathbf{x}$ is called *primitive pure morphic*, if such a primitive morphism φ exists. A word is said

to be *morphic* if it is a morphic image of a pure morphic word. In other words, $\mathbf{y} \in \Sigma^{\mathbb{N}}$ is morphic if there exist a pure morphic word $\mathbf{x} \in \Delta^{\mathbb{N}}$ and a morphism $\gamma : \Delta \to \Sigma^*$ such that $\mathbf{y} = \gamma(\mathbf{x})$.

We make extensive use of the so-called *de Bruijn graphs*. For any $k \geqslant 1$ and alphabet Σ, the *de Bruijn graph* $dB_\Sigma(k)$ *of order* k *over* Σ is defined as a directed graph for which $V(dB_\Sigma(k)) = \Sigma^k$. There is an edge $(x, y) \in E(dB_\Sigma(k))$ if there exists a letter $a \in \Sigma$ such that the word $xa \in \Sigma^{k+1}$ ends with y. In this case (x, y) is denoted by (x, a) (since y is uniquely defined by the word xa). We shall often omit Σ from the subscript, as this is usually clear from context.

We note that any word $u = a_0 \cdots a_n$, where $n \geqslant k - 1$ and $a_i \in \Sigma$ for each $i = 0, \ldots, n$, defines a walk $W_u = (e_i)_{i=0}^{n-k-1}$ in $dB(k)$. Here $e_i = (u[i, i + k), a_{i+k})$ for each $i = 0, \ldots, n - k - 1$. Conversely, any walk $W = ((x_i, a_i))_{i=1}^{t}$ in $dB(k)$ defines the word $x_1 \cdot a_1 \cdots a_t \in \Sigma^{k+t}$ of length $k + t$. Thus a (long enough) word $u \in \Sigma^*$ is often identified as a walk in $dB(k)$ and vice versa. Now each edge in the de Bruijn graph $dB(k)$ is uniquely defined by the initial vertex and its label. Thus, any walk $W = ((x_i, a_i))_{i=1}^{t}$ is uniquely defined by the initial vertex $x_1 \in \Sigma^k$ and the labels of the edges; we may write $W = (x_1, a_1 \cdots a_t)$.

We observe that, for example, for a primitive word x of length k, the walk x^2 defines the cycle (x, x) in the de Bruijn graph $dB(k)$. The vertices of this cycle equals the necklace defined by x. For the history of de Bruijn graphs, see, e.g., [15] and the references therein.

We refer the reader to [23, 63, 64] for more on basic notions in combinatorics on words.

2.2.4 Semigroups and Equations

A semigroup is a set S equipped with an associative binary operation $\cdot$. A semigroup is called a *monoid* if S contains an *identity element* e, that is, an element e satisfying $e \cdot x = x = x \cdot e$ for each $x \in S$. The set of finite, non-empty words Σ^+ equipped with the operation of concatenation forms a semigroup called the *free semigroup* on Σ. Thus the set of finite words Σ^* forms a *monoid*, when equipped with the concatenation operation. An equivalence relation $\sim$ is called a refinement of another equivalence relation $\equiv$, if $x \sim y$ implies $x \equiv y$.

Definition 2.12. An equivalence relation $\mathcal{R}$ on a semigroup S is called a *congruence*, if for all $x, y, z \in S$ it holds that $x\mathcal{R}y$ implies $zx\mathcal{R}zy$ and $xz\mathcal{R}yz$.

It is straightforward to see that an equivalence relation $\mathcal{R}$ on a semigroup S is a congruence if and only if $x_1\mathcal{R}y_1$ and $x_2\mathcal{R}y_2$ implies that $x_1x_2\mathcal{R}y_1y_2$. Let $x\mathcal{R}$ denote the equivalence class represented by x. Then, for a congruence $\mathcal{R}$, the quotient $S/\mathcal{R} = \{x\mathcal{R} \colon x \in S\}$ is a semigroup when equipped with the operation $x\mathcal{R} \cdot y\mathcal{R} = (xy)\mathcal{R}$. In this thesis we only consider congruences on the free monoids Σ^* for a finite alphabet Σ. In fact we only consider two congruences in depth:

Proposition 2.13. *For any $k \geqslant 1$, the k-abelian and the k-binomial equivalence relations are congruences.*

Proof. The k-abelian equivalence being a congruence is a straightforward corollary of Lemma 3.3. The k-binomial equivalence is shown to be a congruence in [93]. □

Note that $\Sigma^*/\equiv_k$ is *cancellative* as a monoid, that is, for $x, y, z \in \Sigma^*/\equiv_k$, $xy \equiv_k xz$ implies $y \equiv_k z$ and $yx \equiv_k zx$ implies $y \equiv_k z$. This can be seen by a straightforward induction utilizing the first point of Proposition 2.4. The monoid $\Sigma^*/\sim_k$ is, on the other hand, not cancellative when $k \geqslant 2$ (consider the words $x = a$, $y = a^{k-2}ba^k$, and $z = a^{k-1}ba^{k-1}$).

We introduce the terminology of semigroups to discuss particular aspects of the k-abelian equivalence, namely, equations. Let Ξ be a finite non-empty set of *variables* and S a semigroup (we assume that $\Xi^+ \cap S = \emptyset$). An element $(u, v) \in (\Xi \cup S)^+ \times (\Xi \cup S)^+$ is called an *equation* over S with variables Ξ. A *solution* to an equation (u, v) over S with variables Ξ is a morphism $\alpha : \Xi \to S$ such that $\alpha(u) = \alpha(v)$ (α is the identity morphism on S). An equation $e = (u, v)$ is often denoted by $e : u = v$. The set of solutions to the equation e is denoted by $\mathrm{Sol}(e)$.

The following properties of words inferred from the following example are used throughout the thesis without explicit mention.

Example 2.14 ([64])**.** For two word $u, v \in \Sigma^*$ we have $uv = vu$ if and only if there exists $r \in \Sigma^*$ such that $u, v \in r^*$. Thus the set $\mathrm{Sol}(xy = yx)$ of solutions to the equation $xy = yx$ in Σ^* equals $\{\alpha : x \mapsto r^i, y \mapsto r^j : r \in \Sigma^*, i, j, \geqslant 0\}$.

For words $x, y, z \in \Sigma^*$ we have $xz = zy$ if and only if there exist $p, q \in \Sigma^*$ such that $x = pq$, $y = qp$, and $z \in (pq)^*p$. In the free monoid, we thus have $\mathrm{Sol}(xz = zy) = \{(x, y, z) \mapsto (pq, qp, (pq)^r p,) : p, q \in \Sigma^*, r \in \mathbb{N}\}$.

We also consider *systems* $E \subseteq \Xi^+ \times \Xi^+$ of equations and the set

$$\mathrm{Sol}(E) = \bigcap_{e \in E} \mathrm{Sol}(e)$$

of solutions to E. We say that two systems E_1 and E_2 of equations are *equivalent* if $\mathrm{Sol}(E_1) = \mathrm{Sol}(E_2)$. Further, we say that a system of equations E is *independent* if E is not equivalent to any of its *finite* proper subsystems $E' \subseteq E$.

There are still interesting open problems regarding independent systems of equations over the free semigroup. Very recently, two longstanding open problems were solved by D. Nowotka and A. Saarela in [74]. Let us state one of them explicitly.

Theorem 2.15 ([74])**.** *Let E be a independent system of constant-free equations on three variables for the free semigroup. Then the system E contains at most 18 equations.*

The problem of upper bounding the number of equations in such a system was open for quite some time. It was conjectured by J. Karhumäki and K. Culik in [25] that the bound is three. In fact, before giving a uniform upper bound of 18, the previous best bound by Nowotka and Saarela was that the number of equations in such a system is logarithmically bounded with respect to the length of the shortest equation [75]. We refer the reader to [74] for more on the development of the solution of the above theorem.

In Chapter 7 we consider equations over the monoids defined by the k-abelian equivalence and the k-binomial equivalence over Σ^*, and consider independent systems of equations in these monoids. We discuss some further theory in that chapter.

2.3 Notions and Terminology from Language Theory

2.3.1 Automata and Formal Languages

Regular expressions over an alphabet Σ are the finite expressions constructed recursively by using the following operations. The symbol $\emptyset$, and each $a \in \Sigma \cup \{\varepsilon\}$ are expressions. If E and E' are expressions, then so are $(E \cdot E')$, $(E + E')$, and (E^*). Each expression E defines a language, denoted by $L(E)$ as follows: Each $a \in \Sigma \cup \{\varepsilon\}$ defines the singleton language $L(a) = \{a\}$ and $\emptyset$ defines the empty language. For expressions E and E', the expressions $(E \cdot E')$, $(E + E')$ and (E^*) define the languages $L(E) \cdot L(E')$, $L(E) \cup L(E')$, and $\cup_{n \geqslant 0} L(E)^n$, respectively.

A *deterministic finite automaton* (DFA) $\mathcal{A}$ over Σ is a tuple (Q, q_0, δ, F), where Q is a finite set of states, q_0 is the initial state, δ is a partial function $\delta : Q \times \Sigma \to Q$ called the *transition function*, and $F \subseteq Q$ is the set of final states. Given a word $w = a_0 \cdots a_n \in \Sigma^*$, the automaton operates on w using δ starting from q_0 by the rule $\delta(q, au) = \delta(\delta(q, a), u)$ for all $u \in \Sigma^+$. If $\delta(q_0, w) \in F$ we say that $\mathcal{A}$ *accepts* w, otherwise $\mathcal{A}$ rejects w. We let $L(\mathcal{A})$ denote the language *recognized* by $\mathcal{A}$; $L(\mathcal{A}) = \{w \in \Sigma^* \mid \mathcal{A} \text{ accepts } w\}$. We identify a DFA $\mathcal{A}$ with a directed labeled multigraph graph $G_{\mathcal{A}}$, where $V(G_{\mathcal{A}}) = Q$ and $(q_1, q_2) \in E(G_{\mathcal{A}})$ with label a if and only $\delta(q_1, a) = q_2$.

The languages defined by regular expressions are exactly the languages recognized by finite automata; that is, these models are equivalent. Such languages are called *regular languages*. Another equivalent model for regular languages considered in this thesis are *non-deterministic finite automata* with *ε-transitions* (ε-NFA), in which case the transition function may be multi-valued and we may have transitions for the empty word.

We need knowledge of several closure properties of regular languages. First, given two regular languages L and $L' \subseteq \Sigma^*$, the union and concatenation of these languages is regular (this is by definition of regular expressions). Also, the intersection $L \cap L'$ is regular. Further, the complement $\Sigma^* \setminus L$ and the difference $L \setminus L'$ are regular languages. We also need the knowledge of the following closure properties. Given a regular language $L \subseteq \Sigma^*$, the language

- $\varphi(L) = \{\varphi(x) \colon x \in L\}$ (morphic image);
- $\varphi^{-1}(L) = \{u \in \Delta^* \colon \varphi(u) \in L\}$, $\varphi : \Delta \to \Sigma^*$ (morphic preimage);
- $u^{-1}L$, Lu^{-1} (left/right quotient);

is regular. We refer to [45] for these facts, and on more of equivalent models and closure properties of regular languages. All the above properties are used without being explicitly mentioned.

For a regular language L there exists a DFA having the least number of states among deterministic finite automata recognizing L. We call such an automaton a *minimal DFA*. For a regular language, there exists a unique minimal automaton recognizing L. We recall a well-known results from the literature concerning minimal DFA. To this end, we define the following. Let p be a state of a DFA $\mathcal{A}$, F be the set of accepting states of $\mathcal{A}$, and δ be the transition function of $\mathcal{A}$. We

define the language L_p as the set of words u for which $\delta(p,u) \in F$. We call two states p and q *equivalent*, if $L_p = L_q$.

Theorem 2.16. *Let $\mathcal{A}$ be a minimal DFA. Then two states p and q are equivalent if and only if $p = q$.*

For a proof and related properties of regular languages, see [45].

2.3.2 Generating Functions and Rational Sequences

We now turn to the *generating functions* of the automata described above. For a general treatment on the topic of generating functions, see [37]. We shall briefly recall results concerning formal languages. To this end, let $L \subseteq \Sigma^*$ be a language. The *(ordinary) generating function* G_L of L is defined as the formal power series

$$G_L(x) = \sum_{k=0}^{\infty} \mathcal{C}_L(k) x^k.$$

We often omit the summation bounds to avoid cluttering the text; we always have $k = 0, \ldots, \infty$. For two generating functions $G_1(x) = \sum_k a_k x^k$ and $G_2(x) = \sum_k b_k x^k$, the sum $G_1(x) + G_2(x)$ and the product $G_1(x) \cdot G_2(x)$ are defined as

$$G_1(x) + G_2(x) = \sum_k (a_k + b_k) x^k \quad \text{and}$$
$$G_1(x) \cdot G_2(x) = \sum_k \Big(\sum_{i+j=k} a_i b_j \Big) x^k.$$

For disjoint languages $L_1, L_2 \subseteq \Sigma^*$, we have

$$G_{L_1 \cup L_2}(x) = G_{L_1}(x) + G_{L_2}(x),$$

as can straightforwardly be verified. If the product

$$L_1 L_2 = \{u_1 u_2 \mid u_1 \in L_1,\ u_2 \in L_2\}$$

of languages L_1 and L_2 is *unambiguous*, that is, $u_1 v_1 = u_2 v_2$ for some $u_1, u_2 \in L_1$, $v_1, v_2 \in L_2$ implies $u_1 = u_2$ and $v_1 = v_2$, then $G_L(x) = G_{L_1}(x) \cdot G_{L_2}(x)$, as can be readily verified.

In addition to classical language theoretical properties, we make use of the theory of *languages with multiplicities*. This counts how many times a word occurs in a language. This leads to the theory of $\mathbb{N}$-*rational sets*. Using the terminology of [98], a multiset over Σ^* is called $\mathbb{N}$-*rational* if it is obtained from finite multisets by applying finitely many times the rational operations *product*, *union*, and taking *quasi-inverses*, i.e., *iteration* restricted to ε-free languages. Further, a unary $\mathbb{N}$-rational subset is referred to as an $\mathbb{N}$-*rational sequence*. We refer to [98] for more on this topic. The basic result we need is (see [98]):

Proposition 2.17. *Let $\mathcal{A}$ be a non-deterministic finite automaton over the alphabet Σ. The function $f_{\mathcal{A}} : \Sigma^* \to \mathbb{N}$ defined as*

$$f_{\mathcal{A}}(w) = \#\textit{of accepting paths of } w \textit{ in } \mathcal{A}$$

is $\mathbb{N}$*-rational. In particular, the function* $\ell_{\mathcal{A}} : \mathbb{N} \to \mathbb{N}$*,*

$$\ell_{\mathcal{A}}(n) = \#\textit{of accepting paths of length } n \textit{ in } \mathcal{A} \tag{2.2}$$

is an $\mathbb{N}$*-rational sequence. Consequently, the* generating function *for* $\ell_{\mathcal{A}}$ *is a rational function.*

The above theorem implies that the generating function $G_L(x)$ of an arbitrary regular language L is an $\mathbb{N}$-rational sequence: We may take a DFA recognizing L and modify it into a unary automaton by identifying all the letters in the automaton. Then we may apply the above theorem to the modified automaton. Thus the sequence ℓ as in the above theorem equals, for each n, the number of words of length n in L, since the original DFA was deterministic. Now, in particular, the generating function $G_L(x)$ as a formal power series has a rational expression $\frac{p(x)}{q(x)}$ for some polynomials p, q.

Example 2.18. The regular language L_1 defined by the expression a^* has the generating function

$$G_{L_1}(x) = \sum_k x^k,$$

which has the rational expression $\frac{1}{1-x}$. The language L_2 defined by the regular expression $(ab)^*$ has the generating function

$$G_{L_2}(x) = \sum_k x^{2k}$$

which has the rational expression $\frac{1}{1-x^2}$. Finally, the regular language L_3 defined by the expression a^*b^* has the generating function

$$G_{L_3}(x) = (\sum_k x^k)(\sum_k x^k) = \sum_k (k+1)x^k$$

which has the rational expression $\frac{1}{(1-x)^2}$.

Chapter 3

Characterizations of k-abelian equivalence

In this chapter we present several characterizations of the k-abelian equivalence relation from the literature. We also introduce a characterization, which has not been previously published, using matrix semigroups. We learn that there are several equivalent definitions of the k-abelian equivalence from different points of view, all of which contribute to the broader view of this equivalence relation.

The first definition of k-abelian equivalence (Definition 2.1) implies that to check the k-abelian equivalence of two words, one needs to compute the number of occurrences of all factors of length at most k. In the next section, we present further characterizations which compare the numbers of occurrences of factors, but for which the number of factors needing comparisons is decreased. These characterizations are also very useful in other aspects, for example, in understanding the structure of k-abelian equivalence classes, and determining the asymptotic number of k-abelian equivalence classes.

The second section of this chapter presents a characterization based on a word rewriting rule. This point of view is totally different to the previous characterizations. As a consequence, this characterization opens the way to a language theoretic approach on the k-abelian equivalence, and this area is explored in subsequent chapters.

In the third section, we consider the k-abelian equivalence in de Bruijn graphs. Using tools from the rich theory of graphs, further properties of k-abelian equivalence classes are obtained. One immediate consequence is a formula for computing the cardinality of a k-abelian equivalence class defined by a given word. The graph theoretic interpretation of the k-abelian equivalence is also used at several other occasions, mainly in Chapters 4 and 6.

This chapter is closed off with the section concerning a matrix representation of the k-abelian equivalence. More precisely, a finitely generated matrix monoid is shown to be isomorphic to that of the monoid $\Sigma^*/\sim_k$. This point of view gives yet another glimpse of k-abelian equivalence in the setting of monoids and, in particular, equations over k-abelian equivalence classes.

3.1 Characterizations by Counting Occurrences of Factors

In [57] a combinatorial characterization of the k-abelian equivalence is obtained. The proof is quite straightforward. We repeat this characterization here, as it will be used on several occasions further on.

Proposition 3.1. *Let $u, v \in \Sigma^*$ be of length at least $k-1$ for some positive integer k. Assume that $|u|_x = |v|_x$ for all $x \in \Sigma^k$. The following are equivalent.*

1) $|u|_t = |v|_t$ *for all* $t \in \Sigma^{\leqslant k-1}$ *(that is, $u \sim_k v$);*
2) $|u|_t = |v|_t$ *for all* $t \in \Sigma^{k-1}$*;*
3) $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ *and* $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$*;*
4) $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$*;*
5) $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$*;*
6) $\mathrm{pref}_i(u) = \mathrm{pref}_i(v)$ *and* $\mathrm{suff}_{k-i-1}(u) = \mathrm{suff}_{k-i-1}(v)$ *for some* $0 \leqslant i \leqslant k-1$.

Observe that Item 3) above is included in both Items 4) and 5). A useful characterization of k-abelian equivalence is thus immediately obtained:

Lemma 3.2. *Let $k \geqslant 1$. For all $u, v \in \Sigma^*$ we have $u \sim_k v$ if and only if $|u|_x = |v|_x$ for all $x \in \Sigma^k$ and either*

$$\mathrm{pref}_{\min\{|u|,k-1\}}(u) = \mathrm{pref}_{\min\{|v|,k-1\}}(v) \text{ or}$$
$$\mathrm{suff}_{\min\{|u|,k-1\}}(u) = \mathrm{suff}_{\min\{|v|,k-1\}}(v).$$

Using this characterization the following properties of the k-abelian equivalence are easy to see.

Lemma 3.3 ([57])**.** *Let $k \geqslant 1$.*

- *For words $u, v \in \Sigma^*$ with $|u|, |v| \leqslant 2k-1$, if $u \sim_k v$ then $u = v$.*
- *For words $u_1, u_2, v_1, v_2 \in \Sigma^*$, if $u_1 \sim_k v_1$ and $u_2 \sim_k v_2$, then $u_1u_2 \sim_k v_1v_2$. In other words, $\sim_k$ is a congruence relation (see Definition 2.12.)*

Observe that the second item above implies that $\Sigma^*/\!\sim_k$ is a monoid. We consider this aspect in Chapter 7 when dealing with k-abelian equations. Indeed, to have a meaningful theory of equations, a monoid structure is required.

Remark 3.4. For any $k \geqslant 1$, the following relation ρ_k is an equivalence relation on Σ^*: $u \rho_k v$ if and only if $|u|_x = |v|_x$ for each $x \in \Sigma^k$. For $k = 1$ this is the abelian equivalence. For $k \geqslant 2$ we note that ρ does not define a congruence: $a^{k-2}ba^kb \ \rho_k \ a^{k-1}ba^k$, but $|a \cdot a^{k-2}ba^kb|_{a^k} \neq |a \cdot a^{k-1}ba^k|_{a^k}$.

Observe that $\sim_k$ is a *refinement* of ρ_k defined above. As a brief sidestep, we make precise the relation between ρ_k and $\sim_k$.

Lemma 3.5. *The k-abelian congruence relation is the coarsest congruence refining ρ_k, i.e., any congruence R refining ρ_k is a refinement of $\sim_k$.*

Proof. Let R be a congruence refining ρ_k. First of all we show that uRv implies $|u| = |v|$. Indeed, since R is a congruence, then $u^r R v^r$, where r is an integer such that $r|u| \geqslant k$. It follows that $r|u| = r|v|$, since now $u^r \rho_k v^r$ and, for all $w \in \Sigma^*$ with $|w| \geqslant k$, we have $\sum_{x\in\Sigma^k} |w|_x = |w| - k + 1$. Next we show that if uRv for some $u, v \in \Sigma^*$, where $|u| < k$, then $u = v$. Indeed, assuming that $|u| < k$, let $t \in \Sigma^*$ such that $|tu| = k$. Then, since R is a congruence refining ρ_k, $tuRtv$ implies that $|tu|_{tu} = |tv|_{tu}$ and hence $u = v$. Assume now, for a contradiction, that there exist $u, v \in \Sigma^*$ such that uRv, but $u \not\sim_k v$. It follows that $|u| = |v| \geqslant k$ by the above observations. Further, since $u\rho_k v$, by the above characterization of $u \sim_k v$ it follows that $\mathrm{pref}_{k-1}(u) \neq \mathrm{pref}_{k-1}(v)$. Let $p = \mathrm{pref}_{k-1}(u)$ and $q = \mathrm{pref}_{k-1}(v)$. Then $auRav$ but $|au|_{ap} = |u|_{ap} + 1 = |v|_{ap} + 1 = |av|_{ap} + 1$, so that R is not a refinement of ρ_k. This contradiction shows that R is a refinement of $\sim_k$. □

The definition of k-abelian equivalence as in Definition 2.1 implies that, in order to check whether two words u and v are k-abelian equivalent, one needs to check the occurrences of $\sum_{i=1}^{k} |\Sigma|^i = \frac{|\Sigma|^{k+1}-|\Sigma|}{|\Sigma|-1}$ many words in both u and v. However, the above characterization implies that one needs to check the occurrences of only $|\Sigma|^k$ factors (together with coinciding prefixes or suffixes).

Example 3.6. Let $u = aaba$ and $v = abaa$. Since $|u|_x = |v|_x$ for each $x \in \mathbb{B}^2$ and $\mathrm{pref}_1(u) = \mathrm{pref}_1(v) = a$, it follows that u and v are k-abelian equivalent.

For $x = aaba$ and $y = baab$, we see that $\mathrm{pref}_1(x) \neq \mathrm{pref}_1(y)$, that is, $x \not\sim_2 y$.

An interesting improvement of the above characterization can be found in [20]. In fact, this characterization of k-abelian equivalence gives a nice argument for a proof of Theorem 1.1 presented in [20]. Let us first recall this characterization.

Lemma 3.7. *Let $a \in \Sigma$. For each $k \geqslant 1$ define $Z_{a,k} = (\Sigma^{\leqslant k} \setminus a\Sigma^*) \setminus \Sigma^* a$. Then $u \sim_k v$ if and only if*

- $|u| = |v|$,
- $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ *and* $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$, *and*
- $|u|_x = |v|_x$ *for all* $x \in Z_{a,k}$.

To check whether two words $u, v \in \Sigma^*$ are k-abelian equivalent over an m-letter alphabet Σ, it suffices to check the number of occurrences of

$$\#(Z_{a,k} \setminus \{\varepsilon\}) = (m-1) + (m-1)^2 \sum_{i=0}^{k-2} m^i = m^{k-1}(m-1)$$

factors (together with checking the prefixes and suffixes of length $k-1$ and the lengths of u and v).

Example 3.8. Recall Example 3.6, where the 2-abelian equivalence of $u = aaba$ and $v = abaa$ was asserted using Lemma 3.2. Let us repeat this task using Lemma 3.7. One checks that $\mathrm{pref}_1(u) = \mathrm{pref}_1(v)$, $\mathrm{suff}_1(u) = \mathrm{suff}_1(v)$, $|u|_b = |v|_b$, and $|u|_{bb} = |v|_{bb}$. Thus $u \sim_k v$.

Let us briefly sketch the proof of Theorem 1.1 given in [20]. The upper bound $\mathcal{O}(n^{m^{k-1}(m-1)})$ is obtained by a straightforward combinatorial argument from the above lemma. For the lower bound $\Omega(n^{m^{k-1}(m-1)})$, let $\Sigma = \{0, \ldots, m-1\}$ be an

m-letter alphabet and let $\Sigma_M = \{0, \dots, M-1\}$, where $M = m^{k-1}(m-1)+1$. Let us define an ordering of the set $Z_{0,k} = \{y_0, \dots, y_{M-1}\}$ (over Σ) and, finally, define $h : \Sigma_M \to \Sigma$, $h(i) = y_i 0^{2k-1-|y_i|}$. Now, for two words $u, v \in \Sigma_M^*$, we have $h(u) \sim_k h(v)$ if and only if $u \sim_1 v$ and $\mathrm{pref}_{k-1}(h(u)) = \mathrm{pref}_{k-1}(h(v))$ [20, Lemma 3.3]. It is then straightforward to exhibit, for n large enough, $\Theta(n^{m^{k-1}(m-1)})$ distinct words, no two of which being k-abelian equivalent [20, Lemma 4.2].

3.2 A Characterization by Rewriting

In this subsection we describe rewriting rules of words which preserve k-abelian equivalence classes. This provides a different characterization of k-abelian equivalence which opens new automata theoretic aspects of the equivalence relation. This section is based on the article [55].

Definition 3.9. Let $k \geqslant 1$ and let $u = u_0 \cdots u_{n-1}$. Suppose further that there exist indices i, j, ℓ and m, with $0 \leqslant i < j \leqslant \ell < m \leqslant n-k+1$, such that $u[i, i+k-1) = u[\ell, \ell+k-1) = x$ and $u[j, j+k-1) = u[m, m+k-1) = y$ for some $x, y \in \Sigma^{k-1}$. We thus have

$$u = u[0, i) \cdot u[i, j) \cdot u[j, \ell) \cdot u[\ell, m) \cdot u[m..],$$

where $u[i..]$ and $u[\ell..]$ begin with x and $u[j..]$ and $u[m..]$ begin with y. Note here that we allow $\ell = j$ (in this case $y = x$). We define a *k-switching* on u, denoted by $S_{u,k}(i, j, \ell, m)$, as

$$S_{u,k}(i, j, \ell, m) = u[0, i) \cdot u[\ell, m) \cdot u[j, \ell) \cdot u[i, j) \cdot u[m..]. \tag{3.1}$$

In other words, for a word $u = a_0 \cdots a_{n-1}$, a k-switching $S_{u,k}(i, j, \ell, m) = v$ can be seen as a permutation π on the set $\{0, \dots, n-1\}$:

$$(0, \dots, n-1) \overset{\pi}{\mapsto} (0, \dots, i-1, \ell, \dots, m-1, j, \dots, \ell-1, i, \dots, j-1, m, \dots, n-1),$$

and $S_{u,k}(i, j, \ell, m) = a_{\pi(0)} \cdots a_{\pi(n-1)}$.

A k-switching operation is illustrated in Figure 3.1. We also give a simple example of applying k-switchings on a word.

Example 3.10. Let $u = aababababaaabab$ and $k = 4$. Let then $x = aba$, $y = bab$, $i = 1$, $j = 2$, $\ell = 3$ and $m = 10$. We then have

$$\begin{aligned} u &= a \cdot a \cdot b \cdot ababaaa \cdot bab \\ S_{u,4}(i, j, \ell, m) &= a \cdot ababaaa \cdot b \cdot a \cdot bab. \end{aligned}$$

One can check that $u \sim_4 S_{u,4}(i, j, \ell, m)$. Note that in this example the occurrences of x and y are overlapping.

Roughly speaking, the idea is to switch the positions of two factors who both begin and end with the same factors of length $k-1$, and we allow the situation where the factors can all overlap. We remark that, in the case of $j = \ell$, k-switchings were considered in a different context in [16].

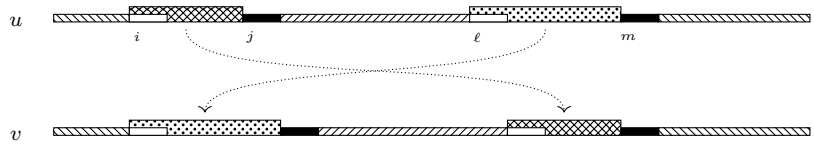

Figure 3.1: Illustration of a k-switching. Here $v = S_{k,u}(i, j, \ell, m)$; the white rectangles symbolize x and the black ones symbolize y. [18, Fig. 1], [19, Figure 1].

Let now $u = u_1u_2u_3u_4u_5$ and $v = u_1u_4u_3u_2u_5$, where $x \in \Sigma^{k-1}$ occurs at positions $|u_1|$ and $|u_1u_2u_3|$, and $y \in \Sigma^{k-1}$ occurs at positions $|u_1u_2|$ and $|u_1u_2u_3u_4|$ in u. Here v is a k-switching on u. We now show that x occurs at positions $|u_1|$ and $|u_1u_4u_3|$, and that y occurs at positions $|u_1u_4|$ and $|u_1u_4u_3u_2|$ in v. Indeed, recall that u_5 is assumed to begin with y and that $u_2 \neq \varepsilon \neq u_4$. Since we allow the occurrences of x and y to overlap, we have no other assumptions on the words u_i, $i = 1, \ldots, 4$. Firstly, since $\operatorname{pref}_{k-1}(u_5) = y = \operatorname{pref}_{k-1}(u_3u_4u_5)$, we have

$$\operatorname{pref}_{k-1}(u_2u_5) = \operatorname{pref}_{k-1}(u_2y) = \operatorname{pref}_{k-1}(u_2u_3u_4u_5) = x.$$

Secondly, by the above, we have

$$\operatorname{pref}_{k-1}(u_3u_2u_5) = \operatorname{pref}_{k-1}(u_3x) = \operatorname{pref}_{k-1}(u_3u_4u_5) = y.$$

Finally, by the previous observation, we have

$$\operatorname{pref}_{k-1}(u_4u_3u_2u_5) = \operatorname{pref}_{k-1}(u_4y) = \operatorname{pref}_{k-1}(u_4u_5) = x.$$

Using these observations, we show that k-switchings preserve k-abelian equivalence.

Lemma 3.11. *Let $u \in \Sigma^*$ and $v = S_{u,k}(i, j, \ell, m)$ be a k-switching on u. Then $u \sim_k v$.*

Proof. Let $u = u_1u_2u_3u_4u_5$ and $v = u_1u_4u_3u_2u_5$, where v is the k-switching with the indices $i = |u_1|$, $j = i + |u_2|$, $\ell = j + |u|_3$, and $m = \ell + |u_4|$. For ease of notation, define $s_k(x) = \operatorname{suff}_{\min\{|u|,k\}}(u)$ for any $x \in \Sigma^*$. Observe now that

$$\begin{aligned}
\Psi_k(u) &= \textstyle\sum_{i=1}^{5} \Psi_k(u_i) + \sum_{i=1}^{4} \Psi_k(s_{k-1}(u_i) \operatorname{pref}_{k-1}(u_{i+1} \cdots u_5)) \\
&= \textstyle\sum_{i=1}^{5} \Psi_k(u_i) + \Psi_k(s_{k-1}(u_1)x) + \Psi_k(s_{k-1}(u_2)y) \\
&\quad + \Psi_k(s_{k-1}(u_3)x) + \Psi_k(s_{k-1}(u_4)y) \\
&= \textstyle\sum_{i=1}^{5} \Psi_k(u_i) + \Psi_k(s_{k-1}(u_1) \operatorname{pref}_{k-1}(u_4u_3u_2u_5)) \\
&\quad + \Psi_k(s_{k-1}(u_4) \operatorname{pref}_{k-1}(u_3u_2u_5)) + \Psi_k(s_{k-1}(u_3) \operatorname{pref}_{k-1}(u_2u_5)) \\
&\quad + \Psi_k(s_{k-1}(u_4) \operatorname{pref}_{k-1}(u_5)) \\
&= \Psi_k(v),
\end{aligned}$$

where, in the third equality, we have used the observation preceding this lemma. Further, since the suffixes of u and v of length $k-1$ are equal, we have $u \sim_k v$ by Proposition 3.1. □

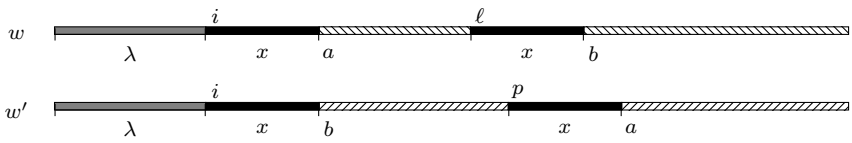

Figure 3.2: Illustration of the proof of Claim 3.13. [55, Fig. 2].

Let us define a relation R_k of Σ^* with uR_kv if and only if $v = S_{u,k}$ for some k-switching on u. Now R_k is clearly symmetric, so that the reflexive and transitive closure R_k^* of R_k is an equivalence relation. It is not hard to see that R_k^* is actually a congruence which refines ρ_k as defined in Remark 3.4. Thus, by Lemma 3.5, R_k^* refines $\sim_k$. In other words, uR_k^*v implies $u \sim_k v$. We now prove the converse, so that the relations $\sim_k$ and R_k^* actually coincide.

Theorem 3.12. *For two words $u, v \in \Sigma^*$, we have $u \sim_k v$ if and only if uR_k^*v.*

For the proof of the the above theorem, we need the following technical claim which will be used also later:

Claim 3.13. *Let $w \sim_k w'$, $w \neq w'$. Let λx be the longest common prefix of w and w' with $\lambda \in \Sigma^*$, $x \in \Sigma^{k-1}$, whence $w = \lambda xa\mu$ and $w' = \lambda xb\mu'$ for some $\mu, \mu' \in \Sigma^*$, $a, b \in \Sigma$, $a \neq b$. Then there exist $y \in \Sigma^{k-1}$ and indices j, ℓ, m, with $|\lambda| \leqslant j \leqslant \ell < m$, such that*

- $w[j, j+k-1) = y$,
- $w[\ell, \ell+k) = xb$, *and*
- $w[m, m+k-1) = y$.

Proof. It follows from $w \sim_k w'$ that w' has an occurrence of xa and w has an occurrence of xb after the common prefix λ. We let $i = |\lambda|+1$ be the position (i.e., the starting index of the occurrence) of xa in w and let ℓ be the minimal position (leftmost occurrence) of xb in w with $\ell > i$. Let p be a position of xa in w' with $p > i$ (see Figure 3.2).

Consider then the set $F_k(w'[i..])$; each word in this set occurs somewhere in $w[i..]$, since $w \sim_k w'$. Let then q, with $q \geqslant i$, be the minimal index such that the factor $w'[q, q+k)$ occurs in $w[i, \ell+k-1)$. Such an index exists since, for example, $w'[p, p+k) = w[i, i+k)$. Moreover, by the minimality of ℓ, we have $q > i$. Let $y = w'[q, q+k-1)$ and let j', where $i \leqslant j' \leqslant \ell-1$, be a position of y in w. We shall now choose the index j in the claim. If $j' > i$ we choose $j = j'$. If $j' = i$, then necessarily $x = y$ and we choose $j = \ell$.

We shall now choose the index m in the claim. By the choice of q, we have that $w'[q-1, q+k-1)$, an element of $F_k(w'[i..])$, occurs at some position m', $m' \geqslant \ell$, in w. It follows that y occurs in w at position $m = m'+1$, with $m > \ell$. We have now obtained the factor y and the positions of y and xb as claimed. □

Proof of Theorem 3.12. It is enough to show that $u \sim_k v$ implies uR_k^*v, since the converse follows from Lemma 3.11.

Assume that $u \sim_k v$ but $u \neq v$. Let ν be the longest common prefix of u and v, denoted by $\mathrm{lcp}(u, v)$. Applying Claim 3.13 to $w = u$ and $w' = v$, with $\nu = \lambda x$,

we obtain indices i, j, ℓ, m which give rise to a k-switching $S_{u,k}(i, j, \ell, m) = z$, such that $|\operatorname{lcp}(z, v)| > |\nu|$. We have $zR_k u$ and, by Lemma 3.11, $z \sim_k u$ and thus $z \sim_k v$. Again we may apply Claim 3.13 to z and v to obtain a word $zR_k z_1$ and $z_1 \sim_k v$ with $|\operatorname{lcp}(z_1, v)| > |\operatorname{lcp}(z, v)|$. Repeating these observations finitely many times, we obtain a sequence $u\ R_k\ z\ R_k\ z_1 \cdots R_k\ v$, and thus $uR_k^* v$. □

The characterization of k-abelian equivalence using k-switchings has a different flavour to the characterizations obtained previously, namely, it gives means for constructing k-abelian equivalent words.

3.3 k-abelian Equivalence Classes as Eulerian Walks

This section is based on [55]. We repeat an observation made in [57] connecting k-abelian equivalence with Eulerian paths in the multigraphs. Let $f \in \mathbb{N}^{\Sigma^k}$ be an arbitrary vector. We modify the de Bruijn graph $dB(k-1)$ with respect to f into a multigraph $G_f = (V, E)$ as follows. We define V as the set of words $x \in \Sigma^{k-1}$ such that x is a prefix or a suffix of a word $z \in \Sigma^k$ for which $f[z] > 0$. We define the set of edges as follows: for each $z \in \Sigma^k$ with $f[z] > 0$, we take the edge from u to v with multiplicity $f[z]$, where u is the length $k-1$ prefix of z, and v is the length $k-1$ suffix of z.

Note that for $f = \Psi_k(w)$, the graph G_f resembles the *Rauzy graph* of w of order $k-1$ (see [89]), with $V = F_{k-1}(w)$ and the edges of G_f corresponding to the set $F_k(w)$ with multiplicities.

In the following, for $u, v \in \Sigma^{k-1}$, we denote by $\Sigma(u, v)$ the set of words which begin with u and end with v: $\Sigma(u, v) = u\Sigma^* \cap \Sigma^* v$.

Lemma 3.14 ([57, Lemma 2.12.])**.** *For a vector $f \in \mathbb{N}^{\Sigma^k}$ and words $u, v \in \Sigma^{k-1}$, the following are equivalent:*

1. *there exists a word $w \in \Sigma(u, v)$ such that $f = \Psi_k(w)$,*
2. *G_f has an Eulerian path starting from u and ending at v,*
3. *the underlying graph of G_f is connected, and $d^-(s) = d^+(s)$ for every vertex s, except that if $u \neq v$, then $d^-(u) = d^+(u) - 1$ and $d^-(v) = d^+(v) + 1$.*

The above lemma is articulated in terms of the k-abelian equivalence in [55] as a corollary.

Corollary 3.15. *For a word $w \in \Sigma(u, v)$ and $k \geqslant 1$, we have that $w' \sim_k w$ if and only if w' induces an Eulerian path from u to v in $G_{\Psi_k(w)}$.*

Given a word w with length at least $k-1$, we may thus identify the k-abelian equivalence class $[w]_k$ as a weighted Eulerian subgraph of the de Bruijn graph of order $k-1$ together with the fixed start and end vertices.

Example 3.16. Let $u \in \Sigma^*$ and $x \in F_{k-1}(u)$ such that $|u|_x \geqslant 3$. We may then write $W_u = W_1 W_2 W_3 W_4$ for some walks W_i with $\operatorname{head}(W_i) = x = \operatorname{tail}(W_{i+1})$ for each $i \in \{1, 2, 3\}$. Then, by the above corollary, we have $u \sim_k v$, where v is defined by the walk $W_v = W_1 W_3 W_2 W_4$. Indeed, W_v is well-defined due to the choice of the extremal vertices of the walks W_i and the same edges are traversed equally many times as in W_u.

An almost immediate consequence of the above corollary is that we may express the cardinality $\#[w]_k$ of the k-abelian equivalence class defined by a given long enough word w. Let $w \in \Sigma(u, v)$, where $u, v \in \Sigma^{k-1}$, and let $f = \Psi_k(w)$. The value $\#[w]_k$ is now the number of *Eulerian walks* of G_f starting at u and ending at v. Here we consider two cycles to be distinct if the vertices are traversed in different orders.

Let us briefly recall a relevant result from the literature.

Definition 3.17. Let $G = (V, E)$ be a directed multigraph. The *Laplacian matrix* $\Delta(G)$ of G is defined as

$$\Delta(G)_{uv} = \begin{cases} -m(u, v), & \text{if } u \neq v, \\ d^+(u) - m(u, v), & \text{if } u = v. \end{cases}$$

For the Laplacian $\Delta(G)$ of a graph G and a vertex v of G, we denote by $\Delta(G)^{(v)}$ the matrix obtained by removing from $\Delta(G)$ the row and column corresponding to v.

Remark 3.18. We note that for a directed multigraph G and a vertex v, $\det(\Delta(G)^{(v)})$ counts the number of *rooted spanning trees with root* v in G. This result is known as *Kirchhoff's matrix tree theorem* (for a proof, see [1]).

We recall the *BEST theorem*, first discovered by C.A.B. Smith and W.T. Tutte in 1941 and later generalized by T. van Aardenne-Ehrenfest and N.G. de Bruijn (see [1]). For this, let $\epsilon(G)$ denote the number of distinct Eulerian cycles in an Eulerian graph G. Here two cycles are considered to be the same, if one is a cyclic shift of the other. Equivalently, $\epsilon(G)$ counts the number of distinct Eulerian cycles beginning from a fixed edge e.

Theorem 3.19 (BEST theorem)**.** *Let G be a connected directed Eulerian multigraph. Then*

$$\epsilon(G) = \det(\Delta(G)^{(u)}) \prod_{v \in V} (d^+(v) - 1)!,$$

where u is any vertex of G.

Now the number of Eulerian walks up to the order of vertices traversed is distinct to the number of Eulerian walks, that is, $\epsilon(G_f)$ (in the case of G_f being Eulerian). Nonetheless, we still employ the BEST theorem (Theorem 3.19) to obtain the desired value. Note that in G_f we have $d^+(x) = |w|_x$ for all $x \neq v$ and $d^+(v) = |w|_v - 1$.

Proposition 3.20. *Let $k \geqslant 1$ and $w \in \Sigma(u, v)$ for some $u, v \in \Sigma^{k-1}$. Then*

$$\#[w]_k = \det(\Delta(G)^{(v)}) \prod_{x \in F_{k-1}(w)} \frac{(|w|_x - 1)!}{\prod_{a \in \Sigma} |w|_{xa}!}, \tag{3.2}$$

where $G = G_{\Psi_k(w)}$.

Proof. Let $V = V(G)$, that is, $V = F_{k-1}(w)$, and let $f = \Psi_k(w)$. Suppose first that $u = v$, so that G contains an Eulerian cycle. We shall first count the number

of distinct Eulerian cycles starting from *vertex* v. Note here that two cycles are considered distinct if the *edges* are traversed in a different order.

It follows from the BEST theorem, that the number of Eulerian cycles starting from vertex v equals

$$d^+(v)\det(\Delta(G)^{(v)})\prod_{x\in V}(d^+(x)-1)! = \det(\Delta(G)^{(v)})\prod_{x\in V}(|w|_x-1)!. \tag{3.3}$$

Now two Eulerian cycles are induced by the same word z if and only if the *vertices* are traversed in the same order. The claim follows by dividing the right hand side of Equation (3.3) by the number of different ways to order the individual edges between two vertices x and y for all $x, y \in V$:

$$\prod_{(x,y)\in E} m(x,y)! = \prod_{x\in V}\prod_{a\in\Sigma} f[xa]!.$$

Suppose then that $u \neq v$. We shall now add to G a new edge $e = (v,u)$ to obtain H, an Eulerian graph. Observe that $d^+_H(v) = d^+_G(v)+1 = |w_v|$, the rest of the out-degrees remain the same. Furthermore, the number of Eulerian paths from u to v in G equals the number of Eulerian cycles beginning with e in H. We again invoke the BEST theorem: the number of Eulerian cycles beginning from edge e is

$$\begin{aligned}\det(\Delta(H)^{(v)})\prod_{x\in V}(d^+_H(x)-1)! &= \det(\Delta(H)^{(v)})\; d^+_G(v)!\prod_{\substack{x\in V\\ x\neq v}}(d^+_G(x)-1)!\\ &= \det(\Delta(H)^{(v)})\prod_{x\in V}(|w|_x-1)!,.\end{aligned} \tag{3.4}$$

Note that $\det(\Delta(H)^{(v)}) = \det(\Delta(G)^{(v)})$, since the Laplacians of G and H differ only in the row and column corresponding to v.

Similar to the previous case, we are not interested in which order the edges from x to y are traversed, with one exception: we have fixed the starting edge e. The right hand side of equation (3.4) should thus be divided by

$$(m_H(v,u)-1)!\prod_{\substack{(x,y)\in E\\ (x,y)\neq(v,u)}} m_H(x,y)! = \prod_{(x,y)\in E} m_{G_f}(x,y)! = \prod_{x\in V}\prod_{a\in\Sigma} f[xa]!.$$

The claim follows. □

Example 3.21. Let $w = ababaaaa$ and $f = \Psi_2(w)$. We have

$$f = (|w|_{aa}, |w|_{ab}, |w|_{ba}, |w|_{bb}) = (3,2,2,0).$$

The Laplacian of G_f is $\left(\begin{smallmatrix} 2 & -2 \\ -2 & 2\end{smallmatrix}\right)$, from which we obtain $\det(\Delta(a)) = 2$. The above proposition then gives us:

$$\#[w]_2 = \det(\Delta(a))\cdot\frac{(|w|_a-1)!(|w|_b-1)!}{|w|_{aa}!|w|_{ab}!|w|_{ba}!} = 2\cdot\frac{5!\cdot 1!}{3!\cdot 2!\cdot 2!} = 10.$$

3.4 Equivalence Classes as Matrices

This section is based on unpublished work of the manuscript [110]. Recall that two words are abelian equivalent if and only if their Parikh vectors are equal. This property may be interpreted in terms of monoids. Namely, the monoid $\Sigma^*/\sim_1$ is isomorphic to the monoid $\mathbb{N}^{|\Sigma|}$ equipped with coordinate-wise summation. We may also embed the monoid $\Sigma^*/\sim_1$ into the multiplicative monoid $(\mathbb{Z}^{2|\Sigma|\times 2|\Sigma|}, \cdot)$ of $2|\Sigma| \times 2|\Sigma|$ integer matrices. Indeed, for each letter $a \in \Sigma$ and $u \in \Sigma^*$, we define the 2×2-matrix $A_a(u) = \left(\begin{smallmatrix} 1 & |u|_a \\ 0 & 1 \end{smallmatrix}\right)$. By a simple computation one sees that $A_a(v)A_a(w) = A_a(vw)$ for all words $v, w \in \Sigma^*$. For $v \in \Sigma^*$ we define the matrix $\mathbf{A}(v) \in \mathbb{Z}^{2|\Sigma|\times 2|\Sigma|}$ as the block diagonal matrix

$$\mathbf{A}(v) = \begin{pmatrix} A_{a_1}(v) & 0 & \cdots & 0 \\ 0 & A_{a_2}(v) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{a_{|\Sigma|}}(v) \end{pmatrix},$$

where $\Sigma = \{a_1, \ldots, a_{|\Sigma|}\}$. It is clear that $u \sim_1 v$ if and only if $\mathbf{A}(u) = \mathbf{A}(v)$, as the numbers of occurrences of letters are indicated in the matrix. Moreover, it is straightforward to verify that $\mathbf{A}(u)\mathbf{A}(v) = \mathbf{A}(uv)$ for all words $u, v \in \Sigma$. Thus the multiplicative monoid generated by the matrices $\{\mathbf{A}(a) \colon a \in \Sigma\}$ is isomorphic to $\Sigma^*/\sim_1$. Note that by replacing the first diagonal block $A_{a_1}(v)$ by the scalar $2^{|v|}$ in $\mathbf{A}(v)$, we still have the above properties, and the dimension of the matrix is reduced by one.

Our aim is to generalize this construction for k-abelian equivalence for arbitrary k. Some technicalities are involved to ensure that the multiplication of matrices corresponds to the multiplication of k-abelian equivalence classes. More precisely, it is crucial to accommodate the property that, for all $v, w \in \Sigma^*$, $u \in \Sigma^k$,

$$|vw|_u = |v|_u + |w|_u + |\operatorname{suff}_{k-1}(v) \operatorname{pref}_{k-1}(w)|_u.$$

For the sake of readability we define, for two words $u, v \in \Sigma^*$, the characteristic function δ_v^u as $\delta_v^u = 1$ if $u = v$ and otherwise $\delta_u^v = 0$. Observe that for any $u, v, w \in \Sigma^*$ we have $\delta_{vw}^u = \sum_{u=u_1u_2} \delta_v^{u_1}\delta_w^{u_2}$, where the sum goes over all factorizations of u into two words.

Definition 3.22. Let $u \in \Sigma^+$ and write $u = u_0 \cdots u_{L-1}$. For all $i \geqslant 1$ and $v \in \Sigma^*$, let $s_i(v) = \operatorname{suff}_{\min\{i,|v|\}}(v)$ and $p_i(v) = \operatorname{pref}_{\min\{i,|v|\}}(v)$. Define, for any word $v \in \Sigma^*$, the $(L+1) \times (L+1)$ matrix

$$A_u(v) = \begin{pmatrix} 1 & |v|_u & \delta_{s_{L-1}(v)}^{\operatorname{pref}_{L-1}(u)} & \delta_{s_{L-2}(v)}^{\operatorname{pref}_{L-2}(u)} & \cdots & \delta_{s_2(v)}^{\operatorname{pref}_2(u)} & \delta_{s_1(v)}^{\operatorname{pref}_1(u)} \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \delta_{p_1(v)}^{\operatorname{suff}_1(u)} & \delta_v^\varepsilon & 0 & \cdots & 0 & 0 \\ 0 & \delta_{p_2(v)}^{\operatorname{suff}_2(u)} & \delta_v^{u[L-2,L-1)} & \delta_v^\varepsilon & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \delta_{p_{L-2}(v)}^{\operatorname{suff}_{L-2}(u)} & \delta_v^{u[2,L-1)} & \delta_v^{u[2,L-2)} & \cdots & \delta_v^\varepsilon & 0 \\ 0 & \delta_{p_{L-1}(v)}^{\operatorname{suff}_{L-1}(u)} & \delta_v^{u[1,L-1)} & \delta_v^{u[1,L-2)} & \cdots & \delta_v^{u[1,2)} & \delta_v^\varepsilon \end{pmatrix}.$$

More precisely, we define $A_u(v) = (a_{i,j})_{i,j=1}^{L+1}$ such that

- $a_{1,1} = a_{2,2} = 1$; $a_{1,2} = |v|_u$; $a_{2,j} = 0$ and $a_{i,1} = 0$ for all $j \neq 2$, $i \geqslant 2$;
- $a_{i+2,2} = \delta_{p_i(v)}^{\mathrm{suff}_i(u)}$ for all $1 \leqslant i < L$;
- $a_{1,j+2} = \delta_{s_{k-j}(v)}^{\mathrm{pref}_{k-j}(u)}$ for all $1 \leqslant j < L$;
- $a_{i+2,i+2} = \delta_v^{\varepsilon}$ for all $1 \leqslant i < L$;
- $a_{i+2,j+2} = \delta_v^{u[L-i,L-j)}$ if $i > j$ and $a_{i+2,j+2} = 0$ if $i < j$ for all $1 \leqslant i, j < L$.

We aim to show that $A_u(v)A_u(w) = A_u(vw)$ for all $u, v, w \in \Sigma^*$. Before doing so, we give an example of the construction of the above matrix for two values of the parameter u.

Example 3.23. We give examples on constructing the matrices $A_u(v)$. Let $a, b \in \Sigma$ be distinct letters. By the definition above, we have

$$A_{ab}(a) = \begin{pmatrix} 1 & 0 & \delta_a^a \\ 0 & 1 & 0 \\ 0 & \delta_a^b & \delta_a^{\varepsilon} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and}$$

$$A_{ab}(b) = \begin{pmatrix} 1 & 0 & \delta_b^a \\ 0 & 1 & 0 \\ 0 & \delta_b^b & \delta_b^{\varepsilon} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Observe now that

$$A_{ab}(a)A_{ab}(b)A_{ab}(a) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & |aba|_{ab} & \delta_a^a \\ 0 & 1 & 0 \\ 0 & \delta_a^b & \delta_{aba}^{\varepsilon} \end{pmatrix} = A_{ab}(aba).$$

We give another example of constructing the matrix $A_u(v)$. Again by definition we have

$$A_{aba}(a) = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_{aba}(b) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We make a similar observation as above:

$$A_{aba}(a)A_{aba}(b)A_{aba}(a) = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = A_{aba}(aba).$$

Remark 3.24. Observe that, for any $k \geqslant 1$ and $u \in \Sigma^k$, the matrix $A_u(\varepsilon)$ is the identity matrix. For any letter $a \in \Sigma$, the matrix $A_a(v)$ is always invertible, which is straightforward to verify. However, for $k \geqslant 2$ and $u \in \Sigma^k$, the matrix $A_u(v)$ is invertible if and only if $v = \varepsilon$. Indeed, for $v \neq \varepsilon$, the third row is either all zeroes or it equals the second row. Further, if $|v| \geqslant k - 1$, the matrix $A_u(v)$ is of simple form, namely, the lower right $k \times (k-1)$ submatrix is all zeroes.

Proposition 3.25. *For all $u \in \Sigma^L$ and all $v, w \in \Sigma^*$, $A_u(vw) = A_u(v)A_u(w)$.*

Proof. We first note that

$$|vw|_u = |v|_u + |w|_u + \sum_{i=1}^{L-1} \delta^u_{\mathrm{suff}_i(v)\,\mathrm{pref}_{L-i}(w)} = |v|_u + |w|_u + \sum_{i=1}^{L-1} \delta^{\mathrm{pref}_{L-i}(u)}_{\mathrm{suff}_{L-i}(v)} \cdot \delta^{\mathrm{suff}_i(u)}_{\mathrm{pref}_i(w)}.$$

Here $A_u(vw)[1,2] = |vw|_u$ and the right hand sum is the inner product of the first row of $A_u(v)$ and the second column of $A_u(w)$. We then note that for all $1 \leqslant \ell \leqslant L-1$

$$\delta^{\mathrm{pref}_{L-\ell}(u)}_{\mathrm{suff}_{L-\ell}(vw)} = \delta^{u[..L-\ell)}_{\mathrm{suff}_{L-\ell}(w)} + \delta^{u[..L-\ell)}_{\mathrm{suff}_{L-\ell}(v)} \cdot \delta^{\varepsilon}_{w} + \sum_{j=\ell+1}^{L-1} \delta^{u[..L-j)}_{\mathrm{suff}_{L-j}(v)} \cdot \delta^{u[L-j,L-\ell)}_{w}.$$

Here $A_u(vw)[1,\ell+2] = \delta^{\mathrm{pref}_{L-\ell}(u)}_{\mathrm{suff}_{L-\ell}(vw)}$ and the sum on the right hand side equals the inner product the first row of $A_u(v)$ and column $\ell+2$ of $A_u(w)$. Further,

$$\delta^{\mathrm{suff}_\ell(u)}_{\mathrm{pref}_\ell(vw)} = \delta^{\mathrm{suff}_\ell(u)}_{\mathrm{pref}_\ell(v)} + \sum_{j=1}^{\ell-1} \delta^{u[L-\ell,L-j)}_{v} \delta^{u[L-j..)}_{w} + \delta^{\varepsilon}_{v} \cdot \delta^{\mathrm{suff}_\ell(u)}_{\mathrm{pref}_\ell(w)},$$

where $A_u(vw)[\ell+2,2] = \delta^{\mathrm{suff}_{L-\ell}(u)}_{\mathrm{pref}_\ell(vw)}$ and the right hand sum is the inner product of row $\ell+2$ of $A_u(v)$ and the second column of $A_u(w)$. Finally, for all i, j, where $1 \leqslant i < j \leqslant L-1$, we have

$$\delta^{u[L-i,L-j)}_{vw} = \delta^{u[L-i,L-j)}_{v} \delta^{\varepsilon}_{w} + \sum_{\ell=i+1}^{j-1} \delta^{u[L-i,L-\ell)}_{v} \cdot \delta^{u[L-\ell,L-j)}_{w} + \delta^{\varepsilon}_{v} \delta^{u[L-i,L-j)}_{w},$$

where $A_u(vw)[i+2,j+2] = \delta^{u[L-i,L-j)}_{vw}$, and the sum on the right hand side is the inner product of row $i+2$ of $A_u(v)$ and column $j+2$ of $A_u(w)$. The claim now follows straightforwardly. □

Next we construct matrices corresponding to words using the above defined matrices $A_u(v)$ as building blocks. This gives rise to a matrix representation of the k-abelian equivalence.

Definition 3.26. Let $a \in \Sigma$ and let $Z_{a,k} = \Sigma^{\leqslant k} \setminus (a\Sigma^* \cup \Sigma^* a)$ (as in Lemma 3.7). Let $k \geqslant 2$. We define for all $v \in \Sigma^*$ the block diagonal matrix

$$\mathbf{A}_k(v) = \begin{pmatrix} 2^{|v|} & 0 & 0 & \cdots & 0 \\ 0 & A_{u^{(1)}}(v) & 0 & \cdots & 0 \\ 0 & 0 & A_{u^{(2)}}(v) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & A_{u^{(\#Z)}}(v) \end{pmatrix},$$

where $u^{(1)}, u^{(2)}, \ldots, u^{(\#Z)}$ are all the words of $Z_{a,k}$ in lexicographic order.

Let us give an example of the above definition.

Example 3.27. Let $k = 2$ and $\Sigma = \mathbb{B} = \{a, b\}$. Now $Z_{a,2} = \{b, bb\}$ and, taking $v = baab$,

$$\mathbf{A}_2(v) = \begin{pmatrix} 2^{|v|} & 0 & 0 \\ 0 & A_b(v) & 0 \\ 0 & 0 & A_{bb}(v) \end{pmatrix} = \begin{pmatrix} 2^4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Once again we note that

$$\mathbf{A}_2(v)^2 = \begin{pmatrix} 2^8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} = \mathbf{A}_2(v^2).$$

For any fixed $k \geqslant 1$, the dimension of the matrix $\mathbf{A}_k(v)$ is constant for all words $v \in \Sigma^*$. We denote this common value by $\dim(\mathbf{A}_k)$. This value is straightforward to compute, as the dimension of $A_u(v)$, where $u \in Z_{a,k}$, is $|u| + 1$. Thus, for an m-letter alphabet Σ,

$$\begin{aligned} \dim(\mathbf{A}_k) = 1 + \sum_{u \in Z_{a,k}} |u| + 1 &= 1 + 2(m-1) + (m-1)^2 \sum_{i=0}^{k-2} m^i(i+3) \\ &= (k+1)m^k - (k+2)m^{k-1} + 2. \end{aligned} \tag{3.5}$$

Proposition 3.28. *The mapping $\Sigma^*/\!\sim_k \to \mathbb{N}^{\dim(\mathbf{A}_k)}$, where $[v]_k \mapsto \mathbf{A}_k(v)$, is an injective morphism. In other words, the monoid $\Sigma^*/\!\sim_k$ is isomorphic to the multiplicative matrix monoid generated by $\{\mathbf{A}_k(a) \mid a \in \Sigma\}$.*

Proof. The claim is straightforward to verify when $k = 1$. Indeed, the length of a word together with the number of occurrences of all but one letters determines the 1-abelian equivalence class. Assume thus that $k \geqslant 2$.

To see that this mapping is well-defined, that is, $v \sim_k w$ implies $\mathbf{A}_k(v) = \mathbf{A}_k(w)$, we observe the following. First of all, v and w are of the same length so that the elements in position $[1, 1]$ of the matrices are equal. If $|v| < k$, then $v = w$ and there is nothing to prove. Assume then that $|v| \geqslant k$ so that the length $k - 1$ prefixes (resp., suffixes) of v and w are equal by Lemma 3.7. Furthermore $|v|_u = |w|_u$ for all $u \in Z$. These observations imply that the submatrices $A_u(v)$ and $A_u(w)$, for each $u \in Z$, in the definitions of $\mathbf{A}_k(v)$ and $\mathbf{A}_k(w)$ are equal. It follows that $\mathbf{A}_k(v) = \mathbf{A}_k(w)$.

To see that the mapping is a morphism, that is, $\mathbf{A}_k(v)\mathbf{A}_k(w) = \mathbf{A}_k(vw)$ for all $v, w \in \Sigma^*$, we simply observe that $2^{|v|}2^{|w|} = 2^{|vw|}$ and that the equality of the rest of the elements reduces to Proposition 3.25.

Finally, we show that the mapping is one-to-one. Assume that $\mathbf{A}_k(v) = \mathbf{A}_k(w)$. It immediately follows that $|v| = |w|$ and that $|v|_u = |w|_u$ for all $u \in Z$. For the claim it suffices to show that $p_{k-1}(v) = p_{k-1}(w)$ and $s_{k-1}(v) = s_{k-1}(w)$. Let

$v_i = \mathrm{pref}_i(v)$ for all $i = 1, \ldots, \min\{k-1, |v|\}$. Let $b \in \Sigma \setminus \{a\}$ and observe that $bv_i \in Z$ if and only if v_i does not end with a. Now if $bv_i \in Z$, then element $[|bv_i|+1, 2]$ of the submatrix $A_{bv_i}(w)$ of $\mathbf{A}_k(w)$ equals 1 implying that $\mathrm{pref}_i(w) = v_i$ and, in particular, that the ith letter of v equals the ith letter of w. Otherwise bv_i must end with a and, for all $u \in Z$ of length $i+1$, the element $[i+2, 2]$ of the submatrix $A_u(w)$ of $\mathbf{A}_k(w)$ equals zero. This implies that $\mathrm{pref}_i(w)$ also ends with a. We have thus shown that for each $i = 1, \ldots, \min\{k-1, |v|\}$ the ith letter of v equals the ith letter of w so that $p_{k-1}(v) = p_{k-1}(w)$. Symmetric arguments show that $s_{k-1}(v) = s_{k-1}(w)$ and the claim follows. □

Remark 3.29. Another matrix representation, only of larger dimension, could be constructed as follows. In the definition of $\mathbf{A}_k(v)$, instead of having the blocks $A_u(v)$, with $u \in Z$, on the diagonal, we instead have the blocks $A_u(v)$ for all $u \in \Sigma^k$. In this construction, the length of $|v|$ does not need to be recorded. The dimension of this representation is $(k+1)|\Sigma|^k$, as opposed to $\dim(\mathbf{A}_k)$ in (3.5).

Remark 3.30. The matrix representation obtained here is inspired by the matrix representation of k-binomial equivalence obtained in [93].

There are several equivalence relations related to finite words using matrices. We mention the so-called *Parikh matrices* introduced in [67] by Mateescu et. al. which define the so-called *M-equivalence* of words. This notion has been extensively studied in the literature (see, e.g., [101] and the references therein).

We have reviewed several characterizations of the k-abelian equivalence relation. These characterizations are put to work in the consequent chapters of this thesis.

Chapter 4

Representatives of Equivalence Classes

When considering equivalence classes over finite words, it is often of use to consider certain representatives of equivalence classes. A natural candidate for a representative is the lexicographically least word in the equivalence class. This approach turns out to be extremely fruitful in studying the language theoretic aspects of the k-abelian equivalence, which in turn gives new tools to study quantitative aspects of the equivalence relation. Moreover, these tools are then put into use in the following chapter.

4.1 Lexicographically Least Elements

Let $\lhd$ denote a total order on Σ and the corresponding lexicographic order on Σ^*. We make use of the following language:

$$L_{k,\Sigma,\lhd} = \{w \in \Sigma^* \mid w \trianglelefteq u \text{ for all } u \in [w]_k\}. \tag{4.1}$$

In other words, $L_{k,\Sigma,\lhd}$ is the language of lexicographically least elements (with respect to $\lhd$) of the k-abelian equivalence classes over Σ. We omit k and Σ from the subscript whenever they are clear from context or have no importance (in which case they are assumed to be fixed but arbitrary).

Remark 4.1. Observe that the language $L_{k,\Sigma,\lhd}$ is *factorial* (a language L is said to be factorial if, for every $u \in L$, any factor w of u is in L). Indeed, suppose that $u = u_1u_2u_3 \in L_{k,\Sigma,\lhd}$ with $u_2 \notin L_{k,\Sigma,\lhd}$. We may take u_2' for which $u_2' \sim_k u_2$ and $u_2' \lhd u_2$ (thus $|u_2| \geqslant 2k-1$). By replacing u_2 by u_2' we obtain $u' = u_1u_2'u_3 \lhd u$ and $u' \sim_k u$, since $\sim_k$ is a congruence relation. This is a contradiction, and thus our statement holds.

4.1.1 A Combinatorial Characterization of Lexicographically Least Representatives

We give a combinatorial characterization of lexicographically least elements of k-abelian equivalence classes. This characterization has a graph theoretic flavor, which we discuss later on.

Definition 4.2. Let $u = a_1 \cdots a_n \in \Sigma^*$, where $n \geqslant k$. For each $i \in [1, n-k+1]$ we define the *extension history* $\Delta_u^i \subseteq \Sigma^{k-1} \times \Sigma$ *of* u *at position* i recursively as follows. For $i = 1$ we add, for each $x \in \Sigma^{k-1}$ and $a \in \Sigma$, the pair $(x, a) \in \Delta_u^1$ if and only if the first occurrence of x in u is followed by the letter $a \in \Sigma$. Assume that Δ_u^i is defined, where $i \leqslant n-k$, and that $u[i+1, i+k] = xb$ for some $x \in \Sigma^{k-1}$, $b \in \Sigma$. It follows that $(x, a) \in \Delta_u^i$ for some $a \in \Sigma$. We set $\Delta_u^{i+1} = (\Delta_u^i \setminus \{(x, a)\}) \cup \{(x, b)\}$. In the case of $a \neq b$ we call (x, b) an *update* and say that position $i+1$ *defines* the update (x, b).

