Contents

i

Introduction (version française)

Cette thèse est principalement consacrée à l'étude du caractère reconnaissable d'ensembles d'entiers par automate fini. Nous nous intéresserons donc aux représentations d'entiers dans différents types de sytèmes de numération.

Tout d'abord, les systèmes de numération de position, tels que le système classique décimal, le système binaire ou encore la numération de Zeckendorf, ont été beaucoup étudiés. Citons par exemple [Fra85, Lot02, BBLT06]. Dans un tel système de numération, un entier positif n peut être décomposé de manière unique par l'algorithme d'Euclide, également appelé algorithme glouton, comme combinaison linéaire

$$\sum_{i=0}^{N} d_i U_{N-i}, \quad d_0, \ldots, d_N \in \mathbb{N}, \quad d_0 \neq 0$$

d'éléments d'une suite strictement croissante $U = (U_m)_{m \geq 0}$ d'entiers telle que $U_0 = 1$. La représentation de n est alors $d_0 \cdots d_N$. Pour que l'alphabet des chiffres d_i soit fini, on suppose le quotient $\frac{U_{m+1}}{U_m}$ borné. Si la suite U est la suite $(k^m)_{m \geq 0}$, on obtient les systèmes classiques de numération en base entière k.

Pour la numération de Zeckendorf [Zec72], la suite U est la suite de Fibonacci définie par

$$U_0 = 1, \quad U_1 = 2, \quad U_{n+2} = U_{n+1} + U_n, \quad \forall n \geq 0.$$

Dans ce cadre, des questions comme les critères de divisibilité, triviales dans le cas d'une base entière, demandent déjà des développements plus élaborés. Par exemple, pour la numération de Zeckendorf, l'algorithme glouton impose que les écritures gloutonnes soient les mots sur $\{0, 1\}$ ne commençant pas par 0 et ne contenant pas deux 1 consécutifs. Fait qui pourrait s'avérer utile en théorie des codes.

Lorsqu'on se donne un système de numération, le fait de représenter les nombres permet de passer des entiers aux mots sur un alphabet fini et donc, de transposer des questions issues de l'arithmétique à des problèmes

de théorie des langages formels. Dans ce contexte, les langages réguliers, i.e.,
acceptés par les "machines les plus simples de la hiérarchie de Chomsky",
les automates finis, sont particulièrement intéressants. Les automates finis
sont des machines abstraites capables uniquement de lire des mots de façon
séquentielle et de décider, en un temps linéaire par rapport à l'entrée, si le
mot est accepté ou non. Ils sont également intéressants du point de vue de la
décidabilité, puisque bon nombre de leurs propriétés peuvent être testées al-
gorithmiquement : universalité, acceptation d'un langage vide ou fini, égalité
ou inclusion de langages, ... Alors que ces mêmes problèmes se révèlent le
plus souvent indécidables pour les langages algébriques.

Ainsi, si l'on veut tirer parti des langages réguliers, quand on s'intéresse
aux ensembles de nombres, il est naturel d'introduire la notion suivante. Un
ensemble d'entiers est dit *reconnaissable* selon un système de numération
donné si l'ensemble des représentations de ses éléments dans ce système
est régulier. En particulier, si l'ensemble des représentations des éléments
d'un ensemble d'entiers dans un système de numération de position U est
régulier, cet ensemble est dit *U-reconnaissable.* Les ensembles reconnais-
sables peuvent être étudiés sous plusieurs angles. En effet, si la suite U
satisfait une récurrence linéaire dont le polynôme caractéristique est le po-
lynôme minimum d'un nombre de Pisot, une généralisation d'un théorème
de Büchi montre que les ensembles U-reconnaissables sont aussi les en-
sembles définissables dans la logique de premier ordre $\langle \mathbb{N}, +, V_U \rangle$ où on pose
$V_U(0) = U_0 = 1$ et où, pour tout entier strictement positif n, $V_U(n)$ est le
plus petit U_i apparaissant dans la décomposition gloutonne de n avec un co-
efficient non nul. Les ensembles U-reconnaissables sont encore les ensembles
dont la suite caractéristique est engendrée par un type particulier de substi-
tution [BH97].

En particulier, dans le cas de la base entière, A. Cobham a notamment
prouvé qu'un ensemble est reconnaissable en base k si et seulement si sa suite
caractéristique est l'image par un codage d'un point fixe d'un morphisme k-
uniforme [Cob72]. Il a également montré que le caractère reconnaissable d'un
ensemble d'entiers dépend de la base de numération envisagée et que les seuls
ensembles d'entiers reconnaissables dans toute base entière sont les ensembles
ultimement périodiques [Cob69]. Ce dernier résultat est connu sous le nom de
premier théorème de Cobham et a donné lieu à un grand nombre de travaux
dans le but d'étendre ce résultat à d'autres types de systèmes de numération
et à d'autres contextes, citons par exemple [Sem77, Vil92a, Vil92b, BHMV94,
Fab94, MV96, Bès97, PB97, Dur98, Han98, Bès00, Dur02a, Dur02b, RW06,
BB07, AB08, Dur08, BB09, AB11, Dur11].

Si on veut calculer de manière effective au sein d'une numération, la

première des opérations réalisables par automate que l'on est en droit d'attendre est de tester si un mot donné est une représentation valide d'un entier ou non. Dès lors, il s'avère également intéressant de chercher des conditions nécessaires ou suffisantes pour que l'ensemble $\mathbb{N}$ des entiers positifs ou nuls soit reconnaissable. En effet, dans ce cas, on peut tester en un temps linéaire, grâce à un automate fini, si un mot donné est une représentation valide ou non.

Tous les systèmes de numération de position mentionnés ci-dessus sont *monotones*, i.e., l'application qui associe à un entier sa représentation est croissante pour l'ordre naturel sur $\mathbb{N}$ et pour l'ordre généalogique induit par l'ordre des chiffres. Pour ces raisons, P. Lecomte et M. Rigo [LR01] ont introduit les *systèmes de numération abstraits* $S = (L, \Sigma, <)$ constitués d'un langage régulier L sur l'alphabet fini Σ et d'un ordre total sur Σ. Dans ces systèmes, un entier positif n est représenté par le $(n + 1)$-ème mot du langage L. La notion de caractère reconnaissable peut être étendue au cas des systèmes abstraits ainsi que le premier théorème de Cobham. En effet, P. Lecomte et M. Rigo [LR01] ont prouvé que les ensembles ultimement périodiques sont toujours S-reconnaissables, laissant entrevoir une généralisation possible du premier théorème de Cobham. Ensuite, M. Rigo et A. Maes [RM02] ont généralisé le second théorème de Cobham en prouvant qu'un ensemble est S-reconnaissable si et seulement si sa suite caractéristique est morphique. Une généralisation du premier théorème de Cobham s'obtient alors grâce à ce dernier résultat et aux travaux de F. Durand [Dur98, Dur02a, Dur02b, Dur08, Dur11].

Les ensembles reconnaissables peuvent être naturellement définis dans un cadre multidimensionnel. Ainsi, O. Salon [Sal87] a montré que les parties de $\mathbb{N}^d$ qui sont k-reconnaissables pour une base entière k sont celles dont la suite caractéristique est générée par un morphisme k-uniforme. É. Charlier *et al.* [CKR10] ont généralisé ce résultat aux systèmes de numération abstraits, démontrant que les parties de $\mathbb{N}^d$ qui sont S-reconnaissables pour un système abstrait S sont celles dont la suite caractéristique est engendrée par un morphisme multidimensionnel satisfaisant une propriété du type "shape-symmetry" introduite par A. Maes [Mae99].

Cette thèse s'inscrit ainsi dans la lignée des travaux sur le caractère reconnaissable menés à Liège depuis quinze ans. Nous y développons les résultats obtenus dans [KLR09], [KLR10], [LRRV12], [LR13] et [CLR12] de la manière suivante.

Dans le premier chapitre, nous rappelons les notions de base utiles à la compréhension de cette dissertation. Tout d'abord, sont rappelés certaines définitions et certains résultats utiles concernant les mots, les langages et

les automates. Ensuite, nous introduisons les systèmes de numération de position et les systèmes de numération abstraits ainsi que les notions de suites automatique et régulière. Enfin, nous introduisons les premières définitions relatives aux morphismes multidimensionnels.

Dans le deuxième chapitre, nous étudions le caractère reconnaissable des ensembles *auto-générés* introduits initialement par C. Kimberling [Kim00]. Ces ensembles sont définis à partir d'un ensemble I d'entiers et d'un ensemble fini F de fonctions affines $\varphi_i : n \mapsto k_i n + \ell_i$. Par exemple, si l'on considère $I = \{1\}$ et $F = \{\varphi_0 : n \mapsto n, \varphi_1 : n \mapsto 2n, \varphi_2 : n \mapsto 4n - 1\}$, on obtient l'ensemble auto-généré $\mathcal{K}_1 := F^\omega(I) = \{1, 2, 3, 4, 6, 7, 8, 11, 12, 14, \ldots\}$. Nous montrons que, d'un point de vue technique, pour l'étude du caractère reconnaissable, on peut considérer uniquement des ensembles d'entiers positifs et se ramener au cas où $k_i \geq 2$ et $\ell_i \geq 0$ pour tout i. Nous prouvons que si les constantes multiplicatives k_i sont multiplicativement dépendantes, i.e., si les k_i sont tous puissance d'un même entier k, alors l'ensemble auto-généré $F^\omega(I)$ est reconnaissable pour le système de numération en base entière associé à k. Avec nos techniques, nous redémontrons également un résultat obtenu initialement par C. Kimberling [Kim00] et reprouvé par J.P. Allouche *et al.* [ASS05], donnant un lien fort entre l'ensemble $\mathcal{K}_1$ et la suite de Fibonacci définie comme le point fixe $\varphi^\omega(0) = 01001010\cdots$ du morphisme $\varphi : 0 \mapsto 01, 1 \mapsto 0$. Enfin, résolvant une conjecture de J.P. Allouche *et al.* [ASS05], nous donnons une condition suffisante sur les constantes multiplicatives k_i pour que l'ensemble auto-généré $F^\omega(I)$ ne soit reconnaissable pour aucun système de numération en base entière.

Rappelons que le premier théorème de Cobham montre que les seuls ensembles reconnaissables dans des bases multiplicativement indépendantes sont les ensembles ultimement périodiques. Dans ce contexte, le problème de décision suivant est naturel : *étant donné un système de numération abstrait S et un automate fini déterministe acceptant les représentations d'un ensemble X dans le système S, décider si l'ensemble X est ultimement périodique ou non.* Au vu du second théorème de Cobham et de sa généralisation par M. Rigo et A. Maes, ce problème se réénonce aussi en termes de mots morphiques : *étant donnés un morphisme et un codage générant un mot infini, décider si ce mot est ultimement périodique ou non.* Ce problème de décision a reçu beaucoup d'attention, d'abord pour les bases entières dès 1986 [HL86, Hon86, Pan86] et récemment, V. Marsault et J. Sakarovitch [MS13] ont donné une procédure de décision en $O(n \log n)$ dans le cas de la base entière et le cas général a été résolu indépendamment par F. Durand [Dur11, Dur12] et I. Mitrofanov [Mit11]. Ainsi, dans le troisième chapitre, nous explorons une nouvelle piste afin de donner une procédure de

décision alternative à ce problème. Cette voie pouvant à terme mener à des procédures de décision efficaces, nous étudions la *complexité syntaxique* des ensembles (ultimement) périodiques, i.e., le nombre de classes d'équivalence de la congruence syntaxique du langage des représentations des éléments ces ensembles. La congruence syntaxique d'un langage L sur Σ est définie par

$$u \leftrightarrow_L v \quad \Leftrightarrow \quad (\forall x, y \in \Sigma^*, xuy \in L \Leftrightarrow xvy \in L).$$

Nous donnons tout d'abord une formule exacte pour la complexité syntaxique de l'ensemble des représentations en base entière des éléments d'un ensemble périodique de période m dans trois cas particuliers : lorsque la période m est première avec la base, lorsque m est une puissance de la base et lorsque m s'écrit comme le produit d'une puissance de la base et d'un entier premier avec la base. Ensuite, nous donnons une borne inférieure pour la complexité syntaxique de l'ensemble des représentations en base entière des éléments d'un ensemble périodique d'entiers dans le cas général. Enfin, grâce à cette borne inférieure pour la complexité syntaxique, nous obtenons des bornes supérieures pour les possibles prépériode et période de l'ensemble considéré dans le problème de décision et la procédure de décision consiste alors à passer en revue les possibles combinaisons "prépériode-période" qui sont en nombre fini.

Dans les premiers chapitres, les langages réguliers jouent un rôle prépondérant puisqu'un ensemble est reconnaissable selon un système de numération donné si le langage de ses représentations dans ce système est régulier. L'automaticité d'un langage permet de mesurer à quel point un langage non régulier est "proche" ou non, en un certain sens, d'un langage régulier. Cette notion a été introduite par B. Trakhtenbrot [Tra64] et a donné lieu à plusieurs travaux dans le but de borner l'automaticité de certains langages non réguliers. Par exemple, J. Shallit [Sha96] a donné une borne inférieure pour l'automaticité des nombres premiers représentés en base entière, quelle que soit la base. Dans le quatrième chapitre, nous étudions l'automaticité de trois langages particuliers : les mots primitifs sur un alphabet fini, i.e., les mots qui ne s'écrivent pas comme une puissance d'un plus petit mot, les mots sans bords sur un alphabet fini, i.e., les mots qui n'ont pas de période non triviale, et les représentations des polynômes moniques irréductibles à coefficients dans un champ fini. On peut montrer qu'une borne inférieure pour l'automaticité d'un langage est donnée par le cardinal d'un ensemble de mots non-semblables pour ce langage, où deux mots sont dits semblables pour un langage L donné si, lorsqu'on leur ajoute le même suffixe, tous deux appartiennent à L ou aucun n'appartient à L. Pour les trois langages que nous étudions ici, la méthode consiste donc à construire un ensemble de

mots non-semblables pour ces langages et à en déduire une borne inférieure pour leur automaticité.

Dans le cinquième et dernier chapitre, répondant à une question de J. Sakarovitch, nous caractérisons les ensembles multidimensionnels qui sont reconnaissables selon tout système de numération abstrait en montrant que les parties de $\mathbb{N}^d$ reconnaissables selon tout système de numération abstrait sont exactement les ensembles 1-reconnaissables, i.e., les ensembles reconnaissables selon le système de numération abstrait construit sur le langage a^*. Nous montrons par exemple qu'un ensemble 1-reconnaissable comme $X = \{(n, 2m) : n, m \in \mathbb{N} \text{ and } n > 2m\}$ dont les représentations unaires sont données par $((a,a)^2)^*(a, \#)^*(a, \#)$ et dont la suite caractéristique est l'image par le codage $\sigma : a, b, d \mapsto 0, c \mapsto 1$ du point fixe

$$
\varphi^\omega(a) =
\begin{array}{cc|cc|cc|c}
\vdots & & & & & & \reflectbox{$\ddots$} \\
b & b & b & b & b & b & a \\
\hline
b & b & b & b & b & a & d \\
b & b & b & b & a & \mathbf{c} & \mathbf{c} \\
\hline
b & b & b & a & d & d & d \\
b & b & a & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} \\
\hline
b & a & d & d & d & d & d \\
a & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} \cdots
\end{array}
$$

du morphisme "shape-symmetric"

$$
\varphi : a \mapsto \begin{array}{|c|c|}\hline b & a \\\hline a & c \\\hline\end{array}, \quad b \mapsto \begin{array}{|c|c|}\hline b & b \\\hline b & b \\\hline\end{array}, \quad c \mapsto \begin{array}{|c|c|}\hline d & d \\\hline c & c \\\hline\end{array}, \quad d \mapsto \begin{array}{|c|c|}\hline d & d \\\hline c & c \\\hline\end{array}
$$

peut également s'écrire $X = \{(2m + k + 1, 2m) : m, k \in \mathbb{N}\}$, ou encore $X = Y_1 \cap Y_2$ avec $Y_1 = \{(n_2 + n_1 + 1)\mathbf{e_1} + n_2\mathbf{e_2} : n_1, n_2 \in \mathbb{N}\}$ et $Y_2 = \{n_1\mathbf{e_1} + 2n_2\mathbf{e_2} : n_1, n_2 \in \mathbb{N}\}$. Nous construisons alors un automate montrant que les ensembles du type Y_j sont reconnaissables pour tout système de numération abstrait.

En conclusion, cette thèse approfondit les liens entre ensembles d'entiers, mots morphiques, langages réguliers et procédures de décision, tant pour les systèmes en base entière (dans le cas des ensembles auto-générés et de la complexité syntaxique) que pour les numération abstraites. Elle laisse également entrevoir plusieurs pistes à explorer. En particulier, dans la continuité du Chapitre 3, on peut envisager de poursuivre l'étude de la complexité syntaxique dans un cadre plus général (systèmes de Bertrand, numérations linéaires, numérations abstraites, ...), d'étudier les propriétés algébriques du monoïde syntaxique associé, de tenter d'obtenir une procédure effective de décision, ...

Introduction (english version)

This thesis is mainly dedicated to the study of the recognizability of integer sets by finite automata. Then we will be interested in representations of integers in different numeration systems.

First, positional numeration systems, such as the classical decimal system, the binary system or the Zeckendorf numeration system, are well studied. Let us cite for example [Fra85, Lot02, BBLT06]. In such numeration systems, a non-negative integer n can be uniquely decomposed by Euclid's algorithm, also called the greedy algorithm, as a linear combination

$$\sum_{i=0}^{N} d_i U_{N-i}, \quad d_0, \ldots, d_N \in \mathbb{N}, \quad d_0 \neq 0$$

of elements of an increasing sequence $U = (U_m)_{m \geq 0}$ of integers such that $U_0 = 1$. Then the representation of n is $d_0 \cdots d_N$. To ensure that the alphabet of the digits d_i is finite, we suppose that the quotient $\frac{U_{m+1}}{U_m}$ is bounded. If the sequence U is the sequence $(k^m)_{m \geq 0}$, we obtain the classical integer base numeration systems.

For the Zeckendorf numeration [Zec72], the sequence U is the Fibonacci sequence defined by

$$U_0 = 1, \quad U_1 = 2, \quad U_{n+2} = U_{n+1} + U_n, \quad \forall n \geq 0.$$

In this case, questions, such as division criteria, that were trivial in the case of integer bases, need more elaborate developments. For example, in the Zeckendorf numeration, the greedy algorithm imposes that the greedy representations are the words over $\{0, 1\}$ not beginning with 0 and not containing two successive 1's. This could be useful in coding theory, for instance.

When we consider a numeration system, representing numbers allows us to pass from integers to words over finite alphabets and then, to transpose some questions from arithmetic to problems in formal language theory. In this context, regular languages, i.e., languages accepted by the "simplest machines of Chomsky's hierarchy", the finite automata, are of particular

interest. Finite automata are abstract machines only able to read words sequentially and to decide, in linear time according to the input, whether the word is accepted or not. They are also interesting from the point of view of decidability, since many of their properties can be tested algorithmically: universality, acceptance of an empty or finite set, equality or inclusion of languages, ... Whereas these problems are generally undecidable for context-free languages.

So if we want to take regular languages into account when we are interested in sets of numbers, it is natural to introduce the following notion. An integer set is said to be *recognizable* for a given numeration system if the set of representations of its elements in this system is regular. In particular, if the set of representations of the elements of an integer set in a positional numeration system U is regular, this set is said to be *U-recognizable*. The recognizable sets can be studied from different points of view. Indeed, if the sequence U satisfies a linear recurrence whose characteristic polynomial is the minimal polynomial of a Pisot number, a generalisation of Büchi's theorem shows that the U-recognizable sets are also the sets that are definable in the first order logic $\langle \mathbb{N}, +, V_U \rangle$ where $V_U(0) = U_0 = 1$ and where, for all positive integers n, $V_U(n)$ is the smallest U_i appearing in the greedy decomposition of n with a non-zero coefficient. The U-recognizable sets are also the sets whose characteristic sequence is generated by a particular type of substitution [BH97].

In particular, in the integer base case, A. Cobham proved that a set is recognizable in base k if and only if its characteristic sequence is the image under a coding of a fixed point of a k-uniform morphism [Cob72]. He also proved that the recognizability of an integer set depends on the base of the considered numeration and that the only integer sets that are recognizable for all integer bases are the ultimately periodic sets [Cob69]. This result is known as *Cobham's first theorem* and gave rise to a lot of works whose aim is to extend this result to some other numeration systems and other contexts; let us cite for example [Sem77, Vil92a, Vil92b, BHMV94, Fab94, MV96, Bès97, PB97, Dur98, Han98, Bès00, Dur02a, Dur02b, RW06, BB07, AB08, Dur08, BB09, AB11, Dur11].

If we want to compute effectively in a numeration system, the first operation realisable by an automaton that one is entitled to expect is to test whether a given word is a valid representation of an integer or not. Then it is also interesting to find some necessary or sufficient conditions under which the whole set $\mathbb{N}$ of non-negative integers is recognizable. Indeed, in this case we can test in linear time, thanks to an automaton, if a given word is a valid representation or not.

All positional numeration systems mentioned above are *monotone*, i.e., the application mapping an integer to its representation is increasing for the natural order over $\mathbb{N}$ and for the genealogical order induced by the order on the digits. For these reasons, P. Lecomte and M. Rigo [LR01] introduced the *abstract numeration systems* $S = (L, \Sigma, <)$ composed of a regular language L over the finite alphabet Σ and a total order over Σ. In these numeration systems, an integer n is represented by the $(n + 1)$-th word of the language L. The notion of recognizability can be extended to the case of abstract numeration systems as well as Cobham's first theorem. Indeed, P. Lecomte and M. Rigo proved that the ultimately periodic sets are always S-recognizable, suggesting a possible generalization of Cobham's first theorem. Then, M. Rigo and A. Maes [RM02] generalized Cobham's second theorem by proving that a set is S-recognizable if and only if its characteristic sequence is morphic. A generalization of Cobham's first theorem can be obtain thanks to this result and F. Durand's works [Dur98, Dur02a, Dur02b, Dur08, Dur11].

Recognizable sets can be naturally defined in the multidimensional case. O. Salon [Sal87] proved that the subsets of $\mathbb{N}^d$ that are k-recognizable for an integer base k are those whose characteristic sequence is generated by a k-uniform morphism. É. Charlier *et al.* [CKR10] generalized this result to the abstract numeration systems, proving that the subsets of $\mathbb{N}^d$ that are S-recognizable for an abstract numeration system S are those whose characteristic sequence is generated by a multidimensional morphism satisfying a property of "shape-symmetry" introduced by A. Maes [Mae99].

This thesis is part of the works about recognizability carried out in Liège for fifteen years. We develop results obtained in [KLR09], [KLR10], [LRRV12], [LR13] and [CLR12] as follows.

In the first chapter, we recall basic notions that are useful for the comprehension of this dissertation. First, we recall some definitions and some useful results about words, languages and automata. Then, we introduce positional numeration systems, abstract numeration systems and notions of automatic and regular sequences. Finally, we introduce the first definitions about multidimensional morphisms.

In the second chapter, we study the recognizability of *self-generating sets* that were initially introduced by C. Kimberling. These sets are defined by a set I of integers and a finite set F of affine maps $\varphi_i : n \mapsto k_i n + \ell_i$. For example, if we consider $I = \{1\}$ and $F = \{\varphi_0 : n \mapsto n, \varphi_1 : n \mapsto 2n, \varphi_2 : n \mapsto 4n - 1\}$, we obtain the self-generating set $\mathcal{K}_1 := F^\omega(I) = \{1, 2, 3, 4, 6, 7, 8, 11, 12, 14, \ldots\}$. We prove that, from a technical point of view, for the study of recognizability, we only need to consider sets of non-

negative integers and consider $k_i \geq 2$ and $\ell_i \geq 0$ for all i. We prove that if the multiplicative constants k_i are multiplicatively dependent, i.e., if all the k_i are powers of the same integer k, then the self-generating set $F^\omega(I)$ is recognizable for the integer base k numeration system. With our techniques, we reprove a result initially obtained by C. Kimberling [Kim00] and reproved by J.P. Allouche *et al.* [ASS05], giving a strong link between the set $\mathcal{K}_1$ defined above and the Fibonacci sequence defined as the fixed point $\varphi^\omega(0) = 01001010\cdots$ of the morphism $\varphi : 0 \mapsto 01, 1 \mapsto 0$. Finally, answering a conjecture of J.P. Allouche *et al.* [ASS05], we give a sufficient condition on the multiplicative constants k_i under which the self-generated set $F^\omega(I)$ is never recognizable for an integer base numeration system.

Recall that Cobham's first theorem shows that the only sets that are recognizable in multiplicatively independent bases are the ultimately periodic sets. In this context, the following decision problem is natural: *given an abstract numeration system S and a deterministic finite automaton accepting the representations of elements of a set X in S, decide whether the set X is ultimately periodic or not.* Given Cobham's second theorem and its generalization by M. Rigo and A. Maes, this problem can be rewritten in terms of morphic words: *given a morphism and a coding generating an infinite word, decide whether this word is ultimately periodic or not.* This decision problem has received lots of attention, first for integer bases since 1986 [HL86, Hon86, Pan86] and recently, V. Marsault and J.Sakarovitch [MS13] gave a $O(n \log n)$ decision procedure in the case of integer bases and the general case has been positively answered independently by F. Durand [Dur11, Dur12] and I. Mitrofanov [Mit11]. Then, in the third chapter, we explore a new approach to give an alternative decision procedure for this problem. This way may eventually lead to efficient decision procedures, so we study the *syntactic complexity* of any (ultimately) periodic set, i.e., the number of equivalence classes of the syntactic congruence of the language of representations of the elements of these sets. The syntactic congruence of a language L over Σ is defined by

$$u \leftrightarrow_L v \quad \Leftrightarrow \quad (\forall x, y \in \Sigma^*, xuy \in L \Leftrightarrow xvy \in L).$$

We first give an exact formula for the syntactic complexity of the set of integer base representations of the elements of a periodic set of period m in three particular cases: when the period m is coprime with the base, when m is a power of the base and when m can be written as a power of the base times an integer coprime with the base. Next, we give a lower bound for the syntactic complexity of the sets of integer base representations of the elements of a periodic set of integers in the general case. Finally, thanks to

this lower bound on the syntactic complexity, we obtain upper bounds for the possible preperiod and period of the set that is considered in the decision problem. Then the decision procedure consists in testing the possible pairs "preperiod-period" that are finite in number.

In first chapters, regular languages play a dominant role since a set is recognizable for a given numeration system if the language of its representations in this system is regular. Automaticity of a language is a measure of how close a non-regular language is to being regular. This notion has been introduced by B. Trakhtenbrot [Tra64] and gave rise to several works whose aim is to give bounds on the automaticity of some non-regular languages. For example, J. Shallit [Sha96] gave a lower bound on the automaticity of prime numbers represented in any integer base numeration system. In the fourth chapter, we study the automaticity of three particular languages: the language of primitive words over a finite alphabet, i.e., words that are not powers of a smaller word, the language of unbordered words over a finite alphabet, i.e., words that have no non-trivial period, and the language of representations of monic irreducible polynomials over a finite field. It can be proved that a lower bound on the automaticity of a language is given by the size of a set of dissimilar words for this language, where two words are said to be similar for a language L if, when we add the same suffix to them, either they both belong to L or they both do not belong to L. For the three considered languages, the method consists of building a set of dissimilar words for these languages and deducing a lower bound on their automaticity.

In the fifth and last chapter, answering a question of J. Sakarovitch, we characterize the multidimensional sets that are recognizable in all abstract numeration systems by proving that the subsets of $\mathbb{N}^d$ that are recognizable in all abstract numeration systems are exactly the 1-recognizable sets, i.e., the sets that are recognizable in the abstract numeration system built on the language a^*. We show for example that a 1-recognizable set such as $X = \{(n, 2m) : n, m \in \mathbb{N} \text{ and } n > 2m\}$ whose unary representations are given by $((a, a)^2)^*(a, \#)^*(a, \#)$ and whose characteristic sequence is the image under the coding $\sigma : a, b, d \mapsto 0, c \mapsto 1$ of the fixed point

$$
\varphi^\omega(a) =
\begin{array}{ccccccc}
\vdots & & & & & & \iddots \\
b & b & b & b & b & b & a \\
\hline
b & b & b & b & b & a & d \\
b & b & b & b & a & \mathbf{c} & \mathbf{c} \\
\hline
b & b & b & a & d & d & d \\
b & b & a & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} \\
\hline
b & a & d & d & d & d & d \\
a & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} & \mathbf{c} \quad \cdots
\end{array}
$$

of the shape-symmetric morphism

$$\varphi : a \mapsto \begin{array}{|c|c|} \hline b & a \\ \hline a & c \\ \hline \end{array}, \quad b \mapsto \begin{array}{|c|c|} \hline b & b \\ \hline b & b \\ \hline \end{array}, \quad c \mapsto \begin{array}{|c|c|} \hline d & d \\ \hline c & c \\ \hline \end{array}, \quad d \mapsto \begin{array}{|c|c|} \hline d & d \\ \hline c & c \\ \hline \end{array}$$

can also be written $X = \{(2m + k + 1, 2m) : m, k \in \mathbb{N}\}$, or $X = Y_1 \cap Y_2$ with $Y_1 = \{(n_2 + n_1 + 1)\mathbf{e_1} + n_2\mathbf{e_2} : n_1, n_2 \in \mathbb{N}\}$ and $Y_2 = \{n_1\mathbf{e_1} + 2n_2\mathbf{e_2} : n_1, n_2 \in \mathbb{N}\}$. Then we build an automaton showing that the sets of type Y_j are recognizable in all abstract numeration systems.