Roughly speaking, the extension history Δ_u^i tells us, for each factor of length $k-1$, by which letter it was most recently followed by. The technicalities come in when, in fact, the factor of length $k-1$ in question has not yet occurred. Our main focus is on the updates, and the reader should observe that the first occurrence of a factor (together with the letter following it) do not constitute as an update.

The following definition might seem quite technical, but the reader should compare this to the definition of k-switchings and keep in mind what sort of k-switchings may be performed on lexicographically least representatives. Indeed, our aim is to characterize the words of $L_{k,\Sigma,\lhd}$ by the following notion.

Definition 4.3. A sequence of extension histories $(\Delta_u^i)_{i=1}^t$ of u is called *increasing* if it satisfies the following property: for each $x \in \Sigma^{k-1}$, if

- $(x, a) \in \Delta_u^i$ and x occurs at some position $i' \leqslant i$,
- ℓ defines the update (x, b), and
- there exists $y \in \Sigma^{k-1}$ occurring at positions j and m,

where $i < j \leqslant \ell < m$, then it follows that $a \trianglelefteq b$. Otherwise the sequence is called *non-increasing*.

We illustrate the above definition with an example.

Example 4.4. Let $u = aababba$ and consider $k = 2$. For each $i = 1, \ldots, 6$, the extension history Δ_u^i thus consists of two elements. For $i = 1$, the first occurrence of a is followed by a and the first occurrence of b is followed by a. By the definition above, $\Delta_u^1 = \{(a, a), (b, a)\}$. At position 2 we have a followed by b, so we get an update (a, b); $\Delta_u^2 = \{(a, b), (b, a)\}$. At position 3 we have b followed by a, so that $\Delta_u^3 = \Delta_u^2$. At position 4 we have a followed by b, whence $\Delta_u^4 = \Delta_u^2$. At position 5 we have b followed by b, so an update occurs: $\Delta_u^5 = \{(a, b), (b, b)\}$. Finally, at position 6 we have b followed by a, so that $\Delta_u^6 = \{(a, b), (b, a)\}$. The sequence of extension histories is increasing, since, even though position 6 defines the update (b, a) where the extension of b decreases, the factors of length 1 between position 4 and 6, and factors occurring after position 6 do not intersect.

We obtain a characterization of words $u \in L_{\lhd}$ using extension histories. The proof is almost immediate due to the evident connection to k-switchings. We still give a rigorous proof to illustrate the connection.

Lemma 4.5. *Let $u = a_1 \cdots a_n$, with $n \geqslant k-1$. Then $u \in L_{k,\Sigma,\lhd}$ if and only if $(\Delta_u^i)_{i=1}^{n-k+1}$ is increasing.*

Proof. If $u \notin L_{k,\Sigma,\lhd}$, then there is a k-switching on u which gives a lexicographically smaller element in the k-abelian equivalence class. Thus there exist words $x, y' \in \Sigma^{k-1}$ and indices i', j', ℓ', m' satisfying $i' < j' \leqslant \ell' < m'$ and letters $a \lhd b$ such that xb occurs at i', xa occurs at ℓ' and y' occurs at positions j and m. Let $i = i'$ so that $(x, b) \in \Delta_u^i$ and x occurs at i. Take ℓ to be the leftmost occurrence of xa satisfying $\ell > i$ so that ℓ defines the update (x, a). If y' occurs at j, $i < j \leqslant \ell$, then we may take $y = y'$ and $m = m'$ so that $(\Delta_u^n)_n$ is non-increasing and we are done. If y' does not occur between i and ℓ, then x occurs at position $\ell' > \ell$ and we may take $y = x$, $j = \ell$, and $m = \ell'$ so that $(\Delta_u^n)_n$ is non-increasing.

Assume then that $u \in L_{k,\Sigma,\lhd}$. Assume that $(x, a) \in \Delta_u^i$ and x occurs at some position $i' \leqslant i$, and that ℓ defines the update (x, b). Observe that it follows that there exists an index i', where $i' \leqslant i$, such that x occurs at i', followed by the letter a and x occurs at ℓ, followed by the letter b. If there exists a word $y \in \Sigma^{k-1}$ such that y occurs at some positions j and m with $i < j \leqslant \ell < m$, then a possibility for a k-switching on u arises. Since $u \in L_{k,\Sigma,\lhd}$, then necessarily $a \trianglelefteq b$. It follows that $(\Delta_u^n)_n$ is increasing. □

4.1.2 Lexicographically Least Representatives in the de Bruijn Graph

We describe the lexicographically least representatives as walks in the de Bruijn graphs. To this end we need the following definition.

Definition 4.6. Let G be a graph and W a walk in G. We say that W is *cycle-deterministic* if, for each cycle C occurring along W, the walk enters C at a unique position. The set of distinct cycles (up to the order of the edges) along a cycle-deterministic walk W is denoted by $\mathrm{Cyc}(W)$.

Note that for a cycle-deterministic walk W and a cycle $C \in \mathrm{Cyc}(W)$, some of the vertices and the edges occurring in C may be traversed by W after leaving C. Cycle-determinism means that the full cycle C is not traversed contiguously later on. For a cycle-deterministic walk W, we may write

$$W = P_0 C_1^{\alpha_1} P_1 \cdots C_r^{\alpha_r} P_r \tag{4.2}$$

for some $r \geqslant 0$, where C_i is a cycle and $\alpha_i \geqslant 1$ for each $i = 1, \ldots, r$, P_i is a (possibly empty) path for each $i = 0, \ldots, r$, and W enters C_i at position $1 + |P_0| + \sum_{j<i} \alpha_j |C_j| + |P_j|$ for each i. Observe that W leaves the cycle C_i before (or at the same time as) entering C_{i+1}. Further, $C_i \neq C_j$ for each $i \neq j$. Here $\mathrm{Cyc}(W) = \{C_1, \ldots, C_r\}$. To clarify the notion, we give the following example.

Example 4.7. Consider the de Bruijn graph $dB_{\mathbb{B}}(2)$. The walk $W_u = (e_i)_{i=1}^{10}$, defined by the word $u = aaaabaabaaba$, has two distinct cycles occurring along it, namely the loop $C_1 = (aa, a)$ and the cycle $C_2 = ((aa, b), (ab, a), (ba, a))$. The walk W_u enters C_1 at position 1 and leaves C_1 at position 2 (both via the vertex aa), and does not enter C_1 later on. It does not enter the loop C_1 at another position. Further, W_u enters the cycle C_2 at position 3 (via the vertex aa) and W leaves the cycle via (ba, a) at position 10 (via the vertex ba). We may write $W = C_1^2 \cdot C_2^2 \cdot ((aa, b), (ab, a))$. Note that $u \in L_{3,\mathbb{B},\lhd}$.

On the other hand, the walk W_{uaa} in $dB(k-1)$ defined by the word uaa is not cycle-deterministic, as we may write $W_{uaa} = W_u \cdot ((ba, a), (aa, a)) = C_1^2 \cdot C_2^3 \cdot C_1$,

whence W_{uaa} enters the cycle C_1 at positions 1 and 12. The cycle C_2 is now left from at position 11. Note that now $uaa \notin L_{3,\mathbb{B},\lhd}$.

We show that walks in the de Bruijn graph defined by lexicographically least representatives are cycle-deterministic. We go on further to compute the number of cycles occurring along such a walk. In doing so, we employ the combinatorial characterization of lexicographically least representatives obtained in Lemma 4.5. Note that the elements of the extension histories can be seen as edges in $dB(k-1)$.

Lemma 4.8. *Let $u \in L_{k,\Sigma,\lhd}$ with $|u| \geqslant k-1$. Then the walk W_u in $dB(k-1)$ is cycle-deterministic.*

Proof. Suppose that this is not true for some $u \in L_{k,\Sigma,\lhd}$. We may thus write

$$W_u = W_1 \cdot C \cdot W_2 \cdot C' \cdot W_3,$$

where C is a cycle and C' is the same cycle up to the order of the edges. We assume here that W_u leaves C at position $|W_1C|$ via the vertex $\mathrm{tail}(C) = x \in \Sigma^{k-1}$ and that W_u enters the cycle C at position $|W_1CW_2|+1$ via the vertex $\mathrm{tail}(C') = y \in \Sigma^{k-1}$. Consequently, W_2 is not the empty walk. Let us write $C' = P_1 \cdot P_2$, where P_1 is the path starting from y and ending in x, and P_2 is the path starting with x and ending with y. (If $x = y$, then we set P_1 as the empty walk and $C' = P_2$.) We may write

$$W_u = W_1 \cdot C \cdot W_2 \cdot P_1 \cdot P_2 \cdot W_3.$$

Assume $(x, a) \in C$ and that W_2 begins with (x, b) for some $b \in \Sigma \setminus a$. In particular, $|W_1C| + 1$ defines the update $(x, b) \in \Delta_u^{|W_1C|+1}$ and x occurs at position $|W_1CW_2P_1| + 1$. It follows that $a \lhd b$ by Lemma 4.5.

Consider now the word u' defined by the walk

$$W' = W_1 \cdot C \cdot P_2 \cdot P_1 \cdot W_2 \cdot W_3 = W_1 \cdot C^2 \cdot W_2 \cdot W_3.$$

Note that W' is a valid walk, since $\mathrm{tail}(W_2) = x = \mathrm{tail}(P_2)$ and $\mathrm{head}(W_2) = y = \mathrm{head}(P_2)$. Now $u' \sim_k u$ by Corollary 3.15 but $u' \lhd u$, a contradiction. □

We further describe the connection between extension histories of lexicographically least representatives and the walks they induce in the de Bruijn graph. We determine an upper bound on the number of cycles in $\mathrm{Cyc}(W)$ for a walk W corresponding to a lexicographically least representative. This value has some importance in our future considerations, especially in the following two chapters. We make use of the following technical lemma.

Lemma 4.9. *Let $W = (e_i)_{i=1}^t$ be a walk in $dB(k-1)$ and u the word corresponding to W. Let $r \in [1, t]$ be fixed. Consider the graph $G = (V(W), \Delta_u^r)$. Then, for any $\ell \leqslant r$, there is a unique path from $\mathrm{tail}(e_\ell)$ to $\mathrm{head}(e_r)$ in G.*

Proof. We prove this by induction starting from $\ell = r$. Now if e_r is not a loop, then $(e_r) \in \Delta_u^r$ is a path and is unique by the definition of extension histories. In the case $\mathrm{head}(e_r) = \mathrm{tail}(e_r)$, we are satisfied with the empty path which is unique. Suppose the claim is true for some ℓ and all indices i with $\ell \leqslant i \leqslant r$. Let $x = \mathrm{tail}(e_{\ell-1})$. If $x = \mathrm{tail}(e_i)$ for some $i \in [\ell, r]$, then the claim follows by

the induction hypothesis. Otherwise, since $\ell - 1$ is now the last occurrence of the vertex x along the walk $(e_j)_{j=1}^{r}$, we must have $e_{\ell-1} \in \Delta_u^r$. Thus, there is a unique simple path $e_{\ell-1}$ from x to $\mathrm{tail}(e_\ell)$ in G. By the induction hypothesis, we may extend this path uniquely all the way to $\mathrm{head}(e_r)$. □

Proposition 4.10. *Let $u = a_1 \cdots a_n \in L_\lhd$ with $n \geqslant k-1$ and let $W = (e_i)_{i=1}^{n-k+1}$ be the corresponding walk in $dB(k-1)$. Then, between two distinct consecutive cycles occurring along W, there is an edge which defines an update in the sequence of extension histories of u which has not occurred along W before. In particular,*

$$\#Cyc(W) \leqslant 1 + \left|\bigcup_{i=1}^{n-k} \Delta_u^{i+1} \setminus \Delta_u^i\right|.$$

Here $\bigcup_{i=1}^{n-k} \Delta_u^{i+1} \setminus \Delta_u^i$ is the set of distinct updates in the sequence of extension histories of u.

Proof. Now W is a cycle-deterministic walk, so that we may write $W = P_0C_1P_1 \cdots C_tP_t$, where $t = \#\mathrm{Cyc}(W)$ and W leaves the cycle C_i at position $s_i = |P_0C_1P_1 \cdots C_i|$ for each $i \in [1, t-1]$. It follows that, for each $i \in [1, t-1]$, the position $s_i + 1$ defines an update in the sequence of extension histories. We claim that, for each $i \in [1, t-1]$, the greatest index $\ell \in [s_i + 1, s_i + |P_i| + 1]$ defining an update, defines an update that has not previously occurred along W.

Suppose, to the contrary, that this is not the case, for some $i \in [1, t-1]$. Let ℓ be the greatest position in $[s_i + 1, s_i + |P_i| + 1]$ which defines an update. By our assumption, $e_\ell = e_j = (x, a)$ for some $j < \ell$, $x \in \Sigma^{k-1}$, and $a \in \Sigma$. Note that there exists a position r, where $j < r < \ell$, defining the update $e_r = (x, b)$, for some $b \neq a$ (otherwise ℓ would not define an update). By Lemma 4.5 we necessarily have $b \rhd a$. We claim that x occurs at some position $m > \ell$. This is a contradiction by Lemma 4.5.

If $\mathrm{head}(e_\ell) = \mathrm{tail}(e_\ell)$, then x occurs at position $m = \ell + 1$. Suppose this is not the case. Observe first that

$$\Delta_u^\ell \cap \Delta_u^{\ell-1} = \Delta_u^\ell \setminus \{e_\ell\}.$$

Let now $G = (V(W), \Delta_u^{\ell-1})$. By Lemma 4.9, there exists a path P from $\mathrm{head}(e_\ell) = \mathrm{tail}(e_{j+1})$ to $\mathrm{head}(e_{\ell-1}) = x$ using only edges from $\Delta_u^{\ell-1}$ (note that $\ell - 1 \geqslant j + 1$). Since we do not use the edge e_ℓ along such a path, we are using only edges from Δ_u^ℓ. Since W enters C_{i+1} at some position $\ell' \geqslant \ell$, we continue using edges from Δ_u^ℓ until C_{i+1} occurs along W. Thus W continues by $e_\ell \cdot P$ which is a cycle having $\mathrm{head}(e_\ell P) = x$, whence x occurs at $m = \ell + |P| + 2$ in u. This concludes the proof. □

Corollary 4.11. *Let $u \in L_{k,\Sigma,\lhd}$ and let W be the corresponding walk in $dB(k-1)$. Then $\#Cyc(W) \leqslant m^{k-1}(m-1) + 1$.*

Proof. For each factor x of length $k-1$ occurring in u there can be at most $(m-1)$ updates, since the pairs $(x, a) \in \Delta_u^1$ are not counted as updates. There are m^{k-1} such factors x. The claim now follows from the above proposition. □

We make a further observation regarding the cycles along a walk defined by $u \in L_\lhd$. If some cycle occurring along u occurs at least twice, then we may construct a whole family of words in $L_\lhd$.

Lemma 4.12. *Let C be a cycle and let W, W' be walks in $dB(k-1)$ with* $\mathrm{head}(W) = \mathrm{tail}(W') = \mathrm{tail}(C)$. *For each r let $u^{(r)} \in \Sigma^*$ be defined by the walk $W(r) = W \cdot C^r \cdot W'$. Then $u^{(r)} \in L_{k,\Sigma,\lhd}$ for all $r \geqslant 0$ if and only if $u^{(2)} \in L_{k,\Sigma,\lhd}$.*

Proof. Notice that the "left to right" direction is trivial: if the claim is true for all values, then it is true for some value. Let thus $u = u^{(2)} \in L_{k,\Sigma,\lhd}$ and let us first show that $u^{(r)} \in L_{k,\Sigma,\lhd}$, for $r \leqslant 1$. Let $(x,a) \in \Delta^i_{u^{(0)}}$, and assume that x occurs at some position prior to i. In fact, we may assume that xa occurs at position i. Assume further that ℓ defines the update (x,b) and that there exists $y \in \Sigma^{k-1}$ which occurs at position j and m, where $i < j \leqslant \ell < m$. It is clear that similar occurrences of xa, xb, and y occur in u (e.g., if $i \leqslant |W|$ then xa occurs at position i in u, and if $i > |WC^r|$, then xa occurs at position $i + |C^{2-r}|$ in u). It follows that $a \trianglelefteqslant b$, since the sequence $(\Delta^i_u)_i$ of extension histories of u is increasing by Lemma 4.5.

We then focus on the case $r \geqslant 2$. Now the sequence of extension histories of $u^{(r)}$ is defined as follows:

$$\Delta^i_{u^{(r)}} = \begin{cases} \Delta^i_u & \text{if } i \leqslant |WC| \\ \Delta^{|WC|}_u & \text{if } |WC| < i < |WC^r| \\ \Delta^{i-(r-2)|C|}_u & \text{if } i \geqslant |WC^r| \end{cases}.$$

Assume that $(x,a) \in \Delta^i_{u^{(r)}}$ and x occurs in $u^{(r)}$ at some position $i' \leqslant i$, and that ℓ defines the update (x,b) for some $i < \ell$. Assume further that there exists $y \in \Sigma^{k-1}$ occurring at positions j and m with $i < j \leqslant \ell < m$. We may assume that either i defines the update (x,a) or that $(x,a) \in \Delta^1_{u^{(r)}}$ and x occurs for the first time at i, since such an index (of magnitude at most i) exists. Since no index n in the interval $[|WC|, |WC^r|)$ defines an update, we see that $i, \ell \notin [|WC|, |W|C^r)$. Observe now that if none of the indices i, j, ℓ, or m occur in the interval $[|WC|, |WC^r|)$, then we may argue as in the case $r \leqslant 1$ to show that $a \trianglelefteqslant b$.

Thus we may consider the case $i < |WC|$, as otherwise none of the indices occur in the interval $[|WC|, |WC^r|)$. Now xa occurs at position i in u also. Assume first that $\ell < |WC|$. Then xb occurs at position ℓ, and y occurs at position j in u as well. Now m is in the interval $[|WC|, |WC^r|)$ and hence y is a vertex of C. There exists an index m' such that $|WC| < m' \leqslant |WC^2|$ and y occurs in u at position m'. Again, $a \trianglelefteqslant b$ since the sequence of extension histories of u is increasing. Assume then that $\ell \geqslant |WC^r|$. Now xb occurs at position $\ell - (r-2)|C|$ and y at position $m - (r-2)|C|$ in u as well. Now j is in the interval $[|WC|, |WC^r|)$ so that y is a vertex of C. Let j' be an index such that $|WC| \leqslant j' < |WC^2|$ and y occurs at j' in u. Again it follows that $a \trianglelefteqslant b$.

We have shown that the sequence of extension histories of $u^{(r)}$ is increasing, that is, $u^{(r)} \in L_{k,\Sigma,\lhd}$ by Lemma 4.5. □

The requirement that the cycle C is repeated twice in the above lemma is necessary. Indeed, consider the following example:

Example 4.13. Let $u = aabbcaacbbab$ with $k = 3$ and $a \lhd b \lhd c$. Now we may write $W_u = CW$, where C is the cycle $((aa,b),(ab,b),(bb,c),(bc,a),(ca,a))$ and $W = ((aa,c),(ac,b),(cb,b),(bb,a),(ba,b))$. Now the sequence of extension histories of u is increasing, since, even though $(bb,c) \in \Delta^3_u$ and $(bb,a) \in \Delta^9_u$, the factors ba

and ab, which are the only factors of length 2 occurring after position 9, do not occur at any position in the interval $[3, 9]$. Thus $u \in L_{k,\Sigma,\lhd}$. But now the sequence of extension histories of the word $aabbc \cdot u = u^{(2)}$ corresponding to the walk C^2W is not increasing. Indeed, ab occurs at positions 7 and 16, while $(bb, c) \in \Delta^3_{u^{(2)}}$ and $(bb, a) \in \Delta^{14}_{u^{(2)}}$. Thus $u^{(2)}$ is not in $L_{k,\Sigma,\lhd}$.

4.2 k-abelian Singletons

We also consider properties of particular k-abelian equivalence classes, namely, singleton classes. Let us formally define words defining such classes.

Definition 4.14. A word $w \in \Sigma^*$ is called a *k-abelian singleton*, if the equivalence class represented by w is a singleton set. We define the language $L_{k,\Sigma,\text{sing}}$ as the language of k-abelian singletons over alphabet Σ:

$$L_{k,\Sigma,\text{sing}} = \{w \in \Sigma^* : |[w]_k| = 1\}.$$

Observe that the language $L_{k,\Sigma,\text{sing}}$ is a subset of the language $L_{k,\Sigma,\lhd}$. Furthermore each word of length at most $2k-1$ is a k-abelian singleton by Lemma 3.3.

Remark 4.15. Observe that the language $L_{k,\Sigma,\text{sing}}$ is also factorial (recall Remark 4.1). This may be seen with a similar proof as in the case of $L_{k,\Sigma,\lhd}$.

Example 4.16. It is not difficult to verify that the set of 2-abelian singletons over $\mathbb{B} = \{a, b\}$ beginning with a is the (regular) language

$$a^+b^* + ab^*a + (ab)^*\{\varepsilon, a\}.$$

As the number of singleton classes beginning with b are the same up to switching a's with b's, the total number of 2-abelian singleton classes of length n over a binary alphabet is $2n + 4$ for $n \geqslant 4$.

Similar to the previous section, we give a combinatorial characterization of k-abelian singletons, and then describe the walks they define in the de Bruijn graphs.

4.2.1 A Combinatorial Characterization of k-abelian Singletons

We characterize k-abelian singletons in terms of generalized return words using k-switchings. For this we say that x is a *proper factor* of w if x occurs in $w[1, |w|-1)$.

Definition 4.17. Let $u \in \Sigma^*$ and let $x, y \in \Sigma^+$ be of the same length. A *return from x to y in u* is a word $v \in \Sigma^+$ such that vy is a factor of u, x is a prefix of vy and x and y do not occur as proper factors of vy. Recall that if $x = y$, then vy is simply a complete first return to x in w.

Note that if v and v', $|v| \leqslant |v'|$, are distinct returns from x to y in a word w, then vy cannot be a factor of $v'y$, as otherwise $v'y$ would contain either x or y as a proper factor. The following characterization appears in the article [55].

Lemma 4.18. *A word $w \in \Sigma^*$ is a k-abelian singleton if and only if for each pair $x, y \in F_{k-1}(w)$ there is at most one return from x to y in w.*

Proof. Let $w \in \Sigma^*$ and suppose it contains two distinct returns v and v' from x to y. Let $vy = w[i,j)y$ and $v'y = w[\ell, m)y$ with $i < \ell$. Note that $j < m$ as otherwise vy contains $v'y$ as a factor. In fact, by definition, we necessarily have $i < j$ and $\ell < m$ (since $v, v' \in \Sigma^+$) and $j \leqslant \ell$, (since otherwise vy contains x as a proper factor). But now

$$w \sim_k w' = w[1,i)w[\ell,m)w[j,\ell)w[i,j)w[m..] \neq w$$

since w begins with $w[1,i)vy$ and w' with $w[1,i)v'y$. It follows that w cannot be a k-abelian singleton.

Suppose then that for each pair $x, y \in F_{k-1}(w)$ there is at most one return from x to y in w. Suppose, for the sake of contradiction, that $w' \sim_k w$ with $w' \neq w$. Let $w = \lambda x a \mu$ and $w' = \lambda x b \mu'$ with $a, b \in \Sigma$, $a \neq b$, $\lambda \in \Sigma^*$ and $x \in \Sigma^{k-1}$. Since $w' \sim_k w$, we have that xa occurs in $w'[|\lambda| + 1..]$ and xb occurs in $w[|\lambda| + 1..]$. Let $\ell > |\lambda| + 1$ be a position of xb in w. Now we must have $y \in \Sigma^{k-1}$ which occurs both in $w[|\lambda x a|, \ell + k - 1)$ and $w[\ell + 1..]$, say at positions j and m, respectively (compare to the proof of Theorem 3.12). We can assume that j and m are minimal.

Now there exists a return to x in w which begins with xa. Therefore, x cannot occur in $w[\ell + 1..]$, as otherwise w contains also a return to x which begins with xb. It follows that $w[\ell, m)y$ is a return from x to y which begins with xb. Now $w[i,j)y$ cannot be a return from x to y, since it begins with xa. We thus have a position of x in between i and j and if we take it to be maximal, we obtain a return from x to y; it has to begin with xb. But this is a final contradiction, since w has now also a return to x which begins with xb. □

4.2.2 k-abelian Singletons as Cycle Decompositions

We now interpret k-abelian singletons as walks in the de Bruijn graph of order $k-1$. Observe that, since k-abelian singletons are elements of $L_{k,\Sigma,\lhd}$, we immediately have that a walk W_u in $dB(k-1)$ is cycle-deterministic by Lemma 4.8. The results in this subsection are based on results in the article [55], though the approach is different. We apply the approach using notions from the previous section.

Lemma 4.19. *Let $u \in L_{k,sing}$ and let W_u be the corresponding walk in $dB(k-1)$. If a vertex x occurs in a cycle C occurring along W_u, then x does not occur along W_u before entering C or after leaving C. Furthermore, if $x \in \Sigma^{k-1}$ occurs at least three times in u, then x occurs (as a vertex) in a cycle occurring along W_u.*

Proof. Let C be a cycle occurring along W_u and let $x \in V(C)$. Let us write

$$W_u = W_1 \cdot C \cdot W_2 \cdot W_3 = W_1 \cdot W_2' \cdot C' W_3,$$

where C' is a shifted version of C, W_u enters C at position $|W_1| + 1$ and leaves C at position $|W_1 W_2' C'|$. Assume for a contradiction that we may write $W_3 = VV'$ for some walks V, V' such that V is non-empty, $\mathrm{head}(V) = x$, and x is not an internal vertex of V. Let us write $C' = P_1 P_2$, where $\mathrm{head}(P_1) = x$ and P_1 is non-empty. Now P_1 corresponds to a return from y to x in u. If y is not an internal vertex of V, then V also corresponds to a return from y to x in u. This is

a contradiction, since P_1 and V begin with distinct edges, implying that the two returns mentioned above are distinct. Now if y is an internal vertex of V, then we may write $V = P_1P_2$, where P_1 is non-empty with $\text{head}(P_1) = y$, and y is not an internal vertex of P_1. Now C' and P_1 correspond to distinct returns to y in u, a contradiction. Thus x does not occur in W_3 (except possibly as the first vertex).

Similarly, assume that x occurs in W_1 with $W_1 = V'V$, V non-empty, $\text{tail}(V) = x$ and x is not an internal vertex of V. Let $\text{tail}(C) = y$. Write $C = P_1P_2$ where P_2 is non-empty and $\text{tail}(P_2) = x$. Then P_2 corresponds to a return from x to y in u. If y does not occur in V, then V corresponds to a different return from x to y in u, a contradiction. If y occurs in V as an internal vertex, then we may write $V = S_1S_2$ with S_2 non-empty and $\text{tail}(S_2) = y$. Hence S_2 corresponds to a return to y and the last edges of C and S_2 are distinct, implying that the returns are distinct. This is again a contradiction.

Assume now that x occurs along W_u at least three times. Factorize $W_u = W_1W_2W_3W_4$, where W_2 and W_3 are non-empty walks and $x = \text{head}(W_1) = \text{head}(W_2) = \text{head}(W_3)$. Assume further that x is not an internal vertex of neither W_2 nor W_3. Since W_2 and W_3 correspond to returns to x in u, they must be equal; let $W = W_2 = W_3$. Assume that W is not cycle. Thus there exists some internal vertex y of W which occurs twice. Then we may write $W = V_1V_2V_3V_4$ where $\text{head}(V_1) = \text{tail}(V_2) = \text{head}(V_2) = \text{tail}(V_3) = y$ and y is not an internal vertex of neither V_1, V_3, nor V_4. We assume that V_3 is non-empty, but V_2 is allowed to be empty. Now V_3 corresponds to a return to y in u which does not contain x as a factor. On the other hand, V_4V_1 corresponds to a return to y in u which contains x as a factor, a contradiction. □

Corollary 4.20. *Let $u \in L_{k,\Sigma,sing}$ and let W_u be the walk in $dB(k-1)$ corresponding to u. Then the cycles in $Cyc(W)$ are vertex-disjoint.*

The previous corollary allows us to consider k-abelian singletons as certain kinds of walks in cycle-semi-decompositions of the de Bruijn graphs. We formalize this concept now. Let $G = (V, E)$ be a graph and let $\mathcal{C} = \{C_1, \ldots, C_m\}$ be a set of vertex-disjoint cycles of G. Let $V_i = V(C_i)$ for each $i = 1, \ldots, m$, and let $V_\otimes = V \setminus \bigcup_{i=1}^{m} V_i$. The set consisting of the partitions V_i, $i = 1, \ldots, m$ and $\{v\}, v \in V_\otimes$, is called a *cycle-semi-decomposition* of G, denoted by V/C.

Definition 4.21. Let $G = (V, E)$ and let $\mathcal{C}$ be a set of vertex disjoint cycles of G. We define the quotient graph $G/\mathcal{C} = (V/\mathcal{C}, E')$ with respect to $\mathcal{C}$ as follows. For $X, Y \in V/C$, $X \neq Y$, we have $(X, Y) \in E'$ if and only if there exist $x \in X$ and $y \in Y$ such that $(x, y) \in E$.

Remark 4.22. Let now u be a k-abelian singleton and consider the graph G_u obtained from $G_{\Psi_k(u)}$ by removing multiplicities of edges. The graph $G_u/\text{Cyc}(W_u)$ then contains a walk which traverses through each vertex $V(C)$, where $C \in \text{Cyc}(W_u)$, once and through each $\{v\}$, where $v \in V_\otimes$, at least once and at most twice.

Example 4.23. Consider the graph G_u induced by the 4-abelian singleton $u = 2(01)^3 2(0110)^3 01^4$. The vertex-disjoint cycles corresponding to u are defined by the sets of vertices $V_1 = \{010, 101\}$, $V_2 = \{011, 110, 100, 001\}$ and $V_3 = \{111\}$. The set of factors occurring at most twice in u is $V_\otimes = \{201, 012, 120\}$. The quotient graph G_u/C_u is displayed in Figure 4.1.

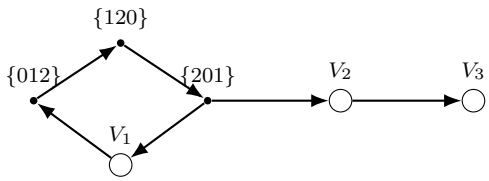

Figure 4.1: The quotient graph $G_u/\mathrm{Cyc}(W_u)$ induced by the 4-abelian singleton $u = 2(01)^3 2(0110)^3 01^4$. The cycles defining V_1, V_2, and V_3 are $\{(010,1),(101,0)\}$, $\{(011,0),(110,0),(100,1),(001,1)\}$, and $\{(111,1)\}$, respectively. Modification of [55, Fig 3].

We give an upper bound on the number of cycles in $\mathrm{Cyc}(W_u)$ for any $u \in L_{\mathrm{sing}}$ and the corresponding walk W_u in the de Bruijn graph. Recall that the number of necklaces of length ℓ over an m-letter alphabet is denoted by $N_m(\ell)$ and that the formula Equation 2.1 is for counting these values. We also recall a related result from the literature, namely, the following theorem, originally conjectured by Lempel and later proved to be true by J. Mykkeltveit (see [73] and references therein).

Theorem 4.24 ([73])**.** *The minimum number of vertices which, if removed from $dB_\Sigma(n)$, will leave a graph with no cycles, is $N_{|\Sigma|}(n)$ (defined by (2.1)).*

The above theorem implies that the size of any set $\mathcal{C}$ of vertex disjoint cycles in the de Bruijn graph $dB_\Sigma(n)$ is at most $N_{|\Sigma|}(n)$. Indeed, if $|\mathcal{C}|$ were larger than $N_{|\Sigma|}(n)$, then removing any set of $N_{|\Sigma|}(n)$ vertices from the graph $dB(n)$ would leave some cycle of $\mathcal{C}$ in $dB(n)$, contradicting the above theorem. Since the cycles in $\mathrm{Cyc}(W_u)$, where u is a k-abelian singleton and W_u is the corresponding walk in $dB(k-1)$, are vertex disjoint by Corollary 4.20, we have the following proposition.

Proposition 4.25. *Let $u \in L_{k,\Sigma,sing}$ and let W_u be the corresponding walk in $dB_\Sigma(k-1)$. Then $\#Cyc(W_u) \leqslant N_m(k-1)$.*

We may prove a similar result to Lemma 4.12 for singletons.

Lemma 4.26. *Let C be a cycle and W and W' be walks in $dB(k-1)$. For each $r \geqslant 0$ let $u^{(r)}$ be defined by the walk $W(r) = WC^rW'$. Then we have $u^{(r)} \in L_{k,\Sigma,sing}$ for all $r \geqslant 0$ if and only if $u^{(1)} \in L_{k,\Sigma,sing}$.*

Proof. Without loss of generality we may assume that $W(r)$ enters C at position $|W|_1$ and leaves C at position $|WC^{(r)}|+t$ for some $t < |C|$. Let us write $C = P_1P_2$, where $|P_1| = t$, whence $WC^{(r)}W' = WC^rP_1W''$, and the cycle C is left at position $|WC^rP_1|$.

Again the "left to right" direction is clear. Assume thus that $u = u^{(1)} \in L_{k,\Sigma,\mathrm{sing}}$ and let r be arbitrary. We show that for each pair $x, y \in \Sigma^{k-1}$, there is at most one return from x to y in $u^{(r)}$.

Assume the contrary, namely that there exist x, $y \in \Sigma^{k-1}$ and two distinct returns from x to y in $u^{(r)}$. These returns correspond to subwalks R_1 and R_2 of $W(r)$ starting from x and ending in y and x and y occur only as external vertices. It follows that we may write $u = T_0R_1T_1R_2T_2$. If $x \in V(C)$ then, by Lemma 4.19, x does not occur in W or W'' except possibly as the last (resp., first) vertex. The

same holds for y. We conclude that $x \in V(C)$ if and only if $y \in V(C)$. If $x \in V(C)$ then the only return from x to y follows the cycle C, and there is only one such return. Assume thus that $x \notin V(C)$. Now if R_1 and R_2 both occur along either W or W'', then the corresponding returns occur in $u \in L_{k,\Sigma,\mathrm{sing}}$, which is not possible. We are left with the case that either R_1 occurs along W or R_2 occurs along W'' (both cannot contain parts of $C^r P_1$). Assume the former, the latter leads to a similar contradiction. Now R_2 is of the form $VC^r P_1 V'$. Thus R_1 and VCP_1V' correspond to distinct returns in u, which is again a contradiction. This concludes the proof. □

4.3 Representatives of Classes of Fixed Size

In the previous section we considered k-abelian singleton classes. In this section we describe a slightly more general set of k-abelian classes. These classes were considered in the work [59], where the results of this section appear.

Definition 4.27. Let $k \geqslant 1$ and $r \geqslant 1$. Define the language $L_{k,\Sigma,r} = \{w \in \Sigma^* \colon |[w]_k| = r\}$. In other words, $L_{k,\Sigma,r}$ consists of the words which represent k-abelian equivalence classes of cardinality r.

We consider these languages in a language theoretic setting. We remark that, for $r = 1$, the language $L_{k,\Sigma,r}$ coincides with the language $L_{k,\Sigma,\mathrm{sing}}$. For $r \geqslant 2$, we observe that $L_{k,\Sigma,r}$ is not a subset of $L_{k,\Sigma,\lhd}$. Let us make a brief observation on words in $L_{k,\Sigma,r}$.

Lemma 4.28. *Let $k \geqslant 1$ and $r \geqslant 2$. There exists an integer $\mathcal{B}_{k,r}$ such that, for each $u \in L_{k,\Sigma,r}$, if $\#\Re_u(x) \geqslant 2$ for some $x \in F_{k-1}(u)$ then $|u|_x \leqslant \mathcal{B}_{k,r}$.*

Proof. Let $u \in L_{k,\Sigma,r}$ and assume $\#\Re_u(x) \geqslant 2$ for some $x \in F_{k-1}(u)$. Let us write W_u in terms of complete first returns of x in u;

$$W_u = W_0 W_1 \cdots W_{|u|_x - 1} W_{|u|_x},$$

where $\mathrm{tail}(W_i) = \mathrm{head}(W_i) = x$ for all $i = 1, \ldots, |u|_x - 1$. Observe now that each walk W_i, where $i = 1, \ldots, |u|_x - 1$, corresponds to complete first return to x in u, and thus x is not an internal vertex of any of the walks W_i, where $i = 0, \ldots, |u|_x$. Furthermore, W_0 and $W_{|u|_x}$ contain the vertex x only as the last and first vertex, respectively. Now, for any permutation σ of $[1, |u|_x)$, we have that $u \sim_k v_\sigma$, where v_σ is defined by the walk

$$W_\sigma = W_0 W_{\sigma(1)} \cdots W_{\sigma(|u|_x - 1)} W_{|u|_x}$$

(by Corollary 3.15). The number of distinct words obtained by this method is $\binom{|u|_x - 1}{m_1, \ldots, m_L} = \frac{(|u|_x - 1)!}{m_1! \cdots m_L!}$, where $L = \#\Re_u(x)$ and $(m_i)_i = (|u|_y)_{y \in \Re_u(x)}$. This is the number of distinct permutations of words $y \in \Re_u(x)$ with multiplicities $|u|_y$, where $y \in \Re_u(x)$. It is clear that all these words are distinct by the definition of complete first returns. We now have, by assumption, $L \geqslant 2$ whence $\binom{|u|_x - 1}{m_1, \ldots, m_L} \geqslant |u|_x - 1$. This implies $r = \#[u]_k \geqslant |u|_x - 1$, or in other words, $|u|_x \leqslant r + 1$. We may take $\mathcal{B}_{k,r} = r + 1$. □

Example 4.29. In the case $r = 1$ and $u \in L_{k,\Sigma,1}$, we have $\#\Re_u(x) \leqslant 1$ for all $x \in \Sigma^{k-1}$ by Lemma 4.18. Thus we may set $\mathcal{B}_{k,1} = 0$ by convention. On the other hand, for $r \geqslant 2$, we have $\mathcal{B}_{k,r} \geqslant 2$. Indeed, the word $u = a^{k+r-2}ba^{k-1}$ defines the k-abelian equivalence class

$$[u]_k = \{a^{k-1+t}ba^{k+r-2-t} \colon 0 \leqslant t \leqslant r-1\}$$

as is straightforward to verify. Thus $u \in L_{k,\Sigma,r}$. We also note that $\#\Re_u(a^{k-1}) = 2$.

Definition 4.30. Let $k \geqslant 1$ and $L \subseteq \Sigma^*$. Let further $\lhd$ be a lexicographic ordering on Σ^*. We define the language $\lhd\text{-Min}_k(L)$ as

$$\lhd\text{-Min}_k(L) = \{u \in L \colon u \trianglelefteqslant v \text{ for all } v \in L \cap [u]_k\}.$$

In other words, $\lhd\text{-Min}_k(L)$ is the language of elements of L which are lexicographically smaller than all other words in the same k-abelian equivalence class. We may say that $\lhd\text{-Min}_k(L)$ is the language of *lexicographically least representatives with respect to* $\lhd$ *and* L. We omit the prefix $\lhd$ whenever it is not of importance. In this case it should be considered to be fixed, but arbitrary.

Observe now that $\text{Min}_k(L_{k,\Sigma,r}) = L_{k,\Sigma,\lhd} \cap L_{k,\Sigma,r}$. This, of course, does not hold for general languages L; we shall consider the language $\text{Min}_k(L)$ for regular languages L in the following chapter. Here we only make the following remark.

Lemma 4.31. *Let $k \geqslant 2$ and let $u \in Min_k(L_{r,k})$ and assume $|u|_x > \mathcal{B}_{k,r}$. Then we may write $W_u = W \cdot C^{|u|_x - 1} \cdot W'$ for some cycle C with $\text{tail}(C) = x$.*

Proof. Assume that $|u|_x > \mathcal{B}_{k,r}$ for some $x \in \Sigma^{k-1}$. Since $|u|_x > \mathcal{B}_{k,r}$, we have $\Re_u(x) = 1$ by the above lemma. Let y be the single element in $\Re_u(x)$. We may write $W_u = W \cdot W_y^{|u|_x - 1} \cdot W'$, where $\text{head}(W) = \text{tail}(W_y) = \text{head}(W_y) = \text{tail}(W') = x$ and W_y is the walk in $dB(k-1)$ corresponding to y. We claim that W_y is a cycle. Assume the converse; then there exists a vertex z in W_y such that there are two distinct edges both with tail z along W_u. As in the proof of Lemma 4.19, we deduce that $\#\Re_u(z) \geqslant 2$. Furthermore, $|u|_z \geqslant 2(|u|_x - 1) > B_{k,r}$ which is contrary to the above lemma. □

We conclude this chapter with yet again a result similar to Lemma 4.12 for words in $\text{Min}_k(L_{k,\Sigma,r})$.

Lemma 4.32. *Let $r \geqslant 2$. Let $u^{(s)}$, for each $s \geqslant 0$, denote the word defined by the walk $W(s) = W \cdot C^s \cdot W'$ (in $dB(k-1)$) for some cycle C and walks W and W'. Then $u^{(s)} \in Min_k(L_{k,\Sigma,r})$ for some $s \geqslant \mathcal{B}_{k,r}$ if and only if $u^{(s)} \in Min_k(L_{k,\Sigma,r})$ for all $s \in \mathbb{N}$.*

Proof. The other implication is immediate, so assume $u^{(s)} \in \text{Min}_k(L_{k,\Sigma,r})$ for some $s \geqslant \mathcal{B}_{k,r}$. The fact that $u^{(s)} \in L_{k,\Sigma,\lhd}$ for all $s \in \mathbb{N}$ follows by Lemma 4.12, so it is enough to show that $u^{(s)} \in L_{k,\Sigma,r}$ for all $s \in \mathbb{N}$ (recall that for $r \geqslant 2$, we have $\mathcal{B}_{k,r} \geqslant 2$ by Example 4.29). Without loss of generality we may assume that $W(s)$ enters $C = (e_i)_{i=0}^{\ell-1}$ ($|C| = \ell \geqslant 1$) at position $|W| + 1$ and that s is maximal, i.e., $W(s)$ leaves C before position $|W| + (s+1)|C|$ and, further, that $W(s)$ leaves C via vertex $y = \text{tail}(e_o)$, where $0 \leqslant o \leqslant \ell - 1$.

Observe now that $|u^{(s)}|_x \geqslant s$ for all $x \in V(C)$. It follows by the above lemma that for each vertex $x \in V(C)$ we have $\#\Re_{u^{(s)}}(x) = 1$ (if there were another complete first return to x in u for some $x \in V(C)$, we would have $|u^{(s)}|_x > \mathcal{B}_{k,r}$, a contradiction). Consequently, by the maximality of s, $|u^{(s)}|_{\mathrm{tail}(e_i)} = s+1$ for all $i \in [0, o]$, and $|u^{(s)}|_{\mathrm{tail}(e_i)} = s$ for all $i \in (o, \ell)$. Further, each (except possibly the last) occurrence of x is followed by the same letter a_x in $u^{(s)}$. Moreover, the only vertex $y \in V(C)$ followed by a letter $b \neq a_y$ in u is the vertex $y = \mathrm{tail}(e_o)$ via which $W(s)$ leaves C from (the only exception is that W' is a subpath of C).

Consider the graph $G_s = G_{u^{(s)}}$ in light of Proposition 3.20. Let κ_s denote $\kappa_{\mathrm{head}(W(s))}$ for each $s \geqslant 0$. By the above observations we conclude that any rooted spanning tree with root $\mathrm{head}(W(s))$ of G_s contains one of the (multiple copies of the) edge e_i for each $i = [0, \ell) \setminus \{o\}$, and the edge $(y, b) \notin E(C)$ (unless $\mathrm{head}(W(s)) = y$ whence no edge from y exists in such a tree). Let us compute κ_s in terms of κ_0 and s. Adding s copies to an edge e_i, where $0 \leqslant i < o$, to G_0 increases the number of trees $(s+1)$-fold, as each tree must contain exactly one copy of this edge and there are $s+1$ to choose from. For the remainder of the vertices $z \in V(C) \setminus y$, any tree in G_s must contain some copy of the path $(e_j)_{j=o+1}^{\ell-1}$ which connects to a copy of a tree defined by G_0. Given s copies of each edge along this path, there are altogether $s^{\ell-o-1}$ choices for the path. We conclude that $\kappa_s = \kappa_0 \cdot (s+1)^o s^{\ell-o-1}$. This may be expressed as $\kappa_s = \kappa_0 \cdot \prod_{x \in V(C)\setminus y}(s + |u^{(0)}|_x)$. Further, we observe that, in the product

$$\prod_{x \in F_{k-1}(u^{(s)})} \frac{(|u^{(s)}|_x - 1)!}{\prod_{a \in \Sigma} |u^{(s)}|_{|xa|}!},$$

the only values that vary according to s are certain values corresponding to the vertices of C. In particular, $|u^{(s)}|_x = |u^{(s)}|_{xa_x} = s + |u^{(s)}|_{x_0} \in \{s, s+1\}$ for each $x \in V(C) \setminus y$, $|u^{(s)}|_y = s+1$, and $|u^{(s)}|_{ya_y} = s$. Recall that $|u^{(s)}|_{yb} = 0$ or 1 depending on whether $\mathrm{head}(W(s)) = y$ or not. Plugging these values in (3.2), we find that the following ratio equals 1 for any $s \geqslant 1$:

$$\frac{\#[u^{(s)}]_k}{\#[u^{(0)}]_k} = \prod_{x \in V(C)\setminus y} (s + |u^{(0)}|_x) \cdot \prod_{x \in V(C)\setminus y} \frac{1}{(s + |u^{(0)}|_x)} = 1.$$

Thus, for any choice of s, the obtained word $u^{(s)}$ has $\#[u^{(s)}] = r$. The claim follows. □

Remark 4.33. We close this chapter with some concluding remarks.

The aim of this chapter has been to explore the structure of minimal representatives of k-abelian equivalence classes. The considerations here are quite technical and involved, but all have graph theoretical flavour due to the main tool used—the de Bruijn graphs. This proves to be of particular use when considering how the factors of length k are situated in such words. Of course, it is the factors of length k together with the prefixes and suffixes of length $k-1$ which determine the k-abelian equivalence class. The involved work done in this chapter allow us to translate certain properties into finite automata, as the graph structure of automata permits certain interpretations using the de Bruijn graph. This is precisely what we do in the following two chapters.

Chapter 5

Automata Theoretic Aspects of k-Abelian Equivalence

We begin the exposition to language theoretic aspects of k-abelian equivalence by noting that the languages $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\text{sing}}$ are regular. This is a straightforward observation from the characterization of k-abelian equivalence by k-switchings and especially Claim 3.13. We also give another proof of the regularity of these languages in the first section of this chapter. This proof is based on the work done in the previous chapter.

In the second section we develop a k-switching operation on languages. We show that regular languages are closed under this operation. In the third section we use this tool to show that the languages $L_{k,\Sigma,r}$ are regular for any $r \geqslant 1$. We conclude this chapter with some language theoretic observations related to the concepts introduced in this chapter.

5.1 The Languages $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\textbf{sing}}$ are Regular

This section consists of two separate proofs (each) of the facts that $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\text{sing}}$ are regular. The first proofs appear in [19]. The second proofs are inspired by the results of [59]. This section contains work mainly from the article [19].

We start by giving a straightforward proof of the fact that $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\text{sing}}$ are regular.

Theorem 5.1 ([19]). *The languages $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,sing}$ are regular for any integer $k \geqslant 1$, any alphabet Σ, and any lexicographic ordering $\lhd$.*

Proof. Let u be the lexicographically least element of $[u]_k$ with respect to $\lhd$. If there exists a k-switching on u which yields a new element, it has to be lexicographically greater than u. In particular, u does not contain factors from the language

$$\big((xb\Sigma^* \cap \Sigma^* y)\,\Sigma^* \cap \Sigma^* x\big) a\Sigma^* \cap \Sigma^* y,$$

for each pair $x, y \in \Sigma^{k-1}$ and each pair of letters $a, b \in \Sigma$ with $a \lhd b$. On the other hand, any word u avoiding such factors is lexicographically least in the k-abelian

equivalence class it represents. Indeed, this follows from Claim 3.13. We thus have

$$L_{k,\Sigma,\lhd} = \bigcap_{\substack{x,y\in\Sigma^{k-1}\\ a,b\in\Sigma,\ a\lhd b}} \overline{\Sigma^*\Big(\big((xb\Sigma^*\cap\Sigma^*y)\,\Sigma^*\cap\Sigma^*x\big)a\Sigma^*\cap\Sigma^*y\Big)\Sigma^*}, \tag{5.1}$$

where, for a regular expression R, $\overline{R}$ denotes the *complement* language $\Sigma^* \setminus L(R)$.

Let $\rhd$ be the inverse ordering of $\lhd$. Since $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\rhd}$ are both regular, then so is their intersection. Observe that a word u is in this intersection if and only if $[u]_k$ is a singleton set. Thus $L_{k,\Sigma,\text{sing}}$ is regular.

A representation of $L_{k,\Sigma,\text{sing}}$ similar to (5.1) is straightforward to construct. Indeed, a k-abelian singleton avoids all k-switchings that yield a different word. Thus, the expression for $L_{k,\Sigma,\text{sing}}$ is the same as in (5.1) except that we require $a \neq b$ instead of $a \lhd b$ in the intersection. □

For a fixed k and Σ it is easy, though time consuming, to construct a DFA recognizing $L_{k,\Sigma,\lhd}$ or $L_{k,\Sigma,\text{sing}}$ using (5.1). This may be done by using known methods for converting the regular expressions in (5.1) to automata and, further, by using known constructions for intersections of automata. Using the expression (5.1) (and the similar expression for k-abelian singletons), in [19] the minimal DFA recognizing the languages $L_{k,\mathbb{B},\lhd}$ for $k = 2, 3, 4$, the languages $L_{k,\mathbb{B},\text{sing}}$ for $k = 2, 3$, and the language $L_{2,\{a,b,c\},\lhd}$ are computed. We recall these in Figures 5.1–5.6. The transition tables are presented in Appendix C. We make a few observations and remarks.

Example 5.2. In the automata of Figures 5.1–5.6 the sink states are omitted (as are all the transitions leading to the sink state). All other states are accepting. This is because the languages are defined by avoiding certain patterns, so that the languages are factor closed. The number of states (including the sink state) in the minimal DFA of $L_{k,\Sigma,\lhd}$ for $k = 2, 3, 4$ are 10, 49, and 936 states, respectively. The number of states in the minimal DFA of $L_{k,\Sigma,\text{sing}}$ for $k = 2, 3$, and 4 are 15, 87, and 1011, respectively. Finally, the number of states in the minimal DFA recognizing $L_{2,\{a,b,c\},\lhd}$ and $L_{2,\{a,b,c\},\text{sing}}$ are 66 and 84, respectively.

Remark 5.3. It seems to be the case that the number of states in the minimal DFA recognizing $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\text{sing}}$ grow rapidly when k or Σ grows. It can be shown that the minimal DFA recognizing $L_{k,\Sigma,\lhd}$ must contain at least $m^{k-1}(m-1)+1$ vertex-disjoint cycles (see Section 6.3), and thus at least that many states. For the languages $L_{k,\Sigma,\text{sing}}$ no such lower bound is known (see discussion in Section 6.4).

Example 5.4. By observing the DFA in Figure 5.1, on the right hand side, we obtain the following regular expression for $L_{2,\mathbb{B},\text{sing}} \cap a\mathbb{B}^*$:

$$a(\varepsilon + aa^* + aa^*bb^* + b + ba(ba)^* + ba(ba)^*b + bb^* + bb^*a)$$

Recall that a more compact expression was obtained in Example 4.16.

We remark that a crude upper bound on the number of states in the minimal DFA recognizing $L_{k,\Sigma,\lhd}$ (and $L_{k,\Sigma,\text{sing}}$) can be obtained from the above regular expression using well-known conversions between various models of regular languages (see, [41, 45]). Indeed, e.g., *Glushkov's algorithm* outputs an equivalent NFA

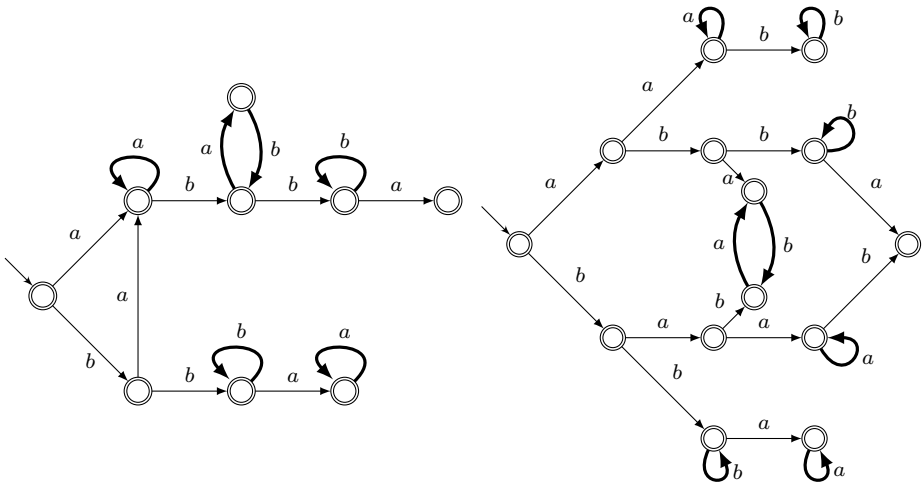

Figure 5.1: On the left the minimal DFA recognizing the language $L_{2,\mathbb{B},\lhd}$ $(a \lhd b)$. On the right the minimal DFA recognizing the language $L_{2,\mathbb{B},\mathrm{sing}}$. The sink states are not illustrated. [19, Figure 3], modification of [18, Fig. 3].

of $n+1$ states, given a regular expression of n occurrences of alphabet symbols. The determinization of an n state NFA can, in the worst case, give a DFA with 2^n states. The minimal DFA of the intersection of two regular languages, having n_1 and n_2 states respectively, can have $n_1 n_2$ states in the worst case. We have not attempted to make any precise estimations of the number of states in the minimal DFA recognizing $L_{k,\Sigma,\lhd}$ for general k.

Next we give an alternative proof of the regularity of the languages $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\mathrm{sing}}$. The first proof given here is quite straightforward and clean, which gives an algorithmic procedure to construct corresponding automata. The alternative proof, on the other hand, is not quite as straightforward, but it gives, as an immediate corollary, the fact that the languages have polynomial growth and provides bounds on the degree of the polynomial using an well-known result from the literature. Observe that the expression given in the first proof of Theorem 5.1 does not, at least not immediately, imply that this should be the case.

Theorem 5.5. *The language $L_{k,\Sigma,\lhd}$ (resp., $L_{k,\Sigma,sing}$) may be expressed by a finite union of regular expressions of the form $z_0 y_1^* z_1 \cdots y_t^* z_t$, where $t \leqslant m^{k-1}(m-1)+1$ (resp., $t \leqslant N_m(k-1)$).*

Proof. The following proof works for both languages with only slight differences, which we mention explicitly. Let thus L be either $L_{k,\Sigma,\lhd}$ or $L_{k,\Sigma,\mathrm{sing}}$. Recall Lemma 4.8, that is, for any long enough $u \in L$ we have that the corresponding cycle-deterministic walk W_u in $dB(k-1)$ is of the form

$$W_u = P_0 C_1^{\alpha_1} P_1 \cdots C_t^{\alpha_t} P_t, \tag{5.2}$$

where $\{C_1, \ldots, C_t\} = \mathrm{Cyc}(W_u)$, $\alpha_i \geqslant 1$ for each $i = 1, \ldots, t$, and P_i is a path for each $i = 0, \ldots, t$. Furthermore, W_u enters C_i at position $|P_0| + \sum_{j=1}^{i-1} |C_j^{\alpha_j} P_j|$ for each $i = 1, \ldots, t$. Up to varying the exponents α_i, there are finitely many such

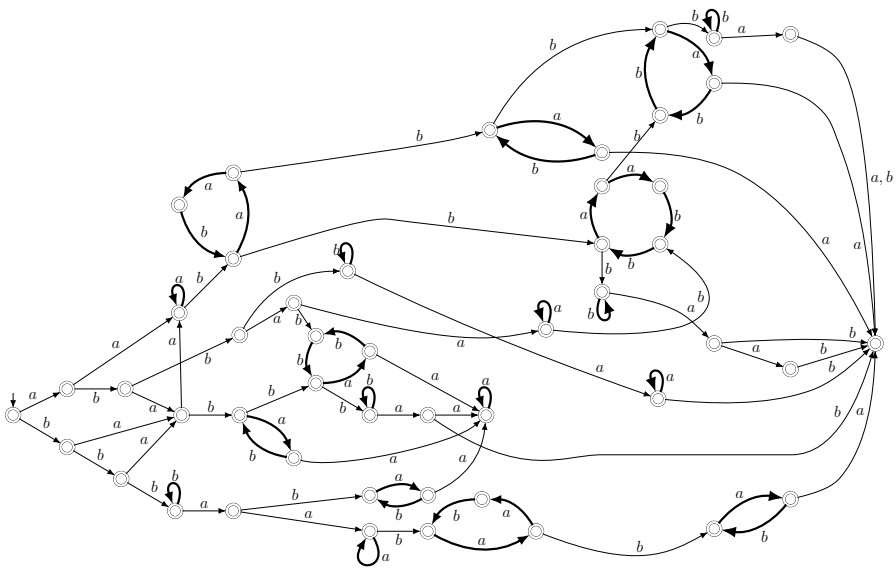

Figure 5.2: The minimal DFA recognizing the language $L_{3,\mathbb{B},\lhd}$ with $a \lhd b$. [19, Figure 4].

representations. Indeed, there are finitely many cycles in $dB(k-1)$, and W_u is cycle-deterministic. Furthermore, there are finitely many choices for the paths P_i. (In fact, by Corollary 4.11 we have $t \leqslant m^{k-1}(m-1)+1$ for $L_{k,\Sigma,\lhd}$. For $L_{k,\Sigma,\mathrm{sing}}$, we have $t \leqslant N_{|\Sigma|}(k-1)$ by Proposition 4.25.)

Let B be the (finite) set of words having a presentation of the form (5.2), where $\alpha_i \leqslant 2$ for all $i = 1, \ldots, t$. Let $v \in B$ and let W_v be of the form (5.2). Let $x = \mathrm{tail}(W_v)$, $z_i = \mathrm{label}(P_i)$ for each $i = 0, \ldots, t$, and $y_i = \mathrm{label}(C_i)$ for each $i = 1, \ldots, t$. We thus have $v = xz_0y_1^{\alpha_1}z_1 \cdots y_t^{\alpha_t}z_t$. Let L_v be the regular language defined by the expression

$$xz_0y_1^{\epsilon_1}z_1 \cdots y_t^{\epsilon_t}z_t, \tag{5.3}$$

where $\epsilon_i = 1$ if $\alpha_i = 1$; otherwise we let ϵ_i be the Kleene star $*$. (If v is a k-abelian singleton, we may take $\epsilon_i = *$ for each i.) We claim that $L = \cup_{v \in B} L_v$. We first show that $L_v \subseteq L$ for each $v \in B$. Indeed, let $u \in L_v$ for some $v \in B$. We have

$$u = xz_0y_1^{\beta_1}z_1 \cdots y_t^{\beta_t}z_t,$$

where $\beta_i \geqslant 0$ for each $i = 1, \ldots, t$. The walk W_u in $dB(k-1)$ corresponding to u is thus of the form

$$W_u = P_0C_1^{\beta_1}P_1 \cdots C_t^{\beta_t}P_t.$$

By repeatedly using Lemma 4.12 in the case of $L_{k,\Sigma,\lhd}$ (Lemma 4.26 in the case of $L_{k,\Sigma,\mathrm{sing}}$), we see that $u \in L$.

We then show that $L \subseteq \cup_{v \in B} L_v$. Let $u \in L$ have a walk in $dB(k-1)$ of the form (5.2). Then the word v has the same representation, only with exponents β_i defined by $\beta_i = 1$ if $\alpha_i = 1$, otherwise $\beta_i = 2$. By Lemma 4.12 in the case of

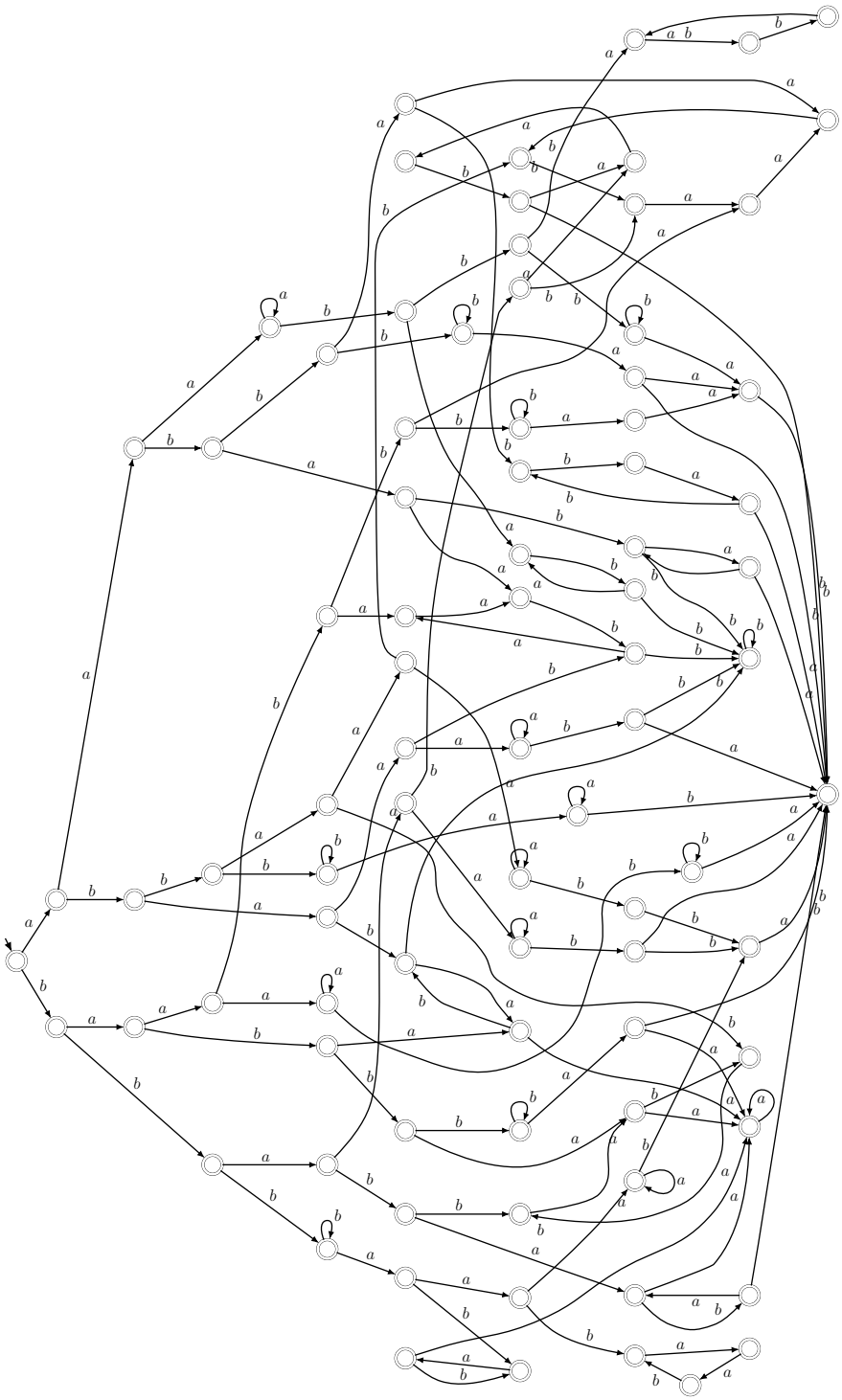

Figure 5.3: The minimal DFA recognizing the language $L_{3,\mathbb{B},\mathrm{sing}}$. Modification of [19, Figure 6].

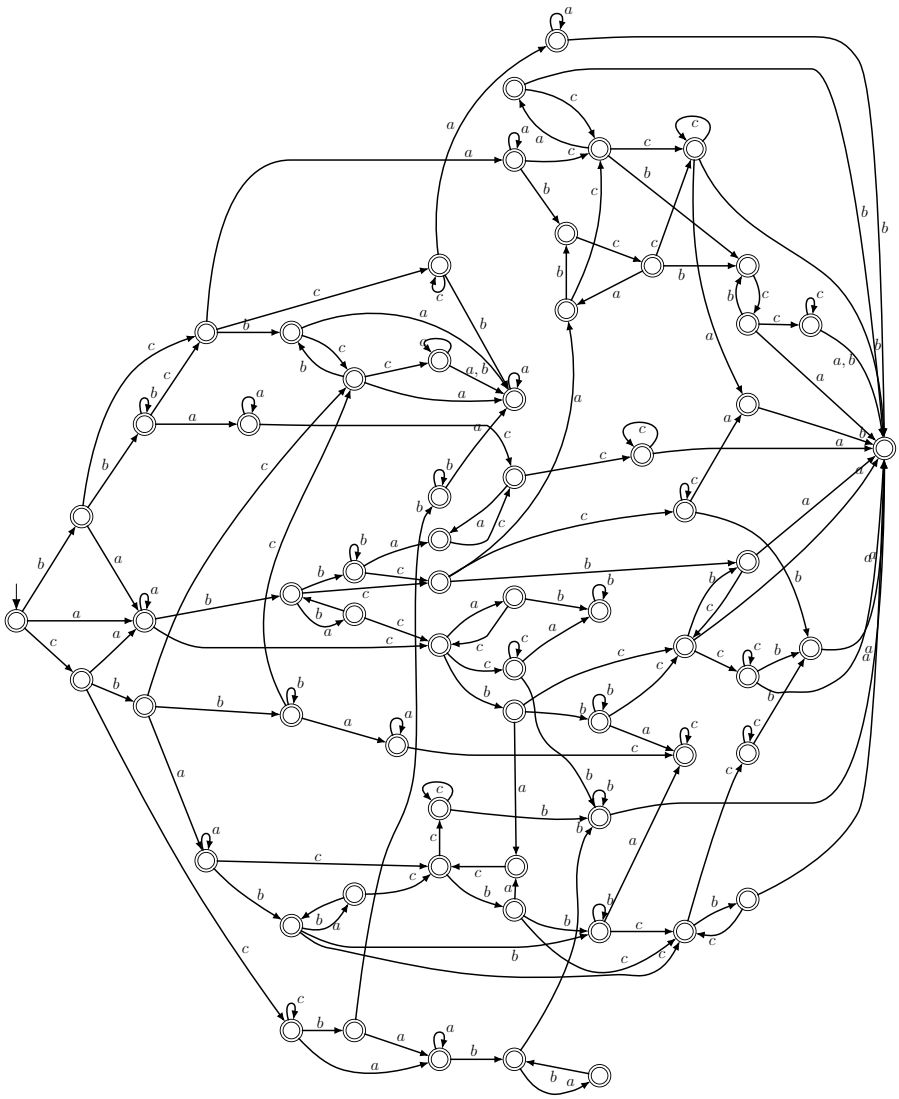

Figure 5.4: The minimal DFA recognizing the language $L_{2,\{a,b,c\},\lhd}$ with $a \lhd b \lhd c$. Modification of [19, Figure 5].

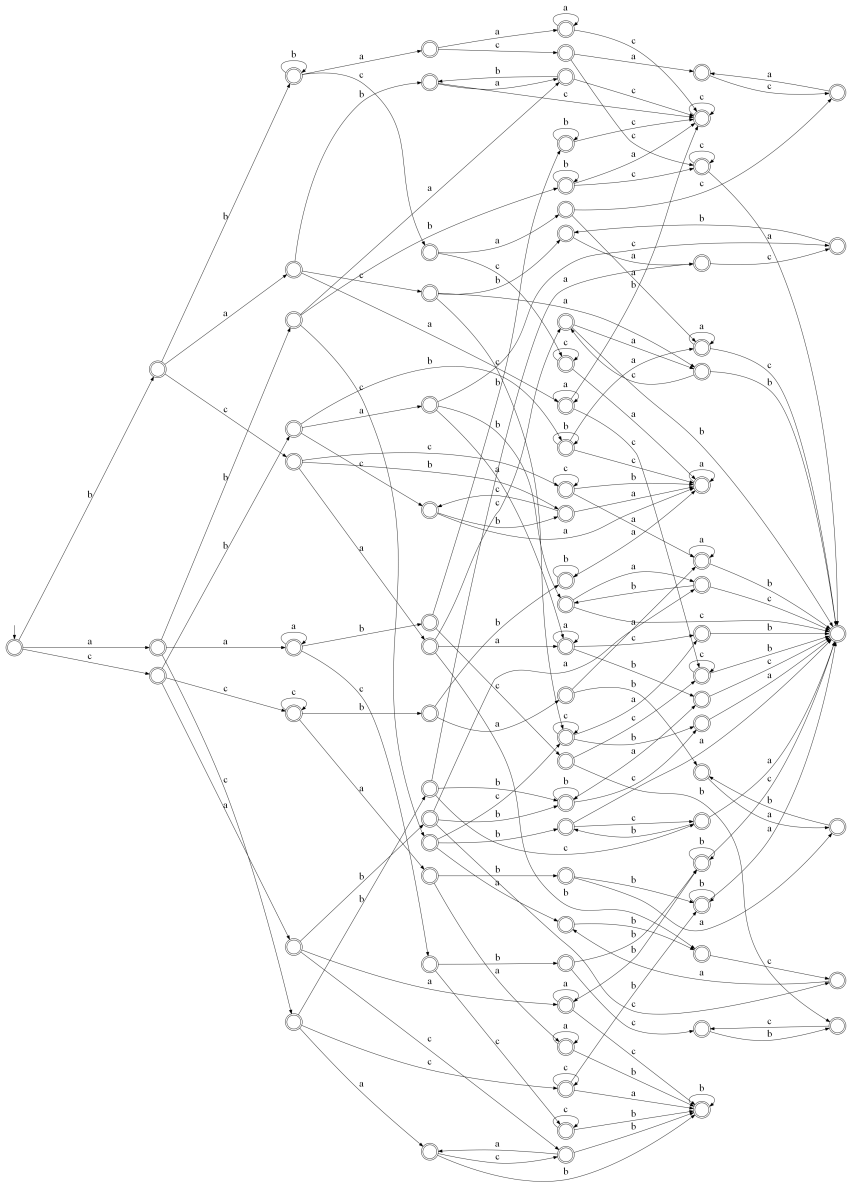

Figure 5.5: The minimal DFA recognizing $L_{2,\{a,b,c\},\mathrm{sing}}$. [19, Figure 7].

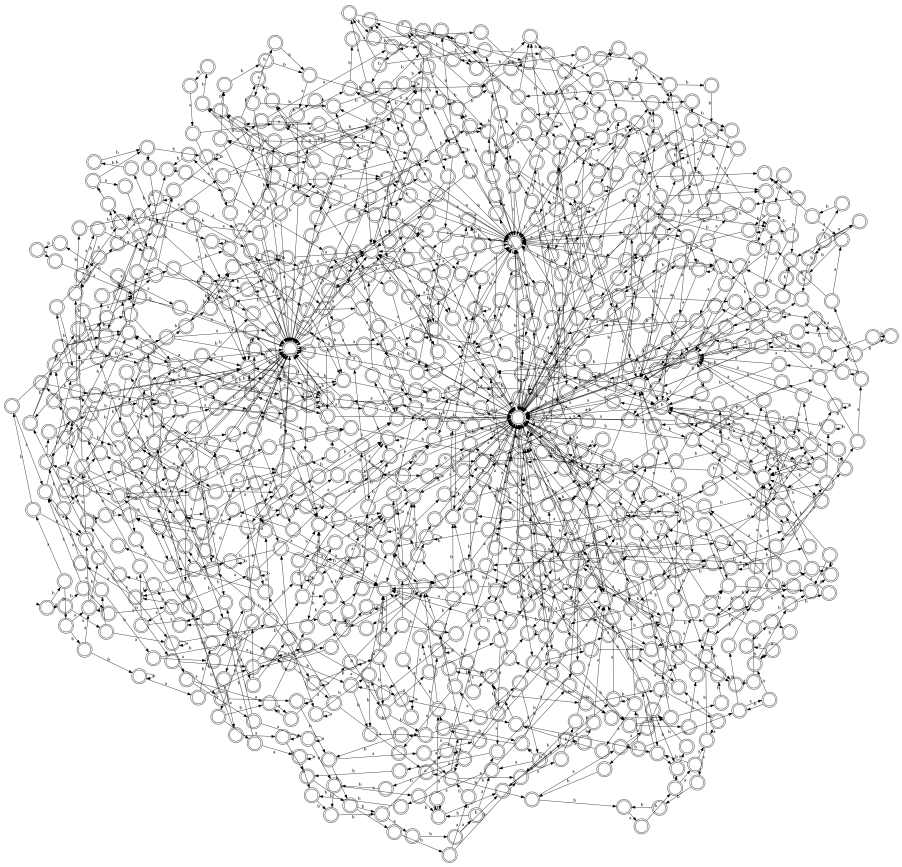

Figure 5.6: The minimal DFA recognizing the language $L_{4,\mathbb{B},\lhd}$ with $a \lhd b$. [19, Figure 8].

$L_{k,\Sigma,\lhd}$, and Lemma 4.26 in the case of $L_{k,\Sigma,\mathrm{sing}}$, we have $v \in L$. Thus $v \in B$ and clearly $u \in L_v$. This concludes the proof. □

We make the following observation which is a consequence of the above theorem.

Theorem 5.6. *For all $k \geqslant 1$ and Σ, where $|\Sigma| = m \geqslant 1$, we have*

$$\mathcal{C}_{L_{k,\Sigma,\lhd}}(n) = \mathcal{O}(n^{m^{k-1}(m-1)}) \text{ and}$$

$$\mathcal{C}_{L_{k,\Sigma,sing}}(n) = \mathcal{O}(n^{N_m(k-1)-1}).$$

This follows from Theorem 5.5 together with the following well-known result of [102] (for related considerations, see [40]):

Theorem 5.7. *For a regular language L, we have $\mathcal{C}_L(n) = \mathcal{O}(n^k)$ for some $k \geqslant 0$ if and only if L can be represented as a finite union of regular expressions of the form $z_0 y_1^* z_1 \cdots y_t^* z_t$ with a non-negative integer $t \leqslant k+1$, where $z_0, y_i, z_i \in \Sigma^*$ for all $i = 1, \ldots, t$.*

We consider more of these quantitative aspects connected to k-abelian equivalence classes and k-abelian singletons in the following chapter. We just mention that the first of the results is a weaker version of Theorem 1.1, and that the second result is actually the same as Theorem 6.25 (see discussion in the following chapter).

5.2 The k-switching as a Language Operation

We proceed to describe a k-switching operation on languages. In this section we show that this language operation preserves regularity. In other words, given a regular language L, the language obtained by this operation is also regular. This result is used in the following section. This section is based on the publications [18, 19]. Let us now define the k-switching on languages.

Definition 5.8. For a language $L \subset \Sigma^*$, we define the k-switching of L, denoted by $R_k(L)$, as the language

$$R_k(L) = \{w \in \Sigma^* \mid wR_kv \text{ for some } v \in L\}.$$

Similarly, we define $R_k^*(L) = \bigcup_{n\in\mathbb{N}} R_k^n(L) = \bigcup_{w\in L}[w]_k$.

Note that, from a regular language L, it is straightforward to identify all words that admit a k-switching (i.e., the words on the top row of Figure 3.1). It is not at all clear that, by performing all possible k-switchings on all words of L (i.e., taking the union of all words on the bottom row of Figure 3.1), the obtained language is also regular. The proof given here appears in [18].

Theorem 5.9. *Let L be a regular language. Then $R_k(L)$ is also regular.*

The proof is constructive. Given a DFA $\mathcal{A}$ recognizing the language L, we construct an ε-NFA $\mathcal{A}'$ which recognizes the language obtained by performing all possible k-switchings. We then have that $R_k(L) = L(\mathcal{A}')$. Let us first briefly sketch the construction of $\mathcal{A}'$ before defining it rigorously. To this end, let $u \in L$ and $v = S_{u,k}(i, j, \ell, m)$. The computation of $\mathcal{A}$ on u may be split into five stages: the computations of

1. $u[1, i)$ starting from the initial state, ending in some state q_i;
2. $u[i, j)$ starting from q_i, ending in some state q_j;
3. $u[j, \ell)$ starting from q_j, ending in some state q_ℓ;
4. $u[\ell, m)$ starting from q_ℓ, ending in some state q_m;
5. $u[m..]$ starting from q_m, ending in an accepting state f.

The word v has the same computations with the exception that stages 2. and 4. are interchanged. An accepting computation of $\mathcal{A}'$ on v performs this interchange by ε-transitions: the automaton $\mathcal{A}'$ guesses at which indices the k-switching has been performed by guessing the states q_t, where $t \in \{i, j, \ell, m\}$, which correspond to the states the automaton is in after reading $u[1, i)$, $u[1, j)$, $u[1, \ell)$, and $u[1, m)$, respectively. The automaton then performs four ε-transitions non-deterministically,

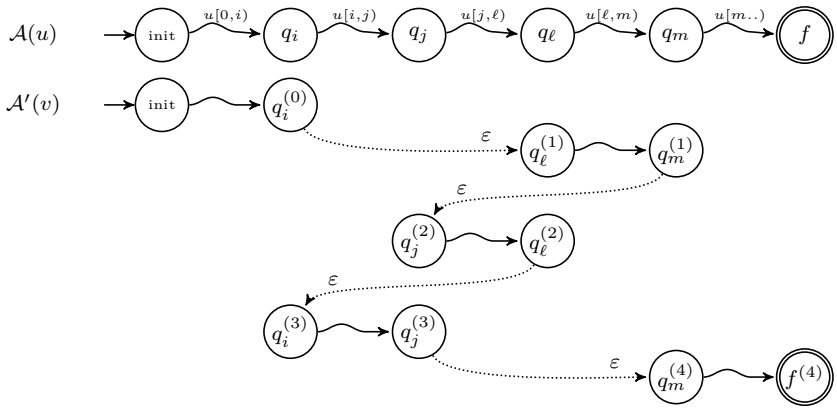

Figure 5.7: The computation of $\mathcal{A}$ on a word u and a computation of $\mathcal{A}'$ on $v = S_{k,u}(i,j,\ell,m)$. The automaton $\mathcal{A}'$ non-deterministically guesses the states q_t, where $t = 1,\ldots,4$, and performs the ε-transitions non-deterministically. The number of ε-transitions performed is encoded into the states of $\mathcal{A}'$. The factors of length $k-1$ starting at the first and third ε-transitions have to be equal, and this is checked in parallel. The same is done for the factors occurring at the second and fourth ε-transitions. We have abbreviated the states $q_r^{(c,(q_i,q_\ell),(q_j,q_m))}$ of the proof of Theorem 5.9 by $q_r^{(c)}$ (for $c \in \{0,\ldots,4\}$ and $r \in \{\text{init}, i, j, \ell, m\}$). [19, Figure 2], modification of [18, Fig. 2].

the order of which is important. The transitions are from q_i to q_ℓ, q_m to q_j, q_ℓ to q_i, and from q_j to q_m. Thus the word v is accepted if it is obtained by a k-switching on u. The computation of $\mathcal{A}'$ on v is depicted in Figure 5.7.

In order to ensure that any word accepted by $\mathcal{A}'$ is in $R_k(L)$, the automaton checks, in parallel, that the factors of length $k-1$ starting from the first and third ε-transitions are equal. Similar verifications are performed for the factors starting at the second and fourth ε-transitions. Furthermore, the automaton checks that, after the first and third ε-transitions, at least one letter is read before the next ε-transition is performed. (These verifications correspond to the requirements that $u[i,j)$ and $u[\ell,m)$ are non-empty.) Thus, any word accepted by $\mathcal{A}'$ can be obtained by a k-switching on some word in L.

Remark 5.10. It is worth noticing that, in a k-switching, the word v is obtained from u by changing the order of the factors $u[i,j)$ and $u[\ell,m)$. They are of unbounded length and hence cannot be remembered by a finite automaton. Instead, in the proof, only the corresponding states at positions of i, j, ℓ, and m in the automaton $\mathcal{A}$ recognizing u are remembered.

Proof of Theorem 5.9. For a language L and fixed words $x, y \in \Sigma^{k-1}$, consider the language

$$\begin{aligned} R_{x,y}(L) = \{w \in \Sigma^* \mid w = & S_{k,u}(i,j,\ell,m) \text{ for some } i < j \leqslant \ell < m, u \in L, \\ & \text{with } u[i,i+k-1) = u[\ell,\ell+k-1) = x \text{ and} \\ & u[j,j+k-1) = u[m,m+k-1) = y\}. \end{aligned}$$

We construct, for a regular language L recognized by a deterministic finite automaton $\mathcal{A} = (Q, \Sigma, \delta, p_{\text{init}}, F)$, an ε-NFA $\hat{\mathcal{A}}$ which recognizes $R_{x,y}(L)$. The claim then follows for $R_k(L)$, as $R_k(L) = \bigcup_{x,y\in\Sigma^{k-1}} R_{x,y}(L)$ is a finite union of regular languages.

In essence, $\hat{\mathcal{A}}$ is a Cartesian product of form $\hat{\mathcal{A}} = \mathcal{A}_1 \times \mathcal{A}_x \times \mathcal{A}_y \times \mathcal{A}_x \times \mathcal{A}_y$. The first component automaton $\mathcal{A}_1$ consists of $5|Q|^4$ copies of $\mathcal{A}$, some of which are connected by ε-transitions. The second and fourth components are copies of an automaton $\mathcal{A}_x$ recognizing the language $x\Sigma^*$ and the third and fifth components are copies of an automaton $\mathcal{A}_y$ recognizing the language $y\Sigma^*$. The components $2, 3, 4$, and 5 are initiated according to the computations performed in $\mathcal{A}_1$. We shall now make this construction more formal.

We first construct $\mathcal{A}_1 = (Q_1, \Sigma, \delta_1, \tilde{p}_{\text{init}}, F_1)$ as follows. For each state $p \in Q$, we have $p^{(c,(p_1,p_2),(p_3,p_4))} \in Q_1$ for all $c = 0, \ldots, 4$ and $p_r \in Q$, $r = 1, \ldots, 4$. We also add the initial state $\tilde{p}_{\text{init}}$, from which we have ε-transitions to all the states of form $p_{\text{init}}^{(1,(p_1,p_2),(p_3,p_4))}$, $p_1, p_2, p_3, p_4 \in Q$. Thus the computation of $\mathcal{A}_1$ begins with an ε-transition. We then add the following ε-transitions for all $p_1, p_2, p_3, p_4 \in Q$:

$$
\begin{aligned}
p_1^{(0,(p_1,p_2),(p_3,p_4))} &\xrightarrow{\varepsilon} p_2^{(1,(p_1,p_2),(p_3,p_4))} \\
p_3^{(1,(p_1,p_2),(p_3,p_4))} &\xrightarrow{\varepsilon} p_4^{(2,(p_1,p_2),(p_3,p_4))}, \\
p_2^{(2,(p_1,p_2),(p_3,p_4))} &\xrightarrow{\varepsilon} p_1^{(3,(p_1,p_2),(p_3,p_4))}, \\
p_4^{(3,(p_1,p_2),(p_3,p_4))} &\xrightarrow{\varepsilon} p_3^{(4,(p_1,p_2),(p_3,p_4))}.
\end{aligned}
$$

Otherwise the computation of $\mathcal{A}_1$ respects the original automaton, that is,

$$\delta_1(p^{(i,(p_1,p_2),(p_3,p_4))}, a) = q^{(i,(p_1,p_2),(p_3,p_4))}$$

if and only if there is a transition $\delta(p, a) = q$ in $\mathcal{A}$. Finally, F_1 consists of all states of form $f^{(5,(p_1,p_2),(p_3,p_4))}$, where $f \in F$ and $p_1, p_2, p_3, p_4 \in Q$.

We remark the following about $\mathcal{A}_1$. Firstly, once the first ε-transition is taken, the states p_1, p_2, p_3, and p_4 are fixed for the remainder of the computation. Secondly, the states p_r, $r = 1, \ldots, 4$, determine between which states an ε-transition can be performed. Furthermore, the parameter c counts the number of ε-transitions performed (after the first ε-transition which starts the computation). The parameters c, p_1, p_2, p_3, and p_4 together determine at which time and between which states an ε-transition can be performed.

We now describe the behavior of the rest of the component automata of $\hat{\mathcal{A}}$. For $s \in \{2, \ldots, 5\}$, the sth component automaton of $\hat{\mathcal{A}}$ is initiated during the sth ε-transition performed in $\mathcal{A}_1$ (the first ε-transition being the first computation step of $\mathcal{A}_1$). We also require from $\hat{\mathcal{A}}$ that, after the second and fourth ε-transition performed in $\mathcal{A}_1$, at least one letter is read before performing the next ε-transition. This is not required after the third ε-transition. Note that these requirements can be encoded, e.g., into the parameter c of the states in A_1. Finally, $\hat{\mathcal{A}}$ accepts if and only if all its components are in accepting states.

We first show that $R_{x,y}(L) \subseteq L(\hat{\mathcal{A}})$. In order to see this, let $u \in L$ and let $v = S_{k,u}(i, j, \ell, m) \in R_{x,y}(L)$. Let q_t, $t = 1, \ldots, |u|$, denote the state $\delta(p_{\text{init}}, u[1, t))$ (note that some of the states q_t can be the same). We then find an accepting computation of $\mathcal{A}_1$ for v as follows. We first take the ε-transition from $\tilde{p}_{\text{init}}$ to the state $p_{\text{init}}^{(0,(q_i,q_\ell),(q_j,q_m))}$. After this, the computation is as in Figure 5.7 by

following the dashed lines. The computation of $\mathcal{A}$ on u follows the continuous lines. Note that the other components of $\hat{\mathcal{A}}$ also end up in accepting states, since by the definition of the k-switching $S_{k,u}(i,j,\ell,m)$, x and y have positions in v corresponding to the initiations of the copies of the automata $\mathcal{A}_x$ and $\mathcal{A}_y$. Thus $R_{x,y}(L) \subseteq L(\hat{\mathcal{A}})$.

We now show the converse. For this, let $v \in L(\hat{\mathcal{A}})$ and consider an accepting path of $\hat{\mathcal{A}}$ on v. By construction, the automaton $\mathcal{A}_1$ starts with an ε-transition to a state $p_{\text{init}}^{(0,(p_1,p_2),(p_3,p_4))}$. After this, the computation contains four more ε-transitions; suppose they occur just before reading the ith, jth, ℓth and mth letter, with $i < j \leqslant \ell < m$, respectively. (Here we use the requirement of not allowing an ε-transition immediately after the second and fourth ε-transitions.) Furthermore, by the acceptance of the other component automata of $\hat{\mathcal{A}}$, x has positions i and ℓ, and y has positions j and m in v. We claim that $u = S_{k,v}(i,j,\ell,m) \in L$. It then follows, by the symmetry of the k-switching relation, that $v \in R_{x,y}(L)$. Indeed, turning back to the computation of $\mathcal{A}_1$ on v, we obtain the following walks in $\mathcal{A}$:

1. a path from p_{init} to p_1 labeled by $v[0,i)$,
2. a path from p_2 to p_3 labeled by $v[i,j)$,
3. a path from p_4 to p_2 labeled by $v[j,\ell)$,
4. a path from p_1 to p_4 labeled by $v[\ell,m)$, and
5. a path from p_3 to an accepting state of $\mathcal{A}$ labeled by $v[m..]$.

Thus $u = v[1,i)v[\ell,m)v[j,\ell)v[i,j)v[m..] \in L$, as was claimed. □

In contrast to the above theorem, the following example shows that the family of regular languages is not closed under the language operation R_k^*.

Example 5.11. Fix $k \geqslant 1$ and let $L = (ab^k)^+$. It is straightforward to verify, e.g., by comparing the number of occurrences of factors of length k, that

$$R_k^*(L) = \{ab^{r_1}ab^{r_2}\cdots ab^{r_n} : n \geqslant 1, r_i \geqslant k-1, \textstyle\sum_{i=1}^n r_i = nk\}.$$

Let now h be a morphism defined by $h(a) = ab^{k-1}$ and $h(b) = b$. It is again straightforward to show that $h^{-1}(R_k^*(L)) = \{w \in a\{a,b\}^* : |w|_a = |w|_b\}$, which is clearly not regular. It follows that $R_k^*(L)$ is not regular.

5.3 The Regularity of Classes of Constant Cardinality

The established regularity of the languages $L_\lhd$ and L_{sing} raises questions for the structure of larger equivalence classes. We are thus interested in the k-abelian equivalence classes of fixed cardinality, that is the languages $L_{k,\Sigma,r}$, where $r \geqslant 2$. We first consider the case $r \geqslant 2$. We show that $L_{k,\Sigma,2}$ is regular by employing Theorem 5.9. In the following, we say that $y \in \Sigma^*$ is *extremal* (with respect to $\lhd$) if y is in the (regular) language $L_{\text{ext}} = L_{k,\Sigma,\lhd} \cup L_{k,\Sigma,\rhd}$. In other words, y is extremal if it is either the lexicographically least or the lexicographically maximal element of $[y]_k$.

Theorem 5.12. *The language $L_{k,\Sigma,2}$ is regular.*

Proof. Consider the regular language $L = \Sigma^* \setminus L_{\text{ext}}$: we have

$$L = \{w \in \Sigma^* \mid |[w]_k| \geqslant 3 \text{ and } w \text{ is not extremal}\},$$

since all classes containing at most two elements are removed. Let us apply the language operation R_k defined previously to L: take $L' = R_k(L) \cup L$. Note that L' is regular since regular languages are closed under union and the operation R_k. It is not hard to see that $L' = \{w \in \Sigma^* \mid |[w]_k| \geqslant 3\}$. Indeed, note that one operation of R_k is sufficient to fill in the equivalence classes: by Claim 3.13, each word $x \in L$ admits at least two distinct switchings, one decreasing and the other increasing lexicographically since x is not extremal. Finally, the complement of L' is the language $\{w \in \Sigma^* \mid |[w]_k| \leqslant 2\}$. Consequently

$$L_{k,\Sigma,2} = (\Sigma^* \setminus L') \setminus L_{k,\Sigma,\text{sing}}$$

is a regular language. □

The use of R_k gives a nice proof of the above result. For $r \geqslant 3$ this approach does not seem to work, at least in a straightforward manner (see Proposition 5.15 and consequent discussion for the case $r = 3$). In the following section we describe a (failing) approach which gives an interesting view on the problem. The rest of this section is devoted to proving the following generalization.

Theorem 5.13. *For any $k \geqslant 1$, alphabet Σ, and integer $r \geqslant 1$, the language $L_{k,\Sigma,r}$ is regular.*

We show that $L_{k,\Sigma,r}$ is regular for any $r \geqslant 3$ by the use of $\text{Min}_k(L_{k,\Sigma,r})$ and the language operation R_k. We first show that $\text{Min}_k(L_{k,\Sigma,r})$ is regular. The proof is very much similar to the proof of Theorem 5.5. The main ingredients here are Lemma 4.28 and Lemma 4.32. The regularity of $L_{k,\Sigma,r}$ then follows from Proposition 5.14 by applying r times the regularity preserving language operation R_k on $\text{Min}_k(L_{k,\Sigma,r})$.

Proposition 5.14. *The language $Min_k(L_{k,\Sigma,r})$ is regular for any integers r, k, and alphabet Σ.*

Proof. We claim that $\text{Min}_k(L_{k,\Sigma,r})$ is a finite union of languages defined by regular expressions of the form $z_0 y_1^* z_1 \cdots y_t^* z_t$.

Consider now a word $u \in \text{Min}_k(L_{k,\Sigma,r})$. Since $u \in L_{k,\Sigma,\lhd}$, by Lemma 4.8, we may write

$$W_u = W_0 C_1^{s_1} W_1 \cdots C_t^{s_t} W_t$$

for some paths W_i, where $i = 0, \ldots, t$, and some repetitions $C_i^{s_i}$ of cycles C_i, where $i = 1, \ldots, t$, such that W_u enters cycle C_i at position $|W_0| + \sum_{j=1}^{i-1} |C_j^{s_j} W_j| + 1$, for all $1 \leqslant i \leqslant t$, and leaves C_i before entering C_{i+1}. Now u may be written as

$$u = \text{tail}(W_0) \cdot \text{label}(W_0 C_1^{s_1} W_1 \cdots C_t^{s_t} W_t) = z_0 y_1^{s_1} z_1 \cdots y_t^{s_t} z_t,$$

where $z_0 = \text{tail}(W_0) \cdot \text{label}(W_0)$, $y_i = \text{label}(C_i)$, and $z_i = \text{label}(W_i)$ for $1 \leqslant i \leqslant t$. Now if $s_i > \mathcal{B}_{k,r}$, then Lemma 4.28 ensures that

$$L(z_0 y_1^{s_1} z_1 \cdots y_i^* z_i \cdots y_t^{s_t} z_t) \subseteq L_{k,\Sigma,r}. \tag{5.4}$$

By repeating the above, we may replace all exponents s_j satisfying $s_j > \mathcal{B}_{k,r}$ with $*$ in (5.4).

Let L be the union of all the languages obtained as above from words $u \in \mathrm{Min}_k(L_{r,k})$ satisfying $|u|_x \leqslant \mathcal{B}_{k,r} + 1$ for all $x \in \Sigma^{k-1}$. These words are bounded in length, so that the union is finite. Clearly $L \subseteq \mathrm{Min}_k(L_{r,k})$ by the above observation. We claim that $\mathrm{Min}_k(L_{r,k}) \subseteq L$.

Indeed, let $u \in \mathrm{Min}_k(L_{r,k})$. If $|u|_x > B_{k,r} + 1$ Lemma 4.31 ensures that we have $W_u = W_0 W_y^{|u|_x - 1} W_1$ for y the unique complete first return to x in u. Further W_y is a cycle. If W_u does not enter W_y at position $|W_0| + 1$, we may extend the cycle to the left and right to obtain $W_0' W_{y'}^t W_1'$, where $t \in \{|u|_x - 1, |u|_x\}$. By the above lemma we may reduce the number of repetitions of W_y' to obtain a word u' for which $|u'|_x \leqslant \mathcal{B}_{k,r} + 1$ and u is in the language defined by u' as in (5.4). If $|u'|_{x'} > \mathcal{B}_{k,r} + 1$ for some $x' \in \Sigma^{k-1}$, we may repeat the above for u' to obtain a word u'' having $|u''|_{x'} \leqslant \mathcal{B}_{k,r} + 1$ and that u and u' are in the language defined by u'' as in (5.4). This can be continued until we obtain a word v such that $|v|_x \leqslant \mathcal{B}_{k,r} + 1$ for all $x \in \Sigma^{k-1}$ and u is contained in the language defined by v as in (5.4). We thus have $u \in L$, which concludes the proof. □

Proof of Theorem 5.13. The language $\mathrm{Min}_k(L_{r,k})$ is regular. Since the operation R_k preserves regularity by Theorem 5.9 and thus, by applying finitely many iterations of R_k, we have that $L_{k,\Sigma,r} = R_k^r(\mathrm{Min}_k(L_{k,\Sigma,r}))$ is regular. □

5.4 Some Related (Non-)Closure Properties

In this section we consider language operations related to R_k. We consider the question whether regular languages are closed under this operation. The motivation for studying these language operations comes from different approaches to proving the regularity of $L_{k,\Sigma,r}$, $r \geqslant 1$. Unfortunately, these approaches either fail or seem to become too involved. We nonetheless consider these as having independent interest.

The language operation R_k can be modified, e.g., to obtain the language operation $R_{k,\neq}$ defined by

$$R_{k,\neq}(L) = \{u \in \Sigma^* \mid \exists v \in L : u \in R_k(\{v\}) \setminus \{v\}\}.$$

This latter operation performs k-switchings that actually give another word. A straightforward modification of the proof of Theorem 5.9 shows that regular languages are closed under this new operation as well. A similar observation holds also for the operation $R_{k,\lhd}(L)$ which gives all words that are obtained by a k-switching on a lexicographically larger element of L.

Using the newly defined operation $R_{k,\neq}$, we can show that the language

$$K = \{x \in \Sigma^* \colon R_{k,\neq}(y) = [x]_k \cap L_{\mathrm{ext}}\ \forall y \in [x]_k \setminus L_{\mathrm{ext}}\}$$

is a regular language. This is the language of words x for which any $y \in [x]_k$, with y not extremal, admits exactly two (non-trivial) k-switchings: one giving the least element of $[x]_k$, the other giving the maximal element of $[x]_k$. Note that the language $L_{k,\Sigma,3}$ is included in K, but that there also exist other classes in K. For example, $[u]_k \subseteq K$, where $u = a^k b a^{k-1} c^k d c^{k-1}$ for which $\#[u]_k = 4$.

Proposition 5.15. *The language K is regular.*

Proof. Let again $L = \Sigma^* \setminus L_{\text{ext}} = \{w \in \Sigma^* : \ w \text{ not extremal}\}$. We obtain another regular language K_1 defined by

$$K_1 = R_{k,\neq}(L) \setminus L_{\text{ext}} = \{w \in \Sigma^* \setminus L_{\text{ext}} : R_{k,\neq}(\{w\}) \setminus L_{\text{ext}} \neq \emptyset\}.$$

The language K_1 consists of all the words $w \in \Sigma^* \setminus L_{\text{ext}}$ for which $R_k(w)$ contains non-extremal elements. Indeed, let $w \in \Sigma^* \setminus L_{\text{ext}}$ and assume that there is a k-switching giving a non-extremal element $w' \in L$. Then there is a k-switching on w' giving w. Thus $w \in R_{k,\neq}(w') \setminus L_{\text{ext}} \subseteq K_1$. On the other hand, if $R_k(w)$ contains only extremal elements, then $w \notin R_{k,\neq}(L) \setminus L_{\text{ext}} = K_1$. We further observe that, for the language $K_2 = R_k(K_1) \cup K_1$, we have

$$K_2 = K_1 \cup \big(L_{\text{ext}} \cap \{[w]_k \mid w \in K_1\}\big).$$

It is clear that if $w \in K_2$ is not extremal, then $w \in K_1$ from the above. To see that $L_{\text{ext}} \subset K_2$, let $w \in K_1$ and consider the lexicographically greatest element w' of $[w]_k \cap K_1$. Since $w' \notin L_{\text{ext}}$, there exists a k-switching on w' which is lexicographically greater. Since w' was maximal in $[w_k] \cap K_1$, it follows that this switching must be the lexicographically greatest element of $[w]_k$. The case of the lexicographically least element is similar.

Consider the language $K_3 = R_k(K_2) \cup K_2$. We claim that it consists of k-abelian equivalence classes $[x]_k$ for which there exists $x' \in [x]_k \setminus L_{\text{ext}}$ such that $R_k(x') \setminus L_{\text{ext}} \neq \emptyset$. Let $x \in K_2$ and assume that $x' \in [x]_k$. We show that $x' \in K_3$. Since $x' \notin K_2$, it follows that x' is not extremal but $R_k(x') = [x]_k \cap L_{\text{ext}}$. It follows that $x' \in R_k([x]_k \cap L_{\text{ext}}) \subseteq K_3$. We conclude by noting that K is the complement of K_3. Since all the operations performed above are regularity preserving and we start from a regular language, we conclude that K is a regular language. □

Considering the problem of showing that $L_{k,\Sigma,3}$ is regular this way, it seems that separating k-abelian equivalence classes of size 3 from other classes occurring in K could be quite involved using k-switchings alone.

We then consider another approach generalizing that of Theorem 5.12. If the class of regular languages was closed under Min_k, the regularity of $L_{k,\Sigma,r}$ would then easily be proved using the same idea as in Theorem 5.12. Indeed, by setting $K_{i+1} = K_i \setminus \text{Min}_k(K_i)$, $K_0 = \Sigma^*$, we obtain that $\Sigma^* \setminus R_k^r(K_r) = \cup_{i \leqslant r} L_{k,\Sigma,r}$ for each $r \in \mathbb{N}$. If Min_k preserved regularity, then $L_{k,\Sigma,r}$ would be regular since a finite sequence of regularity preserving operations would be used. Unfortunately, as it will soon be shown, Min_k does not preserve regularity. The approach of removing (in a regular way) one element from each k-abelian equivalence class at a time does not seem to extend to the languages $L_{k,\Sigma,r}$, $r \geqslant 3$. One reason to think that Min_k would preserve regularity is the following observation. We may extend the definition of k-abelian equivalence to the case of $k = 0$, the relation being the "equal length" relation. An old result gives a positive answer to the above question when $k = 0$.

Theorem 5.16 ([6, Theorem 4.1], see also [11])**.** *For every regular language L, the language* $\min(L) = \{w \in L \mid w \trianglelefteq u \text{ for every } u \in L, |u| = |w|\}$ *is regular, and a regular grammar (and thus an automaton) for it can be effectively constructed.*

For $k \geqslant 1$, we see that Min_k does not preserve regularity.

Example 5.17. Let $k \geqslant 1$ and $L = (ab^k)^* \cup ab^{k-1}b^*(ab^{k-1})^*$. When n is not a multiple of $k+1$, we see that the words of length n of L are all in distinct k-abelian equivalence classes. On the other hand, when $n = s(k+1)$, $s \geqslant 2$, we see that $(ab^k)^s$ and $ab^{k-1}b^s(ab^{k-1})^{s-1}$ are k-abelian equivalent. Again, all other words are in distinct k-abelian equivalence classes. Since $(ab^k)^s \lhd ab^{k-1}b^s(ab^{k-1})^s$ we deduce that

$$L \setminus \mathrm{Min}_k(L) = \{(ab^{k-1})b^{r+1}(ab^{k-1})^r \mid r \geqslant 1\}.$$

The language $h^{-1}(L \setminus \mathrm{Min}_k(L))$, where $a \mapsto ab^{k-1}$ and $b \mapsto b$, equals $\{abb^r a^r \mid r \geqslant 1\}$ which is clearly not regular. Since all other operations preserve regularity, we conclude that Min_k does not preserve regularity.

There might still be other ways to remove one element at a time from each equivalence class in a regular way. The above examples show that this operation would be quite involved, if such an operation existed.

In this chapter we considered language theoretic aspects of the k-abelian equivalence classes. The results obtained here are utilized in the following chapter, where we consider the asymptotic numbers of k-abelian equivalence classes of length n. In particular, we utilize well-known results for computing the asymptotic growth rates of regular languages.

Chapter 6

Quantitative Aspects of k-Abelian Equivalence

In this chapter we focus on quantitative aspects of k-abelian equivalence classes. We mainly consider the number of k-abelian equivalence classes of words of length n over an m-letter alphabet, and the number length n k-abelian singletons over an m-letter alphabet.

In the first section of this chapter we consider the exact number of k-abelian equivalence classes of length n and the exact number of k-abelian singletons of length n. We show that these numbers, interpreted as sequences, are $\mathbb{N}$-*rational*. We then compute explicit expressions giving, for each $n \in \mathbb{N}$, the number of k-abelian equivalence classes of length n over an m-letter alphabet for small values of k and m. We recall relevant tools from the literature. The basis for computing closed formulae for these sequences is based on the automata constructed in the previous chapter.

In the second section we consider a more general setting. In particular, we prove a sufficient condition for an regular language having polynomial complexity to have asymptotically polynomial complexity. That is, we give a condition such that if a regular language L has $\mathcal{C}_L(n)$ of the order $\mathcal{O}(n^t)$ but $\mathcal{C}_L(n)$ is not of the order $\mathcal{O}(n^{t-1})$ for some $t \geqslant 1$, and satisfies the above mentioned condition, then L has complexity $\mathcal{C}_L(n) = C \cdot n^t + \mathcal{O}(n^{t-1})$ for some constant C. The main tools used here are the generating functions of such regular languages.

In the third section we consider the asymptotic growth of the number of k-abelian equivalence classes of length n over an m-letter alphabet. We show that this sequence is asymptotic to a polynomial using results from the second section of this chapter. In the fourth section we consider the numbers of k-abelian singletons and show a connection to *Gray codes for necklaces*.

6.1 Exact Numbers of Equivalence Classes and Singletons

Let $\mathcal{P}_m^{(k)}(n)$ be the number of k-abelian equivalence classes of length n words over an m-letter alphabet. Let similarly $\mathcal{S}_m^{(k)}(n)$ be the number of k-abelian singletons

of length n words over an m-letter alphabet. In this section we study the sequences $(\mathcal{P}_m^{(k)}(n))_{n\geqslant 0}$ and $(\mathcal{S}_m^{(k)}(n))_{n\geqslant 0}$. We show that these sequences are $\mathbb{N}$-rational and how, in principle, an explicit formula can be constructed for each of them. We illustrate this method by giving explicit formulae for $\mathcal{P}_m^{(k)}(n)$ and $\mathcal{S}_m^{(k)}(n)$ for small values of k and m. For example, we show that $\mathcal{P}_2^{(2)}(n) = n^2 - n + 2$ for all $n \geqslant 1$ and $\mathcal{S}_2^{(2)}(n) = 2n+4$ for all $n \geqslant 4$. These results were previously proved in [49] and [55], respectively, using different methods. The results of this section appear in [19], though the methods were only briefly discussed in the article. We elaborate on these methods in this section. We also provide novel results (Propositions 6.6 and 6.7), which have not appeared previously.

We start by observing that $\mathcal{P}_m^{(k)}(n) = \mathcal{C}_{L_{k,\Sigma,\lhd}}(n)$ and $\mathcal{S}_m^{(k)}(n) = \mathcal{C}_{L_{k,\Sigma,\mathrm{sing}}}(n)$, where Σ is any m-letter alphabet. Indeed, each word in $L_{k,\Sigma,\lhd}$ corresponds to a unique k-abelian equivalence class and vice versa. Thus $\mathcal{C}_{L_{k,\Sigma,\lhd}}(n)$ gives the number of k-abelian equivalence classes of length n for each $n \geqslant 0$. Similar arguments hold for k-abelian singletons. As an immediate consequence (by Proposition 2.17) we have the following.

Proposition 6.1 ([19]). *The sequences $(\mathcal{P}_m^{(k)}(n))_{n\geqslant 0}$ and $(\mathcal{S}_m^{(k)}(n))_{n\geqslant 0}$ are $\mathbb{N}$-rational for any k and m.*

Next we study the question of how to obtain explicit formulae for $\mathcal{P}_m^{(k)}(n)$ and $\mathcal{S}_m^{(k)}(n)$. Due to the connection to $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\mathrm{sing}}$, respectively, this task reduces to finding the number of distinct words of length n of a given regular language. Several strategies are known for doing this, and we mention two of them used here to obtain the formulae in Propositions 6.4 and 6.5. We consider here the function $\mathcal{P}_m^{(k)}(n)$, the case of $\mathcal{S}_m^{(k)}(n)$ being analogous. Given k and m, we may construct a DFA $\mathcal{A}$ recognizing $L_{k,\Sigma,\lhd}$ for some m-letter alphabet Σ and an ordering $\lhd$ as described in the previous chapter.

We then construct a unary automaton $\mathcal{A}'$ by identifying all the letters. Now the number of length n words in $L_{k,\Sigma,\lhd}$ equals the number of accepting paths in $\mathcal{A}'$ of length n. Let M be the adjacency matrix of $\mathcal{A}'$. It is known that, for all large enough n,

$$\ell_{\mathcal{A}'}(n) = \sum_{\lambda\in \mathrm{Eig}(M)} p_\lambda(n)\lambda^n = \sum_{i=0}^{d} \Big(\sum_{\lambda\in \mathrm{Eig}(M)} \alpha_{\lambda,i}\lambda^n \Big) n^i, \tag{6.1}$$

where, the first summation is taken over all distinct eigenvalues $\mathrm{Eig}(M)$ of M, and p_λ is a polynomial with complex coefficients of degree at most $\mu_\lambda - 1$ for each eigenvalue λ. Here μ_λ is the multiplicity of λ as a root of the *minimal polynomial* of M. (See, e.g., [34, 107].) In the second summation we have $d = \max_{\lambda\in \mathrm{Eig}(M)\setminus 0} \mu_\lambda - 1$, the coefficients $\alpha_{\lambda,i}$ are some complex numbers, and $\alpha_{\lambda,i} = 0$ when $i \geqslant \mu_\lambda$. When considering our sequences $\mathcal{P}_m^{(k)}(n)$ and $\mathcal{S}_m^{(k)}(n)$, it is evident that the polynomial $p_\lambda \equiv 0$ whenever $|\lambda| > 1$, since in this case $\ell(n)$ is bounded from above by a polynomial.

In Tables 6.1 and 6.2 we list the minimal polynomials of the minimal DFA recognizing $L_{k,\Sigma,\lhd}$ and $L_{k,\Sigma,\mathrm{sing}}$ in Figures 5.1–5.6. Using these minimal polynomials we describe two methods to find an explicit expression of $\mathcal{P}_m^{(k)}$ and $\mathcal{S}_m^{(k)}$ for the corresponding values k and m.

$(k, \lvert\Sigma\rvert)$	Minimal polynomial associated to $L_{k,\Sigma,\lhd}$
$(2,2)$	$(x-2)\,(x-1)^3\,x^2\,(x+1)$
$(2,3)$	$(x-3)\,(x-1)^7\,x^3\,(x+1)^3\,(x^2+x+1)$
$(3,2)$	$(x-2)\,(x-1)^5\,x^5\,(x+1)\,(x^2+1)\,(x^2+x+1)$
$(4,2)$	$(x-2)\,(x-1)^9\,x^{10}\,(x+1)^5\,(x^2-x+1)\,(x^2+1)^3\,(x^2+x+1)^3\cdot$ $\cdot\,(x^4+1)\,(x^4+x^3+x^2+x+1)\,(x^6+x^5+x^4+x^3+x^2+x+1)$

Table 6.1: The minimal polynomials associated to the automata recognizing $L_{k,\Sigma,\lhd}$ in Figures 5.1, 5.2, 5.4, and 5.6.

$(k, \lvert\Sigma\rvert)$	Minimal polynomial associated to $L_{k,\Sigma,\mathrm{sing}}$
$(2,2)$	$(x-2)\,(x-1)^2\,x^4\,(x+1)$
$(2,3)$	$(x-3)\,(x-1)^3\,x^6\,(x+1)\,(x^2+x+1)$
$(3,2)$	$(x-2)\,(x-1)^3\,x^9\,(x+1)\,(x^2+1)\,(x^2+x+1)$
$(4,2)^*$	$(x-2)\,(x-1)^4\,x^{18}\,(x+1)^2\,(x^2-x+1)\,(x^2+1)\,(x^2+x+1)^2\cdot$ $\cdot\,(x^4+1)\,(x^4+x^3+x^2+x+1)\,(x^6+x^5+x^4+x^3+x^2+x+1)$

Table 6.2: The first three rows are the minimal polynomials associated to the automata recognizing $L_{k,\Sigma,\mathrm{sing}}$ in Figures 5.1, 5.3, and 5.5, respectively. In the fourth row we have the minimal polynomial of the automaton recognizing $L_{k,\Sigma,\mathrm{sing}} \setminus b\Sigma^*$ (see the transition table in Appendix C).

Method I. We start by identifying the vector e_{init} corresponding to the initial state. We also construct the vector e_F corresponding to all accepting states. We then have that $\mathcal{P}_m^{(k)}(n) = e_{\mathrm{init}} M^n e_F^T$.

To calculate $\mathcal{P}_m^{(k)}(n)$ explicitly, the *Jordan decomposition* (see, e.g., [107]) $M = SJS^{-1}$ of M is computed. A closed form for J^n, $n \geqslant \mu_0$, can be easily computed, from which we obtain a closed form for $M^n = SJ^nS^{-1}$. (Note that the dimension of the largest *Jordan block* corresponding to the eigenvalue 0 of M equals μ_0.) Thus $e_{\mathrm{init}} M^n e_F^T$ may be computed.

Example 6.2. We illustrate this method to obtain an explicit expression for $\mathcal{P}_2^{(2)}(n)$. Consider the DFA of Figure 5.1 which recognizes the language $L_{2,\mathbb{B},\lhd}$. Let us then identify all the letters to obtain a unary NFA. Let then M be the adjacency matrix of this transformed NFA:

$$M = \begin{pmatrix} 0&1&1&0&0&0&0&0&0&0\\ 0&1&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&1&0&1&0&0\\ 0&0&0&0&1&0&1&0&0&0\\ 0&0&0&1&0&0&0&0&0&1\\ 0&0&0&0&0&0&1&0&0&1\\ 0&0&0&0&0&0&0&1&1&0\\ 0&0&0&0&0&0&0&0&0&2\\ 0&0&0&0&0&0&0&0&0&2 \end{pmatrix}. \tag{6.2}$$

Here we order the states as follows: the initial state corresponds to the first row and the sink state corresponds to the last row. The rest of the states are arranged using the radix ordering as in the following table.

state obtained by reading:	a	b	ab	bb	aba	bba	abb	$abba$
row corresponding to state:	2	3	4	5	6	7	8	9

The vector e_{init}, corresponding to the initial state, equals $(1, 0, \ldots, 0)$, and the vector e_F, corresponding to all accepting states, equals $(1, \ldots, 1, 0)$. We then compute the Jordan decomposition $M = SJS^{-1}$ of M:

$$J = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}, \text{ and}$$

$$S = \begin{pmatrix} 0 & 0 & 1 & 0 & -1 & 1 & 2 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 2 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

A closed form for J^n, $n \geqslant 2$, is straightforward to compute:

$$J^n = \begin{pmatrix} (-1)^n & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & n & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & n & \frac{n(n-1)}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & n & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^n \end{pmatrix},$$

from which we obtain a closed form for $M^n = SJ^nS^{-1}$. Finally, we have that, for $n \geqslant 2$, $e_{\text{init}} S J^n S^{-1} e_F^T = n^2 - n + 2$.

Method II. The construction of the Jordan decomposition of a matrix becomes quite time consuming computationally, which leads us to consider the method

of curve fitting. This is done by first computing the eigenvalues $\{\lambda\}$, the minimal polynomial $m(x)$ of M, and the multiplicities $\{\mu_\lambda\}$ of the eigenvalues $\{\lambda\}$ as roots of $m(x)$. In general, numerical approximation could be a problem when computing the eigenvalues. But for the small cases that we studied, all eigenvalues are simple algebraic numbers: 0, -1, $e^{\pm\frac{2i\pi}{3}}$, etc. The computations can thus be made completely formally. In theory, this can be done with arbitrary algebraic numbers. We then continue to compute the coefficients of the polynomials $p_\lambda(x)$ in the expression (6.1) for each $\lambda \in \mathrm{Eig}(M)$ by fitting a curve to the data points $\mathcal{P}_m^{(k)}(\mu_0), \ldots, \mathcal{P}_m^{(k)}(n_M)$, where $n_M = \mu_0 + \sum_{\lambda\in\mathrm{Eig}(M)\setminus 0} \mu_\lambda - 1 \leqslant \dim(A)$. These values should be computed separately. There are thus $\sum_{\lambda\in\mathrm{Eig}(M)\setminus 0} \mu_\lambda$ functions and we fit them using equally many points.

Let us be more precise. We first compute the minimal polynomial of the adjacency matrix M to obtain the eigenvalues and their multiplicities. We then construct a vector function $\vec{\mathbf{x}}(n)$ of dimension $\sum_{\lambda\in\mathrm{Eig}(M)\setminus 0} \mu_\lambda$. For each eigenvalue λ of M, $\vec{\mathbf{x}}(n)$ contains the components $n^i\lambda^n$, $i = 0, \ldots, \mu_\lambda - 1$. Let then $\vec{\mathbf{c}} = (\mathcal{P}_m^{(k)}(\mu_0 + i))_{i=0}^{\dim(\vec{\mathbf{x}}(n))-1}$. We then construct a $\dim(\vec{\mathbf{x}}(n))$ square matrix B, where the ith row equals $\vec{\mathbf{x}}(\mu_0 + i)$ for each $i = 0, \ldots, \dim(\vec{\mathbf{x}}(n)) - 1$. Let $\vec{\mathbf{y}}$ be a vector containing for each $n^i\lambda^n$ the corresponding (unknown) coefficient in the polynomial p_λ. We have the linear system $B\vec{\mathbf{y}} = \vec{\mathbf{c}}$. We may solve for $\vec{\mathbf{y}}$ since B is invertible. Thus, we may obtain a closed formula $\mathcal{P}_m^{(k)}(n) = (\vec{\mathbf{y}}, \vec{\mathbf{x}}(n))$, where $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ denotes the inner product of the vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$.

Example 6.3. We compute the polynomial of the matrix M defined in (6.2). We obtain $m(x) = (x-2)(x-1)^3x^2(x+1)$, from which we infer that $\mathrm{Eig}(M) = \{2, 1, 0, -1\}$ and that $\mu_2 = 1$, $\mu_1 = 3$, $\mu_0 = 2$, and $\mu_{-1} = 1$. We then construct a vector function $\vec{\mathbf{x}}(n) = (2^n, 1, n, n^2, (-1)^n)$ and the matrix B having as rows $\vec{\mathbf{x}}(n)$, $n = 2, \ldots, 6$:

$$B = \begin{pmatrix} 4 & 1 & 2 & 4 & 1 \\ 8 & 1 & 3 & 9 & -1 \\ 16 & 1 & 4 & 16 & 1 \\ 32 & 1 & 5 & 25 & -1 \\ 64 & 1 & 6 & 36 & 1 \end{pmatrix}.$$

The vector $\vec{\mathbf{c}} = (\mathcal{P}_2^{(2)}(n))_{n=2}^6$ is computed to be $(4, 8, 14, 22, 32)$. We may solve for $\vec{\mathbf{y}} = B^{-1}\vec{\mathbf{c}} = (0, 2, 1, 1, 0)$. Finally $\mathcal{P}_2^{(2)}(n) = (\vec{\mathbf{y}}, \vec{\mathbf{x}}(n)) = n^2 - n + 2$.

Using the described methods the following explicit formulae for $\mathcal{P}_m^{(k)}(n)$ and $\mathcal{S}_m^{(k)}(n)$ were computed in [19]. Here we represent the formulae in a slightly different manner to the expression in (6.1). In all these cases we see that each $\lambda \in \mathrm{Eig}(M)$ having non-zero coefficient polynomial is a root of unity. Thus the coefficients of the terms n^i (in the second summation in the expression (6.1)) may be viewed as periodic functions, where the period depends on the degree of the roots of unity $\lambda \in \mathrm{Eig}(M)$. We denote a periodic function f with period m by $f(n) = \langle a_0, \ldots, a_{m-1}\rangle_n$, where $a_i = f(i)$, $0 \leqslant i < m$. For example, let

$$f(n) = e^{-\pi i/3}(e^{2\pi i/3})^n + e^{\pi i/3}(e^{-2\pi i/3})^n.$$

Then f may be expressed as $f(n) = \langle 1, 1, -2\rangle_n$. We recall that, for a periodic function $f(n) = \langle a_0, \ldots, a_{m-1}\rangle_n$, we may express f as $f(n) = \sum_{\lambda^m=1} \alpha_\lambda\lambda^n$, where

the summation goes through all the mth roots of unity. The coefficients α_λ are determined by the values $a_0, \ldots, a_{m-1}$.

We are now in the position to state the explicit formulae.

Proposition 6.4 ([19])**.**

$$\begin{aligned}
&\textit{For all } n \geqslant 1,\ \mathcal{P}_2^{(2)}(n) = n^2 - n + 2;\\
&\textit{for all } n \geqslant 2,\ \mathcal{P}_2^{(3)}(n) = \tfrac{1}{18}n^4 - \tfrac{5}{18}n^3 + \tfrac{65}{36}n^2 - \tfrac{23}{6}n + \tfrac{1307}{216}\\
&\qquad\qquad - \tfrac{1}{8}\langle 1,-1\rangle_n + \tfrac{2}{27}\langle 1,1,-2\rangle_n;\ \textit{and}\\
&\textit{for all } n \geqslant 1,\ \mathcal{P}_3^{(2)}(n) = \tfrac{1}{960}n^6 + \tfrac{7}{320}n^5 + \tfrac{67}{384}n^4 - \tfrac{19}{32}n^3 + \tfrac{1457}{480}n^2\\
&\qquad\qquad - \tfrac{1569}{640}n + \tfrac{741}{256} - \tfrac{3}{128}\langle 1,-1\rangle_n \cdot n + \tfrac{27}{256}\langle 1,-1\rangle_n.
\end{aligned}$$

Proposition 6.5 ([19])**.**

$$\begin{aligned}
&\textit{For all } n \geqslant 4,\ \mathcal{S}_2^{(2)}(n) = 2n + 4;\\
&\textit{for all } n \geqslant 9,\ \mathcal{S}_2^{(3)}(n) = \tfrac{1}{2}n^2 + 16n - \tfrac{535}{12} + \tfrac{2}{3}\langle 2,-1,-1\rangle_n - \tfrac{3}{4}\langle 1,-1\rangle_n;\ \textit{and}\\
&\textit{for all } n \geqslant 6,\ \mathcal{S}_3^{(2)}(n) = 3n^2 + 27n - 63.
\end{aligned}$$

The formulae for $\mathcal{P}_2^{(2)}$ and $\mathcal{S}_2^{(2)}$ were proved in [49] and [55], respectively, using different methods.

The values given by the formulae for $\mathcal{P}_3^{(2)}$ and $\mathcal{P}_2^{(3)}$ obtained here coincide with values previously computed by Eero Harmaala ($n = 2, \ldots, 18$ and $n = 4, \ldots, 28$, respectively) (personal communication), and to values (see Appendix B) computed using an algorithm suggested by Julien Cassaigne (see Appendix A). Similarly, the values given by the formulae obtained here for $\mathcal{S}_2^{(3)}$ and $\mathcal{S}_3^{(2)}$ coincide with the values (see Appendix B) computed using an algorithm suggested by Julien Cassaigne (see Appendix A).

We obtain formulae for $\mathcal{P}_2^{(4)}$ and $\mathcal{S}_2^{(4)}$.

Proposition 6.6. *For all $n \geqslant 3$ we have*

$$\begin{aligned}
\mathcal{P}_2^{(4)}(n) = {}& \tfrac{283}{512\cdot 243\cdot 25\cdot 49}n^8 + \tfrac{223}{32\cdot 243\cdot 25\cdot 49}n^7 + \tfrac{2\,657}{256\cdot 243\cdot 25\cdot 7}n^6\\
&- \tfrac{731}{8\cdot 243\cdot 25\cdot 7}n^5 + \tfrac{14\,111}{32\cdot 243\cdot 25}n^4 - \tfrac{1\,609}{4\cdot 27\cdot 25}n^3 + \tfrac{1\,850\,177\,503}{512\cdot 729\cdot 25\cdot 49}n^2\\
&- \tfrac{3\,779\,893}{64\cdot 729\cdot 7}n + \tfrac{81\,883\,529\,107}{1\,024\cdot 729\cdot 125\cdot 49}\\
&+ \tfrac{1}{512}\langle 1,-1\rangle_n \cdot n^2 - \tfrac{5}{64}\langle 1,-1\rangle_n \cdot n + \tfrac{489}{1\,024}\langle 1,-1\rangle_n\\
&+ \tfrac{1}{2\cdot 729}\langle -7,5,2\rangle_n \cdot n^2 + \tfrac{2}{729}\langle 38,-7,-31\rangle_n \cdot n\\
&+ \tfrac{1}{4\cdot 729}\langle -1\,853, 571, 1\,282\rangle_n + \tfrac{1}{16}\langle -2,1,2,-1\rangle_n\\
&+ \tfrac{2}{125}\langle 21,6,-4,-14,-9\rangle_n + \tfrac{1}{8}\langle -1,1,1,1,-1,-1\rangle_n\\
&+ \tfrac{4}{49}\langle 2,1,-1,-4,-1,1,2\rangle_n.
\end{aligned}$$

Proof. Let us denote the expression on the right hand side by $f(n)$. Observe that the periodic coefficients of each term n^i of $f(n)$ may be expressed as a linear combination of the eigenvalues of the adjacency matrix associated to the language $L_{4,\mathbb{B},\lhd}$ (see Table 6.1). Further, we have $f(n) = \mathcal{P}_2^{(4)}(n)$ for $n = 10, \ldots, 52$, which

were computed separately. (Notice that $\mu_0 = 10$). We have thus shown that $f(n) = \mathcal{P}_2^{(4)}(n)$ for $n \geqslant 10$, since $\sum_{\lambda \in \mathrm{Eig}(M)\setminus 0} \mu_\lambda = 43$. The values of $f(n)$ and $\mathcal{P}_2^{(4)}(n)$ also coincide for the values $n \geqslant 3$. This concludes the proof. □

Proposition 6.7. *For all $n \geqslant 18$ we have*

$$\begin{aligned} S_2^{(4)}(n) = &\tfrac{2}{27}n^3 + \tfrac{1823}{60}n^2 - \tfrac{28651}{270}n - \tfrac{665587}{504} \\ &+ \tfrac{2}{27}\langle 2,-1,-1\rangle_n \cdot n + \tfrac{1}{27}\langle -66,109,-43\rangle_n - \tfrac{25}{24}\langle 1,-1\rangle_n \\ &+ \tfrac{4}{7}\langle -1,-1,-1,-1,6,-1,-1\rangle_n + \tfrac{1}{3}\langle -4,-5,-1,4,5,1\rangle_n \\ &+ \tfrac{2}{5}\langle 5,0,-4,-2,1\rangle_n + \tfrac{1}{2}\langle 0,-1,0,1\rangle_n. \end{aligned}$$

Proof. Similar to the proof of the above proposition, we let $f(n)$ denote the expression on the right hand side. Observe that the number of 4-abelian singletons *not* beginning with b equals $\frac{1}{2}\mathcal{S}_2^{(4)}(n)$ for all $n \geqslant 1$. Indeed, a word u is a k-abelian singleton if and only if $E(u)$ is a k-abelian singleton, where E is the complementation morphism $a \leftrightarrow b$. Further $u = E(u)$ if and only if $u = \varepsilon$.

It thus suffices to show that $\mathcal{C}_L(n) = \frac{1}{2}f(n)$, where $L = L_{4,\mathbb{B},\mathrm{sing}} \setminus b\mathbb{B}^*$. This language is regular, and the minimal DFA recognizing it has 664 states (see Appendix C for the transition table). The corresponding minimal polynomial is given in the fourth row of Table 6.2. The periodic coefficients of the terms n^i in the expression of $\frac{1}{2}f(n)$ may be expressed using the roots of this polynomial. Further, we notice that $\mu_0 = 18$. Now the values $\mathcal{C}_L(n)$ and $\frac{1}{2}f(n)$ coincide for the values $n = 18, \ldots, 46$, where $46 = \mu_0 + \sum_{\lambda \in \mathrm{Eig}(M)} \mu_\lambda - 1$. We conclude that $\mathcal{S}_2^{(4)}(n) = f(n)$ for all $n \geqslant 18$. □

Proposition 6.6 was already conjectured in [19] and Proposition 6.7 was conjectured by Julien Cassaigne (personal communication). The expressions for $\mathcal{P}_2^{(4)}(n)$ and $\mathcal{S}_2^{(4)}(n)$ were computed by Julien Cassaigne using only the values $\mathcal{P}_2^{(4)}(n)$, $n = 4, \ldots, 49$, and $\mathcal{S}_2^{(4)}(n)$, $n = 18, \ldots, 100$, respectively. In particular, he did not use the constructed automata. These values were computed using the algorithm in Appendix A for independent implementations of this algorithm by Julien Cassaigne and the author. The values $\mathcal{P}_2^{(4)}(n)$, $n = 51, \ldots, 55$ were computed by the author.

Remark 6.8. The *On-Line Encyclopedia of Integer Sequences* (`http://oeis.org`, accessed November 5, 2018) contains the following sequences from Propositions 6.4 and 6.5: $(\mathcal{P}_2^{(2)}(n))_{n\geqslant 0}$ (shifted), $(\mathcal{P}_2^{(3)}(n))_{n\geqslant 0}$, $(\mathcal{P}_3^{(2)}(n))_{n\geqslant 0}$, and $(\mathcal{S}_2^{(2)}(n))_{n\geqslant 0}$ (shifted) as the sequences A014206, A289657, A289658, and A005843, respectively. The other sequences do not appear in the OEIS. We remark that the sequences were not indexed in the OEIS prior to our work.

6.2 On the Asymptotic Growth of Regular Languages

Observing Propositions 6.4 and 6.6, it seems that $\mathcal{P}_m^{(k)}(n) \sim C_{k,m} n^{m^{k-1}(m-1)}$ for some constant $C_{k,m}$ depending on k and m. In other words $\lim_{n\to\infty} \frac{\mathcal{P}_m^{(k)}(n)}{n^{m^{k-1}(m-1)}} =$

$C_{k,m}$. This turns out to be true for all $m, k \geqslant 1$, which we show in the following section. The proof relies on the connection of $\mathcal{P}_m^{(k)}(n)$ to regular languages, and in this section, we consider the slightly more general question: Given a regular language L having $\mathcal{C}_L(n)$ of order $\mathcal{O}(n^k)$ for some $k \geqslant 0$, when does $\mathcal{C}_L(n) \sim Cn^k$ for some constant C (depending on L)? We give a sufficient condition, regarding the structure of a DFA recognizing L, for this to happen.

We first consider certain kinds of polynomial regular languages defined by regular expressions of a very special form. We consider their generating functions, from which we deduce a condition for these kind of languages to have complexity asymptotic to a polynomial. Finally, we apply these considerations to general polynomial regular languages. This section is based on the article [19], in which the main results of this section appear.

6.2.1 The Asymptotic Complexities of Simple Polynomial Languages

Let us first set some terminology. The longest common prefix of the words x and y is denoted by $\operatorname{lcp}(x, y)$. We say that $\operatorname{lcp}(x, y)$ is *proper*, if $y \neq \operatorname{lcp}(x, y) \neq x$.

Definition 6.9. Let L be a regular language defined by a regular expression $z_0 y_1^* z_1 \cdots y_t^* z_t$, where, for each $i = 1, \ldots, t-1$, the longest common prefix of y_i and z_i is proper. Then L is called a *simple polynomial language of degree t*.

It is easy to show that a simple polynomial language L of order t has $\mathcal{C}_L(n)$ of order $\mathcal{O}(n^{t-1})$ but not of order $\mathcal{O}(n^{t-2})$. Indeed, it is straightforward to show that each word $u \in L$ has exactly one factorization of the form $z_0 y_1^{\alpha_1} z_1 \cdots y_t^{\alpha_t} z_t$ (this is done in the proof of Lemma 6.12). A similar counting argument to [102, Lemma 1] shows that $C_L(n) \neq \mathcal{O}(n^{t-2})$. The main result in this subsection is the following:

Proposition 6.10 ([19])**.** *Let L be a simple polynomial language having a regular expression as in Definition 6.9. Assume further that* $\gcd(|y_1|, |y_2|, \ldots, |y_t|) = 1$. *Then*

$$\mathcal{C}_L(n) \sim \frac{1}{\prod_{i=1}^{t} |y_i|} \frac{1}{(t-1)!} n^{t-1}.$$

We proceed to prove the above proposition by considering the generating functions of simple polynomial languages. Recall Subsection 2.3.2 for the basic notions and results regarding this notion. We give a general treatment of the asymptotic growth of simple polynomial languages, from which we deduce the above theorem.

Example 6.11. Let L_1 be a regular language defined by the regular expression $z_0 y_1^* z_1$ for some $z_0, z_1, y_1 \in \Sigma^*$, where $y_1 \neq \varepsilon$. It is readily verified that the generating function G_{L_1} of L_1 may be written as

$$G_{L_1}(x) = \sum_k x^{|z_0|+k|y_1|+|z_1|} = x^{|z_0 z_1|} \sum_k x^{k|y_1|} = \frac{x^{|z_0 z_1|}}{1 - x^{|y_1|}}.$$

Let L_2 be the language defined by the expression $y_2^* z_2$ for some $y_2, z_2 \in \Sigma^*$, where $y_2 \neq \varepsilon$. Then, similar to the above, $G_{L_2}(x) = \frac{x^{|z_2|}}{1-x^{|y_2|}}$. Assume further

that $\mathrm{lcp}(z_1, y_1)$ is proper. Consider then the generating function of the language $L = L_1 \cdot L_2$: Now L has the property that each $u \in L$ has a unique factorization of the form $u = u_1 u_2$, where $u_1 \in L_1$, $u_2 \in L_2$. Indeed, this follows from the assumption that $\ell cp(y_1, z_1)$ is proper (see the proof of the following lemma). It follows that $G_L(x)$ is the product of $G_{L_1}(x)$ and $G_{L_2}(x)$:

$$G_L(x) = G_{L_1}(x) \cdot G_{L_2}(x) = \frac{x^{|z_0 z_1 z_2|}}{(1-x^{|y_1|})(1-x^{|y_2|})}.$$

We generalize the above example.