In conclusion, this thesis deepens the links between sets of numbers, morphic words, regular languages and decision procedures, for integer base systems (in the case of self-generating sets and syntactic complexity) as well as for abstract numeration systems. It also suggests several trails to explore. In particular, in line with Chapter 3, we can continue the study of the syntactic complexity in a more general case (Bertrand numeration systems, linear numerations, abstract numerations, ...), study the algebraic property of the associated syntactic monoid, obtain an effective decision procedure, ...

Chapter 1

Basics

In this chapter we introduce the basic notions that are needed in this thesis.

First, we recall some usual definitions and results about formal languages and automata theory. The interested reader can find more details in [Eil74, Sak03].

Next, we give definitions of positional numeration systems, as in particular the well-known integer base k numeration systems and we define abstract numeration systems initially introduced by Lecomte and Rigo [LR01]. We give some basic results about these two types of numeration systems. See for instance [Lot02] for details about positional numeration systems.

We also define automatic and regular sequences. These notions are linked to the numeration systems mentioned above. More details can be found in [AS03].

Finally, we extend the notion of morphisms to the multidimensional case. This notion will be useful for our last chapter about recognizability of sets of $\mathbb{N}^d$.

1.1   Words and languages

Definition 1.1.1. An *alphabet* is a non-empty finite set. The elements of an alphabet are called *letters*. A *word* over an alphabet Σ is a (finite or infinite) sequence of letters of Σ. The *empty word*, denoted by ε, is the empty sequence. The *length* of a finite word w, denoted by $|w|$, is the number of letters making up w. If w is a non-empty finite (resp. infinite) word, then for any $n \in \{0, \ldots, |w| - 1\}$ (resp. $n \in \mathbb{N}$), the $(n+1)$-th letter of w is denoted by w_n. If $w = w_0 \cdots w_r$, the *mirror* of w is the word $\tilde{w} = w_r \cdots w_0$.

Example 1.1.2. Let $\Sigma = \{a, b\}$ be an alphabet. The finite word $w = abba$

over Σ is of length $|w| = 4$ and its 4th letter is $w_3 = a$. Consider now an infinite word $v = ababbaab\cdots$. The 6th letter of v is $v_5 = a$.

Definition 1.1.3. The set of finite (resp. infinite) words over an alphabet Σ is denoted by Σ^* (resp. Σ^ω). For a unary alphabet $\{a\}$, we usually write a^* instead of $\{a\}^*$. A *language* (resp. ω-language) over an alphabet Σ is a subset of Σ^* (resp. Σ^ω).

Definition 1.1.4. If u and v are two finite words over an alphabet Σ, then the *concatenation* of u and v, denoted by $u \cdot v$ (or simply uv), is the finite word w satisfying $w_n = u_n$ for all $n \in \{0, \ldots, |u| - 1\}$ and $w_n = v_{n-|u|}$ for all $n \in \{|u|, \ldots, |u| + |v| - 1\}$. For a finite word u over an alphabet Σ and a non-negative integer n, we let u^n denote the concatenation of n copies of u, which is defined inductively by $u^0 = \varepsilon$ and $u^{n+1} = u^n u$ for all $n \in \mathbb{N}$.

Example 1.1.5. The concatenation of the words *lady* and *bug* gives the word *ladybug*.

Definition 1.1.6. Let u be a finite word over an alphabet Σ and v be an infinite word over Σ. The *concatenation* of u and v, denoted by $u\cdot v$ (or simply uv), is the infinite word w satisfying $w_n = u_n$ for all $n \in \{0, \ldots, |u| - 1\}$ and $w_n = v_{n-|u|}$ for all $n \geq |u|$.

Definition 1.1.7. Let L and M be two languages. The *concatenation of L and M* is the language $LM = \{uv \mid u \in L, v \in M\}$. For all $n \in \mathbb{N}$, we let L^n denote the concatenation of n copies of L, which is defined by

$$L^0 = \{\varepsilon\}$$

and $\forall n \in \mathbb{N}\backslash\{0\}$,

$$L^n = \{u^{(1)} \cdots u^{(n)} \mid \forall i \in \{1, \ldots, n\}, u^{(i)} \in L\}.$$

For all $n \in \mathbb{N}$, we define $L^{\leq n} = \bigcup_{i=0}^{n} L^i$. The *Kleene closure* of L is the language $L^* = \bigcup_{n \geq 0} L^n$. For a language $L = \{w\}$ containing only one element, we usually write w^* instead of $\{w\}^*$.

Note that, equipped with the concatenation of words, Σ^* is the free monoid generated by Σ having ε as neutral element. We can thus define morphisms from Σ^* to Δ^* for two alphabets Σ and Δ.

Definition 1.1.8. Let Σ and Δ be two alphabets. A *morphism* is a map $\varphi : \Sigma^* \to \Delta^*$ satisfying $\varphi(uv) = \varphi(u)\varphi(v)$ for all $u, v \in \Sigma^*$. If $\Sigma = \Delta$, we say that φ is a *morphism on* Σ. If there is a constant k such that $|\varphi(a)| = k$ for all $a \in \Sigma$, then we say that φ is a *k-uniform* morphism. A 1-uniform morphism is called a *coding*. If $\varphi(a) \neq \varepsilon$ for all $a \in \Sigma$, then φ is *nonerasing*. If $\varphi(a) = \varepsilon$ for all $a \in \Sigma$, the we say that φ is *trivial*. If there exists an integer $j \geq 1$ such that $\varphi^j(a) = \varepsilon$, then the letter a is said to be *mortal*. The set of mortal letters associated with a morphism φ is denoted by M_φ. A finite or infinite word w such that $\varphi(w) = w$ is said to be a *fixed point* of φ. If there exists a letter $a \in \Sigma$ such that $\varphi(a) = ax$, and $x \notin M_\varphi^*$, we say φ is *prolongable* on a. In this case, the sequence of words $a, \varphi(a), \varphi^2(a), \ldots$ converges[1], in the limit, to the infinite word

$$\varphi^\omega(a) := \lim_{n \to +\infty} \varphi^n(a) = ax\varphi(x)\varphi^2(x) \cdots ,$$

which is a fixed point of φ, that is, $\varphi(\varphi^\omega(a)) = \varphi^\omega(a)$. Furthermore, it is easy to see that $\varphi^\omega(a)$ is the unique fixed point of φ which starts with a. If $w = \varphi^\omega(a)$, then we call w a *pure morphic* sequence . If there is a coding $\tau : \Sigma \to \Delta$ and $w = \tau(\varphi^\omega(a))$, then we call w a *morphic* sequence.

Example 1.1.9. Let $\Sigma = \Delta = \{0, 1\}$. The Thue-Morse morphism is defined as follows

$$\mu : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 10. \end{cases}$$

Then we have $\mu(0) = 01$, $\mu^2(0) = 0110$, $\mu^3(0) = 01101001, \ldots$ and the Thue-Morse word is the fixed point of the morphism μ starting with 0,

$$\mu^\omega(0) = 0110100110010110 \cdots .$$

The next theorem can be found in [AS03, Theorem 7.7.5] and has been initially proved by Cobham [Cob68].

Theorem 1.1.10. *The image of a morphic sequence by any morphism is either finite or morphic.*

Definition 1.1.11. Let w be a word over an alphabet Σ. A *factor* of w is a finite word u such that there exist $x \in \Sigma^*$ and $y \in \Sigma^* \cup \Sigma^\omega$ satisfying $w = xuy$. For any non-negative integers m and n satisfying $m \leq n$, we let $w[m, n]$ denote the factor $w_m \cdots w_n$ of w. For any $n \in \mathbb{N}$, the *prefix*

---

[1]for the product topology of the discrete topology over the alphabet

of length n of w is the factor $w[0, n-1]$, where, by convention, if $m > n$ we set $w[m, n] = \varepsilon$. The *suffix of length n* of a finite word w is the factor $w[|w| - n + 1, |w|]$. We let $\mathrm{Pref}(w)$ denote the set of all prefixes of w:

$$\mathrm{Pref}(w) = \{x \in \Sigma^* \mid \exists y \in \Sigma^* \cup \Sigma^\omega, w = xy\}$$

and $\mathrm{Suff}(w)$ denotes the set of all suffixes of w.

Note that we have $w_n = w[n, n]$ for any non-empty word w and any non-negative integer n.

Definition 1.1.12. Let u be a finite word over an alphabet Σ. We let u^ω denote the concatenation of infinitely many copies of u, which is defined by $u^\omega[n|u|, (n+1)|u| - 1] = u$ for all $n \in \mathbb{N}$.

Definition 1.1.13. The *prefix-closure* of a language L over an alphabet Σ, which is denoted by $Pref(L)$, is the language of the prefixes of its words:

$$Pref(L) = \{x \in \Sigma^* \mid \exists y \in \Sigma^*, xy \in L\}.$$

A language L is *prefix-closed* if it satisfies $L = Pref(L)$.

Example 1.1.14. Let us consider the word $w = abbaba$ over the alphabet $\Sigma = \{a, b\}$. The prefix of length 3 of w is *abb* and its suffix of length 3 is *aba*. The set of its factors of length 3 is $\{abb, bba, bab, aba\}$ and we have $Pref(w) = \{\varepsilon, a, ab, abb, abba, abbab, abbaba\}$.

Definition 1.1.15. Let $(\Sigma, <)$ be a totally ordered alphabet. The order $<$ on Σ extends to an order on Σ^ω, called the *lexicographical order*, as follows. If u and v are two distinct infinite words over Σ, then u is said to be *lexicographically less* than v and we write $u <_{lex} v$, if there exist $p \in \Sigma^*$, $s, t \in \Sigma^\omega$, and $a, b \in \Sigma$ such that we have $u = pas$, $v = pbt$, and $a < b$. This order extends to $\Sigma^\omega \cup \Sigma^*$ by replacing finite words z over Σ by $z\#^\omega \in (\Sigma \cup \{\#\})^\omega$, where $\#$ is a letter not belonging to the alphabet Σ which is assumed to satisfy $\# < a$ for all $a \in \Sigma$. We write $u \leq_{lex} v$ for two words u and v satisfying either $u <_{lex} v$ or $u = v$.

Definition 1.1.16. Let $(\Sigma, <)$ be a totally ordered alphabet. The order $<$ on Σ extends to a total order on Σ^*, called the *genealogical order* (or *radix order*), as follows. If u and v are two finite words over Σ, then u is said to be *genealogically less* than v, and we write $u <_{gen} v$, if they satisfy either

$|u| = |v|$ and $u <_{lex} v$ or $|u| < |v|$. We write $u \leq_{gen} v$ for two words u and v satisfying either $u <_{gen} v$ or $u = v$.

Example 1.1.17. Consider the alphabet $\Sigma = \{a, b, c\}$ with $a < b < c$. In the lexicographic order, the first words of Σ^* are $\varepsilon, a, a^2, a^3, \ldots$. In the genealogical order, the first words of Σ^* are $\varepsilon, a, b, c, aa, ab, ac, ba, \ldots$.

1.2 Automata and regular languages

Definition 1.2.1. A *deterministic finite automaton* (or DFA) is a 5-tuple

$$\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$$

where

- Q is a non-empty finite set, called the set of *states*;

- Σ is an alphabet;

- $\delta : Q \times \Sigma \to Q$ is the (possibly partial) *transition function*;

- q_0 is a distinguished element of Q, called the *initial state*;

- $F \subseteq Q$ is the set of *final states*.

The transition function δ naturally extends to a function on $Q \times \Sigma^*$ as follows:

$$\delta(q, \varepsilon) = q$$

and

$$\delta(q, aw) = \delta(\delta(q, a), w), \quad a \in \Sigma, w \in \Sigma^*.$$

If the function δ is total, the automaton is said to be *complete*.

If the set of states is infinite, we said that $\mathcal{A}$ is a *deterministic infinite automaton*.

Remark 1.2.2. If the context is clear, we write $q.w$ instead of $\delta(q, w)$.

Definition 1.2.3. The language *accepted* (or *recognized*) by the DFA $\mathcal{A}$ is

$$L(\mathcal{A}) = \{w \in \Sigma^* | \delta(q_0, w) \in F\}.$$

If $w \in L(\mathcal{A})$, we say that $\mathcal{A}$ *accepts* (or *recognizes*) the word w (or that w is *accepted* by $\mathcal{A}$).

Automata can be represented by oriented graphs. The vertices are the states of the automaton $\mathcal{A}$ and there is an edge labelled by $a \in \Sigma$ between two vertices q and q' if the transition function δ of the automaton is such that $\delta(q, a) = q'$. The initial state is distinguished by an incoming arrow and final states by an outgoing arrow.

Example 1.2.4. Consider the DFA $\mathcal{A} = (\{I, II, III\}, \{0, 1\}, \delta, \{I\}, \{II\})$ such that the transition function δ is given by the following table.

δ	0	1
I	II	I
II	II	III
III	III	III

This automaton is represented in Figure 1.1 and accepts the language 1^*0^*0.
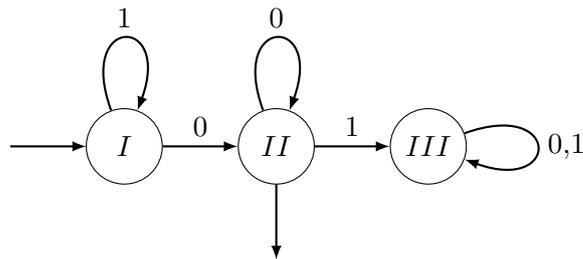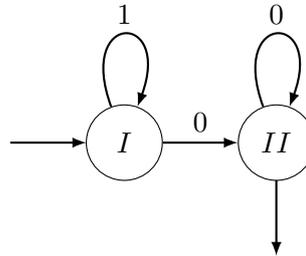

Figure 1.1: DFA accepting 1^*0^*0.

Definition 1.2.5. Let $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$ be a deterministic automaton. A state q in Q is *accessible* if it can be reached from the initial state, i.e., if there exists a word $w \in \Sigma^*$ such that we have $\delta(q_0, w) = q$ and is *coaccessible* if one can reach a final state from it, i.e., if there exists $w \in \Sigma^*$ such that we have $\delta(q, w) \in F$. The automaton $\mathcal{A}$ is *accessible* (resp. *coaccessible*) if all its states are accessible (resp. coaccessible) and is *trim* if it is both accessible and coaccessible.

Example 1.2.6. The automaton depicted in Figure 1.2 is a trim DFA recognizing the same language as the DFA of Figure 1.1 . Note that its transition function is partial : for instance, $\delta(II, 1)$ is not defined.

Definition 1.2.7. A *non-deterministic finite automaton* (or NDFA) is a

Figure 1.2: Trim DFA accepting 1^*0^*0.

5-tuple

$$\mathcal{A} = (Q, I, F, \Sigma, \Delta)$$

where

- Q is a non-empty finite set, called the set of *states*;

- $I \subseteq Q$ is the set of *initial states*;

- $F \subseteq Q$ is the set of *final states*;

- Σ is the alphabet of the automaton;

- $\Delta \subset Q \times \Sigma^* \times Q$ is a *transition relation*.

Definition 1.2.8. A word w is accepted by a NDFA $\mathcal{A}$ if there exists $q_0 \in I$, $l \in \mathbb{N} \backslash \{0\}$, $v_1, \ldots, v_l \in \Sigma^*$, $q_1, \ldots, q_l \in Q$ such that

$$(q_0, v_1, q_1), (q_1, v_2, q_2), \ldots, (q_{l-1}, v_l, q_l) \in \Delta,$$

$$w = v_1 \cdots v_l \text{ and } q_l \in F,$$

i.e., there exists a path in the graph associated to $\mathcal{A}$ beginning in an initial state and ending in a final state. The language *accepted* by a NDFA $\mathcal{A}$ is the set of words accepted by $\mathcal{A}$ and is denoted by $L(\mathcal{A})$.

Example 1.2.9. The NDFA $\mathcal{A} = (\{I, II\}, \{I\}, \{II\}, \{0, 1\}, \Delta)$ where

$$\Delta = \{(I, 0, II), (I, 1, I), (I, 0, II), (II, 0, II)\} \subset \{I, II\} \times \{0, 1\} \times \{I, II\}$$

is represented by the oriented graph of Figure 1.3. This automaton accepts the language $\{0, 1\}^*0$.
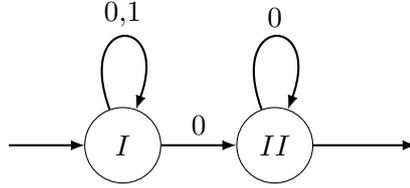
Figure 1.3: NDFA accepting $\{0,1\}^*0$.

The following result is well known.

Proposition 1.2.10 ([RS59])**.** *Every language accepted by a NDFA is accepted by a DFA.*

Definition 1.2.11. Let $L \subseteq \Sigma^*$ be a language. If w is a word over Σ, we denote by $w^{-1}.L$ the set of words which belong to L when they are concatenated with w, i.e.,

$$w^{-1}.L = \{u \in \Sigma^* | wu \in L\}.$$

We define the *Nerode congruence* on Σ^*, denoted by $\sim_L$, as follows: for all $x, y \in \Sigma^*$,

$$x \sim_L y \Leftrightarrow x^{-1}.L = y^{-1}.L.$$

Lemma 1.2.12. *Let $L \subseteq \Sigma^*$ be a language and let u, v be two words over Σ. We have*

$$(uv)^{-1}.L = v^{-1}.(u^{-1}.L).$$

Definition 1.2.13. The minimal automaton of L is the deterministic finite automaton

$$\mathcal{A}_L = (Q_L, \Sigma, \delta_L, q_{0,L}, F_L)$$

where

- $Q_L = \{w^{-1}.L | w \in \Sigma^*\}$,

- $\delta_L(q, a) = a^{-1}.q$, for all $q \in Q_L$, $a \in \Sigma$,

- $q_{0,L} = \varepsilon^{-1}.L = L$,

- $F_L = \{w^{-1}.L | w \in L\} = \{q \in Q_L | \varepsilon \in q\}$.

The *trim minimal automaton* of a language L is the minimal automaton of L from which the only possible non-coaccessible state, called the *sink state*, is removed.

Example 1.2.14. The automata depicted in Figure 1.1 and Figure 1.2 are the minimal automaton and the trim minimal automaton of the language 1*0*0.

The following result is clear.

Proposition 1.2.15. *The minimal automaton of a language $L \subseteq \Sigma^*$ accepts L.*

The term "minimal automaton" is justified by the following proposition.

Proposition 1.2.16. *Let L be a language over an alphabet Σ and let $\mathcal{A}$ be a deterministic automaton accepting L and having Q as set of states. Then we have a surjective morphism $\Phi : Q \to Q_L$ such that*

- $\Phi(F) \subseteq F_L$,

- $\Phi(q_0) = q_{0,L}$,

- $\Phi(\delta(q, w)) = \delta_L(\Phi(q), w)$.

The morphism Φ in Proposition 1.2.16 is called the *canonical morphism* of automata.

Definition 1.2.17. A language is *regular* if it is accepted by a finite automaton.

The two following results give characterizations of regular languages.

Proposition 1.2.18 ([Kle56]). *The set of regular languages over Σ is the smallest family of languages containing the empty set and the singletons and closed under union, concatenation and Kleene closure.*

Theorem 1.2.19 ([Ner58]). *A language L is regular if and only if the equivalence $\sim_L$ is of finite index (i.e., has a finite number of equivalence classes), i.e., if its minimal automaton $\mathcal{A}_L$ is finite.*

The following result is often used to prove the non-regularity of a language.

Lemma 1.2.20 (Pumping Lemma)**.** *Let $L \subseteq \Sigma^*$ be a regular language. There exists an integer l such that for all words w of L satisfying $|w| \geq l$, there exist $x, y, z \in \Sigma^*$ such that $w = xyz$ and*

- $|xy| \leq l$,

- $y \neq \varepsilon$,

- $xy^*z \subset L$.

Definition 1.2.21. A *deterministic finite automaton with output* (or DFAO) is defined to be a 6-tuple

$$\mathcal{M} = (Q, \Sigma, \delta, q_0, \Delta, \tau)$$

where Q, Σ, δ and q_0 are as in the definition of DFA, Δ is the *output alphabet*, and $\tau : Q \rightarrow \Delta$ is the *output function*. In particular, the output for a word $w \in \Sigma^*$ is $\tau(\delta(q_0, w))$.

We can represent a DFAO in the same way we did for DFA. The only difference is that a state labelled q/a indicates that the output associated with the state q is the symbol a, i.e., $\tau(q) = a$.

Example 1.2.22. Consider the DFAO defined by $Q = \{q_0, q_1\}$, $\Sigma = \Delta = \{0, 1\}$, $\delta(q_0, 0) = q_0$, $\delta(q_0, 1) = q_1$, $\delta(q_1, 0) = q_1$, $\delta(q_1, 1) = q_0$, $\tau(q_0) = 0$ and $\tau(q_1) = 1$. This deterministic finite automaton with output is depicted in Figure 1.4.
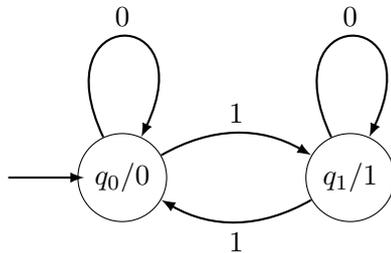

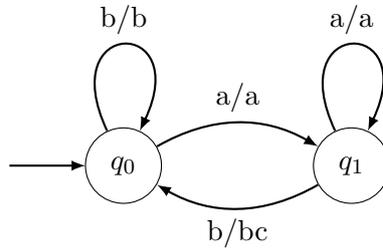Figure 1.4: A deterministic finite automaton with output.

Figure 1.5: A finite-state transducer.

Another model of automata with output can be defined as follows.

Definition 1.2.23. A *transducer* $\mathcal{T}$ (from Σ^* to Δ^*) can be viewed as an automaton whose transitions are labelled by pairs of words (elements of $\Sigma^* \times \Delta^*$). The set of labels of successful computations is a subset of $\Sigma^* \times \Delta^*$ and is called the *relation realized by* $\mathcal{T}$. A *letter-to-letter transducer* is an automaton whose transitions are labelled by elements of $\Sigma \times \Delta$. If there exists a path from the state q_0 to a finite state and labelled by $(w, w') \in \Sigma^* \times \Delta^*$, we say that w' is an *output* for the word w.

Example 1.2.24. Consider the deterministic transducer depicted in Figure 1.5. This transducer inserts a c after every occurrence of ab in the input string. For example, the output for the word *aab* is *aabc*.

A proof of the following result can be found in [AS03, Theorem 4.3.6].

Proposition 1.2.25. *If a regular language is an input of a transducer then the output language is also regular.*

Corollary 1.2.26. *Let $L \subseteq \Sigma^*$ be a regular language, and let $\varphi : \Sigma^* \to \Delta^*$ be a morphism. Then $\varphi(L) = \{\varphi(w) : w \in L\}$ is regular.*

1.3 Positional numeration systems

Let $1 = u_0 < u_1 < \cdots < u_N < \cdots$ be an increasing sequence of integers. Every positive integer n can be decomposed as $n = d_0 u_N + \cdots + d_{N-1} u_1 + d_N u_0$ where $d_0, \ldots, d_N$ are non-negative integers. The aim is to decompose any positive integer as $d_0 u_N + \cdots + d_{N-1} u_1 + d_N u_0$ with some prescribed

conditions like, for example, greediness.

Definition 1.3.1. A *positional numeration system* is a strictly increasing sequence $U = (u_m)_{m \geq 0}$ of integers such that $u_0 = 1$ and the ratio $\frac{u_{m+1}}{u_m}$ is bounded. If the sequence U satisfies a linear recurrence relation with integer coefficients, the numeration system is said to be *linear*. Let $\Sigma \subset \mathbb{Z}$ be an alphabet and $w = d_0 \cdots d_N$ be a word over Σ. We denote by $\mathrm{val}_U(w)$ the integer

$$n = d_0 u_N + \cdots + d_{N-1} u_1 + d_N u_0.$$

A *U-representation* of an integer n is a word $d_0 \cdots d_N$ such that

$$n = \sum_{i=0}^{N} d_i u_{N-i}.$$

Note that an integer can have more than one U-representation. A special representation of n can be obtained thanks to a greedy algorithm as follows. Let N be such that $u_N \leq n < u_{N+1}$ and let d_0 and r_0 be the quotient and the remainder of the Euclidean division of n by u_N. For $i = 1, \ldots, N$, we denote by d_i and r_i respectively the quotient and the remainder of the Euclidean division of r_{i-1} by u_{n-i}. As the ratio $\frac{u_{m+1}}{u_m}$ is bounded by a constant k_U, the d_i's belong to the alphabet $\Sigma_U := \{0, \ldots, k_U - 1\}$.

Definition 1.3.2. The word $d_0 \cdots d_N \in \Sigma_U^*$ obtained by the greedy algorithm is unique, with the condition that $d_0 \neq 0$, and is called the *normal U-representation* (or *greedy representation*) of n and is denoted by $\mathrm{rep}_U(n)$. The normal U-representation of 0 is $\mathrm{rep}_U(0) = \varepsilon$. The elements of Σ_U are called the *digits*. The set $\mathrm{rep}_U(\mathbb{N})$ is called the *numeration language*.

Example 1.3.3. Consider the Fibonacci sequence $F = (F_m)_{m \geq 0}$ defined by $F_0 = 1$, $F_1 = 2$ and $F_m = F_{m-1} + F_{m-2}$ for all $m \geq 2$. Hence $F = (1, 2, 3, 5, 8, 13, \ldots)$. The numeration system built on this sequence is called the *Fibonacci numeration system*. Zeckendorf [Zec72] proved that $\Sigma_F = \{0, 1\}$ and that the numeration language is the set $\mathrm{rep}_F(\mathbb{N}) = 1\{0, 01\}^* \cup \{\varepsilon\}$, i.e., the set of words over $\{0, 1\}$ not containing the factor 11 and not beginning with 0. For instance, we have $\mathrm{rep}_F(17) = 100101$ because $17 = 13 + 3 + 1$ and $\mathrm{val}_F(1010) = 5 + 2 = 7$.

Definition 1.3.4. Let $k \geq 2$ be an integer. If the sequence U is the sequence $(k^m)_{m \geq 0}$, the obtained numeration system is the *integer base k numeration system*. The normal k-representation of an integer n is denoted by $\mathrm{rep}_k(n)$

and the value of a word w over $\Sigma_k := \{0, \ldots, k-1\}$ is denoted by $\mathrm{val}_k(w)$. In this case, the numeration language is $\mathrm{rep}_k(\mathbb{N}) = \Sigma_k^* \backslash 0 \Sigma_k^*$.

Example 1.3.5. If $k = 3$, we have $\Sigma_3 = \{0, 1, 2\}$. Then $\mathrm{rep}_3(17) = 122$ because $17 = 9 + 2.3 + 2.1$ and $\mathrm{val}_3(1010) = 27 + 3 = 30$.

Definition 1.3.6. Let Σ be a finite alphabet of integers that can be positive or negative. The *numerical value* in base k on Σ^* is the function

$$\mathrm{val}_k : \Sigma^* \to \mathbb{Z}$$

that maps a word $w = c_0 \cdots c_N$ of Σ^* onto $\sum_{i=0}^{N} c_i k^{N-i}$. Note that $\mathrm{val}_k(uv) = \mathrm{val}_k(u) k^{|v|} + \mathrm{val}_k(v)$ for all $u, v \in \Sigma_k$. The *normalization* on Σ^* is the partial function

$$\nu_\Sigma : \Sigma^* \to \Sigma_k^*$$

that maps a word $w = c_0 \cdots c_N$ of Σ^*, such that $n = \mathrm{val}_k(w)$ is non-negative, onto its normal representation $\mathrm{rep}_k(n)$.

As stated by Frougny's normalization theorem [Lot02, Fro92], the normalization is computable by transducer.

Theorem 1.3.7. *In base k, for every finite alphabet $\Sigma \subset \mathbb{Z}$, the normalization on Σ^* is computable by a finite transducer.*

As stated by the next proposition, considering normal representations allows us to work with order-preserving positional numeration systems. More precisely, in this case, the natural order on the set of integers corresponds to the genealogical order on the numeration language.

Proposition 1.3.8. *Let U be a positional numeration system. For all non-negative integers m and n, we have*

$$m < n \Leftrightarrow \mathrm{rep}_U(m) <_{gen} \mathrm{rep}_U(n)$$

where the genealogical order $<_{gen}$ is induced by the natural order on the alphabet Σ_U.

Definition 1.3.9. Let U be a positional numeration system. A set X of $\mathbb{N}$ is *U-recognizable* if the language $\mathrm{rep}_U(X)$ over Σ_U is regular. In particular, a set X of $\mathbb{N}$ is *k-recognizable* if the language $\mathrm{rep}_k(X)$ is regular.

A similar definition can be given for the U-recognizable subsets of $\mathbb{Z}$ using convenient conventions to represent negative numbers, like adding a symbol "-" to the alphabet or considering the positive and negative elements separately.

Example 1.3.10. The set X of powers of 2, whose normal representations in base 2 are the words of the language 10^*, is 2-recognizable. Indeed the numeration language $\mathrm{rep}_2(X)$ is accepted by the automaton depicted in Figure 1.6.
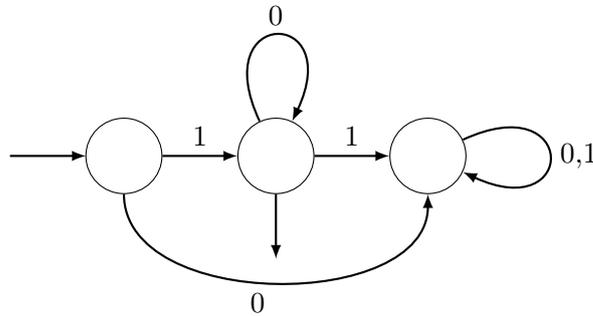
Figure 1.6: DFA accepting $\mathrm{rep}_2(X)$.