Lemma 6.12. *Let L be a regular language defined by the regular expression $z_0 y_1^* z_1 \cdots y_t^* z_t$, where, for each $i = 1, \ldots, t-1$, $\mathrm{lcp}(y_i, z_i)$ is proper. Then*

$$G_L(x) = x^z \prod_{i=1}^{t} \frac{1}{1 - x^{|y_i|}}, \text{ where } z = |z_0 z_1 \cdots z_t|. \tag{6.3}$$

Proof. The case $t = 1$ was handled in the above example. Assume that the claim is true for t and consider the case of $t+1$. Let L_1 and L_2 be the languages defined by the expressions $z_0 y_1^* z_1 \cdots y_t^* z_t$ and $y_{t+1}^* z_{t+1}$ respectively, so that $L = L_1 \cdot L_2$. We claim that

$$G_{L_1 \cdot L_2}(x) = G_{L_1}(x) \cdot G_{L_2}(x),$$

that is, each element of L has a unique factorization into a word of L_1 concatenated with a word of L_2. Suppose the contrary: there exist $i_r, j_r \in \mathbb{N}$, $r = 1, \ldots, t+1$, such that

$$z_0 y_1^{i_1} z_1 \cdots y_{t+1}^{i_{t+1}} z_{t+1} = z_0 y_1^{j_1} z_1 \cdots y_{t+1}^{j_{t+1}} z_{t+1}$$

and there exists a minimal index $\ell \geqslant 0$ such that $i_r \neq j_r$ (if no such index existed, the factorizations would be the same). Observe that $\ell \leqslant t$, as otherwise the lengths implied by the factorizations would not coincide. We may assume that $i_\ell > j_\ell$, from which it follows that

$$y_\ell^{i_\ell - j_\ell} z_\ell \cdots y_{t+1}^{i_{t+1}} z_{t+1} = z_\ell \cdots y_{t+1}^{j_{t+1}} z_{t+1}.$$

This is a contradiction, since $\mathrm{lcp}(y_\ell, z_\ell)$ is assumed to be proper. The claim now follows by the induction hypothesis, since

$$G_{L_1}(x) \cdot G_{L_2}(x) = x^z \prod_{i=1}^{t} \frac{1}{1 - x^{|y_i|}} \cdot \frac{x^{|z_{t+1}|}}{1 - x^{|y_{t+1}|}} = x^{z + |z_{t+1}|} \prod_{i=1}^{t+1} \frac{1}{1 - x^{|y_i|}},$$

where $z = |z_0 z_1 \cdots z_t|$. □

For the rest of this subsection we fix L to be a language as in Lemma 6.12 so that the generating function $G_L(x)$ of L has the rational expression (6.3). Writing $G_L(x) = \sum a_k x^k$ as a formal power series, we consider the asymptotic behavior of the coefficients a_k, for $k \geqslant 0$, by performing certain manipulations to the rational expression of $G_L(x)$.

For the rest of the subsection, let $R_L(x) = x^z \prod_{i=1}^{t} \frac{1}{1-x^{|y_i|}}$ be the rational function defined by expression (6.3) of $G_L(x)$. Further, let $q_L(x) = \prod_{i=1}^{t} 1 - x^{|y_i|}$

denote the denominator of $R_L(x)$. Assume further that $q_L(x)$ has d distinct roots $\lambda_1, \ldots, \lambda_d$ for some $d \geqslant 1$. (Observe that each λ_i is a root of unity.)

Now $q_L(x)$ has the decomposition $q_L(x) = \prod_{i=1}^{d}(x-\lambda_i)^{m_i}$, where $\lambda_i \neq \lambda_j$ for $i \neq j$, and $m_i \geqslant 1$ for each $i = 1, \ldots, d$. We may express $R_L(x)$ using the partial fraction decomposition

$$R_L(x) = r_0(x) + \sum_{i=1}^{d}\sum_{j=1}^{m_i} \frac{C_{ij}}{(\lambda_i - x)^j}, \tag{6.4}$$

where $r_0(x)$ is a polynomial of degree $\deg p - \deg q$ if this number is non-negative, or otherwise $r_0(x) = 0$, and C_{ij} are constants for each $i = 1, \ldots, d$ and $j = 1, \ldots, m_i$. We may now express $G_L(x)$ as a sum of formal power series by recalling the expression for the *negative binomial series* $\frac{1}{(\lambda-x)^t} = \sum_k \binom{k+t-1}{t-1}\lambda^{-t-k}x^k$ in (6.4):

$$G_L(x) = \sum_k a_k x^k = r_0(x) + \sum_{i=1}^{d}\sum_{j=1}^{m_i} C_{ij} \sum_k \binom{k+j-1}{j-1}\lambda_i^{-j-k}x^k. \tag{6.5}$$

It is clear that one of these roots is 1; we set $\lambda_1 = 1$. Let also $m = \max\{m_i\}$, that is, m is the maximal multiplicity of the roots of $q_L(x)$. In the following, we call λ a *maximal root* of $q_L(x)$ if the multiplicity of λ as a root of $q_L(x)$ equals m. We shall now consider the asymptotic behaviour of the coefficients a_k in $G_L(x) = \sum_k a_k x^k$ using (6.5).

Lemma 6.13. *For all $k > \deg r_0$, where r_0 is as in the formal power series (6.5),*

$$a_k = \sum_{i=1}^{d} \tfrac{C_i}{(m-1)!}\lambda_i^{-m-k}k^{m-1} + \mathcal{O}(k^{m-2}),$$

where $C_i = C_{im}$ as defined in (6.4) if λ_i is a maximal root of $q_L(x)$, otherwise $C_i = 0$.

Proof. As discussed above, each root λ_i of $q_L(x)$ is a root of unity whence the values λ_i^k, for $k \in \mathbb{Z}$, are uniformly bounded for each i. Further, we note that $\binom{k+j-1}{j-1} = \frac{1}{(j-1)!}(k+1)\cdots(k+j-1) = \frac{1}{(j-1)!}k^{j-1} + \mathcal{O}(k^{j-2})$. Let then $k > \deg r_0$. By the equality (6.5),

$$\begin{aligned}
a_k &= \sum_{i=1}^{d}\sum_{j=1}^{m_i} C_{ij}\binom{k+j-1}{j-1}\lambda_i^{-j-k} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m_i} \Big(\tfrac{C_{ij}}{(j-1)!}k^{j-1} + \mathcal{O}(k^{j-2})\Big)\lambda_i^{-j-k} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m_i} \Big(\tfrac{C_{ij}}{(j-1)!}\lambda_i^{-j-k}k^{j-1} + \mathcal{O}(k^{j-2})\Big) \\
&= \sum_{i=1}^{d} \Big(\tfrac{C_{im_i}}{(m_i-1)!}\lambda_i^{-m_i-k}k^{m_i-1} + \mathcal{O}(k^{m_i-2})\Big) \\
&= \sum_{i=1}^{d} \tfrac{C_i}{(m-1)!}\lambda_i^{-m-k}k^{m-1} + \mathcal{O}(k^{m-2}).
\end{aligned}$$

Here we employ the facts that the values λ_i^k, for $k \in \mathbb{Z}$, are uniformly bounded, and that $m_i \leqslant m$ for each $i = 1, \ldots, d$. □

For the polynomial $q_L(x)$ we see that the root $\lambda_1 = 1$ has multiplicity t. Since each polynomial of the form $1-x^r$, where $r \geqslant 1$, has r distinct roots, it follows that the maximum multiplicity m of a root of the polynomial $q_L(x) = \prod_{i=1}^{t}(1 - x^{|y_i|})$ equals t. It thus follows that 1 is a maximal root of $q_L(x)$. In particular, in the partial fraction decomposition (6.4) of $R_L(x)$, we have $m_i \leqslant t$ for each i, and $m_1 = m = t$, where $\lambda_1 = 1$. Furthermore, $C_1 = C_{1t}$ in the above lemma. Let us compute the exact value of C_1.

Lemma 6.14. *Let C_1 be as defined above. Then $C_1 = \prod_{i=1}^{t} \frac{1}{|y_i|}$.*

Proof. Let $H(x)$ be defined by $q_L(x) = (1-x)^t H(x)$, that is,

$$H(x) = \frac{q_L(x)}{(1-x)^t} = \prod_{i=1}^{t} \sum_{j=0}^{|y_i|-1} x^j.$$

Note that $H(1) = \prod_{i=1}^{t} |y_i|$. By combining all other terms in (6.4), we may express $R_L(x)$ as

$$R_L(x) = \frac{x^z}{q_L(x)} = \frac{C_1}{(1-x)^t} + \frac{P(x)}{(1-x)^{t-1}H(x)} = \frac{C_1 H(x) + (1-x)P(x)}{q_L(x)},$$

where $P(x)$ is the polynomial defined by the above equality. This implies that $C_1 H(x) + (1-x)P(x) = x^z$. Evaluating both sides at $x = 1$ yields $C_1 = 1/H(1)$. The claim follows. □

We are in the position to prove the main result of this subsection.

Proof of Proposition 6.10. We show that the root $\lambda_1 = 1$ of $q_L(x)$ is the unique maximal root of $q_L(x)$. The claim then follows by Lemma 6.13 together with the above lemma.

First note that an mth root of unity λ, that is, $\lambda^m = 1$, is a root of the polynomial $1 - x^r$ if and only if m divides r. We already observed that $\lambda_1 = 1$ is a maximal root of $q_L(x)$, and has multiplicity t. Suppose then that an mth root of unity $\lambda \neq 1$ is also a maximal root of $q_L(x)$. It follows that λ is a root of each of the polynomials $1 - x^{|y_i|}$, $i = 1, \ldots, t$, whence m divides $|y_i|$ for each $i = 1, \ldots, t$. This is a contradiction, since it would follow that $\gcd(|y_1|, |y_2|, \ldots, |y_t|) \geqslant m > 1$. Thus λ has multiplicity less than t and λ_1 is the unique maximal root of $q_L(x)$. □

6.2.2 A Sufficient Condition for Languages Having Growth Asymptotic to a Polynomial

Let us then consider the general case of a regular language L having $\mathcal{C}_L(n) = \mathcal{O}(n^k)$ for some $k \geqslant 0$. Consider any DFA $\mathcal{A}$ recognizing L and any walk $\mathcal{W}$ in the underlying multigraph of $\mathcal{A}$ starting from the initial state and ending in some

accepting state. Any two distinct cycles[1] occurring along $\mathcal{W}$ must be vertex-disjoint, as otherwise $\mathcal{C}_L(n)$ is not of the order $\mathcal{O}(n^k)$ for any $k \geqslant 0$ (see [102]).[2] In particular, the only strongly connected components in the underlying graph of $\mathcal{A}$ are simple cycles (after removing non-reachable states and states from which no accepting state is reachable). For example, see the automata in Figures 5.1–5.5. Thus any such walk $\mathcal{W}$ is cycle-deterministic. We may write

$$\mathcal{W} = \mathcal{P}_0\mathcal{C}_1^{\alpha_1}\mathcal{P}_1 \cdots \mathcal{C}_t^{\alpha_t}\mathcal{P}_t, \tag{6.6}$$

where $\mathcal{P}_0\mathcal{P}_1 \cdots \mathcal{P}_t = \mathcal{P}$ is a path, $\mathcal{C}_i$ is a cycle and $\alpha_i \geqslant 1$ for each $i = 1, \ldots, t$. Further, $\mathcal{W}$ enters the cycle $\mathcal{C}_i$ at position $|\mathcal{P}_0\mathcal{C}_1^{\alpha_1}\mathcal{P}_1 \cdots \mathcal{P}_{i-1}| + 1$ for each i. Not only is $\mathcal{W}$ cycle-deterministic, but it has an even stronger property: once $\mathcal{W}$ leaves the cycle $\mathcal{C}_i$, none of the previously visited vertices are visited afterwards.

Definition 6.15. Let $\mathcal{W}$ be a walk in the underlying graph of the DFA $\mathcal{A}$, such that tail($\mathcal{W}$) is the initial state of $\mathcal{A}$, and head($\mathcal{W}$) is an accepting state of $\mathcal{A}$. Then $\mathcal{W}$ defines the sub-automaton $\mathcal{A}_\mathcal{W}$ of $\mathcal{A}$, called a *walk-automaton* (defined by $\mathcal{W}$) of $\mathcal{A}$, as follows. The initial state of $\mathcal{A}_\mathcal{W}$ is the initial state of $\mathcal{A}$, and the only accepting state is head($\mathcal{W}$). The states of $\mathcal{A}_\mathcal{W}$ are the vertices occurring in $\mathcal{W}$ and the transitions are as in $\mathcal{A}$ restricted to the states in $V(\mathcal{W})$. If there is a transition leading out of $V(\mathcal{W})$ in $\mathcal{A}$, then this transition is directed to a sink state in $\mathcal{A}_\mathcal{W}$.

Remark 6.16. A word accepted by a walk-automaton of a DFA $\mathcal{A}$ is called *r-tiered* in [102]. Here r represents the number of cycles the computation of $\mathcal{A}$ enters.

Clearly the language recognized by a walk-automaton $\mathcal{A}_\mathcal{W}$ is a sublanguage of $L(\mathcal{A})$. Observe that several walks may define the same walk-automaton $\mathcal{B}$. Now if two walks $\mathcal{W}_1$ and $\mathcal{W}_2$ define the same walk-automaton $\mathcal{B}$, the corresponding sets of cycles occurring along the walks are equal: $\mathrm{Cyc}(\mathcal{W}_1) = \mathrm{Cyc}(\mathcal{W}_2)$. Thus, for a walk-automaton $\mathcal{B}$, we let $\mathrm{Cyc}(\mathcal{B})$ denote the set of cycles occurring along any walk defining $\mathcal{B}$. We call the walk-automaton $\mathcal{B}$ *saturated*, if the number of cycles $\#\mathrm{Cyc}(\mathcal{B})$ is maximal among all walk-automata of $\mathcal{A}$.

Example 6.17. Consider the automaton in Figure 5.2 recognizing the language $L_{3,\mathbb{B},\lhd}$. Let $\mathcal{W}$ be the walk defined by the word $w = aabaabaababaaba$. We may write $\mathcal{W} = \mathcal{P}_0\mathcal{C}_1^2\mathcal{P}_1\mathcal{C}_2\mathcal{P}_2$, where $\mathrm{label}(\mathcal{P}_0) = aab = \mathrm{label}(\mathcal{C}_1)$, $\mathrm{label}(\mathcal{P}_1) = ab = \mathrm{label}(\mathcal{C}_2)$, and $\mathrm{label}(\mathcal{P}_2) = a$. The walk-automaton $\mathcal{B}$ defined by $\mathcal{W}$ thus recognizes the language $aab(aab)^*ab(ab)^*a$. Observe that the walk $\mathcal{W}'$ defined by the word $w' = aabaababaaba$ also defines $\mathcal{B}$. Now $\mathrm{Cyc}(\mathcal{B}) = \{\mathcal{C}_1, \mathcal{C}_2\}$. On the other hand, the walk $\mathcal{W}''$ defined by the word $aaabaababaaba$ defines a different walk-automaton, which recognizes the language $aaa^*b(aab)^*ab(ab)^*a$.

The maximum number of cycles occurring along a walk in $\mathcal{A}$ is 5, as can be verified from the figure. (The cycles have emboldened edges for convenience). The union of the saturated walk-automata of $\mathcal{A}$ is defined by the regular expression

$$\begin{aligned}(aa + abaa + baa + bbaa) \cdot a^*b(aab)^* \cdot \\ \cdot \big(ab(ab)^*b(abb)^* + b(aabb)^*ab(bab)^*b\big) \cdot bb^*(\varepsilon + a + aa + ab),\end{aligned}$$

as can be verified by carefully inspecting Figure 5.2.

[1] Two cycles of a multigraph are distinct if the sets of (labeled) edges are distinct.

[2] If two cycles shared a common vertex, L would contain a language of the form $x(y_1 + y_2)^*z$ for some words $x, y_1, y_2, z \in \Sigma^*$, $y_1 \neq \varepsilon \neq y_2$.

For the remainder of this section we let L be a regular language having polynomial complexity, and let $\mathcal{A}$ be a DFA recognizing L.

Let now $\mathcal{A}_\mathcal{W}$ be a walk-automaton of $\mathcal{A}$ defined by the walk $\mathcal{W}$ of the form (6.6). By the structure of $\mathcal{A}$, $L(\mathcal{A}_\mathcal{W})$ is defined by the regular expression $z_0 y_1^* z_1 \cdots y_t^* z_t$, where

- The word z_0 is the label of the path $\mathcal{P}_0$ from $\mathrm{tail}(\mathcal{W})$ to $\mathrm{tail}(\mathcal{C}_1)$;
- The word y_i is the label of $\mathcal{C}_i$ for each $i = 1, \ldots, t$;
- The word z_i is the label of the path $\mathcal{P}_i$ from $\mathrm{tail}(\mathcal{C}_i)$ to $\mathrm{tail}(\mathcal{C}_{i+1})$ for each $i = 1, \ldots, t-1$;
- The word z_t is the label of the path $\mathcal{P}_t$ from $\mathrm{tail}(\mathcal{C}_t)$ to f.

Moreover, since $\mathcal{A}$ is deterministic, the longest common prefix $\mathrm{lcp}(y_i, z_i)$ of y_i and z_i is proper for each $i = 1, \ldots, t-1$. In particular, $z_i \in \Sigma^+$ for each $i = 1, \ldots, t-1$. It follows that $L(\mathcal{A}_\mathcal{W})$ is a simple polynomial language. For a walk $\mathcal{W}$ we call the above regular expression the *canonical* expression defined by $\mathcal{W}$.

Lemma 6.18. *Let $\mathcal{W}$ and $\mathcal{W}'$ define distinct saturated walk-automata of $\mathcal{A}$. Then $L(\mathcal{A}_\mathcal{W}) \cap L(\mathcal{A}_{\mathcal{W}'}) = \emptyset$.*

Proof. Assume, for a contradiction, that both automata accept the same word w. Without loss of generality, we may take

$$\mathcal{W} = \mathcal{P}_0 \mathcal{C}_1 \mathcal{P}_1 \cdots \mathcal{C}_t \mathcal{P}_t \quad \text{and} \quad \mathcal{W}' = \mathcal{P}'_0 \mathcal{C}'_1 \mathcal{P}'_1 \cdots \mathcal{C}'_t \mathcal{P}'_t.$$

Let further $z_0 y_1^* z_1 \cdots y_t^* z_t$ and $u_0 v_1^* u_1 \cdots v_t^* u_t$ be the canonical expressions defined by $\mathcal{W}$ and $\mathcal{W}'$ respectively. Now, for some α_i, $\beta_i \geqslant 0$, we have

$$w = z_0 y_1^{\alpha_1} z_1 \cdots y_t^{\alpha_t} z_t = u_0 v_1^{\beta_1} u_1 \cdots v_t^{\beta_t} u_t.$$

Let $i \geqslant 1$ be the minimum index for which $z_{i-1} \neq u_{i-1}$ or $y_i \neq v_i$. Such an index exists, since the walks $\mathcal{W}$ and $\mathcal{W}'$ define distinct walk-automata and $\mathcal{A}$ is deterministic. It follows that $\alpha_j = \beta_j$ for $j < i$ since the longest common prefix of y_j and z_j is proper. Thus $z_{i-1} y_i^{\alpha_i} z_i \cdots y_t^{\alpha_t} z_t = u_{i-1} v_i^{\beta_i} u_i \cdots v_t^{\beta_t} u_t$. If $z_{i-1} = u_{i-1}$, then $y_i \neq v_i$. It follows that two distinct cycles start from the state $\mathrm{tail}(\mathcal{C}_i) = \mathrm{tail}(\mathcal{C}'_i)$ in $\mathcal{A}$, which is not possible. Thus $z_{i-1} \neq u_{i-1}$ and, without loss of generality, we may assume $|z_{i-1}| \geqslant |u_{i-1}|$. We deduce that u_{i-1} is a proper prefix of z_{i-1}. Let $\mathcal{P}_{i-1} = \mathcal{P}'_{i-1} \mathcal{P}''$, and consider the walk

$$\mathcal{W}'' = \mathcal{P}_0 \mathcal{C}_1 \mathcal{P}_1 \cdots \mathcal{C}_{i-2} \mathcal{P}_{i-2} \mathcal{C}_{i-1} \mathcal{P}'_{i-1} \mathcal{C}'_i \mathcal{P}'' \mathcal{C}_i \mathcal{P}_i \cdots \mathcal{C}_t \mathcal{P}_t.$$

It is a well-defined walk in $\mathcal{A}$, and thus defines some walk-automaton $\mathcal{B}''$. But now $\#\mathrm{Cyc}(\mathcal{B}'') = t + 1$, which contradicts the assumption that $\mathcal{A}_\mathcal{W}$ is saturated. This concludes the proof. □

Theorem 6.19. *Let L be a regular language with $\mathcal{C}_L(n)$ of order $\mathcal{O}(n^k)$, but not of order $\mathcal{O}(n^{k-1})$ for some $k \geqslant 0$. Let $\mathcal{A}$ be a DFA recognizing L. Assume that for each $\mathcal{B} \in Sat(\mathcal{A})$, we have $\gcd\{|\mathcal{C}| \mid \mathcal{C} \in Cyc(\mathcal{B})\} = 1$. Then $\mathcal{C}_L(n) \sim Dn^k$, where*

$$D = \frac{1}{k!} \sum_{\mathcal{B} \in Sat(\mathcal{A})} \prod_{\mathcal{C} \in Cyc(\mathcal{B})} \frac{1}{|\mathcal{C}|}.$$

Proof. Let $\mathcal{B}$ be any walk-automaton of $\mathcal{A}$. Then $\mathcal{B}$ recognizes a simple polynomial language K having complexity $\mathcal{C}_K(n) = \mathcal{O}(n^{\mathrm{Cyc}(\mathcal{B})-1})$ by Lemma 6.13. Let $M = \#\mathrm{Cyc}(\mathcal{B})$ for a saturated walk-automaton $\mathcal{B}$. Since $\mathcal{A}$ is the union of all its walk-automata, we deduce that $\mathcal{C}_L(n) = \mathcal{O}(n^{M-1})$, in particular, $M \geqslant k+1$. On the other hand, since $C_{L(\mathcal{B})}(n) \sim \alpha n^{M-1}$ for some α by Proposition 6.10, it follows that $M = k+1$. Let $L' = \cup_{\mathcal{B}\in\mathrm{Sat}(\mathcal{A})} L(\mathcal{B})$. We have that $\mathcal{C}_L(n) = \mathcal{C}_{L'}(n) + \mathcal{O}(n^{k-1})$, since all other walk-automata contribute at most $\mathcal{O}(n^{k-1})$ words. Since, by the above lemma, $L(\mathcal{B}) \cap L(\mathcal{B}') = \emptyset$ for distinct walk-automata $\mathcal{B}$, $\mathcal{B}' \in \mathrm{Sat}(\mathcal{A})$, it follows that $\mathcal{C}_{L'}(n) \sim \left(\sum_{\mathcal{B}\in\mathrm{Sat}(\mathcal{A})} \frac{1}{k!} \prod_{\mathfrak{C}\in\mathrm{Cyc}(\mathcal{B})} \frac{1}{|\mathfrak{C}|}\right) n^k$ by Proposition 6.10. The claim follows. □

The above result is crucial in our analysis of k-abelian equivalence classes in the following sections. Before moving towards the analysis of $L_{k,\Sigma,\lhd}$ we give a clarifying example concerning the notions discussed above.

Example 6.20. Recall the automaton $\mathcal{A}$ in Figure 5.2. In Example 6.17 a regular expression was obtained for the union L' of the languages recognized by the saturated walk-automata of $\mathcal{A}$. Since these languages are disjoint, the generating function of L equals the sum of the generating functions of these languages. We thus obtain the rational expression for $G_L(x)$:

$$\begin{aligned}
&(x^2+x^3+2x^4)\cdot\tfrac{1}{1-x}\cdot x\cdot\tfrac{1}{1-x^3}\cdot \\
&\qquad\cdot\left(x^2\tfrac{1}{1-x^2}x\tfrac{1}{1-x^3}+x\tfrac{1}{1-x^4}x^2\tfrac{1}{1-x^3}x\right)\cdot x\tfrac{1}{1-x}\cdot(1+x+2x^2)\\
&= \frac{x^7(1+x+2x^2)^2(1+x+x^2)}{(1-x)^5(1+x)(1+x^2)(1+x+x^2)^2}\\
&= \frac{4}{3}\frac{1}{(1-x)^5} - \frac{12x^{10}+\mathcal{O}(x^9)}{3(1-x)^4(1+x)\,(1+x^2)\,(1+x+x^2)}.
\end{aligned}$$

In the corresponding generating function $\sum a_k x^k$, the dominating term of the coefficient a_k is contributed from $\frac{4}{3}\frac{1}{(1-x)^5} = \frac{4}{3}\frac{1}{4!}\sum_k^\infty (k+4)(k+3)(k+2)(k+1)x^k$. Thus $a_k = \frac{4}{3}(\frac{k^4}{4!}+\mathcal{O}(k^3)) = \frac{1}{18}k^4+\mathcal{O}(k^3)$. Another way to find the coefficient of k^4 is by inspecting the lengths of the cycles of saturated automata by Theorem 6.19. We again find $\frac{1}{4!}16(\frac{1}{1\cdot3\cdot2\cdot3\cdot1}+\frac{1}{1\cdot3\cdot4\cdot3\cdot1}) = \frac{1}{18}$. This coincides with the coefficient computed in Proposition 6.4, as expected.

Remark 6.21. Assume that a regular language L has $\mathcal{C}_L(n)$ of order $\mathcal{O}(n^t)$ but not of order $\mathcal{O}(n^{t-1})$. It is not hard to see that then $\mathcal{C}_L(n) = \langle a_0,\ldots,a_m\rangle_n n^t + \mathcal{O}(n^{t-1})$, where $a_0,\ldots,a_m$ are some non-negative constants and at least one of them is positive. (For example, this can be seen by Lemma 6.13 and Lemma 6.18). If, further, L is factor closed, then each a_i is positive, so that $\mathcal{C}_L(n) = \Theta(n^t)$. Indeed, in this case $\mathcal{C}_L(n) \leqslant \mathcal{C}_L(n+1) \leqslant |\Sigma|\mathcal{C}_L(n)$. This does not imply that $\mathcal{C}_L(n) \sim Cn^t$ for some constant C as shown by the following example.

Example 6.22. Let a_0, a_i, b_i, c_i be distinct letters for each $i = 1,\ldots,t$. Let $K = a_0(b_1c_1)^*a_1\cdots(b_tc_t)^*a_t$ be a simple polynomial language of degree t. Let further L be the factor closure $\mathrm{Fact}(K)$ of K. We claim that $\mathcal{C}_L(n) = \Theta(n^{k-1})$ but $\mathcal{C}_L(n) \neq C'n^{t-1}$ for any constant C'. Let now $K' = \mathrm{Fact}(a_0^{-1}L)\cup\mathrm{Fact}(La_t^{-1})$. Note that K' is also factor closed as the union of factor closed languages. We see

that $K = K' \cup L$ and that this union is disjoint, as none of the words of K' both begin with a_0 and end with a_t, while each word in L does so. It is not hard to convince oneself that $\mathcal{C}_{K'}(n)$ has complexity $\Theta(n^{t-1})$. For example, after some consideration, we see that

$$\begin{aligned} K' = &(\varepsilon + c_1)(b_1c_1)^*a_1 \cdots (b_tc_t)^*(\varepsilon + b_t + a_t) \\ &\cup (\varepsilon + a_0 + c_1)(b_1c_1)^*a_1 \cdots (b_tc_t)^*(\varepsilon + b_t) \\ &\cup \mathrm{Fact}(a_1(b_2c_2)^*a_2 \cdots (b_{t-1}c_{t-1})^*a_{t-1}), \end{aligned}$$

so, by induction, K' is a finite union of simple polynomial languages of degree at most t.

Assume that $\mathcal{C}_{K'}(n) \sim C'n^{t-1}$. Now $\mathcal{C}_K(t + 2m) = 0$ for all $m \geqslant 0$. On the other hand, $\mathcal{C}_K(t + 1 + 2m) = \binom{m+t-1}{t-1} \sim \frac{1}{(t-1)!}m^{t-1}$, so that $\mathcal{C}_K(n) = \langle 0, \frac{1}{2^{t-1}(t-1)!}\rangle_{n+t}n^{t-1} + \mathcal{O}(n^{t-2})$. But now $\mathcal{C}_L(n) = \langle C, C + \frac{1}{2^{t-1}(t-1)!}\rangle_{n+t}n^{t-1} + \mathcal{O}(n^{t-2})$ so that $\mathcal{C}_L(n) \neq C'n^{t-1}$ for any constant C'.

6.3 The Asymptotic Number of k-Abelian Equivalence Classes

In this section we show that $\mathcal{P}_m^{(k)}(n)$ is asymptotic to a polynomial. This is a sharpening of Theorem 1.1, which states that $\mathcal{P}_m^{(k)}(n) = \Theta(n^{m^{k-1}(m-1)})$ for all $m, k \geqslant 1$ (the constants implied by Θ depend on m and k). The considerations of this section appear in the article [19].

Theorem 6.23. *For all $k, m \geqslant 1$ we have $\mathcal{P}_m^{(k)}(n) \sim C_{k,m}n^{m^{k-1}(m-1)}$ for some rational constant $C_{k,m}$ depending on k and m.*

We prove this theorem by showing that the minimal DFA recognizing $L_{k,\Sigma,\lhd}$, where $|\Sigma| = m$, satisfies the assumptions of Theorem 6.19. Namely we aim to show that each saturated walk-automaton of the minimal DFA contains a cycle of length 1, that is, a loop.

We begin with a connection between cycles in walk-automata of a DFA recognizing $L_{k,\Sigma,\lhd}$, and cycles in the de Bruijn graph.

Let $\mathcal{A}$ be a DFA recognizing $L_{k,\Sigma,\lhd}$ and let $\mathcal{B}$ be a walk-automaton of $\mathcal{A}$. Assume $\mathcal{B}$ has the canonical expression $z_0y_1^*z_1 \cdots y_t^*z_t$. Take $u = z_0y_1^{\alpha_1}z_1 \cdots y_t^{\alpha_t}z_t$, where α_i is the minimal integer such that $\alpha_i|y|_i \geqslant k - 1 + 2|y_i|$. We have $y_i^{\alpha_i} = p_iy_i'y_iy_i$, where $p_iy_i' = y^{\alpha_i-2}$, $|p_i| = k - 1$, and y_i' is a proper suffix of y_i. Now $p_iy_i'y_i^2 = y_ip_iy_i'y_i = y_i^2p_iy_i'$. Consider the walk W_i defined by $y_i^2p_i$ in $dB(k-1)$. We may write $W_i = P_i^2$, where $\mathrm{tail}(P_i) = \mathrm{head}(P_i) = p_i$ and $\mathrm{label}(P_i) = y_i'y_i'' = x_i$, where $y_i = y_i''y_i'$. We claim that P_i is a repetition of a cycle; $P_i = C_i^{r_i}$. Indeed, since $\mathrm{tail}(P_i) = \mathrm{head}(P_i)$, a cycle C occurs along P_i, that is, $P_i = XCX'$. Now $P_i^2 = XCX'XCX'$ and, since P_i is cycle-deterministic,[3] we have $X'X = C^r$. But then $XX' = D^r$, where D is the cycle C traversed starting from a different vertex, and thus $P_i^2 = D^{2r+2}$, that is, $P_i = D^{r+1}$. Thus P_i is a repetition of a cycle.

Using the above notation, we write the walk W_u defined by u in $dB(k-1)$ as

$$W_u = V_0C_1^{2r_1}V_1 \cdots C_t^{2r_t}V_t,$$

[3] Recall that the language $L_{k,\Sigma,\lhd}$ is factor closed.

where $\mathrm{tail}(W_u) = \mathrm{pref}_{k-1}(u)$, C_i is a cycle labeled by x_i with $\mathrm{tail}(C_i) = p_i$, V_0 is a walk from $\mathrm{pref}_{k-1}(u)$ to p_1 labeled by $u[k-1, k-1+|z_0|)$, V_i is a walk from p_i to p_{i+1} labeled by $y_i' z_i p_{i+1}$, and V_t is a walk from p_t to $\mathrm{suff}_{k-1}(v)$ labeled by $y_t' z_t$. Recall now that $\mathrm{lcp}(y_i, z_i)$ is proper. It follows that the walk $C_i^{r_i} V_i$ leaves the cycle C_i at some point before reaching the end. We conclude that $\mathrm{Cyc}(W_u) \geqslant \mathrm{Cyc}(\mathcal{B})$.

Assume now that $\mathcal{B}$ above is saturated. Since $\mathcal{P}_m^{(k)}(n) = \Theta(n^{m^{k-1}(m-1)})$, it follows that $\#\mathrm{Cyc}(\mathcal{B}) = m^{k-1}(m-1)+1$. By the above and Corollary 4.11, we have $\#\mathrm{Cyc}(\mathcal{B}) \leqslant \#\mathrm{Cyc}(W_u) \leqslant m^{k-1}(m-1)+1$ so that these numbers are equal. We conclude that there is a one-to-one correspondence between the labels of the cycles in $\mathrm{Cyc}(\mathcal{B})$ and the labels of the cycles in $\mathrm{Cyc}(W_u)$.

Proposition 6.24 ([19]). *Let $\mathcal{A}$ be the minimal DFA recognizing $L_{k,\Sigma,\lhd}$. Then each saturated walk-automaton $\mathcal{B}$ of $\mathcal{A}$ contains a cycle of length 1.*

Proof. Let $T = m^{k-1}(m-1)+1$ and let $\mathcal{B}$ be defined by the walk

$$\mathcal{W} = \mathcal{P}_0 \mathcal{C}_1 \mathcal{P}_1 \cdots \mathcal{C}_T \mathcal{P}_T,$$

where $\mathcal{P}_0 \cdots \mathcal{P}_T$ is a path, $\mathcal{C}_i$ is a cycle, and $\mathcal{W}$ enters the cycle at position $|\mathcal{P}_0 \mathcal{C}_1 \mathcal{P}_1 \cdots \mathcal{C}_{i-1} \mathcal{P}_{i-1}| + 1$. Let $z_0 y_1^* z_1 \cdots y_T^* z_T$ be the canonical expression of $\mathcal{B}$ and let $u = z_0 y_1^{\alpha_1} z_1 \cdots y_T^{\alpha_T} z_T$, where α_i is the minimal integer for which $\alpha_i |y_i| \geqslant k - 1 + 2|y_i|$. Then, letting W_u be the walk defined by u in $dB(k-1)$, we have $W_u = V_0 C_1^{2r_1} V_1 \cdots C_T^{2r_T} V_T$ for some cycles C_i, by the above discussion. Further, the walk $C_i^{2r_i} V_i$ leaves the cycle C_i before it ends. Observe that V_{i-1} could end with a repetition of C_i, and V_i can begin with a repetition of C_i. Let us rewrite

$$W_u = P_0 D_1^{\beta_1} P_1 \cdots D_T^{\beta_T},$$

where D_i is a cycle, $\beta_i \geqslant 2r_i$, P_0 is a path from $\mathrm{pref}_{k-1}(u)$ to $\mathrm{tail}(D_1)$, P_i is a path from $\mathrm{head}(D_i)$ to $\mathrm{tail}(D_{i+1})$ and W_u enters the cycle D_i at position $|P_0 D_1^{\beta_1} P_1 \cdots D_{i-1}^{\beta_{i-1}} P_{i-1}| + 1$.

Recall Proposition 4.10: Between two consecutive cycles occurring along W_u, there is an update in the sequence of extension histories of u which has not occurred along W_u before. Since there are $m^{k-1}(m-1)$ possible updates,[4] it follows that each of these updates occur along W_u and, moreover, after each update, W enters a cycle before another update occurs. Further, there is a cycle which occurs along W before the first update. Let $a \in \Sigma$. Now the vertex a^{k-1} occurs at some index j with $(a^{k-1}, a) \in \Delta_u^j$, so that W_u enters the cycle (a^{k-1}, a). We conclude that one of the cycles D_i equals this cycle. It follows that the label of the corresponding cycle $\mathcal{C}_i$ along the defining walk $\mathcal{W}$ of $\mathcal{B}$ equals a^{r_i}.

We claim that $r_i = 1$, that is, the cycle $\mathcal{C}_i$ is a loop. Let e be the first edge of $\mathcal{C}_i$. We show that $\mathrm{tail}(e)$ is equivalent to $\mathrm{head}(e)$, which implies that $\mathrm{tail}(e) = \mathrm{head}(e)$ by Theorem 2.16, since $\mathcal{A}$ is assumed to be the minimal DFA recognizing $L_{k,\Sigma,\lhd}$.

Let $v_0 = z_0 z_1 \cdots z_{i-1} a^{r_i \alpha_i}$. Notice that since $\mathcal{A}$ is deterministic, after reading v_0 (resp., $v_0 a$), $\mathcal{A}$ reaches the state $\mathrm{tail}(e)$ (resp., $\mathrm{head}(e)$) when starting from the initial state. Now the walks W_{v_0} and $W_{v_0 a}$ in $dB(k-1)$ are of the form $X(a^{k-1}, a)^2$ and $X(a^{k-1}, a)^3$, respectively. Further, for any word $v \in \Sigma^*$ the

[4] Recall that the elements of Δ_u^1 are not counted as updates.

corresponding walks W_{v_0v} and W_{v_0av} in $dB(k-1)$ are of the form $X(a^{k-1},a)^2X'$ and $X(a^{k-1},a)^3X'$. By Lemma 4.12, we have $v_0v \in L_{k,\Sigma,\lhd}$ if and only if $v_0av \in L_{k,\Sigma,\lhd}$, which implies that $v \in L_{\mathrm{tail}(e)}$ if and only if $v \in L_{\mathrm{head}(e)}$, that is, $\mathrm{tail}(e)$ and $\mathrm{head}(e)$ are equivalent. This concludes the proof. □

Theorem 6.23 follows immediately from the above proposition by Theorem 6.19.

6.4 On the Asymptotic Number of k-abelian Singletons

An immediate consequence of Theorem 5.6 is the following:

Theorem 6.25. *For k and m fixed, we have $\mathcal{S}_{k,m}(n) = \mathcal{O}(n^{N_m(k-1)-1})$, where the constants implied by $\mathcal{O}$ depend on k and m.*

Recall that $N_m(\ell)$ is the number of necklaces of length ℓ over an m-letter alphabet. In this section we show that for each $m, k \geqslant 1$ there exists an integer $s_{k,m} \leqslant N_m(k-1) - 1$ such that $\mathcal{S}_m^{(k)}(n) = \Theta(n^{s_{k,m}})$. We give a characterization of this s in terms of cycle semi-decompositions of the de Bruijn graph of order $k-1$. We show for some small values of k and m that $s_{k,m} = N_m(k-1) - 1$. We actually conjecture that $s_{k,m} = N_m(k-1) - 1$ for any $k, m \geqslant 1$. We discuss some supporting evidence and tie, in some cases of m and k, this conjecture to a conjecture on so-called *Gray codes* for necklaces stated in the literature (see Conjecture 6.35). We show that whenever $s_{k,m} = N_m(k-1) - 1$, then, in fact, we have $\mathcal{S}_m^{(k)}(n) \sim D_{k,m} n^{N_m(k-1)-1}$ for some constant $D_{k,m}$ depending on k and m.

Our approach is similar to the approach to Theorem 6.23.

Theorem 6.26. *There exists an integer $s_{k,m}$ such that $\mathcal{S}_m^{(k)}(n) = \Theta(n^{s_{k,m}})$.*

Proof. Let $\mathcal{A}$ be a DFA recognizing $L_{k,\Sigma,\mathrm{sing}}$, where $|\Sigma| = m$, and let $\mathcal{B}$ be a saturated walk-automaton of $\mathcal{A}$. It follows that $L_{k,\Sigma,\mathrm{sing}}$ may be expressed as a finite union of expressions of the form $z_0 y_1^* z_1 \cdots y_t^* z_t$ where $t \leqslant \#\mathrm{Cyc}(\mathcal{B})$, whence $\mathcal{S}_m^{(k)}(n) = \mathcal{O}(n^{\#\mathrm{Cyc}(\mathcal{B})-1})$. Now $\mathcal{C}_{L(\mathcal{B})}(n)$ is not of the order $\mathcal{O}(n^{\#\mathrm{Cyc}(\mathcal{B})-2})$ which implies that $\mathcal{S}_m^{(k)}(n)$ is not of the order $\mathcal{O}(n^{\#\mathrm{Cyc}(\mathcal{B})-2})$. Finally, since $L_{k,\Sigma,\mathrm{sing}}$ is factorial by Remark 4.15, we conclude that $\mathcal{S}_m^{(k)}(n) = \Theta(n^{\#\mathrm{Cyc}(\mathcal{B})-1})$ by Remark 6.21. Thus $s_{k,m} + 1$ equals the number of cycles in a saturated walk-automaton of any DFA recognizing $L_{k,\Sigma,\lhd}$. □

The above theorem is not explicitly stated in [55], though it can be deduced from the discussion in the paper.

The following characterization of $s_{k,m}$ in terms of cycle semi-decompositions of the de Bruijn graph occurs in [55]. The proof given here is different, though in the same spirit.

Proposition 6.27. *Let $k, m \geqslant 1$. Then $s_{k,m} + 1$ equals the maximal number s such that there exists a set $\mathcal{C}$ of cardinality s of vertex-disjoint cycles of $dB(k-1)$ for which $dB(k-1)/\mathcal{C}$ contains a walk which traverses each vertex corresponding to a cycle precisely once and each vertex of $dB(k-1) \setminus V(\mathcal{C})$ at most twice.*

Proof. Let u be a k-abelian singleton and let W_u be the corresponding word in $dB(k-1)$. In the alternative proof of Theorem 5.1, it is shown that such a walk induces a polynomial language $L \subseteq L_{k,\Sigma,\text{sing}}$ having $\mathcal{C}_L(n) = \mathcal{O}(n^{\#\text{Cyc}(W_u)-1})$, when applied to a k-abelian singleton. It is not hard to see that this language is a simple polynomial language, so that the complexity of $L_{k,\Sigma,\text{sing}}$ is not of the order $\mathcal{O}(n^{\#\text{Cyc}(W_u)-2})$. By Remark 4.22, u defines a cycle-semi-decomposition with $\#\text{Cyc}(W_u)$ cycles. We conclude that $s_{k,m}$ is the claimed quantity. □

In pursuit of finding bounds for $s_{k,m}$, we consider the cycle-semi-decompositions of the de Bruijn graph due to the above characterization of $s_{k,m}$.

6.4.1 On Maximal Cycle-Decompositions

As concluded in Proposition 4.25, $s_{k,m}+1$ is bounded above by the number of necklaces of length $k-1$ over an m-letter alphabet. We explore the possibility when $s_{k,m} = N_m(k-1)-1$ and define the following.

Definition 6.28. A cycle-semi-decomposition $V/\mathcal{C}$ of $dB_\Sigma(n)$ is called *maximal* if $\mathcal{C}$ contains $N_{|\Sigma|}(n)$ cycles.

Note that a maximal cycle-semi-decomposition of $dB_\Sigma(n)$ exists for any $n \in N$: take the cycles induced by necklaces. The corresponding quotient graph then contains only vertices corresponding to cycles. This actually holds for any maximal cycle-semi-decomposition of $dB_\Sigma(n)$ by the following proposition.

Proposition 6.29 ([55])**.** *For any maximal cycle-semi-decomposition $V/\mathcal{C}$ of $dB_\Sigma(n)$, each vertex occurs in one of the cycles of $\mathcal{C}$. In particular, $V/\mathcal{C}$ is a cycle-decomposition.*

Proof. We first recall the following: Let G be an Eulerian graph and $\tilde{C}$ a set of edge-disjoint cycles of G. Then there exists a decomposition $\tilde{D}$ of G into edge-disjoint cycles such that $\tilde{C} \subseteq \tilde{D}$. Indeed, since G is Eulerian, each vertex v has the property $d^+_G(v) = d^-_G(v)$. If G' is the graph obtained from G by removing edges occurring in $\tilde{C}$, then each vertex $v \in G'$ has the property $d^+_{G'}(v) = d^-_{G'}(v)$. It follows from Veblen's theorem ([105]) for directed graphs (see, e.g., [13], exercise 2.4.2) that there exists a decomposition $\tilde{E}$ of G' into edge-disjoint cycles. Now $\tilde{E} \cup \tilde{C} = \tilde{D}$ is a decomposition of G into edge-disjoint cycles satisfying the claim.

Let then C be a set of $N(n)$ vertex-disjoint cycles in $dB(n)$. The vertices of $dB(n)$ correspond to the edges of $dB(n-1)$, so that C can be seen as a set $\tilde{C}$ of *edge*-disjoint cycles of $dB(n-1)$, an Eulerian graph. Suppose there is a vertex v in $dB(n)$ not included in any of the cycles of C. Then v corresponds to an edge e in $dB(n-1)$ which does not occur in $\tilde{C}$. By the above, $\tilde{C}$ can be extended to a decomposition of $dB(n-1)$ into edge-disjoint cycles $\tilde{D}$ such that $\tilde{C} \subset \tilde{D}$. But now $\tilde{D}$ can be seen as a set of vertex-disjoint cycles of $dB(n)$ with more than $N(n)$ cycles, a contradiction. □

Observe that, for a maximal cycle-decomposition $\mathcal{C}$ of $dB(n)$, the quotient graph can be seen as an undirected graph. To see this, let (X,Y) be an edge of $G = dB(n)/\mathcal{C}$, that is, there exist $a,b \in \Sigma$, $u \in \Sigma^{n-1}$ such that $au \in V(X)$ and $ub \in Y$ whence (au, ub) is an edge in $dB(n)$. By Proposition 6.29, X and Y are

cycles, so that there exist $c, d \in \Sigma$ such that $(au, uc) \in E(X)$ and $(du, ub) \in E(Y)$, whence (du, uc) is an edge in $dB(n)$. By definition, $(Y, X) \in G$ as well.

The following proposition is immediate, as noted in [55].

Proposition 6.30. *Let $k, m \geqslant 1$. Then $\mathcal{S}_m^{(k)}(n) = \Theta(n^{N_m(k-1)-1})$ if and only if there exists a maximal cycle-decomposition of $dB_\Sigma(k-1)$, where $|\Sigma| = m$, such that the induced quotient graph contains a Hamiltonian path, i.e., a path traversing through each vertex precisely once.*

When inspecting propositions 6.5 and 6.6, one observes that actually, in those cases, $\mathcal{S}_m^{(k)}(n) \sim D_{k,m} n^{N_m(k-1)-1}$ for some constant $D_{k,m}$ depending on k and m. Indeed, for these values of k and m there exist maximal cycle-decompositions of $dB(k-1)$ such that the quotient graph contains a Hamiltonian path. For any k and m, the existense of such a decomposition implies the existense of such a constant $D_{k,m}$ by the following results.

Lemma 6.31. *Let $n \geqslant 1$ and let Σ be an alphabet. Then any maximal cycle-decomposition of $dB_\Sigma(n)$ contains the cycle (loop) (a^n, a) for any $a \in \Sigma$.*

Proof. Let $\mathcal{C}$ be a maximal cycle-decomposition of $dB(k-1)$. Now the vertex a^n occurs in some cycle $C \in \mathcal{C}$ by Proposition 6.29. Assume that $C \neq (a^n, a)$. Since the cycles in $\mathcal{C}$ are vertex-disjoint, the set $\mathcal{C}' = (\mathcal{C} \setminus C) \cup (a^n, a)$ consists of $N_m(n)$ vertex-disjoint cycles. But now the vertices in $V(C) \setminus \{a^n\}$ do not occur in any cycle, which contradicts Proposition 6.29. □

The following conditional result is based on work presented in this thesis.

Proposition 6.32. *Assume that $\mathcal{S}_m^{(k)}(n) = \Theta(n^{N_m(k-1)-1})$. Then $\mathcal{S}_m^{(k)}(n) \sim D_{k,m} n^{N_m(k-1)-1}$ for some constant $D_{k,m}$ depending on k and m.*

Proof. The proof is analogous to that of Theorem 6.23. Namely, we show, similar to Proposition 6.24, that each saturated walk-automaton of the minimal DFA recognizing $L_{k,\Sigma,\mathrm{sing}}$ contains a cycle of length 1.

Let $\mathcal{A}$ be the minimal DFA recognizing $L_{k,\Sigma,\mathrm{sing}}$. Let $\mathcal{B}$ be a saturated walk-automaton of $\mathcal{A}$. By assumption we have $\#\mathrm{Cyc}(\mathcal{B}) = N_m(k-1)$. As in the proof of Proposition 6.24, we may construct a corresponding walk W in $dB(k-1)$ containing at least $N_m(k-1)$ cycles. Further, the labels of these cycles correspond to conjugates of the labels of the cycles in $\mathrm{Cyc}(W)$. On the other hand, since the word corresponding to W is a k-abelian singleton, the cycles occurring along W are vertex-disjoint, so that $\#\mathrm{Cyc}(\mathcal{B}) = \#\mathrm{Cyc}(W)$. The cycles $\mathrm{Cyc}(W)$ now define a maximal cycle-decomposition of $dB(k-1)$, whence, by the above lemma, one of the cycles occurring along W is (a^{k-1}, a). It follows that one of the cycles of $\mathrm{Cyc}(\mathcal{B})$ has label a^{r_i} for some integer r_i. It can be shown that $r_i = 1$, using the assumption that $\mathcal{A}$ is minimal (using the same argument as in Proposition 6.24). Thus $\mathrm{Cyc}(\mathcal{B})$ contains a cycle of length 1. □

We conclude this chapter by considering concrete cycle-decompositions of the de Bruijn graphs. The rest of the considerations of this chapter appear in [55].

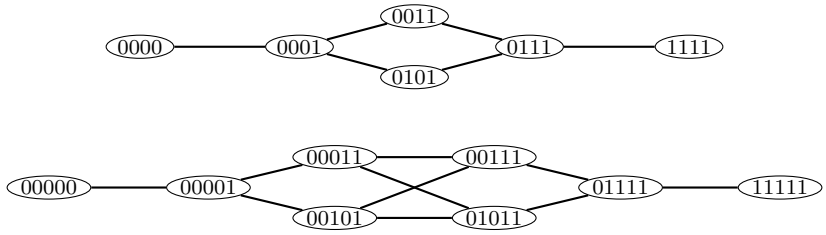

Figure 6.1: The binary necklace graphs $NG(4)$ and $NG(5)$. A vertex represents the necklace induced by its label. [55, Fig 4].

On Long Paths in Necklace Graphs

We consider the cycle-decomposition of the de Bruijn graph given by the cycles defined by necklaces. Note that the length of such a cycle necessarily divides the order of the de Bruijn graph. We begin with a definition.

Definition 6.33. Let $\mathcal{C}_\Sigma(n)$ be the set of cycles induced by necklaces of length n over the alphabet Σ. The quotient graph $NG_\Sigma(n) = dB_\Sigma(n)/\mathcal{C}_\Sigma(n)$ is called the *necklace graph* of order n.

Note that $NG(n)$ does not always contain a Hamiltonian path:

Example 6.34. The binary necklace graphs of order 4 and 5 are illustrated in Figure 6.1. A longest path in $NG(4)$ contains 5 vertices out of a total of 6. On the other hand, one can easily find a Hamiltonian path in $NG(5)$.

The problem of finding a Hamiltonian path in necklace graphs has been extensively studied in terms of *Gray codes* ([28, 97, 99, 108, 111]). A *Gray code for necklaces* of length n is defined as a sequence of all necklaces of length n such that two consecutive necklaces have representatives which differ in one bit. One can easily see that Gray codes for necklaces correspond to Hamiltonian paths in necklace graphs. The following conjecture was stated over 20 years ago. To the best of our knowledge it remains unsolved to this day.

Conjecture 6.35 ([99, Section 7])**.** *Let $n \in \mathbb{N}$ be odd and let Σ be a binary alphabet. Then there exists a Gray code for necklaces of length n over Σ. In other words, $NG(n)$ contains a Hamiltonian path.*

The conjecture has previously been verified for necklaces up to length 9 in [111]. Michael Rao verified the conjecture computationally up to $n = 15$ [55]. For completeness, we give Gray codes for binary words of odd lengths 5–15 in Table 6.3. In the table, a Gray code is represented as a word $a_1 \dots a_{N(n)-1}$ over the *hexadecimals* to be read as follows. The first necklace in the ordering is 0^n. The $(i+1)$st, $i \geqslant 0$, necklace is then obtained by complementing the a_ith letter ($0 \leftrightarrow 1$) of the lexicographically least representative of the ith necklace. For example, the coding 1114111 corresponds to the following ordering of necklaces of length 5:

$$00000, 00001, 00011, 00111, 00101, 01011, 01111, 11111.$$

n	Gray code
5	1114111
7	1116165614521341111
9	111818781678767815768765678618786781859515754158137541132 11
11	111a1a9a189a989a179a989a9798a9891679a89a978a98a6789a89789a871579a978a98a67a89a8a7898b6167a9a879697a89a97a5856896378a2789a87269a78a9896787984892789 79a62a8a7292474525452711858 4a82a346181811
13	111c1cbc1abcbabc19bcbabcb9bacbab189cbcabcb9abacba89bcbabcb9bcaba1789abcab9acbcac98abcbabc78abcab9acbcac98abcbab97acbcac7abcba16789cbc9abcbab89cabcac9abcba789acbac9acbc89cbacbc89b65189abcb9abcac89cabcac9879acabc8abcab96acba6915677ca1579bacba9a8cbcab1ca78b9cabcb8797b8cba89cbacb9b768c9bc9acb89bcabcb9856ab8abcb9acbac89cbca97c7ba7568cb89acabc9789cba9babc98acabda7acb188acabc78cbc8ca967cb7ab76a8bacba8a57c5a8bacba614ccb2bc89aca9aba789bc9acb8b67963acab567bc9acbab789b9abca98ca9ba678ba89abc7ac859c97897ca98ba4a9ca2c67cb98747a158ca9bc89ca7a89b8d6167bc74a674ca47c45956c89aca6a42c11b18bc92c8ba86ca657b52865628ba392952178a712ca139c78197a4411
15	111e1ede1cdedcde1bdedcdedbdcedcd1abedecdedbcdcedcabdedcdedbdecdc19abcedecedbcedeceabcdedcde9acedecedbcedeced17dedcd9bcedecebcdedc189abcdecdbcedecebacdedcde9acedecedbcedeced89bcdecdbcedecebacdedcdebc8edeced9bcedecebcdedc89cdc9abcedecedacedeced16edece178abedebcdedcdabcedeced9acdecdbcedecebacdedcdbc9edece9cdca89bcdecdbcedecebacdedcde9acdecdbcedecebacdedcdb9cedece9cded8c8789abdedcbdedabdedcded9acedceb9edacdece9acedecbacdedbdedcf9cded18eb89abedcecbcdedcd9bcdecdbcdecba9cdedcde167bdcedcbdecede9edecd1679edbecdedbdedca8bde3cdedbdcedcd9becdecea8cdcedcbdcedca8dedced89bcdecdbce9bdecded9bec9dcda789bedbcedecabcdcedcbded9badcedcbdece9ae9bcbedbcdbd98abdedcbdedabedcedebadeaced9f7c9edece179aedecdecded8bab8cdcedcbadcedc678becdedbdcedcda9becdecebcdedc9bcedce89aedcbdedcbacedcdbadeacedec9abcecdebcecda67d96f819bcdece8bcded8a789dedcbacdcedcbdedcdaedad9abedcecbedceaebdecdedbcec86dce67ec79abedebcdedcd9bedcecbedca5dec19bcdcedcbded971579adedbcedceabecdecbcedcda9bcecdebcdecdabecdecbcedcd789acdcedcabedcede9acecdebcdecda9ecd9bdcedcbc98abedbcdeabedcedeba9867bceedcecdedbcedce9aecd9bdcedcbcde8c86789decb9adebcdecda9ed9bcdce89eadebcedceaecbdeceba9bcbdcbedbca8b8edcdeb879adedcabdcedcb9aebcdedbdeca9ce9becdece9b9c789ec7abebcdceabecdecebcdc97875f467abebcdceabeced2edeb9decded15bcbdcbedbc9a9bedcedeb7569bcec9adedcbdcedc9aedc98cdecd8bedcebed789dbecdedbadaeda9bdedcbdceabedcecbc98aedcedebacecdebcdc98a8768bcedccdedbc867edcbedbdabedcecb98acedceabedabacbec9adedaceb8edcd79a9bdedbdecda68adcea968bedcecbcedc87c9adecd5a7a59ecdeceabcecde9f46be68ecded8adedbcedc7abebd67bdeb81689de71edaedcbe1ebeabcecdbcdc89acdcabdcecbacaed9adecbedbd78bdbcdecba9ecbdcb9dceb9ce9cdc87bdcecbce79aecedabdcebce968dc1ebedbcdeabecd5679dce2bcedcd98cd38569bedbcde9aedcbdebea98acabcdba97bdebcedadbcebc79ca9bed6dabdcbd89dacdabdc9caeb8ab89ba7ecbcdc96bdcbd8965958adecbebd87cdea879c8b6986ecbc7897bcbe745eddcbddeab9aebcec8de6ed4c92ecbedbdaeb676459dabdc9ca78bda9dbab45dcbdd9bcb635f4de51676b2dab76ba6997ec89a9ce7ec9ea7978bc8ce45e6cbce6746bec1e78aecbacbd9b74319db8de9a8cec9acecb986511398b7959edbda796c8cd8a867a8dc1aeceb1d425975ae7ec425527e11d1dbadcac1642aca8acd8abd89bd96e729a8595dbeb7265bd18a71dcdc3c4562b8676d7b74169b3115278631cca414f52ed1db62ad416187cb111

Table 6.3: Gray codes for binary necklaces of odd lengths up to 15. [55, Table 1].

Remark 6.36. By Proposition 6.32 and the above discussion, for each even $k \leqslant 16$, there exists a constant $D_{k,2}$ such that $\mathcal{S}_2^{(k)}(n) \sim D_{k,2} n^{N_2(k-1)-1}$.

On the other hand, binary necklace graphs are bipartite. When $n \geqslant 4$ is even, the difference of cardinalities of the partitions is greater than 1 so the graph cannot contain a Hamiltonian path. In fact, it is not hard to calculate an upper bound on the length of the longest path:

Proposition 6.37. *For the binary alphabet and n even, the number of vertices in a longest path in $NG(n)$ is at most*

$$BPL(n) = \frac{1}{n} \sum_{\substack{d \mid n \\ 2 \nmid d}} \varphi(d) 2^{n/d} + 1. \tag{6.7}$$

Remark 6.38. The first few terms of $(BPL(2n))_{n=1}^{\infty}$ are

$$3,\ 5,\ 13,\ 33,\ 105,\ 345,\ 1173,\ 4097,\ 14573,\ 52433,\ 190653,\ 699073, \ldots$$

The sequence $(BPL(n))_{n=1}^{\infty}$ equals $(a(n)+1)_{n=1}^{\infty}$ where $(a(n))_{n=1}^{\infty}$ is the sequence A063776 in the *On-Line Encyclopedia of Integer Sequences* (`http://oeis.org`, accessed November 5, 2018).

Proof. Let A (resp., B) be the set of necklaces containing an even (resp., odd) number of 1s; $NG(n)$ is then bipartite with respect to the partition into A and B. Now the number $N(n,\ell)$ of necklaces of length n containing precisely ℓ 1s equals $\frac{1}{n}\sum_{d \mid \gcd(\ell,n)} \varphi(d) \binom{n/d}{\ell/d}$ (see, e.g., [97]). We thus have

$$n|A| = \sum_{\ell=0}^{n/2} nN(n,2\ell) = \sum_{\ell=0}^{n/2} \sum_{d \mid \gcd(2\ell,n)} \varphi(d) \binom{n/d}{2\ell/d}.$$

Let us count the above sum in a different order. Consider first even divisors d of n. In the above sum, we count $\varphi(d)\binom{n/d}{2\ell/d}$ for each $0 \leqslant 2\ell \leqslant n$ such that $d \mid 2\ell$, i.e., $2\ell = d\ell'$ for some ℓ' (since $2 \mid d$). We thus count $\varphi(d)\binom{n/d}{\ell'}$ for each $0 \leqslant \ell' \leqslant \frac{n}{d}$.

Consider then a fixed divisor d of n such that $2 \nmid d$. Similar to the above, we count $\varphi(d)\binom{n/d}{2\ell/d}$ for each $0 \leqslant 2\ell \leqslant n$ such that $d \mid 2\ell$, that is, $2\ell = 2d\ell'$ for some ℓ' (since $2 \nmid d$). We thus count $\varphi(d)\binom{n/d}{2\ell'}$ for each $0 \leqslant \ell' \leqslant \frac{1}{2}\frac{n}{d}$.

Combining the above calculations we obtain

$$\begin{aligned} n|A| &= \sum_{\substack{d \mid n \\ 2 \mid d}} \varphi(d) \sum_{\ell'=0}^{n/d} \binom{n/d}{\ell'} + \sum_{\substack{d \mid n \\ 2 \nmid d}} \varphi(d) \sum_{\ell'=0}^{\frac{1}{2}n/d} \binom{n/d}{2\ell'} \\ &= \sum_{\substack{d \mid n \\ 2 \mid d}} \varphi(d) 2^{n/d} + \sum_{\substack{d \mid n \\ 2 \nmid d}} \varphi(d) 2^{n/d-1} = \tfrac{n}{2} N_2(n) + \tfrac{1}{2} \sum_{\substack{d \mid n \\ 2 \mid d}} \varphi(d) 2^{n/d}, \end{aligned}$$

so that $|B| = \frac{1}{2}N_2(n) - \frac{1}{2n}\sum_{\substack{d \mid n \\ 2 \mid d}} \varphi(d) 2^{n/d} = \frac{1}{2n}\sum_{\substack{d \mid n \\ 2 \nmid d}} \varphi(d) 2^{n/d} < |A|$. A longest path in $NG(n)$ can thus contain at most $|B|+1$ vertices from A and $|B|$ vertices from B, that is, $2|B|+1 = \frac{1}{n}\sum_{\substack{d \mid n \\ 2 \nmid d}} \varphi(d) 2^{n/d} + 1 = BPL(n)$ vertices in total. □

n	code
6	111521651511
8	11171767156725671472674521615611

Table 6.4: Path in the binary necklace graph $NG(n)$ of length $BPL(n)$ as defined in (6.7). [55, Table 2].

Example 6.39. For even $n \leqslant 8$ the bound in the above proposition is actually achievable. See Figure 6.1 for the case $n = 4$ and Table 6.4 for $n = 6$ and 8, where the coding is defined as that of the Gray codes in Table 6.3.

In [55] it is conjectured that this bound is achievable for all even n:

Conjecture 6.40. *For even n, the length of a longest path in the (bipartite) binary necklace graph $NG(n)$ is equal to $BPL(n)$ (defined by (6.7)).*

On Other Maximal Cycle-Decompositions of the de Bruijn graph

Next, we briefly discuss the maximal cycle-decompositions not induced by necklaces. In other words, the length of a cycle in such a decomposition need not divide the order of the de Bruijn graph. We give some examples of such decompositions which induce quotient graphs containing Hamiltonian paths.

Example 6.41. The 5-abelian singleton

$$u = 0^4 0^2 1 (10001)^2 1001 (001)^2 0101 (01)^2 1011 (1011)^2 1^4$$

corresponds to a cycle-decomposition of $dB(4)$ with 6 cycles $C_1, \cdots, C_6$ and the resulting quotient graph contains a Hamiltonian path. Here the cycles are defined as (v, ℓ), where v is the starting vertex and ℓ is the label of the cycle:

$$(0^4, 0),\ (0^3 1, 10^3 1),\ (1001, 001),\ ((01)^2, 01),\ (1011, 1011),\ (1^4, 1).$$

Moreover, the cycle decomposition contains $N_2(4) = 6$ cycles, which is the maximal possible by Theorem 4.24. Note here that the second and third cycles have lengths which do not divide 4. The other cycles are defined by necklaces.

Example 6.42. Similarly, we obtain the following $N_2(6) = 14$ vertex-disjoint cycles in $dB(6)$ which induce a quotient graph containing a Hamiltonian path:

$$(0^6, 0),\ (0^5 1, 0^5 1),\ (0^4 11, 0^4 11),\ (0^3 1^3, 0^3 1^3),\ (01^3 01, 01^3 01),\ ((10)^3, 10),$$
$$(010100, 010100),\ (001011, 001011),\ ((100)^2, 100),\ (010011, 010011),$$
$$((011)^2, 011),\ (1101^3, 101^3)^*,\ (1^4 00, 1^5 00)^*,\ (1^6, 1).$$

The ordering of the cycles gives a Hamiltonian path in the quotient graph. The first vertex of each cycle is reached by a vertex in the previous cycle. The cycles marked by $*$ have lengths 5 and 7, respectively. Neither of these lengths divides the order 6 of the de Bruijn graph. All other cycles are defined by necklaces.

Example 6.43. For $n = 8$ we compute the following set of $N_2(8) = 36$ vertex-disjoint cycles of $dB(8)$, the cycles listed in an order yielding a Hamiltonian path in the quotient graph. For each cycle, the first vertex is reached by some vertex in the previous cycle. Here all cycles marked with $*$ have lengths which do not divide the order 8 of the de Bruijn graph. All other cycles are defined by necklaces.

$(0^8, 0)$, $(0^71, 0^71)$, $(0^611, 0^611)$, $(0^51^3, 0^51^3)$, $(10^510, 10^510)$,
$(1010^41, 1010^41)$, $(010^410, 010^410)$, $(10^4101, 10^4101)$, $(110^411, 110^411)$,
$(0^31^5, 0^31^5)$, $(10^31^30, 10^31^30)$, $(010^3110, 010^3110)$, $((0010)^2, 0010)$,
$(010^3101, 010^3101)$, $((01)^4, 01)$, $((10)^311, 1^3(01)^31)^*$, $1(10)^30, 1(10)^30)$,
$(01001100, 01001100)$, $(0110^311, 0110^311)$, $((011)^201, 101)^*$,
$(1(011)^21, (101)^2(011)^21)^*$, $((01)^2(10)^2, (01)^2(10)^2)$, $(10(100)^2, 10(100)^2)$,
$((100)^211, (100)^211)$, $(01001^4, 01001^4)$, $(1001^5, 1001^5)$, $(1101^300, 1101^300)$,
$((1100)^2, 1100)$, $(011001^3, 011001^3)$, $0(01)^31, 0(01)^31)$,
$(11001011, 11001011)$, $(00101^30, 00101^30)$, $(0101^301, 01^301)^*$,
$((1011)^2, 1011)$, $(1^301^4, 1^301^4)$, $(1^8, 1)$.

Remark 6.44. By Proposition 6.32 and the above examples, for each odd $k \leqslant 9$ there exists a constant $D_{k,2}$ such that $\mathcal{S}_2^{(k)}(n) \sim D_{k,2} n^{N_2(k-1)-1}$.

The above observations led to state the following conjecture in [55], which was verified for all odd $n \leqslant 15$ and all even $n \leqslant 8$ over the binary alphabet.

Conjecture 6.45. *For every $n \in \mathbb{N}$ and alphabet Σ, there exists a maximal cycle-decomposition $\mathcal{C}$ of $dB_\Sigma(n)$ for which $dB_\Sigma(n)/\mathcal{C}$ contains a Hamiltonian path.*

This problem is quite intricate at least in the binary case for even n. Indeed, if true, such a cycle-decomposition contains cycles not defined by necklaces. On the other hand, it seems that the problem is quite hard for odd lengths as well, since it is not known whether the necklace graph would give such a cycle-decomposition.

An equivalent formulation, due to Proposition 6.30 together with Proposition 6.32, is the following.

Conjecture 6.46. *For any $k, m \geqslant 1$, there exists a constant $D_{k,m}$ such that $\mathcal{S}_m^{(k)}(n) \sim D_{k,m} n^{N_m(k-1)-1}$.*

Chapter 7

On k-Abelian and k-Binomial Equations

Recall that the k-abelian equivalence is a congruence, whence $\Sigma^*/\sim_k$, is a monoid. Recall also that the k-binomial equivalence defines a congruence, and thus $\Sigma^*/\equiv_k$ is a monoid. We call these monoids the *k-abelian monoid* and the *k-binomial monoid*, respectively.

In this chapter we consider equations over these previously described monoids. The two equivalence relations are quite closely related, though they are incomparable as equivalence relations. In particular, we consider the concrete equations $xy = yx$ and $xz = zy$, that is, commutation and conjugacy, respectively. For the k-abelian monoid we obtain characterizations of the sets $\mathrm{Sol}(xy, yx)$ and $\mathrm{Sol}(xz, zy)$. In contrast, in the k-binomial monoid, we only obtain partial results. Part of the challenges in this case follow from the property that a modification in just one position of a word can have global effects of the distribution of subwords, and thus the structure of the equivalence classes.

We also consider the questions of independent systems of equations over these monoids. We show that both monoids possess the so-called *compactness property* (see Definition 7.17.) Moreover, in both cases, we give an upper bound on the number of equations in an independent system of equations. This chapter is based on the manuscript [110].

7.1 On Commutation in $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$

For a word $x \in \Sigma^*$ we let $\rho(x)$ denote the *primitive root* of x, that is, $\rho(x)$ satisfies $x = \rho(x)^n$, where n is maximal. Thus x is primitive if $x = \rho x$. We let $\mathrm{per}(x)$ denote the set of *periods* of x, that is, $\mathrm{per}(x) = \{y \colon x \in \mathrm{pref}(y^\omega), |y| \leqslant |x|\}$. We let $\widetilde{\mathrm{per}}(x)$ denote the *left periods* of x, that is,

$$\widetilde{\mathrm{per}}(x) = \{y \colon x \in \mathrm{suff}(y^n), |y| \leqslant |x| \leqslant n|y|\}.$$

7.1.1 Commutation in the k-Abelian Monoid

Let us first inspect commutativity in the k-abelian monoid. In the following, we let $p_k(w)$ (resp., $s_k(w)$) denote the word $\mathrm{pref}_{\min\{|w|,k\}}(w)$ (resp., $\mathrm{suff}_{\min\{|w|,k\}}(w)$). We begin with a technical lemma. Recall here that the generalized Parikh vector $\Psi_k(w)$ equals the vector $(|w|_x)_{x\in\Sigma^k}$.

Lemma 7.1. *Let $x, y \in \Sigma^*$ be primitive and $k \geqslant |x|, |y|$. Let further $u \in \mathrm{pref}(x^\omega)$ and $v \in \mathrm{pref}(y^\omega)$ such that $u \neq v$ and $|u| = |v| \geqslant k + \max\{|x|, |y|\} - 1$. Then $\Psi_k(u) = \Psi_k(v)$ if and only if x and y are conjugates (in Σ^*) and $|u| \equiv k - 1 \pmod{|x|}$.*

Proof. Let $X = \mathrm{pref}_{k-1}(x^\omega)$ and $Y = \mathrm{pref}_{k-1}(y^\omega)$. We first observe that, since $|x| \leqslant k$ and x is primitive, we have $\#F_k(x^\omega) = |x|$. In other words, each position $i \in [0, |x|)$ starts a distinct factor of x^ω of length k; we enumerate these factors by $w_i = x^\omega[i, i + k)$, $i \in [0, |x|)$.

Suppose first that x and y are conjugates and $u \in F(x^\omega) = F(y^\omega)$ has length $n|x| + k - 1$ for some $n \geqslant 1$. Then we may write $u = x^n X$. By arguing as above, it is straightforward to see that $|u|_{w_i} = n$ for each $i \in [0, |x|)$. Moreover, this holds for any factor of x^ω of length $|u|$, and thus $\Psi_k(u) = \Psi_k(v)$, since $v \in F(x^\omega)$.

Suppose then that $\Psi_k(u) = \Psi_k(v)$. We first show that x and y are conjugates. Since $|x| \leqslant k$ and $|u| \geqslant k - 1 + |x|$ it follows that xX is a prefix of u. Similarly, yY is a prefix of v. Since $\Psi_k(u) = \Psi_k(v)$, each conjugate of x is a factor of v and thus of y^ω. Similarly, each conjugate of y is a factor of x^ω. It follows that $|x| = |y|$, and thus x and y are conjugates.

We previously showed that if $|u| \equiv k - 1 \pmod{|x|}$ then $\Psi_k(u) = \Psi_k(v)$. Assume thus that $|u| \not\equiv k-1 \pmod{|x|}$. We may write $u = x^m X x_0$ and $v = y^m Y y_0$ for some $m \geqslant 1$, x_0 a factor of x^ω, y_0 a factor of y^ω, and $1 \leqslant |x_0| = |y_0| < |x| - 1$. We now have $\Psi_k(u) = \Psi_k(x^m X) + \Psi_k(X x_0)$ and similarly $\Psi_k(v) = \Psi_k(y^m Y) + \Psi_k(Y y_0)$. Since $\Psi_k(x^m X) = \Psi_k(y^n Y)$ by the above, we obtain

$$\Psi_k(u) - \Psi_k(v) = \Psi_k(X x_0) - \Psi_k(Y y_0).$$

Since $|x_0| < |x|$, we have that $|X x_0|_{w_i} = 1$ for all $i \in [0, |t|)$ and $|X x_0|_{w_i} = 0$ otherwise. Let then $\mathrm{pref}_k(Y y_0) = w_j$. Since $u \neq v$, we have $j > 0$. Similarly $|Y y_0|_{w_i} = 1$ for all $i \in \{j + \ell \pmod{|x|} \mid \ell \in [0, |y_0|)\}$, and otherwise $|Y y_0|_{w_i} = 0$. Consider now the index $n = \min\{j - 1, |x_0| - 1\}$. Now $|u|_{w_n} = 1$, since $n < |x|_0$. On the other hand, $n \notin \{j + \ell \pmod{|x|} \mid \ell \in [0, |y_0|)\}$ as, for each $0 \leqslant \ell < |y_0|$, either $j + \ell \pmod{|x|} > j$ or $j + \ell \pmod{|x|} < |y_0| - 1$. It follows that $|Y y_0|_{w_n} = 0$, and hence $\Psi_k(u) \neq \Psi_k(v)$. This concludes the proof. □

We may now characterize k-abelian commutation.

Proposition 7.2. *Let $x, y \in \Sigma^+$, with $|x| \leqslant |y|$.*

- *If $|xy| < 2k$, then $xy \sim_k yx$ if and only if $x, y \in r^*$ for some $r \in \Sigma^*$.*
- *If $|x| < k - 1$ and $|xy| \geqslant 2k$, then $xy \sim_k yx$ if and only if x is a period of $\mathrm{pref}_{k-1}(y)$ and x is a left period of $\mathrm{suff}_{k-1}(y)$.*
- *If $|x| \geqslant k - 1$ and $|xy| \geqslant 2k$, then $xy \sim_k yx$ if and only if $\mathrm{pref}_{k-1}(x) = \mathrm{pref}_{k-1}(y)$ and $\mathrm{suff}_{k-1}(x) = \mathrm{suff}_{k-1}(y)$.*

Proof. The first point follows trivially from the observation that $xy \sim_k yx$ if and only if $xy = yx$.

Assume thus that $|xy| \geqslant 2k$. We have

$$\Psi_k(xy) = \Psi_k(x) + \Psi_k(s_{k-1}(x)p_{k-1}(y)) + \Psi_k(y).$$

Hence $xy \sim_k yx$ if and only if $\Psi_k(s_{k-1}(x)p_{k-1}(y)) = \Psi_k(s_{k-1}(y)p_{k-1}(x))$, $\mathrm{pref}_{k-1}(xy) = \mathrm{pref}_{k-1}(yx)$, and $\mathrm{suff}_{k-1}(xy) = \mathrm{suff}_{k-1}(yx)$. For the third point, observe that this implies $xy \sim_k yx$ if and only if $\mathrm{pref}_{k-1}(x) = \mathrm{pref}_{k-1}(y)$ and $\mathrm{suff}_{k-1}(x) = \mathrm{suff}_{k-1}(y)$, as it was claimed.

We are thus left with the second point, that is, $|x| < k-1$ and $|xy| \geqslant 2k$. Observe that now $p_{k-1}(x) = x = s_{k-1}(x)$. Suppose first that x is a period of $Y = \mathrm{pref}_{k-1}(y)$ and x is a left period of $Y' = \mathrm{suff}_{k-1}(y)$. It follows that $Y = x^n p$ and $Y' = sx^n$ for some $n \in \mathbb{N}$, $p \in \mathrm{pref}(x)$, and $s \in \mathrm{suff}(x)$. Thus

$$\begin{aligned}\mathrm{pref}_{k-1}(xy) = \mathrm{pref}_{k-1}(xY) = \mathrm{pref}_{k-1}(x^{n+1}p) = Y = \mathrm{pref}_{k-1}(yx) \text{ and}\\ \mathrm{suff}_{k-1}(yx) = \mathrm{suff}_{k-1}(Y'x) = \mathrm{suff}_{k-1}(sx^{n+1}) = Y' = \mathrm{suff}_{k-1}(xy).\end{aligned}$$

Furthermore, $|sx^{n+1}| = |x^{n+1}p| = k - 1 + |x|$ so that $|xY| \equiv k-1 \pmod{|\rho(x)|}$. By the above lemma we have $\Psi_k(xY) = \Psi_k(Y'x)$. It thus follows that $xy \sim_k yx$.

Suppose then that $xy \sim_k yx$. Now $\mathrm{pref}_{k-1}(xy) = xy_1 = \mathrm{pref}_{k-1}(yx) = Y = y_1y_2$ for some $y_1, y_2 \in \Sigma^*$. It follows that $y_1 \in (pq)^*p$, $x = pq$, and $y_2 = qp$ for some $p, q \in \Sigma^*$. Thus x is a period of Y.

Similarly $\mathrm{suff}_{k-1}(yx) = y_3x = \mathrm{suff}_{k-1}(xy) = Y' = y_4y_3$ for some $y_3, y_4 \in \Sigma^*$, so that $x = rs$, $y_3 \in (sr)^*s$, and $y_4 = sr$ for some $r, s \in \Sigma^*$. We thus see that x is a left period of Y', as was claimed. □

This concludes the characterization of the set $\mathrm{Sol}(xy, yx)$ of solutions in the k-abelian monoid.

7.1.2 On Commutation in the k-Binomial Monoid

Next we consider commutation in the k-binomial monoid. In this case we only obtain partial results, and, in particular, are not able to characterize the set $\mathrm{Sol}(xy, yx)$ of solutions in the k-binomial monoid, for general k. However, we manage to do so in the case of the 2-binomial monoid:

Proposition 7.3. *For all $x, y \in \Sigma^+$, $xy \equiv_2 yx$ if and only if $\Psi(x) = \frac{|x|}{|y|}\Psi(y)$.*

Proof. Suppose first that $xy \equiv_2 yx$. It follows that for all $a, b \in \Sigma$

$$\tbinom{x}{ab} + \tbinom{x}{a}\tbinom{y}{b} + \tbinom{y}{ab} = \tbinom{xy}{ab} = \tbinom{yx}{ab} = \tbinom{x}{ab} + \tbinom{y}{a}\tbinom{x}{b} + \tbinom{y}{ab},$$

which is equivalent to $|x|_a|y|_b = |y|_a|x|_b$. By summing both sides over all $b \in \Sigma$, we obtain $|x|_a|y| = |y|_a|x|$ for all $a \in \Sigma$, which is equivalent to $\Psi(x) = \frac{|x|}{|y|}\Psi(y)$ as claimed.

For the converse, we observe that the property $|x|_a = \frac{|x|}{|y|}|y|_a$ for all $a \in \Sigma$ implies that $|x|_a|y|_b = \frac{|x|}{|y|}|y|_a\frac{|y|}{|x|}|x|_b = |y|_a|x|_b$ for all $a, b \in \Sigma$, which, in turn, is equivalent to $xy \equiv_2 yx$ as was seen above. □

By allowing x and y to be empty words, we may state the above proposition in other words: The elements x and y commute in $\Sigma^*/\equiv_2$ if and only if there exist a word $r \in \Sigma^*$ and non-negative integers ℓ and n such that $x \equiv_1 r^\ell$ and $y \equiv_1 r^n$. In this case r is a *common* 1*-binomial root*. In the following, we consider generalizing this to larger k. The following proposition says that sharing common $(k-1)$-binomial roots implies k-binomial commutation.

Proposition 7.4. *Let $k \geqslant 2$ be an integer, $r \in \Sigma^*$, and $m, n \geqslant 0$. For any $x \equiv_{k-1} r^m$ and $y \equiv_{k-1} r^n$ we have $xy \equiv_k yx$.*

Proof. For all $a \in \Sigma$, we clearly have $|xy|_a = |yx|_a$. Further, for each word $e \in \Sigma^{\leqslant k}$ of length at least two,

$$\begin{aligned}\binom{xy}{e} - \binom{x}{e} - \binom{y}{e} &= \sum_{\substack{e_1e_2=e\\ e_1,e_2\in\Sigma^+}} \binom{x}{e_1}\binom{y}{e_2} \qquad = \sum_{\substack{e_1e_2=e\\ e_1,e_2\in\Sigma^+}} \binom{r^\ell}{e_1}\binom{r^n}{e_2} \\ &= \binom{r^{\ell+n}}{e} - \binom{r^\ell}{e} - \binom{r^n}{e} = \sum_{\substack{e_1e_2=e\\ e_1,e_2\in\Sigma^+}} \binom{r^n}{e_1}\binom{r^\ell}{e_2} \\ &= \sum_{\substack{e_1e_2=e\\ e_1,e_2\in\Sigma^+}} \binom{y}{e_1}\binom{x}{e_2} \qquad = \binom{yx}{e} - \binom{x}{e} - \binom{y}{e},\end{aligned}$$

where the second and fifth equalities above follow from $x \equiv_{k-1} r^m$ and $y \equiv_{k-1} r^n$ and the observation that $e_1, e_2 \in \Sigma^{\leqslant k-1}$ in the summations. □

The converse of the above proposition does not hold. In other words, for $k \geqslant 3$, $xy \equiv_k yx$ does not necessarily imply that x and y share a common $(k-1)$-binomial root.

Example 7.5. Let $x = aba$ and $y = baaaab$. Now $y \equiv_2 x^2$, by simply counting the occurrences of subwords of length at most two:

$$a : 4,\ b : 2,\ aa : 6,\ ab : 4,\ ba : 4,\ bb : 1$$

By the above proposition we have $xy \equiv_3 yx$, as can also be verified by counting the occurrences of subwords.

We shortly show in Example 7.7 that the converse of the above proposition does not always hold. Namely, we show that $xy \equiv_k yx$ does not imply that x and y necessarily share a common $(k-1)$-binomial root. We first make the following observation.

Lemma 7.6. *Let $s, r \in \Sigma^*$ and $\ell \geqslant 1$ such that $s^\ell \equiv_k r^\ell$. Then $s \equiv_k r$.*

Proof. Let $e \in \Sigma^k$. Notice first that for all $\ell \geqslant 1$ and $r \in \Sigma^*$

$$\binom{r^\ell}{e} = \sum_{i=1}^{|e|} \binom{\ell}{i} \sum_{\substack{e=e_1\cdots e_i\\ e_j\in\Sigma^+}} \binom{r}{e_1}\cdots\binom{r}{e_i},$$

where $\binom{\ell}{i} = 0$ whenever $i > \ell$. We prove the claim by induction on k. The claim is trivially true for $k = 1$. Assume that $s^\ell \equiv_{k+1} r^\ell$, so that $s \equiv_k r$ by the induction hypothesis. For all $e \in \Sigma^{k+1}$ we have

$$\begin{aligned}\binom{r^\ell}{e} &= \binom{\ell}{1}\binom{r}{e} + \sum_{i=2}^{|e|}\binom{\ell}{i}\sum_{\substack{e=e_1\cdots e_i\\ e_j\in\Sigma^+}}\binom{r}{e_1}\cdots\binom{r}{e_i}\\ &= \binom{\ell}{1}\binom{r}{e} + \sum_{i=2}^{|e|}\binom{\ell}{i}\sum_{\substack{e=e_1\cdots e_i\\ e_j\in\Sigma^+}}\binom{s}{e_1}\cdots\binom{s}{e_i}\\ &= \binom{s^\ell}{e} - \binom{\ell}{1}\left(\binom{s}{e} - \binom{r}{e}\right),\end{aligned}$$

where, in the second equality, we use the induction hypothesis, and observe that $|e_j| \leqslant k$ in the summation. Since $\binom{r^\ell}{e} = \binom{s^\ell}{e}$, we conclude that $\binom{r}{e} = \binom{s}{e}$. □

We are in the position to give a counter-example to the converse of Proposition 7.4.

Example 7.7. Let $x = aaababbaba$, $y = aaabbaaabbbaaab$, and $r = aabab$. We see that $x \equiv_2 r^2$ and $y \equiv_2 r^3$, whence $xy \equiv_3 yx$ by Proposition 7.4. One can further check that $xy \equiv_4 yx$, as the values $\binom{xy}{w} = \binom{yx}{w}$ for $w \in \{a,b\}^4$ are as follows:

$aaaa$: 1365,	$aaab$: 1302,	$aaba$: 1294,	$aabb$: 897,
$abaa$: 1241,	$abab$: 1106,	$abba$: 880,	$abbb$: 571,
$baaa$: 713,	$baab$: 700,	$baba$: 644,	$babb$: 447,
$bbaa$: 498,	$bbab$: 453,	$bbba$: 329,	$bbbb$: 210.

Furthermore $x \not\equiv_3 r^2$, as $\binom{x}{bba} = 9$ while $\binom{r^2}{bba} = 5$. Assume for a contradiction that x and y have a common 3-binomial root s. It follows that $|s|$ divides $\gcd(|x|,|y|) = 5$. Since x and y contain both letters, it follows that $|s| = 5$, and $x \equiv_3 s^2$, $y \equiv_3 s^3$. By the above lemma we have $s \equiv_2 r$. As $\binom{r}{ba} = 1$ and $|r|_b = 2$ we have that s ends with bab. It follows that $r = s$, which is not possible. Thus x and y do not have a common 3-binomial root.

It is unknown to us whether $xy \equiv_3 yx$ implies the existence of a common 2-binomial root of x and y.

Further, we present a characterization of commutation among words of equal length.

Proposition 7.8. *Let $x, y \in \Sigma^*$ with $|x| = |y|$. Then $xy \equiv_k yx$ if and only if $x \equiv_{k-1} y$.*

Proof. Note that $x \equiv_{k-1} y$ implies $xy \equiv_k yx$ by Proposition 7.4. We shall prove the converse by induction on k. Note that the case of $k = 2$ follows from applying Proposition 7.3 with $|x| = |y|$. Assume that the claim holds for some $k \geqslant 2$ and suppose $xy \equiv_{k+1} yx$. It follows that $xy \equiv_k yx$ so that $x \equiv_{k-1} y$ by induction. Let

then $a, b \in \Sigma$ and $e \in \Sigma^{k-1}$. We have

$$\binom{xy}{aeb} = \binom{x}{aeb} + \binom{y}{aeb} + \binom{x}{a}\binom{y}{eb} + \binom{x}{ae}\binom{y}{b} + \sum_{\substack{e_1e_2=e \\ e_1,e_2\in\Sigma^+}} \binom{x}{ae_1}\binom{y}{e_2b} \text{ and}$$

$$\binom{yx}{aeb} = \binom{y}{aeb} + \binom{x}{aeb} + \binom{y}{a}\binom{x}{eb} + \binom{y}{ae}\binom{x}{b} + \sum_{\substack{e_1e_2=e \\ e_1,e_2\in\Sigma^+}} \binom{y}{ae_1}\binom{x}{e_2b}.$$

Putting $\binom{xy}{aeb} = \binom{yx}{aeb}$ and noting that $\binom{y}{ae_1}\binom{x}{e_2b} = \binom{x}{ae_1}\binom{y}{e_2b}$ for all terms in the summation (as $x \equiv_{k-1} y$), we obtain, after rearranging,

$$|x|_a \left(\binom{y}{eb} - \binom{x}{eb}\right) = |x|_b \left(\binom{y}{ae} - \binom{x}{ae}\right).$$

Note that the above equation holds for all $a, b \in \Sigma$ and $e \in \Sigma^{k-1}$. Assume without loss of generality that $|x|_a \neq 0$. Letting $e = e_0 \cdots e_{k-2}$ and repeatedly applying the above (to possibly different letters a, b and words $e \in \Sigma^{k-1}$), we obtain

$$\begin{aligned}
\binom{y}{eb} - \binom{x}{eb} &= \left(\binom{y}{ae_0\cdots e_{k-2}} - \binom{x}{ae_0\cdots e_{k-2}}\right) \frac{|x|_b}{|x|_a} \\
&= \left(\binom{y}{aae_0\cdots e_{k-3}} - \binom{x}{aae_0\cdots e_{k-3}}\right) \frac{|x|_{e_{k-2}}}{|x|_a} \frac{|x|_b}{|x|_a} \\
&= \cdots \\
&= \left(\binom{y}{a^\ell} - \binom{x}{a^\ell}\right) \frac{|x|_b \prod_{i=0}^{k-2} |x|_{e_i}}{|x|_a^{k-1}} = 0,
\end{aligned}$$

since $\binom{y}{a^k} = \binom{x}{a^k}$, as $x \equiv_1 y$. It thus follows that $\binom{y}{eb} = \binom{x}{eb}$ for all $b \in \Sigma$ and $e \in \Sigma^{k-1}$, and therefore, $x \equiv_k y$. □

Corollary 7.9. *Let $k \geqslant 2$ and $x, y \in \Sigma^*$. If $xy \equiv_k yx$, then there exist $\ell, n \in \mathbb{N}$ such that $x^\ell \equiv_{k-1} y^n$.*

Proof. Since $xy \equiv_k yx$ it follows that $x^\ell y^n \equiv_k y^n x^\ell$ for all $\ell, n \in \mathbb{N}$. We may choose $\ell = |y|$ and $n = |x|$, whence $|x^{|y|}| = |x||y| = |y^{|x|}|$. By the above proposition we have that $x^\ell \equiv_{k-1} y^n$ as was claimed. □

It seems that the full characterization of $\mathrm{Sol}(xy, yx)$ is quite a tricky problem. We are left with the open problem:

Problem 7.10. *Characterize when, for two words $x, y \in \Sigma^*$, we have $xy \equiv_k yx$.*

7.2 On Conjugacy in $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$

Next we consider conjugacy in our monoids of interest. We are thus interested in the set of solutions to the equation $xz = zy$ in the k-abelian and k-binomial monoids. For the k-abelian monoid, we obtain a satisfactory characterization of when there exists $z \in \Sigma^*$ such that $xz \sim_k zy$. In the k-binomial monoid we find the situation to be quite complicated, and only briefly consider solutions to this equation.

Note that $xz \sim_k zy$ and $xz \equiv_k zy$ both imply that $|x| = |y|$.

7.2.1 Conjugacy in $\Sigma^*/\sim_k$

Recall that in the monoid Σ^* we have $xz = zy$ if and only if there exist words $p, q \in \Sigma^*$ such that $x = pq$, $y = qp$, and $z \in (pq)^*p$.

We first make a straightforward observation. Let $k \geqslant 1$ and $x, y \in \Sigma^*$. Assume further that there exists $z \in \Sigma^*$ such that $xz \sim_k zy$. Then there exists a word $z' \in \Sigma^*$ of length at least $k-1$ with $xz' \sim_k z'y$. Indeed, we may take $z' = x^n z$ for some suitably large n, whence $x \cdot x^n z = x^n \cdot xz = x^n z \cdot y$.

Proposition 7.11. *Let $x, y \in \Sigma^+$. Then there exists $z \in \Sigma^*$ such that $xz \sim_k zy$ if and only if y has a conjugate y' (in Σ^*) such that $y'p_{k-1}(y') \sim_k xp_{k-1}(x)$.*

Proof. Assume first that there exists a conjugate y' of y such that $y'p_{k-1}(y') \sim_k xp_{k-1}(x)$. There thus exist $p, q \in \Sigma^*$ such that $y = qp$ and $y' = pq$. Let $z = pqp$, whence $y'z = zy$. We may write $z = p_{k-1}(y')Z = p_{k-1}(x)Z$ for some $Z \in \Sigma^*$, since z begins with y'. Now $y'z = y'p_{k-1}(y')Z \sim_k xp_{k-1}(x)Z = xz$, so that $zy \sim_k xz$.

Assume then that there exists a word $z \in \Sigma^*$ such that $xz \sim_k zy$. We may assume that $|z| \geqslant k-1$ by the discussion in the beginning of this subsection. Let $Z = \mathrm{pref}_{k-1}(z)$ and $Z' = \mathrm{suff}_{k-1}(z)$. Now $xz \sim_k zy$ implies that

$$\Psi_k(xZ) = \Psi_k(Z'y),$$

$\mathrm{pref}_{k-1}(xZ) = Z$, and $\mathrm{suff}_{k-1}(Z'y) = Z'$. If $|x| \leqslant k-1$, then x is a period of Z and y is a left period of Z'. It follows that there is a conjugate y_0 of y which is a period of Z'. Let us write $y_0Y = Z'y$. Now $\Psi_k(xZ) = \Psi_k(Z'y) = \Psi_k(y_0Y)$. By applying Lemma 7.1 to $u = xZ$ and $v = y_0Y$, we have that x and y_0, and thus y, are conjugates. We may thus take $y' = x$ in the claim.

Assume then that $|x| \geqslant k-1$, and let $X = \mathrm{pref}_{k-1}(x)$ and $Y' = \mathrm{suff}_{k-1}(y)$. We thus have $Z = X$ and $Z' = Y'$. Further, $\Psi_k(xX) = \Psi_k(Y'y)$. We infer that X is a factor of $Y'y$ and thus there exists a conjugate y' of y beginning with X, so that $y'\,\mathrm{pref}_{k-1}(y') = y'X$. Now $\mathrm{suff}_{k-1}(xX) = \mathrm{suff}_{k-1}(y'X)$. Further, $\Psi_k(Y'y) = \Psi_k(y'X)$, since all factors of y^ω of length k occur equally many times in $Y'y$, and the same holds for $y'X$. It follows that $\Psi_k(y'X) = \Psi_k(xX)$ so that $y'X \sim_k xX$, as claimed. This concludes the proof. □

7.2.2 On Conjugacy in $\Sigma^*/\equiv_k$

As observed previously, $xz \equiv_k zy$ implies that $x \equiv_1 y$ in any case. For $k \geqslant 2$ and assuming that such a word z exists, then z cannot contain any letters not occurring in x and y. Indeed, if $|x|_c = |y|_c = 0$, $|z|_c \geqslant 1$, and $|x|_a \geqslant 1$, then $\binom{xz}{ac} > \binom{z}{ac}$, but $\binom{zy}{ac} = \binom{z}{ac}$. In this subsection, we consider conjugacy only in $\Sigma^*/\equiv_2$.

Proposition 7.12. *Let $x, y \in \mathbb{B}^*$. Then there exists $z \in \mathbb{B}^*$ such that $xz \equiv_2 zy$ if and only if $x \equiv_1 y$ and $\gcd(|x|_a, |x|_b)$ divides $\binom{x}{ab} - \binom{y}{ab}$.*

Proof. Indeed, assume first there exists z such that $xz \equiv_2 zy$. It immediately follows that $x \equiv_1 y$. We also have

$$\binom{x}{ab} + \binom{x}{a}\binom{z}{b} + \binom{z}{ab} = \binom{xz}{ab} = \binom{zy}{ab} = \binom{y}{ab} + \binom{z}{a}\binom{y}{b} + \binom{z}{ab},$$

which implies that $\binom{x}{ab} - \binom{y}{ab} = |z|_a|y|_b - |z|_b|x|_a = |z|_a|x|_b - |z|_b|x|_a$. It now follows that $\gcd(|x|_a, |x|_b)$ divides $\binom{x}{ab} - \binom{y}{ab}$.

Let $d = \gcd(|x|_a, |x|_b)$ and assume that $x \equiv_1 y$ and $\binom{x}{ab} - \binom{y}{ab} = kd$ for some $k \in \mathbb{Z}$. By Bezout's identity there exist $i, j \in \mathbb{Z}$, such that $kd = i|x|_b - j|x|_a$. Here we may assume that $i, j \geqslant 0$ since otherwise we may replace i with $h|x|_a + i$ and j with $h|x|_b + j$ for a suitably large h. We claim that $z = a^i b^j$ satisfies $\binom{xz}{ab} = \binom{zy}{ab}$. Indeed, $\binom{x}{ab} - \binom{y}{ab} = i|x|_b - j|x|_a$ which is equivalent to

$$\binom{x}{ab} + |z|_b|x|_a + \binom{z}{ab} = \binom{y}{ab} + |z|_a|x|_b + \binom{z}{ab}.$$

The latter is equivalent to $\binom{xz}{ab} = \binom{zy}{ab}$ as seen above. By Lemma 2.5, we have $\binom{xz}{ba} = \binom{zy}{ba}$ and, since $y \equiv_1 x$, we have $xz \equiv_2 zy$ as claimed. □

Example 7.13. Let $x = aabaaaabbbab$ and $y = bbaababaaaba$. As $y \equiv_1 x$ and $\gcd(|x|_a, |x|_b) = 1$, there exists $z \in \Sigma^*$ such that $xz \equiv_2 zy$. Now $\binom{x}{ab} - \binom{y}{ab} = 16$ and $3 \cdot |x|_b - 2 \cdot |x|_a = 1$; therefore, the proof above gives us, for example, $z = a^{48}b^{32}$. Note that also $z' = b^2$ satisfies $xz \equiv_2 zy$.

On the other hand, if $x = aabb$ and $y = abab$, we have $x \equiv_1 y$ and $\gcd(|x|_a, |x|_b) = 2$, but 2 does not divide $\binom{x}{ab} - \binom{y}{ab} = 1$. Thus x and y are not 2-binomial conjugate, in other words, $xz \not\equiv_2 zy$ for all $z \in \Sigma^*$.

We now discuss the generalization of the above characterization for larger alphabets. As seen in Example 2.6, we have $xz \equiv_2 zy$ if and only if $x \equiv_1 y$ and $\binom{xz}{ab} = \binom{zy}{ab}$ for each pair of letters $a, b \in \Sigma$ with $a < b$. An equivalent form of the latter requirement is, by the above example, that $\binom{x}{ab} - \binom{y}{ab} = |z|_a|x|_b - |z|_b|x|_a$ for all pairs of letters $a, b \in \Sigma$, $a \lhd b$. We thus obtain a system of linear equations.

Let us formalize what we mean by the above. Let $x, y \in \Sigma^*$ and assume that $x \equiv_1 y$. Assume further that each letter of Σ occurs in x. Fix an ordering on Σ and define the vector $\mathbf{D}_{x,y}$ indexed by pairs of letters $a, b \in \Sigma$, $a \lhd b$, defined as follows: $\mathbf{D}_{x,y}[(a,b)] = \binom{x}{ab} - \binom{y}{ab}$. Let then M_x be a $\binom{|\Sigma|}{2} \times |\Sigma|$-matrix, where the rows are indexed by pairs $a, b \in \Sigma$, $a < b$, and the columns by Σ, defined as $M_x[(a,b), a] = |x|_b$, $M_x[(a,b), b] = -|x|_a$, and $M_x[(a,b), c] = 0$ for $c \neq a, b$. Let $\vec{\mathbf{X}}$ be a vector of $|\Sigma|$ unknowns indexed by the letters $a \in \Sigma$. We consider solutions to the equation

$$M_x \vec{\mathbf{X}} = \mathbf{D}_{x,y}. \tag{7.1}$$

Let us give a brief example of the entities defined above.

Example 7.14. Let $\Sigma = \{0, 1, 2\}$ and let $x, y \in \Sigma^*$ such that $x \equiv_1 y$ and $|x|_a \geqslant 1$ for each $a \in \Sigma$. Then Equation (7.1) is defined as

$$\begin{pmatrix} |x|_1 & -|x|_0 & 0 \\ |x|_2 & 0 & -|x|_0 \\ 0 & |x|_2 & -|x|_1 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{X}}[0] \\ \vec{\mathbf{X}}[1] \\ \vec{\mathbf{X}}[2] \end{pmatrix} = \begin{pmatrix} \binom{x}{01} - \binom{y}{01} \\ \binom{x}{02} - \binom{y}{02} \\ \binom{x}{12} - \binom{y}{12} \end{pmatrix}.$$

Observe that for any word $z \in \Sigma^*$ we have

$$M_x \Psi(z)^T = \sum_{c \in \Sigma} M_x[(a,b), c] \cdot |z|_c = (|x|_b|z|_a - |x|_a|z|_b)_{(a,b), a \lhd b}. \tag{7.2}$$

Now, for x and y as defined above, if there exists $z \in \Sigma^*$ such that $xz \equiv_k zy$, then $\vec{\mathbf{X}} = \Psi(z)^T$ is a solution to Equation (7.1). Indeed, recall that

$$|x|_b|z|_a - |x|_a|z|_b = \binom{x}{ab} - \binom{y}{ab} = D_{x,y}[(a,b)].$$

On the other hand, if $\vec{\mathbf{X}}$ is a solution to Equation (7.1) having non-negative entries, then the word $z = \prod_{a\in\Sigma} a^{\vec{\mathbf{X}}[a]}$ is a solution to $xz \equiv_2 zy$.

We may now characterize 2-binomial conjugacy over arbitrary alphabets.

Proposition 7.15. *Let $x, y \in \Sigma^*$ and assume that each letter of Σ occurs in x. Then there exists $z \in \Sigma^*$ such that $xz \equiv_2 zy$ if and only if $x \equiv_1 y$ and Equation (7.1) has solution $\vec{\mathbf{X}}$ having integer entries.*

Proof. If there exists z such that $xz \equiv_2 zy$, then $\Psi(z)^T$ is an (non-negative) integer solution to the equation, as was observed above.

Conversely, assume that $\vec{\mathbf{X}}$ is an integer solution to Equation (7.1). *A priori*, some entries of $\vec{\mathbf{X}}$ could be negative. Observe now that plugging $z = x$ in Equation (7.2), we have $M_x\Psi(x)^T = \vec{\mathbf{0}}$.[1] Thus, for each $n \geqslant 0$, $\vec{\mathbf{X}} + n\Psi(x)^T$ is also an integer solution to the equation. Moreover, taking n large enough, each entry is a non-negative integer, since all entries of $\Psi(x)$ are assumed to be positive. Now the word $z = \prod_{a\in\Sigma} a^{\vec{\mathbf{X}}[a]+n|x|_a}$ satisfies $xz \equiv_2 zx$ (and is well-defined), as was observed above. □

As seen by the above result, characterizing k-binomial conjugacy of two words is already quite involved even for $k = 2$. We have not considered word combinatorial characterizations for this case. Neither have we attempted characterizing k-binomial conjugacy for larger values of k. We leave open the following problem.

Problem 7.16. *Characterize when, for words $x, y, z \in \Sigma^*$, we have $xz \equiv_k zy$.*

7.3 On Systems of Equations

In this section we consider systems of equations over the monoids $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$. The basic notion studied here is the so-called *compactness property* of semigroups, defined as follows.

Definition 7.17. A semigroup S is said to possess the *compactness property* if any system of equations E over a finite number of variables has a finite equivalent subsystem E'.

Famously, the free monoid Σ^* possesses the compactness property, as was proved in [4] and [42] independently. The latter also shows that free groups possess the compactness property. In [44] it is shown, employing Redei's Theorem [90] among other arguments, that all commutative semigroups possess the compactness property. Thus, for example, for each $x, y \in \Sigma^{k-1}$, the subsemigroup $(x\Sigma^* \cap \Sigma^* y)/\sim_k$ of $\Sigma^*/\sim_k$ possesses the compactness property, since it is commutative by Proposition 7.2.

Nevertheless, not all semigroups possess compactness property. For example, neither the monoid of finite languages, nor the so-called *bicyclic monoid*, nor the *Baumslag–Solitar group* possess the compactness property. For the first result, see [61], the latter two are shown in [44].

We remark that the compactness property must cover also the *inconsistent* systems of equations, that is, systems that admit no solutions. In other words, a

[1] In fact, it is not hard to verify (compare to Proposition 7.3) that $\mathrm{Ker}(M_x) = \mathrm{Span}(\Psi(x))$.

system E of equations, whether it has a solution or not, must have an equivalent finite subsystem. There exist semigroups for which each consistent system of equations has an equivalent finite subsystem, but for which there exists an inconsistent system, all finite subsystems of which are consistent [100].

For our main purpose, the following result gives us what we need.

Theorem 7.18 ([64, Chapter 13])**.** *Let R be a commutative* Noetherian ring[2] *containing an identity element. If a semigroup S can be embedded into a subsemigroup of matrices over R, then S possesses the compactness property.*

As the ring $\mathbb{Z}$ of integers is Noetherian, we see that the monoids $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$ possess the compactness property. (The former follows from Proposition 3.28, while the latter from [93, Corollary 18].)

Theorem 7.19. *For any $k \geqslant 1$, the monoid $\Sigma^*/\sim_k$ has the compactness property.*

Theorem 7.20. *For any $k \geqslant 1$, the monoid $\Sigma^*/\equiv_k$ has the compactness property.*

Now if a semigroup possesses the compactness property, then any independent system of equations is finite. We turn to the interesting question on how large such a system over our monoids can be. The aspect of considering sizes of independent systems of equations in semigroups has been previously treated, e.g., in the paper [54]. See also [75], and references therein, concerning the free semigroup.

Let us recall some relevant results from the literature. For this, let S possess the compactness property and define $F_S(n) = c$, where c is the maximum size of an independent system $E \subseteq \Xi^+ \times \Xi^+$ of equations without constants, and with the number $|\Xi|$ of variables n, if this constant exists. Otherwise $F_S(n) = UB$ (short for unbounded).

There exists a commutative group G for which $F_G(1) = UB$ (see the following example). On the other hand, for a finitely generated commutative group G, there exists a constant c_G depending on G such that $F_G(n) = c_G n$. These two results appear in [54]. For free groups G, we have $F_G(6) = UB$ [64, Chapter 13]. For the free monoids, it is not known if $F_{\Sigma^*}(n) = UB$ for some n. There exists independent systems of equations with n variables having size $\Theta(n^4)$ [54]. Note that Theorem 2.15 gives $F_{\Sigma^*}(3) \leqslant 18$ and it is conjectured that $F_{\Sigma^*}(3) = 3$ [25].

Example 7.21 ([54])**.** For distinct prime numbers $p_1, \ldots, p_k$, let $n = p_1 \cdots p_k$ and $n_i = n/p_i$ for each $i = 1, \ldots k$. Consider the system E of k equations $e_i : x^{n_i} = 1$, where $i = 1, \ldots, k$. Now, for a fixed j, $e^{2\pi i n_j/n}$ is a solution to each e_i with $i \neq j$, but not for e_j. Thus E is independent.

The rest of this section is devoted to showing that, for both $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$, there exists a uniform upper bound on the number of equations of an independent system of equations. For k fixed, the size of an independent system of equations has a polynomial upper bound with respect to the number of unknowns. On the other hand, the upper bound is exponential when the number $|\Xi|$ of unknowns Ξ is fixed and k is allowed to vary. We remark that these bounds do not depend on the size of the alphabet Σ, when the equations have no constants, that is, the system of equations is a subset of $\Xi^+ \times \Xi^+$.

[2] A ring R is Noetherian if for each chain $I_0 \subseteq I_1 \subseteq I_2 \cdots$ of ideals there exists $n_0 \geqslant 0$ such that $I_n = I_{n_0}$ for each $n \geqslant n_0$. For us it suffices to know that $\mathbb{Z}$ is a Noetherian ring.

Remark 7.22. Let S be a semigroup with a finite generating set G. Any system E of equations (with or without constants) over S may be modified into a system without constants by identifying each generator $g \in G$ with a new variable X_g. The set of solutions of the original system are obtained from the solutions to the modified system by choosing the solutions where $X_g \mapsto g$ for each generator. Further, if the number of equations in an independent system of equations without constants using n variables is at most $f(n)$ for each n, then the number of equations in an independent system of equations is at most $f(n + \#G)$.

In particular, the monoids $\Sigma^*/\sim_k$ and $\Sigma^*/\equiv_k$ are finitely generated with $|\Sigma|$ generators. Thus, by considering systems of equations without constants, we obtain bounds for systems of equations where constants are allowed.

7.3.1 On Independent Systems of Equations over $\Sigma^*/\sim_k$

We set some notation. As before, for any $k \geqslant 0$, we set $s_k(u) = \mathrm{suff}_{\min\{|u|,k\}}(u)$ and $p_k(u) = \mathrm{pref}_{\min\{|u|,k\}}(u)$. Let now $u \in \Xi^+$. We define the $\frac{|\Xi|^{k+1}-|\Xi|}{|\Xi|-1}$-dimensional vector $\vec{\mathbf{u}}$ (whose components are indexed by non-empty words in $\Xi^{\leqslant k}$) associated to u as

$$\vec{\mathbf{u}} = (|u|_X)_{X\in\Xi} \times (|u|_{XYZ})_{X,Z\in\Xi,Y\in\Xi^{\leqslant k-2}}.$$

Let then $h : \Xi \to \Sigma^*$ be a non-erasing morphism. For any word $w \in \Sigma^k$ we define the $\frac{|\Xi|^{k+1}-|\Xi|}{|\Xi|-1}$-dimensional vector $\vec{\mathbf{h}}_w$ (the entries indexed by the non-empty words in $\Xi^{\leqslant k}$) associated to h and w as follows: For each $X \in \Xi$ we set $\vec{\mathbf{h}}_w[X] = |h(X)|_w$. For each $X, Z \in \Xi$, $Y \in \Xi^{\leqslant k-2}$ we set

$$\vec{\mathbf{h}}_w[XYZ] = \left|s_{k-1}(h(XY))p_{k-1-|h(Y)|}(h(Z))\right|_w$$

if $|h(Y)| < k-1$, else we set $\vec{\mathbf{h}}_w[XYZ] = 0$.

The following observation is crucial to our endeavors towards showing an upper bound on the size of independent systems of equations.

Lemma 7.23. *Let $u \in \Xi^+$, $h : \Xi \to \Sigma^+$ be a non-erasing morphism, and $w \in \Sigma^k$. Then $|h(u)|_w = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{u}}$.*

Proof. Let $u = X_0 \cdots X_n$, where $X_i \in \Xi$ for each $i = 0, \ldots, n$. To avoid cluttering the text, we let $\widehat{Y} = h(Y)$ for each $Y \in \Xi^{\leqslant k}$. Now

$$\begin{aligned}
|h(u)|_w &= \sum_{i=0}^{n} |\widehat{X}_i|_w + \sum_{i=1}^{n} \left|s_{k-1}(\widehat{X}_{i-1}) \cdot p_{k-1}(\widehat{X}_i \cdots)\right|_w \\
&= \sum_{X\in\Xi} |u|_X \cdot |\widehat{X}|_w + \sum_{i=1}^{n} \left|s_{k-1}(\widehat{X}_{i-1}) \cdot p_{k-1}(\widehat{X}_i \cdots)\right|_w. \\
&= \sum_{X\in\Xi} \vec{\mathbf{u}}[X] \cdot \vec{\mathbf{h}}_w[X] + \sum_{i=1}^{n} \left|s_{k-1}(\widehat{X}_{i-1}) \cdot p_{k-1}(\widehat{X}_i \cdots)\right|_w. \qquad (7.3)
\end{aligned}$$

The first sum counts the number occurrences of w occurring in the images $\widehat{X}$ of letters X occurring in u, while the second sum counts the number of occurrences

of w not appearing in an image of a letter. We focus on the second sum in Equation (7.3). Observe that, for each i, we may write $p_{k-1}(\widehat{X}_i \cdots) = p_{k-1}(\widehat{X}_i \cdots \widehat{X}_{r_i})$, where r_i equals the minimal index $s \geqslant i$ such that $|\widehat{X}_i \cdots \widehat{X}_s| \geqslant k-1$ if such an index s exists, otherwise $r_i = n$. As h is non-erasing, we have that $|X_i \cdots X_{r_i}| \leqslant k-1$, whence $X_{i-1}X_i \cdots X_{r_i} \in \Xi^{\leqslant k}$ for all $i = 1, \ldots, n$. We may thus rewrite the second sum in the above as

$$\sum_{i=1}^{n} \left|s_{k-1}(\widehat{X}_{i-1}) \cdot p_{k-1}(\widehat{X}_i \cdots \widehat{X}_{r_i})\right|_w = \sum_{\substack{X,Z \in \Xi, Y \in \Xi^* \\ |\widehat{Y}\widehat{Z}| \geqslant k-1 \\ |\widehat{Y}| < k-1}} |u|_{XYZ} \cdot \left|s_{k-1}(X) p_{k-1}(YZ)\right|_w.$$

Let now $Y \in \Xi^{\leqslant k-2}$ and $Z \in \Xi$ such that $|\widehat{Y}| < k-1$ and $|\widehat{Y}\widehat{Z}| \geqslant k-1$. Consider the occurrences of w occurring as a *proper overlap* of $\widehat{X} \cdot \widehat{Y} \cdot \widehat{Z}$, for some $X \in \Xi$. By a proper overlap we mean that the occurrence of w starts in $\widehat{X}$ and ends in $\widehat{Z}$. The number of such occurrences is easily seen to equal $|s_{k-1}(\widehat{X}\widehat{Y}) p_{k-1-|\widehat{Y}|}(\widehat{Z})|_w$. Now each occurrence of w that does not occur in an image $\widehat{X}$ of a letter $X \in \Xi$ occurs as a proper overlap of $\widehat{X} \cdot \widehat{Y} \cdot \widehat{Z}$ for some $X, Z \in \Xi$, $Y \in \Xi^{\leqslant k-2}$ with $|\widehat{Y}| < k-1$. Thus, to count the number of occurrences of w as an overlap, we count the number of occurrences of w as a proper overlap in all images of factors of u of length at most k:

$$\sum_{\substack{X,Z \in \Xi \\ Y \in \Xi^{\leqslant k-2}}} |u|_{XYZ} \cdot \left|s_{k-1}(\widehat{X}\widehat{Y}) p_{k-1-|\widehat{Y}|}(\widehat{Z})\right|_w \cdot \delta_{|\widehat{Y}|<k-1} = \sum_{\substack{X,Z \in \Xi \\ Y \in \Xi^{\leqslant k-2}}} \vec{\mathbf{u}}[XYZ] \cdot \vec{\mathbf{h}}_w[XYZ]$$

Note that the summation goes through all words $W \in \Xi^{\leqslant k}$ with $|W| \geqslant 2$. Thus, plugging the above into Equation (7.3), we obtain $|h(u)|_w = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{u}}$, as claimed. □

For an equation $e : u = v$, we define $\vec{\mathbf{e}} = \vec{\mathbf{u}} - \vec{\mathbf{v}}$. In the following, we let $\vec{\mathbf{e}}^{\perp}$ denote the orthogonal complement (subspace) of the vector $\vec{\mathbf{e}}$. We characterize the solutions to an equation in $\Sigma^*/\sim_k$.

Lemma 7.24. *Let $e : u = v$ be an equation over Ξ and $h : \Xi \to \Sigma^+$ a non-erasing morphism. Then h is a solution to e in $\Sigma^*/\sim_k$ if and only if $p_{k-1}(h(p_{k-1}(u))) = p_{k-1}(h(p_{k-1}(v)))$ and $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}^{\perp}$ for all $w \in \Sigma^*$.*

Proof. Since h is non-erasing we have $p_{k-1}(h(u)) = p_{k-1}(h(v))$ if and only if $p_{k-1}(h(p_{k-1}(u))) = p_{k-1}(h(p_{k-1}(v)))$. Furthermore by the above lemma we have, for all $w \in \Sigma^k$, $|h(u)|_w - |h(v)|_w = 0$ if and only if $\vec{\mathbf{h}}_w \cdot \vec{\mathbf{u}} - \vec{\mathbf{h}}_w \cdot \vec{\mathbf{v}} = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{e}} = 0$ for all $w \in \Sigma^*$. Thus $h(u) \sim_k h(v)$ if and only if $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}^{\perp}$ and $p_{k-1}(h(p_{k-1}(u))) = p_{k-1}(h(p_{k-1}(v)))$. The claim follows. □

Theorem 7.25. *The number of equations in an independent system of equations over the semigroup $\Sigma^+/\sim_k$ with variables Ξ is at most $|\Xi^{\leqslant k}| + \binom{|\Xi^{\leqslant k-1}|-1}{2}$.*

Proof. Let Ξ be a finite set of variables and let E be an independent system of equations with variables Ξ; $E = \{e_i : u_i = v_i \mid u_i, v_i \in \Xi^+\}_{i \in I}$. (Note that E is finite by Theorem 7.19). Assume first that the set $\mathrm{Sol}(E)$ is not empty. By the above lemma, if h is a solution to E, then $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}_i^{\perp}$ for all $w \in \Sigma^k$ and $i \in I$. As $\bigcap_{i \in I} \vec{\mathbf{e}}_i^{\perp} = U$ is a finite dimensional vector space, there exists a finite

set $F = \{e_1, \ldots, e_f\}$ of equations, where $f < |\Xi^{\leqslant k}|$, such that $U = \bigcap_{i=1}^{f} \vec{\mathbf{e}}_i^{\perp}$. Let further $F' \subseteq E$ be a set chosen so that, for each pair of distinct non-empty words $z, z' \in \Xi^{\leqslant k-1}$, precisely one equation $e : u = v$ from E having $p_{k-1}(u) = z$ and $p_{k-1}(v) = z'$ occurs in F'. There are at most $\binom{|\Xi^{\leqslant k-1}|-1}{2}$ such equations.

We now claim that the subsystem $E' = F \cup F'$ is equivalent to E (and thus $E' = E$). Assume that h is a solution to E' over $\Sigma^*/\sim_k$ and let $e : u = v$ be an equation in E. Since h is a solution to the system of equations F, it follows that $\vec{\mathbf{h}}_w \in U \subseteq \vec{\mathbf{e}}^{\perp}$, where U is defined as above, for all $w \in \Sigma^k$. Further, since there exists an equation $e' : u' = v'$ in F' having $p_{k-1}(u) = p_{k-1}(u')$, $p_{k-1}(v) = p_{k-1}(v')$, (and $h(u') \sim_k h(v')$) it follows that $p_{k-1}(h(p_{k-1}(u))) = p_{k-1}(h(p_{k-1}(v)))$ (if $p_{k-1}(u) = p_{k-1}(v)$ then the previous conclusion is trivial). It follows that $h(u) \sim_k h(v)$ by the above lemma, and thus h is a solution to e.

Assume then that the system E has no solutions. Let $e \in E$ and let $G = E \setminus \{e\}$. Now there exists a solution to G, since otherwise E is not independent (or G is empty). Applying the above to G, we obtain the finite, equivalent system of equations E' as defined above. Now $E' \cup e$ has no solutions (otherwise a solution to $E' \cup e$ would be a solution to E), and thus $E = E' \cup e$. The claim follows. □

We have not attempted giving a lower bound on the maximal number of equations in an independent system of equations for the k-abelian monoid. We propose the following question.

Question 7.26. What is the maximal size of an independent system of equations (without constants) with n variables in the k-abelian monoid?

7.3.2 On Independent Systems of Equations over $\Sigma^*/\equiv_k$

Our approach for upper bounding the size of an independent system of equations over $\Sigma^*/\equiv_k$ is identical to the approach for $\Sigma^*/\sim_k$. The considerations here turn out to be slightly simpler.

Let us fix some notation. Let $k \geqslant 1$ be fixed. Consider a word $u \in \Xi^+$ and define the $\frac{|\Xi|^{k+1}-|\Xi|}{|\Xi|-1}$-dimensional vector $\vec{\mathbf{u}}$ as

$$\vec{\mathbf{u}} = \left(\binom{u}{Y}\right)_{Y \in \Xi^{\leqslant k} \setminus \{\varepsilon\}}.$$

For any non-erasing morphism $h : \Xi \to \Sigma^*/\equiv_k$ we define, for each word $w \in \Sigma^{\leqslant k}$, the $\frac{|\Xi|^{k+1}-|\Xi|}{|\Xi|-1}$-dimensional vector $\vec{\mathbf{h}}_w$ (components indexed by non-empty words in $\Xi^{\leqslant k}$) as

$$\vec{\mathbf{h}}_w[Y] = \sum_{\substack{w = w_1 \cdots w_\ell \\ w_j \in \Sigma^+}} \binom{h(Y_1)}{w_1} \cdots \binom{h(Y_\ell)}{w_\ell},$$

for each $Y = X_1 \cdots X_\ell \in \Xi^{\leqslant k}$ with $X_i \in \Xi$ for all $i = 1, \ldots, \ell$. Note that $\vec{\mathbf{h}}_e[Y] = 0$ for all Y for which $|Y| > |e|$, as e does not have a factorization into $|Y|$ non-empty words.

The following lemma is crucial in the following endeavors.

Lemma 7.27. *Let $h : \Xi \to \Sigma^*/\equiv_k$ be a non-erasing morphism, $u \in \Xi^+$, and $w \in \Sigma^{\leqslant k}$. Then $\binom{h(u)}{w} = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{u}}$.*

Proof. Again, to avoid cluttering the text, we set $h(X) = \widehat{X}$ for each $X \in \Xi$. Let $u = X_1 \cdots X_n$, where $X_i \in \Xi$ for each $i = 1, \ldots, n$. For any subset S of $\{1, \ldots, n\}$, by the sequence $S_1, \ldots, S_{|S|}$ we mean the sequence of elements of S arranged in increasing order. Now, for each $w \in \Sigma^{\leqslant k}$, we observe that

$$\binom{h(u)}{w} = \sum_{\substack{S \subseteq [1,n] \\ |S| \leqslant |w|}} \sum_{\substack{w = w_1 \cdots w_{|S|} \\ w_j \in \Sigma^+}} \binom{\widehat{X}_{S_1}}{w_1} \cdots \binom{\widehat{X}_{S_{|S|}}}{w_{|S|}}.$$

Indeed, for each occurrence of w as a subword, there exists a subset $S \subseteq [1, n]$ of length at most k such that $w = w_1 \cdots w_{|S|}$, where $w_i \in \Sigma^+$ and the indices of w_i in u are a subset of the indices of $\widehat{X}_{S_i}$ in $h(u)$. For each subset S of $[1, n]$ having $|S| \geqslant |e|$, there exists no such factorization, and thus the corresponding sum contributes nothing to the total sum. Now for two subsets $S, S' \subseteq [1, n]$ having $X_{S_1} \cdots X_{S_{|S|}} = X_{S'_1} \cdots X_{S'_{|S'|}} = Y$, the corresponding sums contribute the same value. The number of distinct such sets equals $\binom{u}{Y}$. We may thus rewrite the above equation as

$$\sum_{Y \in \Xi^{\leqslant k}} \binom{u}{Y} \sum_{\substack{w = w_1 \cdots w_{|Y|} \\ w_j \in \Sigma^+}} \binom{\widehat{X}_1}{w_1} \cdots \binom{\widehat{X}_{|Y|}}{w_{|Y|}} = \sum_{Y \in \Xi^{\leqslant k}} \vec{\mathbf{h}}_w[Y] \cdot \vec{\mathbf{u}}[Y] = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{u}},$$

as claimed. □

Lemma 7.28. *Let $w : u = v$ be an equation and let $h : \Xi \to \Sigma^*/\equiv_k$ be a non-erasing morphism. Then h is a solution to e over $\Sigma^*/\equiv_k$ if and only if $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}^{\perp}$ for all $w \in \Sigma^{\leqslant k}$, where $\vec{\mathbf{e}} = \vec{\mathbf{u}} - \vec{\mathbf{v}}$.*

Proof. We have $h(u) \equiv_k h(v)$ if and only if $\binom{h(u)}{w} - \binom{h(v)}{w} = 0$ for all non-empty $w \in \Sigma^{\leqslant k}$ if and only if $\vec{\mathbf{h}}_w \cdot (\vec{\mathbf{u}} - \vec{\mathbf{v}}) = \vec{\mathbf{h}}_w \cdot \vec{\mathbf{e}} = 0$ for each $w \in \Sigma^{\leqslant k}$, by the lemma above. □

We may now bound the number of equations in an independent system.

Theorem 7.29. *The number of equations in an independent system of equations (without constants) over the semigroup $\Sigma^+/\equiv_k$ with variables Ξ is at most $|\Xi^{\leqslant k}|$.*

Proof. Let $E = \{e_i : u_i = v_i\}_{i \in I}$ be an independent system of equations over Ξ. Assume again that $\mathrm{Sol}(E)$ is not empty. The case of $\mathrm{Sol}(E)$ having no solutions is analogous to the k-abelian case. Now h is a solution to E if and only if $\vec{\mathbf{h}}_w \in \bigcap_{e \in E} \vec{\mathbf{e}}^{\perp} = U$ for all $w \in \Sigma^{\leqslant k}$. Since U is a finite dimensional vector space, there exist equations $e_1, \ldots, e_f \in E$ such that $U = \cap_{i=1}^{f} \mathbf{e_i}^{\perp}$, where $f \leqslant |\Xi^{\leqslant k}| - 1$. We claim that $E' = \{e_1, \ldots, e_f\}$ is an equivalent subsystem of E.

Let $e \in E$. Let then h be a solution to E'. It follows that $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}_i^{\perp}$ for all $i = 1, \ldots, f$, so that $\vec{\mathbf{h}}_w \in U$ for all $w \in \Sigma^*$. Furthermore $\vec{\mathbf{h}}_w \in \vec{\mathbf{e}}^{\perp}$ which is equivalent to h being a solution to e by the above lemma. □

We again leave open the following question:

Question 7.30. What is the maximal number of equations in an independent system of equations in the k-binomial monoid?

Chapter 8

Asymptotic Abelian Complexities

The last two chapters of this thesis focus on the k-abelian equivalence in infinite words. This chapter focuses on the 1-abelian complexity, that is, abelian complexity of infinite words, or more precisely, of certain morphic binary words. The analysis of binary words is already quite intricate, even though the abelian complexity in this case is strongly related to the so-called *balancedness* of the infinite word (see Definition 8.1).

We first consider the pure morphic binary words. We recall relevant results from the literature and observe that the limit superior abelian complexities of pure morphic binary words are almost all known. Our first main result is the analysis of the asymptotic abelian complexities of the remaining unknown cases of pure morphic binary words. The words studied here admit *fluctuating* abelian complexity, that is, the limit inferior and limit superior abelian complexities are of different order. In studying limit superior abelian complexities, we make use of the notion of *derivated words* of uniformly recurrent words (see Definition 8.9). We associate the limit superior abelian complexity of a word to the limit superior balance function of one of its derivated words. This allows us to invoke the classification result of [2] dealing with the *balance function* (defined in the following) of primitive pure morphic words. We thus complete the classification of the limit superior abelian complexities of pure morphic binary words: given a binary morphism φ admitting a fixed point $\mathbf{x} = \varphi(a)^\omega$, the asymptotic behaviour of the limit superior abelian complexity of $\mathbf{x}$ can be easily computed.

We then consider morphic binary words which are not pure morphic, having asymptotic abelian complexity of the order $\mathcal{O}(n^r)$ for some rational r for which $0 < r < 1$. The main result here is the following. For each pair $p, q \in \mathbb{N}$, $p < q$, we give a sequence $(\mathbf{w}_n)_{n \geqslant 0}$ of morphic binary words each having asymptotic abelian complexity $\Theta(n^{p/q})$, while the factor complexities satisfy $\mathcal{C}_{\mathbf{w}_{i+1}}(n) = o(\mathcal{C}_{\mathbf{w}_i}(n))$ for each $i \geqslant 0$. The morphic words in these sequences are defined as binary images of pure morphic words having larger and larger alphabet sizes. Consequently, the analysis of the factor complexities and abelian complexities become quite intricate.

This results of this chapter appear in the article [109].

8.1 Background

Let us first recall relevant results from the literature. We define a complexity function closely related to the abelian complexity. For this we need the following notation. For an infinite word $\mathbf{w} \in \Sigma^{\mathbb{N}}$ and a letter $a \in \Sigma$, we define

$$\max_{\mathbf{w},a}(n) = \max\{|u|_a \mid u \in F_n(\mathbf{w})\}.$$

The function $\min_{\mathbf{w},a} : \mathbb{N} \to \mathbb{N}$ is defined analogously.

Definition 8.1. Let $\mathbf{w} \in \Sigma^{\mathbb{N}}$. The *balance function* $B_{\mathbf{w}}$ of $\mathbf{w}$ is defined as $B_{\mathbf{w}}(n) = \max\{\max_{\mathbf{w},a}(n) - \min_{\mathbf{w},a}(n) \mid a \in \Sigma\}$.

It is straightforward to verify that, for $\mathbf{w} \in \mathbb{B}^{\mathbb{N}}$, we have $\mathcal{C}^{\mathrm{ab}}_{\mathbf{w}}(n) = B_{\mathbf{w}}(n) + 1$ for all $n \in \mathbb{N}$.

The following deep result of B. Adamczewski [2] is the first stepping stone of the classification of the abelian complexities of morphic words. The result classifies the asymptotic growth of the balance function of primitive pure morphic words. The asymptotic behaviour of $\mathcal{U}^{\mathrm{ab}}_{\mathbf{x}}$ for binary words $\mathbf{x}$ can be extracted from the above, as was done in [12]. We state the result here because we make use of it in our later considerations. Before we do so, however, we recall some basic notions of linear algebra.

When talking about eigenvalues of a morphism φ, we mean eigenvalues of A_φ. The multiplicity of the eigenvalue λ in the minimal polynomial of A_φ is denoted by α_λ. We let $\theta_1, \theta_2, \ldots, \theta_n$ be the distinct eigenvalues of φ ordered in such a way that $|\theta_i| \geqslant |\theta_{i+1}|$ and if $|\theta_i| = |\theta_{i+1}|$, then $\alpha_{\theta_i} \geqslant \alpha_{\theta_{i+1}}$. For a primitive φ, the Perron–Frobenius theorem (see, e.g., [85]) implies that $\theta_1 \in \mathbb{R}$, $\theta_1 > 1$, $\theta_1 > |\theta_2|$, and $\alpha_{\theta_1} = 1$. The eigenvalue θ_1 is called the *Perron-eigenvalue* of φ. We also make use of the eigenvalue θ_2, which can be seen as the second most significant eigenvalue of φ.

In the following we let $\alpha_2 = \alpha_{\theta_2} - 1$.

Theorem 8.2. *([2] (as formulated in [12])) Let $\mathbf{x}$ be a fixed point of a primitive morphism φ. Then the following hold:*

1. *If $|\theta_2| < 1$ then $B_{\mathbf{x}}(n) = (\mathcal{O} \cap \hat{\Omega})(1)$;*
2. *If $|\theta_2| > 1$ then $B_{\mathbf{x}}(n) = (\mathcal{O} \cap \hat{\Omega})((\log n)^{\alpha_2} n^{\log_{\theta_1} |\theta_2|})$;*
3. *If $|\theta_2| = 1$ and θ_2 is not a root of unity, then $B_{\mathbf{x}}(n) = (\mathcal{O} \cap \hat{\Omega})((\log n)^{\alpha_2+1})$;*
4. *If $|\theta_2| = 1$ and θ_2 is a root of unity, then either*
   - *$B_{\mathbf{x}}(n) = (\mathcal{O} \cap \hat{\Omega})((\log n)^{\alpha_2+1})$, or*
   - *$B_{\mathbf{x}}(n) = (\mathcal{O} \cap \hat{\Omega})((\log n)^{\alpha_2})$,*

   according to whether a certain constant $A_{\varphi,\mathbf{x}}$ equals zero or not, respectively.

We refer the interested reader to [2] for more on computing the constant $A_{\varphi,\mathbf{u}}$ mentioned in the last item above.

From the above we immediately infer that $\mathcal{U}^{\mathrm{ab}}_{\mathbf{x}}(n)$, for a primitive pure morphic binary word $\mathbf{x}$, is of order $\Theta(1)$, $\Theta(\log n)$, or $\Theta(n^{\log_{\theta_1} \theta_2})$ (since $\alpha_2 = 0$). In [12]

Blanchet-Sadri, Fox, and Rampersad go on to study abelian complexities of fixed points of non-primitive binary morphisms. Before stating their result, we recall a straightforward characterization of such morphisms.