However we can prove that the set of powers of 2 is not 3-recognizable. So the recognizability of a set depends on the base. Proofs of the two following results can be found in [BHMV94].

Proposition 1.3.11. *Let $k \geq 2$ and $l \geq 1$ be integers. A set X of integers is k-recognizable if and only if X is k^l-recognizable.*

Proposition 1.3.12. *Let $k \geq 2$ and $m \geq 0$ be integers. A set $X \subseteq \mathbb{N}$ is k-recognizable if and only if $X + m$ is k-recognizable.*

Proposition 1.3.13. *Let $X \subseteq \mathbb{N}$ be a k-recognizable set. Then the set bX is k-recognizable for any integer $b \geq 0$.*

Definition 1.3.14. A set of integers X is *ultimately periodic* (or *periodic*) if there exist $a, p \in \mathbb{N}$ with $p > 0$, such that, for all $i \geq a$ we have $i \in X \Leftrightarrow i + p \in X$. If the integers a and p are minimal for the latter property, then we say that they are the *preperiod* and the *period* of X respectively.

Proposition 1.3.15. *If a set of integers X is ultimately periodic, X is k-recognizable for all integers $k \geq 2$.*

The celebrated first theorem of Cobham can be stated as follows.

Theorem 1.3.16 ([Cob69])**.** *Let k, l be two multiplicatively independent integers, i.e., $\frac{\log k}{\log l}$ is irrational. If a set of integers X is both k-recognizable and l-recognizable, then X is ultimately periodic.*

Thanks to the previous two results, it is clear that the sets that are k-recognizable for all integer base $k \geq 2$ are exactly the ultimately periodic sets.

1.4 Abstract numeration systems

Definition 1.4.1. An *abstract numeration system* (or ANS) is a triple $S = (L, \Sigma, <)$ where L is an infinite regular language over a totally ordered alphabet $(\Sigma, <)$. The map $\text{rep}_S : \mathbb{N} \to L$ is the one-to-one correspondence mapping $n \in \mathbb{N}$ onto the $(n+1)$th word of L for the genealogical order. This word is called the *S-representation* of n. The S-representation of 0 is the first word in L. The inverse map is denoted by $\text{val}_S : L \to \mathbb{N}$. If w is a word in L, $\text{val}_S(w)$ is its *S-numerical value*.

Example 1.4.2. Consider the abstract numeration system

$$S = (a^* b^*, \{a, b\}, a < b).$$

If we enumerate the words of $a^* b^*$ in the genealogical order we obtain the following correspondence between words and integers.

w	$\text{val}_S(w)$
ε	0
a	1
b	2
aa	3
ab	4
bb	5
aaa	6
$\vdots$	$\vdots$

Thanks to Proposition 1.3.8, positional numeration systems whose language of numeration is regular are special cases of abstract numeration systems. However, abstract numeration systems are not always positional numeration systems but we can define S-recognizability in a way similar to U-recognizability.

Definition 1.4.3. Let X be a set of integers. We denote by $\mathrm{rep}_S(X)$ the *set of S-representations* of the elements of X, i.e.,

$$\mathrm{rep}_S(X) = \{\mathrm{rep}_S(n) | n \in X\}.$$

A set X of integers is said to be *S-recognizable* if $\mathrm{rep}_S(X)$ is a regular language. In particular, a set of non-negative integers is *1-recognizable* if this set is S-recognizable where S is the abstract numeration system built on the language a^*. These sets will be of particular interest in the sequel .

We collect here some basic results on S-recognizable sets.

Theorem 1.4.4 ([LR01])**.** *A set $X \subseteq \mathbb{N}$ is S-recognizable for all abstract numeration systems S if and only if it is ultimately periodic.*

Theorem 1.4.5 ([Eil74])**.** *A set of non-negative integers is 1-recognizable if and only if it is a finite union of arithmetic progressions, i.e., ultimately periodic.*

Proposition 1.4.6 ([LR01])**.** *[Translation by a constant] Let $S = (L, \Sigma, <)$ be an abstract numeration system and $t \in \mathbb{N}$. If $X \subseteq \mathbb{N}$ is S-recognizable, then $X + t$ is S-recognizable.*

Theorem 1.4.7 ([Rig01])**.** *Let $L \subseteq \Sigma^*$ be a regular language such that $u_L(n) = \Theta(n^k)$ for some $k \in \mathbb{N}$, where $u_L(n)$ denotes the number of words of length n, and $S = (L, \Sigma, <)$. Preservation of S-recognizability after multiplication by a constant λ holds only if $\lambda = \beta^{k+1}$ for some $\beta \in \mathbb{N}$.*

Proposition 1.4.8 ([Rig00])**.** *The set $\mathcal{P}$ of prime numbers is never S-recognizable.*

1.5   Automatic and regular sequences

Definition 1.5.1. Let $u = (u_n)_{n \in \mathbb{N}}$ be an infinite sequence over a finite alphabet Σ and $k \geq 2$ be an integer. The sequence u is *k-automatic* if there

exists a DFAO $\mathcal{M} = (Q, \Sigma_k, \delta, q_0, \Sigma, \tau)$ such that $u_n = \tau(\delta(q_0, w))$ for all $n \geq 0$ and all w with $\mathrm{val}_k(w) = n$.

Example 1.5.2. The Thue-Morse sequence $\mu^\omega(0) = 0110100110010110 \cdots$ is 2-automatic. Indeed, the DFAO represented in Figure 1.4 of Example 1.2.22 is the automaton generating this sequence.

The *k-kernel* of a sequence u is the set

$$K_k(u) = \{u_{k^i n + j} \;:\; i \geq 0, \; 0 \leq j \leq k^i\}.$$

It is well known that the sequence u is k-automatic if and only if its k-kernel is finite [Eil74].

Definition 1.5.3. A sequence u over a ring is *k-regular* if the $\mathbb{Z}$-module generated by its k-kernel is finitely generated.

It is clear that a k-automatic sequence is necessarily k-regular.

The notion of k-automatic sequence is also related to k-recognizable sets. Recall that the *characteristic sequence* $(\mathbf{c}_X(n))_{n \geq 0}$ of a set $X \subseteq \mathbb{N}$ is defined by $\mathbf{c}_X(n) = 1$, if $n \in X$ and $\mathbf{c}_X(n) = 0$, otherwise. In particular, X is k-recognizable (resp. ultimately periodic) if and only if $(\mathbf{c}_X(n))_{n \geq 0}$ is k-automatic (resp. an ultimately periodic infinite word).

We give below some basic properties of automatic sequences. These results can be found in [AS03, Section 5.4].

Theorem 1.5.4. *If a sequence v differs only in finitely many terms from a k-automatic sequence u, then it is k-automatic.*

This result is similar to the fact that adding or deleting a finite set of words to a regular language gives again a regular language. The following result is equivalent to Proposition 1.3.15

Theorem 1.5.5. *If u is an ultimately periodic sequence, then this sequence is k-automatic for all $k \geq 2$.*

Theorem 1.5.6. *Let u be a k-automatic sequence, and let ρ be a coding. Then the sequence $\rho(u)$ is also k-automatic.*

Let $u = u_0 u_1 u_2 \cdots$ and $v = v_0 v_1 v_2 \cdots$ be two sequences with values in Σ

and Σ'. The product $u \times v$ is defined by

$$u \times v = (u, v) = (u_0, v_0)(u_1, v_1)(u_2, v_2) \cdots \in (\Sigma \times \Sigma')^{\omega}.$$

Theorem 1.5.7. *Let u and v be two k-automatic sequences with values in Σ and Σ' respectively. Then $u \times v$ is k-automatic.*

The following result is known as Cobham's second theorem.

Theorem 1.5.8 ([Cob72])**.** *Let $k \geq 2$ be an integer. An infinite sequence is k-automatic if and only if it is the image under a coding of a fixed point of k-uniform morphism.*

Allouche, Scheicher and Tichy [AST00] extended the notions of k-regular and k-automatic sequences to number systems based on linear recurrence sequences. For example, if we consider the Fibonacci sequence $F = (F_n)_{n \in \mathbb{N}}$ introduced in Example 1.3.3, we can define Fibonacci-regularity and Fibonacci-automaticity as follows. Let $\mathcal{F}$ be the set of F-representations, i.e., $\mathcal{F} = \mathrm{rep}_F(\mathbb{N})$. Shallit [Sha88] proved that this set is regular. For each word $v \in \Sigma_2^*$, consider $\mathcal{L}_v = \{u \in \Sigma_2^* \; : \; uv \in \mathcal{F}\}$. As $\mathcal{F}$ is regular there is a finite number of $\mathcal{L}_v$ and it is easy to prove that for each v, $\mathcal{L}_v$ is either empty or infinite. For each v such that $\mathcal{L}_v \neq \emptyset$, let $i_v(n)$ be the sequence such that $\{(i_v(n)) : n \geq 0\} = \mathrm{val}_F(\mathcal{L}_v)$ where the elements of $\mathcal{L}_v$ are sorted in increasing order and $i_\varepsilon(n) = 0$.

Example 1.5.9. Consider the first F-representations of integers given in the following table.

$\mathbb{N}$	$\mathrm{rep}_F(\mathbb{N})$
0	ε
1	1
2	10
3	100
4	101
5	1000
6	1001
7	1010
8	10000
9	10001
10	10010
$\vdots$	$\vdots$

Then to each suffix corresponds a subsequence and we have for example

$$\{i_0(n) \ : \ n \in \mathbb{N}\} = \{2, 3, 5, 7, 8, 10, \ldots\}$$

and

$$\{i_{10}(n) \ : \ n \in \mathbb{N}\} = \{7, 10, \ldots\}.$$

Definition 1.5.10. Let $(u_n)_{n \in \mathbb{N}}$ be a sequence. The *Fibonacci-kernel* of u is the set

$$K_F(u) = \{u_{i_v(n)} : v \in \Sigma_2^*\}.$$

A sequence u is *Fibonacci-regular* if the module generated by its Fibonacci-kernel is finitely generated.

A sequence u is *Fibonacci-automatic* if u_n is a finite state function of the Fibonacci-representations of n.

1.6 Multidimensional morphisms

In this section, we generalize words and morphisms to the case of multidimensional sequences. These definitions will be useful in Chapter 5 which deals with multidimensional sets recognizable in abstract numeration systems. For $i \leq j$, we denote by $[\![i, j]\!]$ the interval of integers $\{i, i+1, \ldots, j-1, j\}$.

Definition 1.6.1. Let $s_1, \ldots, s_d$ be positive integers or ∞. A *d-dimensional array* over the alphabet Σ is a map x with domain $[\![0, s_1 - 1]\!] \times \cdots [\![0, s_d - 1]\!]$ taking values in Σ. By convention, if we have $s_i = \infty$ for some i, then we set $[\![0, s_i - 1]\!] = \mathbb{N}$. If x is such an array, we write $|x|$ for the d-tuple $(s_1, \ldots, s_d) \in (\mathbb{N} \cup \infty)^d$, which is called the *shape* of x. We let ε_d denote the d-dimensional array of shape 0. Note that we have $\varepsilon_1 = \varepsilon$. A d-dimensional array x is said to be *bounded* if we have $|x|_i < \infty$ for all $i \in [\![1, d]\!]$. The set of d-dimensional bounded arrays over Σ is denoted by $B_d(\Sigma)$. A bounded array x is a *square* of size $c \in \mathbb{N}$ if $|x| = (c, \ldots, c)$.

Let x be a d-dimensional array. If we have $\mathbf{0} \leq \mathbf{s} \leq \mathbf{t} \leq |x| - \mathbf{1}$, where $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$, then $x[\mathbf{s}, \mathbf{t}]$ is said to be a *factor* of x and is defined as the array y of shape $\mathbf{t} - \mathbf{s} + \mathbf{1}$ given by $y(\mathbf{n}) = x(\mathbf{n} + \mathbf{s})$ for all $\mathbf{n} \in \mathbb{N}^d$ such that $\mathbf{n} \leq \mathbf{t} - \mathbf{s}$. For any $\mathbf{v} \in \mathbb{N}^d$, the set of factors of x of shape $\mathbf{v}$ is denoted by $\mathrm{Fact}_{\mathbf{v}}(x)$.

Example 1.6.2. Consider the bidimensional bounded array of shape $(2, 5)$,

$$x = \begin{array}{|c|c|c|c|c|} \hline a & b & a & a & b \\ \hline c & d & b & c & d \\ \hline \end{array}.$$

We have

$$x[(0,0)(1,1)] = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array} \quad \text{and} \quad x[(0,2)(1,4)] = \begin{array}{|c|c|c|} \hline a & a & b \\ \hline b & c & d \\ \hline \end{array}$$

We have also $\mathrm{Fact}_{\mathbf{1}}(x) = \{a,b,c,d\}$ and

$$\mathrm{Fact}_{(2,3)}(x) = \left\{ \begin{array}{|c|c|c|} \hline a & b & a \\ \hline c & d & b \\ \hline \end{array}, \begin{array}{|c|c|c|} \hline b & a & a \\ \hline d & b & c \\ \hline \end{array}, \begin{array}{|c|c|c|} \hline a & a & b \\ \hline b & c & d \\ \hline \end{array} \right\}$$

Definition 1.6.3. Let x be a d-dimensional array of shape $\mathbf{s} = (s_1, \ldots, s_d)$. For all $i \in [\![1,d]\!]$ and $k < s_i$, we let $x_{|i,k}$ denote the $(d-1)$-dimensional array of shape

$$|x|_{\hat{i}} = \mathbf{s}_{\hat{i}} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_d)$$

defined by setting the ith coordinate equal to k in x, that is,

$$x_{|i,k}(n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_d) = x(n_1, \ldots, n_{i-1}, k, n_{i+1}, \ldots, n_d)$$

for all $n_j \in [\![0, s_j - 1]\!]$ with $j \in [\![1,d]\!] \setminus \{i\}$.

Definition 1.6.4. Let x, y be two d-dimensional arrays. If for some $i \in [\![1,d]\!]$, $|x|_{\hat{i}} = |y|_{\hat{i}} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_d)$, then we define the concatenation of x and y in the direction i to be the d-dimensional array $x \odot^i y$ of shape

$$(s_1, \ldots, s_{i-1}, |x|_i + |y|_i, s_{i+1}, \ldots, s_d)$$

satisfying

(i) $x = (x \odot^i y)[\mathbf{0}, |x| - \mathbf{1}]$

(ii) $y = (x \odot^i y)[(0, \ldots, 0, |x|_i, 0, \ldots, 0), (0, \ldots, 0, |x|_i, 0, \ldots, 0) + |y| - \mathbf{1}).$

The d-dimensional empty word ε_d is a word of shape $\mathbf{0}$. We extend the definition to the concatenation of ε_d and any d-dimensional word x in the direction $i \in [\![1,d]\!]$ by

$$\varepsilon_d \overset{i}{\odot} x = x \overset{i}{\odot} \varepsilon_d = x.$$

In particular, $\varepsilon_d \odot^i \varepsilon_d = \varepsilon_d$.

Example 1.6.5. Consider the two bidimensional arrays

$$x = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array} \quad \text{and} \quad y = \begin{array}{|c|c|c|} \hline a & a & b \\ \hline b & c & d \\ \hline \end{array}$$

of shape respectively $|x| = (2,2)$ and $|y| = (2,3)$. Since $|x|_{\hat{2}} = |y|_{\hat{2}} = 2$, we get

$$x \overset{2}{\odot} y = \begin{array}{|c|c|c|c|c|} \hline a & b & a & a & b \\ \hline c & d & b & c & d \\ \hline \end{array}.$$

However $x \odot^1 y$ is not defined because $2 = |x|_{\hat{1}} \neq |y|_{\hat{2}} = 3$.

Definition 1.6.6. Let $\mu : \Sigma \to B_d(\Sigma)$ be a map and x be a d-dimensional array such that

$$\forall i \in [\![1,d]\!], \forall k < |x|_i, \forall a, b \in \mathrm{Fact}_{\mathbf{1}}(x_{|i,k}) : |\mu(a)|_i = |\mu(b)|_i. \tag{1.1}$$

Then the image of x by μ is the d-dimensional array defined by

$$\mu(x) = \overset{1}{\underset{0 \leq n_1 \leq |x|_1}{\odot}} \left(\cdots \left(\overset{d}{\underset{0 \leq n_d \leq |x|_d}{\odot}} \mu(x(n_1, \ldots, n_d)) \right) \cdots \right).$$

Note that the ordering of the products in the different directions is unimportant.

Definition 1.6.7. Let $\mu : \Sigma \to B_d(\Sigma)$ be a map. If for all $a \in \Sigma$ and all $n \geq 1$, $\mu^n(a)$ is inductively well defined from $\mu^{n-1}(a)$, that is, $\mu^{n-1}(a)$ satisfies (1.1), then μ is said to be a d-dimensional morphism. Let a be a letter such that $(\mu(a))(\mathbf{0}) = a$. We say that μ is prolongable on a. Then the limit

$$w = \mu^\omega(a) := \lim_{n \to +\infty} \mu^n(a)$$

is well defined and $w = \mu(w)$ is a fixed point of μ.

Chapter 2

Self-generating sets

2.1 Introduction

Let I be a finite set of integers and F be a finite set of maps of the form $\varphi_i : n \mapsto k_i\, n + \ell_i$ with integer coefficients. For an integer base $k \geq 2$, we study the k-recognizability of the minimal set X of integers containing I and satisfying $\varphi(X) \subseteq X$ for all $\varphi \in F$. The material of this chapter can be found in [KLR09, KLR10] .

Self-generating sets of integers were introduced by Kimberling [Kim00]. They can be defined as follows. Let $r \geq 1$, I be a finite set of integers and $G = \{\varphi_1, \varphi_2, \ldots, \varphi_r\}$ be a set of affine maps where $\varphi_i : n \mapsto k_i\, n + \ell_i$ with $k_i, \ell_i \in \mathbb{Z}$ and $2 \leq k_1 \leq k_2 \leq \cdots \leq k_r$. The set *generated by G and I* is the minimal subset X of $\mathbb{Z}$ containing I and such that $\varphi_i(X) \subseteq X$ for all $i = 1, \ldots, r$. For any subset $S \subseteq \mathbb{Z}$, we set $G(S) := \{\varphi(s) \mid s \in S, \varphi \in G\}$, $G^0(S) := S$ and $G^{m+1}(S) := G(G^m(S))$ for all $m \geq 0$. Otherwise stated, $X = \bigcup_{m \geq 0} G^m(I)$ is the set of all integers n such that there exist $m \geq 0$, $a \in I$ and a finite sequence $(\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_m})$ of maps in G such that

$$n = \varphi_{i_m} \circ \varphi_{i_{m-1}} \circ \cdots \circ \varphi_{i_1}(a) = \varphi_{i_m}(\varphi_{i_{m-1}}(\cdots \varphi_{i_1}(a) \cdots)). \tag{2.1}$$

Example 2.1.1. Let $G = \{n \mapsto 2n, n \mapsto 4n - 1\}$ and $I = \{1\}$. The corresponding self-generating set is the set

$$\mathcal{K}_1 = \{1, 2, 3, 4, 6, 7, 8, 11, 12, 14, 15, 16, \ldots\}$$

studied by Kimberling in [Kim00]. Notice that for $I = \{0\}$, we get a subset containing negative integers: $\mathcal{K}_0 = \{0, -1, -2, -4, -5, -8, -9, \ldots\}$. In particular, for $I = \{0, 1\}$, the corresponding self-generating set is $\mathcal{K}_0 \cup \mathcal{K}_1$.

Kimberling [Kim00] asks the following question: *How are the even numbers in $\mathcal{K}_1$ distributed among the odds ?*   To answer this question, Kimberling considers the increasing sequence S of the elements of $\mathcal{K}_1$ and, numbering the positions of these elements to the right of the initial 1, he constructs the rank-past-1 sequences s_1 and s_2 given by the even numbers of S and the odd numbers respectively as follows.

$\mathcal{K}_1$	1	2	3	4	6	7	8	11	12	14	15	16	...
s_1	1		3	4		6		8	9		11		...
s_2		2			5		7			10			...

Then Kimberling proves the following result.

Theorem 2.1.2. *The rank-past-1 sequence for even terms of S is given by*

$$s_1(n) = \lfloor n\tau \rfloor$$

and the rank-past-1 sequence for odd terms of S is given by

$$s_2(n) = n + \lfloor n\tau \rfloor$$

where τ is the Golden mean.

This theorem proves that the set $\mathcal{K}_1$ is closely related to the Fibonacci word defined as the fixed point $\varphi^\omega(0) = 01\varphi(1)\varphi^2(1)\cdots = 01001010\cdots$ of the morphism $\varphi\colon 0 \mapsto 01, 1 \mapsto 0$.

Corollary 2.1.3. *The rank-past-1 sequences s_1 and s_2 give the position of 1's and 0's in the Fibonacci word.*

A particular case of self-generated sets is studied by Kimberling [Kim04]. The set of initial integers I is the singleton $\{1\}$ and the set of generating functions G is a set of maps of the form $\varphi_i\colon n \mapsto k_i\, n + \ell_i$, where $k_i, \ell_i \in \mathbb{N}$ for $i = 1, \ldots, r$. These sets are called *affinely recursive sets*. It is clear that each $x \in X = \bigcup_{m \geq 0} G^m(I)$ except 1 is given by

$$x = \varphi_{i_m}(\varphi_{i_{m-1}}(\cdots \varphi_{i_1}(1) \cdots)) \tag{2.2}$$

for some finite sequence $i_1 \cdots i_m$. Let s be the increasing sequence of elements of X. The relation (2.2) induces a mapping from $X \backslash \{1\}$ to the language $L(A_r)$ of non-empty words $i_1 \cdots i_m$ over $A_r = \{1, \ldots, r\}$. Kimberling gives some conditions under which the sets $\varphi_i(X)$ are pairwise disjoint, inducing that this mapping is a bijection, which in turn implies that the natural

ordering on the integers induces an ordering on the set of non-empty words over A_r providing a kind of abstract numeration system.

Example 2.1.4. Consider the self-generating set $X = \bigcup_{m \geq 0} G^m(I)$ defined by $G = \{\varphi_0 : n \mapsto 2n, \ \varphi_1 : n \mapsto 4n + 3\}$ and $I = \{1\}$. It is clear that $\varphi_0(X)$ and $\varphi_1(X)$ are disjoint. We can define a kind of abstract numeration system $S = (\Sigma_2^*, \Sigma_2, <_X)$ where the order $<_X$ is induced by the natural order on X. The first words of the language Σ_2^* with respect to the order $<_X$ are given in the following array

X	Σ_2^*
1	ε
$2 = \varphi_0(1)$	0
$4 = \varphi_0(\varphi_0(1))$	00
$7 = \varphi_1(1)$	1
$8 = \varphi_0(\varphi_0(\varphi_0(1)))$	000
$11 = \varphi_1(\varphi_0(1))$	10
$14 = \varphi_0(\varphi_1(1))$	01
$16 = \varphi_0(\varphi_0(\varphi_0(\varphi_0(1))))$	0000
$17 = \varphi_1(\varphi_0(\varphi_0(\varphi_0(1))))$	1000
$\vdots$	$\vdots$

and are the S-representations of the first non-negative integers as follows.

$\mathbb{N}$	$\mathrm{rep}_S(\mathbb{N})$
0	ε
1	0
2	00
3	1
4	000
5	10
6	01
7	0000
8	1000
$\vdots$	$\vdots$

Allouche, Shallit and Skordev [ASS05] consider a general framework for self-generating sets. The k-ary representations of the elements in some self-generating sets are related to words over $\Sigma_k = \{0, 1, \ldots, k-1\}$ where some fixed block of digits is missing. As an illustration, one can notice that the set $\mathcal{K}_1 - 1 = \{0, 1, 2, 3, 5, 6, 7, 10, \ldots\}$ introduced in Example 2.1.1 consists of all

integers whose binary expansion does not contain "00" as factor. Allouche *et al.* give some other examples and ask whether or not the sets of integers with missing blocks are always particular cases of affinely recursive sets.

They also give a new proof of Kimberling's result and they study self-generating sets from the point of view of automatic and morphic sequences as well as in relation to non-standard numeration systems. In particular, they proved the following two results.

Proposition 2.1.5. *The characteristic sequence of $\mathcal{K}_1 - 1$ is 2-automatic. In other words, $\mathcal{K}_1 - 1$ is 2-recognizable.*

Proposition 2.1.6. *Let s be the increasing sequence of elements of $\mathcal{K}_1 - 1$. The sequence s reduced modulo any integer greater or equal to 2 is morphic.*

Similarly, they prove that s is Fibonacci-regular and that this sequence reduced modulo any integer greater or equal to 2 is Fibonacci-automatic (for the definitions, see Chapter 1 Section 1.5). Moreover, Allouche *et al.* ask the following question: *Under what conditions is the characteristic sequence of a self-generating set k-automatic?* They also present the following conjecture.

Conjecture 1. With "mixed base" rules, such as $G = \{n \mapsto 2n+1, n \mapsto 3n\}$, the set generated from $I = \{1\}$ is not k-recognizable for any integer base $k \geq 2$.

Garth and Gouge [GG07] partially answer the previous question by studying a particular case of self-generating sets. Let r and k be two positive integers. Let $I = \{1\}$ and $G = \{\varphi_1, \ldots, \varphi_r\}$ be such that $\varphi_i \colon n \mapsto k^{e_i} n + \ell_i$ with $1 = e_1 \leq e_2 \leq \cdots \leq e_r$, $\ell_1 = 0$ and $0 \leq \ell_i \leq k^{e_i} - 1$. Garth and Gouge prove that the increasing sequence s of elements of the set X generated by G and I reduced modulo a positive integer is the image under a coding of an infinite word that is a fixed point of a morphism over a finite alphabet. They also prove that the characteristic sequence of X is k-automatic.

Let us fix the notation once and for all.

Definition 2.1.7. In this chapter, instead of considering a set G of maps as described above, we will consider the extended set of $r + 1 \geq 2$ maps

$$F = G \cup \{\varphi_0\} = \{\varphi_0, \varphi_1, \ldots, \varphi_r\}$$

where $\varphi_0 \colon n \mapsto n$ and $\varphi_i \colon n \mapsto k_i\, n + \ell_i$ with $k_i, \ell_i \in \mathbb{Z}$ and

$$2 \leq k_1 \leq k_2 \leq \cdots \leq k_r.$$

Having the identity function φ_0 at our disposal, for any set $S \subseteq \mathbb{Z}$, we have $F^m(S) \subseteq F^{m+1}(S)$. Therefore, for any finite set I of integers, the set

$$F^\omega(I) := \lim_{m \to \infty} F^m(I)$$

is exactly the *self-generating set* with respect to G and I.

The content of the chapter is the following. In Section 2.2 we give some simple observations on self-generating sets. For example, if we add to F an extra map $\psi \colon n \mapsto n + \ell$ with $\ell \neq 0$, then the corresponding self-generating set $F^\omega(I)$ is ultimately periodic and therefore k-recognizable for all $k \geq 2$. We also show that we can restrict our considerations to subsets of $\mathbb{N}$ and assume that all additive constants ℓ_i for the maps $\varphi_i \in F$ are non-negative.

In sections 2.3 and 2.4 we consider the multiplicatively dependent case. The results are based on Frougny's normalization transducer. If all multiplicative constants k_i are pairwise multiplicatively dependent, then we give a general method to build a finite automaton recognizing $\mathrm{rep}_k(F^\omega(I))$ for any k that is multiplicatively dependent on every k_i. This allows us to generalize a recognizability result considered in [GG07]. Moreover, a new proof of the relation between the Kimberling set $\mathcal{K}_1$ and the infinite Fibonacci word is given in Section 2.4; for other proofs, see [Kim00, ASS05].

In the multiplicatively independent case of Section 2.5 we study differences and ratios of consecutive elements in the considered self-generating set. The results rely on a classical gap theorem; see Theorem 2.5.1. We prove that if there exist i, j such that k_i and k_j are multiplicatively independent and if $\sum_{i=1}^{r} k_i^{-1} < 1$, then $F^\omega(I)$ is not k-recognizable for any $k \geq 2$. In particular, this condition always holds for sets F where $r = 2$ and $k_1 < k_2$ are multiplicatively independent, answering Conjecture 1 in the affirmative.

2.2 Some reductions

First we show that assuming $k_i \geq 2$ for every $i = 1, 2, \ldots, r$ is not a true restriction from the point of view of recognizability.

Lemma 2.2.1. *If we add to F in Definition 2.1.7 an extra map $\psi \colon n \mapsto n + \ell$ with $\ell \neq 0$, then the corresponding self-generating set $F^\omega(I)$ is ultimately periodic of period ℓ.*

Proof. Denote by $F^j(I) \mod \ell$ the set $\{n \mod \ell \mid n \in F^j(I)\}$. Recall that the identity function φ_0 belongs to F. Since there are finitely many congruence classes modulo ℓ and $F^j(I) \mod \ell \subseteq F^{j+1}(I) \mod \ell$, there must

exist an integer J such that $F^{J+1}(I) \mod \ell = F^J(I) \mod \ell$. Moreover, this means that $F^j(I) \mod \ell = F^J(I) \mod \ell$ for every $j \geq J$, and, consequently,

$$F^\omega(I) \mod \ell = F^J(I) \mod \ell. \qquad (2.3)$$

On the other hand, if $n \in F^\omega(I)$, then $\psi^t(n) = n + t\,\ell \in F^\omega(I)$. Since $n + t\,\ell \equiv n \mod \ell$, we conclude by (2.3) that, for any $n \geq \max F^J(I)$,

$$\mathbf{c}_{F^\omega(I)}(n) = \left\{ \begin{array}{ll} 1, & \text{if } n \mod \ell \in F^J(I) \mod \ell; \\ 0, & \text{otherwise.} \end{array} \right.$$

Hence, the characteristic sequence of $F^\omega(I)$ is ultimately periodic with preperiod $\max F^J(I)$ and period ℓ. $\qquad\square$

Remark 2.2.2. In Definition 2.1.7 and in what follows, we always assume that all multiplicative constants k_i of the affine maps $\varphi_1, \ldots, \varphi_r$ in F are at least 2. This condition does not guarantee that the corresponding self-generating set is not ultimately periodic. For example, if $\varphi_i(x) = r\,x + i$ for $i = 1, \ldots, r$, then we easily see that $F^\omega(\{0\}) = \mathbb{N}$.

The next lemma justifies that we may restrict our consideration to non-negative integers.

Lemma 2.2.3. *Let $F^\omega(I)$ be a self-generating set as in Definition 2.1.7. One can effectively construct two finite sets of non-negative integers I^+ and I^- such that*

$$F^\omega(I) \cap \mathbb{N} = F^\omega(I^+) \cap \mathbb{N} \quad \text{and} \quad F^\omega(I) \cap -\mathbb{N} = -(\overline{F}^\omega(I^-) \cap \mathbb{N}),$$

where $-\mathbb{N}$ is the set of all non-positive integers and $\overline{F} = \{\varphi_0, \overline{\varphi}_1, \overline{\varphi}_2, \ldots, \overline{\varphi}_r\}$ with $\overline{\varphi}_i \colon n \mapsto k_i\,n - \ell_i$ for $i = 1, 2, \ldots, r$.

Proof. Let $m = \max\{|\ell_i| \mid i = 1, 2, \ldots, r\}$ and denote by M the interval of integers $[\![-m, m]\!]$. Define $I_j := F^j(I) \cap M$ for $j \geq 0$. Since $k_i \geq 2$ for all $i \in \{1, 2, \ldots, r\}$, it follows that if n does not belong to M, then $\varphi_i(n) \notin M$ for all $i \in \{0, 1, \ldots, r\}$. By this property and since $F^j(I) \subseteq F^{j+1}(I)$, there must exist an integer J such that $I_j = I_J$ for all $j \geq J$. Hence, the integers of $F^\omega(I)$ falling into the interval M are exactly the ones in I_J and we can find the set $I^+ := ((F^\omega(I) \cap M) \cup I) \cap \mathbb{N}$ effectively.

Next we show that $F^\omega(I) \cap \mathbb{N} = F^\omega(I^+) \cap \mathbb{N}$. Since $I^+ \subseteq F^\omega(I)$, it is clear by definition that $F^\omega(I^+) \cap \mathbb{N} \subseteq F^\omega(I) \cap \mathbb{N}$. Assume now that there exists an integer x belonging to $(F^\omega(I) \cap \mathbb{N}) \setminus (F^\omega(I^+) \cap \mathbb{N})$. Since I^+ contains all non-negative elements of I, the element x must be generated from some negative

element $a \in I$. In other words, there exists a finite sequence $(\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_t})$ of maps in F such that $x = \varphi_{i_t} \circ \varphi_{i_{t-1}} \circ \cdots \circ \varphi_{i_1}(a)$. Since a is negative and x is positive, there exists j such that $\varphi_{i_{j-1}} \circ \varphi_{i_{j-2}} \circ \cdots \circ \varphi_{i_1}(a) = y < 0$ and $\varphi_{i_j}(y) = z \geq 0$. Hence, we have $k_{i_j} y < 0$ and $z = k_{i_j} y + \ell_{i_j} < m$. This means that $z \in (F^\omega(I) \cap M) \cap \mathbb{N}$ and therefore $x = \varphi_{i_t} \circ \varphi_{i_{t-1}} \circ \cdots \circ \varphi_{i_{j+1}}(z) \in F^\omega(I^+) \cap \mathbb{N}$. This is a contradiction.

Similarly, by defining $I^- := -(((F^\omega(I) \cap M) \cup I) \cap -\mathbb{N})$, we obtain $F^\omega(I) \cap -\mathbb{N} = F^\omega(-I^-) \cap -\mathbb{N}$. If $\overline{F} = \{\varphi_0, \overline{\varphi}_1, \overline{\varphi}_2, \ldots, \overline{\varphi}_r\}$, where $\overline{\varphi}_i \colon n \mapsto k_i\, n - \ell_i$ for $i = 1, 2, \ldots, r$, then we clearly have $F^\omega(I) \cap -\mathbb{N} = -(\overline{F}^\omega(I^-) \cap \mathbb{N})$, which concludes the proof. $\qquad\square$

Let $y \geq 0$. Recall that a set $Y \subseteq \mathbb{N}$ is k-recognizable if and only if $Y + y$ is k-recognizable (see Proposition 1.3.12 page 14). As explained by the following lemma, from the point of view of recognizability of subsets of $\mathbb{N}$, one can also assume that all additive constants ℓ_i are non-negative.

Lemma 2.2.4. *Let $F^\omega(I)$ be a self-generating set as in Definition 2.1.7. There exist a non-negative integer y and a self-generating set $\widehat{F}^\omega(I - y)$ such that $F^\omega(I) = \widehat{F}^\omega(I - y) + y$ and $\widehat{F} = \{\varphi_0, \widehat{\varphi}_1, \ldots, \widehat{\varphi}_r\}$, where $\widehat{\varphi}_i \colon n \mapsto k_i\, n + \widehat{\ell}_i$ for every $i = 1, 2, \ldots, r$ with some non-negative constants $\widehat{\ell}_i$ completely determined by F.*

Proof. Assume that at least for some function $\varphi_i \in F$ the constant ℓ_i is negative. Otherwise, the claim is trivial. Let $y = \max\{|\ell_i| \mid \ell_i < 0\}$ and set

$$\widehat{\ell}_i := \ell_i + (k_i - 1)y$$

for $i = 1, 2, \ldots, r$. Since $k_i \geq 2$, the constants $\widehat{\ell}_i$ are non-negative for all i. Let $\widehat{F} = \{\varphi_0, \widehat{\varphi}_1, \ldots, \widehat{\varphi}_r\}$ where $\widehat{\varphi}_i \colon n \mapsto k_i\, n + \widehat{\ell}_i$ for $i = 1, \ldots, r$. We show by induction on the number of applied maps m that x belongs to $F^m(I)$ if and only if $x - y$ belongs to $\widehat{F}^m(I - y)$.

First, for any $x \in I$, it is obvious that $x - y$ belongs to $I - y$ and vice versa. Assume now that $x \in F^m(I)$ for some $m \geq 1$. In other words, there exist $z \in F^{m-1}(I)$ and $i \in \{0, \ldots, r\}$ such that $x = \varphi_i(z)$. By the induction hypothesis, $z - y$ belongs to $\widehat{F}^{m-1}(I - y)$. If $\varphi_i = \varphi_0$, then $x = z$ and $x - y \in \widehat{F}^{m-1}(I - y) \subseteq \widehat{F}^m(I - y)$. Hence, assume that $\varphi_i \neq \varphi_0$. We have $\varphi_i(z) = k_i\, z + \ell_i$ and $\widehat{\varphi}_i(z - y) = k_i(z - y) + \ell_i + (k_i - 1)y = \varphi_i(z) - y$. This proves that $x - y$ belongs to $\widehat{F}^m(I - y)$.

Next assume that $x - y \in \widehat{F}^m(I - y)$ for some $m \geq 1$, i.e., $x - y = \widehat{\varphi}_i(z)$ for some $z \in \widehat{F}^{m-1}(I - y)$ and $i \in \{0, \ldots, r\}$. As above, we may assume that

$\varphi_i \neq \varphi_0$. Then we have

$$
\begin{aligned}
x &= \widehat{\varphi}_i(z) + y \\
&= k_i z + \widehat{\ell}_i + y \\
&= k_i z + \ell_i + (k_i - 1)y + y \\
&= k_i(z + y) + \ell_i \\
&= \varphi_i(z + y),
\end{aligned}
$$

where $z + y$ belongs to $F^{m-1}(I)$ by induction hypothesis. Hence, x belongs to $F^m(I)$. $\square$

Example 2.2.5. Consider the set $\mathcal{K}_1$ of Example 2.1.1 generated from $\{1\}$ by the maps $n \mapsto 2n$ and $n \mapsto 4n - 1$. Applying the construction given in the previous proof, set $y = 1$ and consider the maps $n \mapsto 2n + 1$ and $n \mapsto 4n + 2$. These two maps generate from $\{1\} - 1 = \{0\}$, the set $\{0, 1, 2, 3, 5, 6, 7, 10, \ldots\}$ which is equal to $\mathcal{K}_1 - 1$.

2.3   The Multiplicatively Dependent Case

In this section we assume that the multiplicative coefficients k_i appearing in Definition 2.1.7 are all pairwise multiplicatively dependent, i.e., for every pair (i, j), there exist positive integers e_i and e_j such that $k_i^{e_i} = k_j^{e_j}$. Note that k_i and k_j are multiplicatively dependent if and only if there exists an integer $n \geq 2$ and two integers $d_i, d_j \geq 1$ such that $k_i = n^{d_i}$ and $k_j = n^{d_j}$. By this characterization, it is easy to see that if the coefficients k_i are pairwise multiplicatively dependent, then there exists an integer k such that every k_i is a power of k. Our aim is to build a finite automaton showing that the set $F^\omega(I)$ is k-recognizable.

Theorem 2.3.1. *Let F given in Definition 2.1.7 be such that the multiplicative coefficients $k_1, \ldots, k_r$ are all pairwise multiplicatively dependent. Let k be such that k_i is a power of k for every $i = 1, 2, \ldots, r$. For any finite $I \subset \mathbb{Z}$, the self-generating set $F^\omega(I)$ is k-recognizable.*

We give two proofs for this theorem. First we give a proof relying on Frougny's normalization theorem, then we give an effective proof.

Proof 1 (relying on Frougny's normalization theorem). Let us assume that the maps in F are of the kind $\varphi_i \colon n \mapsto k^{e_i} n + \ell_i$ with $e_i \geq 1$ for all $i \in \{1, \ldots, r\}$. Since in the constructions of $\overline{F}$ and $\widehat{F}$ of Lemma 2.2.3 and

Lemma 2.2.4 in Section 2.2 the multiplicative constants k_i are not modified, it suffices to consider only non-negative elements of $F^\omega(I)$ and, moreover, we may assume that all initial values in I and all additive constants ℓ_i are non-negative. Thus, we assume $F^\omega(I) \subseteq \mathbb{N}$ and show that this self-generating set is k-recognizable.

Let n be an element of $F^\omega(I)$. In other words, there exists a finite sequence $(\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_m})$ of maps in F such that

$$n = \varphi_{i_m}(\varphi_{i_{m-1}}(\cdots \varphi_{i_1}(a) \cdots))$$

for some $a \in I$. With that integer, we associate the word

$$w = a\, 0^{e_{i_1}-1} \ell_{i_1} \cdots 0^{e_{i_m}-1} \ell_{i_m}$$

over the finite alphabet $\Sigma = I \cup \{0, \ell_1, \ldots, \ell_r\} \subset \mathbb{N}$. One can notice that $\mathrm{val}_{\Sigma,k}(w) = n$ and $\mathrm{val}_{\Sigma,k}(I\{0^{e_1-1}\ell_1, \ldots, 0^{e_r-1}\ell_r\}^*) = F^\omega(I)$. Frougny's normalization theorem says that normalization over Σ is computable by a finite transducer T. By Proposition 1.2.25, if a regular language L is an input of a transducer then the output language is also regular. Hence, feeding the transducer T with the language $I\{0^{e_1-1}\ell_1, \ldots, 0^{e_r-1}\ell_r\}^*$ gives us the regular language $\mathrm{rep}_k(F^\omega(I))$, which proves the claim. $\qquad\square$

We give below another proof which is independent from Frougny's normalization theorem. It describes a way to build an automaton recognizing the k-ary representations of $F^\omega(I)$.

Proof 2. From Lemma 2.2.3, we may assume without loss of generality that I and X are subsets of $\mathbb{N}$. To preserve the effectiveness of the proof, we do not restrict the constants ℓ_i to be non-negative integers. Let k be an integer such that all the coefficients k_i are powers of k. Note that $\mathrm{val}_k^{-1}(n)$ contains all the representations of the integer n in base k over Σ_k^*, including those with leading zeros. We define a non-deterministic finite automaton $\mathcal{A} = (Q, \{q_0\}, \Sigma_k, \Delta, T)$ accepting the mirror of the elements in $\mathrm{val}_k^{-1}(F^\omega(I))$, so we may allow leading zeros in front of the most significant digit. The transition relation Δ is a finite subset of $Q \times \Sigma_k^* \times Q$. If (p, w, q) belongs to Δ, we write $p \xrightarrow{w} q$. An input $x \in \Sigma_k^*$ is accepted if and only if there is a sequence of states $q_0, q_1, \ldots, q_i$ such that $q_i \in T$, x can be factorized as $u_1 \cdots u_i$ and $(q_0, u_1, q_1), (q_1, u_2, q_2), \ldots (q_{i-1}, u_i, q_i) \in \Delta$.

Let M be the maximal element in $I \cup \{k_1, \ldots, k_r, |\ell_1|, \ldots, |\ell_r|\}$ and $m = |\mathrm{rep}_k(M)|$. Define $Q = \{q_0\} \cup (\{-1, 0, +1\} \times \Sigma_k^{m+1})$. A state $q = (c, x) \in Q \setminus \{q_0\}$ is final if and only if $c = 0$ and $x \in \mathrm{val}_k^{-1}(I)$. From the initial state q_0, we have all the transitions

$$q_0 \xrightarrow{w} (0, \widetilde{w})$$

where $w \in \Sigma_k^{m+1}$ and $\widetilde{w}$ is the mirror of w. Recall that entries are read in $\mathcal{A}$ the least significant digit first, that is from right to left. This explains why we consider the mirrors in the encoding. From each state $Q \setminus \{q_0\}$ there are transitions corresponding to the maps φ_i, $i = 1, 2, \ldots, r$. The idea is to guess the sequence of maps $(\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_m})$ that was used to obtain the integer corresponding to the input belonging to $\mathrm{val}_k^{-1}(F^\omega(I))$ and apply the inverses of these maps in reversed order to get back the representation of one of the initial values in I. The first component of a state $q = (c, x_m x_{m-1} \cdots x_0)$ corresponds to a carry bit and the second component represents the last $m+1$ digits of a number, x_0 being the least significant one. We show how to simulate the multiplications and additions in the successive applications of the affine functions φ_i using only the carry bit c and the digits $x_m x_{m-1} \cdots x_0$.

First consider a state $p = (0, x_m x_{m-1} \cdots x_0)$ and φ_i-transitions, where $\varphi_i : n \mapsto k_i\, n + \ell_i$ and $\ell_i \geq 0$. For the inverse of φ_i, we want to subtract ℓ_i and then divide by $k_i = k^t$ for some positive integer t. We do this using the classical paper-and-pencil method as illustrated in Figure 2.1(a),

$$
\begin{array}{r}
x x_m x_{m-1} \cdots x_1 x_0 \\
- \quad y_m y_{m-1} \cdots y_1 y_0 \\
\hline
z_m z_{m-1} \cdots z_1 z_0 \\
\text{(a) subtraction}
\end{array}
\qquad
\begin{array}{r}
x_m x_{m-1} \cdots x_1 x_0 \\
+ \quad y_m y_{m-1} \cdots y_1 y_0 \\
\hline
z z_m z_{m-1} \cdots z_1 z_0 \\
\text{(b) addition}
\end{array}
$$

Figure 2.1: The paper-and-pencil subtraction and addition.

where $\mathrm{val}_k(y_m y_{m-1} \cdots y_0) = \ell_i$ and $x = 1$, if $\mathrm{val}_k(x_m x_{m-1} \cdots x_0) < \ell_i$ and $x = 0$, otherwise. Note that, by definition of m, we have $y_m = 0$. Hence, a "carry" bit x might be needed only if $x_m = 0$. Multiplying an integer n by k^t corresponds to adding t zeros at the end of the k-ary representation of n. Hence, if φ_i is the correct guess, $z_m z_{m-1} \cdots z_0$ should have at least t zeros as suffix. If this is not the case, we choose to have no φ_i-transitions starting from p. If $x = 0$, then φ_i-transitions are of the form

$$p \xrightarrow{w} (0, \widetilde{w} z_m \cdots z_t), \tag{2.4}$$

where w is any word over Σ_k of length t. If $x = 1$, then we have two cases depending on the form of $w \in \Sigma_k^t$:

1. If $w = 0^t$, then the transition is

$$p \xrightarrow{w} \left(-1, (k-1)^t z_m \cdots z_t\right), \tag{2.5}$$

   where the first component -1 indicates that a carry was needed in a "previous" subtraction and it must be borrowed from the first non-zero digit of the input that will be read in the future.

2. Otherwise $\widetilde{w} = vu0^s$, where $s < t$, $u \in \{1, 2, \ldots, k-1\}$ and $v \in \Sigma_k^{t-s-1}$, then the transition is

$$p \xrightarrow{w} (0, v(u-1)(k-1)^s z_m \cdots z_t). \qquad (2.6)$$

Here the carry $x = 1$ was borrowed from u and no carry is postponed to future calculations.

Consider next a state $p = (0, x_m x_{m-1} \cdots x_0)$ and φ_i-transitions, where $\varphi_i : n \mapsto k_i\, n + \ell_i$ and $\ell_i < 0$. Instead of subtraction, we consider now addition by the paper-and-pencil method where $\mathrm{val}_k(y_m y_{m-1} \cdots y_0) = |\ell_i|$. This is illustrated in Figure 2.1(b). Note that since $y_m = 0$ by the definition of m, a carry $z = 1$ can occur only if $x_m = k - 1$. As above, the φ_i-transitions exist only if the last t digits of $z_m z_{m-1} \cdots z_0$ are zeros. This holds also for any transition considered in the sequel. If $z = 0$, then we have the transitions of the form (2.4). If $z = 1$ we have again two cases depending on the digits of $w \in \Sigma_k^t$:

1. If $w = (k-1)^t$, then the carry is shifted to future calculations.

$$p \xrightarrow{w} \left(+1, 0^t z_m \cdots z_t\right). \qquad (2.7)$$

2. If $\widetilde{w} = vu(k-1)^s$, where $s < t$, $u \in \{0, 1, \ldots, k-2\}$ and $v \in \Sigma_k^{t-s-1}$, then

$$p \xrightarrow{w} (0, v(u+1)0^s z_m \cdots z_t). \qquad (2.8)$$

Here the carry is added to the digit u and no carry is postponed to future calculations.

Secondly, consider a state of the form $p = (-1, x_m x_{m-1} \cdots x_0)$ and assume that $\ell_i \geq 0$. The carry component -1 means that we have borrowed a carry in a subtraction and after the subtraction we have read only zeros, which have been turned into digits $k - 1$. Otherwise, if non-zero digits were read, there would be no longer a carry -1. Hence, we can be sure that $x_m = k - 1$ in Figure 2.1(a), and consequently, we have $x = 0$, since $y_m = 0$. This is important, since it means that no "new" carry is borrowed. Again, assume that $z_{t-1} \cdots z_0 = 0^t$. If $w = 0^t$, then the transition is of type (2.5). Otherwise, the transitions are of type (2.6).

If $\ell_i < 0$, then we perform addition as in Figure 2.1(b) and assume $z_{t-1} \cdots z_0 = 0^t$. If $z = 0$ and $w = 0^t$, no new carries occur and the transition is again of type (2.5). If $z = 0$ and $w \neq 0^t$, then the transitions are of type (2.6). If $z = 1$, then this positive carry and the negative carry borrowed

in a previous calculation annihilate each other. Hence, the transition is of type (2.4).

Finally, consider the states of the form $p = (+1, x_m x_{m-1} \cdots x_0)$. The carry component $+1$ means that we obtained a carry in an addition and after the addition we have read only digits $k - 1$, which have been turned into zeros. Otherwise, if a digit $u \neq k-1$ were read, it would have been turned into $u + 1$ and there would be no longer a carry $+1$. Hence, we conclude that $x_m = 0$ in Figure 2.1(a) and Figure 2.1(b). If $\ell_i \geq 0$, then consider Figure 2.1(a) and assume $z_{t-1} \cdots z_0 = 0^t$. If $x = 0$ and $w = (k-1)^t$, then the transition is of type (2.7). If $x = 0$ and $w \neq (k-1)^t$, then the transitions are of type (2.8). If $x = 1$, then the negative and positive carry annihilate each other and the transitions are of type (2.4). Assume now that $\ell_i < 0$ and $z_{t-1} \cdots z_0 = 0^t$. In Figure 2.1(b) no new carry $z = 1$ can occur, since both $x_m = 0$ and $y_m = 0$. Hence, we have only two cases. If $w = (k-1)^t$, then the transition is of type (2.7). Otherwise, it is of type (2.8).

If $n = \varphi_{i_m}(\varphi_{i_{m-1}}(\cdots \varphi_{i_1}(a) \cdots))$ for some $a \in I$, then using the above transitions and k-ary representations, we are able to correctly simulate the calculation $n \mapsto \varphi_{i_m}^{-1}(n) \mapsto \varphi_{i_{m-1}}^{-1}(\varphi_{i_m}^{-1}(n)) \mapsto \cdots \mapsto a$ as long as the k-ary representation of n given as input contains enough leading zeros. However, we may fix this by replacing the set of final states T by an enlarged set T'. A state $q' \in Q$ belongs to T' if there exists a path with label 0^t, $t \geq 0$, from q' to some state $q \in T$. Hence, with the modified final states the automaton $\mathcal{A}$ accepts all the mirrors of the words in $\mathrm{val}_k^{-1}(F^\omega(I))$. On the other hand, it cannot accept any other word. Namely, for any word w accepted by $\mathcal{A}$ there is a sequence $(\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_m})$ such that (2.1) holds for $n = \mathrm{val}_k(w)$. It is well known that any non-deterministic finite automaton can be turned into a DFA, e.g., by the subset construction. Hence, $F^\omega(I)$ is k-recognizable. $\quad\square$
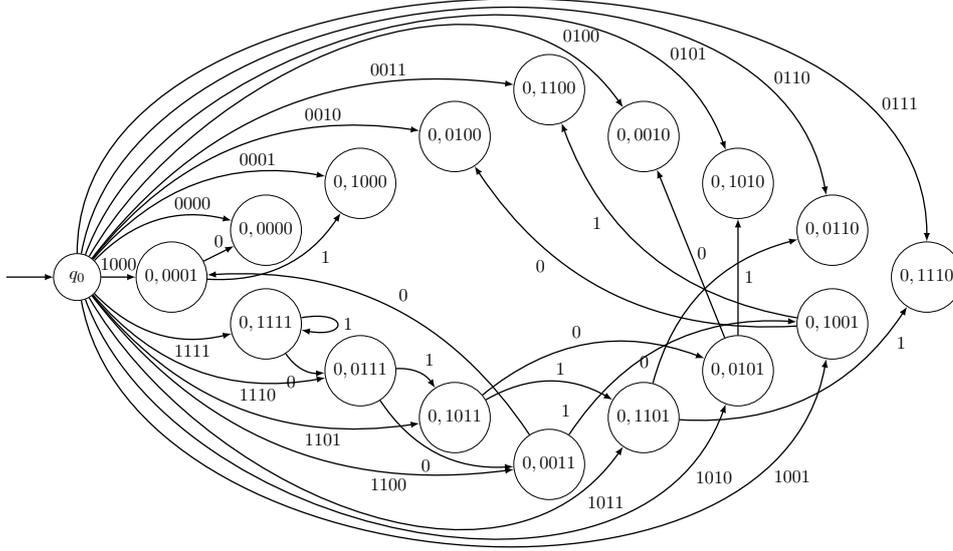
Example 2.3.2. If we consider the set $\mathcal{K}_1 - 1$ introduced in Example 2.1.1 and studied by Allouche *et al.* [ASS05], we have

$$G = \{\varphi_1 : n \mapsto 2n + 1, \varphi_2 : n \mapsto 4n + 2\} \text{ and } I = \{0\}$$

and we obtain $k = 2$, $M = \max\{I \cup \{k_1, k_2, |\ell_1|, |\ell_2|\} = 4$, $m = |\mathrm{rep}_2(4)| = 3$, $Q = \{q_0\} \cup (\{-1, 0, +1\} \times \Sigma_2^4)$, $\Delta \subseteq Q \times \Sigma_2^* \times Q$.

To construct the automaton accepting the mirror of $val_2^{-1}(\mathcal{K}_1 - 1)$ we have first transitions from q_0 to $(0, w^R)$ labelled by w for all $w \in \Sigma_2^4$. These first transitions are depicted in Figure 2.2.

Next we simulate the inverse of $\varphi_1 : n \mapsto 2n + 1$ on each state. Consider first the state $(0, 0001)$. Subtracting 1 from 0001 gives $(0)0000$ (where (0) means that there is no carry) and dividing 0000 by 2 gives 000. Then we

Figure 2.2: Transitions from q_0.

have two transitions simulating the inverse of φ_1 from $(0, 0001)$. However, if we consider $(0, 0000)$ and subtract 1 from 0000 we obtain $(-1)1111$ and it is not possible to divide 1111 by 2. So there is no transition simulating φ_1 from the state $(0, 0000)$. The transitions simulating φ_1 are represented in Figure 2.3.

Then we simulate the inverse of φ_2 on each state in a similar way and, since a state $q = (c, s)$ is final if and only if $c = 0$ and $s \in \mathrm{val}_k^{-1}(I)$, the unique final state is $(0, 0000)$. Then the automaton accepting the mirror of $val_2^{-1}(\mathcal{K}_1 - 1)$ is the automaton depicted in Figure 2.4.

Remark 2.3.3. The set $F^\omega(I)$ in the above theorem is k-recognizable and therefore k^n-recognizable for all $n \geq 1$; again, see Proposition 1.3.11. But usually this set is not ultimately periodic and therefore, by Cobham's first Theorem, not ℓ-recognizable for any $\ell \geq 2$ such that k and ℓ are multiplicatively independent. Indeed, if Theorem 2.5.3 described below can be applied, then $F^\omega(I)$ contains arbitrarily large gaps.

Figure 2.3: Transitions simulating φ_1.

Remark 2.3.4. Garth and Gouge [GG07] consider the sequence S_F which is the increasing sequence of the elements in $F^\omega(I)$ in the case where $I = \{1\}$, $k_i = k^{e_i}$, $1 = e_1 \leq e_2 \leq \cdots \leq e_r$, $\ell_1 = 0$ and $-k^{e_i} < \ell_i \leq 0$ for each $i = 1, 2, \ldots, r$. They prove that this sequence reduced modulo a positive integer m is *morphic*. In other words, there exists a morphism f satisfying $f(a) = ax$ for some letter a and some word $x \neq \varepsilon$ such that $S_F \bmod m$ is the image under a coding of the infinite word

$$f^\omega(a) = \lim_{n \to \infty} f^n(a) = axf(x)f^2(x)\cdots,$$

which is a fixed point of f. Moreover, they show that the characteristic sequence of $F^\omega(I)$ is k-automatic.