Proposition 8.3. *Let φ be a non-primitive binary morphism which admits an infinite fixed point $\mathbf{y} = \varphi^\omega(a)$. Then either $\varphi(a) \in aa^+$ and $\varphi(b) \in \Sigma_2^*$, or φ is of the form*

$$\varphi(a) \in a\Sigma^* b\Sigma^* \text{ and } \varphi(b) \in b^*, \tag{8.1}$$

where, if $\varphi(b) = \varepsilon$, then $|\varphi(a)|_a \geqslant 2$. Further, $\mathbf{y}$ is ultimately periodic if and only if $\varphi(a) \in aa^+$ or φ is of the form (8.1) and satisfies one of the following conditions: $\varphi(a) \in ab^+$, $\varphi(b) = \varepsilon$, or $\varphi(a) = (ab^r)^s a$ and $\varphi(b) = b$ for some $r, s \geqslant 1$.

Theorem 8.4. *([12]) Let φ be a non-primitive binary morphism as in (8.1) with $\varphi(b) = b^k$ for some $k \geqslant 1$. Suppose further that φ admits an aperiodic infinite fixed point $\mathbf{y} = \varphi^\omega(a)$. Then the following holds:*

1. *If $k = 1$ and $\varphi(a)$ ends with b, then $\mathcal{C}_{\mathbf{y}}^{ab}(n) = \Theta(n)$;*
2. *If $k \geqslant 2$ then*
   - $\mathcal{C}_{\mathbf{y}}^{ab}(n) = \Theta(n)$ *if* $|\varphi(a)|_a > k$,
   - $\mathcal{C}_{\mathbf{y}}^{ab}(n) = \Theta(n/\log n)$ *if* $|\varphi(a)|_a = k$, *and*
   - $\mathcal{C}_{\mathbf{y}}^{ab}(n) = \Theta\left(n^{\log_k |\varphi(a)|_a}\right)$ *if* $|\varphi(a)|_a < k$.

Observe that if $\mathbf{x} \in \Sigma^{\mathbb{N}}$ is (ultimately) periodic, then $\mathcal{C}_{\mathbf{x}}^{\mathrm{ab}} = \Theta(1)$. It is straightforward to check that the words fixed by non-primitive morphisms whose asymptotic (upper) abelian complexities are not yet classified are as in (8.1), where $k = 1$ and $\varphi(a)$ ends with a. More precisely, φ is of the form

$$\varphi(a) = ab^{k_1}ab^{k_2}\cdots ab^{k_s}a, \quad \varphi(b) = b, \tag{8.2}$$

where $k_i \geqslant 0$ for all $i = 1, \ldots, s$ and there exist i, j such that $k_i < k_j$. Our aim is to complete the classification by proving the following in Section 8.2:

Theorem 8.5. *Let φ be as in (8.2) and $\mathbf{y} = \varphi^\omega(a)$. Then $\mathcal{U}_{\mathbf{y}}^{ab}(n) = \Theta(\log n)$ and $\mathcal{L}_{\mathbf{y}}^{ab}(n) = \Theta(1)$.*

In particular, morphisms of the form (8.2) are the only non-primitive binary morphisms whose fixed points have upper and lower abelian complexities of different orders of growth.

8.2 Completing a Classification of Pure Morphic Binary Words

In this section we prove Theorem 8.5. We first consider the lower abelian complexities of words fixed by morphisms of the form (8.2). After this, we focus on the upper abelian complexity. We achieve this by finding a connection between

the upper abelian complexities and the balance functions of some *derivated words* of these words (see Subsection 8.2.1).

Let us fix the notation for the rest of this section. We let φ be a morphism as in (8.2) and we let $\mathbf{Y} = \varphi^\omega(a)$. We also let k_m (resp., k_M) denote the minimal (resp., maximal) of the exponents k_i, $i = 1, \dots, s$, in (8.2).

We start with some elementary properties of the word $\mathbf{Y}$.

Lemma 8.6. *The word $\mathbf{Y}$ has the following properties.*

1. *The set $\mathfrak{R}_{\mathbf{Y}}(a)$ equals $\mathfrak{R}_{\varphi(a)}(a) = \{ab^{k_i}a \mid i = 1, \dots, s\}$.*
2. *For any fixed $m \in \mathbb{N}$, we have $\mathbf{Y} \in \{\varphi^m(a)b^{k_i} \mid i = 1, \dots, s\}^\omega$.*
3. *The word $\mathbf{Y}$ is linearly recurrent (so, in particular, uniformly recurrent).*

Proof. 1. Suppose this is not the case, $ab^ra \in F(\mathbf{Y}) \setminus F(\varphi(a))$ for some $r \in \mathbb{N}$. Suppose that $ab^ra \in \varphi(w)$, where $w \in F(\varphi^t(a))$, $t \geqslant 1$, t is the least such integer, and w is the shortest such factor of $\mathbf{Y}$. Clearly w is not a letter, and since $|ab^ra|_a = 2$, we have $w = ab^sa$ for some $s \geqslant 0$. Now $ab^ra \in F(\varphi(a)b^s\varphi(a))$. Since $ab^ra \notin \mathfrak{R}_{\varphi(a)}(a)$, it follows that $s = r$, that is, $ab^ra \in F(\varphi^{t-1}(a))$, a contradiction.

2. The claim is true for $m = 0$ by the previous item. Suppose then that the claim is true for some $m \geqslant 0$; $\mathbf{Y} = \prod_{i=1}^{\infty} \varphi^m(a)b^{r_i}$, $r_i \in \{k_1, \dots, k_s\}$ for all $i \geqslant 1$. But then $\mathbf{Y} = \varphi(\mathbf{Y}) = \prod_{i=1}^{\infty} \varphi^{m+1}(a)b^{r_i}$.

3. Let $u_m = \varphi^m(a)$ for each $m \geqslant 0$. It is straightforward to conclude that u_{m+1} contains each factor of length $|u_m|$ of $\mathbf{Y}$ for each $m \in \mathbb{N}$. Further, by item 2, any factor of length $2|u_{m+1}| + k_M$ contains u_m as a factor. The claim follows since $|u_{m+1}| \leqslant |u_1||u_m|$. □

Remark 8.7. Observe that $\lim_{n\to\infty} \frac{\min_{\mathbf{Y},c}(n)}{n} > 0$ for both $c \in \mathbb{B}$. This is immediate by item 2 (case $m = 1$) in the above lemma together with $|\varphi(a)|_a, |\varphi(a)|_b > 0$. Observe that the limit always exists as the sequence $(\min_{\mathbf{Y},c}(n))_{n\geqslant 0}$ is *subadditive*. In fact, since $\mathbf{Y}$ is linearly recurrent, we have $\lim\limits_{n\to\infty} \min\limits_{v\in F_n(\mathbf{Y})} \frac{|v|_u}{n} = \lim\limits_{n\to\infty} \max\limits_{v\in F_n(\mathbf{Y})} \frac{|v|_u}{n}$ for any $u \in F(\mathbf{Y})$ ([32, Theorem 15] and [35, Proposition 7.2.10]). For us however, the first observation above is enough.

We are ready to show that the lower abelian complexity of $\mathbf{Y}$ is bounded.

Lemma 8.8. *Let φ and $\mathbf{Y}$ be as fixed previously. Then $\mathcal{L}^{ab}_{\mathbf{Y}}(n) = \Theta(1)$.*

Proof. Let $u_m = \varphi^m(a)$ for each $m \in \mathbb{N}$. We claim that $\mathcal{C}^{\mathrm{ab}}_{\mathbf{Y}}(|u_m|)$ is bounded by a constant depending only on φ. Let now $v \in F(\mathbf{Y})$ have length $|u_m|$. By item 2 in the above lemma, it follows that v is a factor of $u_mb^{k_i}u_m$ for some $i \in \{1, \dots, s\}$. In other words we have $v = qb^rp$, where p (resp., q) is a (possibly empty) prefix (resp., suffix) of u_m and $r \leqslant k_M$. On the other hand, we have $u_m = puq$, for some $u \in F_r(\mathbf{Y})$. We thus conclude that

$$|u_m|_b = |p|_b + |u|_b + |q|_b \leqslant |p|_b + r + |q|_b = |v|_b \leqslant |u_m|_b + r \leqslant |u_m|_b + k_M.$$

It follows that $\mathcal{C}^{\mathrm{ab}}_{\mathbf{Y}}(|u_m|) \leqslant k_M + 1$ for all $m \in \mathbb{N}$, and the claim follows. □

The rest of this section is devoted to the upper abelian complexity of $\mathbf{Y}$. We develop the tools needed in the following.

8.2.1 On Derivated Words of Uniformly Recurrent Words

We recall the definition of a *derivated word* of a uniformly recurrent word from [31]. We then study the derivated words of uniformly recurrent pure morphic words, and remark a slight generalization of a result from [31].

Definition 8.9. Let $\mathbf{x} \in \Sigma^{\mathbb{N}}$ be uniformly recurrent and $p \in \operatorname{pref}(\mathbf{x})$ be non-empty. By the above discussion, we may write uniquely $\mathbf{x} = \prod_{i=0}^{\infty} q_i$, where $q_i \in \Re_{\mathbf{x}}(p)p^{-1}$ for each $i \in \mathbb{N}$. Let $\Delta_{p,\mathbf{x}}$ be an alphabet with $|\Delta_{p,\mathbf{x}}| = |\Re_{\mathbf{x}}(p)|$, and let $\pi_{p,\mathbf{x}} : \Delta_{p,\mathbf{x}} \to \Re_{\mathbf{x}}(p)p^{-1}$ be a bijection. The *derivated word of* $\mathbf{x}$ *with respect to* p, denoted by $D_p(\mathbf{x})$, is defined as $D_p(\mathbf{x}) = \prod_{i=0}^{\infty} \pi_{p,\mathbf{x}}^{-1}(q_i) \in \Delta^{\omega}$.

In the following, we order the elements of $\Re_{\mathbf{x}}(p) = \{p_1, \ldots, p_d\}$ in the order they occur for the first time in $\mathbf{x}$. We then set $\Delta_{p,\mathbf{x}} = \{\delta_1, \ldots, \delta_d\}$ and fix $\pi_{p,\mathbf{x}}$ by $\pi_{p,\mathbf{x}}(\delta_i) = p_i$, $i = 1, \ldots, d$. We often omit the subscripts from $\Delta_{p,\mathbf{x}}$ and $\pi_{p,\mathbf{x}}$ whenever the word $\mathbf{x}$ and prefix p are clear from context.

Note that $\pi_{p,\mathbf{x}}$ can be interpreted as a morphism $\pi_{p,\mathbf{x}} : \Delta_{p,\mathbf{x}}^* \to \Sigma^*$, whence $x = \pi_{p,\mathbf{x}}(D_p(\mathbf{x}))$. Note also that, since $\mathbf{x}$ is uniformly recurrent, then so is $D_p(\mathbf{x})$. The following result is a minor generalization of [31, Proposition 5.1].

Proposition 8.10. *Let $\rho : \Sigma^* \to \Sigma^*$ be a morphism admitting a uniformly recurrent fixed point $\mathbf{x} = \rho^{\omega}(a)$. Let p be a non-empty prefix of $\mathbf{x}$. Then $D_p(\mathbf{x})$ is primitive pure morphic. Moreover, a primitive morphism μ fixing $D_p(\mathbf{x})$ may be constructed when $\Re_{\mathbf{x}}(p)$ is known.*

For the sake of completeness, we give a proof, which is essentially the same as for Durand's result, as suggested by Jarkko Peltomäki (personal communication).

To this end we partially define $\pi_{p,\mathbf{x}}^{-1} : \Sigma^* \to \Delta_{p,x}^*$ as follows. Let $q \in F(\mathbf{x})$ be a return to p in $\mathbf{x}$, that is, $q \in \Sigma^+$, $qp \in F(\mathbf{x})$, and $p \in \operatorname{pref}(qp)$. We can then write uniquely (by Proposition 2.7) $qp = q_1 \cdots q_n p$, where $q_i \in \Re_{\mathbf{x}}(p)p^{-1}$ for each $i = 1, \ldots, n$. We define $\pi_{p,\mathbf{x}}^{-1}(q) = \pi_{p,\mathbf{x}}^{-1}(q_1) \cdots \pi_{p,\mathbf{x}}^{-1}(q_n) \in \Delta^*$. Otherwise we leave $\pi_{p,\mathbf{x}}^{-1}$ undefined.

Proof. For ease of notation, let $\Delta = \Delta_{p,\mathbf{x}}$ and $\pi = \pi_{p,\mathbf{x}}$. We first define a mapping $\mu : \Delta \to \Delta^*$ as follows. For each $\delta \in \Delta$, we set $\mu(b) = \pi^{-1}\theta\pi(\delta)$.

We first verify that μ is well-defined, that is, we show that $\theta(q)$ is a return to p in x whenever $q \in \Re_{\mathbf{x}}(p)p^{-1}$. By definition, $qp \in F(\mathbf{x})$, so that $\theta(q)\theta(p) \in F(\mathbf{x})$. Note that $\theta(q) \neq \varepsilon$, as q begins with a. Furthermore, we have $p \in \operatorname{pref}(\theta(p))$, as can easily be deduced using the properties of θ. Thus $\theta(q)p \in F(\mathbf{x})$, $p \in \operatorname{pref}(\theta(q)p)$, and $|\theta(q)p| > |p|$. In other words, $\theta(q)p$ is a complete return to p in $\mathbf{x}$. Thus π^{-1} is defined on $\theta(q)$, as was to be shown.

We now consider μ as the morphism $\mu : \Delta^* \to \Delta^*$. Our next step is to show that $\mu^{\omega}(\delta_1) = D_p(\mathbf{x})$. For this, we first note that $\theta\pi = \pi\mu$. Indeed, both mappings are morphisms (as compositions of morphisms) and they agree on the letters of Δ. Further, $\theta\pi(\delta_1) = \theta(p_1) \in \operatorname{pref}(\mathbf{x})$. It follows that $\mu(\delta_1) = \pi^{-1}\theta\pi(\delta_1)$ begins with δ_1. Moreover, for all $n \in \mathbb{N}$ and $s = 1, \ldots, d$,

$$\pi\mu^n(\delta_i) = \theta\pi\mu^{n-1}(\delta_i) = \cdots = \theta^n\pi(\delta_i) = \theta^n(p_i),$$

which implies that $\lim_{n\to\infty} \pi(\mu^n(\delta_1)) = \mathbf{x}$. Hence, there exists a unique fixed point $\mathbf{z} = \mu^{\omega}(\delta_1)$ satisfying, by the above, $\pi(\mathbf{z}) = \mathbf{x}$. Since the factorization of $\mathbf{x}$ into complete first returns to p is unique by Proposition 2.7, we have $\mathbf{z} = D_p(\mathbf{x})$.

Finally, we show that μ is primitive. Note that $D_p(\mathbf{x})$ is uniformly recurrent, so it suffices to show that $|\mu^n(\delta_i)|$ obtains arbitrarily large values for each $\delta_i \in \Delta$. Now $|\pi\mu^n(\delta_i)| = |\theta^n\pi(\delta_i)|$ and $\pi(\delta_i) = p_i$ always begins with the letter a. Thus $|\theta^n\pi(\delta_i)| \geqslant |\theta^n(a)|$ so that $|\pi\mu^n(\delta)|$ obtains arbitrarily large values, and hence so does $|\mu^n(\delta_i)|$. □

The ingredients of the proof of the above result are essential to our later considerations. In particular, the construction of the primitive morphism $\mu : \Delta_{p,\mathbf{x}} \to \Delta_{p,\mathbf{x}}^*$ satisfying $\mu^\omega(\delta_1) = D_p(\mathbf{x})$ is analyzed in our setting. We clarify the above construction by an example.

Example 8.11. Let $\varphi(a) = aabab^2a$ and $\varphi(b) = b$ so that φ is of the form (8.2). Let $\mathbf{y} = \varphi^\omega(a)$. By Lemma 8.6, $\Re_{\mathbf{y}}(a) = \{aa, aba, ab^2a\}$ so we set $\Delta_{a,\mathbf{y}} = \{\delta_1, \delta_2, \delta_3\}$. Now π is defined by $\pi(\delta_i) = ab^{i-1}$ for each $i = 1, 2, 3$. The primitive morphism μ is now defined by

$$\mu(\delta_i) = \pi^{-1}\varphi\pi(\delta_i) = \pi^{-1}(aabab^2ab^{i-1}) = \delta_1\delta_2\delta_3\delta_i$$

for each $i = 1, 2, 3$. The incidence matrix A_μ of μ is thus

$$A_\mu = \begin{pmatrix} |\mu(\delta_1)|_{\delta_1} & |\mu(\delta_2)|_{\delta_1} & |\mu(\delta_3)|_{\delta_1} \\ |\mu(\delta_1)|_{\delta_2} & |\mu(\delta_2)|_{\delta_2} & |\mu(\delta_3)|_{\delta_2} \\ |\mu(\delta_1)|_{\delta_3} & |\mu(\delta_2)|_{\delta_3} & |\mu(\delta_3)|_{\delta_3} \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Note that, for example, the word fixed by the morphism $a \mapsto ab^2abab^3a$, $b \mapsto b$, has the same derivated word $D_a(\mathbf{y})$.

In the above example the constructed morphism μ is uniform with length $4 = |\varphi(a)|_a$. This is no coincidence when the morphism φ is of the form (8.2). Indeed, consider the construction of μ for our word $\mathbf{Y}$ and prefix a. We have $\Delta = \Delta_{a,\mathbf{Y}} = \{\delta_1, \ldots, \delta_d\}$. Now $\Re_{\mathbf{Y}}(a)a^{-1} \subseteq ab^*$ so we may define $\pi = \pi_{a,\mathbf{Y}}$ by $\pi(\delta_i) = ab^{r_i} \in \Re_{\varphi(a)}(a)a^{-1}$, for each $i = 1, \ldots, d$. By the definition of μ in the above construction, we obtain

$$\mu(\delta_i) = \pi^{-1}\varphi\pi(\delta_i) = \pi^{-1}(\varphi(a)b^{r_i}) = \pi^{-1}(\varphi(a)a^{-1}ab^{r_i}) = p\delta_i, \tag{8.3}$$

where $p = \pi^{-1}(\varphi(a)a^{-1})$. The morphism μ is thus uniform with length $|\varphi(a)|_a$.

Proposition 8.12. *We have* $B_{D_a(\mathbf{Y})}(n) = \mathcal{O}(\log n)$.

Proof. We aim to show that μ has eigenvalues $|\varphi(a)|$ and 1, both with multiplicities 1 (as roots of the minimal polynomial of A_μ). The claim then follows by the fourth point of Theorem 8.2. Indeed, the incidence matrix A_μ is of the form

$$A_\mu = \left(\Psi(p)^T \mid \Psi(p)^T \mid \cdots \mid \Psi(p)^T\right) + \mathbf{I}_{d\times d} = \mathbf{A} + \mathbf{I}_{d\times d},$$

where $\Psi(p)$ is the *Parikh vector* of p in (8.3), $\mathbf{I}_{d\times d}$ is the $d \times d$ *identity matrix*, and $\mathbf{A}$ is a $d \times d$ matrix, where each column is the same vector $\Psi(p)^T$.

Now let λ be an eigenvalue of A_μ. This implies that

$$0 = \det(A_\mu - \lambda\mathbf{I}_{d\times d}) = \det(\mathbf{A} + \mathbf{I}_{d\times d} - \lambda\mathbf{I}_{d\times d}) = \det(\mathbf{A} - (\lambda - 1)\mathbf{I}_{d\times d}).$$

It is readily verified that the only eigenvalues of $\mathbf{A}$ are $\sum_{i=1}^{d} |p|_{\delta_i} = |p|$ and 0 from which it follows that the eigenvalues of A_μ are $|\varphi(a)|_a$ and 1.

We now claim that the minimal polynomial of A_μ is $x^2 - (|p|+2)x + |p| + 1$. Indeed, it is straightforward to check that $\mathbf{A}^2 = |p|\mathbf{A}$, from which it follows that $A_\mu^2 - (|p|+2)A_\mu + (|p|+1)\mathbf{I}_{d\times d} = \mathbf{0}$. Now the minimal polynomial of A_μ is of degree 2 implying that the eigenvalue 1 has multiplicity 1, as was to be shown. □

8.2.2 The Upper Abelian Complexity of Y

We are now ready to prove that the upper abelian complexity of $\mathbf{Y}$ is of order $\mathcal{U}_{\mathbf{Y}}^{\mathrm{ab}} n = \Theta(\log n)$. For this, we bound the upper abelian complexity of $\mathbf{Y}$ in terms of the asymptotic balance function of $D_a(\mathbf{Y})$. We then show that for infinitely many m, we have $\mathcal{C}_{\mathbf{Y}}^{\mathrm{ab}}(m) = \Theta(\log m)$.

Lemma 8.13. *We have* $\mathcal{U}_{\mathbf{Y}}^{ab}(n) = \mathcal{O}(\log n)$.

Proof. Let μ be the primitive morphism having $D_a(\mathbf{Y})$ as a fixed point, and let $\pi = \pi_{a,\mathbf{Y}}$ as defined in the previous subsection. Let $n \in \mathbb{N}$ with $\mathcal{C}^{\mathrm{ab}}(n) > 2$. Now there exists a factor $u_M \in F(\mathbf{Y})$ of length n such that $|u_M|_a = \max_{\mathbf{Y},a}(n)$ and u_M begins with a. Indeed, if $|v|_a = \max_{\mathbf{Y},a}(n)$ with $v \in b^r a\Sigma^*$, then, by considering a factor $vw \in F(\mathbf{Y})$ with $|w| = r$, we have $|\operatorname{suff}_n(vw)|_a = |v|_a + |w|_a$, and $\operatorname{suff}_n(vw)$ begins with a. By a similar argument, there exists a factor $u_m \in F(\mathbf{Y})$ of length n such that $|u_m|$ begins with a and $|u_m|_a \leqslant \min_{\mathbf{Y},a}(n) + 1$. Observe that now $|u_M|_a > |u_m|_a$ by the choice of n.

As u_m begins with a, there exists a factor x of $D_a(\mathbf{Y})$ of length $|u_m|_a$ such that u_m is a prefix of $\pi(x)$, whence $|u_m|_a = |\pi(x)|_a$ and $|u_m|_b \leqslant |\pi(x)|_b \leqslant |u_m|_b + k_M$. (Recall that between consecutive occurrences of a in $\mathbf{Y}$, there are at most k_M bs.) Similarly, we may write $u_M = \pi(z)v$, where z is factor of $D_a(\mathbf{Y})$ of length $|u_m|_a$, and v begins with a. We now have $|u_M|_a - |u_m|_a = |v|_a$ so that $\mathcal{C}_{\mathbf{Y}}^{\mathrm{ab}}(n) \leqslant |v|_a + 2$. We claim that $|v| = \mathcal{O}(\log n)$ to conclude the proof. As $|\pi(z)| + |v| = |u_M| \leqslant |\pi(x)|$ and $|\pi(z)|_a = |\pi(x)|_a$, we have $|v| \leqslant |\pi(x)| - |\pi(z)| = |\pi(x)|_b - |\pi(z)|_b$. Moreover, x and z are factors of $D_a(\mathbf{Y})$ of equal length. By Proposition 8.12,

$$|\pi(x)|_b - |\pi(z)|_b = \sum_{i=1}^{d} r_i(|x|_{a_i} - |z|_{a_i}) \leqslant dk_M B_{D_a(\mathbf{Y})}(|x|) = \mathcal{O}(\log |x|),$$

where r_i is defined as $\pi_{a,\mathbf{Y}}(a_i) = b^{r_i}$, and $r_i \leqslant k_M$ by definition. Finally, $|x| = |u_m|_a \leqslant n$ and thus $|v| = \mathcal{O}(\log n)$. The claim follows. □

We now proceed to show that $\mathcal{U}_{\mathbf{Y}}^{\mathrm{ab}}(n) = \Omega(\log n)$. To this end, let $\varphi(a) = gauah$ for some $g, u, h \in \Sigma^*$. We construct a sequence of factors of y as follows. Let $u_0^{g,h} = a$ and, for all $n \geqslant 0$, define $u_{n+1}^{g,h} = g^{-1}\varphi(u_n^{g,h})h^{-1}$. Note that the sequence is well-defined, as $\varphi(u_n) \in ga\Sigma_2^* ah$ for each $n \geqslant 1$. We now make some observations on the words in the sequence. In the following we let $\alpha = |\varphi(a)|_a$ and $\beta = |\varphi(a)|_b$ for ease of notation.

Lemma 8.14. *For all* $n \in \mathbb{N}$

- $|u_n^{g,h}|_a = (1 - \frac{|gh|_a}{\alpha - 1})\alpha^n + \frac{|gh|_a}{\alpha - 1}$,
- $|u_n^{g,h}|_b = \frac{\beta}{\alpha-1}(1 - \frac{|gh|_a}{\alpha-1})(\alpha^n - 1) + (\frac{\beta}{\alpha-1}|gh|_a - |gh|_b)n$.

Proof. Define $\Phi_{g,h}(v) = (|v|_a, |v|_b, -|gh|_a, -|gh|_b)^T \in \mathbb{Z}^4$ for all $v \in \Sigma^*$. Consider the following 4×4 matrix (in block form)

$$\widehat{A}_\varphi = \begin{pmatrix} A_\varphi & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I} = \mathbf{I}_{2\times 2}$ and $\mathbf{0} = \mathbf{0}_{2\times 2}$ are the 2×2 identity matrix and zero matrix, respectively. It is readily verified that for any $v \in a\Sigma^* a$, we have $\widehat{A}_\varphi \Phi_{g,h}(v) = \Phi_{g,h}(g^{-1}\varphi(v)h^{-1})$. This implies that $\widehat{A}^n_\varphi \Phi_{g,h}(a) = \Phi_{g,h}(u^{g,h}_n)$ for all $n \in \mathbb{N}$. We then have, for all $n \in \mathbb{N}$, that

$$\widehat{A}^n_\varphi = \begin{pmatrix} A^n_\varphi & \sum_{i=0}^{n-1} A^i_\varphi \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

where

$$A^n_\varphi = \begin{pmatrix} \alpha^n & 0 \\ \beta\frac{\alpha^n-1}{\alpha-1} & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=0}^{n-1} A^i_\varphi = \begin{pmatrix} \frac{\alpha^n-1}{\alpha-1} & 0 \\ \beta\frac{\alpha^n-1-n(\alpha-1)}{(\alpha-1)^2} & n \end{pmatrix},$$

by straightforward induction. We finally have, for all $n \in \mathbb{N}$, that

$$\widehat{A}^n_\varphi \Phi_{g,h}(a)[1,2] = \begin{pmatrix} \alpha^n - \frac{\alpha^n-1}{\alpha-1}|gh|_a \\ \beta\frac{\alpha^n-1}{\alpha-1} - \beta\frac{\alpha^n-1-n(\alpha-1)}{(\alpha-1)^2}|gh|_a - n|gh|_b \end{pmatrix}.$$

Rearranging the terms gives our claims. □

Recall that there exist $g_m, h_m, g_M, h_M \in \Sigma^*$ such that $\varphi(a) = g_m a b^{k_m} a h_m = g_M a b^{k_M} a h_M$, where k_m and k_M are fixed as in the beginning of this section. Let then $(u_n) = (u^{g_m,h_m}_n)$ and $(v_n) = (u^{g_M,h_M}_n)$ be sequences constructed as above. Note that $|g_m h_m|_a = |g_M h_M|_a = \alpha - 2$ and $|g_m h_m|_b - |g_M h_M|_b = k_M - k_m$ whence, by the above lemma,

$$|v_n| - |u_n| = |v_n|_b - |u_n|_b = -n|g_M h_M|_b + n|g_m h_m|_b = n(k_M - k_m) \tag{8.4}$$

for all $n \in \mathbb{N}$. We are now in the position to complete the proof of Theorem 8.5.

Proof of Theorem 8.5. By Lemma 8.8 and Lemma 8.13 it is enough to show that $\mathcal{U}^{\mathrm{ab}}_{\mathbf{Y}} n = \Omega(\log n)$.

Let $(u_n)_n$ and $(v_n)_n$ be the sequences as discussed above. Let then $u_n f_n \in F(\mathbf{Y})$, so that $|u_n f_n| = |v_n|$, that is, $|f_n| = n(k_M - k_m)$. Now, by Remark 8.7, for all large enough n there exists $\gamma < 1$ such that $|f_n|_b \leqslant \gamma |f_n|$. From (8.4) we obtain

$$\mathcal{C}^{\mathrm{ab}}_{\mathbf{Y}}(|v_n|) \geqslant |v_n|_b - |u_n f_n|_b \geqslant (1-\gamma)(k_M - k_m)n.$$

Further, from the above lemma, we have

$$|v_n| = |v_n|_a + |v_n|_b = \tfrac{|\varphi(a)|-1}{(\alpha-1)^2}\alpha^n + \mathcal{O}(n) = \Theta(\alpha^n).$$

We thus conclude that $\mathcal{C}^{\mathrm{ab}}_{\mathbf{Y}}(|v_n|) = \Omega(\log |v_n|)$, whence $\mathcal{U}^{\mathrm{ab}}_{\mathbf{Y}} n = \Omega(\log n)$. □

We have shown that aperiodic words fixed by a morphism of the form (8.2) have abelian complexity which fluctuates between constant and logarithmic growth. This completes the classification of the abelian complexities of pure morphic words fixed by non-primitive binary morphisms. Further, the classification of upper abelian complexities of pure morphic binary words is completed. Observe that the classification of lower abelian complexities remains open for primitive pure morphic binary words. We mention that the lower abelian complexities of a large family of uniform binary morphisms is obtained in [12].

The following problem is left open.

Question 8.15. Is there a classification similar to the one in Theorem 8.2 for the asymptotic orders of growth of $\mathcal{U}^{\mathrm{ab}}_{\mathbf{x}}(n)$ and $\mathcal{L}^{\mathrm{ab}}_{\mathbf{x}}(n)$ for pure morphic words $\mathbf{x}$?

Classifying the abelian complexities for primitive pure morphic words over larger alphabets remains totally open. The methods used here are specific to binary words and cannot be applied to larger alphabets directly. More precisely, the techniques rely on the equivalence of the balance function and the abelian complexity of binary words. For larger alphabets, the link is not that clear.

8.3 On Morphic Words with Common Abelian Complexities

In this section we extend our analysis to morphic binary words. In particular, we are interested in morphic words having abelian complexity of the order $\Theta(n^{p/q})$, where $p < q$.

We note that, in the case of pure morphic binary words, one can achieve such abelian complexity with both primitive and non-primitive morphisms. Primitive morphisms having their adjacency matrix of the form $\left(\begin{smallmatrix} 2^q-1 & 2^q-2^p-1 \\ 1 & 2^p+1 \end{smallmatrix}\right)$, where $p < q$, yield words with $\mathcal{U}^{\mathrm{ab}}(n) = \Theta\left(n^{p/q}\right)$ (item 2 of Theorem 8.2). On the other hand, non-primitive morphisms having their adjacency matrix of the form $\left(\begin{smallmatrix} 2^p & 0 \\ s & 2^q \end{smallmatrix}\right)$ with $p < q$ and $s \geqslant 1$, yield fixed points having $\mathcal{C}^{\mathrm{ab}}(n) = \Theta\left(n^{p/q}\right)$ (item 2 of Theorem 8.4, third point). What is worth noting is that both of the above types of words give $\Theta(n)$ factor complexity:

Lemma 8.16. *Let $\mathbf{y} = \varphi^\omega(a)$ for some binary morphism φ. If $\mathcal{U}^{ab}_{\mathbf{y}}(n) = \Theta(n^r)$ for some $r \in \mathbb{Q}$ with $0 < r < 1$, then $\mathcal{C}_{\mathbf{y}}(n) = \Theta(n)$.*

Proof. Note that $\mathbf{y}$ is necessarily aperiodic. We show that the morphism φ is *everywhere-growing* and *quasi-uniform* (for definitions see [76, 77], or [21, Definitions 4.7.35 and 4.7.39]), that is, there exists $\beta > 1$ such that $|\varphi^n(a)|, |\varphi^n(b)| = \Theta(\beta^n)$. The claim follows, as aperiodic fixed points of everywhere-growing quasi-uniform morphisms have linear factor complexity by Pansiot's result [76].

It is a simple exercise to show that a primitive morphism φ is everywhere-growing and quasi-uniform. If φ is non-primitive, then, by Theorems 8.4 and 8.5, $|\varphi(a)|_a < |\varphi(b)|_b$. In this case it is simple to see that β above equals $|\varphi(b)| > 1$. □

The main result of this section is the following.

Theorem 8.17. *For each pair $p, q \in \mathbb{N}$ with $p < q$, there exists a sequence of morphic binary words $(\mathbf{y}_s)_{s\in\mathbb{N}}$ satisfying $\mathcal{C}^{ab}_{\mathbf{y}_s}(n) = \Theta(n^{p/q})$ and $\mathcal{C}_{\mathbf{y}_{s+1}}(n) = o\left(\mathcal{C}_{\mathbf{y}_s}(n)\right)$.*

There thus exists a family of morphic binary words having the same asymptotic abelian complexities (up to a constant) while the asymptotic factor complexities are different. We shall construct such sequences for each pair $p, q \in \mathbb{N}$.

8.3.1 Words of Interest and Some Initial Observations

We first fix the notation of the remainder of the section. For convenience we use the infinite alphabet $\Sigma_{\mathbb{N}} = \{a_i \mid i \geqslant 0\}$ indexed by the natural numbers. Let also $\Sigma_s = \{a_0, \ldots, a_s\}$ and $\Gamma_s = \{a_i \mid i \geqslant s\}$. Define then the morphism $\gamma : \Sigma_{\mathbb{N}} \to \Sigma_{\mathbb{N}}^*$ by $\gamma(a_0) = a_0$ and $\gamma(a_r) = a_r a_{r-1}$ for $r \geqslant 1$. Further, for each $s \in \mathbb{N}$, we define the morphism $\sigma_s : \Sigma_{\mathbb{N}} \to \mathbb{B}^*$ by $\sigma_s(a_r) = b$ if $a_r \in \Gamma_s$ and $\sigma_s(a_r) = a$ otherwise.

Now, for each $r \geqslant 1$, the word $\gamma^\omega(a_r)$ exists. For the remainder of this section we set, for each $r \geqslant 1$, $\mathbf{X}_r = \gamma^\omega(a_r)$. Further, we let $\mathbf{X}_{s,r}$ denote the morphic word $\sigma_s(\mathbf{X}_r)$ for all $s \geqslant 0$ and $r \geqslant 1$.

Example 8.18. We illustrate the words defined above. (Here we identify a_i with i, $i = 0, 1, 2, 3$).

$$\begin{aligned}
\mathbf{X}_3 &= 3221211021101002110100100021101001000100002\cdots\\
\mathbf{X}_{1,3} &= bbbbbbbabbbabaabbbabaabaaabbbabaabaaabaaaab\cdots\\
\mathbf{X}_{2,3} &= bbbabaaabaaaaaabaaaaaaaaaaabaaaaaaaaaaaaaaab\cdots.
\end{aligned}$$

We remark that the words $\mathbf{X}_{s,s+1}$ appear in the literature as examples of morphic, but not pure morphic, words ([77], see also [21, Subsection 4.7.1]).

The aim is to prove the following proposition.

Proposition 8.19. *Let $r, s \in \mathbb{N}$ with $1 \leqslant s < r$, and let $\mathbf{x} = \mathbf{X}_{s,r}$. Then*

1. $\mathcal{C}_{\mathbf{x}}(n) = \Theta\left(n^{1+1/s}\right)$ *and*
2. $\mathcal{C}^{ab}_{\mathbf{x}}(n) = \Theta\left(n^{1-s/r}\right)$.

Let us first see how Theorem 8.17 follows from the above proposition.

Proof of Theorem 8.17. Let us fix $p, q \in \mathbb{N}$ with $1 \leqslant p < q$. For each $s \geqslant 1$, let $y_s = \mathbf{X}_{s(q-p),sq}$. By Proposition 8.19, $\mathbf{y}_s$ has $\mathcal{C}_{\mathbf{y}_s}(n) = \Theta\left(n^{1+1/s(q-p)}\right)$ and $\mathcal{C}^{\mathrm{ab}}_{\mathbf{y}_s}(n) = \Theta\left(n^{1-s(q-p)/sq}\right) = \Theta\left(n^{p/q}\right)$. It is now clear that $\mathcal{C}_{\mathbf{y}_{s+1}}(n) = o\left(\mathcal{C}_{\mathbf{y}_s}(n)\right)$. □

We need to make some observations concerning γ and the words $\mathbf{X}_r$, $r \geqslant 1$, before proving the above proposition. To tidy up notation, we define $\Lambda_{m,\ell,r} = \prod_{i=m}^{\ell} \gamma^i(a_{r-1})$ for all $m, r \in \mathbb{N}$ and $\ell \in \mathbb{N} \cup \{\infty\}$, with $r \geqslant 1$ and $\ell \geqslant m$. For technical reasons we also allow $\ell = m - 1$, and we set $\Lambda_{m,m-1,r} = \varepsilon$.

Lemma 8.20. *The following properties hold for all $r \geqslant 1$.*

1. $\mathbf{X}_r \in a_r \Sigma^\omega_{r-1}$. *In particular,* $\mathbf{X}_1 = a_1 a_0^\omega$ *and* $\mathbf{X}_{r,r} = ba^\omega$.
2. *For all $n, m \in \mathbb{N}$ with $n \geqslant m \geqslant 0$, we have*

$$\gamma^n(a_r) = a_r \Lambda_{0,n-1,r} = \gamma^m(a_r)\Lambda_{m,n-1,r} \quad \text{and} \quad \mathbf{X}_r = \gamma^n(a_r)\Lambda_{n,\infty,r}.$$

3. $F(\mathbf{X}_t) \subseteq F(\mathbf{X}_r)$ *and* $F(\mathbf{X}_{s,t}) \subseteq F(\mathbf{X}_{s,r})$ *for all $t, r, s \in \mathbb{N}$ with $1 \leqslant t \leqslant r$.*

Proof. Item 1 is clear by the definition of γ and item 2 is easily shown by induction. Item 3 is immediate by item 2. □

To further simplify notation, we define, for all $r \geqslant 1$ and $s \geqslant 0$, the functions $p_r, p_{s,r} : \mathbb{N} \to \mathbb{N}$ by $p_r(n) = |\gamma^n(a_r)|$, and $p_{s,r}(n) = |\gamma^n(a_r)|_{a_s}$. Thus, for example, $p_r(n) = \sum_{i=0}^{r} p_{i,r}(n)$.

Lemma 8.21. *Let $r, s \in \mathbb{N}$ with $r \geqslant s \geqslant 0$ and $r > 0$. Then,*

1. $p_{s,r}(n) = \binom{n}{r-s} = \frac{1}{(r-s)!} n^{r-s} + \mathcal{O}(n^{r-s-1})$ *and*
2. $p_r(n) = \sum_{i=0}^{r} \binom{n}{i} = \frac{1}{r!} n^r + \mathcal{O}(n^{r-1})$.

Proof. For each $s \in \Sigma_r$, let $\mathbf{I}_s$ denote the $(r+1) \times (r+1)$ matrix having the entry $a_{ij} = 1$ if $j = i + s$ and $a_{ij} = 0$ otherwise. It is easy to check that $\mathbf{I}_1^t = \mathbf{I}_t$ for each $t = 1, \ldots, r$, that $\mathbf{I}_1^t = \mathbf{0}$ for $t \geqslant r$, and that $\mathbf{I}_0$ is the identity matrix.

Consider then the adjacency matrix $A_{\gamma,r}$ of γ restricted to the alphabet Σ_r (the top-left entry being $|\gamma(a_0)|_{a_0}$ while the bottom-right entry being $|\gamma(a_r)|_{a_r}$). We have $A_{\gamma,r} = \mathbf{I}_0 + \mathbf{I}_1$ so that

$$A_{\gamma,r}^n = \sum_{i=0}^{r} \binom{n}{i} \mathbf{I}_1^i = \sum_{i=0}^{r} \binom{n}{i} \mathbf{I}_i.$$

The rightmost column contains the entries $|\gamma^n(a_r)|_{a_i}$ for $i = 0, \ldots, r$, whence

$$p_{s,r}(n) = A_{\gamma,r}^n[s, r] = \binom{n}{r-s} = \frac{1}{(r-s)!} n^{r-s} + \mathcal{O}(n^{r-s-1}).$$

Finally, $p_r(n) = \sum_{i=0}^{r} \binom{n}{i} = \frac{1}{r!} n^r + \mathcal{O}(n^{r-1})$. The claims follow. □

We move on to prove Proposition 8.19 in two parts.

8.3.2 Analyzing the Factor Complexity

We first analyze the factor complexity of $\mathbf{X}_{s,r}$ for any pair $1 \leqslant s < r$. Our aim is to prove Proposition 8.19, Part 1. We start with an observation, after which we give a straightforward proof of the desired result.

Lemma 8.22. *Let $0 \leqslant s \leqslant r$. Then $\sigma_s(\gamma^n(a_r))$ ends with $ba^{p_s(n-r+s)-1}$ for all $n > r - s$.*

Proof. Let s be fixed. We shall prove the claim by induction on r. The base case $r = s$ is trivial, as $\sigma_s(\gamma^n(a_s)) = ba^{p_s(n)-1}$ for all $n \geqslant 1$. Suppose the claim is true for r and consider the case of $r + 1$. Let $n > r + 1 - s$. By item 2 of Lemma 8.20, $\gamma^n(a_{r+1})$ ends with $\gamma^{n-1}(a_r)$. As $n - 1 > r - s$, the induction hypothesis asserts that $\sigma_s(\gamma^{n-1}(a_r))$ ends with $ba^{p_s(n-(r+1)+s)-1}$. We have thus completed the induction step. □

We are in the position to analyze the factor complexity. The proof is quite crude and heavily uses the structure of the defined words.

Proof of Proposition 8.19, item 1. Let $s \geqslant 1$ be fixed. We prove, by induction on r, that $\mathbf{X}_{s,r}$ has the claimed factor complexity. The base case $r = s + 1$ is a result in [77] (see also [21, Proposition 4.7.2]). Suppose then that the claim is true for

a fixed r and consider the word $\mathbf{X}_{s,r+1}$. Let us fix n and estimate the size of $F_n(\mathbf{X}_{s,r+1}) \setminus F_n(\mathbf{X}_{s,r})$. Factorize $\mathbf{X}_{s,r+1}$ into three parts

$$\mathbf{X}_{s,r+1} = \sigma_s\left(a_{r+1}\Lambda_{0,k_1,r+1}\right) \cdot \sigma_s\left(\Lambda_{k_1+1,k_2,r+1}\right) \cdot \sigma_s\left(\Lambda_{k_2+1,\infty,r+1}\right),$$

where k_1 is minimal in the sense that $p_r(k_1) \geqslant n$ and k_2 is minimal in the sense that $\sigma_s(\gamma^{k_2}(a_r))$ ends with at least n a's. By Lemma 8.21, $p_t(x) = \frac{1}{t!}x^t + \mathcal{O}(x^{t-1})$ for each $t \in \mathbb{N}$, so that $k_1 = \Theta(n^{1/r})$ and, by the above lemma, $k_2 = \Theta(n^{1/s})$.

Consider the prefix. We first note that, by Lemma 8.20, $a_{r+1}\Lambda_{0,k_1,r+1} = \gamma^{k_1+1}(a_{r+1})$. Trivially $|F_n\left(\sigma_s\left(a_{r+1}\Lambda_{0,k_1,r+1}\right)\right)| \leqslant |\gamma^{k_1+1}(a_{r+1})|$ and we obtain, by Lemma 8.21, the rough upper bound

$$|\gamma^{k_1+1}(a_{r+1})| = \tfrac{1}{(r+1)!}k_1^{r+1} + \mathcal{O}(k_1^r) = \mathcal{O}\left(n^{1+1/r}\right).$$

Consider next the factors occurring in $\sigma_s(\Lambda_{k_1+1,k_2,r+1})$. Any factor occurring in $\sigma_s(\gamma^i(a_r))$ occurs already in $\mathbf{X}_{s,r}$. By the choice of k_1, it suffices to consider factors that are of the form $\sigma_s(u_1u_2)$, where $u_1 \in \operatorname{suff}(\gamma^i(a_r))$ and $u_2 \in \operatorname{pref}(\gamma^{i+1}(a_r))$ for some i satisfying $k_1 \leqslant i < k_2$. For each such i, there are at most $n-1$ choices of u_1 and u_2, and we obtain the upper bound

$$\textstyle\sum_{i=k_1}^{k_2} n = n\mathcal{O}(n^{1/s}) = \mathcal{O}(n^{1+1/s}).$$

Finally, the factors of length n occurring in the infinite tail have already been counted previously, either as factors of $\mathbf{X}_{s,r}$, or as a prefix of $\mathbf{X}_{s,r}$ preceded by a block of a's. We conclude, by the induction hypothesis, that

$$\mathcal{C}_{\mathbf{X}_{s,r+1}}(n) = \mathcal{C}_{\mathbf{X}_{s,r}}(n) + \mathcal{O}(n^{1+1/s}) + \mathcal{O}(n^{1+1/r}) = \Theta\left(n^{1+1/s}\right).$$

□

8.3.3 Analyzing the Abelian Complexity

We now analyze the abelian complexity of $\mathbf{X}_{s,r}$ for $1 \leqslant s < r$. Our aim is to prove Proposition 8.19, Part 2, the following lemma being crucial in doing so. In what follows, for $w \in \Sigma_{\mathbb{N}}^*$ and $s \in \mathbb{N}$, we let $|w|_{\Gamma_s} = \sum_{a\in\Gamma_s} |w|_a$.

Lemma 8.23. *Let $1 \leqslant s \leqslant r$ and let $n, m \in \mathbb{N}$. Then $|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m,\infty,r})|_{\Gamma_s}$ for all $v \in F_n(\Lambda_{m,\infty,r})$. Further, $|\operatorname{pref}_n(\mathbf{X}_r)|_{\Gamma_s} = \max_{v\in F_n(\mathbf{X}_r)} |v|_{\Gamma_s}$ for all $n \in \mathbb{N}$.*

Proof. We prove these claims, for any fixed $s \geqslant 1$, by induction on r. Both of these are trivial for the base case $r = s$. Suppose the claims are true for some $r \geqslant s$, and consider the case of $r+1$. Let n be fixed. We start by proving the following:

Claim 8.24. *If $v \in F_n(\Lambda_{m,\infty,r+1})$ is of the form*

$$v = e\Lambda_{m+1,\ell,r}f \tag{8.5}$$

for some $\ell, m \in \mathbb{N}$ with $\ell \geqslant m \geqslant 0$, $e \in \operatorname{suff}(\gamma^m(a_r))$, and $f \in \operatorname{pref}(\gamma^{\ell+1}(a_r))$, then $|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s}$.

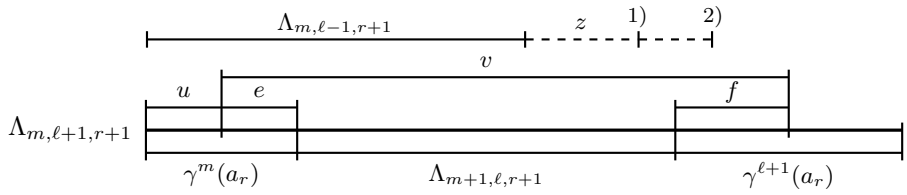

Figure 8.1: The words v and $\Lambda_{m,\ell-1,r+1}z$ in the proof of Claim 8.24. Here z ends at point 1) if $|z| < |\gamma^\ell(a_r)|$, otherwise z ends at point 2). If $\ell = m$, then $\Lambda_{m+1,\ell,r+1} = \Lambda_{m,\ell-1,r+1} = \varepsilon$, $v = ef$, and z is a prefix of uef. [109, Figure 1].

Proof. Let $v \in F_n(\Lambda_{m,\infty,r+1})$ be as in (8.5). Let $z \in \mathrm{pref}(\gamma^\ell(a_r)f)$ so that $|\Lambda_{m,\ell-1,r+1}z| = |v|$. Thus $\gamma^m(a_r) = ue$ for some $u \in \Sigma_r^*$ and $|z| = |\gamma^\ell(a_r)| + |f| - |u|$. Note that these notations are valid for the technical case $\ell = m$ also. The situation is illustrated in Figure 8.1.

Suppose first that $|z| \geqslant |\gamma^\ell(a_r)|$ whence $|u| \leqslant |f|$ and thus u is a prefix of f. In Figure 8.1, this corresponds to z ending at point 2). Let $v' = \mathrm{suff}_{|u|}(f)$. We have

$$|\Lambda_{m,\ell-1,r+1}z|_{\Gamma_s} - |v|_{\Gamma_s} = |u|_{\Gamma_s} - |v'|_{\Gamma_s} \geqslant 0,$$

by applying the induction hypothesis to $u \in \mathrm{pref}(\mathbf{X}_r)$ and $v' \in F(\mathbf{X}_r)$.

Suppose then that $|z| < |\gamma^\ell(a_r)|$ whence $|f| < |u|$ and f is a proper prefix of u. In Figure 8.1, this corresponds to z ending at point 1). If $e = \varepsilon$ and $\ell = m$, then $z = v = f \in \mathrm{pref}(\gamma^m(a_r))$ and there is nothing to prove. Assume then that either $e \neq \varepsilon$ or $\ell > m$. Let $v' = \mathrm{suff}_{|u|-|f|}(\gamma^\ell(a_r))$. If $f = \varepsilon$, then we have $|\Lambda_{m,\ell-1,r+1}z|_{\Gamma_s} - |v|_{\Gamma_s} = |u|_{\Gamma_s} - |v'|_{\Gamma_s} \geqslant 0$ by applying the induction hypothesis to $v' \in F(\mathbf{X}_r)$ and $u \in \mathrm{pref}(\mathbf{X}_r)$.

We are left with the case of f being a non-empty proper prefix of u. Write $u = fu'$ for some $u' \in \Sigma_r^+$, whence $|\Lambda_{m,\ell-1,r+1}z|_{\Gamma_s} - |v|_{\Gamma_s} = |u'|_{\Gamma_s} - |v'|_{\Gamma_s}$. Hence, to conclude the proof, it suffices to show that $|u'|_{\Gamma_s} \geqslant |v'|_{\Gamma_s}$. There exist $m_1 \in \mathbb{N}$, $0 \leqslant m_1 < m$, and words $g_1, g_2 \in \Sigma_r^*$ such that

$$f = \gamma^{m_1}(a_r)g_1 \quad \text{and} \quad \gamma^{m_1+1}(a_r) = \gamma^{m_1}(a_r)\gamma^{m_1}(a_{r-1}) = fg_2,$$

that is, $g_1g_2 = \gamma^{m_1}(a_{r-1})$. Now, by item 2 of Lemma 8.20, $\gamma^\ell(a_r) = \gamma^{m_1}(a_r)\Lambda_{m_1,\ell-1,r}$. We may thus write $g_1u' = \mathrm{pref}_{|g_1u'|}(\Lambda_{m_1,\ell-1,r}) \in F(\mathbf{X}_r)$. Observe now that $\gamma^{\ell+1}(a_r) = \gamma^\ell(a_r)\gamma^\ell(a_{r-1})$. Since $v' \in \mathrm{suff}(\gamma^\ell(a_r))$ and $g_1 \in \mathrm{pref}(\gamma^{m_1}(a_{r-1})) \subseteq \mathrm{pref}(\gamma^\ell(a_{r-1}))$, it follows that we may write $v'g_1 = e'\Lambda_{m_2,\ell-1,r}g_1 \in F(\Lambda_{m_1,\infty,r})$, where m_2 is minimal and $e' \in \mathrm{suff}(\gamma^{m_2-1}(a_r))$. Note that $m_2 > m_1$ since $e \neq \varepsilon$ or $\ell > m$. We apply the induction hypothesis on $v'g_1$ and g_1u to obtain $|v'g_1|_{\Gamma_s} \leqslant |g_1u'|_{\Gamma_s}$, from which it follows that $|u'|_{\Gamma_s} \geqslant |v'|_{\Gamma_s}$. This concludes the proof of Claim 8.24. □

From Claim 8.24 it follows that $|\mathrm{pref}_n(\Lambda_{m',\infty,r+1})|_{\Gamma_s} \leqslant |\mathrm{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s}$ for all $m' > m$. Indeed, since $\mathrm{pref}_n(\Lambda_{m',\infty,r+1})$ has a factorization of the form (8.5) (with m' in the role of $m+1$ and $e = \varepsilon$), we obtain

$$|\mathrm{pref}_n(\Lambda_{m',\infty,r+1})|_{\Gamma_s} \leqslant |\mathrm{pref}_n(\Lambda_{m'-1,\infty,r+1})|_{\Gamma_s} \leqslant \ldots \leqslant |\mathrm{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s}.$$

Assume now that $v \in F_n(\Lambda_{m,\infty,r+1})$ has a factorization of the form $v = e\Lambda_{m'+1,\ell',r+1}f$ for some $\ell' \geqslant m' \geqslant m$, $e \in \operatorname{suff}(\gamma^{m'}(a_r))$, and $f \in \operatorname{pref}_n(\gamma^{\ell'+1}(a_r))$. By Claim 8.24 and the previous observation, we have

$$|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m',\ell',r+1}f)|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s}.$$

If, on the other hand, $v \in F_n(\Lambda_{m,\infty,r+1})$ has no factorization of the form (8.5), then $v \in F(\mathbf{X}_r)$. By the induction hypothesis and the above observation, we have

$$|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\mathbf{X}_r)|_{\Gamma_s} = |\operatorname{pref}_n(\Lambda_{m',\infty,r+1})|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s},$$

where m' is minimal such that $|\gamma^{m'}(a_r)| \geqslant n$. We have proved that, for all $v \in F(\Lambda_{m,\infty,r+1})$, $|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{m,\infty,r+1})|_{\Gamma_s}$, that is, the first part of Lemma 8.23. It remains to prove that $|\operatorname{pref}_n(\mathbf{X}_{r+1})|_{\Gamma_s} = \max_{v\in F_n(\mathbf{X}_{r+1})} |v|_{\Gamma_s}$. But this is trivial since for all $v \in F_n(\mathbf{X}_{r+1}) \setminus \{\operatorname{pref}_n(\mathbf{X}_{r+1})\} = F_n(\Lambda_{0,\infty,r+1})$, we have

$$|v|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\Lambda_{0,\infty,r+1})|_{\Gamma_s} = |\operatorname{pref}_n(a_{r+1}^{-1}\mathbf{X}_{r+1})|_{\Gamma_s} \leqslant |\operatorname{pref}_n(\mathbf{X}_{r+1})|_{\Gamma_s}.$$

We have thus completed the induction step, completing the proof of Lemma 8.23. □

We are finally in the position to complete the proof of Proposition 8.19.

Proof of Proposition 8.19, item 2. We now complete the proof by analyzing the abelian complexity of $\mathbf{X}_{s,r}$. Note that $\mathcal{C}^{\mathrm{ab}}_{\mathbf{X}_{s,r}}$ is monotonously increasing, since $\min_{\mathbf{X}_{s,r},b}(n) = 0$ for all $n \in \mathbb{N}$. By Lemmas 8.21 and 8.23, we have $\mathcal{C}^{\mathrm{ab}}_{\mathbf{X}_{r,s}}(p_r(k)) = |\gamma^k(a_r)|_{\Gamma_s} + 1 = \frac{1}{(r-s)!}k^{r-s} + \mathcal{O}(k^{r-s-1})$. We thus have $\mathcal{C}^{\mathrm{ab}}_{\mathbf{X}_{r,s}}(n_k) = \Theta(n_k^{1-s/r})$ for a sequence (n_k) of indices. Note also that there exists $\alpha \in \mathbb{R}$ such that $n_{k+1} \leqslant \alpha n_k$ for all large enough k. Let now $n \in \mathbb{N}$, such that $n_k < n \leqslant n_{k+1}$ for some large enough $k \in \mathbb{N}$. Now there exist $C_1, C_2 \in \mathbb{R}$ such that

$$\mathcal{C}^{\mathrm{ab}}(n) \leqslant \mathcal{C}^{\mathrm{ab}}(n_{k+1}) \leqslant C_1 n_{k+1}^{1-s/r} \leqslant C_1\alpha^{1-s/r}n^{1-s/r} \text{ and}$$
$$\mathcal{C}^{\mathrm{ab}}(n) \geqslant \mathcal{C}^{\mathrm{ab}}(n_k) \geqslant C_2 n_k^{1-s/r} \geqslant \tfrac{C_2}{\alpha^{1-s/r}}n^{1-s/r}.$$

Thus $\mathcal{C}^{\mathrm{ab}}_{\mathbf{X}_{r,s}}(n) = \Theta(n^{1-s/r})$. □

We already noted that Theorem 8.17 follows from Proposition 8.19. This concludes our considerations of abelian complexities of morphic words. We have only considered binary words. We note that the following question posed by J.-J. Pansiot in [77] is still open.

Question 8.25. What can the factor complexity of a morphic word be?

Some progress has recently been made:

Theorem 8.26 ([30])**.** *The factor complexity of a morphic word is either of the order $\Theta(n^{1+1/k})$ for some $k \geqslant 1$, or is of the order $\mathcal{O}(n \log n)$.*

We are naturally interested in the following question.

Question 8.27. What can the abelian complexity of a morphic word be?

It is straightforward to see that the abelian complexity of a morphic word is of the order $\mathcal{O}(n^2)$ ($\mathcal{O}(n)$ among binary words). A sharper answer would seem to be quite intricate, as implied by Theorem 8.2 applied to the binary case.

Chapter 9

On the k-Abelian Equivalence in Sturmian Words

The study of the k-abelian complexity of infinite words was first initiated in [57]. Relating the complexity of an infinite word to the complexity of its finite factors, this time the k-abelian equivalence, again turned out to be a fruitful notion. Recall that for an infinite word $\mathbf{w}$, the k-abelian complexity function $\mathcal{C}_{\mathbf{w}}^{(k)}(n)$ of $\mathbf{w}$ is defined as the number of k-abelian equivalence classes of the factors of length n of $\mathbf{w}$. Define, for each $k \geqslant 1$, the function $q^{(k)} : \mathbb{N} \to \mathbb{N}$ by $q^{(k)}(n) = n + 1$ for $n < 2k$, $q^{(k)}(n) = 2k$ for $n \geqslant 2k$. Now if an infinite word $\mathbf{w} \in \Sigma^{\mathbb{N}}$ has $\mathcal{C}_{\mathbf{w}}^{(k)}(n_0) < q^{(k)}(n_0)$ for some $k, n_0 \geqslant 1$ or $k = \infty$, then $\mathbf{w}$ is ultimately periodic (but the converse is not necessarily true for $k \neq \infty$, $k \neq 1$) [57]. Moreover, Sturmian words (recall Definition 2.10, we also give another definition in this section) may be characterized:

Theorem 9.1 ([57, Theorem 4.1])**.** *For any $k \geqslant 1$ (or $k = \infty$), the k-abelian complexity of any Sturmian word equals $q^{(k)}$. Conversely, if the k-abelian complexity function of an aperiodic word $\mathbf{w} \in \Sigma^{\mathbb{N}}$ equals $q^{(k)}$ for some fixed $k \geqslant 1$, then $\mathbf{w}$ is Sturmian.*

Other aspects of k-abelian equivalence in infinite words were also considered in [57]. Among these aspects, the authors investigated various questions related to k-abelian repetitions, such as connection of repetitions to the k-abelian complexity. We do not repeat the results here, but we recall an immediate corollary applied to Sturmian words, which is of interest to us. A word of the form $u_0 u_1 \cdots u_{n-1}$, where $u_0 \sim_k u_1 \sim_k \cdots u_{n-1}$, is called a k-abelian power of exponent n.

Theorem 9.2 ([57, Corollary 5.7])**.** *For any $k \geqslant 1$ and $N \geqslant 1$, a Sturmian word contains a k-abelian power of exponent N.*

In this chapter we consider several aspects of k-abelian equivalence in Sturmian words. We give alternative proofs for several results related to Sturmian words obtained in [57]. In particular, we give a new proof of the k-abelian complexity

of Sturmian words, and we prove a result slightly sharpening Theorem 9.2. We also generalize related research done in [36, Proposition 3.3], where abelian repetitions in Sturmian words were considered. In particular, we define the so-called *k-Lagrange spectrum* for each $k \geqslant 1$, which generalizes the well-known Lagrange spectrum from number theory. The basis of the considerations in this chapter is the interpretation of Sturmian words as certain dynamical systems, which we recall from [64, Chapter 2] next.

The results of this chapter appear in the work [81].

For a real number x, let us define the *fractional part* of x as $\langle x \rangle = x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.[1] Let us identify the unit interval $[0, 1)$ with the torus $\mathbb{T}$, and let α be a positive irrational number. The mapping $R\colon \mathbb{T} \to \mathbb{T}$, $x \mapsto \langle x + \alpha \rangle$ defines a rotation on $\mathbb{T}$. Partition the torus $\mathbb{T}$ into two half-open intervals I_a and I_b defined by the endpoints 0 and $1 - \alpha$. Now we have two choices: $I_a = [0, 1 - \alpha)$ or $I_a = (0, 1 - \alpha]$ (in either case $I_b = \mathbb{T} \setminus I_a$). This choice is represented by the choice of whether $0 \in I_a$ or not. For our considerations this choice does not matter, as we consider only the interior points of the intervals, therefore we just consider it fixed. Define a coding function $\nu : \mathbb{T} \to \{a, b\}$ by $\nu(x) = a$ if $x \in I_a$ and $\nu(x) = b$ if $x \in I_b$. Define now the infinite word $\mathbf{s}_{x,\alpha}$ as the word obtained by setting its nth, $n \geqslant 0$, letter to $\nu(R^n(x))$. The word $\mathbf{s}_{x,\alpha}$ is called the *Sturmian word of slope α and intercept x*.[2] Note that this word is unique after fixing the choice of whether $0 \in I_a$ or not.

Recall the previously mentioned example of a Sturmian word, namely, the Fibonacci word $\mathbf{f}$. Its slope is $1/\varphi^2 \approx 0.38$, where φ is the golden ratio, and its intercept equals its slope. We have

$$\mathbf{f} = abaababaabaababaababaabaababaaba \cdots .$$

Let $x, y \in \mathbb{T}$ with $x < y$. Then by both $I(x, y)$ and $I(y, x)$ we mean the interval $[x, y)$ if $0 \in I_a$ and the interval $(x, y]$ if $0 \notin I_a$. We let $\|x\|$ be the distance of x to the nearest integer, that is, $\|x\| = \min\{\langle x \rangle, 1 - \langle x \rangle\}$.

It is well-known that the sequence $(\langle n\alpha \rangle)_{n \geqslant 0}$ is dense in the interval $[0, 1)$. In particular, Sturmian words of slope α have the same finite factors; for a fixed α we let F_α denote the set of factors of a Sturmian word of slope α. Let $w = a_0 a_1 \cdots a_{n-1}$ be a word in F_α having length n. Then there exists a unique subinterval $[w]$ of $\mathbb{T}$ such that the Sturmian word $\mathbf{s}_{x,\alpha}$ begins with w if and only if $x \in [w]$. It is not hard to see that $[w] = I_{a_0} \cap R^{-1}(I_{a_1}) \cap \ldots \cap R^{-(n-1)}(I_{a_{n-1}})$ (here the choice of whether or not $0 \in I_a$ matters, but we only consider interior points of these intervals). The points 0, $\langle -\alpha \rangle$, $\langle -2\alpha \rangle$, …, $\langle -n\alpha \rangle$ partition the torus into $n + 1$ subintervals which are exactly the intervals $[w]$ for factors of length n. We call these $n + 1$ intervals the *level n intervals*, and we denote the set containing them by $L(n)$.

Continued fractions are extremely useful in studying Sturmian words, so let us recall continued fraction expansions of irrational numbers. For an overview on this connection, see [80, Chapter 4].

[1] Note that the fractional part of a number x is usually denoted by $\{x\}$. We shall be considering explicit sets of fractional parts of numbers, so we use the notation $\langle x \rangle$ instead to avoid misinterpretations.

[2] This is a characterization of Sturmian words. These words are exactly the words previously defined in Definition 2.10.

Every irrational real number α has a unique infinite continued fraction expansion:

$$\alpha = [a_0; a_1, a_2, a_3, \ldots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}} \tag{9.1}$$

with $a_0 \in \mathbb{Z}$ and $a_k \geqslant 1$ for all $k \geqslant 1$. The term a_k is called the *kth partial quotient* of α. The number $[a_0; a_1, a_2, \ldots, a_k]$, which we call the *kth convergent* of α, is a rational number, and we set $[a_0; a_1, a_2, \ldots, a_k] = p_k/q_k$ for each $k \geqslant 1$. The convergent p_k/q_k, for any $k \geqslant 1$, of α satisfies the property

$$\|q_k\alpha\| = \min_{0<m<q_{k+1}} \|m\alpha\|,$$

which is called the *best approximation property*.

In the rest of chapter, we fix the irrational number $\alpha \in (0, 1)$ with continued fraction expansion $[a_0; a_1, a_2, \ldots]$. When talking about the convergents p_k/q_k, the level n intervals $L(n)$, the rotation R, etc., we implicitly mean these notions defined by this fixed α.

9.1 k-Abelian Equivalence and Repetitions in Sturmian Words

In this section we consider the k-abelian complexity of Sturmian words. We give an alternative proof of Sturmian words having k-abelian complexity function equal to $q^{(k)}$ as defined in the beginning of this chapter (one direction of Theorem 9.1). We do this by showing that the k-abelian equivalence classes of factors of length n of a Sturmian word correspond to $q^{(k)}(n)$ intervals on the torus $\mathbb{T}$. Moreover, we characterize the endpoints of these intervals (Theorem 9.7). We also prove a result slightly sharpening the following result of [57] (applied to Sturmian words) (see Theorem 9.13).

Proposition 9.3 ([57, Proposition 2.8])**.** *Let u and v be two factors of length n occurring in a Sturmian word. Then $u \sim_k v$ if and only if u and v share common prefixes and common suffixes of length $\min\{n, k-1\}$ and $u \sim_1 v$.*

Remark 9.4. We remark that this result is quite interesting in the sense that a rather weak condition is enough to determine k-abelian equivalence of factors of a Sturmian word. This is not unique to Sturmian words: it holds for the so-called *episturmian words* [57, Proposition 2.8], and the so-called *Cantor word*, the fixed point $\varphi^\omega(a)$ of the morphism $\varphi : a \mapsto aba$, $b \mapsto bbb$, [22, Theorem 1]. We return to related questions later on.