We explain below that these results follow easily from Theorem 2.3.1 for $F^\omega(I) \subseteq \mathbb{N}$ with arbitrary additive constants ℓ_i and any multiplicative constants k_i which are powers of some k. Namely, as mentioned in the introduction, the set of non-negative integers $F^\omega(I)$ is k-recognizable if and only if its characteristic sequence $(\mathbf{c}_{F^\omega(I)}(n))_{n \geq 0}$ is k-automatic. Note that
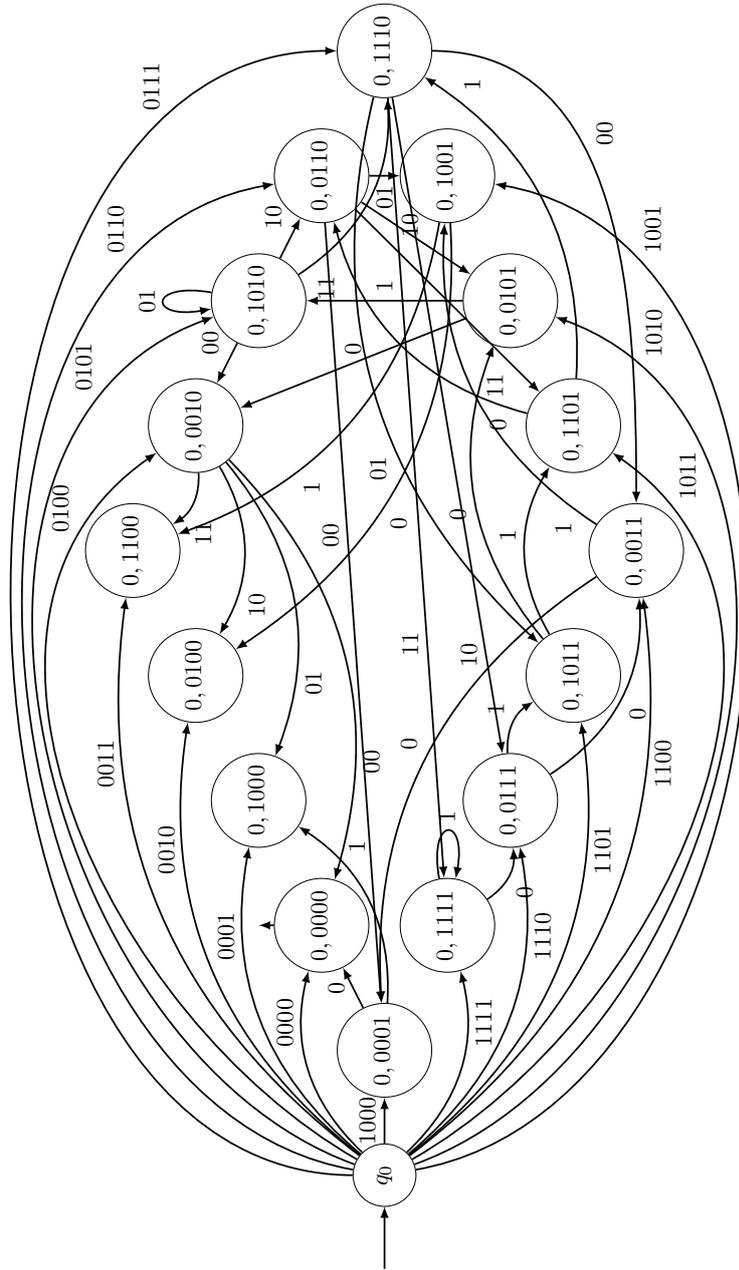
Figure 2.4: A DFA accepting the mirror of $\mathrm{val}_2^{-1}(\mathcal{K}_1 - 1)$.

in the general case $F^\omega(I) \subseteq \mathbb{Z}$ we should consider two-sided k-automatic sequences and two-sided infinite fixed points (see [AS03, Section 5.3 and Section 7.4] for more information) or consider non-negative and non-positive integers separately. In any case, by Lemma 2.2.3, the general case can be reduced to subsets of $\mathbb{N}$.

Hence, let us consider a self-generating set $F^\omega(I) \subseteq \mathbb{N}$ where the multiplicative constants k_i are powers of some k. By Theorem 2.3.1, the characteristic sequence $(\mathbf{c}_{F^\omega(I)}(n))_{n \geq 0}$ is k-automatic. Since $(n \mod m)_{n \geq 0}$ is clearly k-automatic for any $k \geq 2$, then also the sequence

$$\mathbf{u} = ([\mathbf{c}_{F^\omega(I)}(n), n \mod m])_{n \geq 0}$$

over the alphabet $\Sigma_2 \times \Sigma_m$ is k-automatic. Thus, by Cobham's second theorem (Theorem 1.5.8), it is the image under a coding of a fixed point of a k-uniform morphism. Define a morphism $f \colon (\Sigma_2 \times \Sigma_m)^* \to \Sigma_m^*$ by

$$f([a,b]) = \begin{cases} \varepsilon, & \text{if } a = 0; \\ b, & \text{otherwise.} \end{cases}$$

Since the image of a morphic sequence by any morphism is either finite or morphic [Cob68] (see also [AS03, Corollary 7.7.5]) and $(\mathbf{c}_{F^\omega(I)}(n))_{n \geq 0}$ contains infinitely many ones, we conclude that $f(\mathbf{u})$ is morphic. Since $f(\mathbf{u})$ is clearly the sequence S_F reduced modulo m, we have answered an open question in [GG07] by generalizing the result of Garth and Gouge for any additive constants ℓ_i.

Remark 2.3.5. Sequences with missing blocks are considered in [ASS05, GG07, Kim04]. For example, if $\varphi_1 \colon n \mapsto 2n+1, \varphi_2 \colon n \mapsto 4n+2$ and $I = \{0\}$, then the set $F^\omega(I)$ is the set of integers that do not contain the block "00" in their normalized binary expansion. Recall that this set is $\mathcal{K}_1 - 1$; see Example 2.2.5. In [ASS05] the authors ask whether or not the sequences with missing blocks are always particular cases of affinely recursive sets. Observe that, if $F^\omega(I)$ is a sequence with missing blocks, then all constants k_i must be multiplicatively dependent. Otherwise, assume that k_1 and k_2 are multiplicatively independent. Consider now the subset $X_i \subseteq F^\omega(I)$ generated from I by only applying the map φ_i. By Theorem 2.3.1, this subset is k_i-recognizable. Consider now the language $0^* \operatorname{rep}_k(X_i)$, where k is multiplicatively independent to k_i. It is known that this language is *right dense* meaning that every word over the alphabet Σ_k appears as a prefix of some word in $0^* \operatorname{rep}_k(X_i)$; for a proof, see [AS03, Lemma 11.1.1]. Hence, it follows that any block of digits over Σ_k is a factor of $\operatorname{rep}_k(n)$ for some integer $n \in X_i$. For any integer $k \geq 2$, either k_1 or k_2 is multiplicatively independent with k,

and therefore the set X_1 or X_2, and consequently also $F^\omega(I)$, cannot be a set of integers that do not have a particular block of digits in their normalized base k representation.

2.3.1 A close inspection of Frougny's normalization transducer

In order to obtain a self-contained proof for Theorem 2.3.1, we may tailor Frougny's normalization transducer for the language

$$0^* I \{0^{e_1-1}\ell_1, \ldots, 0^{e_r-1}\ell_r\}^*$$

and directly conclude that the output language over Σ_k is regular. Next we describe this in more detail. The following construction is needed to prove the result relating $\mathcal{K}_1$ and the infinite Fibonacci word in the next section. By Lemma 2.2.3, it suffices to consider the set $F^\omega(I) \cap \mathbb{N}$.

Let $C \subset \mathbb{Z}$ be a finite input alphabet and let Σ_k be the output alphabet. Denote $m = \max\{|c - a| \mid c \in C, a \in \Sigma_k\}$ and let $\gamma = m/(k - 1)$. Note that by the Euclidean division, for every $s \in \mathbb{Z}$ and $c \in C$, there exists a unique $a \in \Sigma_k$ and $s' \in \mathbb{Z}$ such that $s + c = s'k + a$. Moreover, if $|s| < \gamma$, then $|s'| \leq (|s| + |c - a|)/k < (\gamma + m)/k = \gamma$. This justifies that we may define a finite right subsequential transducer, where the set of states $Q = \{s \in \mathbb{Z} \mid |s| < \gamma\}$ corresponds to possible carries, the initial state is 0 and the set of edges is

$$E = \{s \xrightarrow{c/a} s' \mid s + c = s'k + a\}. \tag{2.9}$$

A *right subsequential* transducer is a transducer that reads the input from right to left and the underlying automaton where only inputs are considered is deterministic. Moreover, we have a partial terminal function $\omega \colon Q \to \Sigma_k^*$ mapping a state $s \geq 0$ onto its normalized representation $\mathrm{rep}_k(s)$. Let $w = c_n c_{n-1} \cdots c_0 \in C^* \setminus 0C^*$ be a representation of an integer $N = \mathrm{val}_{C,k}(w) \geq 0$. If we enter w into the transducer, there is a unique path

$$0 = s_0 \xrightarrow{c_0/a_0} s_1 \xrightarrow{c_1/a_1} s_2 \xrightarrow{c_2/a_2} \cdots \xrightarrow{c_n/a_n} s_{n+1}$$

such that $N = \sum_{i=0}^n c_i k^i = \sum_{i=0}^n a_i k^i + s_{n+1} k^{n+1}$. Hence, $\omega(s_{n+1}) a_n a_{n_1} \cdots a_0$ is the normalized representation in base k of the integer N. This transducer is Frougny's normalization transducer for an input that do not contain leading zeros.

Next we adapt the previous construction to our specific case of self-generating sets. Let $C = I \cup \{0, \ell_1, \ldots, \ell_r\}$ be the input alphabet. We want

to restrict the accepted input to the words $w \in 0^*I\{0^{e_1-1}\ell_1, \ldots, 0^{e_r-1}\ell_r\}^*$ such that $\mathrm{val}_{C,k}(w) \geq 0$. As shown in the proof of Theorem 2.3.1, these words represent exactly the numbers in $F^\omega(I) \cap \mathbb{N}$. Hence, we build a transducer $\mathcal{T}$ such that from each carry state $q \in Q = \{s \in \mathbb{Z} \mid |s| < \gamma\}$ we may read only words of the form $0^{e_i-1}\ell_i$ from right to left, output the corresponding output of Frougny's transducer and end up in some carry state $q' \in Q$. This can be achieved by introducing chains of intermediate states where each state has only one incoming and outgoing edge simulating the behavior of Frougny's transducer. For example, assume that $k = 2$, $q = 1$ and we want to read 003 from right to left. This corresponds to the map $\varphi\colon n \mapsto 8n + 3$. By the construction, in our modified transducer there is a path

$$1 \xrightarrow{3/0} \hat{2} \xrightarrow{0/0} \hat{1} \xrightarrow{0/1} 0,$$

where $\hat{2}$ and $\hat{1}$ are additional intermediate states and the starting state 1 and the ending state 0 belong to the original set Q. From each state $q \in Q$ there are exactly r paths of this kind corresponding to the r maps $\varphi_i \in F$.

In addition, we need transitions corresponding to the initial values I. Thus, for each $q \in Q$ and $a \in I$ such that $q + a \geq 0$ we make a path from q to a unique final state q_f such that it simulates Frougny's transducer with input $0^i a$, where i is the maximum integer satisfying $k^i \leq q + a$. Padding with sufficiently many zeros insures that the carry is 0 after entering the final state q_f. Here, if needed, intermediate states are added as above. Note that since we consider only non-negative elements of $F^\omega(I)$, there is no path from q to the final state q_f for $a \in I$ such that $q + a < 0$. For example, in the case $k = 2$, $q = 1$ and $a = 5$ we have $i = 2$, since $k^2 < q + a = 6 < k^3$, and the path from q to q_f is

$$1 \xrightarrow{5/0} \hat{3} \xrightarrow{0/1} \hat{1} \xrightarrow{0/1} q_f,$$

where $\hat{3}$ and $\hat{1}$ are new intermediate states. There is also a loop from the final state q_f onto itself with input 0 and output 0. This corresponds to allowing leading zeros after the most significant non-zero digit.

By our construction, each path from the initial state 0 to the final state q_f corresponds to reading some word of the language $0^*I\{0^{e_1-1}\ell_1, \ldots, 0^{e_r-1}\ell_r\}^*$. Therefore, the output of such accepted path in our transducer $\mathcal{T}$ corresponds to some normalized representation (with possibly leading zeros) of a number in the self-generating set $F^\omega(I)$. Conversely, the normalized representation of a number in $F^\omega(I)$ padded with sufficiently many zeros corresponds to the input of some accepted path in our transducer $\mathcal{T}$. Therefore, we may forget the input and consider a finite automaton $\mathcal{A}$ where the edges are labelled only with the output. Moreover, let us define that if in $\mathcal{A}$ there is a path

from a state q to the state q_f with a label belonging to 0^*, then the set q is an accepting state. This allows us to accept all normalized representations with an arbitrary number of leading zeros. We may also change the reading direction by turning the arrows and changing the roles of the initial and final states. Of course, the automaton obtained this way does not need to be complete and deterministic, but it can be made complete by adding missing edges which end up in a sink state and it can be made deterministic by the subset construction. Hence, we have constructed in this way a deterministic finite automaton $\mathcal{B}$ which recognizes $0^* \operatorname{rep}_k(F^\omega(I) \cap \mathbb{N})$ and, by Lemma 2.2.3, we conclude that $F^\omega(I)$ is k-recognizable.

2.4 The Kimberling set and the Fibonacci word

In this section we show a result connecting the Kimberling set $\mathcal{K}_1$ considered in Example 2.1.1 and the infinite Fibonacci word. Recall that $\mathcal{K}_1 = F^\omega(I)$, where $F = \{\varphi_0, \varphi_1, \varphi_2\}$, $\varphi_1\colon n \mapsto 2n$, $\varphi_2\colon n \mapsto 4n - 1$ and $I = \{1\}$.

Theorem 2.4.1. *Let S be the increasing sequence of elements of $\mathcal{K}_1$. Omitting the first term, the sequence S reduced modulo 2, is the Fibonacci word $\varphi^\omega(0)$.*

This is the main result in [Kim00] and it is reproved in [ASS05]. Here we give a third proof based on the transducer construction of the previous section and on some technical manipulation of morphisms.

Proof. Let us first build the transducer $\mathcal{T}$ for the set $\mathcal{K}_1 = F^\omega(I)$ as explained at the end of Section 2.3. This transducer and the corresponding reduced automaton $\mathcal{A}$ are illustrated in Figure 2.5. Using the same notation as above, we have $k = 2$, $C = \{1, 0, -1\}$, $m = 2$ and $\gamma = 2$. Since we never reach a carry state 1 from the initial state 0, our set $Q = \{-1, 0, 1\}$ can be reduced to $\{-1, 0\}$. The input 0 corresponds to the map φ_1 and the input $0(-1)$ corresponds to the map φ_2. When we read $0(-1)$ from right to left starting from either state 0 or -1, we introduce an intermediate state $\widehat{-1}$. Namely, for $s = 0$ and $c = -1$, we have $s + c = (-1) \cdot k + 1$ and, for $s = -1$ and $c = -1$, we have $s + c = (-1) \cdot k + 0$. Then from the state $\widehat{-1}$ we must read 0 and, since $-1 + 0 = -1 \cdot k + 1$, we output 1 and end up in $-1 \in Q$. Moreover, we can read the initial value $1 \in I$ starting from any state in Q. For example, there is an edge with label $1/0$ from -1 to q_f, since $-1 + 1 = 0 \cdot k + 0$.

Using standard techniques we may easily build from $\mathcal{A}$ a deterministic automaton $\mathcal{B}$ accepting $0^* \operatorname{rep}_2(\mathcal{K}_1)$ when reading digits from left to right.
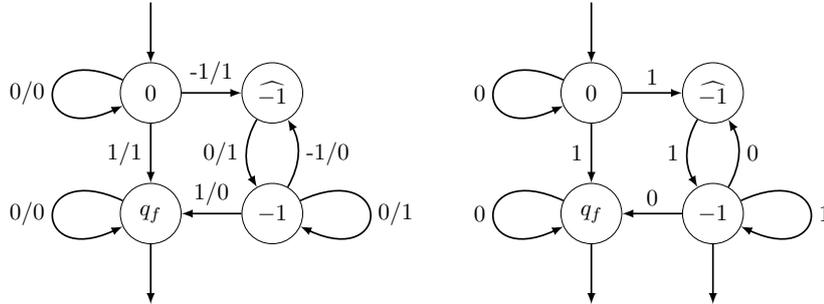
Figure 2.5: Transducer $\mathcal{T}$ and automaton $\mathcal{A}$ corresponding to the Kimberling set.

This automaton is described in Figure 2.6. A number in $\mathcal{K}_1$ such that its binary representation is accepted by b (the corresponding path ends in the final state b) must be odd, since all incoming edges of b are labeled by 1. Similarly, we conclude that a number having a binary representation accepted by c or d must be even. Hence, with an output function $\tau \colon A^* \mapsto \Sigma_2^*$, where A denotes the set of states of $\mathcal{B}$ and

$$\tau(x) = \begin{cases} 1, & \text{if } x = b; \\ 0, & \text{if } x = c \text{ or } x = d; \\ \varepsilon, & \text{otherwise,} \end{cases}$$

the automaton $\mathcal{B}_\tau$ generates the sequence $S \bmod 2$, where S is the increasing sequence of elements of $\mathcal{K}_1$.
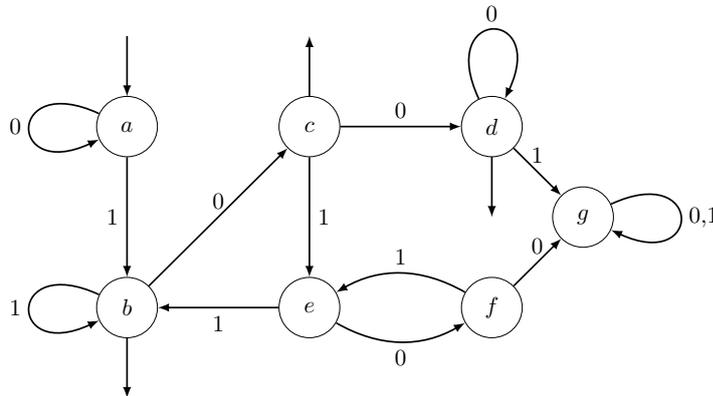
Figure 2.6: A finite deterministic automaton $\mathcal{B}$ accepting $0^* \operatorname{rep}_k(\mathcal{K}_1)$.

The 2-uniform morphism corresponding to $\mathcal{B}$ is $\sigma\colon A^* \to A^*$ defined by

$$a \mapsto ab, \quad b \mapsto cb, \quad c \mapsto de, \quad d \mapsto dg, \quad e \mapsto fb, \quad f \mapsto ge, \quad g \mapsto gg.$$

By the above reasoning, it is clear that $\tau(\sigma^\omega(a)) = S \bmod 2$. Let B be the set $\{a, b, c, d, e, f, g, h\}$. Using the techniques described in the proof of [AS03, Theorem 7.6.1 and Theorem 7.7.4], we obtain a coding $\nu\colon B^* \to \Sigma_2^*$ and a nonerasing morphism $\mu\colon B^* \to B^*$ such that $\nu(\mu^\omega(a)) = \mathcal{S}(\tau(\sigma^\omega(a)))$, where $\mathcal{S}$ is the shift function deleting the first element of the infinite word. The morphism μ is defined by

$$a \mapsto abcdbeb, \quad b \mapsto cdb, \quad c \mapsto fgb, \quad d \mapsto eb, \quad e \mapsto fh,$$
$$f \mapsto f, \quad g \mapsto gbeb, \quad h \mapsto hcdb$$

and

$$\nu(x) = \begin{cases} 1, & \text{if } x = b \text{ or } x = h; \\ 0, & \text{otherwise.} \end{cases}$$

Our aim is to show that $\nu(\mu^\omega(a))$ is the infinite Fibonacci word. For this purpose, let us first simplify the morphism μ. Since $\mu(fgb) = fgbebcdb = \mu(cdb)$, we conclude that $\nu(\mu^i(fgb)) = \nu(\mu^i(cdb))$ for every $i \geq 0$ and, consequently, we may set $\mu(c) = cdb$ without changing $\nu(\mu^\omega(a))$. Similarly, $\mu(fh) = fhcdb = \mu(eb)$ and therefore $\nu(\mu^i(e)) = \nu(\mu^i(d))$ for $i \geq 0$. Thus, we may set $e = d$ and replace the morphism μ by a simpler morphism on a four-letter alphabet $\{a, b, c, d\}$:

$$a \mapsto abcdbdb, \quad b \mapsto cdb, \quad c \mapsto cdb, \quad d \mapsto db.$$

Note that b and c have a different role with respect to the coding, i.e., $\nu(b) \neq \nu(c)$. Since b is always preceded by d except at the very beginning, we finally redefine the morphism $\mu\colon \{a, b, c, d\}^* \to \{a, b, c, d\}^*$ by

$$a \mapsto abcdbdbc, \quad b \mapsto db, \quad c \mapsto cdb, \quad d \mapsto dbc.$$

Hence, the sequence obtained by reducing S modulo 2 and omitting the first element can be obtained as the image of a coding ν of the fixed point $\mu^\omega(a)$.

Let us next modify the morphism generating the Fibonacci word. First, note that we may replace φ by φ^2, since clearly

$$\lim_{n\to\infty} \varphi^n(0) = \lim_{n\to\infty} (\varphi^2)^n(0).$$

Since $\varphi^2(0) = 010$ and $\varphi^2(1) = 01$, we notice that there are two types of zeros in the Fibonacci word: those followed by 0 will be denoted by c and

those followed by 1 will be denoted by d. Let us also replace every 1 by b. Hence, we have $\varphi^\omega(0) = \nu(\phi^\omega(d))$, where ν is the coding defined above and $\phi\colon \{b,c,d\}^* \to \{b,c,d\}^*$ is a morphism such that

$$b \mapsto db, \quad c \mapsto dbc, \quad d \mapsto dbc.$$

We denote

$$(f_n)_{n\geq 0} = \phi^\omega(d) = dbcdb\cdots$$

and

$$(s_n)_{n\geq 0} = \mu^\omega(a) = abcdbdbcdb\cdots.$$

In order to prove the result of Kimberling, we have to show that $\nu(\phi^\omega(d)) = \nu(\mu^\omega(a))$. Since $\nu(f_0) = \nu(d) = 0 = \nu(a) = \nu(s_0)$, it suffices to show that $f_n = s_n$ for all $n \geq 1$. We do this by induction.

First observe that if $s_n = f_n$ for all $n = 1, 2, \ldots, k$, then

$$|\mu(s_0 \cdots s_k)| = |\phi(f_0 \cdots f_k)| + 5. \tag{2.10}$$

This holds because $|\mu(x)|_y = |\phi(x)|_y$ for every x and y in $\{b,c,d\}$ and $|\mu(s_0)| = |\mu(a)| = |\phi(f_0)| + 5$. Here $|w|_y$ denotes the number of letters y occurring in the word w.

Now assume that $s_n = f_n$ for $1 \leq n \leq l$ and l is such that $\phi(f_0 \cdots f_k) = f_0 f_1 \cdots f_l$ for some $k > 1$ satisfying $f_k = b$. This implies that $\phi(f_0 \cdots f_k) = u f_{l-1} f_l = udb$ and, by (2.10) and by the assumption, we have

$$\mu(s_0 \cdots s_k) = udbs_{l+1}s_{l+2}s_{l+3}s_{l+4}s_{l+5} = udb \cdot dbc \cdot db, \tag{2.11}$$

where $s_{l+4}s_{l+5} = \mu(s_k) = \mu(b)$ and $s_{l+1}s_{l+2}s_{l+3} = \mu(s_{k-1}) = \mu(d)$, since $s_k = b$ must be preceded by d if $k > 1$. We have two possibilities, either $f_{k+1}f_{k+2} = db$ or $f_{k+1}f_{k+2}f_{k+3} = cdb$.

If $f_{k+1}f_{k+2} = db$, then $\phi(f_0 \cdots f_{k+2}) = udb\phi(f_{k+1})\phi(f_{k+2}) = udb \cdot dbc \cdot db$ and, by comparing this to (2.11), we conclude that the claim $s_n = f_n$ holds for $1 \leq n \leq l + 5$.

Assume next that $f_{k+1}f_{k+2}f_{k+3} = cdb$. Now $f_1 \cdots f_{k+3} = s_1 \cdots s_{k+3}$, since we must have $k + 3 \leq l$. Hence, we obtain

$$\begin{aligned}
\phi(f_0 \cdots f_{k+3}) &= udb \cdot dbc \cdot dbc \cdot db, \\
\mu(s_0 \cdots s_{k+3}) &= udb \cdot dbc \cdot db \cdot cdb \cdot dbc \cdot db,
\end{aligned}$$

which implies that $s_n = f_n$ for $1 \leq n \leq l + 8$.

Since in the first case $f_{k+2} = b$ and in the second case $f_{k+3} = b$, we may proceed by induction. This concludes the proof, since the claim clearly holds for small values of $n \geq 1$. $\qquad\square$

2.5 The Multiplicatively Independent Case

In this section our aim is to show that $F^\omega(I) \subseteq \mathbb{N}$ given in Definition 2.1.7 is not recognizable in any base $k \geq 2$ provided that $\sum_{i=1}^r k_i^{-1} < 1$ and that there are at least two multiplicatively independent coefficients k_i. For the proof, we introduce the following notation. Let $X = \{x_0 < x_1 < x_2 < \cdots\}$ be an infinite ordered subset of $\mathbb{N}$. Then we denote

$$R_X = \limsup_{i \to \infty} \frac{x_{i+1}}{x_i} \text{ and } D_X = \limsup_{i \to \infty} (x_{i+1} - x_i),$$

where for a sequence $(u_i)_{i \in \mathbb{N}}$, $\limsup_{i \to \infty} u_i = \lim_{i \to \infty} \sup\{u_j : j \geq i\}$. In order to prove that a set is not k-recognizable for any base $k \geq 2$, we use the following result from [Cob72], see also Eilenberg's book [Eil74, Chapter V, Theorem 5.4].

Theorem 2.5.1 (Gap Theorem)**.** *Let $k \geq 2$. If X is a k-recognizable infinite subset of $\mathbb{N}$, then either $R_X > 1$ or $D_X < \infty$.*

Example 2.5.2. Consider the set $X = \{n^2 : n \in \mathbb{N}\}$. We have

$$D_X = \limsup_{n \to \infty} ((n+1)^2 - n^2) = \infty$$

and

$$R_X = \limsup_{n \to \infty} \frac{(n+1)^2}{n^2} = 1.$$

Then, by Theorem 2.5.1, X is never k-recognizable.

Note that $D_X < \infty$ means that X is *syndetic*, i.e., there exists a constant C such that the gap $x_{i+1} - x_i$ between any two consecutive elements x_i, x_{i+1} in X is bounded by C. Let us first show that if $\sum_{i=1}^r k_i^{-1} < 1$, then the set $F^\omega(I)$ given in Definition 2.1.7 contains arbitrarily large gaps.

Theorem 2.5.3. *Let $X = F^\omega(I)$ be a self-generating subset of $\mathbb{N}$ given in Definition 2.1.7. If $\sum_{i=1}^r k_i^{-1} < 1$, then X is not syndetic.*

Proof. Let $n \geq 1$ and $K = k_1 k_2 \cdots k_r$. Let

$$g = g_1 \circ g_2 \circ \cdots \circ g_n \tag{2.12}$$

be a composite function, where g_j belongs to $\{\varphi_1, \varphi_2, \ldots, \varphi_r\}$ for every $j = 1, 2, \ldots, n$ and n_i is the number of indices $j \in \{1, \ldots, n\}$ such that

$g_j = \varphi_i$ in (2.12). Note that $n_1 + n_2 \cdots + n_r = n$. By definition, we have $g(x) = k_1^{n_1} k_2^{n_2} \cdots k_r^{n_r} x + c_g$, where c_g is some constant depending on g. Since $k_1^{n_1} k_2^{n_2} \cdots k_r^{n_r}$ divides K^n, we get

$$\#\{g(x) \mod K^n \mid x \in \mathbb{Z}\} = k_1^{n-n_1} k_2^{n-n_2} \cdots k_r^{n-n_r}.$$

The set $F^n(I)$ contains exactly the integers obtained by at most n applications of maps in F. For any interval of integers $[\![N, N + K^n - 1]\!]$ where $N > \max F^n(I)$, the elements of X belonging to this interval have been obtained by applying at least $n+1$ maps. Hence, in the interval $[\![N, N + K^n - 1]\!]$ there can be at most $k_1^{n-n_1} k_2^{n-n_2} \cdots k_r^{n-n_r}$ integers $x \in X$ such that the last n maps which produce x correspond to the composite function g, i.e., such that there exists $y \in X$ satisfying $g(y) = x$. For fixed numbers n_i, $i = 1, 2, \ldots, r$, there are $n!/(n_1! n_2! \cdots n_r!)$ functions g of the type described above. Thus, the number of integers in $X \cap [\![N, N + K^n - 1]\!]$ for any large enough N is at most

$$\sum_{n_1, n_2, \ldots, n_r} \left(\frac{n!}{n_1! n_2! \cdots n_r!} \right) k_1^{n-n_1} k_2^{n-n_2} \cdots k_r^{n-n_r}$$

$$= K^n \left(\frac{1}{k_1} + \frac{1}{k_2} + \cdots + \frac{1}{k_r} \right)^n.$$

where the sum is over $n_1, n_2, \ldots, n_r \geq 0$ satisfying $n_1 + n_2 + \cdots + n_r = n$. Hence, the biggest gap $x_{i+1} - x_i$ between two consecutive elements $x_i, x_{i+1} \in X$ in the interval $[\![N, N + K^n - 1]\!]$ is at least

$$d(n) = \frac{K^n}{K^n \left(\frac{1}{k_1} + \frac{1}{k_2} + \cdots + \frac{1}{k_r} \right)^n} = \left(\frac{1}{k_1} + \frac{1}{k_2} + \cdots + \frac{1}{k_r} \right)^{-n}.$$

Since $\sum_{i=1}^{r} k_i^{-1} < 1$, the function $d(n)$ tends to infinity as n tends to infinity. This means that there are arbitrarily large gaps in X. In other words, the self-generating set X is not syndetic. $\square$

Before showing that $R_X = 1$ let us first recall a density property of multiplicatively independent integers. A set S is *dense* in an interval $I \subseteq \mathbb{R}$ if every subinterval of I contains an element of S.

The following result is a consequence of Kronecker's theorem, which states that for any irrational number θ the sequence $(\{n\theta\})_{n\geq 0}$ is dense in the interval $[0, 1)$. Here $\{x\}$ denotes the fractional part of the real number x. The proof of Kronecker's theorem as well as the proof of Theorem 2.5.4 can be found in [AS03, Section 2.5] or [HW79].

Theorem 2.5.4. *If $k, \ell \geq 2$ are multiplicatively independent, the set*

$$\{k^p/\ell^q \mid p, q \geq 0\}$$

is dense in $[0, \infty)$.