Our Theorem 9.7 can be seen as a generalization of the following result, which characterizes, in a dynamical sense, abelian equivalence among factors of the same length in a Sturmian word.

Proposition 9.5 ([94, Theorem 19], see also [36, Proposition 3.3])**.** *Let u and v be two factors of length n occurring in a Sturmian word of slope α. Then $u \sim_1 v$ if and only if $[u], [v] \subseteq I(0, \langle -n\alpha \rangle)$ or $[u], [v] \subseteq I(\langle -n\alpha \rangle, 1)$.*

The above result says that there are two possible abelian equivalence classes for factors of length n. These classes correspond to $q^{(1)}(n) = 2$ intervals on the torus and, furthermore, the endpoints of these intervals are 0 and $\langle -n\alpha \rangle$. It is this interpretation we generalize for $k \geqslant 2$.

9.1.1 k-Abelian Equivalence on the Torus

In light of Proposition 3.1, for two words of equal length to be k-abelian equivalent, they must share common prefixes and common suffixes of length $k-1$. We interpret this property in Sturmian words in the dynamical sense, that is, in the torus.

Let $m \geqslant 1$, and define $\mathcal{D}_{k,m} = \{0, \langle -\alpha \rangle, \langle -2\alpha \rangle, \ldots, \langle -\min\{m, k-1\}\alpha \rangle\}$. These points divide the torus into $\min\{m+1, k\}$ intervals (which are the level $\min\{m, k-1\}$ intervals), and two points x and y belong to the same interval if and only if the prefixes of $\mathbf{s}_{x,\alpha}$ and $\mathbf{s}_{y,\alpha}$ of length $\min\{m, k-1\}$ are equal. Now if $m \geqslant k-1$, then $\#\mathcal{D}_{k,m} = k$ and

$$R^{-(m-(k-1))}(\mathcal{D}_{k,m}) = \{\langle -(m-(k-1))\alpha \rangle, \ldots, \langle -m\alpha \rangle\}.$$

These points also divide the torus into k intervals. Again, two points x and y belong to the same interval if and only if the prefixes of $\mathbf{s}_{x,\alpha}$ and $\mathbf{s}_{y,\alpha}$ of length m have a common suffix of length $k-1$. Let us now set $\mathbb{P}_{k,m} = \mathcal{D}_{k,m} \cup R^{-(m-(k-1))}(\mathcal{D}_{k,m})$ if $m \geqslant k-1$; otherwise set $\mathbb{P}_{k,m} = \mathcal{D}_{k,m}$.

Definition 9.6. Order the points x_i of $\mathbb{P}_{k,m}$: $0 = x_0 < x_1 < \ldots < x_{\ell-1} < x_\ell = 1$, where $\ell = \#\mathbb{P}_{k,m}$, and set $I_i = [x_i, x_{i+1})$ if $0 \in I_a$ and $I_i = (x_i, x_{i+1}]$ if $0 \notin I_a$ for $0 \leqslant i < \ell$. We define $\mathbb{I}_{k,m}$ as the set of the intervals I_i, $i = 0, \ldots, \ell - 1$.

Note that $\mathbb{P}_{k,m}$ is always a subset of the points defining the level m intervals $L(m)$. This implies that the interval $[u]$ corresponding to a factor of length m is always contained entirely in some interval of $\mathbb{I}_{k,m}$. Observe that, when $m \geqslant 2k$, $\mathbb{I}_{k,m}$ consists of $2k$ intervals. Furthermore, when $m < 2k$, the intervals of $\mathbb{I}_{k,m}$ coincide with the level m intervals. Indeed, if $m < k-1$ this is so by definition, and if $k - 1 \leqslant m < 2k$, we have

$$\begin{aligned}\mathbb{P}_{k,m} &= \{0, \langle -\alpha \rangle, \ldots, \langle -(k-1)\alpha \rangle\} \cup \{\langle -(m-(k-1))\alpha \rangle, \ldots, \langle -m\alpha \rangle\} \\ &= \{0, \langle -\alpha \rangle, \ldots, \langle -m\alpha \rangle\}.\end{aligned}$$

We claim that the intervals $\mathbb{I}_{k,m}$ determine the k-abelian equivalence classes.

Theorem 9.7. *Let $k \geqslant 1$ and let u and v be two factors of length m occurring in a Sturmian word of slope α. Then $u \sim_k v$ if and only if there exists $J \in \mathbb{I}_{k,m}$ such that $[u], [v] \subseteq J$.*

Let us first analyze the statement. In the case of $k = 1$, this is exactly Proposition 9.5. Thus, for the rest of these considerations, we may assume $k \geqslant 2$. Now, if $m \leqslant 2k-1$ in the above theorem, then $u \sim_k v$ is equivalent to $u = v$. Further, by the observation above, the intervals of $\mathbb{I}_{k,m}$ coincide with the level m intervals. Thus the claim reduces to $u = v$ if and only if $[u] = [v]$, which is, of course, trivial. We proceed to prove Theorem 9.7 in parts. First we show the only if direction.

Lemma 9.8. *Let u and v be k-abelian equivalent factors of length m of a Sturmian word of slope α. Then there exists an interval $J \in \mathbb{I}_{k,m}$ such that $[u], [v] \subseteq J$.*

Proof. Assume for a contradiction that $[u]$ and $[v]$ are contained in distinct intervals of $\mathbb{I}_{k,m}$. We thus deduce that $u \neq v$ whence $m \geqslant k-1$. By Proposition 3.1, u and v share a common prefix and a common suffix of length $k-1$. For all $x \in [u]$ we have either $x > y$ for all $y \in [v]$ or $x < y$ for all $y \in [v]$. Thus, without loss of generality, we assume that $\sup[u] \leqslant \inf[v]$. Let K be the interval $\{z\colon \sup[u] \leqslant z \leqslant \inf[v]\}$. (If $\sup[u] = \inf[v]$, then K is the singleton set containing the common endpoint of the intervals $[u]$ and $[v]$.) Since $[u]$ and $[v]$ are contained in distinct intervals of $\mathbb{I}_{k,m}$, there exists a point x in $\mathbb{P}_{k,m}$ such that $x \in K$. Let $\mathcal{S}$ denote the set $R^{-(m-(k-1))}(\mathcal{D}_{k,m})$. Now the point x cannot be in $\mathcal{D}_{k,m}$, as otherwise $[u] \subseteq I(0,x)$ and $[v] \subseteq I(x,1)$ implying that u and v have distinct prefixes of length $k-1$, contradicting our assumption. Thus we must have $x \in \mathcal{S}$. Let y be an arbitrary point in $\mathcal{S}$. If $y \in \mathbb{T} \setminus ([u] \cup [v] \cup K)$, then either $[u] \subseteq I(x,y)$ and $[v] \cap I(x,y) = \emptyset$ or symmetrically $[v] \subseteq I(x,y)$ and $[u] \cap I(x,y) = \emptyset$. Then, by the definition of the points $\mathcal{S}$, we see that u and v have distinct suffixes of length $k-1$, which is impossible. We conclude that $\mathcal{S} \subseteq K$ (see Example 9.12 for this situation). Since $\langle -m\alpha\rangle \in \mathcal{S}$, it follows by Proposition 9.5 that u and v are not abelian equivalent, and thus cannot be k-abelian equivalent. This is a contradiction. □

For the if direction of Theorem 9.7, we offer two proofs. The first, and clearly the shorter, makes use of Proposition 9.3. We give a second proof, which only uses the dynamical aspects of Sturmian words (and combinatorial properties of k-abelian equivalence), to let us claim an alternative proof of Proposition 9.3.

Lemma 9.9. *Let u and v be two factors of length m of a Sturmian word of slope α. Assume there exists an interval $J \in \mathbb{I}_{k,m}$ such that $[u], [v] \subseteq J$. Then $u \sim_k v$.*

Proof 1. By the definition of the intervals $\mathbb{I}_{k,m}$, the words u and v share a common prefix and a common suffix of length $k-1$. Moreover they are abelian equivalent by Proposition 9.5 because the point $\langle -m\alpha\rangle$ separating the two abelian equivalence classes is among the points $\mathbb{P}_{k,m}$. Now Proposition 9.3 implies that $u \sim_k v$. □

Before giving an alternative proof of the above lemma, we make an observation.

Lemma 9.10. *Let u be a factor of length $2k$ of a Sturmian word $\mathbf{x}$ of slope α. Assume that $[u]$ has $\langle -k\alpha\rangle$ as an endpoint. Then $|u|_w = 1$ for each factor of length k of $\mathbf{x}$.*

Proof. We show that each factor of length k occurs at least once in u. Since $|u| = 2k$, it follows that each factor occurs exactly once. Let K denote the point $\langle -k\alpha\rangle$, whence $[u] = I(K,x)$ for some $x = \langle -r\alpha\rangle$, $r \leqslant 2k$. Assume that $K < x$, the other case being analogous.

Let v be a factor of length k, whence $[v] = I(y,z)$, where $y = \langle -i\alpha\rangle$ and $z = \langle -j\alpha\rangle$ for some $i, j \leqslant k$ and $y < z$. Observe now that $R^{k-i}([u]) = I(y, \langle (k-i-r)\alpha\rangle)$. Now if $R^{k-i}([u]) \not\subseteq [v]$ then $\langle -j\alpha\rangle$ occurs in the interval $R^{k-i}([u])$. But then $[u]$ would contain the point $R^{-(k-i)}(\langle -j\alpha\rangle) = \langle -(k-j+i)\alpha\rangle$, where $0 \leqslant k-j+i \leqslant 2k$, since $-k \leqslant i-j \leqslant k$. Moreover $[u]$ would not be a level $2k$ interval, which is absurd. Thus $R^{k-i}([u]) \subseteq [v]$, which implies that v occurs as a factor of u, since $0 \leqslant i \leqslant k$. □

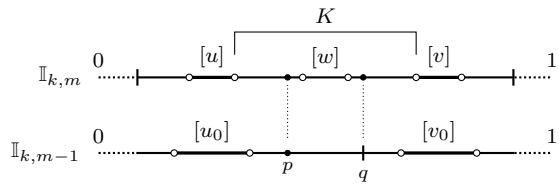

Figure 9.1: Illustration of the situation in Proof II of Lemma 9.9. The unit line $[0,1)$ is partitioned by the intervals of $\mathbb{I}_{k,m}$, in one of which both $[u]$ and $[v]$ are assumed to rest in. The endpoints of this interval in $\mathbb{P}_{k,m}$ are denoted by bars. It is assumed that $[u_0]$ and $[v_0]$ rest in distinct intervals of $\mathbb{I}_{k,m-1}$. The only possible point of $\mathbb{P}_{k,m-1}$ separating $[u_0]$ and $[v_0]$ is $q = -(m-k)\alpha$. Here $[u] \subseteq [u_0]$ and $[v] \subseteq [v_0]$. Similarly, $[u_1]$ and $[v_1]$ are assumed to lie in distinct intervals of $\mathbb{I}_{k,m-1}$, and the only point of $\mathbb{P}_{k,m-1}$ that separates these points is $R(p)$, where $p = \langle -k\alpha \rangle$. In this figure $p < q$.

Proof II of Lemma 9.9. The claim is true for all $m \leqslant 2k-1$ by previous discussion. We proceed to prove the claim for $m \geqslant 2k$ by induction. Assume the claim is true for $m-1$. It follows that the prefixes (resp., suffixes) of length $k-1 \geqslant 1$ of u and v coincide. Let $u = u_0 c = d u_1$ and $v = v_0 c = d v_1$ for some letters $c, d \in \{a, b\}$. Now if $[u_0]$ and $[v_0]$ are contained in a common interval of $\mathbb{I}_{k,m-1}$, then $u_0 \sim_k v_0$ by induction, and thus $u = u_0 c \sim_k v_0 c = v$. Similarly, if $[u_1]$ and $[v_1]$ are contained in a common interval of $\mathbb{I}_{k,m-1}$, then $u_1 \sim_k v_1$ and thus $u \sim_k v$. We are left with the case that both $[u_0]$ and $[v_0]$, and $[u_1]$ and $[v_1]$ lie in distinct intervals of $\mathbb{I}_{k,m-1}$.

We fix some technical notation. Without loss of generality we may assume that $\sup[u] \leqslant \inf[v]$. Let us set K to be the set $\{z\colon \sup[u] \leqslant z \leqslant \inf[v]\}$. See Figure 9.1 for an illustration of the intervals $[u]$, $[v]$, and K. The figure is also helpful for following the subsequent arguments. Observe that $[u] \subseteq [u_0]$ and $[v] \subseteq [v_0]$, and that $R([u]) \subseteq [u_1]$ and $R([v]) \subseteq [v_1]$. Since $[u_0]$ and $[v_0]$ lie in distinct intervals of $\mathbb{I}_{k,m-1}$, there are some points of $\mathbb{P}_{k,m-1}$ occurring in K. Of these points, only $\langle -(m-k)\alpha \rangle = q$ can lie in K, since otherwise $[u]$ and $[v]$ would reside in distinct intervals of $\mathbb{I}_{k,m}$. Similarly, since $[u_1]$ and $[v_1]$ rest in distinct intervals of $\mathbb{I}_{k,m-1}$, we deduce that $R(K)$ contains some points of $\mathbb{P}_{k,m-1}$, and thus K contains some points of $R^{-1}(\mathbb{P}_{k,m-1})$. We conclude that the only possible point is $\langle -k\alpha \rangle = p$, as otherwise $[u]$ and $[v]$ would reside in distinct intervals of $\mathbb{I}_{k,m}$. See again Figure 9.1 for a depiction of the situation.

If $p = q$ it follows that $m = 2k$, since α is irrational. In this case the intervals $[u]$ and $[v]$ are the two level $2k$ intervals having $p = \langle -k\alpha \rangle$ as a common endpoint. By Lemma 9.10, we have $u \sim_k v$.

Assume now that $p \neq q$. Consider the case $p < q$, the other case being symmetric. Let w be a factor of length m with $[w] \subseteq I(p,q)$. Since w lies in the same interval of $\mathbb{I}_{k,m}$ as both $[u]$ and $[v]$, we may write $w = w_0 c = d w_1$. Now $[w_0]$ and $[u_0]$ rest in a common interval of $\mathbb{I}_{k,m-1}$ (both intervals are to the left of q). We showed that in this case $u \sim_k w$. Similarly, $[w_1]$ and $[v_1]$ lie an a common interval of $\mathbb{I}_{k,m-1}$. We showed that in this case $v \sim_k w$. Thus, by the transitivity of $\sim_k$, we have $u \sim_k v$, as claimed. □

Notice that $\mathbb{I}_{k,m}$ contains $q^{(k)}(m)$ intervals for all m. We thus have proved the

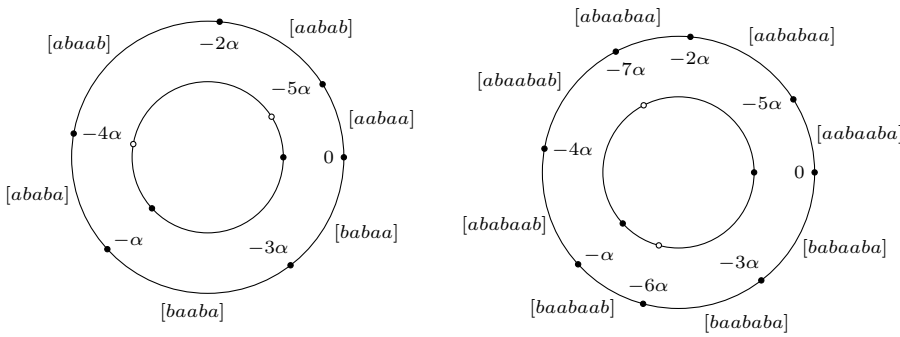

Figure 9.2: Factors of length 5 and 7 of the Fibonacci word on the unit circle. The outer circles illustrate the level 5 and 7 intervals and the inner circles the 2-abelian equivalence classes of length 5 and 7. [81, Figure 1].

following proposition:

Proposition 9.11. *For any $k \geqslant 1$, the k-abelian complexity function of an arbitrary Sturmian word equals $q^{(k)}$.*

As mentioned previously, the above proposition is one direction of Theorem 9.1.

Example 9.12. Let us consider the 2-abelian equivalence classes of length 5 of the Fibonacci word $\mathbf{f}$. Recall that the slope α of $\mathbf{f}$ is $1/\varphi^2$. On the left in Figure 9.2, there are two concentric circles. The outer circle represents the level 5 intervals separated by the points 0, $\langle -\alpha \rangle (\approx 0.62)$, $\langle -2\alpha \rangle$ (≈ 0.24), $\langle -3\alpha \rangle$ (≈ 0.85), $\langle -4\alpha \rangle$ (≈ 0.47), and $\langle -5\alpha \rangle$ (≈ 0.09). The inner circle shows the endpoints of the 2-abelian equivalence classes. The points 0 and $\langle -\alpha \rangle$ of $\mathcal{D}_{2,5}$ are shown in black while the points $\langle -4\alpha \rangle$ and $\langle -5\alpha \rangle$ of $R^{-4}(\mathcal{D}_{2,5})$ are represented by circles filled with white. The concentric circles on the right in Figure 9.2 give the corresponding intervals and points when $m = 7$.

We have four 2-abelian equivalence classes for length 5: $\{aabaa\}$, $\{aabab, abaab\}$, $\{ababa\}$, and $\{baaba, babaa\}$. The singleton classes are special. At the end of the proof of Lemma 9.8, we had to take some extra steps because factors corresponding to two distinct intervals of $\mathbb{I}_{k,m}$ could share prefixes and suffixes of length $k-1$. Indeed here $aabaa$ and $ababa$ have common prefixes and suffixes of length 1, but this does not guarantee abelian equivalence.

9.1.2 A Sharpening of Proposition 9.3

Let us now translate Theorem 9.7 from the torus into the word combinatorial setting. This allows us to state the following strengthening of Proposition 9.3.

Theorem 9.13. *Let u and v be two factors of equal length occurring in a Sturmian word of slope α. Then $u \sim_k v$ if and only if they share a common prefix and a common suffix of length $\min\{|u|, k-1\}$ and $u \sim_1 v$. Moreover, the condition $u \sim_1 v$ may be omitted if $(2k-2)\|\alpha\| > 1$.*

Proof. The claim is trivial when $k = 1$, so we may assume $k \geqslant 2$. Let $|u| = m$ and assume that u and v share common prefixes and common suffixes of length $\min\{m, k-1\}$. (Let us not yet assume $u \sim_1 v$ or $\|\alpha\| > 1/(2k-2)$.) Assume further that $[u]$ and $[v]$ lie in distinct intervals of $\mathbb{I}_{k,m}$ which is equivalent to $u \not\sim_k v$. This will lead to contradictions in both cases. We now have $m \geqslant 2k-1$. We may proceed as in the proof of Lemma 9.8, to obtain a situation where $\mathcal{S} \subseteq K$ (using the notation of Lemma 9.8). (Instead of invoking Proposition 3.1 for u and v to have common prefixes and common suffixes of length $k-1$, we have this by assumption.)

To conclude the first part of the theorem, we now further assume that $u \sim_1 v$, so we may continue in a way similar to the proof of Lemma 9.8, to reach a contradiction. Thus $[u]$ and $[v]$ rest in a common interval of $\mathbb{I}_{k,m}$, whence, by Lemma 9.9, $u \sim_k v$.

For the second part, that is, assuming that $\|\alpha\| > 1/(2k-2)$ with $k \geqslant 2$, in addition to u and v having common prefixes and suffixes of length $k-1$, we obtain a contradiction as follows.

Notice that the set K is fully contained in some level $(k-1)$ interval J (u and v have the same prefixes of length $k-1$), and the points $\mathcal{S}$ all rest in J properly (i.e., not as endpoints). Since $k \geqslant 2$ and $m \geqslant k-1$, $\mathcal{S}$ contains two points which have distance $\|\alpha\|$ (e.g., $\langle -(m-1)\alpha\rangle$ and $\langle -m\alpha\rangle$), so we deduce that the length of J is greater than $\|\alpha\|$.

We first claim that $k - 1 < \lfloor 1/\|\alpha\| \rfloor$. Indeed, if $k - 1 \geqslant \lfloor 1/\|\alpha\| \rfloor$, then the longest interval of level $k-1$ has length at most $\|\alpha\|$ (recall how the points $0, \langle -\alpha\rangle, \ldots, \langle -(k-1)\alpha\rangle$ are drawn on the torus: the distance of $\langle -i\alpha\rangle$ and $\langle -(i+1)\alpha\rangle$ is $\|\alpha\|$). Since J, a level $(k-1)$ interval, has length greater than $\|\alpha\|$, we conclude that $k - 1 < \lfloor 1/\|\alpha\| \rfloor$.

Now the points $\{0, \langle -\alpha\rangle, \ldots, \langle -(k-1)\alpha\rangle\}$ are exactly the points $\mathcal{P} = \{0, 1 - \|\alpha\|, \ldots, 1 - (k-1)\|\alpha\|\}$. In particular, J must be the interval having length $(k-1)\|\alpha\|$, as the other intervals have length $\|\alpha\|$. Now $\mathcal{S}$ consists of the points $R^{-(m-(k-1))}(\mathcal{P})$. Since R is an isometry, these points define, on the torus, $k-1$ intervals of length $\|\alpha\|$ together with one interval having length equal to that of J. Since $\mathcal{S} \subseteq J$, we conclude that J must have length at least $(k-1)\|\alpha\|$. Putting these together, we have $(k-1)\|\alpha\| < 1 - (k-1)\|\alpha\|$, that is, $\|\alpha\| < 1/(2k-2)$, which directly contradicts our assumption $\|\alpha\| > 1/(2k-2)$. □

Notice that, in the above theorem, $\|\alpha\|$ is always less than $1/2$. Thus, in order to have $(2k-2)\|\alpha\| > 1$ (to omit the requirement $u \sim_1 v$ in the k-abelian equivalence), we need $k \geqslant 3$.

Example 9.14. The slope of the Fibonacci word is approximately 0.38 which is greater than $1/4 = 1/(2\cdot 3 - 2)$, so Theorem 9.13 says that two equal length factors u and v of the Fibonacci word are k-abelian equivalent, for $k \geqslant 3$, if and only if they share common prefixes and suffixes of length $k-1$.

It is rather surprising that such a weak condition is sufficient to establish the k-abelian equivalence of two factors in an infinite word. This raises the question of whether it is possible to improve on the Fibonacci word: does there exist an infinite word such that, for all $k \geqslant 2$, two factors u and v are k-abelian equivalent if and only if $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ and $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$? We study this question, and related ones, in the final section of this chapter.

9.1.3 k-Abelian Repetitions in Sturmian Words

If u_0, u_1, …, u_{n-1} are k-abelian equivalent words of length m, then their concatenation $u_0u_1\cdots u_{n-1}$ is a *k-abelian power of exponent n and period m*. In this section, we consider k-abelian powers in Sturmian words. We are only interested in nondegenerate powers, that is, we assume that $n \geqslant 2$. Let $\mathcal{A}exp_{k,\alpha}(m)$ be the maximum exponent of k-abelian powers of period m occurring in a Sturmian word of slope α. We give a formula to compute $\mathcal{A}exp_{k,\alpha}(m)$ in terms of the intervals $\mathbb{I}_{k,m}$ and $\|m\alpha\|$. To this end, we need the following lemma. We let $\mathrm{len}(J)$ denote the length of an interval J.

Lemma 9.15. *Let u be a length m factor of a Sturmian word $\mathbf{s}$ of slope α. Let $k \geqslant 1$, $J \in \mathbb{I}_{k,m}$ such that $u \subseteq J$. Then $\mathbf{s}$ contains a k-abelian power $u_1\cdots u_n$ of period m and exponent n with $u_1 \sim_k u$ if and only if $n \leqslant \left\lfloor \frac{len(J)}{\|m\alpha\|} \right\rfloor + \delta$, where $\delta = 1$ if $len(J)$ is not an integral multiple of $\|m\alpha\|$, and otherwise $\delta = 0$.*

Proof. Take the Sturmian word $\mathbf{s}_{x,\alpha}$ of slope α and intercept x. Assume that $\mathbf{s}_{x,\alpha}$ begins with a k-abelian power of period m and exponent n. The prefix of $\mathbf{s}_{x,\alpha}$ of length m and the factor of $\mathbf{s}_{x,\alpha}$ of length m starting after this prefix are k-abelian equivalent so, by Theorem 9.7, the points x and $\langle x+m\alpha\rangle$ both lie in J. The distance between these points is $\|m\alpha\|$. Continuing this line of thought, we see that the points $x = x_0, x_1, \ldots, x_{n-1}$, where $x_i = \langle x + im\alpha\rangle$ for each $i = 0, \ldots, n-1$, all rest J. In particular, $\mathrm{len}(J) \geqslant (n-1)\|m\alpha\|$, and actually $\mathrm{len}(J) > (n-1)\|m\alpha\|$, since J is half-open. Conversely, given a point x such that the points $R^{im}(x)$, where $i = 0, \ldots, n-1$, are all included in J implies that that the word $\mathbf{s}_{x,\alpha}$ begins with a k-abelian power of period m and exponent n. □

In the following, we abuse the notation by letting $\max \mathbb{I}_{k,m}$ be the maximum length of an interval in $\mathbb{I}_{k,m}$. The following result is immediate from the above lemma.

Proposition 9.16. *We have $\mathcal{A}exp_{k,\alpha}(m) = \left\lfloor \frac{\max \mathbb{I}_{k,m}}{\|m\alpha\|} \right\rfloor + \delta$, where $\delta = 1$ if $\max \mathbb{I}_{k,m}$ is not an integral multiple of $\|m\alpha\|$ and otherwise $\delta = 0$.*

Example 9.17 (Example 9.12 continued)**.** Let intervals $\mathbb{I}_{2,5}$ have lengths $\|\alpha\|$, $\|3\alpha\|$, and $\|5\alpha\|$ in decreasing order. Observe now that $\|5\alpha\| = \|-5\alpha\| \approx 0.09$. Thus, by Proposition 9.16, the maximal 2-abelian power with period 5 in the Fibonacci word equals $\lfloor \alpha/\|5\alpha\| \rfloor + 1 = 5$. The interval of the class $\{baaba, babaa\}$ has length α which means, by Lemma 9.15, that using the words in this class, a 2-abelian power of exponent 5 and period 5 can be formed. Indeed, it is straightforward to check that $(babaa)^2(baaba)^3$ is a factor of the Fibonacci word. Using words from the class $\{ababa\}$ only 2-abelian powers of exponent $\lfloor \|3\alpha\|/\|5\alpha\| \rfloor + 1 = 2$ can be formed. The word $(aabaa)^2$ is not a factor of the Fibonacci word since the Fibonacci word does not contain the factor aaa. Indeed, we see using Lemma 9.15 that the exponent for this class is 1.

Interestingly, $\mathcal{A}exp_{2,\alpha}(7) = 1$. Indeed, it may be computed that $\|m\alpha\| = \|7\alpha\|$ is large: we have $\|7\alpha\| \approx 0.33$. This is large compared to the length of the longest interval of $\mathbb{I}_{2,7} = \|7\alpha\|$. The k-abelian equivalence relation for $k > 1$ differs in this respect from abelian equivalence: it follows from [36, Theorem 4.7] that in any Sturmian word there exists an abelian square of period m for each $m \geqslant 1$.

Our next target is to give an alternative proof of Theorem 9.2. As the value $\max \mathbb{I}_{k,m}$ is generally difficult to find, let us argue next that, when m is chosen suitably, then, in order to approximate $\mathcal{A}exp_{k,\alpha}(m)$, it is sufficient to study the level $2k-2$ intervals instead of the intervals of $\mathbb{I}_{k,m}$. Recall that the points $\mathcal{D}_{k,m} = \{0, \langle -\alpha\rangle, \langle -2\alpha\rangle, \ldots, \langle -(k-1)\alpha\rangle\}$ together with the points $\mathcal{S} = R^{-(m-(k-1))}(\mathcal{D}_{k,m}) = \{\langle -(m-(k-1))\alpha\rangle, \ldots, \langle -m\alpha\rangle\}$ determine the intervals $\mathbb{I}_{k,m}$. Suppose now that $\|m\alpha\|$ is very small (to be bounded above later). Then the points $R^m(\mathcal{S}) = R^{k-1}(D_{k,m})$ are very close to the points $\mathcal{S}$; indeed, the distance of $x_i = \langle -(m-(k-i))\alpha\rangle$ and $R^m(x_i) = \langle (k-i)\alpha\rangle$ is exactly $\|m\alpha\|$ for each $i = 0, \ldots, k-1$. Now compare the intervals $\mathbb{I}_{k,m}$ defined by the points $\mathcal{D}_{k,m} \cup \mathcal{S}$ to those intervals defined by the points $\mathcal{D}_{k,m} \cup R^{k-1}(\mathcal{D}_{k,m})$, we see that some intervals of $\mathbb{I}_{k,m}$ might be shortened or lengthened by $\|m\alpha\|$, but the order of the points is the same whenever $\|m\alpha\|$ is small enough. The points $\langle -m\alpha\rangle$ and 0 however merge, but, since the interval has length $\|m\alpha\|$ and is thus assumed to be very short, we do not care. Now

$$\mathcal{D}_{k,m} \cup R^{k-1}(\mathcal{D}_{k,m}) = \{\langle -(k-1)\alpha\rangle, \ldots, \langle -\alpha\rangle, 0, \alpha, \ldots, \langle (k-1)\alpha\rangle\}.$$

Furthermore, since we are interested in lengths of intervals and R is an isometry, we can study the set $R^{-(k-1)}(\mathcal{D}_{k,m} \cup R^{k-1}(\mathcal{D}_{k,m}))$ instead. This is the set of endpoints of the level $2k-2$ intervals. It is quite obvious from the preceding that the above considerations may be performed, that is, $\|m\alpha\|$ is small enough, whenever $\|m\alpha\|$ is less than the length of the shortest interval of level $2k-2$.

We abuse the notation by letting $\min L(n)$ (resp., $\max L(n)$) denote the length of the shortest (resp., longest) interval of level n. We have thus argued that whenever $\|m\alpha\| < \min L(2k-2)$, we have

$$|\max \mathbb{I}_{k,m} - \max L(2k-2)| \leqslant \|m\alpha\|.$$

Therefore we have proved the following lemma.

Lemma 9.18. *Let m be a positive integer and suppose that $\|m\alpha\| < \min L(2k-2)$. Then*

$$\left| \left\lfloor \frac{\max L(2k-2)}{\|m\alpha\|} \right\rfloor - \mathcal{A}exp_{k,\alpha}(m) \right| \leqslant 1.$$

Theorem 9.2 follows now by observing that $\|m\alpha\|$ can be made as small as we wish.

Let us now consider the general case for $\mathcal{A}exp_{k,\alpha}(m)$. It is possible that $\mathcal{A}exp_{k,\alpha}(m)$ is large, but we may bound this with respect to $\mathcal{A}exp_{k,\alpha}(q_t)$, where q_t refers to the denominator of the tth convergent of α.

Proposition 9.19. *For all large enough t, we have $\mathcal{A}exp_{k,\alpha}(m) \leqslant \mathcal{A}exp_{k,\alpha}(q_t) + 2$ for all $1 \leqslant m < q_{t+1}$.*

Proof. Let $t \geqslant 1$, and assume that t is so large that $\|q_t\alpha\| < \min L(2k-2)$. Suppose that m is an integer such that $1 \leqslant m < q_{t+1}$. By the best approximation property of the convergents, we have $\|m\alpha\| > \|q_t\alpha\|$. Suppose first that $\|m\alpha\| < \min L(2k-2)$. Then by Lemma 9.18, we have

$$\mathcal{A}exp_{k,\alpha}(m) \leqslant \frac{\max L(2k-2)}{\|m\alpha\|} + 1 < \frac{\max L(2k-2)}{\|q_t\alpha\|} + 1,$$

so, by the same lemma, we have $\mathcal{A}exp_{k,\alpha}(m) \leqslant \mathcal{A}exp_{k,\alpha}(q_t)+2$. Suppose next that $\|m\alpha\| \geqslant \min L(2k-2)$. Then

$$\frac{\max L(m)}{\|m\alpha\|} \leqslant \frac{\max L(m)}{\min L(2k-2)} \leqslant \frac{1}{\min L(2k-2)},$$

so $\mathcal{A}exp_{k,\alpha}(m)$ is bounded by a constant. Thus $\mathcal{A}exp_{k,\alpha}(m) < \mathcal{A}exp_{k,\alpha}(q_t)$ for all large enough t. The sequence $(\mathcal{A}exp_{k,\alpha}(q_i))_i$ reaches arbitrarily high values due to Lemma 9.18. □

Observe that it is very well possible that $\mathcal{A}exp_{k,\alpha}(q_t) > \mathcal{A}exp_{k,\alpha}(q_{t+1})$. For example, let $k = 2$ and take $\alpha = [0; 3, 1, 1, 1, 100, \overline{1}]$. The sequence of denominators of convergents of α begins with $1, 3, 4, 7, \ldots$. It is readily computed that $\mathcal{A}exp_{k,\alpha}(4) = 6 > 5 = \mathcal{A}exp_{k,\alpha}(7)$. On the other hand, if $k = 1$, then we have $\mathcal{A}exp_{k,\alpha}(m) < \mathcal{A}exp_{k,\alpha}(q_t)$ for all t and $1 \leqslant m < q_t$ as can be readily observed from [36, Lemma 4.7].

9.2 Generalizations of the Lagrange Spectrum

Recall that the critical exponent of an infinite word $\mathbf{w}$ is the supremum of the set $\{\alpha \in \mathbb{Q} : u^\alpha \in F(\mathbf{w})\}$. A similar notion could be defined for the k-abelian repetitions. For Sturmian words, however, this notion would not lead to anything interesting, because by Theorem 9.2, Sturmian words contain k-abelian powers of arbitrarily large exponents for any $k \geqslant 1$. Instead, we consider the following related notion.

Definition 9.20. We define the *k-abelian critical exponent of slope α*, denoted by $\mathcal{A}crit_k(\alpha)$, as

$$\limsup_{m\to\infty} \frac{\mathcal{A}exp_{k,\alpha}(m)}{m}.$$

The value $\mathcal{A}crit_k(\alpha)$ measures the maximal ratio between the exponent and period of a k-abelian power in a Sturmian word of slope α. The notion was introduced in the case $k = 1$ in [36]. For the case of $k = 1$ this leads to a very interesting notion in the theory of continued fractions called the *Lagrange spectrum*. We recall some terminology from the literature.

Let α be an irrational number, and define the *Lagrange constant* $\lambda(\alpha)$ of α as the infimum of real numbers λ such that for every $c > \lambda$ the inequality

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{cq^2} \tag{9.2}$$

has only finitely many rational solutions p/q. Hurwitz's Theorem [51] states that $\lambda(\alpha) \geqslant \sqrt{5}$ for any irrational α, and there exist numbers with $\lambda(\alpha) = \sqrt{5}$. The numbers with finite Lagrange constant are often called *badly approximable numbers*. The Lagrange constant of α with continued fraction expansion as in (9.1) is computed as follows:

$$\lambda(\alpha) = \limsup_{t\to\infty}([a_{t+1}; a_{t+2}, \ldots] + [0; a_t, a_{t-1}, \ldots, a_1]). \tag{9.3}$$

Two numbers with continued fraction expansions $[a_0; a_1, \ldots]$ and $[b_0; b_1, \ldots]$ are called *equivalent* if there exists integers N and M such that $a_{N+i} = b_{M+i}$ for all $i \geqslant 0$. From the formula above, it is clear that two equivalent numbers have the same Lagrange constant.

Definition 9.21. The *Lagrange spectrum* is defined as the set of the finite Lagrange constants.

For the above details and more on the Lagrange spectrum, see [26] or [3].

We are now in the position to state the connection to the critical abelian exponent of Sturmian words.

Theorem 9.22 ([36, Theorem 5.10])**.** *The set of finite values $\mathcal{A}crit_1(\alpha)$, $\alpha \in \mathbb{R}$, equals the Lagrange spectrum.*

Now the set of finite values of $\mathcal{A}crit_k(\alpha)$, $k \geqslant 2$, can be viewed as a combinatorial generalization of the Lagrange spectrum. Thus we give the following definition.

Definition 9.23. We call the set $\{\mathcal{A}crit_k(\alpha) \colon \alpha \text{ is irrational and positive}\} \cap \mathbb{R}$ of finite k-abelian critical exponents the *k-Lagrange spectrum*, denoted by $\mathcal{L}_k$.

This section is devoted to studying some properties of the k-Lagrange spectrum. We first characterize the value $\mathcal{A}crit_k(\alpha)$ in terms of $\mathcal{A}crit_1(\alpha)$. Using this characterization, we are able to consider some topological aspects of the k-Lagrange spectra in the spirit of what is known about the ordinary Lagrange spectrum. There are similarities and stark differences between the generalized spectra for $k > 1$ compare to the ordinary spectrum, as we shall shortly see.

Using the results from the previous section we may state the following.

Theorem 9.24. *We have $\mathcal{A}crit_k(\alpha) = \max L(2k-2) \cdot \mathcal{A}crit_1(\alpha)$ for all $k \geqslant 1$.*

Proof. Let t be large and let m be an integer such that $q_t \leqslant m < q_{t+1}$. It follows from Proposition 9.19 that

$$\frac{\mathcal{A}exp_{k,\alpha}(m)}{m} \leqslant \frac{\mathcal{A}exp_{k,\alpha}(q_t) + 2}{q_t}.$$

This observation leads us, by Lemma 9.18, to

$$\mathcal{A}crit_k(\alpha) = \limsup_{m\to\infty} \frac{\mathcal{A}exp_{k,\alpha}(m)}{m} = \limsup_{t\to\infty} \frac{\mathcal{A}exp_{k,\alpha}(q_t)}{q_t} = \limsup_{t\to\infty} \frac{\max L(2k-2)}{q_t \|q_t\alpha\|}.$$

When $k = 1$, we have, by Theorem 9.22, $\mathcal{A}crit_1(\alpha) = \limsup_{t\to\infty} \frac{1}{q_t\|q_t\alpha\|}$. This implies our claim:

$$\mathcal{A}crit_k(\alpha) = \max L(2k-2) \cdot \mathcal{A}crit_1(\alpha).$$

□

Example 9.25 (Example 9.12 continued)**.** Consider the Fibonacci word $\mathbf{f}$. Recall that the abelian critical exponent of $\mathbf{f}$ equals $\sqrt{5}$. It is readily verified that the longest level 2 interval has length $1/\varphi^2$. Thus the 2-abelian critical exponent of the Fibonacci word equals $\frac{1}{\varphi^2}\sqrt{5} \approx 0.85$.

Notice that $\mathcal{A}crit_1(\alpha)$ is finite if and only if α has bounded partial quotients; see (9.3). Therefore, for any $k \geqslant 1$, $\mathcal{A}crit_k(\alpha)$ is finite if and only if α has bounded partial quotients. Furthermore, it is well-known that numbers with bounded partial quotients comprise a set of measure zero (this is a consequence of the *Borel–Bernstein Theorem* [14, 7, 8]).

Recall that equivalent real numbers (that is, numbers with eventually partial quotients) have the same Lagrange constant. By Theorem 9.24, this no longer holds for the k-Lagrange constant when $k > 1$ because $\max L(2k-2)$ depends on α. It is not difficult to see that the points obtained in Theorem 9.24 from a single class of equivalent numbers form a dense set. This is what we prove next. As a corollary we obtain Theorem 9.28, which states that the k-Lagrange spectrum $\mathcal{L}_k$ is itself dense when $k > 1$. In the statement of the following lemma, by $\max L_\beta(\ell)$ we mean the maximal length of a level ℓ interval of slope β.

Lemma 9.26. *Let α be irrational. The set $\{\max L_\beta(\ell)\colon \beta$ is equivalent to $\alpha\}$ is contained in $(\frac{1}{\ell+1}, 1)$ and is dense there for all $\ell > 1$.*

Proof. First of all, clearly 1 is an upper bound of the set, as the lengths of the intervals on the torus are less than 1 for any ℓ. Let us argue why $\frac{1}{\ell+1}$ is an unachievable lower bound on the set. The value $\max L_\beta(\ell)$ is the maximum length of the intervals on the torus when partitioned with $\ell+1$ points. There are thus $\ell+1$ intervals, so the longest of the intervals must have length at least $\frac{1}{\ell+1}$. Since the intervals cannot all have length $\frac{1}{\ell+1}$ (as α is irrational) we conclude that $\max L_\beta(\ell) > \frac{1}{\ell+1}$ for all β.

Let $\gamma \in (\frac{1}{\ell+1}, 1)$, and suppose without loss of generality that it is irrational. By cutting the continued fraction expansion of $1-\gamma$, after finitely many partial quotients, we obtain a fraction that is as close to $1-\gamma$ as we desire. Thus we can find a rational β such that $\ell\beta$ is arbitrarily close to $1-\gamma$ (from either side).

Now form an irrational β' by continuing the continued fraction expansion of β in such a way that it is equivalent to α. By selecting the partial quotients appropriately, we find that $\ell\beta'$ is arbitrarily close to $1-\gamma$. Consider now the level ℓ intervals of slope β'. The longest such interval clearly has length $1-\ell\beta'$ since $\gamma > \frac{1}{\ell+1}$. As $1-\ell\beta'$ is as close to γ as we like, the claim follows. □

Now the smallest element of the Lagrange spectrum is $\sqrt{5}$. Interestingly, the Lagrange constant corresponding to the Fibonacci word defined previously, equals $\sqrt{5}$. Thus, by Theorem 9.24 and Lemma 9.26 we have the following:

Theorem 9.27. *Let $k > 1$. Then $\mathcal{L}_k \subseteq (\frac{\sqrt{5}}{2k-1}, \infty)$ and $\frac{\sqrt{5}}{2k-1}$ is the least accumulation point of $\mathcal{L}_k$. In particular, the set $\mathcal{L}_k$ is not closed.*

The above theorem should be compared with the fact that $\mathcal{L}_1$ is closed [26, Theorem 2 of Chapter 3]. Notice that it also follows that when $k > 1$, the Fibonacci word no longer has minimal critical k-abelian exponent among all Sturmian words. Indeed, there exist Sturmian words that have k-abelian critical exponent arbitrarily close to $\sqrt{5}/(2k-1)$. In particular, for $k = 2$, there are Sturmian words having 2-abelian critical exponent arbitrarily close to $\sqrt{5}/3 \approx 0.75$.

Let us now recall some remarkable facts about the Lagrange spectrum. *Hall's ray* is the largest half-line contained in $\mathcal{L}_1$. It was proven by M. Hall that the half-line $[6, \infty)$ is contained in $\mathcal{L}_1$ [43]. Several improvements of the left endpoint

were made by several researchers, and finally, in [38], G. Freiman determined that Hall's ray equals $[c_F, \infty)$, where c_F is the *Freiman constant*

$$c_F = \frac{2221564096 + 283748\sqrt{462}}{491993569} = 4.5278295661\ldots$$

The detailed history and references can be found in [26, Chapter 4].

We now show that $\mathcal{L}_k$ forms a dense set when $k > 1$.

Theorem 9.28. *The k-Lagrange spectrum $\mathcal{L}_k$ is dense in $(\frac{\sqrt{5}}{2k-1}, \infty)$ when $k > 1$.*

Proof. By Lemma 9.26, the intervals $(\frac{\sqrt{5}}{2k-1}, \sqrt{5})$ and $(\frac{c_F}{2k-1}, c_F)$ are dense with points of $\mathcal{L}_k$. Now c_F is less than 6, so $\frac{c_F}{2k-1} < 2 < \sqrt{5}$ meaning that these dense sets overlap. Thus the interval $(\frac{\sqrt{5}}{2k-1}, c_F]$ is dense with points of $\mathcal{L}_k$. For each point θ in Hall's ray $[c_F, \infty)$ there exists an irrational α such that $\mathcal{A}crit_1(\alpha) = \theta$. Now θ is an accumulation point of the set $\theta \cdot \{\max L_\beta(2k-2) \colon \beta \text{ is equivalent to } \alpha\} \subseteq \mathcal{L}_k$, so that θ is in the set of accumulation points of $\mathcal{L}_k$. The claim follows. □

The above should be compared to the fact that the (1-)Lagrange spectrum is not dense between $\sqrt{5}$ and c_F. In fact, a substantial amount of research has been performed on searching for maximal gaps occurring in this interval, see, e.g., [26, Chapter 5]. It is known for example that the set $[\sqrt{5}, 3] \cap \mathcal{L}_1$ is discrete and that the interior of the interval $[\sqrt{12}, \sqrt{13}]$ does not include any points of $\mathcal{L}_1$ while its endpoints are in $\mathcal{L}_1$. It is unknown if $\mathcal{L}_1$ contains an interval below c_F.

We do not know whether a corresponding half-line exists in $\mathcal{L}_k$ when $k > 1$. Indeed, we only have that the spectrum is dense in the half-line $(\frac{\sqrt{5}}{2k-1}, \infty)$. As an example, take for each real number $\theta > 0$ the set $\mathbb{Q}_{k,\theta} = (\frac{\theta}{2k-1}, \theta) \cap \mathbb{Q}$. Clearly the union of the sets $\mathbb{Q}_{k,\theta}$, $\theta > 0$, does not contain any intervals. It is not clear whether a similar phenomenon happens with the k-Lagrange spectra for $k > 1$. As another example, it is possible for uncountably many numbers to have the same Lagrange constant (one such number is 3; it is the Lagrange constant of uncountably many numbers [96, Theorem 3, Chapter IV§6]). From such numbers it could be possible to construct an interval. Indeed, we have that the set $(\frac{3}{2k-1}, 3) \cap \mathcal{L}_k$ contains a disjoint union of uncountably many dense disjoint sets. We do not suggest that this gives an interval in $\mathcal{L}_k$, but it cannot be ruled out without further inspection.

Let us state the underlying question above explicitly:

Question 9.29. Does the k-Lagrange spectrum $\mathcal{L}_k$ contain an interval? Does it contain a half-line?

Let us also point out that it is easy to come up with numbers greater than $\sqrt{5}/(2k-1)$ that are not in $\mathcal{L}_k$. The two smallest elements of $\mathcal{L}_1$ are $\sqrt{5}$ and $\sqrt{8}$, so any point in $\mathcal{L}_k$ between $\sqrt{5}/(2k-1)$ and $\sqrt{8}/(2k-1)$ is of the form $\max L_\alpha(2k-2) \cdot \sqrt{5}$ for some α equivalent to the golden ratio. The number $\max L_\alpha(2k-2)$ is always irrational, so rational multiples of $\sqrt{5}$ between $\sqrt{5}/(2k-1)$ and $\sqrt{8}/(2k-1)$ are not in $\mathcal{L}_k$. In particular, the k-Lagrange spectrum is not a half-line for any $k \geqslant 1$.

We conclude these considerations by stating the following questions.

Question 9.30. Is there an arithmetical characterization or interpretation for the k-Lagrange spectra, $k \geqslant 2$?

Considering the k-abelian critical exponents of infinite words in general, the following question arises naturally.

Question 9.31. Let $\alpha \in \mathbb{R}$ be non-negative. Is α the critical k-abelian exponent of some infinite word?

We answer this question in the positive in [82], which was recently accepted for publication in the proceedings of the 12th International Conference on Words 2019, Loughborough, UK.

9.3 Words with Rigid Structure on k-Abelian Equivalence

To conclude this chapter, and this thesis, we briefly discuss the topic of k-abelian equivalence among factors of certain infinite words. Our focus is on infinite words where the k-abelian equivalence of two factors is determined with quite weak requirements. We firstly set some terminology.

Definition 9.32. An infinite word $\mathbf{w}$ is said to have *Property* $\mathcal{K}$ if, for all $k \geqslant 1$, two equal length factors u and v of $\mathbf{w}$ are k-abelian equivalent if and only if $u \sim_1 v$ and u and v share common prefixes and common suffixes of length $\min\{|u|, k-1\}$.

An infinite word $\mathbf{w}$ is said to have *Property* $\mathcal{K}'_{k_0}$ if, for any $k \geqslant k_0$, two equal length factors u and v of $\mathbf{w}$ are k-abelian equivalent if and only if u and v share common prefixes and common suffixes of length $\min\{|u|, k-1\}$.

As was observed already in Remark 9.4, several infinite words have Property $\mathcal{K}$, including all Sturmian words, all episturmian words, and the Cantor word. The authors of [22] asked what sort of words have the Property $\mathcal{K}$. In fact, the proof of [57, Proposition 2.8] can be used to prove that all infinite words having at most one right special factor of each length have Property $\mathcal{K}$. Sturmian words and episturmian words are in this family, but there also exist others. Observe that the Cantor word is not in this family, as it contains the right special factors ab and bb. We give the proof here for the sake of completeness.

Proposition 9.33. *Let $\mathbf{x} \in \Sigma^{\mathbb{N}}$ be an infinite word for which there exists at most one right special factor of length n for each $n \in \mathbb{N}$. Then $\mathbf{x}$ has Property $\mathcal{K}$.*

Proof. The claim is of course trivial for $k = 1$, so assume $k \geqslant 2$. Assume u and v share common prefixes and common suffixes of length $k-1$ and that $u \sim_1 v$. By induction we have $u \sim_{k-1} v$. Let axb be a factor of $\mathbf{x}$ of length k for some letters a and b and a word x. We aim to show that $|u|_{axb} = |v|_{axb}$.

If ax is not right special, then ax is always followed by b in $\mathbf{x}$. We deduce that $|u|_{axb} = |u|_{ax} - \delta$, where $\delta = 1$ if ax is a suffix of u and otherwise $\delta = 0$. Since u and v share common suffixes of length $k-1$ and $u \sim_{k-1} v$, we deduce that $|u|_{ax} - \delta = |v|_{ax} - \delta = |v|_{axb}$.

Assume then that ax is right special. For each letter c we define $n_c = |u|_{cxb}$ and $n'_c = |v|_{cxb}$. Observe that for each letter $c \neq a$, cx is not right special, as ax is the unique right special factor of length $k-1$. We already showed that, in this case, $n_c = n'_c$. Now $\sum_{c\in\Sigma} n_c = |u|_{xb} - \delta = |v|_{xb} - \delta = \sum_{c\in\Sigma} n'_c$, where $\delta = 1$ if xb is a prefix of u (and hence a prefix of v) and otherwise $\delta = 0$. Since $n_c = n'_c$ for all $c \neq a$, it follows that $n_a = n'_a$, which concludes the proof. □

Now Theorem 9.13 implies that each Sturmian word $\mathbf{x}$ of slope α has Property $\mathcal{K}'_{k_0}$ for $k_0 = \left\lceil \frac{1}{2\|\alpha\|} \right\rceil + 1$. Thus there exist Sturmian words having the Property $\mathcal{K}'_3$ (e.g., the Fibonacci word). Clearly only words a^ω, where a is a letter, have the property $\mathcal{K}'_1$. At the end of Subsection 9.1.2 it was asked whether there exist words having Property $\mathcal{K}'_2$. The next proposition tells that such binary words exist, but that they are rather uninteresting. (In particular, no Sturmian word has this property).

Proposition 9.34. *Let* $\mathbf{w}$ *be an infinite binary word such that, for each of its factors* u *and* v *of equal length, we have* $u \sim_1 v$ *if they begin with and end with common letters. Then* $\mathbf{w}$ *is ultimately periodic.*

Proof. Assume for a contradiction, that $\mathbf{x} \in \mathbb{B}^{\mathbb{N}}$ is aperiodic. Thus $\mathbf{x}$ contains either aa or bb. By symmetry, we may assume that aa occurs, and, furthermore, it occurs prior to a possible occurrence of bb. Now if bb does occur, $\mathbf{x}$ contains the factors aab and abb. This is impossible, since they begin and end with common letters, but are not abelian equivalent (we call such factors incompatible). Thus each occurrence of b is always preceded and followed by a in $\mathbf{x}$. Observe that aba thus occurs in $\mathbf{x}$ implying that aaa cannot occur in $\mathbf{x}$, as these two factors are incompatible. We conclude that $\mathbf{x}$ contains an occurrence of aa which is followed by an infinite product of the words ba and baa.

If $baa(ba)^n baa$ and $baa(ba)^n baba$ both occur in $\mathbf{x}$, then $\mathbf{x}$ contains the incompatible pair $aa(ba)^n baa$ and $a(ba)^n baba$. Therefore $baa(ba)^n baa$ can occur only for at most one value n. We conclude that $\mathbf{w}$ must have either of the words $(ba)^\omega$ or $(baa(ba)^n)^\omega$ as a suffix. This is a contradiction. □

However, if we allow more than two letters, then aperiodicity is possible as is shown by the next proposition. Before we prove this, we articulate two properties of Sturmian words that have already been alluded to previously. Firstly, in any Sturmian word there exists a unique right special factor of each length. Also, Sturmian words are *balanced*. In other words, for any pair of equal length factors u and v of occurring in a Sturmian word, we have $||u|_a - |v|_a| \leqslant 1$. These properties can be found from [64, Chapter 2].

Let σ be the morphism defined by $\sigma(a) = 02$, $\sigma(b) = 1$. It is easy to see that the word $\sigma(\mathbf{s})$ is aperiodic for any Sturmian word $\mathbf{s}$. We show the following:

Proposition 9.35. *Let* $\mathbf{s}$ *be a Sturmian word containing* aa. *Then* $\sigma(\mathbf{s})$ *has property* $\mathcal{K}'_2$.

Let us first show an intermediate result.

Lemma 9.36. *Let* $\mathbf{s}$ *be a Sturmian word containing* aa. *Then* $\sigma(\mathbf{s})$ *has property* $\mathcal{K}'_2$.

Proof. We show that $\sigma(\mathbf{s})$ has exactly one right special factor of each length. It then follows by Proposition 9.33 that $\sigma(\mathbf{s})$ has the property $\mathcal{K}$.

Let w and w' be two right special factors of equal length occurring in $\sigma(\mathbf{s})$. It is clear that both w and w' must end with 2, since 0 is always followed by 2, and 1 is always followed by 0. By the form of the morphism σ, there exist words $c, d \in \{\varepsilon, 0\}$ and unique factors x and y of $\mathbf{s}$ such that $cw = \sigma(x)$ and $dw' = \sigma(y)$. Without loss of generality we may assume that $|\sigma(x)| \geqslant |\sigma(y)|$. Since w and w'

are right special, so are x and y. Since $\mathbf{s}$ has a unique right special factor of each length we deduce that y is a suffix of x. It follows that both w and w' are suffixes of $\sigma(x)$, and thus $w = w'$, since $|w| = |w'|$. We have thus shown that $\sigma(\mathbf{s})$ has property $\mathcal{K}$. □

Proof of Proposition 9.35. We first show that if u and v are factors of equal length of $\sigma(\mathbf{s})$ beginning and ending with common letters, then $u \sim_1 v$.

By the form of the morphism σ and the fact that b is always preceded and followed by a in $\mathbf{s}$, there exist words $c \in \{\varepsilon, 0\}$, $d \in \{\varepsilon, 2\}$, and factors x and y of $\mathbf{s}$ such that $aub = \sigma(x)$ and $avb = \sigma(y)$. We show that $x \sim_1 y$ from which it follows that $u \sim_1 v$, as claimed.

Observe that $|\sigma(x)| = |\sigma(y)|$. Now the words x and y end in a common letter $c \in \{0, 1\}$ by the above. Now $x \sim_1 y$ if and only if $xc^{-1} \sim_1 yc^{-1}$ so, by replacing x with xc^{-1} and y with yc^{-1} if necessary, we may assume that x and y end with the letter a (recall that b is always preceded by a in $\mathbf{s}$). For each binary word w, we have $|\sigma(w)| = |w| + |w|_a$. Since $|\sigma(x)| = |\sigma(y)|$, we obtain

$$|x| + |x|_a = |y| + |y|_a. \tag{9.4}$$

Suppose, without loss of generality, that $|x| \geqslant |y|$, and write $x = zt$ with $|z| = |y|$. By plugging this into (9.4), we obtain $|t| + |t|_a = |y|_a - |z|_a$. Since $\mathbf{s}$ is balanced, we see that $|t| + |t|_a \leqslant 1$. Thus $t = \varepsilon$ or $t = b$. The latter case is impossible as x ends with a, so $t = \varepsilon$. Thus $|x| = |y|$ and so $|x|_a = |y|_a$ by (9.4). This means that $x \sim_1 y$.

Let us finally show that $\sigma(\mathbf{s})$ has Property $\mathcal{K}'_2$. Let $k \geqslant 2$ and assume that two equal length factors u and v of $\sigma(\mathbf{s})$ share common prefixes and common suffixes of length $\min\{|u|, k-1\}$. We may assume that $|u| \geqslant k-1$. Since $k \geqslant 2$, u and v begin and end with common letters so, by the above, $u \sim_1 v$. But since $\sigma(\mathbf{s})$ has Property $\mathcal{K}$, it follows that $u \sim_k v$. This concludes the proof. □

Finally, we state some obvious open problems in this area are as follows.

Problem 9.37. *Characterize those words having Property $\mathcal{K}$.*

This was in essence already asked in [22]. All words having at most one right special factor for each length were shown to have this property, but other words exists (e.g., the Cantor word). For such words the analysis of the asymptotic k-abelian complexities reduces to studying the asymptotic abelian complexity. Indeed, for such words w, we have $\mathcal{C}^{\mathrm{ab}}_{\mathbf{w}}(n) \leqslant \mathcal{C}^{(k)}_{\mathbf{w}}(n) \leqslant \mathcal{C}_{\mathbf{w}}(k-1)^2 \cdot \mathcal{C}^{\mathrm{ab}}_{\mathbf{w}}(n)$ for all $n \geqslant 2k-2$. For example, by Theorem 8.4, item 3), the Cantor word $\mathbf{C}$ has $\mathcal{C}^{\mathrm{ab}}_{\mathbf{C}}(n) = \Theta(n^{\log_3 2})$, and thus $\mathcal{C}^{(k)}_{\mathbf{C}}(n) = \Theta(n^{\log_3 2})$ for all $k \geqslant 2$.

We may also state the related following problems.

Problem 9.38. *Characterize those words* $\mathbf{w}$ *for which there exists* $k_0 \in \mathbb{N}$ *such that* $\mathbf{w}$ *has Property* $\mathcal{K}'_{k_0}$.

Sturmian words were observed to have this property, and there exist other such words. An interesting particular open problem would be the following.

Problem 9.39. *Characterize those words having property $\mathcal{K}'_2$.*

Proposition 9.35 gives a large family of these words. Such words would necessarily be quite rigid in the k-abelian sense. Clearly, for $k \geqslant 2$, each of the k-abelian complexities of such a word $\mathbf{w}$ would be bounded: $\mathcal{C}_{\mathbf{w}}^{(k)}(n) \leqslant \mathcal{C}_{\mathbf{w}}(k-1)^2$ for all $n \geqslant 2k-2$.

Appendix A

Algorithms

Algorithm 1: Algorithm to check if, for a given minimal representative v and a letter a, va admits a lexicographically smaller k-switching involving the suffix of va of length $k-1$.

```
   input  : A word v ∈ L_{k,Σ,◁}, letter a ∈ Σ.
   output: Returns true if and only if va ∈ L_{k,Σ,◁}.
 1 if |va| < 2k then
 2 |   return true;
 3 end
 4 y ← suff_{k−1}(va),
 5 i ← position of last occurrence of y in va;
 6 while y occurs in va before i do
 7 |   i ← last occurrence of y in va before i;
 8 |   for x ∈ Σ^{k−1} do
 9 |   |   for b ∈ Σ do
10 |   |   |   for c ∈ Σ, b ◁ c do
11 |   |   |   |   if xc occurs before i in va & xb occurs at or after i in va
   |   |   |   |    then
12 |   |   |   |   |   return false;
13 |   |   |   |   end
14 |   |   |   end
15 |   |   end
16 |   end
17 end
18 return true;
```

The above procedure checks whether, given a word $v = a_0 \cdots a_{n-1}$ and a letter a, the word va admits a k-switching $S_{va,k}(i, j, \ell, n-k+2)$. Now if v is a minimal representative of $[v]_k$ and algorithm 1 returns false on the input v, a, then va is the minimal representative of $[va]_k$. Indeed, in this case va avoids k-switchings that give lexicographically smaller words. Observe also that all words of length less than $2k$ are all minimal representatives of their equivalence classes.

Now using the above procedure, the number of minimal representatives of a given length may be computed using the following algorithm.

Algorithm 2: Algorithm to compute $\mathcal{P}_m^{(k)}(n)$ for given integers k, m, n.

input : Integer $k \geqslant 1$, integer $n \geqslant 1$, alphabet Σ.
output: $\mathcal{P}_m^{(k)}(n)$, where $m = |\Sigma|$

```
1  if n ⩽ 2k − 1 then
2  |  return m^n;
3  else
4  |  Q ← Σ^{2k−1} as queue;
5  |  v ← dequeue(Q);
6  |  while |v| < n do
7  |  |  for Letter a ∈ Σ do
8  |  |  |  if algorithm 1(v,a) then
9  |  |  |  |  enqueue(Q, va);
10 |  |  |  end
11 |  |  end
12 |  |  v ← dequeue(Q);
13 |  end
14 end
15 return length(Q) + 1;
```

In the case of singletons, we may define a similar algorithm to algorithm 1. Indeed, consider the algorithm obtained by modifying row 10 to have, instead of $b \lhd c$, we have $b \neq c$. Now it should be clear that the modified algorithm returns true if and only if the word va is a k-abelian singleton, given that v is a k-abelian singleton.

To count the number of k-abelian singletons, we only need to modify the row 8 of algorithm 2, to invoke the newly modified algorithm described above instead of algorithm 1.

Appendix B

Sequences of Numbers of Equivalence Classes

This appendix contains numbers of k-abelian equivalence classes and numbers of k-abelian singletons. These values have been computed using implementations of the algorithms described in Appendix A.