Proof. We show how to get arbitrarily close to any $x > 0$. Let $\theta = \frac{\log k}{\log \ell} = \log_\ell k$. It is easy to prove that $\log_\ell k$ is rational if and only if k and ℓ are multiplicatively dependant. Then θ is irrational. Let $\alpha = \frac{\log x}{\log \ell}$. By Kronecker's theorem, for all $\varepsilon > 0$ there exist integers $a, c \geq 0$ with $|a\theta - \alpha - c| < \varepsilon$. Hence $|a \log k - \log x - c \log l| < \varepsilon \log \ell$, or $a \log k - c \log \ell \in \;]\log x - \varepsilon \log \ell, \log x + \varepsilon \log \ell[$. Thus $k^a/\ell^c \in]x\ell^{-\varepsilon}, x\ell^\varepsilon[$, which can fit inside any open interval around x by taking ε sufficiently small. $\quad\square$

As an easy consequence of the previous theorem, we obtain the following result.

Corollary 2.5.5. *Let $\alpha > 0$ and β be two real numbers. If k and ℓ are multiplicatively independent, then the set $\{(\alpha k^p + \beta)/\ell^q \mid p, q \geq 0\}$ is dense in $[0, \infty)$.*

Proof. We show how to get arbitrarily close to any positive real number x. Let $\varepsilon > 0$. By Theorem 2.5.4, there exist integers p and q such that

$$\left| \frac{x}{\alpha} - \frac{k^p}{\ell^q} \right| < \frac{\varepsilon}{2\alpha} \quad \text{and} \quad \left| \frac{\beta}{\ell^q} \right| < \frac{\varepsilon}{2}.$$

Hence, it follows that

$$\left| x - \frac{\alpha k^p + \beta}{\ell^q} \right| \leq \left| x - \frac{\alpha k^p}{\ell^q} \right| + \left| \frac{\beta}{\ell^q} \right| < \frac{\varepsilon}{2\alpha}\alpha + \frac{\varepsilon}{2} = \varepsilon.$$

$\quad\square$

Let us next consider the ratio R_X of a self-generating set X.

Theorem 2.5.6. *For any self-generating set $X = F^\omega(I) \subseteq \mathbb{N}$ given in Definition 2.1.7 where k_i and k_j are multiplicatively independent for some i and j, we have $R_X = 1$.*

Proof. Without loss of generality, we may assume that $F = \{\varphi_0, \varphi_1, \varphi_2\}$, where $\varphi_1 \colon n \mapsto k_1\, n + \ell_1$, $\varphi_2 \colon n \mapsto k_2\, n + \ell_2$, and k_1 and k_2 are multiplicatively

independent. Namely, for $F \subseteq F'$, it is obvious that $F^\omega(I) \subseteq F'^\omega(I)$ and consequently, $R_{F^\omega(I)} = 1$ implies $R_{F'^\omega(I)} = 1$. By Lemma 2.2.4, we may also assume that ℓ_1 and ℓ_2 are non-negative.

Let $a \in X$ be a positive integer and set $X_n := X \cap [\varphi_1^{n-1}(a), \varphi_1^n(a)]$ for all $n > 0$. Note that $\cup_{n \in \mathbb{N}} X_n = X \cap [a, \infty)$. Recall that $X = \{x_0 < x_1 < x_2 < \cdots\}$ and define

$$r_n := \max\left\{ \frac{x_{i+1} - x_i}{x_i} \mid x_{i+1}, x_i \in X_n \right\}.$$

Note that, for all x and for $j = 1, 2$, if we set $b_j := \ell_j/(k_j - 1)$, then we have

$$\varphi_j^n(x) = k_j^n x + \ell_j \sum_{i=0}^{n-1} k_j^i = (x + b_j) k_j^n - b_j. \tag{2.13}$$

Let $m \geq 0$ and x_i, x_{i+1} be two consecutive elements belonging to the set X_m. By Corollary 2.5.5, there exist infinitely many positive integers p and q such that $\frac{\varphi_2^p(a)}{k_1^q}$ is equal to

$$\frac{(a + b_2)k_2^p - b_2}{k_1^q} \in \left[x_{i+1} + b_1 - \frac{3}{4}(x_{i+1} - x_i), x_i + b_1 + \frac{3}{4}(x_{i+1} - x_i)\right].$$

Therefore $\varphi_2^p(a)$ is an element of X belonging to the interval

$$[c, d] := \left[k_1^q(x_{i+1} + b_1) - \frac{3}{4}k_1^q(x_{i+1} - x_i), k_1^q(x_i + b_1) + \frac{3}{4}k_1^q(x_{i+1} - x_i)\right],$$

which is a sub-interval[1] of the interval $[\varphi_1^q(x_i), \varphi_1^q(x_{i+1})]$. In other words, we have

$$\varphi_1^q(x_i) < c < \varphi_2^p(a) < d < \varphi_1^q(x_{i+1}).$$

Hence, for all $t > q$, the difference $x_{j+1} - x_j$ of any two consecutive elements x_j, x_{j+1} of X in the interval $[\varphi_1^t(x_i), \varphi_1^t(x_{i+1})]$ is at most

$$\max\{\varphi_1^{t-q}(\varphi_1^q(x_{i+1})) - \varphi_1^{t-q}(\varphi_2^p(a)), \varphi_1^{t-q}(\varphi_2^p(a)) - \varphi_1^{t-q}(\varphi_1^q(x_i))\}$$

$$\leq \max\{\varphi_1^t(x_{i+1}) - \varphi_1^{t-q}(c), \varphi_1^{t-q}(d) - \varphi_1^t(x_i)\} = \frac{3}{4}k_1^t(x_{i+1} - x_i) + b_1 k_1^{t-q}.$$

Thus, the ratio $(x_{j+1} - x_j)/x_j$ is at most

$$\frac{3\,k_1^t(x_{i+1} - x_i)}{4\,\varphi_1^t(x_i)} + \frac{b_1 k_1^{t-q}}{\varphi_1^t(x_i)} = \frac{3\,k_1^t(x_{i+1} - x_i)}{4\,\varphi_1^t(x_i)} + \frac{1}{k_1^q}\frac{b_1 k_1^t}{(x_i + b_1)k_1^t - b_1}. \tag{2.14}$$

---

[1] $c - \varphi_1^q(x_i) = \frac{1}{4}k_1^q(x_{i+1} - x_i) + b_1$ and $\varphi_1^q(x_{i+1}) - d = \frac{1}{4}k_1^q(x_{i+1} - x_i) - b_1$ which is positive for large enough q.

The latter term in this sum can be taken as small as possible for q and t large enough ($1/k_1^q$ tends to 0 and the other factor tends to the constant $b_1/(x_i + b_1)$). In particular, for q and t large enough, we have

$$\frac{b_1 k_1^{t-q}}{\varphi_1^t(x_i)} < \frac{x_{i+1} - x_i}{12 x_i}.$$

Moreover, we have

$$\frac{3 k_1^t (x_{i+1} - x_i)}{4 \varphi_1^t(x_i)} = \frac{3 (x_{i+1} - x_i)}{4 (x_i + b_1 - b_1/k_1^t)} < \frac{3 (x_{i+1} - x_i)}{4 x_i} < \frac{10 (x_{i+1} - x_i)}{12 x_i}.$$

Thus, by (2.14), we obtain

$$\frac{x_{j+1} - x_j}{x_j} < \frac{11 (x_{i+1} - x_i)}{12 x_i}. \tag{2.15}$$

Since the above inequality holds for any consecutive elements x_i and x_{i+1} in X_m and since there are only finitely many such pairs, we conclude that there exists an integer N_1 such that (2.15) holds for any consecutive elements $x_j, x_{j+1} \in X_n$ where $n \geq N_1$. Hence, we obtain $r_n < \frac{11}{12} r_m$ for every $n \geq N_1$. Moreover, by repeating this procedure, we conclude that there exists an integer N_k such that

$$r_n < \left(\frac{11}{12}\right)^k r_m$$

for every $n \geq N_k$. This implies that $\limsup_{n \to \infty} r_n = 0$ and, consequently,

$$R_X = 1 + \limsup_{n \to \infty} r_n = 1.$$

$\square$

Our main result is a straightforward consequence of the previous theorems.

Theorem 2.5.7. *Let $X = F^\omega(I) \subseteq \mathbb{N}$ be a self-generating set as in Definition 2.1.7. If $\sum_{t=1}^r k_t^{-1} < 1$ and there exist i, j such that k_i and k_j are multiplicatively independent, then $F^\omega(I)$ is not k-recognizable for any integer base $k \geq 2$.*

Proof. Let $X = F^\omega(I)$ satisfy the assumptions of the theorem. By Theorem 2.5.3, we have $D_X = \infty$ and, by Theorem 2.5.6, we have $R_X = 1$. Thus, Theorem 2.5.1 implies that X is not k-recognizable for any $k \geq 2$. $\square$

As a corollary, we have solved Conjecture 1 presented in [ASS05].

Corollary 2.5.8. *Let $F = \{\varphi_0, n \mapsto k_1 \, n + \ell_1, n \mapsto k_2 \, n + \ell_2\}$, where k_1 and k_2 are multiplicatively independent. Then any infinite self-generating set $F^\omega(I)$ given in Definition 2.1.7 is not k-recognizable for any $k \geq 2$.*

Proof. This follows directly from Theorem 2.5.7. Namely, if k_1 and k_2 are multiplicatively independent, then $k_1 \geq 2$ and $k_2 \geq 3$ and $k_1^{-1} + k_2^{-1} \leq 1/2 + 1/3 = 5/6 < 1$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

The condition $\sum_{t=1}^{r} k_t^{-1} < 1$ is not needed in a very special case of self-generating sets where $\ell_i = 0$ for every $i = 1, 2, \cdots, r$. This situation is related to the so-called y-smooth numbers.

Definition 2.5.9. An integer is *y-smooth* if it has no prime factors greater than y.

The analysis of many number theoretic algorithms turns on the role played by integers which have only small prime factors. To be able to determine which algorithm is faster than the others, it turns out to be important to have accurate estimates for the number of smooth numbers in various sequences. A survey of important estimates has been written by Granville [Gra08]. Smooth numbers also appear in other contexts: for example, integers that we know we can factorize in polynomial time are typically smooth, or a smooth number times a prime.

Theorem 2.5.10. *Let $X = F^\omega(I)$ be as in Definition 2.1.7. If $\ell_i = 0$ for every $i = 1, 2, \cdots, r$ and if there exist i, j such that k_i and k_j are multiplicatively independent, then $F^\omega(I)$ is not k-recognizable for any integer base $k \geq 2$. In particular, for $y \geq 3$, the set of y-smooth numbers is not k-recognizable for any $k \geq 2$.*

Proof. Assume that $\varphi_i \colon n \mapsto k_i n$ for $i = 1, 2, \ldots, r$ and denote $X = F^\omega(I)$. Let $x \geq 2$ be an integer and consider $n \in X \cap [\![0, x]\!]$. By the definition of X, the integer n must be of the form $k_1^{e_1} \cdots k_r^{e_r} a$, where $a \in I$. Since the exponent e_i is at most $\log_2(x)$ for every $i = 1, 2, \ldots, r$, the number of integers in $X \cap [\![0, x]\!]$ is at most $(1 + \log_2(x))^r \# I = O(\log^r(x))$. It follows that $x / \#(X \cap [\![0, x]\!])$ tends to infinity when x tends to infinity. This implies that $F^\omega(I)$ cannot be syndetic, i.e., $D_X = \infty$. If there are two multiplicatively independent constants k_1 and k_2, then $R_X = 1$ by Theorem 2.5.6. Hence,

by Theorem 2.5.1, the self-generating set X is not k-recognizable for any $k \geq 2$. The second claim follows, since the set of y-smooth numbers can be represented as a self-generating set $F^\omega(I)$, where $I = \{1\}$ and $\varphi_i \colon n \mapsto p_i n$ for $i = 1, 2, \ldots, r$. Here p_i is the ith smallest prime and p_r is the largest prime less than or equal to y. $\qquad\square$

Chapter 3

Syntactic complexity

3.1 Introduction

In this chapter, we estimate the syntactic complexity, i.e., the cardinality of the syntactic monoid, of the language $0^* \operatorname{rep}_k(m\mathbb{N})$ consisting of base k-representations of the multiples of the integer m. Next, we study the syntactic complexity of any (ultimately) periodic set of integers written in base k. We apply our results to a special case of the following decision problem. Hence, syntactic complexity could possibly serve as a new tool to deal with this general problem.

Problem 1. Given

- an abstract numeration system S,

- a DFA accepting some S-recognizable set $X \subseteq \mathbb{N}$.

Decide whether or not the set X is ultimately periodic.

The material of this chapter can be found in [RV11] where partial results were given and in [LRRV12] where a complete decision procedure when S is the usual k-ary system is developed.

A typical technique to decide if a k-recognizable set X is ultimately periodic is the following.

- Prove that if a set $X \subseteq \mathbb{N}$ is periodic of period P (resp. of preperiod I), then
$$\#\mathcal{A}_X \geq f(P) \qquad (\text{resp. } \#\mathcal{A}_X \geq g(I))$$

with $\lim_{x\to\infty} f(x) = \infty$ (resp. $\lim_{x\to\infty} g(x) = \infty$) and where $\#\mathcal{A}_X$ is the number of states of the minimal automaton recognizing X.

- These lower bounds on the state complexity of X give some upper bounds on the possible period and preperiod of X. Then, as we know $\mathcal{A}_X$, there is only a finite number of possible ultimately periodic sets Y to compare with X. In order to test whether $X = Y$, we construct the language $L' = \mathrm{rep}_k(Y)$. Since

$$\mathrm{val}_k(L) = \mathrm{val}_k(L') \Leftrightarrow 0^*((0^*)^{-1}.L) = 0^*((0^*)^{-1}.L'),$$

  we only have to test if $0^*((0^*)^{-1}.\mathrm{rep}_k(X)) = 0^*((0^*)^{-1}.L')$.

This technique was used by Honkala [Hon86] to settle Problem 1 positively for integer base numeration systems. In particular, the following result of Alexeev [Ale04] gives exactly the state complexity of the language $0^*\mathrm{rep}_k(m\mathbb{N})$. As usual (m, n) denotes the greatest common divisor of m and n.

Theorem 3.1.1 ([Ale04])**.** *Let $k, m \geq 2$ be integers. Let N, M be such that $k^N < m \leq k^{N+1}$ and $(m, 1) < (m, k) < \cdots < (m, k^M) = (m, k^{M+1}) = (m, k^{M+2}) = \cdots$. The minimal automaton of $0^*\mathrm{rep}_k(m\mathbb{N})$ has exactly*

$$\frac{m}{(m, k^{N+1})} + \sum_{t=0}^{\inf\{N, M-1\}} \frac{k^t}{(m, k^t)} \; states.$$

Recall that a sequence $(x_n)_{n\geq 0}$ is ultimately periodic if there exist integers $P \geq 1$ and $I \geq 0$ such that $x_n = x_{n+P}$ for all $n \geq I$. Allouche, Rampersad and Shallit [ARS09] also answered the integer base k decision problem by constructing an NDFA accepting a pair (I, P) of preperiod and period if and only if $\mathbf{c}_X(n) = \mathbf{c}_X(n + P)$ for all $n \geq I$, where $\mathbf{c}_X$ is the characteristic sequence of X.

Using Büchi's theorem [Büc60], a first order logic approach to this problem can be considered, see for instance [BHMV94] when the base is an integer or [BH97] when the base is a Pisot number. Indeed, they proved that a set X is k-recognizable if and only if X is k-definable, i.e., if there exists a formula φ in $\langle \mathbb{N}, +, V_k \rangle$ such that

$$\varphi(n) \text{ is true } \Leftrightarrow n \in X.$$

Then X is ultimately periodic if the formula

$$\psi \equiv \exists P \exists I \forall n \geq I, \; \varphi(n) \Leftrightarrow \varphi(n + P)$$

is true. For instance, Bruyère *et al.* show how to construct an automaton $\mathcal{A}_\psi$ accepting the set T_ψ of pairs (I, P) such that ψ is true. The emptiness of the set T_ψ is decidable.

Very recently, giving a structural description of minimal automata accepting an ultimately periodic set of integers, Marsault and Sakarovitch [MS13] gave a $O(n \log n)$ decision procedure to this problem in the integer base case.

This decision problem was settled positively in [BCFR09], using the same kind of decision procedure as Honkala, for linear numeration systems $U = (U_i)_{i \geq 0}$ satisfying some weak assumptions: $\lim_{i \to +\infty} U_{i+1} - U_i = +\infty$, $\mathbb{N}$ is U-recognizable and $\lim_{m \to +\infty} N_U(m) = +\infty$, where $N_U(m) \in \{1, \ldots, m\}$ denotes the number of values that occur infinitely often in the sequence $(U_i \mod m)_{i \geq 0}$.

Considering this decision problem for any abstract numeration system turns out to be equivalent to the so-called ω-HD0L ultimate periodicity decision problem [BR10, HR04]:

Problem 2. Given a coding and a morphism generating an infinite word w. Decide whether or not w is ultimately periodic.

For instance, the pure morphic case, i.e., the ω-D0L ultimate periodicity decision problem, was solved by Harju and Linna [HL86] and independently by Pansiot [Pan86]. Relying on combinatorics on words techniques, the general case has very recently received a positive answer thanks to Durand [Dur12, Dur11] and Mitrofanov [Mit11] independantly.

In the framework of numeration systems, syntactic complexity has an advantage in comparison to left or right quotients: we have the opportunity to work simultaneously on prefixes and suffixes of base k-expansions, that is on most and least significant digits. See for instance [BY11] for some background. Since syntactic complexity provides an alternative measure for the complexity of a regular language, one could try to develop new decision procedures based on the syntactic complexity instead of the state complexity of the corresponding languages. A step in that direction is to first consider integer base numeration systems. As a consequence of our results, we present such a procedure in Section 3.5, using a similar approach as in [Hon86].

In the next section, we recall basic definitions, fix notation and discuss the case of n-definite languages. In Section 3.3, we compute the syntactic complexity of the set $m\mathbb{N}$ written in base k, i.e., the cardinality $M_{k,m}$ of the syntactic monoid of the language $0^* \mathrm{rep}_k(m\mathbb{N})$. For the binary system, the first few values of $M_{2,m}$ are given in Table 3.1. Let $k \geq 2$. An explicit formula for $M_{k,m}$ is obtained as a consequence of Theorems 3.3.1, 3.3.4 and 3.3.8 for

m		2	3	4	5	6	7	8	9	10	11	12
$M_{2,m}$		3	6	5	20	13	21	7	54	41	110	20

m	13	14	15	16	17	18	19	20	21	22	23	24
$M_{2,m}$	156	43	60	9	136	109	342	62	126	221	253	27

m	25	26	27	28	29	30	31	32	33	34	35	36
$M_{2,m}$	500	313	486	65	812	121	155	11	330	273	420	164

Table 3.1: The first values of the syntactic complexity in the base 2 case.

the following three cases:

- the constant m and the base k are coprime

- m is a power of k

- $m = k^n q$ with $(q,k) = 1$, $q \geq 2$ and $n \geq 1$.

Next, we provide a lower bound on the number of infinite classes of the syntactic monoid of a periodic set of period $m = k^n q$ with $(q,k) = 1$. In Section 3.4, we provide such a lower bound in the general case where the period m is of the kind $dk^n q$ with $(q,k) = 1$ and the set of prime factors of d is a subset of that of k.

We end the chapter with a procedure for the decision problem described above.

3.2   Definitions and basic properties

3.2.1   Periodic sets

Recall that a set $X \subseteq \mathbb{N}$ is *periodic* of *period* p if for all $n \in \mathbb{N}$, $n \in X \Leftrightarrow n + p \in X$. The period is always understood to be the minimal period of X. In particular, if $X \subseteq \mathbb{N}$ is periodic of period p, then for all $i, j \in \mathbb{N}$,

$$i \not\equiv j \bmod p \Rightarrow \exists r \in [\![0, p-1]\!] : (i+r \in X, j+r \notin X) \text{ or } (i+r \notin X, j+r \in X). \tag{3.1}$$

A set $X \subseteq \mathbb{N}$ is *ultimately periodic* of *period* p and *preperiod* $I > 0$ if, for all $n \geq I$, $n \in X \Leftrightarrow n+p \in X$ and exactly one of the two elements $I-1, I+p-1$ is in X. Again, preperiod and period are always understood to be minimal.

3.2.2 Syntactic complexity and n-definite languages

In the theory of regular languages, there are two fundamental congruences. The first one has already been introduced in Chapter 1, Definition 1.2.11.

Definition 3.2.1. Let L be a language over the finite alphabet Σ. The *Nerode congruence* of L is defined as follows. Let $u, v \in \Sigma^*$. We have

$$
\begin{aligned}
u \sim_L v \quad &\Leftrightarrow \quad (ux \in L \Leftrightarrow vx \in L, \quad \forall x \in \Sigma^*) \\
&\Leftrightarrow \quad u^{-1}.L = v^{-1}.L.
\end{aligned}
$$

Each equivalence class is the set of words that reach the same state from the initial state in the minimal automaton of L. The number of these classes is equal to the number of states of $\mathcal{A}_L$ which is exactly the *state complexity* of the language L.

Definition 3.2.2. Let L be a language over the finite alphabet Σ. The *context* of a word $u \in \Sigma^*$ with respect to L is given by the set of pairs

$$
\mathsf{C}_L(u) = \{(x, y) \in \Sigma^* \times \Sigma^* \mid xuy \in L\}.
$$

If L is clearly understood, we will simply write $\mathsf{C}(u)$.

The *Myhill congruence* (or *syntactic congruence*) of L is defined as follows. Let $x, y \in \Sigma^*$. We set

$$
\begin{aligned}
x \leftrightarrow_L y \quad &\Leftrightarrow \quad (uxv \in L \Leftrightarrow uyv \in L \quad \forall u, v \in \Sigma^*) \\
&\Leftrightarrow \quad \mathsf{C}_L(x) = \mathsf{C}_L(y).
\end{aligned}
$$

As stated by the following proposition, each equivalence class is the set of words that perform the same transformation on the set of states of the minimal automaton of L.

Proposition 3.2.3. *Let L be a language over Σ. Two words $u, v \in \Sigma^*$ are such that $u \leftrightarrow_L v$ if and only if they perform the same transformation on the set of states of the minimal automaton $\mathcal{A}_L = (Q_L, q_{0,L}, F_L, \Sigma, \delta_L)$ of L, i.e., for all $r \in Q_L$, $\delta_L(r, u) = \delta_L(r, v)$. In particular, if u, v are such that $\delta_L(q_{0,L}, u) \neq \delta_L(q_{0,L}, v)$, then $u \not\leftrightarrow_L v$.*

Definition 3.2.4. Let $Q_L = \{q_1, \ldots, q_n\}$ be the set of states of the minimal automaton of L. The *transformation automaton* of L has $(Q_L)^n$ as set of states, $(q_1, \ldots, q_n)$ as initial state and for all $w \in \Sigma^*$, $(r_1, \ldots, r_n) \in (Q_L)^n$,

its transition function $\tau : (Q_L)^n \times \Sigma^* \to (Q_L)^n$ is given by $\tau((r_1, \ldots, r_n), w) = (\delta_L(r_1, w), \ldots, \delta_L(r_n, w))$. We will only consider the accessible part of this automaton. Thanks to Proposition 3.2.3,

$$u \leftrightarrow_L v \text{ if and only if } \tau((q_1, \ldots, q_n), u) = \tau((q_1, \ldots, q_n), v).$$

Example 3.2.5. Consider the set $X = 3\mathbb{N}$ in base 2. The minimal automaton of $L = 0^* \operatorname{rep}_2(X)$ is depicted in Figure 3.1 and its transformation automaton is given in Figure 3.2.
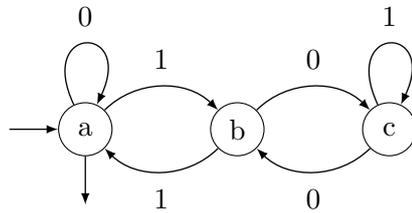
Figure 3.1: The minimal automaton of $0^* \operatorname{rep}_2(3\mathbb{N})$.
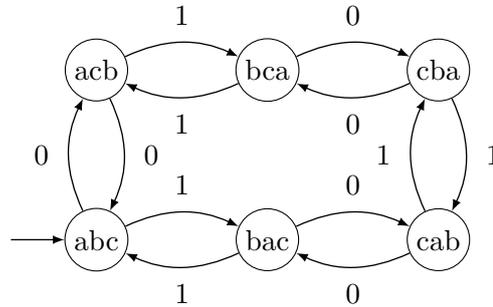
Figure 3.2: The transformation automaton of $0^* \operatorname{rep}_2(3\mathbb{N})$.

Remark 3.2.6. Note that if the transformation automaton of L has a strongly connected component containing l states, we can find l infinite equivalence classes for $\leftrightarrow_L$. Indeed, if there is a path labelled by $u \in \Sigma^*$ from the initial state $(q_1, \ldots, q_n)$ to a state $(r_1, \ldots, r_n)$ belonging to this strongly connected component, there is a cycle on $(r_1, \ldots, r_n)$ labelled by some $v \in \Sigma^+$. Then, if u belongs to an equivalence class for $\leftrightarrow_L$, uv^i belongs to the same class for all $i > 0$ and this class is infinite.

Let X be an ultimately periodic set of preperiod I. This remark ensures that we can find elements greater than I in each class. Then if we find a lower bound on the number of infinite equivalence classes of a periodic set, it is still valid in the ultimately periodic case.

The monoid $\Sigma^*/\leftrightarrow_L$ consisting of the equivalence classes of the relation $\leftrightarrow_L$ is the *syntactic monoid* of L. It is well known that L is a regular language if and only if $\Sigma^*/\leftrightarrow_L$ is finite. The *syntactic complexity* of L is the cardinality of its syntactic monoid. If $X \subseteq \mathbb{N}$ is a k-recognizable set of integers, by extension we define the *syntactic complexity* of X (w.r.t. k) as the syntactic complexity of the language $0^* \operatorname{rep}_k(X)$.

Example 3.2.7. Consider the automata depicted in Figure 3.3. These two automata have the same alphabet, are minimal and have the same state complexity. However, the syntactic complexity of $L(\mathcal{A}_1)$ is 3 and that of $L(\mathcal{A}_2)$ is 9 and the multiplication table of the syntactic monoid of $L(\mathcal{A}_1)$ and $L(\mathcal{A}_2)$ are given in Table 3.2 and Table 3.3 respectively where each equivalence class is given by one of its representatives.
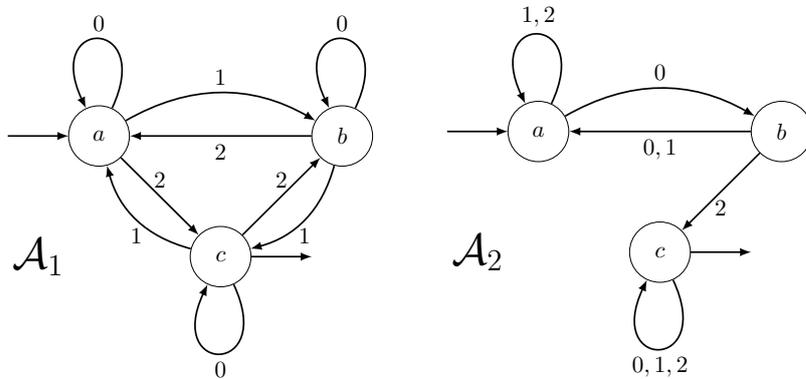

Figure 3.3: Automata $\mathcal{A}_1$ and $\mathcal{A}_2$.

	ε	1	2
ε	ε	1	2
1	1	2	ε
2	2	ε	1

Table 3.2: The multiplication table of the syntactic monoid of $L(\mathcal{A}_1)$.

	ε	0	1	2	02	10	20	020	102
ε	ε	0	1	2	02	10	20	020	102
0	0	ε	1	02	2	10	020	20	102
1	1	10	1	1	102	10	10	102	102
2	2	20	1	2	102	20	20	102	102
02	02	020	02	02	102	020	020	102	102
10	10	1	102	102	1	10	102	10	102
20	20	2	2	102	2	20	102	20	102
020	020	02	02	102	02	020	102	020	102
102	102	102	102	102	102	102	102	102	102

Table 3.3: The multiplication table of the syntactic monoid of $L(\mathcal{A}_2)$.

As shown by Example 3.2.7, two languages can have same state complexity but different syntactic complexity. Then the syntactic complexity gives an alternative measure to the state complexity. The following result gives a link between these two quantities.

Proposition 3.2.8. *For each language $L \subseteq \Sigma^*$ of state complexity $c > 1$, we have*

$$c \leq \mu(L) \leq c^c$$

where $\mu(L)$ is the syntactic complexity of L.

We now give basic definitions and properties about the so-called n-definite languages. This kind of languages will be of particular interest in estimating the syntactic complexity of periodic sets when the period is a power of the base.

Definition 3.2.9. A language $L \subseteq \Sigma^*$ is *weakly n-definite* , if for any $x, y \in \Sigma^*$ satisfying $|x| \geq n$, $|y| \geq n$ and having the same suffix of length n, $x \in L$ if and only if $y \in L$. In other words, L can be written as $G \cup \Sigma^* F$ where F (resp. G) is finite and contains only words of length n (resp. less than n). For more about weakly n-definite languages, see [PRS63, Brz63]. Let $n \geq 1$. A language is *n-definite* if it is weakly n-definite and not weakly $(n-1)$-definite. One also finds the terminology *suffix testable* in the literature [Pin97].

Example 3.2.10. Consider the language $L = 0^* \operatorname{rep}_2(4\mathbb{N}) \subseteq \Sigma_2^*$ of base 2-representations of the multiples of 4. This language is weakly 2-definite since $L = \{\varepsilon\} \cup \Sigma_2^*\{00\}$ but is not weakly 1-definite since 100 and 110 have the

same suffix of length 1 but $100 \in L$ and $110 \notin L$. Then, by definition, the language L is 2-definite.

The following proposition is well known and gives a lower bound on the state complexity of a n-definite language.

Proposition 3.2.11 ([PRS63]). *If a language L is n-definite, then its minimal automaton has at least $n + 1$ states .*

The original arguments of Perles, Rabin and Shamir involved in the proof of Proposition 3.2.11 are of the same kind as the one that we will be developing. This is why we have decided to include the proof.

Proof. Let $\mathcal{A}_L = (Q_L, \Sigma, q_{0,L}, F_L, \delta_L)$ the minimal automaton of L such that L is n-definite. We must prove $\#Q_L \geq n + 1$.

If $n = 0$, there is nothing to prove since we always have $\#Q_L \geq 1$.

Therefore, assume $n \geq 1$. Define, for $q, q' \in Q_L$ and $i \geq 0$,

$$E_i(q, q') \Leftrightarrow (\forall x \in \Sigma)[|x| \geq i \Rightarrow (\delta_L(q, x) \in F_L \Leftrightarrow \delta_L(q', x) \in F_L)].$$

$E_i(q, q')$ means that no word x satisfying $|x| \geq i$ distinguishes the states q and q'. Each E_i is an equivalence relation on Q_L and induces a partition P_i of Q_L into $\#P_i$ equivalence classes. The relation $E_i(q, q')$ always implies $E_{i+1}(q, q')$; the partition P_i is therefore a refinement of P_{i+1} hence, $\#P_i \geq \#P_{i+1}$.

Since L is n-definite and $n \geq 1$, L is weakly n-definite but not weakly $(n-1)$-definite. Therefore, there exist three words x, y and y' (y and y' possibly empty) such that $|x| = |x_0 \cdots x_{n-2}| = n - 1$, $yx \in L$ and $y'x \notin L$. Define, for $0 \leq i \leq n - 1$,

$$q_i = \delta_L(q_{0,L}, yx_0 \cdots x_{i-1}) = \delta_L(\delta_L(q_{0,L}, y), x_0 \cdots x_{i-1})$$

and

$$q'_i = \delta_L(q_{0,L}, y'x_0 \cdots x_{i-1}) = \delta_L(\delta_L(q_{0,L}, y'), x_0 \cdots x_{i-1}).$$

Let $z \in \Sigma^*$ such that $|z| \geq n - i$. Then we have

$$\delta_L(q_i, z) = \delta_L(\delta_L(q_{0,L}, yx_0 \cdots x_{i-1}), z) = \delta_L(q_{0,L}, yx_0 \cdots x_{i-1}z)$$

and similarly

$$\delta_L(q'_i, z) = \delta_L(\delta_L(q_{0,L}, y'x_0 \cdots x_{i-1}), z) = \delta_L(q_{0,L}, y'x_0 \cdots x_{i-1}z).$$

Now, $|x_0 \cdots x_{i-1}z| = i + |z| \geq n$. The words $yx_0 \cdots x_{i-1}z$ and $y'x_0 \cdots x_{i-1}z$ agree in their last n symbols. Since L is n-definite, $yx_0 \cdots x_{i-1}z \in L$ if

and only if $y'x_0 \cdots x_{i-1}z \in L$. In other words, $\delta_L(q_i, z) \in F_L$ if and only if $\delta_L(q'_i, z) \in F_L$, i.e.,

$$E_{n-i}(q_i, q'_i). \tag{3.2}$$

On the other hand, since $|x_i \cdots x_{n-2}| = n - 1 - i$, we get

$$
\begin{aligned}
\delta_L(q_i, x_i \cdots x_{n-2}) &= \delta_L(\delta_L(q_{0,L}, yx_0 \cdots x_{i-1}), x_i \cdots x_{n-2}) \\
&= \delta_L(q_{0,L}, yx) \in F_L.
\end{aligned}
$$

In a similar way, it follows from $y'x \notin L$ that

$$\delta_L(q'_i, x_i \cdots x_{n-2}) = \delta_L(q_{0,L}, y'x) \notin F_L.$$

This implies that

$$q \text{ and } q' \text{ do not satisfy } E_{n-i-1}(q, q'). \tag{3.3}$$

Relations (3.2) and (3.3) show that, for $0 \le i \le n - 1$, there exists q, q' such that we have $E_{n-i}(q, q')$ and not $E_{n-i-1}(q, q')$. The partition P_{n-i-1} is therefore a proper refinement of P_{n-i}, which implies $\#P_{n-i} < \#P_{n-i-1}$. Thus, we have

$$1 \le \#P_n < \#P_{n-1} < \cdots < \#P_1 < \#P_0;$$

hence $\#P_0 \ge n + 1$.

Since each equivalence class in the partition P_0 contains at least one state, we finally get

$$\#Q_L \ge \#P_0 \ge n + 1.$$

$$\square$$

In particular, the syntactic monoid of an n-definite language also has at least $n + 1$ elements. We adapt the previous result to get extra information about such a monoid in the case of sets of integers.

Lemma 3.2.12. *Let X be a set of integers. If $L = 0^* \operatorname{rep}_k(X)$ is n-definite, then there exist arbitrarily large integers $t_1, \ldots, t_{n+1}$ such that the $n+1$ words $\operatorname{rep}_k(t_1), \ldots, \operatorname{rep}_k(t_{n+1})$ belong to different equivalence classes of $\leftrightarrow_L$.*

Proof. Note that there exist ℓ words $s_1, \ldots, s_\ell$ of length n such that a word of length at least n belongs to L if and only if it belongs to $\Sigma_k^* \{s_1, \ldots, s_\ell\}$. Define an accessible DFA $\mathcal{A} = (Q, q_\varepsilon, F, \Sigma_k, \delta)$ where

$$Q_n = \{q_w \mid |w| = n\} \text{ and } Q = Q_n \cup \{q_w \mid |w| < n\}$$

and for all $u \in \Sigma_k^*$ such that $|u| < n$ and $a \in \Sigma_k$, we have $\delta(q_u, a) = q_{ua}$. Now if $|u| = n$, then $u = cx$ for some $c \in \Sigma_k$, $|x| = n - 1$ and we have $\delta(q_u, a) = q_{xa}$. Notice that $\mathcal{A}$ restricted to the states in Q_n is a strongly connected component isomorphic to the de Bruijn graph of order n over Σ_k. The set of final states of $\mathcal{A}$ is easily defined in such a way that the language accepted by $\mathcal{A}$ is L. In particular, a state in Q_n is final if and only if it is of the form q_{s_i} for $i \in [\![1, \ell]\!]$.

Let $\mathcal{A}_L = (Q_L, q_{0,L}, F_L, \Sigma_k, \delta_L)$ denotes the minimal automaton of L and let Φ be the canonical morphism of automata $\Phi : Q \to Q_L$ from $\mathcal{A}$ to $\mathcal{A}_L$ such that $\Phi(\delta(r, w)) = \delta_L(\Phi(r), w)$ for all $r \in Q$ and $w \in \Sigma_k^*$. Let $R := \Phi(Q_n)$. In other words, R is the set of states of $\mathcal{A}_L$ reached by words of length at least n. Using the same arguments as in [PRS63], let us show that $\#R \geq n + 1$. For all $r, r' \in R$ and $i \geq 0$, define

$$E_i(r, r') \Leftrightarrow (\forall x \in \Sigma_k^*)\big[|x| \geq i \Rightarrow (\delta_L(r, x) \in F_L \Leftrightarrow \delta_L(r', x) \in F_L)\big].$$

This equivalence relation E_i over R induces a partition P_i of R into $\#P_i$ equivalence classes. It is clear that $E_i(r, r')$ implies $E_{i+1}(r, r')$ and thus $\#P_i \geq \#P_{i+1}$.

Since L is n-definite, there exists a word $V = v_1 \cdots v_{n-1}$ of length $n - 1$ and two symbols $\sigma, \tau \in \Sigma_k$ such that $\sigma V \in L$ and $\tau V \notin L$. Let $T \geq n$ and $i \in [\![0, n - 1]\!]$. Take the two states $r = \Phi(\delta(q_\varepsilon, 10^T \sigma v_1 \cdots v_{n-i-1}))$ and $r' = \Phi(\delta(q_\varepsilon, 10^T \tau v_1 \cdots v_{n-i-1}))$ in R. By considering the word $v_{n-i} \cdots v_{n-1}$ of length i, the states r and r' do not satisfy $E_i(r, r')$ but for all words u of length at least $i + 1$, we have

$$\delta(q_\varepsilon, 10^T \sigma v_1 \cdots v_{n-i-1} u) = q_S = \delta(q_\varepsilon, 10^T \tau v_1 \cdots v_{n-i-1} u)$$

where S is the suffix of length n of $v_1 \cdots v_{n-i-1} u$ and thus $E_{i+1}(r, r')$. We have just shown that E_i is a refinement of E_{i+1} and

$$\#P_0 > \#P_1 > \cdots > \#P_{n-1} > \#P_n \geq 1.$$

Consequently, $\#R \geq \#P_0 \geq n + 1$.

The minimal automaton $\mathcal{A}_L$ of L contains at least $n + 1$ distinct states of the kind $\Phi(q_{u_1}), \ldots, \Phi(q_{u_{n+1}}) \in R$ for some words $u_1, \ldots, u_{n+1} \in \Sigma_k^*$ of length n. Let $I > 0$. Take a large enough T such that, for all $i \in [\![1, n + 1]\!]$, $\mathrm{val}_k(10^T u_i) > I$ and observe that

$$\Phi(\delta(q_\varepsilon, 10^T u_i)) = \Phi(q_{u_i}) = \delta_L(q_{0,L}, 10^T u_i) \in R.$$

The words $10^T u_i$, $i = 1, \ldots, n + 1$, perform pairwise distinct transformations on the set of states of the minimal automaton $\mathcal{A}_L$ and the syntactic monoid of L contains at least $n + 1$ classes (see Proposition 3.2.3). $\qquad\square$

Example 3.2.13. Consider the ultimately periodic set $X = 4\mathbb{N} + 3$ with preperiod 3 and period 4. The language $L = 0^* \operatorname{rep}_2(X)$ is 2-definite since we have

$$w \in L \Leftrightarrow w \in \Sigma_2^*\{11\}.$$

With the notation of the previous proof, we have $n = 2$ and the automaton $\mathcal{A} = (Q, q_\varepsilon, F, \Sigma_2, \delta)$ with $Q_2 = \{q_{00}, q_{01}, q_{10}, q_{11}\}$, $Q = Q_2 \cup \{q_\varepsilon, q_0, q_1\}$ and $F = \{q_{11}\}$ is represented in Figure 3.4.
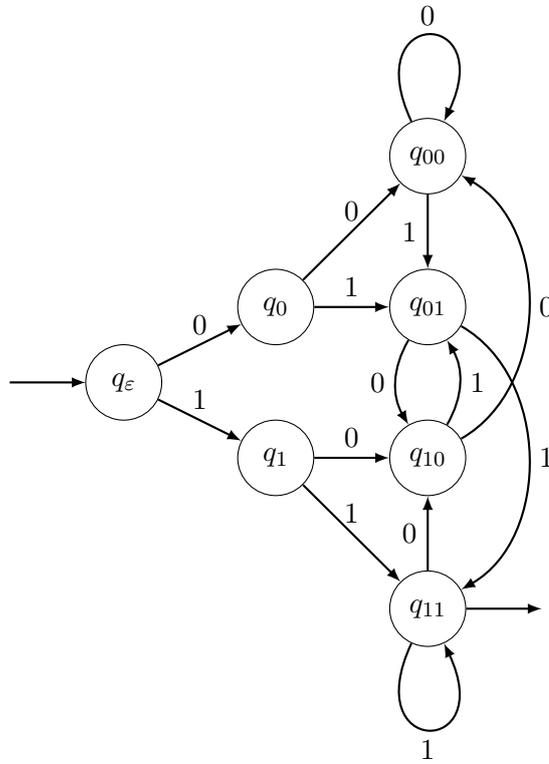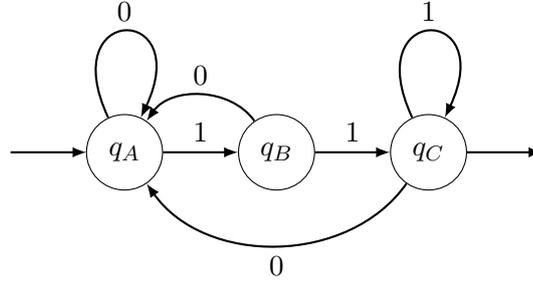

Figure 3.4: A DFA accepting $\operatorname{rep}_2(4\mathbb{N} + 3)$.

The minimal automaton of the language L is represented in Figure 3.5 and we have $R = \{q_A, q_B, q_C\}$. So $\#R = 3 = n + 1$ and R contains $\Phi(q_{00})$, $\Phi(q_{01})$ and $\Phi(q_{11})$.

Let $I = 3$ be the preperiod of X and take for instance $T = 0$. Then we have $\operatorname{val}_2(100) = 4 > 3$, $\operatorname{val}_2(110) = 6 > 3$ and $\operatorname{val}_2(111) = 7 > 3$ and the words 100, 110 and 111 perform pairwise distinct transformations on the set of states. Hence 100, 110 and 111 belong to different classes of $\leftrightarrow_L$.

Figure 3.5: The minimal DFA accepting $\mathrm{rep}_2(4\mathbb{N}+3)$.

3.3 First results on the syntactic complexity

Let $m, x \geq 2$ be integers such that $(m, x) = 1$. We denote by $\mathrm{ord}_m(x)$ the order of x in the multiplicative group $U(\mathbb{Z}/m\mathbb{Z})$ consisting of the invertible elements in $\mathbb{Z}/m\mathbb{Z}$, that is, $\mathrm{ord}_m(x)$ is the smallest positive integer j such that $x^j \equiv 1 \bmod m$. In particular, $\mathrm{ord}_m(x)$ is the period of the sequence $(x^n \bmod m)_{n \geq 0}$.

In the sequel, the integer $m \geq 2$ will denote the period of the considered periodic set and the integer $k \geq 2$ will denote the base of the representations.

Case 1 : $(m, k) = 1$

We first consider the case where the base and the period are coprime. Interestingly, the syntactic complexity only depends on the period and not on the structure of the periodic set. As in Remark 3.2.6, the second part of the following result ensures that we can find elements greater than the preperiod of an ultimately periodic set and will be useful to transpose our decision procedure from periodic sets to ultimately periodic sets.

Theorem 3.3.1. *Let $m, k \geq 2$ be integers such that $(m, k) = 1$. If $X \subseteq \mathbb{N}$ is periodic of (minimal) period m, then the syntactic complexity of $0^* \mathrm{rep}_k(X)$ is given by $m. \mathrm{ord}_m(k)$. In particular, this result holds for $X = m\mathbb{N}$. Moreover, for each class of $\leftrightarrow_{0^* \mathrm{rep}_k(X)}$, there exists an arbitrarily large integer ℓ such that $\mathrm{rep}_k(\ell)$ belongs to this class.*

Proof. Let $X \subseteq \mathbb{N}$ be a periodic set of period m. Let $u, v \in \Sigma_k^*$. Let us first

show that we have

$$u \leftrightarrow_{0^* \operatorname{rep}_k(X)} v \quad \Leftrightarrow \quad \mathtt{C}(u) = \mathtt{C}(v)$$

$$\Leftrightarrow \quad \begin{cases} \operatorname{val}_k(u) \equiv \operatorname{val}_k(v) & \mod m, \\ \quad |u| \equiv |v| & \mod \operatorname{ord}_m(k). \end{cases} \tag{3.4}$$

Let α be a multiple of $\operatorname{ord}_m(k)$ such that $k^\alpha > m$. Since $(k^i \mod m)_{i \geq 0}$ is a purely periodic sequence of period $\operatorname{ord}_m(k)$, it follows that $\operatorname{val}_k(u0^\alpha) \equiv \operatorname{val}_k(u) \mod m$. Indeed we have

$$\begin{aligned} \operatorname{val}_k(u0^\alpha) &= \operatorname{val}_k(u).k^\alpha + \operatorname{val}_k 0^\alpha \\ &= \operatorname{val}_k(u).k^\alpha \\ &\equiv \operatorname{val}_k(u).1 \mod m. \end{aligned}$$

Assume that $\operatorname{val}_k(u) \not\equiv \operatorname{val}_k(v) \mod m$. Since X is periodic of period m, using (3.1) there exists $r \in [\![0, m-1]\!]$ such that $\operatorname{val}_k(u) + r \in X$ and $\operatorname{val}_k(v) + r \notin X$ (the other case is treated similarly). So $(\varepsilon, 0^{\alpha - |\operatorname{rep}_k(r)|} \operatorname{rep}_k(r))$ belongs to $\mathtt{C}(u)$ and not to $\mathtt{C}(v)$ and we have $u \not\leftrightarrow_{0^* \operatorname{rep}_k(X)} v$.

Now assume that $\operatorname{val}_k(u) \equiv \operatorname{val}_k(v) \mod m$ and $|u| \not\equiv |v| \mod \operatorname{ord}_m(k)$. In that case, we have

$$\begin{aligned} \operatorname{val}_k(1u) - \operatorname{val}_k(1v) &= k^{|u|} + \operatorname{val}_k(u) - (k^{|v|} + \operatorname{val}_k(v)) \\ &\equiv k^{|u|} - k^{|v|} \mod m. \end{aligned}$$

and

$$\begin{aligned} & |u| \not\equiv |v| \mod \operatorname{ord}_m(k) \\ \Leftrightarrow \quad & \exists \ell \geq 0, r \in [\![0, \operatorname{ord}_m(k)]\!] \text{ such that } |u| = |v| + \ell \operatorname{ord}_m(k) + r \\ \Leftrightarrow \quad & \exists \ell \geq 0, r \in [\![0, \operatorname{ord}_m(k)]\!] \text{ such that } k^{|u|} = k^{|v|} k^{\ell \operatorname{ord}_m(k)} k^r \\ \Leftrightarrow \quad & \exists r \in [\![0, \operatorname{ord}_m(k)]\!] \text{ such that } k^{|u|} \equiv k^{|v|} k^r \mod m \\ \Leftrightarrow \quad & k^{|u|} \not\equiv k^{|v|} \mod m \end{aligned}$$

as $0 < r < \operatorname{ord}_m(k)$. Thus we obtain that $\operatorname{val}_k(1u) \not\equiv \operatorname{val}_k(1v) \mod m$ and we can proceed as in the first situation. There exists some $r \in [\![0, m-1]\!]$ such that $(1, 0^{\alpha - |\operatorname{rep}_k(r)|} \operatorname{rep}_k(r))$ belongs to $\mathtt{C}(u)$ and not to $\mathtt{C}(v)$. So we have $u \not\leftrightarrow_{0^* \operatorname{rep}_k(X)} v$.

Now proceed to the converse and assume that u, v are such that $\operatorname{val}_k(u) \equiv \operatorname{val}_k(v) \mod m$ and $|u| \equiv |v| \mod \operatorname{ord}_m(k)$. We have $k^{|u|} \equiv k^{|v|} \mod m$ and for all $x, y \in \Sigma_k^*$,

$$\begin{aligned} \operatorname{val}_k(xuy) &= \operatorname{val}_k(x) \, k^{|u|+|y|} + \operatorname{val}_k(u) \, k^{|y|} + \operatorname{val}_k(y) \\ &\equiv \operatorname{val}_k(x) \, k^{|v|+|y|} + \operatorname{val}_k(v) \, k^{|y|} + \operatorname{val}_k(y) \mod m \\ &\equiv \operatorname{val}_k(xvy) \mod m. \end{aligned}$$

So we have $u \leftrightarrow_{0^* \mathrm{rep}_k(X)} v$ and we have again used the fact that the sequence $(k^i \bmod m)_{i \geq 0}$ is purely periodic of period $\mathrm{ord}_m(k)$.

To conclude the proof, by considering words of the kind

$$u_{r,j} = 0^{\alpha - |\mathrm{rep}_k(r)| + j} \, \mathrm{rep}_k(r),$$

for $r \in [\![0, m-1]\!]$ and $j \in [\![1, \mathrm{ord}_m(k)]\!]$, it follows from (3.4) that we have $m \cdot \mathrm{ord}_m(k)$ non-empty classes of $\leftrightarrow_{0^* \mathrm{rep}_k(X)}$. Indeed, for all $r, r' \in [\![0, m-1]\!]$ and $i, j \in [\![1, \mathrm{ord}_m(k)]\!]$ such that $r \neq r'$ and $i \neq j$, we have

$$|u_{r,i}| \not\equiv |u_{r,j}| \mod \mathrm{ord}_m(k)$$

and

$$
\begin{aligned}
\mathrm{val}_k(u_{r,j}) &= \mathrm{val}_k(r) \\
&\not\equiv \mathrm{val}_k(r') \mod m \\
&\not\equiv \mathrm{val}_k(u_{r',j}) \mod m.
\end{aligned}
$$

Moreover each class contains representations of arbitrarily large integers. For all $T \geq 0$ and for all $u \in \Sigma_k^*$, using (3.4) we have

$$u \leftrightarrow_{0^* \mathrm{rep}_k(X)} (1\,0^{\mathrm{ord}_m(k)-1})^{Tm} u.$$

Indeed we have

$$
\begin{aligned}
\mathrm{val}_k((1\,0^{\mathrm{ord}_m(k)-1})^{Tm} u) &= \mathrm{val}_k(u) + T \sum_{j=1}^{m} k^{j \, \mathrm{ord}_m(k) + |u| - 1} \\
&\equiv \mathrm{val}_k(u) + T \sum_{j=1}^{m} k^{|u|-1} \mod m \\
&\equiv \mathrm{val}_k(u) + Tm k^{|u|-1} \mod m \\
&\equiv \mathrm{val}_k(u) \mod m.
\end{aligned}
$$

$\square$

Remark 3.3.2. Note that the characterization of Equation (3.4) implies that two periodic sets of minimal period m coprime with k have isomorphic syntactic monoids.

Example 3.3.3. Consider the set $X = 3\mathbb{N}$ in base 2 as in Example 3.2.5. The minimal automaton of $L = 0^* \mathrm{rep}_2(X)$ is depicted in Figure 3.1 on page 58 and Table 3.4 gives the corresponding multiplication table of the

syntactic monoid of $0^* \operatorname{rep}_2(X)$ where each class is given by one of its representatives.

In this case, with the previous notation, we have $k = 2$, $m = 3$ and $\operatorname{ord}_m(k) = \operatorname{ord}_3(2) = 2$. It is clear that the syntactic complexity of L is $m.\operatorname{ord}_m(k) = 3.2 = 6$.

	ε	0	1	01	10	101
ε	ε	0	1	01	10	101
0	0	ε	01	1	101	10
1	1	10	ε	101	0	01
01	01	101	0	10	ε	1
10	10	1	101	ε	01	0
101	101	01	10	0	1	ε

Table 3.4: The multiplication table of the syntactic monoid of $0^* \operatorname{rep}_2(3\mathbb{N})$.

Case 2 : $m = k^n$

Now consider the case where the period is a power of the base.

Theorem 3.3.4. *Let $k \geq 2$ and $m = k^n$ with $n \geq 1$. Then the syntactic complexity of $0^* \operatorname{rep}_k(m\mathbb{N})$ is given by $2n + 1$.*

Proof. The words $\varepsilon, 0, \ldots, 0^n, 1, 10, \ldots, 10^{n-1}$ have pairwise different contexts with respect to the language $0^* \operatorname{rep}_k(m\mathbb{N})$. For $i = 0, \ldots, n$, $(10^{n-i}, \varepsilon)$ belongs to $\mathtt{C}(0^{i+\ell})$, for all $\ell \geq 0$, since $\operatorname{val}_k(10^{n+l}) = k^n k^l \equiv 0 \mod m$ but does not belong to $\mathtt{C}(0^j)$ for $j < i$ nor to $\mathtt{C}(10^r)$, for $0 \leq r \leq n - 1$. Indeed, we have

$$\begin{aligned} \operatorname{val}_k(10^{n-i}0^j) &= \operatorname{val}_k(10^{n+j-i}) \\ &= k^{n+j-i} \\ &\not\equiv 0 \mod m \end{aligned}$$

since $n + j - i < n$, and

$$\begin{aligned} \operatorname{val}_k(10^{n-i}10^r) &= k^{n-i+r+1} + k^r \\ &= k^r(k^{n-i+1} + 1) \\ &\not\equiv 0 \mod m \end{aligned}$$

since $r < n$. Then, for $i, j \in [\![0, n]\!]$, $i \neq j$, we have $\mathtt{C}(0^i) \neq \mathtt{C}(0^j)$ and for $i \in [\![0, n]\!]$ and $j \in [\![0, n - 1]\!]$, we have $\mathtt{C}(0^i) \neq \mathtt{C}(10^j)$

In the same way, for $i = 1, \ldots, n$, $(\varepsilon, 0^i)$ belongs to $\mathtt{C}(10^{n-j})$, for $0 \leq j \leq i$, because

$$
\begin{aligned}
\mathrm{val}_k(10^{n-j}0^i) &= \mathrm{val}_k(10^{n+i-j}) \\
&= k^{n+i-j} \\
&\equiv 0 \mod m
\end{aligned}
$$

since $n+i-j > n$. But $(\varepsilon, 0^i)$ does not belong to $\mathtt{C}(10^j)$ for $j < n-i$ because

$$
\begin{aligned}
\mathrm{val}_k(10^j 0^i) &= k^{j+i} \\
&\not\equiv 0 \mod m
\end{aligned}
$$

since $j + i < n$. Then, for $i, j \in [\![0, n-1]\!]$, $i \neq j$, we have $\mathtt{C}(10^i) \neq \mathtt{C}(10^j)$. So the syntactic monoid of $0^* \mathrm{rep}_k(m\mathbb{N})$ has at least $2n + 1$ elements.

Now consider some word $u \in \Sigma_k^+$. Write $u = v0^i$ where v is either empty or ends with a non-zero digit.

If $i \geq n$, then $u \leftrightarrow_{0^* \mathrm{rep}_k(m\mathbb{N})} 0^n$. Indeed, we have

$$
\begin{aligned}
\mathrm{val}_k(v0^i) &= \mathrm{val}_k(v)k^i \\
&= \mathrm{val}_k(v)k^{n+i'} \\
&= m \, \mathrm{val}_k(v)k^{i'}.
\end{aligned}
$$

If $v \neq \varepsilon$ and $i < n$, then $u \leftrightarrow_{0^* \mathrm{rep}_k(m\mathbb{N})} 10^i$. Indeed, we have

$$
xuy = xv0^i y \in 0^* \mathrm{rep}_k(m\mathbb{N}) \Leftrightarrow (x \in \Sigma_k \text{ and } y = 0^l \text{ for } l \geq n - i).
$$

If $v = \varepsilon$ and $i < n$, the case $u = 0^i$ was already considered.

Hence the syntactic monoid of $0^* \mathrm{rep}_k(m\mathbb{N})$ has exactly $2n + 1$ elements. $\qquad\square$

As in Theorem 3.3.1, the second part of the following result will be of particular interest in the decision procedure and ensures that we can find elements greater than the preperiod of the ultimately periodic set in each equivalence class.

Proposition 3.3.5. *Let $k \geq 2$. If $X \subseteq \mathbb{N}$ is a periodic set of (minimal) period $m = k^n$ with $n \geq 1$, then the syntactic complexity of $L = 0^* \mathrm{rep}_k(X)$ is greater than or equal to $n + 1$. Moreover there exist arbitrarily large integers $t_1, \ldots, t_{n+1}$ such that the $n + 1$ words $\mathrm{rep}_k(t_1), \ldots, \mathrm{rep}_k(t_{n+1})$ belong to different equivalence classes of $\leftrightarrow_L$.*

Proof. Let $X \subseteq \mathbb{N}$ be a periodic set of period k^n, $n \geq 1$. By minimality of the period there exist $V \in \Sigma_k^*$, $\sigma, \tau \in \Sigma_k$ such that $\sigma \neq \tau$, $|V| = n-1$, and for all $u \in \Sigma_k^*$, we have $\mathrm{val}_k(u\sigma V) \in X$ and, for all $u \in \Sigma_k^*$, $\mathrm{val}_k(u\tau V) \notin X$. If that was not the case, the fact that a word w belongs to L would only depend on its suffix of length $n-1$, so in particular, we would have $\mathrm{val}_k(w) \in X$ if and only if $\mathrm{val}_k(w) + k^{n-1} \in X$ for all words w. This contradicts the fact that k^n is the period of X. Then, in other words, L is a n-definite language. One can conclude using Lemma 3.2.12.

$\square$

Remark 3.3.6. The bound in Proposition 3.3.5 is tight. One can for instance consider the set $5 + 8\mathbb{N}$ written in base 2. The minimal automaton of $0^* \mathrm{rep}_2(5 + 8\mathbb{N})$ is depicted in Figure 3.6. The corresponding transforma-
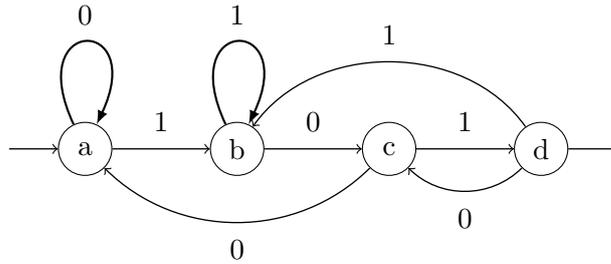
Figure 3.6: The minimal automaton of $0^* \mathrm{rep}_2(5 + 8\mathbb{N})$.

tion automaton (see Definition 3.2.4) is given in Figure 3.7 displaying one strongly connected component with four elements and the syntactic monoid has exactly four infinite equivalence classes as explained in Remark 3.2.6.