B.1 Numbers of Equivalence Classes

The values $\mathcal{P}_2^{(2)}(n)$ for $n = 0, \ldots, 100$

1, 2, 4, 8, 14, 22, 32, 44, 58, 74, 92, 112, 134, 158, 184, 212, 242, 274, 308, 344, 382, 422, 464, 508, 554, 602, 652, 704, 758, 814, 872, 932, 994, 1058, 1124, 1192, 1262, 1334, 1408, 1484, 1562, 1642, 1724, 1808, 1894, 1982, 2072, 2164, 2258, 2354, 2452, 2552, 2654, 2758, 2864, 2972, 3082, 3194, 3308, 3424, 3542, 3662, 3784, 3908, 4034, 4162, 4292, 4424, 4558, 4694, 4832, 4972, 5114, 5258, 5404, 5552, 5702, 5854, 6008, 6164, 6322, 6482, 6644, 6808, 6974, 7142, 7312, 7484, 7658, 7834, 8012, 8192, 8374, 8558, 8744, 8932, 9122, 9314, 9508, 9704, 9902

The values $\mathcal{P}_2^{(3)}(n)$ for $n = 0, \ldots, 81$

1, 2, 4, 8, 16, 32, 60, 106, 176, 280, 426, 626, 892, 1238, 1678, 2230, 2910, 3738, 4734, 5920, 7318, 8954, 10852, 13040, 15546, 18400, 21632, 25276, 29364, 33932, 39016, 44654, 50884, 57748, 65286, 73542, 82560, 92386, 103066, 114650, 127186, 140726, 155322, 171028, 187898, 205990, 225360, 246068, 268174, 291740, 316828, 343504, 371832, 401880, 433716, 467410, 503032, 540656, 580354, 622202, 666276, 712654, 761414, 812638, 866406, 922802, 981910, 1043816, 1108606, 1176370, 1247196, 1321176, 1398402, 1478968, 1562968, 1650500, 1741660, 1836548, 1935264, 2037910, 2144588, 2255404

The values $\mathcal{P}_2^{(4)}(n)$ for $n = 0, \ldots, 55$

1, 2, 4, 8, 16, 32, 64, 128, 250, 478, 886, 1590, 2768, 4680, 7692, 12326, 19286, 29524, 44300, 65256, 94496, 134710, 189270, 262374, 359210, 486124,

650802, 862534, 1132430, 1473700, 1901988, 2435694, 3096338, 3909016, 4902824, 6111338, 7573208, 9332712, 11440372, 13953708, 16937952, 20466812, 24623434, 29501272, 35205068, 41851996, 49572780, 58512844, 68833756, 80714528, 94353070, 109967848, 127799534, 148112650, 171197634, 197372694

The values $\mathcal{P}_2^{(5)}(n)$ for $n = 0, \ldots, 30$

1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1014, 1992, 3874, 7438, 14078, 26266, 48254, 87298, 155474, 272578, 470496, 799794, 1339378, 2210844, 3598784, 5780204, 9165504, 14356150, 22223932, 34019930, 51521788

The values $\mathcal{P}_3^{(2)}(n)$ for $n = 0, \ldots, 50$

1, 3, 9, 27, 75, 186, 414, 840, 1578, 2784, 4662, 7476, 11556, 17313, 25245, 35955, 50157, 68697, 92559, 122889, 161001, 208404, 266808, 338154, 424620, 528654, 652980, 800634, 974970, 1179699, 1418895, 1697037, 2019015, 2390175, 2816325, 3303783, 3859383, 4490526, 5205186, 6011964, 6920094, 7939500, 9080802, 10355376, 11775360, 13353717, 15104241, 17041623, 19181457, 21540309, 24135723

The values $\mathcal{P}_3^{(3)}(n)$ for $n = 0, \ldots, 20$

1, 3, 9, 27, 81, 243, 717, 2073, 5814, 15774, 41250, 103842, 251436, 586056, 1316847, 2858295, 6006132, 12244842, 24270909, 46865127, 88315263

B.2 Numbers of Singletons

The values $\mathcal{S}_2^{(2)}(n)$ for $n = 0, \ldots, 100$

1, 2, 4, 8, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100, 102, 104, 106, 108, 110, 112, 114, 116, 118, 120, 122, 124, 126, 128, 130, 132, 134, 136, 138, 140, 142, 144, 146, 148, 150, 152, 154, 156, 158, 160, 162, 164, 166, 168, 170, 172, 174, 176, 178, 180, 182, 184, 186, 188, 190, 192, 194, 196, 198, 200, 202, 204

The values $\mathcal{S}_2^{(3)}(n)$ for $n = 0, \ldots, 100$

0,1.0 1,2.0 2,4.0 3,8.0 4,16.0 5,32.0 56, 86, 112, 142, 164, 192, 220, 248, 276, 310, 338, 372, 406, 440, 474, 514, 548, 588, 628, 668, 708, 754, 794, 840, 886, 932, 978, 1030, 1076, 1128, 1180, 1232, 1284, 1342, 1394, 1452, 1510, 1568, 1626, 1690, 1748, 1812, 1876, 1940, 2004, 2074, 2138, 2208, 2278, 2348, 2418, 2494, 2564, 2640, 2716, 2792, 2868, 2950, 3026, 3108, 3190, 3272, 3354, 3442, 3524, 3612, 3700, 3788, 3876, 3970, 4058, 4152, 4246, 4340, 4434, 4534, 4628, 4728, 4828, 4928, 5028, 5134, 5234, 5340, 5446, 5552, 5658, 5770, 5876, 5988, 6100, 6212, 6324, 6442, 6554

The values $\mathcal{S}_2^{(4)}(n)$, $n = 0, \ldots, 100$

1, 2, 4, 8, 16, 32, 64, 128, 244, 446, 760, 1202, 1784, 2486, 3272, 4140, 5052, 6012, 7046, 8142, 9300, 10538, 11840, 13210, 14656, 16180, 17756, 19424, 21156, 22958, 24842, 26796, 28820, 30930, 33108, 35360, 37692, 40098, 42578, 45158, 47792, 50504, 53306, 56184, 59136, 62188, 65304, 68494, 71782, 75144, 78586, 82128, 85732, 89428, 93214, 97078, 101020, 105070, 109186, 113392, 117702, 122074, 126538, 131116, 135756, 140490, 145326, 150240, 155238, 160356, 165538, 170814, 176198, 181658, 187218, 192890, 198624, 204460, 210412, 216438, 222562, 228810, 235114, 241530, 248062, 254668, 261374, 268208, 275110, 282114, 289244, 296444, 303750, 311190, 318694, 326316, 334050, 341862, 349788, 357850, 365974

The values $\mathcal{S}_2^{(5)}(n)$ for $n = 0, \ldots, 55$

1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1004, 1938, 3664, 6742, 12032, 20870, 35004, 56914, 89544, 136324, 200778, 286504, 396572, 533934, 700014, 896762, 1125230, 1386962, 1683012, 2015768, 2386114, 2797250, 3250650, 3749452, 4295018, 4891412, 5539324, 6242502, 7002982, 7823982, 8706584, 9655342, 10671144, 11758610, 12919294, 14156478, 15472042, 16871266, 18354570, 19926674, 21589688, 23347734, 25202186, 27159222, 29218694, 31386376,

The values $\mathcal{S}_3^{(2)}(n)$ for $n = 0, \ldots, 100$

1, 3, 9, 27, 69, 135, 207, 273, 345, 423, 507, 597, 693, 795, 903, 1017, 1137, 1263, 1395, 1533, 1677, 1827, 1983, 2145, 2313, 2487, 2667, 2853, 3045, 3243, 3447, 3657, 3873, 4095, 4323, 4557, 4797, 5043, 5295, 5553, 5817, 6087, 6363, 6645, 6933, 7227, 7527, 7833, 8145, 8463, 8787, 9117, 9453, 9795, 10143, 10497, 10857, 11223, 11595, 11973, 12357, 12747, 13143, 13545, 13953, 14367, 14787, 15213, 15645, 16083, 16527, 16977, 17433, 17895, 18363, 18837, 19317, 19803, 20295, 20793, 21297, 21807, 22323, 22845, 23373, 23907, 24447, 24993, 25545, 26103, 26667, 27237, 27813, 28395, 28983, 29577, 30177, 30783, 31395, 32013, 32637

The values $\mathcal{S}_3^{(3)}(n)$ for $n = 0, \ldots, 33$

1, 3, 9, 27, 81, 243, 705, 1965, 5133, 12543, 28347, 59223, 113445, 200055, 325881, 495819, 712575, 981141, 1305345, 1691769, 2146701, 2677611, 3289701, 3991029, 4788819, 5690025, 6703461, 7836837, 9097701, 10496643, 12041667, 13741521, 15607401, 17648571

Appendix C

Transition Tables of Automata

We give the transition tables of the Minimal DFA recognizing $L_{k,\Sigma,\triangleleft}$ illustrated in Figures 5.1–5.6. The states are enumerated starting from 0. The initial state of the automaton is always indexed by 0. All other states are final, except the one with the largest index. This state is the sink state. The transitions are represented by triples $[p, \delta(p,a), \delta(p,b)]$ for the binary case, quadruples $[p, \delta(p,a), \delta(p,b), \delta(p,c)]$ in the ternary case.

C.1 Transition Tables of Minimal Representatives

Transitions of the Minimal DFA Recognizing $L_{2,\{a,b\},\triangleleft}$

[0,1,2], [1,1,3], [2,1,4], [3,5,6], [4,7,4], [5,9,3], [6,8,6], [7,7,9], [8,9,9], [9,9,9]

Transitions of the Minimal DFA Recognizing $L_{3,\{a,b\},\triangleleft}$

[0,1,2], [1,3,4], [2,5,6], [3,3,7], [4,5,8], [5,3,9], [6,5,10], [7,11,12], [8,13,14], [9,15,16], [10,17,10], [11,18,19], [12,20,21], [13,22,23], [14,24,14], [15,25,9], [16,26,27], [17,28,29], [18,48,7], [19,30,31], [20,32,33], [21,34,21], [22,22,35], [23,48,16], [24,24,36], [25,25,48], [26,25,23], [27,37,27], [28,28,38], [29,39,48], [30,36,19], [31,40,41], [32,48,35], [33,48,31], [34,42,36], [35,48,12], [36,48,48], [37,25,36], [38,43,48], [39,25,29], [40,36,33], [41,44,41], [42,48,36], [43,45,46], [44,36,36], [45,48,38], [46,47,48], [47,36,46], [48,48,48]

Transitions of the Minimal DFA Recognizing $L_{2,\{a,b,c\},\triangleleft}$

[0,1,2,3], [1,1,4,5], [2,1,6,7], [3,1,8,9], [4,10,11,12], [5,13,14,15], [6,16,6,7], [7,17,18,19], [8,20,21,22], [9,23,24,9], [10,65,4,5], [11,25,11,12], [12,26,27,28], [13,65,29,5], [14,30,31,32], [15,29,33,15], [16,16,65,34], [17,17,35,36], [18,37,65,22], [19,38,37,19], [20,20,39,40], [21,41,21,22], [22,37,18,42], [23,23,43,65], [24,23,44,65], [25,65,65,34], [26,65,35,36], [27,45,65,32], [28,46,47,28], [29,65,29,65], [30,65,65,40],

[31,48,31,32], [32,45,27,49], [33,45,33,65], [34,25,65,50], [35,65,65,51], [36,52,53,54], [37,37,65,65], [38,38,45,65], [39,55,56,57], [40,65,58,59], [41,41,65,48], [42,37,37,42], [43,60,33,65], [44,37,44,65], [45,65,65,65], [46,65,45,65], [47,45,65,65], [48,65,65,48], [49,45,47,49], [50,45,65,50], [51,26,53,54], [52,65,45,36], [53,65,65,61], [54,46,45,54], [55,65,39,40], [56,48,56,57], [57,65,62,63], [58,30,56,57], [59,65,33,59], [60,65,43,65], [61,45,53,64], [62,45,65,57], [63,65,47,63], [64,45,45,64], [65,65,65,65]

Transitions of the Minimal DFA Recognizing $L_{4,\{a,b\},\lhd}$

[0,1,2], [1,3,4], [2,5,6], [3,7,8], [4,9,10], [5,11,12], [6,13,14], [7,7,15], [8,9,16], [9,11,17], [10,13,18], [11,7,19], [12,9,20], [13,11,21], [14,13,22], [15,23,24], [16,25,26], [17,27,28], [18,29,30], [19,31,32], [20,33,34], [21,35,36], [22,37,22], [23,38,39], [24,40,41], [25,42,43], [26,44,45], [27,46,17], [28,47,48], [29,49,50], [30,51,30], [31,52,53], [32,54,55], [33,56,57], [34,58,59], [35,60,61], [36,62,63], [37,64,65], [38,66,67], [39,68,69], [40,70,71], [41,72,73], [42,74,75], [43,76,77], [44,78,79], [45,80,45], [46,81,82], [47,83,84], [48,85,86], [49,87,88], [50,89,90], [51,91,92], [52,93,19], [53,94,95], [54,96,97], [55,98,99], [56,100,101], [57,102,36], [58,103,104], [59,105,59], [60,106,107], [61,108,109], [62,110,57], [63,111,112], [64,113,114], [65,115,116], [66,935,15], [67,117,118], [68,119,39], [69,120,121], [70,122,123], [71,124,125], [72,126,127], [73,128,73], [74,74,129], [75,130,32], [76,131,61], [77,132,133], [78,134,135], [79,136,137], [80,138,139], [81,81,140], [82,935,141], [83,142,143], [84,144,77], [85,145,146], [86,147,86], [87,87,148], [88,149,150], [89,151,152], [90,935,63], [91,153,154], [92,155,156], [93,93,935], [94,157,53], [95,158,159], [96,93,75], [97,160,161], [98,162,163], [99,164,99], [100,100,165], [101,166,167], [102,110,935], [103,168,169], [104,102,90], [105,170,171], [106,106,172], [107,173,174], [108,175,61], [109,176,177], [110,178,179], [111,110,104], [112,180,112], [113,113,181], [114,182,183], [115,184,185], [116,186,935], [117,187,188], [118,189,190], [119,191,192], [120,193,194], [121,195,196], [122,935,129], [123,130,118], [124,197,198], [125,199,200], [126,201,202], [127,203,204], [128,205,206], [129,130,24], [130,935,207], [131,208,209], [132,210,84], [133,211,212], [134,134,213], [135,130,150], [136,214,152], [137,935,133], [138,138,215], [139,216,156], [140,935,217], [141,218,219], [142,142,220], [143,221,167], [144,222,935], [145,223,224], [146,144,137], [147,225,226], [148,227,228], [149,229,230], [150,935,55], [151,231,232], [152,233,234], [153,153,235], [154,236,156], [155,91,139], [156,935,935], [157,93,82], [158,237,238], [159,239,240], [160,241,242], [161,243,244], [162,93,135], [163,245,246], [164,247,248], [165,249,250], [166,251,252], [167,253,254], [168,168,255], [169,256,257], [170,258,259], [171,102,156], [172,260,261], [173,262,263], [174,264,265], [175,175,266], [176,222,267], [177,268,269], [178,178,270], [179,271,935], [180,110,171], [181,272,273], [182,274,275], [183,276,935], [184,277,278], [185,279,280], [186,110,281], [187,156,67], [188,282,283], [189,284,285], [190,286,287], [191,935,140], [192,935,288], [193,289,290], [194,291,125], [195,292,293], [196,294,196], [197,295,296], [198,297,298], [199,299,194], [200,300,301], [201,935,213], [202,130,302], [203,303,304], [204,935,200], [205,305,215], [206,306,156], [207,130,935], [208,208,307], [209,935,174], [210,210,308], [211,210,146], [212,309,212], [213,130,228], [214,310,311], [215,130,156], [216,312,139], [217,313,314], [218,157,935], [219,315,316], [220,317,250], [221,935,252], [222,222,156], [223,223,318], [224,319,257], [225,225,320], [226,144,156], [227,321,322], [228,935,41], [229,93,88], [230,323,324], [231,231,325], [232,326,327], [233,328,152], [234,935,177], [235,329,156], [236,330,248], [237,93,143], [238,331,161], [239,332,333], [240,334,240], [241,93,209], [242,335,336], [243,337,238], [244,338,339], [245,340,341], [246,935,244], [247,93,215], [248,342,156], [249,343,344], [250,345,346], [251,93,347], [252,935,348], [253,349,350], [254,351,352], [255,353,354], [256,355,356], [257,935,254], [258,258,357], [259,358,156], [260,359,360], [261,361,362], [262,93,107], [263,363,364], [264,935,365], [265,366,367], [266,935,368], [267,144,109], [268,222,369], [269,370,269], [270,371,935], [271,372,935], [272,373,374], [273,375,935], [274,93,114], [275,376,377], [276,378,379], [277,277,380], [278,381,382], [279,383,185], [280,384,935], [281,102,116], [282,385,188], [283,386,387], [284,156,123], [285,388,389], [286,390,391], [287,392,287], [288,393,394], [289,935,220], [290,395,396], [291,397,935], [292,398,399], [293,291,204], [294,400,401], [295,935,307], [296,935,402], [297,403,198], [298,404,405], [299,406,308], [300,299,293], [301,407,301], [302,935,190], [303,408,409], [304,410,411], [305,935,215], [306,412,206], [307,935,261], [308,156,935], [309,210,226], [310,310,413], [311,935,327], [312,312,414], [313,119,935], [314,415,416], [315,417,935], [316,418,316], [317,935,344], [318,419,354], [319,935,356], [320,420,156], [321,421,422], [322,423,424], [323,417,230], [324,935,159], [325,425,426], [326,427,428], [327,935,265], [328,328,429], [329,430,206], [330,93,154], [331,431,935], [332,93,224], [333,331,246], [334,432,433], [335,434,242], [336,435,436], [337,93,308], [338,337,333], [339,437,339], [340,93,311], [341,438,439], [342,440,248], [343,441,442], [344,935,443], [345,444,445], [346,446,447], [347,166,448], [348,449,450], [349,93,451], [350,156,452], [351,453,454], [352,455,352], [353,456,457], [354,935,346], [355,93,458], [356,935,459], [357,460,156], [358,461,414], [359,462,463], [360,464,465], [361,935,466], [362,467,468], [363,434,263], [364,469,470], [365,471,364], [366,935,472], [367,473,367], [368,156,368], [369,144,234], [370,222,226], [371,474,935], [372,93,179], [373,475,476], [374,477,478], [375,479,480], [376,481,275], [377,482,935], [378,93,483], [379,484,485], [380,486,487], [381,488,489], [382,490,935], [383,383,406], [384,222,491], [385,156,192], [386,492,493], [387,494,495], [388,496,497], [389,498,499], [390,156,202], [391,500,501], [392,502,503], [393,385,935], [394,504,505], [395,935,506], [396,507,508], [397,414,156], [398,935,318], [399,509,510], [400,511,320], [401,291,156], [402,512,513], [403,429,266], [404,397,514], [405,515,516], [406,935,308], [407,299,401], [408,935,413], [409,935,517], [410,518,304], [411,935,405], [412,420,414], [413,935,426], [414,935,156], [415,519,935], [416,520,416], [417,93,521], [418,440,935], [419,935,457], [420,935,414], [421,935,148], [422,522,302], [423,519,322], [424,935,121], [425,523,524], [426,935,362], [427,93,232], [428,525,526], [429,935,266], [430,527,528], [431,93,156], [432,93,320], [433,331,156], [434,93,266], [435,431,529], [436,530,531], [437,337,433], [438,532,341], [439,935,436], [440,93,414], [441,935,533], [442,534,535], [443,536,537], [444,538,539], [445,156,540], [446,541,542], [447,543,447], [448,544,545], [449,546,547], [450,548,549], [451,414,167], [452,550,551], [453,93,552], [454,156,553], [455,554,555], [456,556,557], [457,935,558], [458,256,559], [459,935,450], [460,560,414], [461,93,259], [462,935,172], [463,561,402], [464,403,360], [465,562,563], [466,564,465], [467,935,565], [468,566,468], [469,935,567], [470,568,569], [471,241,263], [472,570,526], [473,935,248], [474,571,572], [475,935,181], [476,573,574], [477,575,374], [478,576,935], [479,577,578], [480,579,580], [481,93,581], [482,582,583], [483,130,183], [484,584,585], [485,586,935], [486,587,588], [487,589,935], [488,93,278], [489,590,591], [490,935,592], [491,144,280], [492,156,290], [493,593,389], [494,594,595], [495,596,495], [496,156,296], [497,597,598], [498,599,493], [499,600,601], [500,602,603], [501,935,499], [502,156,215], [503,604,156], [504,605,935], [505,606,505], [506,935,607], [507,608,609], [508,610,611], [509,935,612], [510,935,508], [511,935,320], [512,935,613], [513,614,615], [514,291,298], [515,397,616], [516,617,516], [517,935,513], [518,618,429], [519,619,620], [520,412,935], [521,935,621], [522,622,623], [523,624,625], [524,626,627], [525,532,428], [526,935,470], [527,935,235], [528,628,156], [529,331,336], [530,431,629], [531,630,531], [532,93,429], [533,249,631], [534,632,506], [535,633,634], [536,635,636], [537,637,638], [538,935,639], [539,414,396], [540,640,641], [541,642,643], [542,156,644], [543,645,555], [544,646,547], [545,647,648], [546,93,649], [547,935,650], [548,651,652], [549,653,549], [550,337,350], [551,654,655], [552,414,257], [553,935,551], [554,93,397], [555,156,156], [556,935,656], [557,657,658], [558,935,537], [559,935,545], [560,659,660], [561,661,662], [562,935,663], [563,664,665], [564,197,360], [565,666,627], [566,935,206], [567,331,364], [568,935,667], [569,668,569], [570,340,428], [571,935,270], [572,669,935], [573,670,671], [574,672,935], [575,673,674], [576,675,676], [577,935,677], [578,130,574], [579,678,679], [580,680,935], [581,935,681], [582,93,682], [583,331,485], [584,93,683], [585,684,685], [586,337,583], [587,686,687], [588,688,689], [589,935,690], [590,691,489], [591,692,935], [592,693,591], [593,555,935], [594,156,399], [595,593,501], [596,694,695], [597,696,497], [598,697,698], [599,156,308], [600,599,595], [601,699,601], [602,156,409], [603,700,701], [604,702,503], [605,156,620], [606,702,935], [607,703,704], [608,156,539], [609,156,705], [610,706,707], [611,708,611], [612,935,709], [613,710,711], [614,935,712], [615,713,615], [616,291,411], [617,397,401], [618,935,429], [619,935,714], [620,935,715], [621,935,219], [622,156,422], [623,716,717], [624,935,325], [625,718,517], [626,518,524], [627,935,563], [628,719,503], [629,331,439], [630,431,433], [631,720,721], [632,156,442], [633,722,723], [634,724,725], [635,726,727], [636,935,728], [637,729,730], [638,731,638], [639,414,250], [640,299,445], [641,732,733], [642,935,734], [643,414,510], [644,935,641], [645,735,397], [646,93,736], [647,737,652], [648,738,648], [649,221,448], [650,739,740], [651,93,741], [652,935,742],

[653,432,414], [654,337,454], [655,743,655], [656,353,744], [657,745,612], [658,935,634], [659,935,357], [660,746,156], [661,156,463], [662,747,711], [663,291,465], [664,935,748], [665,749,665], [666,303,524], [667,331,526], [668,935,433], [669,750,935], [670,156,476], [671,751,752], [672,753,754], [673,935,755], [674,935,756], [675,757,758], [676,291,580], [677,130,273], [678,759,760], [679,761,762], [680,299,676], [681,763,935], [682,764,765], [683,935,382], [684,691,585], [685,766,935], [686,935,380], [687,767,768], [688,769,588], [689,770,935], [690,771,689], [691,93,406], [692,935,772], [693,584,489], [694,156,320], [695,593,156], [696,156,266], [697,555,773], [698,774,775], [699,599,695], [700,776,603], [701,935,698], [702,156,414], [703,777,723], [704,778,779], [705,780,781], [706,156,643], [707,156,782], [708,783,555], [709,935,704], [710,496,662], [711,784,785], [712,786,787], [713,935,503], [714,935,788], [715,935,394], [716,605,623], [717,935,387], [718,789,790], [719,156,528], [720,791,636], [721,792,793], [722,156,794], [723,935,795], [724,796,797], [725,798,725], [726,935,799], [727,395,535], [728,800,801], [729,802,803], [730,935,804], [731,400,414], [732,299,542], [733,805,733], [734,414,354], [735,935,397], [736,414,448], [737,93,806], [738,554,414], [739,337,547], [740,807,808], [741,319,559], [742,935,740], [743,337,555], [744,935,721], [745,156,557], [746,809,414], [747,696,662], [748,291,627], [749,935,401], [750,156,572], [751,810,671], [752,811,935], [753,156,578], [754,812,813], [755,935,814], [756,815,935], [757,935,816], [758,817,818], [759,935,819], [760,935,768], [761,769,679], [762,820,935], [763,481,935], [764,935,821], [765,822,935], [766,431,823], [767,824,825], [768,826,935], [769,827,406], [770,935,828], [771,678,588], [772,331,591], [773,593,598], [774,555,829], [775,830,775], [776,156,429], [777,156,727], [778,831,797], [779,832,779], [780,599,609], [781,833,834], [782,935,781], [783,156,397], [784,935,835], [785,836,837], [786,602,790], [787,935,785], [788,935,314], [789,156,625], [790,838,787], [791,839,794], [792,840,730], [793,841,793], [794,414,535], [795,842,843], [796,156,844], [797,935,845], [798,783,414], [799,317,631], [800,299,636], [801,846,847], [802,935,848], [803,509,658], [804,935,801], [805,299,555], [806,414,559], [807,337,652], [808,849,808], [809,156,660], [810,156,674], [811,850,851], [812,852,853], [813,854,935], [814,855,935], [815,810,935], [816,856,857], [817,935,858], [818,859,935], [819,935,487], [820,397,860], [821,935,861], [822,862,863], [823,331,685], [824,156,687], [825,864,865], [826,935,866], [827,935,406], [828,291,689], [829,593,701], [830,555,695], [831,156,803], [832,694,414], [833,599,707], [834,867,834], [835,593,711], [836,935,868], [837,869,837], [838,776,790], [839,935,870], [840,871,844], [841,645,414], [842,599,723], [843,872,873], [844,414,658], [845,935,843], [846,299,730], [847,874,847], [848,419,744], [849,337,414], [850,156,758], [851,593,813], [852,156,760], [853,875,876], [854,599,851], [855,575,935], [856,935,877], [857,878,935], [858,935,879], [859,880,881], [860,291,762], [861,882,935], [862,93,883], [863,156,884], [864,885,825], [865,886,935], [866,887,865], [867,599,555], [868,593,787], [869,935,695], [870,414,631], [871,935,888], [872,599,797], [873,889,873], [874,299,414], [875,885,853], [876,890,935], [877,935,891], [878,892,893], [879,894,935], [880,156,895], [881,156,896], [882,897,898], [883,414,765], [884,899,935], [885,156,406], [886,935,900], [887,852,825], [888,414,744], [889,599,414], [890,555,901], [891,902,935], [892,903,895], [893,156,904], [894,905,906], [895,414,818], [896,907,935], [897,93,908], [898,935,909], [899,337,863], [900,593,865], [901,593,876], [902,910,911], [903,935,912], [904,913,935], [905,156,914], [906,935,915], [907,599,881], [908,764,916], [909,917,935], [910,918,914], [911,935,919], [912,414,857], [913,299,893], [914,817,920], [915,921,935], [916,922,935], [917,337,898], [918,935,923], [919,924,935], [920,925,935], [921,599,906], [922,926,898], [923,856,927], [924,299,911], [925,928,906], [926,93,929], [927,930,935], [928,156,931], [929,414,916], [930,932,911], [931,414,920], [932,933,931], [933,935,934], [934,414,927], [935,935,935]

C.2 Transition Tables for Singletons

Transitions of the Minimal DFA Recognizing $L_{2,\{a,b\},\mathbf{sing}}$

[0,1,2], [1,3,4], [2,5,6], [3,3,7], [4,8,9], [5,10,11], [6,12,6], [7,14,7], [8,14,11], [9,13,9], [10,10,13], [11,8,14], [12,12,14], [13,14,14], [14,14,14]

Transitions of the Minimal DFA Recognizing $L_{3,\{a,b\},\mathbf{sing}}$

[0,1,2], [1,3,4], [2,5,6], [3,7,8], [4,9,10], [5,11,12], [6,13,14], [7,7,15], [8,16,17], [9,18,19], [10,20,21], [11,22,23], [12,24,25], [13,26,27], [14,28,14], [15,29,30], [16,31,32], [17,33,34], [18,35,36], [19,24,37], [20,38,39], [21,40,21], [22,22,41], [23,42,43], [24,44,19], [25,45,46], [26,47,48], [27,49,50], [28,51,52], [29,86,53], [30,54,55], [31,86,36], [32,56,37], [33,57,58], [34,59,34], [35,35,60], [36,42,37], [37,86,37], [38,61,62], [39,86,50], [40,40,63], [41,63,41], [42,31,86], [43,64,65], [44,44,86], [45,44,39], [46,66,46], [47,47,67], [48,68,69], [49,44,70], [50,45,86], [51,71,72], [52,73,86], [53,29,37], [54,86,74], [55,75,55], [56,63,32], [57,86,62], [58,86,76], [59,75,63], [60,63,37], [61,61,77], [62,86,69], [63,86,86], [64,57,86], [65,78,65], [66,44,63], [67,63,79], [68,80,86], [69,64,86], [70,49,63], [71,71,79], [72,81,86], [73,44,52], [74,86,82], [75,86,63], [76,83,86], [77,86,79], [78,75,86], [79,63,86], [80,86,84], [81,85,86], [82,54,86], [83,63,58], [84,68,63], [85,86,72], [86,86,86]

Transitions of the Minimal DFA Recognizing $L_{2,\{a,b,c\},\mathbf{sing}}$

[0,1,2,3], [1,4,5,6], [2,7,8,9], [3,10,11,12], [4,4,13,14], [5,15,16,17], [6,18,19,20], [7,21,22,23], [8,24,8,25], [9,26,27,28], [10,29,30,31], [11,32,33,34], [12,35,36,12], [13,83,37,38], [14,83,39,40], [15,83,22,41], [16,41,16,42], [17,43,44,45], [18,83,46,31], [19,47,48,49], [20,46,50,20], [21,21,41,51], [22,15,83,41], [23,52,53,45], [24,54,83,55],

[25,56,83,57], [26,58,59,60], [27,61,83,34], [28,62,61,28], [29,29,63,46], [30,64,48,65], [31,18,46,83], [32,58,66,67], [33,68,33,61], [34,61,27,83], [35,69,70,83], [36,71,72,83], [37,83,37,41], [38,83,73,51], [39,83,63,74], [40,83,46,40], [41,83,83,41], [42,75,83,42], [43,83,59,83], [44,75,83,49], [45,76,77,45], [46,83,46,83], [47,83,83,67], [48,78,48,77], [49,75,44,83], [50,75,50,83], [51,83,75,51], [52,83,75,60], [53,47,83,83], [54,54,83,41], [55,79,83,42], [56,68,83,80], [57,61,83,57], [58,58,78,76], [59,83,83,65], [60,52,75,83], [61,61,83,83], [62,62,75,83], [63,83,63,75], [64,83,66,75], [65,43,83,83], [66,64,83,75], [67,83,53,83], [68,68,83,75], [69,69,46,83], [70,81,50,83], [71,62,82,83], [72,61,72,83], [73,83,83,74], [74,83,73,83], [75,83,83,83], [76,83,75,83], [77,75,83,83], [78,83,83,75], [79,83,83,80], [80,79,83,83], [81,83,82,83], [82,81,83,83], [83,83,83,83]

Transitions of the Minimal DFA Recognizing $L_{4,\{a,b\},\mathbf{sing}} \setminus b\{a,b\}^*$

[0,1,663], [1,2,3], [2,4,5], [3,6,7], [4,8,9], [5,10,11], [6,12,13], [7,14,15], [8,8,16], [9,17,18], [10,19,20], [11,21,22], [12,23,24], [13,25,26], [14,27,28], [15,29,30], [16,31,32], [17,33,34], [18,35,36], [19,37,38], [20,39,40], [21,41,42], [22,43,44], [23,45,46], [24,47,48], [25,49,50], [26,51,52], [27,53,54], [28,55,56], [29,57,58], [30,59,30], [31,60,61], [32,62,63], [33,64,65], [34,66,67], [35,68,69], [36,70,71], [37,72,73], [38,47,74], [39,75,76], [40,77,78], [41,79,80], [42,81,82], [43,83,84], [44,85,44], [45,45,86], [46,87,88], [47,89,90], [48,91,92], [49,93,94], [50,25,95], [51,96,97], [52,98,99], [53,100,101], [54,102,103], [55,104,105], [56,106,107], [57,108,109], [58,110,111], [59,112,113], [60,663,114], [61,115,116], [62,117,118], [63,119,120], [64,663,73], [65,121,74], [66,122,123], [67,124,125], [68,126,127], [69,128,129], [70,130,131], [71,132,71], [72,72,133], [73,87,74], [74,134,135], [75,75,136], [76,39,137], [77,138,97], [78,139,140], [79,141,142], [80,143,103], [81,144,105], [82,145,107], [83,146,147], [84,148,149], [85,150,151], [86,152,153], [87,154,90], [88,155,156], [89,157,38], [90,663,74], [91,158,159], [92,160,161], [93,162,163], [94,663,164], [95,165,166], [96,167,168], [97,169,82], [98,170,171], [99,172,99], [100,100,173], [101,174,175], [102,176,177], [103,178,107], [104,179,180], [105,181,182], [106,183,184], [107,663,107], [108,185,186], [109,187,188], [110,189,190], [111,663,191], [112,192,193], [113,194,195], [114,196,74], [115,197,198], [116,199,200], [117,663,201], [118,202,203], [119,204,205], [120,206,120], [121,207,90], [122,197,136], [123,66,208], [124,209,210], [125,211,212], [126,663,142], [127,143,213], [128,214,215], [129,216,107], [130,217,218], [131,219,220], [132,221,222], [133,152,74], [134,663,223], [135,224,225], [136,663,226], [137,169,137], [138,227,228], [139,229,171], [140,230,140], [141,141,231], [142,143,175], [143,663,232], [144,233,234], [145,235,97], [146,236,237], [147,143,188], [148,238,190], [149,663,239], [150,150,240], [151,241,195], [152,195,90], [153,242,243], [154,64,663], [155,244,245], [156,246,247], [157,157,663], [158,157,248], [159,249,250], [160,251,252], [161,253,161], [162,162,254], [163,663,255], [164,256,257], [165,258,663], [166,259,260], [167,261,262], [168,263,264], [169,265,663], [170,266,267], [171,169,149], [172,268,269], [173,270,271], [174,272,273], [175,274,107], [176,157,275], [177,276,277], [178,278,232], [179,279,280], [180,281,282], [181,283,105], [182,169,107], [183,284,285], [184,286,56], [185,185,287], [186,288,289], [187,290,291], [188,663,292], [189,293,294], [190,295,296], [191,297,663], [192,298,299], [193,300,195], [194,301,151], [195,663,663], [196,60,90], [197,663,136], [198,115,302], [199,303,304], [200,305,306], [201,143,307], [202,308,309], [203,310,107], [204,663,311], [205,312,313], [206,314,315], [207,195,65], [208,316,208], [209,317,318], [210,316,129], [211,319,320], [212,321,212], [213,322,107], [214,323,324], [215,325,326], [216,327,210], [217,663,237], [218,143,328], [219,329,330], [220,663,331], [221,314,240], [222,332,195], [223,663,333], [224,663,334], [225,335,225], [226,195,226], [227,336,337], [228,263,663], [229,338,339], [230,340,269], [231,143,271], [232,143,663], [233,341,342], [234,663,282], [235,235,343], [236,236,344], [237,143,289], [238,345,346], [239,347,663], [240,143,195], [241,348,151], [242,195,349], [243,350,351], [244,352,663], [245,353,354], [246,355,356], [247,357,247], [248,358,264], [249,359,663], [250,360,107], [251,157,361], [252,362,363], [253,364,365], [254,663,366], [255,367,368], [256,369,663], [257,370,371], [258,372,373], [259,374,663], [260,375,260], [261,261,376], [262,377,378], [263,663,379], [264,380,107], [265,265,195], [266,381,382], [267,383,384], [268,268,385], [269,169,195], [270,195,386], [271,387,107], [272,388,663], [273,389,390], [274,391,232], [275,281,392], [276,393,394], [277,395,107], [278,157,80], [279,279,396], [280,397,398], [281,176,399], [282,400,107], [283,283,399], [284,401,402], [285,403,663], [286,183,663], [287,404,405], [288,406,407], [289,663,408], [290,157,409], [291,410,411], [292,412,663], [293,413,414], [294,415,416], [295,417,190], [296,663,418], [297,183,419], [298,298,420], [299,421,195], [300,422,423], [301,424,425], [302,426,302], [303,663,427], [304,426,203], [305,428,429], [306,430,306], [307,431,107], [308,663,432], [309,433,434], [310,435,304], [311,143,436], [312,437,438], [313,663,439], [314,663,240], [315,440,195], [316,441,663], [317,663,337], [318,442,663], [319,443,444], [320,316,220], [321,445,446], [322,447,232], [323,663,342], [324,663,448], [325,449,215], [326,316,107], [327,435,343], [328,663,450], [329,451,452], [330,453,454], [331,455,663], [332,456,222], [333,134,107], [334,663,457], [335,663,358], [336,336,458], [337,377,663], [338,459,460], [339,383,663], [340,340,461], [341,341,462], [342,663,398], [343,195,663], [344,143,405], [345,463,464], [346,663,416], [347,235,171], [348,348,358], [349,343,465], [350,195,466], [351,467,351], [352,663,468], [353,469,663], [354,470,107], [355,471,663], [356,472,473], [357,474,475], [358,663,195], [359,157,234], [360,476,477], [361,358,384], [362,478,663], [363,663,479], [364,157,441], [365,480,195], [366,343,366], [367,481,663], [368,482,483], [369,157,373], [370,484,663], [371,480,371], [372,485,486], [373,663,487], [374,488,489], [375,348,663], [376,490,491], [377,663,492], [378,493,107], [379,663,277],

[380,158,343], [381,381,494], [382,495,496], [383,663,497], [384,663,498], [385,335,195], [386,270,499], [387,195,232], [388,663,500], [389,501,502], [390,503,107], [391,126,663], [392,358,107], [393,157,399], [394,276,504], [395,505,663], [396,506,507], [397,272,399], [398,508,107], [399,663,509], [400,663,510], [401,401,511], [402,512,663], [403,513,663], [404,195,514], [405,663,515], [406,516,663], [407,517,518], [408,519,663], [409,415,520], [410,521,522], [411,663,523], [412,524,232], [413,413,525], [414,526,527], [415,290,528], [416,663,529], [417,417,528], [418,169,663], [419,286,111], [420,530,195], [421,531,532], [422,157,425], [423,533,195], [424,534,535], [425,536,195], [426,358,663], [427,537,663], [428,663,538], [429,426,313], [430,539,540], [431,117,232], [432,663,541], [433,542,309], [434,426,107], [435,663,343], [436,663,543], [437,663,544], [438,545,546], [439,547,663], [440,335,315], [441,358,195], [442,663,548], [443,663,460], [444,549,663], [445,539,461], [446,316,195], [447,195,127], [448,550,107], [449,542,399], [450,551,663], [451,663,464], [452,663,552], [453,553,330], [454,663,554], [455,327,320], [456,335,358], [457,663,555], [458,490,663], [459,459,556], [460,495,663], [461,335,663], [462,663,507], [463,463,557], [464,663,527], [465,242,107], [466,343,558], [467,195,559], [468,358,378], [469,323,663], [470,560,561], [471,663,562], [472,563,663], [473,663,564], [474,565,663], [475,566,195], [476,157,343], [477,567,250], [478,157,346], [479,568,663], [480,569,663], [481,570,663], [482,571,663], [483,566,483], [484,157,489], [485,485,572], [486,663,573], [487,256,107], [488,574,575], [489,663,576], [490,663,572], [491,577,107], [492,663,390], [493,244,343], [494,578,579], [495,663,580], [496,663,581], [497,663,411], [498,582,663], [499,343,107], [500,397,392], [501,542,663], [502,389,583], [503,584,663], [504,567,107], [505,157,228], [506,195,399], [507,585,107], [508,663,586], [509,195,107], [510,249,663], [511,343,663], [512,587,663], [513,157,285], [514,404,585], [515,387,663], [516,663,588], [517,589,590], [518,663,591], [519,592,232], [520,663,426], [521,157,528], [522,410,593], [523,594,663], [524,157,147], [525,595,596], [526,406,528], [527,663,597], [528,663,435], [529,598,663], [530,195,420], [531,599,663], [532,600,195], [533,569,423], [534,534,601], [535,602,195], [536,422,358], [537,663,603], [538,604,663], [539,663,461], [540,426,195], [541,605,107], [542,663,399], [543,606,663], [544,663,607], [545,608,438], [546,663,609], [547,435,429], [548,663,610], [549,663,611], [550,663,612], [551,613,232], [552,663,614], [553,608,528], [554,316,663], [555,224,663], [556,578,663], [557,663,596], [558,663,615], [559,343,195], [560,435,663], [561,609,354], [562,358,496], [563,451,663], [564,616,663], [565,663,441], [566,461,663], [567,617,663], [568,476,618], [569,157,358], [570,663,486], [571,619,663], [572,663,499], [573,367,107], [574,574,620], [575,663,621], [576,663,622], [577,195,343], [578,663,620], [579,663,623], [580,663,518], [581,624,663], [582,251,343], [583,609,107], [584,317,663], [585,663,511], [586,353,663], [587,625,663], [588,526,520], [589,608,663], [590,517,626], [591,627,663], [592,217,663], [593,663,628], [594,629,663], [595,195,528], [596,663,630], [597,631,663], [598,663,632], [599,663,535], [600,461,532], [601,633,195], [602,531,358], [603,663,634], [604,663,635], [605,663,636], [606,204,232], [607,663,637], [608,663,528], [609,426,663], [610,638,107], [611,663,639], [612,640,663], [613,195,218], [614,641,663], [615,350,663], [616,560,642], [617,157,195], [618,567,363], [619,663,575], [620,663,585], [621,663,643], [622,370,663], [623,577,663], [624,355,343], [625,663,402], [626,663,644], [627,645,663], [628,567,663], [629,157,339], [630,585,663], [631,663,646], [632,362,663], [633,195,358], [634,647,107], [635,663,648], [636,649,663], [637,650,663], [638,651,663], [639,663,652], [640,653,663], [641,663,654], [642,609,473], [643,482,663], [644,609,663], [645,443,663], [646,472,663], [647,303,663], [648,663,655], [649,308,663], [650,663,656], [651,195,318], [652,657,663], [653,195,324], [654,658,663], [655,659,663], [656,660,663], [657,661,663], [658,662,663], [659,428,663], [660,437,663], [661,195,444], [662,195,452], [663,663,663]

Bibliography

[1] T. van Aardenne-Ehrenfest and N.G. de Bruijn. "Circuits and Trees in Oriented Linear Graphs". *Simon Stevin* 28 (1951), pp. 203–217.

[2] B. Adamczewski. "Balances for Fixed Points of Primitive Substitutions". *Theoretical Computer Science* 307.1 (2003), pp. 47–75. DOI: 10.1016/S0304-3975(03)00092-6.

[3] M. Aigner. *Markov's Theorem and 100 Years of the Uniqueness Conjecture. A Mathematical Journey from Irrational Numbers to Perfect Matchings*. Springer, 2013. DOI: 10.1007/978-3-319-00888-2.

[4] M. Albert and J. Lawrence. "A Proof of Ehrenfeucht's Conjecture". *Theoretical Computer Science* 41 (1985), pp. 121–123. DOI: 10.1016/0304-3975(85)90066-0.

[5] J.-P. Allouche and J. Shallit. "The Ubiquitous Prouhet-Thue-Morse Sequence". In: *Sequences and their Applications*. Ed. by C. Ding, T. Helleseth, and H. Niederreiter. London: Springer London, 1999, pp. 1–16.

[6] M. Andrașiu, G. Păun, J. Dassow, and A. Salomaa. "Language-theoretic Problems Arising from Richelieu Cryptosystems". *Theoretical Computer Science* 116.2 (1993), pp. 339–357. DOI: 10.1016/0304-3975(93)90327-P.

[7] F. Bernstein. "Über eine Anwendung der Mengenlehre auf ein aus der Theorie der säkularen Störungen herrührendes Problem". *Mathematische Annalen* 71.3 (1911), pp. 417–439. DOI: 10.1007/BF01456856.

[8] F. Bernstein. "Über geometrische Wahrscheinlichkeit und über das Axiom der beschränkten Arithmetisierbarkeit der Beobachtungen". *Mathematische Annalen* 72.4 (1912), pp. 585–587. DOI: 10.1007/BF01456678.

[9] J. Berstel. "Axel Thue's work on repetitions in words". *Séries Formelles et Combinatoire Algébrique* (1997).

[10] J. Berstel. "Sturmian and episturmian words. A survey of some recent results". In: *Algebraic Informatics. Second International Conference, CAI 2007*. Ed. by S. Bozapalidis and G. Rahonis. Lecture Notes in Computer Science 4728. Springer, 2007, pp. 23–47. DOI: 10.1007/978-3-540-75414-5_2.

[11] J. Berstel and L. Boasson. "The Set of Minimal Words of a Context-free Language is Context-free". *Journal of Computer and System Sciences* 55.3 (1997), pp. 477–488. DOI: 10.1006/jcss.1997.1497.

[12] F. Blanchet-Sadri, N. Fox, and N. Rampersad. "On the Asymptotic Abelian Complexity of Morphic Words". *Advances in Applied Mathematics* 61.C (2014), pp. 46–84. DOI: 10.1016/j.aam.2014.08.005.

[13] J.A. Bondy and U.S.R. Murty. *Graph Theory*. Graduate Texts in Mathematics 244. New York: Springer, 2008. ISBN: 978-3642142789.

[14] E. Borel. "Les probabilités denombrables et leurs applications arithmétiques". *Rendiconti Del Circolo Matematico Di Palermo* 27 (1909), pp. 247–271.

[15] N.G. de Bruijn. *Acknowledgement of Priority to C. Flye Sainte-Marie on the Counting of Circular Arrangements of 2^n Zeros and Ones that Show Each n-letter Word Exactly Once*. Tech. rep. (EUT report. WSK, Dept. of Mathematics and Computing Science; Vol. 75-WSK-06. Technische Hogeschool Eindhoven Nederland, 1975.

[16] A. Carpi and A. de Luca. "Uniform Words". *Advances in Applied Mathematics* 32.3 (2004), pp. 485–522. DOI: 10.1016/S0196-8858(03)00057-5.

[17] J. Cassaigne, J. Karhumäki, and S. Puzynina. "On k-abelian palindromes". *Information and Computation* 260 (2018), pp. 89–98. DOI: 10.1016/j.ic.2018.04.001.

[18] J. Cassaigne, J. Karhumäki, S. Puzynina, and M.A. Whiteland. "k-Abelian Equivalence and Rationality". In: *Developments in Language Theory - 20th International Conference, DLT 2016, Montréal, Canada, July 25–28, 2016, Proceedings*. Ed. by S. Brlek and C. Reutenauer. Vol. 9840. Lecture Notes in Computer Science. Springer, 2016, pp. 77–88. DOI: 10.1007/978-3-662-53132-7_7.

[19] J. Cassaigne, J. Karhumäki, S. Puzynina, and M.A. Whiteland. "k-Abelian Equivalence and Rationality". *Fundamenta Informaticae* 154.1–4 (2017), pp. 65–94. DOI: 10.3233/FI-2017-1553.

[20] J. Cassaigne, J. Karhumäki, and A. Saarela. "On Growth and Fluctuation of k-Abelian Complexity". *European Journal of Combinatorics* 65 (2017), pp. 92–105. DOI: 10.1016/j.ejc.2017.05.006.

[21] J. Cassaigne and F. Nicolas. "Factor Complexity". In: *Combinatorics, Automata and Number Theory*. Vol. 135. Encyclopedia Math. Appl. Cambridge Univ. Press, Cambridge, 2010, pp. 163–247.

[22] J. Chen, X. Lü, and W. Wu. "On the k-abelian complexity of the Cantor sequence". *Journal of Combinatorial Theory, Series A* 155 (2018), pp. 287–303. DOI: 10.1016/j.jcta.2017.11.010.

[23] C. Choffrut and J. Karhumäki. "Combinatorics of Words". In: *Handbook of Formal Languages: Volume 1 Word, Language, Grammar*. Ed. by G. Rozenberg and A. Salomaa. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 329–438. DOI: 10.1007/978-3-642-59136-5_6.

[24] E.M. Coven and G.A. Hedlund. "Sequences with Minimal Block Growth". *Mathematical Systems Theory* 7.2 (1973), pp. 138–153. DOI: 10.1007/BF01762232.

[25] K. Culik II and J. Karhumäki. "Systems of equations over a free monoid and Ehrenfeucht's conjecture". *Discrete Mathematics* 43.2 (1983), pp. 139–153. DOI: https://doi.org/10.1016/0012-365X(83)90152-8.

[26] T.W. Cusick and M.E. Flahive. *The Markoff and Lagrange Spectra*. Mathematical Surveys and Monographs 30. Providence, Rhode Island: American Mathematical Society, 1989.

[27] D. Damanik and D. Lenz. "The index of Sturmian sequences". *European Journal of Combinatorics* 23 (2002), pp. 23–29. DOI: 10.1006/eujc.2000.0496.

[28] C. Degni and A.A. Drisko. "Gray-ordered Binary Necklaces". *Electronic Journal of Combinatorics* 14.1 (2007). URL: http://www.combinatorics.org/Volume_14/Abstracts/v14i1r7.html.

[29] F. Dekking. "Strongly non-repetitive sequences and progression-free sets". *Journal of Combinatorial Theory, Series A* 27.2 (1979), pp. 181–185. DOI: 10.1016/0097-3165(79)90044-X.

[30] R. Devyatov. "On Factor Complexity of Morphic Sequences". *Moscow Mathematical Journal* 18.2 (2018), pp. 211–303.

[31] F. Durand. "A Characterization of Substitutive Sequences Using Return Words". *Discrete Mathematics* 179.1–3 (1998), pp. 89–101. DOI: 10.1016/S0012-365X(97)00029-0.

[32] F. Durand. "Linearly Recurrent Subshifts have a Finite Number of Non-periodic Subshift Factors". *Ergodic Theory and Dynamical Systems* 20.4 (2000), pp. 1061–1078.

[33] T. Ehlers, F. Manea, R. Mercaş, and D. Nowotka. "k-Abelian Pattern Matching". *Journal of Discrete Algorithms* 34 (2015), pp. 37–48. DOI: 10.1016/j.jda.2015.05.004.

[34] S. Eilenberg. *Automata, Languages, and Machines*. Vol. A. New York, New York, USA: Academic Press, Inc., 1974. ISBN: 978-0-12-234001-7.

[35] S. Ferenczi and T. Monteil. "Infinite Words with Uniform Frequencies, and Invariant Measures". In: *Combinatorics, Automata and Number Theory*. Vol. 135. Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, Cambridge, 2010, pp. 373–409.

[36] G. Fici, A. Langiu, T. Lecroq, A. Lefebvre, F. Mignosi, Jarkko Peltomäki, and É. Prieur-Gaston. "Abelian powers and repetitions in Sturmian words". *Theoretical Computer Science* 635 (2016), pp. 16–34. DOI: 10.1016/j.tcs.2016.04.039.

[37] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. 1st ed. New York, NY, USA: Cambridge University Press, 2009. ISBN: 978-0-521-89806-5.

[38] G.A. Freiman. *Diophantine approximation and geometry of numbers (Markov's problem)*. (Russian). Kalininskii Gosudarstvennyi Universitet, Kalinin, 1975.

[39] D.D. Freydenberger, P. Gawrychowski, J. Karhumäki, F. Manea, and W. Rytter. "Testing k-binomial equivalence". *arXiv e-prints* (2015), arXiv:1509.00622. arXiv: 1509.00622 [cs.FL].

[40] P. Gawrychowski, D. Krieger, N. Rampersad, and J. Shallit. "Finding the Growth Rate of a Regular of Context-Free Language in Polynomial Time". In: *Developments in Language Theory: 12th International Conference, DLT 2008, Kyoto, Japan, September 16-19, 2008. Proceedings*. Ed. by M. Ito and M. Toyama. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 339–358. DOI: 10.1007/978-3-540-85780-8_27.

[41] H. Gruber and M. Holzer. "From Finite Automata to Regular Expressions and Back — A Summary on Descriptional Complexity". *International Journal of Foundations of Computer Science* 26.8 (2015), pp. 1009–1040. DOI: 10.1142/S0129054115400110.

[42] V.S. Guba. "Equivalence of Infinite Systems of Equations in Free Groups and Semigroups to Finite Subsystems". *Matematicheskie Zametki* 40.3 (1986). In Russian, pp. 321–324, 428.

[43] M. Hall, Jr. "On the sum and products of continued fractions". *Annals of Mathematics* 48.4 (1947), pp. 966–993. DOI: 10.2307/1969389.

[44] T. Harju, J. Karhumäki, and W. Plandowski. "Compactness of Systems of Equations in Semigroups". *International Journal of Algebra and Computation* 7.4 (1997), pp. 457–470. DOI: 10.1142/S0218196797000204.

[45] J.E. Hopcroft and J.D. Ullman. *Introduction To Automata Theory, Languages, And Computation*. 1st. Addison-Wesley Publishing Co., Inc., 1979. ISBN: 0-201-02988-X.

[46] Huova, Mari. "Existence of an Infinite Ternary 64-Abelian Square-free Word". *RAIRO Theoretical Informatics and Applications* 48.3 (2014), pp. 307–314. DOI: 10.1051/ita/2014012.

[47] M. Huova and J. Karhumäki. *Observations and Problems on k-Abelian Avoidability*. 2011. arXiv: 1104.4273 [math.CO].

[48] M. Huova, J. Karhumäki, and A. Saarela. "Problems in between Words and Abelian Words: k-Abelian Avoidability". *Theoretical Computer Science* 454 (2012), pp. 172–177. DOI: 10.1016/j.tcs.2012.03.010.

[49] M. Huova, J. Karhumäki, A. Saarela, and K. Saari. "Local Squares, Periodicity and Finite Automata". *Rainbow of Computer Science: Dedicated to Hermann Maurer on the Occasion of His 70th Birthday*. Ed. by C.S. Calude, G. Rozenberg, and A. Salomaa. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 90–101. DOI: 10.1007/978-3-642-19391-0_7.

[50] M. Huova and A. Saarela. "Strongly k-Abelian Repetitions". In: *Combinatorics on Words - 9th International Conference, WORDS 2013, Turku, Finland, September 16-20. Proceedings*. 2013, pp. 161–168. DOI: 10.1007/978-3-642-40579-2_18.

[51] A. Hurwitz. "Ueber die angenäherte Darstellung der Irrationalzahlen durch rationale Brüche". *Mathematische Annalen* 39.2 (1891), pp. 279–284. DOI: 10.1007/BF01206656.

[52] J. Justin and G. Pirillo. "Fractional powers in Sturmian words". *Theoretical Computer Science* 255 (2001), pp. 363–376. DOI: 10.1016/S0304-3975(99)90294-3.

[53] J. Karhumäki. "Generalized Parikh Mappings and Homomorphisms". *Information and Control* 47.3 (1980), pp. 155–165. DOI: 10.1016/S0019-9958(80)90493-3.

[54] J. Karhumäki and W. Plandowski. "On the Size of Independent Systems of Equations in Semigroups". *Theoretical Computer Science* 168.1 (1996), pp. 105–119. DOI: 10.1016/S0304-3975(96)00064-3.

[55] J. Karhumäki, S. Puzynina, M. Rao, and M.A. Whiteland. "On Cardinalities of k-Abelian Equivalence Classes". *Theoretical Computer Science* 658, Part A (2017). Formal Languages and Automata: Models, Methods and Application In honour of the 70th birthday of Antonio Restivo, pp. 190–204. DOI: 10.1016/j.tcs.2016.06.010.

[56] J. Karhumäki, S. Puzynina, and A. Saarela. "Fine and Wilf's Theorem for k-Abelian Periods". *International Journal of Foundations of Computer Science* 24.7 (2013), pp. 1135–1152. DOI: 10.1142/S0129054113400352.

[57] J. Karhumäki, A. Saarela, and L.Q. Zamboni. "On a Generalization of Abelian Equivalence and Complexity of Infinite Words". *Journal of Combinatorial Theory, Series A* 120.8 (2013), pp. 2189–2206. DOI: 10.1016/j.jcta.2013.08.008.

[58] J. Karhumäki, A. Saarela, and L.Q. Zamboni. "Variations of the Morse–Hedlund Theorem for k-Abelian Equivalence". *Acta Cybernetica* 23.1 (2017), pp. 175–189. DOI: 10.14232/actacyb.23.1.2017.11.

[59] J. Karhumäki and M.A. Whiteland. "Regularity of k-Abelian Equivalence Classes of Fixed Cardinality". In: *Adventures Between Lower Bounds and Higher Altitudes - Essays Dedicated to Juraj Hromkovič on the Occasion of His 60th Birthday*. 2018, pp. 49–62. DOI: 10.1007/978-3-319-98355-4_4.

[60] V. Keränen. "Abelian Squares Are Avoidable on 4 Letters". In: *Proceedings of the 19th International Colloquium on Automata, Languages and Programming*. ICALP '92. Springer-Verlag, 1992, pp. 41–52.

[61] J. Lawrence. "The Non-existence of Finite Test Sets for Set-equivalence of Finite Substitions". *Bulletin of the EATCS* 28 (1986), pp. 34–36.

[62] M. Lejeune, J. Leroy, and M. Rigo. "Computing the k-binomial complexity of the Thue–Morse word". *arXiv e-prints* (2018), arXiv:1812.07330. arXiv: 1812.07330 [cs.DM].

[63] M. Lothaire. *Combinatorics on Words*. Vol. 17. Encyclopedia of Mathematics and its Applications. Addison-Wesley, Advanced Book Program, World Science Division, 1983. ISBN: 978-0-201-13516-9.

[64] M. Lothaire. *Algebraic Combinatorics on Words*. Vol. 90. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2002. DOI: 10.1017/CBO9781107326019.

[65] M. Lothaire. *Applied Combinatorics on Words*. Vol. 105. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2005. DOI: 10.1017/CBO9781107341005.

[66] B. Madill and N. Rampersad. "The Abelian Complexity of the Paperfolding Word". *Discrete Mathematics* 313.7 (2013), pp. 831–838. DOI: 10.1016/j.disc.2013.01.005.

[67] A. Mateescu, A. Salomaa, K. Salomaa, and S. Yu. "A sharpening of the Parikh mapping". *RAIRO-Theoritcal Informatics and Applications* 35.6 (2001), pp. 551–564. DOI: 10.1051/ita:2001131.

[68] R. Mercaş and A. Saarela. "3-Abelian Cubes Are Avoidable on Binary Alphabets". In: *Developments in Language Theory*. Ed. by M.-P. Béal and O. Carton. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 374–383. DOI: 10.1007/978-3-642-38771-5_33.

[69] R. Mercaş and A. Saarela. "5-Abelian Cubes are Avoidable on Binary Alphabets". *RAIRO - Theoretical Informatics and Applications* 48.4 (2014), pp. 467–478. DOI: 10.1051/ita/2014020.

[70] F. Mignosi and G. Pirillo. "Repetitions in the Fibonacci infinite word". *RAIRO Informatique Théorique et Applications* 26.3 (1992), pp. 199–204.

[71] H.M. Morse. "Recurrent geodesics on a surface of negative curvature". *Transactions of the American Mathematical Society* 22.1 (1921), pp. 84–100. DOI: 10.2307/1988844.

[72] M. Morse and G.A. Hedlund. "Symbolic Dynamics II. Sturmian Trajectories". *American Journal of Mathematics* 62.1 (1940), pp. 1–42. ISSN: 00029327, 10806377. URL: https://www.jstor.org/stable/2371431.

[73] J. Mykkeltveit. "A Proof of Golomb's Conjecture for the de Bruijn graph". *Journal of Combinatorial Theory, Series B* 13 (1972), pp. 40–45. DOI: 10.1016/0095-8956(72)90006-8.

[74] D. Nowotka and A. Saarela. "An Optimal Bound on the Solution Sets of One-Variable Word Equations and its Consequences". In: *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Ed. by I. Chatzigiannakis, C. Kaklamanis, D. Marx, and D. Sannella. Vol. 107. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, 136:1–136:13. DOI: 10.4230/LIPIcs.ICALP.2018.136.

[75] D. Nowotka and A. Saarela. "One-Variable Word Equations and Three-Variable Constant-Free Word Equations". *International Journal of Foundations of Computer Science* 29.5 (2018). DOI: 10.1142/S0129054118420121.

[76] J.-J. Pansiot. "Complexité des Facteurs des Mots Infinis Engendrés par Morphimes Itérés". In: *Automata, Languages and Programming, 11th Colloquium, Antwerp, Belgium, July 16-20, 1984, Proceedings*. Ed. by J. Paredaens. Vol. 172. Lecture Notes in Computer Science. Springer, 1984, pp. 380–389. ISBN: 3-540-13345-3. DOI: 10.1007/3-540-13345-3_34.

[77] J.-J. Pansiot. "Subword Complexities and Iteration". *Bulletin of the EATCS* 26 (1985), pp. 55–62.

[78] A. Parreau, M. Rigo, E. Rowland, and É. Vandomme. "A New Approach to the 2-Regularity of the ℓ-Abelian Complexity of 2-Automatic Sequences". *The Electronic Journal of Combinatorics* 22.1 (2015), P1.27. URL: https://www.combinatorics.org/ojs/index.php/eljc/article/view/v22i1p27.

[79] J. Peltomäki. "Introducing Privileged Words: Privileged Complexity of Sturmian Words". *Theoretical Computer Science* 500 (2013), pp. 57–67. DOI: 10.1016/j.tcs.2013.05.028.

[80] J. Peltomäki. "Privileged Words and Sturmian Words". Ph.D. dissertation. Turku, Finland: Turku Centre for Computer Science, University of Turku, 2016. URL: http://urn.fi/URN:ISBN:978-952-12-3422-4.

[81] J. Peltomäki and M.A. Whiteland. *On k-Abelian Equivalence and Generalized Lagrange Spectra*. (Submitted). 2018. URL: https://arxiv.org/abs/1809.09047v1.

[82] J. Peltomäki and M.A. Whiteland. "Every nonnegative real number is an abelian critical exponent". In: *Proceedings of WORDS 2019*. Lecture Notes in Computer Science. (To appear). Springer, 2019.

[83] E. Prouhet. "Mémoire sur quelques Relations entre les Puissances des Nombres". *Comptes rendus de l'Académie des Sciences Paris*. I 33 (1851), p. 225.

[84] N. Pytheas Fogg. *Substitutions in Dynamics, Arithmetics and Combinatorics*. Lecture Notes in Mathematics 1794. Springer, 2002. DOI: 10.1007/b13861.

[85] M. Queffélec. *Substitution Dynamical Systems — Spectral Analysis*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2010. ISBN: 978-3-642-11212-6.

[86] M. Rao. "On Some Generalizations of Abelian Power Avoidability". *Theoretical Computer Science* 601 (2015), pp. 39–46. DOI: 10.1016/j.tcs.2015.07.026.

[87] M. Rao, M. Rigo, and P. Salimov. "Avoiding 2-binomial squares and cubes". *Theoretical Computer Science* 572 (2015), pp. 83–91. DOI: 10.1016/j.tcs.2015.01.029.

[88] M. Rao and M. Rosenfeld. "Avoidability of Long k-Abelian Repetitions". *Mathematics of Computation* 85.302 (2016), pp. 3051–3060. DOI: 10.1090/mcom/3085.

[89] G. Rauzy. "Suites á Termes dans un Alphabet Fini". *Seminaire de Théorie des Nombres de Bordeaux* 12 (1982-1983), pp. 1–16. URL: http://eudml.org/doc/182163.

[90] L. Rédei. *The Theory of Finitely Generated Commutative Semigroups*. International series of monographs in pure and applied mathematics. Pergamon Press, 1965. DOI: doi.org/10.1016/C2013-0-01797-5.

[91] G. Richomme, K. Saari, and L.Q. Zamboni. "Abelian Complexity of Minimal Subshifts". *Journal of the London Mathematical Society* 83.1 (2011), pp. 79–95. DOI: 10.1112/jlms/jdq063.

[92] M. Rigo. "Relations on Words". *Indagationes Mathematicae* 28.1 (2017), pp. 183–204. DOI: 10.1016/j.indag.2016.11.018.

[93] M. Rigo and P. Salimov. "Another Generalization of Abelian Equivalence: Binomial Complexity of Infinite Words". *Theoretical Computer Science* 601 (2015), pp. 47–57. DOI: 10.1016/j.tcs.2015.07.025.

[94] M. Rigo, P. Salimov, and É. Vandomme. "Some properties of abelian return words". *Journal of Integer Sequences* 16 (2013).

[95] J. Riordan. *An introduction to combinatorial analysis*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1958.

[96] A.M. Rockett and P. Szüsz. *Continued Fractions*. World Scientific Publishing, 1992.

[97] F. Ruskey and J. Sawada. "An Efficient Algorithm for Generating Necklaces with Fixed Density". *SIAM J. Comput.* 29.2 (1999), pp. 671–684. DOI: 10.1137/S0097539798344112.

[98] A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series*. Texts and Monographs in Computer Science. Springer, 1978. DOI: 10.1007/978-1-4612-6264-0.

[99] C. Savage. "A Survey of Combinatorial Gray Codes". *SIAM Review* 39.4 (1997), pp. 605–629. DOI: 10.1137/S0036144595295272.

[100] A. Shevlyakov. "Elements of Algebraic Geometry Over a Free Semilattice". *Algebra and Logic* 54.3 (2015), pp. 258–271. DOI: 10.1007/s10469-015-9345-6.

[101] K.G. Subramanian, A.M. Huey, and A.K. Nagar. "On Parikh Martices". *International Journal of Foundations of Computer Science* 20.02 (2009), pp. 211–219. DOI: 10.1142/S0129054109006528.

[102] A. Szilard, S. Yu, K. Zhang, and J. Shallit. "Characterizing regular languages with polynomial densities". In: *Mathematical Foundations of Computer Science 1992: 17th International Symposium Prague, Czechoslovakia, August 24–28, 1992 Proceedings*. Ed. by I.M. Havel and V. Koubek. Berlin, Heidelberg: Springer, 1992, pp. 494–503. DOI: 10.1007/3-540-55808-X_48.

[103] A. Thue. "Über Unendliche Zeichenreihen". *Skrifter Udgivne af Videnskabsselskabet i Christiania: Mathematisk-naturvidenskabelig Klasse* (1906). Reprinted in "Selected mathematical papers of Axel Thue," T. Nagell, ed., Universitetsforlaget, Oslo, 1977, pp. 139–158., pp. 1–22.

[104] A. Thue. "Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen". *Skrifter Udgivne af Videnskabsselskabet i Christiania: Mathematisk-naturvidenskabelig Klasse* 1 (1912). Reprinted in "Selected mathematical papers of Axel Thue," T. Nagell, ed., Universitetsforlaget, Oslo, 1977, pp. 413–478., pp. 1–67.

[105] O. Veblen. "An Application of Modular Equations in Analysis Situs". *Annals of Mathematics* 14.1/4 (1912), pp. 86–94. ISSN: 0003486X. URL: http://www.jstor.org/stable/1967604.

[106] L. Vuillon. "A Characterization of Sturmian Words by Return Words". *European Journal of Combinatorics* 22.2 (2001), pp. 263–275. DOI: 10.1006/eujc.2000.0444.

[107] S.H. Weintraub. *Jordan Canonical Form: Theory and Practice*. Synthesis Lectures on Mathematics & Statistics. Morgan & Claypool Publishers, 2009. DOI: 10.2200/S00218ED1V01Y200908MAS006.

[108] M. Weston and V. Vajnovszki. "Gray Codes for Necklaces and Lyndon Words of Arbitrary Base". *Pure Mathematics and Applications* 17.1–2 (2006), pp. 175–182.

[109] M.A. Whiteland. "Asymptotic Abelian Complexities of Certain Morphic Binary Words". *Journal of Automata, Languages and Combinatorics* 24.1 (2019), pp. 89–114. DOI: 10.25596/jalc-2019-089.

[110] M.A. Whiteland. *On Equations Over Monoids Defined by Generalizations of Abelian Equivalence*. (in preparation). Parts presented at the Fifth Russian-Finnish Symposium on Discrete Mathematics (RuFiDiM), Veliky Novgorod, Russia, May 19–22. 2019.

[111] I. Zinovik, D. Kroening, and Y. Chebiryak. "Computing Binary Combinatorial Gray Codes Via Exhaustive Search With SAT Solvers". *IEEE Trans. Information Theory* 54.4 (2008), pp. 1819–1823. DOI: 10.1109/TIT.2008.917695.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönnqvist**, Knowledge-Lean Text Mining
228. **Mohammad-Hashem Hahgbayan**, Energy-Efficient and Reliable Computing in Dark Silicon Era
229. **Charmi Panchal**, Qualitative Methods for Modeling Biochemical Systems and Datasets: The Logicome and the Reaction Systems Approaches
230. **Erkki Kaila**, Utilizing Educational Technology in Computer Science and Programming Courses: Theory and Practice
231. **Fredrik Robertsén**, The Lattice Boltzmann Method, a Petaflop and Beyond
232. **Jonne Pohjankukka**, Machine Learning Approaches for Natural Resource Data
233. **Paavo Nevalainen**, Geometric Data Understanding: Deriving Case-Specific Features
234. **Michal Szabados**, An Algebraic Approach to Nivat's Conjecture
235. **Tuan Nguyen Gia**, Design for Energy-Efficient and Reliable Fog-Assisted Healthcare IoT Systems
236. **Anil Kanduri**, Adaptive Knobs for Resource Efficient Computing
237. **Veronika Suni**, Computational Methods and Tools for Protein Phosphorylation Analysis
238. **Behailu Negash**, Interoperating Networked Embedded Systems to Compose the Web of Things
239. **Kalle Rindell**, Development of Secure Software: Rationale, Standards and Practices
240. **Jurka Rahikkala**, On Top Management Support for Software Cost Estimation
241. **Markus A. Whiteland**, On the k-Abelian Equivalence Relation of Finite Words

Turku Centre *for* Computer Science

http://www. tucs.fi

tucs@abo.fi

University of Turku
Faculty of Science and Engineering
- Department of Future Technologies
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3837-6
ISSN 1239-1883

Markus A. Whiteland

On the k-Abelian Equivalence Relation of Finite Words

Markus A. Whiteland

On the k-Abelian Equivalence Relation of Finite Words