Case 3 : $m = k^n q$

For convenience, we set $s = \mathrm{ord}_q(b)$ in all what follows.

Remark 3.3.7. The sequence $(b^i \bmod m)_{i \geq 0}$ is ultimately periodic of period $\mathrm{ord}_q(b)$ and preperiod n. For instance, the sequence

$$(2^i \bmod 24)_{i \geq 0} = (1, 2, 4, (8, 16)^\omega)$$

has preperiod $3 = \log_2 8$ and period $\mathrm{ord}_3(2)$. Indeed, $(b^i \bmod q)_{i \geq 0}$ is purely periodic of period $\mathrm{ord}_q(b)$ and

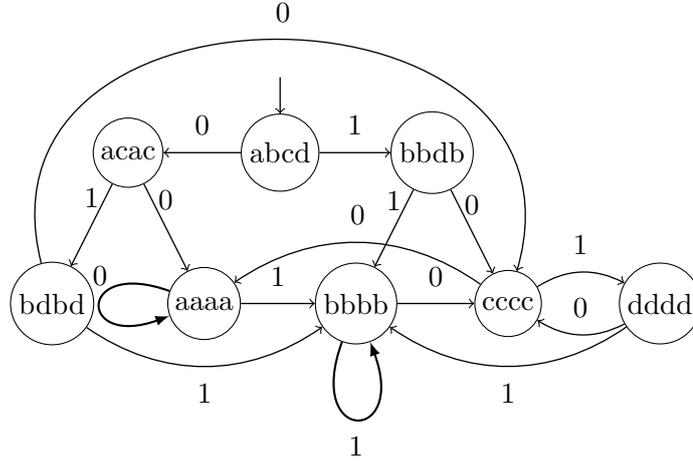$$(b^i \bmod b^n)_{i \geq 0} = (1, b, b^2, \ldots, b^{n-1}, 0, 0, \ldots)$$

Figure 3.7: The transformation automaton of $0^* \operatorname{rep}_2(5 + 8\mathbb{N})$.

is ultimately periodic with preperiod n and period 1.

For the sake of completeness, we restate the following result. See [RV11, Theorem 4].

Theorem 3.3.8. *Let $k \geq 2$ and $m = k^n q$ where $n \geq 1$ and $(q, k) = 1$ and $q \geq 2$. Then the syntactic complexity of $0^* \operatorname{rep}_k(m\mathbb{N})$ is given by $(n + 1).q.\operatorname{ord}_q(k) + n$.*

Before giving the proof of Theorem 3.3.8, we need to recall the following results that were proved in [RV11]. Arguments of the proof of Lemma 3.3.9 are similar to the ones developed in the proof of Theorem 3.3.1 and the proofs of the other results are straightforward.

Lemma 3.3.9 ([RV11]). *Let $k \geq 2$ and $m = k^n q$ where $n \geq 1$ and $(q, k) = 1$ and $q \geq 2$. Let $X \subseteq \mathbb{N}$ be a periodic set of (minimal) period m. For any words two $u, v \in \Sigma_k^*$ of length at least n, the following implications hold*

$$((|u| \equiv |v| \bmod s) \wedge (\operatorname{val}_k(u) \equiv \operatorname{val}_k(v) \bmod m)) \Rightarrow u \leftrightarrow_{0^* \operatorname{rep}_k(X)} v, \quad (3.5)$$

$$((|u| \not\equiv |v| \bmod s) \vee (\operatorname{val}_k(u) \not\equiv \operatorname{val}_k(v) \bmod q)) \Rightarrow u \not\leftrightarrow_{0^* \operatorname{rep}_k(X)} v. \quad (3.6)$$

Corollary 3.3.10 ([RV11]). *Let $k \geq 2$ and $m = k^n q$ where $n \geq 1$ and $(q, k) = 1$ and $q \geq 2$. If $X \subseteq \mathbb{N}$ is a periodic set of (minimal) period m and $u, v \in \Sigma_k^*$ are two words of length at least n, then $u \leftrightarrow_{0^* \mathrm{rep}_k(X)} v$ implies that the following conditions hold*

 i) $|u| \equiv |v| \bmod s$

 ii) either, $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod m$ or, there exists a unique $i \in [\![0, n-1]\!]$ such that $(\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod k^i q) \wedge (\mathrm{val}_k(u) \not\equiv \mathrm{val}_k(v) \bmod k^{i+1} q)$.

Remark 3.3.11. The converse does not hold. Consider $X = 12\mathbb{N}$ and $k = 2$, one can check with $u = 100$ and $v = 10110$ that *i)* and *ii)* are fulfilled but $(1, \varepsilon)$ belongs to $\mathtt{C}(u)$ but not to $\mathtt{C}(v)$.

A word $u \in \Sigma_k^*$ of length at least n has three features: its length modulo s, $|u| \bmod s$, its value modulo q, $\mathrm{val}_k(u) \bmod q$, and its suffix of length n (i.e., its value modulo k^n). Observe that if we know $\mathrm{val}_k(u) \bmod q$ and $\mathrm{val}_k(u) \bmod k^n$, then we obtain $\mathrm{val}_k(u) \bmod k^n q$.

Lemma 3.3.12 ([RV11]). *Let $q, k \geq 2$ be two coprime integers, $w \in \Sigma_k^*$ and $i \in [\![0, q-1]\!]$. There exists some word $x \in (0^{s-1}1)^* 0^*$ such that $\mathrm{val}_k(xw) \equiv i \bmod q$.*

Corollary 3.3.13 ([RV11]). *For all words $w \in \Sigma_k^*$ of length at least n, $i \in [\![0, q-1]\!]$, $\ell \in [\![0, s-1]\!]$ and $I > 0$, there exists a word u having w as suffix and such that $\mathrm{val}_k(u) \equiv i \bmod q$, $|u| \equiv \ell \bmod s$ and $\mathrm{val}_k(u) > I$.*

We are now ready to give the proof of Theorem 3.3.8 from [RV11].

Proof of Theorem 3.3.8. Let $X = m\mathbb{N}$. Notice that $d \in \mathbb{N}$ is a multiple of m if and only if $\mathrm{rep}_k(d) = x0^n$ and $\mathrm{val}_k(x)$ is a multiple of q. We get the first term in the formula. Let u, v be two words of length at least n. From (3.6), if $|u| \not\equiv |v| \bmod s$ or $\mathrm{val}_k(u) \not\equiv \mathrm{val}_k(v) \bmod q$, then u and v belong to two different classes for $\leftrightarrow_{0 \, \mathrm{rep}_k(m\mathbb{N})}$. Now assume that $|u| \equiv |v| \bmod s$ and $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod q$. Observe that if u ends with exactly $i \in [\![0, n-1]\!]$ zeroes and v ends with j zeroes with $j > i$, then $u \not\leftrightarrow_{0 \, \mathrm{rep}_k(m\mathbb{N})v}$. With the same construction as in the proof of Lemma 3.3.12, there exists some word $x \in (0^{s-1}1)0$ such that $\mathrm{val}_k(xu) \equiv \mathrm{val}_k(xv) \equiv 0 \bmod q$ and we can conclude that $(x, 0^{\max n-j, 0})$ belongs to $\mathtt{C}(v)$ and not to $\mathtt{C}(u)$. This proves with Corollary 3.3.13 that words of length at least n provide the syntactic monoid of $0 \, \mathrm{rep}_k(m\mathbb{N})$ with at least $(n+1).q.s$ classes.

Now observe that for any two words u, v of length at least n such that $|u| \equiv |v| \mod s$, $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \mod q$, and u, v either end with exactly the same number i of zeroes with $i < n$, or both end with at least n zeroes, then $u \leftrightarrow_{0\,\mathrm{rep}_k(m\mathbb{N})} v$. This proves that words of length at least n provide the syntactic monoid of $0\,\mathrm{rep}_k(m\mathbb{N})$ with no more than $(n+1).q.s$ classes. Now we take into account words of length less than n and show that they provide the syntactic monoid of $0\,\mathrm{rep}_k(m\mathbb{N})$ with n new classes giving the second term in the expression of $M_{k,m}$. The n words $\varepsilon, 0, \ldots, 0^{n-1}$ have pairwise different contexts: for all $\ell \in [\![0, n-1]\!]$,

$$\mathsf{C}(0^\ell) \cap \{(\mathrm{rep}_k(q), 0^r) \mid r \geq 0\} = \{(\mathrm{rep}_k(q), 0^r) \mid r \geq n - \ell\}.$$

Let $\ell \in [\![0, n-1]\!]$. We show by contradiction that they are indeed n new classes. Assume that there exists $u \in \Sigma_k$ such that $|u| \geq n$ and $u \leftrightarrow_{0\,\mathrm{rep}_k(m\mathbb{N})} 0^\ell$. Since $(\varepsilon, \varepsilon)$ belongs to $\mathsf{C}(0^\ell)$, we deduce that $\mathrm{val}_k(u) \equiv 0 \mod m$. We have

$$\mathrm{val}_k(1u) \equiv k^{|u|} \mod m \text{ and } \mathrm{val}_k(10^\ell) \equiv k^\ell \mod m$$

but since the sequence $(k^r \mod m)_{r \geq 0}$ is ultimately periodic with period s and preperiod n, from $\ell < n \leq |u|$ we conclude that

$$\mathrm{val}_k(1u) \not\equiv \mathrm{val}_k(10^\ell) \mod m.$$

Let $t \geq n$ be a multiple of s. We have

$$\mathrm{val}_k((0^{t-1}1)^m u) \equiv \mathrm{val}_k(u) \equiv 0 \mod m$$

and

$$\mathrm{val}_k((0^{t-1}1)^m 0^\ell) = k^\ell + \sum_{r=1}^{m-1} k^{\ell+rt} \equiv k^\ell + (m-1)k^{\ell+t} \mod m.$$

Since $\ell < n \leq \ell + t$, we have $k^\ell \not\equiv k^{\ell+t} \mod m$. We conclude that

$$\mathrm{val}_k((0^{t-1}1)^m 0^\ell) \not\equiv 0 \mod m$$

proving that $((0^{t-1}1)^m, \varepsilon)$ belongs to $\mathsf{C}(u)$ and not to $\mathsf{C}(0^\ell)$.

To conclude the proof, we have to consider some word $u \not\in 0$ of length less than n and prove that u does not provide any new equivalence class. This comes from the fact that $u \leftrightarrow_{0\,\mathrm{rep}_k(m\mathbb{N})} 0^{rs}u$ where r is chosen large enough such that $rs + |u| \geq n$. It is enough to show that $u \leftrightarrow_{0\,\mathrm{rep}_k(m\mathbb{N})} 0^s u$. Let $x, y \in \Sigma_k^*$. If $|uy| \geq n$ then $\mathrm{val}_k(xuy) \equiv \mathrm{val}_k(x0^s uy) \mod m$ and $|xuy| \equiv |x0^s uy| \mod s$ and we can use (3.5). Otherwise $|uy| < n$ and since $u \not\in 0$, this means that we simultaneously have $\mathrm{val}_k(xuy) \not\equiv 0 \mod m$ and $\mathrm{val}_k(x0^s uy) \not\equiv 0 \mod m$. This means that $xuy \in 0\,\mathrm{rep}_k(m\mathbb{N})$ if and only if $x0^s uy \in 0\,\mathrm{rep}_k(m\mathbb{N})$. $\qquad\square$

3.4   Lower bound in the general case

In this section, we give a lower bound on the syntactic complexity of a periodic set X in a case that did not appear in [RV11]. Indeed, we consider the case where the period m is a power of the base times any integer, i.e., not necessarily an integer coprime with the base b. Recall that such a case is needed to get a decision procedure to test periodicity.

Let $k \geq 2$. Notice that any integer m can be uniquely written as

$$m = dk^n q \quad \text{with } (q, k) = 1 \tag{3.7}$$

such that $n \geq 0$ and $q \geq 1$ are chosen maximal. Hence, if $p_1^{\alpha_1} \cdots p_\ell^{\alpha_\ell}$ denotes the prime decomposition of k with $\alpha_i > 0$, for all $i \in [\![1, \ell]\!]$, then $d = p_1^{\delta_1} \cdots p_\ell^{\delta_\ell}$ with $\delta_i \geq 0$, for all $i \in [\![1, \ell]\!]$ and there exists r such that $\delta_r < \alpha_r$. For convenience, we set $s := \mathrm{ord}_q(k)$ in all what follows. Let

$$\eta := \min\{j \in \mathbb{N} \mid \forall i \in [\![1, \ell]\!],\ n\alpha_i + \delta_i \leq j\alpha_i\}.$$

In other words, η is the smallest j such that $dk^n | k^j$. Note that $\eta \geq n$.

Remark 3.4.1. Since $dk^n | k^\ell$, for all $\ell \geq \eta$, a word has its numerical value modulo dk^n completely determined by its suffix of length η. Indeed, the digits written to the left of this suffix add to its numerical value a multiple of k^η which is a multiple of dk^n. In particular, if two words $u, v \in \Sigma_k^*$ of length at least η have same suffix of length η and are such that $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod q$, then $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod m$ since q and dk^n are coprime.

Lemma 3.4.2. *Let $k \geq 2$ and $m = dk^n q$ be given as in (3.7). Let $X \subseteq \mathbb{N}$ be a periodic set of (minimal) period m. For any words $u, v \in \Sigma_k^*$ of length at least η, we have*

$$((|u| \equiv |v| \bmod s) \wedge (\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod m)) \Rightarrow u \leftrightarrow_{0^* \mathrm{rep}_k(X)} v. \tag{3.8}$$

Proof. Let $u, v \in \Sigma_k^*$ be two words of length at least η. Using the fact that $k^\ell \equiv k^{\ell+s} \bmod q$, for all $\ell \geq 0$ and $k^\ell \equiv k^{\ell+s} \bmod dk^n$, for all $\ell \geq \eta$, we have that $k^\ell \equiv k^{\ell+s} \bmod m$, for all $\ell \geq \eta$. Notice that if u, v are such that $|u| \equiv |v| \bmod s$ and $\mathrm{val}_k(u) \equiv \mathrm{val}_k(v) \bmod m$, then, for all $x, y \in \Sigma_k^*$, $\mathrm{val}_k(xuy) \equiv \mathrm{val}_k(xvy) \bmod m$ which means that $u \leftrightarrow_{0^* \mathrm{rep}_k(X)} v$. Indeed, we

have

$$\text{val}_k(xuy) - \text{val}_k(xvy)$$
$$= \text{val}_k(x)k^{|u|+|y|} + \text{val}_k(u)k^{|y|} + \text{val}_k(y)$$
$$\quad - \text{val}_k(x)k^{|v|+|y|} - \text{val}_k(v)k^{|y|} - \text{val}_k(y)$$
$$= \text{val}_k(x)(k^{|u|+|y|} - k^{|v|+|y|}) + k^{|y|}(\text{val}_k(u) - \text{val}_k(v))$$
$$\equiv \text{val}_k(x)(k^{|u|+|y|} - k^{|v|+|y|}) \mod m$$
$$\equiv \text{val}_k(x)(k^{|u|+|y|} - k^{|u|+ts+|y|}) \mod m \text{ for } t \geq 0$$
$$\equiv \text{val}_k(x)k^{|u|}(k^{|y|} - k^{ts+|y|}) \mod m$$
$$\equiv 0 \mod m.$$

$\square$

Lemma 3.4.3. *Let $k \geq 2$ and $m = dk^n q$ given as in (3.7). Let $X \subseteq \mathbb{N}$ be a periodic set of (minimal) period m. For any two words $u, v \in \Sigma_k^*$, we have*

$$((|u| \not\equiv |v| \mod s) \vee (\text{val}_k(u) \not\equiv \text{val}_k(v) \mod q)) \Rightarrow u \not\leftrightarrow_{0^* \text{rep}_k(X)} v. \quad (3.9)$$

One may notice that the main difference between (3.8) and (3.9) is that congruences of numerical values are considered modulo m and q respectively.

Proof. As a first case, suppose that $\text{val}_k(u) \not\equiv \text{val}_k(v) \mod q$. Take $\alpha \in \mathbb{N}$ such that $k^\alpha > m$. Since $(k, q) = 1$, we get that $\text{val}_k(u)k^\alpha \not\equiv \text{val}_k(v)k^\alpha \mod q$. Hence $\text{val}_k(u)\, k^\alpha \not\equiv \text{val}_k(v)\, k^\alpha \mod m$ and, using (3.1), there exists $r \in [\![0, m-1]\!]$ such that $\text{val}_k(u)k^\alpha + r \in X$ and $\text{val}_k(v)k^\alpha + r \notin X$ (the other case is treated similarly). We can conclude that $(\varepsilon, 0^{\alpha - |\text{rep}_k(r)|} \text{rep}_k(r))$ belongs to $\text{C}(u)$ and not to $\text{C}(v)$. So we have $u \not\leftrightarrow_{0^* \text{rep}_k(X)} v$.

As a second case, suppose that $\text{val}_k(u) \equiv \text{val}_k(v) \mod q$ but that $|u| \not\equiv |v| \mod s$. This implies that $k^{|u|} \not\equiv k^{|v|} \mod q$. Thus $\text{val}_k(1u) \not\equiv \text{val}_k(1v) \mod q$ and we can proceed as in the first case to prove that $1u \not\leftrightarrow_{0^* \text{rep}_k(X)} 1v$. So there exists $r \in [\![0, m-1]\!]$ such that $(1, 0^{\alpha - |\text{rep}_k(r)|} \text{rep}_k(r))$ belongs to $\text{C}(u)$ and not to $\text{C}(v)$ and we have $u \not\leftrightarrow_{0^* \text{rep}_k(X)} v$. $\square$

In the following lemma, the integer I will play the role of the preperiod of the ultimately periodic set we will consider. Again, it permits us to extend the lower bound of the periodic case to the ultimately periodic case.

Lemma 3.4.4. *For all words $w \in \Sigma_k^*$, $i \in [\![0, q-1]\!]$, $\ell \in [\![0, s-1]\!]$ and $I > 0$, there exists a word u having w as suffix and such that $\text{val}_k(u) \equiv i \mod q$, $|u| \equiv \ell \mod s$ and $\text{val}_k(u) > I$.*

Proof. Let $L \geq |w|$ be a multiple of s. We set $z = \text{val}_k(w) \bmod q$. Using the periodicity of $(k^r \bmod q)_{r \geq 0}$, we have $k^{L+rs} \equiv 1 \bmod q$, for all $r \geq 0$, and

$$
\begin{aligned}
\text{val}_k[(0^{s-1}1)^{i+q-z}\,0^{L-|w|}w] &= \sum_{r=0}^{i+q-z-1} k^{L+rs} + \text{val}_k(w) \\
&\equiv \sum_{r=0}^{i+q-z-1} 1 + z \bmod q \\
&\equiv i \bmod q
\end{aligned}
$$

since $k^{L+rs} \equiv 1 \bmod q$.

To conclude the proof, set $v := (0^{s-1}1)^{i+q-z}\,0^{L-|w|}w$ and consider $u = 0^\ell(0^{s-1}1)^{tq}v$. We have

$$
\text{val}_k(u) \equiv \text{val}_k(v) \equiv i \bmod q
$$

and

$$
|u| = l + stq + s(i + q - z) + L \equiv l \bmod s.
$$

To obtain $\text{val}_k(u) > I$, it suffices to take t sufficiently large. $\qquad\square$

Definition 3.4.5. Take a periodic set $X \subseteq \mathbb{N}$ of period $m = dk^n q$ given as in (3.7). Consider the characteristic word $(x_t)_{t \geq 0} \in \{0,1\}^\omega$ of X where $x_t = 1$ if and only if $t \in X$. This infinite word is periodic of period m. The q infinite words $(x_{qt})_{t \geq 0}$, $(x_{qt+1})_{t \geq 0}$, $\ldots$, $(x_{qt+q-1})_{t \geq 0}$ are periodic and each of their periods divides dk^n. As m is the minimal period of X, there exists a non-empty set $J \subseteq [\![0, q-1]\!]$ such that, for all $j \in J$, $(x_{qt+j})_{t \geq 0}$ has exactly period P_j satisfying

$$
P_j := p_1^{\beta_{j,1}} \cdots p_\ell^{\beta_{j,\ell}} \quad \text{where} \quad \max_{r \in [\![1,\ell]\!]} \beta_{j,r} = \max_{r \in [\![1,\ell]\!]}(n\alpha_r + \delta_r). \tag{3.10}
$$

Indeed, assume without loss of generality that $\max_{r \in [\![1,\ell]\!]}(n\alpha_r + \delta_r) = n\alpha_1 + \delta_1$. Using the same arguments as in [Hon86], set $u = \max_{j \in [\![0,q-1]\!]} \beta_{j,r}$ and assume $u < n\alpha_1 + \delta_1$. Then P_j divides $p_1^u p_2^{n\alpha_2+\delta_2} \cdots p_\ell^{n\alpha_\ell+\delta_\ell}$ for all $j \in [\![0, q-1]\!]$ and we have that

$$
m = q.\max_{j \in [\![0,q-1]\!]} P_j \text{ divides } q.p_1^u p_2^{n\alpha_2+\delta_2} \cdots p_\ell^{n\alpha_\ell+\delta_\ell} < m.
$$

This leads to a contradiction and ensures that the assertion (3.10) is true. We set

$$
\gamma_j := \min\{g \in \mathbb{N} \mid \forall r \in [\![1, \ell]\!], \ g\alpha_r \geq \beta_{j,r}\}. \tag{3.11}
$$

In other words, γ_j is the smallest g such that $P_j | k^g$. Now choose $i \in J$ such that the corresponding γ_i is maximal (if several i's fulfill this condition, to avoid ambiguity, we take the smallest such i) and we will denote by P and γ the corresponding P_i and γ_i. Clearly, γ is a function of d, n and q. In particular, $k^{\gamma-1} \not\equiv 0 \bmod P$.

Remark 3.4.6. Note that if $n \to +\infty$ or $d \to +\infty$, then $\gamma \to +\infty$.

Example 3.4.7. Let $k = 75$, i.e., $p_1 = 3$, $p_2 = 5$, $\alpha_1 = 1$ and $\alpha_2 = 2$. Consider two infinite words over $\{0, 1\}$, v having a 1 at multiples of 9, w at multiples of 25. We define x by interleaving these two words. This word x has period $2.9.25$, i.e., $d = 3$, $n = 1$ and $q = 2$. Hence $\delta_1 = 1$, $\delta_2 = 0$ and we get $n\alpha_1 + \delta_1 = n\alpha_2 + \delta_2 = 2$. Observe that $(x_{2t})_{t \geq 0} = v$ (resp. $(x_{2t+1})_{t \geq 0} = w$) has period $P_0 = 3^2$ (resp. $P_1 = 5^2$) which satisfies condition (3.10). We get $J = \{0, 1\}$, $\gamma_0 = 2$ and $\gamma_1 = 1$. So we choose $i = 0$ and set $P = 3^2$ and $\gamma = 2$.

Proposition 3.4.8. *Let $k \geq 2$ and $m = dk^n q$ be given as in (3.7). If $X \subseteq \mathbb{N}$ is a periodic set of (minimal) period m, then the syntactic complexity of $0^* \operatorname{rep}_k(X)$ is greater than or equal to $M = \max\{q.s, \frac{\gamma+1}{q.s}\}$. Moreover there exist arbitrarily large integers $t_1, \ldots, t_M$ such that $\operatorname{rep}_k(t_1), \ldots, \operatorname{rep}_k(t_M)$ belong to different equivalence classes of $\leftrightarrow_{0^* \operatorname{rep}_k(X)}$.*

Proof. Thanks to Lemma 3.4.4, for all $i \in [\![0, q-1]\!]$, $\ell \in [\![0, s-1]\!]$ and $I > 0$, we can construct a word $u \in \Sigma_k^*$ such that $\operatorname{val}_k(u) \equiv i \bmod q$, $|u| \equiv \ell \bmod s$ and $\operatorname{val}_k(u) > I$. Thus, by (3.9), any two such words are in different classes for $\leftrightarrow_{0^* \operatorname{rep}_k(X)}$. Hence we have at least $q.s$ different classes and each of them contains representations of arbitrary large elements.

Consider $i \in [\![0, q-1]\!]$, P and γ given in Definition 3.4.5 such that $(x_{qt+i})_{t \geq 0}$ has exactly period P. Take $T \geq 0$ such that $|\operatorname{rep}_k(qT + i)| \geq \gamma$. Note that if $j \not\equiv j' \bmod P$, then $\operatorname{rep}_k(q(T + j) + i)$ and $\operatorname{rep}_k(q(T + j') + i)$ have different suffixes of length γ. Indeed, assume that $\operatorname{rep}_k(q(T + j) + i)$ and $\operatorname{rep}_k(q(T + j') + i)$ have the same suffix of length γ, then $q(T + j) + i \equiv q(T + j') + i \bmod k^\gamma$ and we have $q(j' - j) \equiv 0 \bmod k^\gamma$. Since $(q, k) = 1$, we get $j' \equiv j \bmod k^\gamma$. As P divides k^γ, we have $j' \equiv j \bmod P$.

Since $(x_{qt+i})_{t \geq 0}$ has (minimal) period P and $k^{\gamma-1} \not\equiv 0 \bmod P$, then $T \not\equiv T + k^{\gamma-1} \bmod P$ and there exists $j \in [\![0, P-1]\!]$ such that $g_1 := q(T + j) + i \in X$ and $g_2 := q(T + j + k^{\gamma-1}) + i \notin X$ (or equivalently $g_1 \notin X$ and $g_2 \in X$). Since $g_2 = g_1 + qk^{\gamma-1}$, then $\operatorname{rep}_k(g_1)$ and $\operatorname{rep}_k(g_2)$ have the same suffix $V = v_1 \cdots v_{\gamma-1}$ of length $\gamma - 1$. But since $g_1 \not\equiv g_2 \bmod P$ and $P | k^\gamma$, $g_1 \not\equiv g_2 \bmod k^\gamma$. Thus $\operatorname{rep}_k(g_1)$ and $\operatorname{rep}_k(g_2)$ do not have the same suffix of length

γ. We can assume that there exist some prefixes $p, p' \in \Sigma_k^*$ and two distinct symbols $\sigma, \tau \in \Sigma_k$ such that

$$\operatorname{rep}_k(g_1) = p\sigma V \in 0^* \operatorname{rep}_k(X) \text{ and } \operatorname{rep}_k(g_2) = p'\tau V \notin 0^* \operatorname{rep}_k(X).$$

We are now ready to show that the minimal automaton of $0^* \operatorname{rep}_k(X \cap (q\mathbb{N}+i))$ has at least $\gamma + 1$ states reached by words of length at least η (this last requirement is always fulfilled by adding leading zeroes if necessary). Let $L = 0^* \operatorname{rep}_k(X)$ and $L' = 0^* \operatorname{rep}_k(q\mathbb{N} + i)$ having respectively $\mathcal{A}_L$ and $\mathcal{A}_{L'}$ as minimal automaton. We proceed as in the proof of Lemma 3.2.12. On the set of states of the minimal automaton $\mathcal{A}_{L \cap L'}$ of $L \cap L'$, for all $s \geq 0$, we define the relation

$$E_s(r, r') \Leftrightarrow (\forall x \in \Sigma_k^*) \big[|x| \geq s \Rightarrow (\delta_{L \cap L'}(r, x) \in F \Leftrightarrow \delta_{L \cap L'}(r', x) \in F) \big]$$

where F is the set of final states of $\mathcal{A}_{L \cap L'}$. Let $s \in [\![0, \gamma - 1]\!]$. Consider the states r and r' reached when reading respectively $p\sigma v_1 \cdots v_{\gamma - s - 1}$ and $p'\tau v_1 \cdots v_{\gamma - s - 1}$. They do not satisfy $E_s(r, r')$ but we can observe that they satisfy $E_{s+1}(r, r')$. Indeed, take any $x \in \Sigma_k^*$ of length $s+1$. Since $\operatorname{val}_k(p\sigma V) \equiv \operatorname{val}_k(p'\tau V) \bmod q$, subtracting $\operatorname{val}_k(V)$ on both sides, we get $\operatorname{val}_k(p\sigma 0^{|V|}) \equiv \operatorname{val}_k(p'\tau 0^{|V|}) \bmod q$. Since $(q, k) = 1$, we can multiply or divide a convenient number of times by k and get $\operatorname{val}_k(p\sigma 0^{|W|}) \equiv \operatorname{val}_k(p'\tau 0^{|W|}) \bmod q$. Hence, one adds $\operatorname{val}_k(W)$ to both sides and, for all words $W \in \Sigma_k^*$, we have $\operatorname{val}_k(p\sigma W) \equiv \operatorname{val}_k(p'\tau W) \bmod q$.

So, in particular,

$$X_s := p\sigma v_1 \cdots v_{\gamma - s - 1} x \text{ and } Y_s := p'\tau v_1 \cdots v_{\gamma - s - 1} x$$

have the same value modulo q and the same suffix of length γ. This means that either they both belong to $L \cap L'$, or they both do not belong to $L \cap L'$. Indeed, if $\operatorname{val}_k(X_s)$ and $\operatorname{val}_k(Y_s)$ are both congruent to a value not equal to i modulo q, then X_s and Y_s do not belong to L'. As a second case, we may assume that $X_s, Y_s \in L'$, i.e., $\operatorname{val}_k(X_s)$ and $\operatorname{val}_k(Y_s)$ are respectively of the form $q(T + j) + i$ and $q(T + j') + i$ for some j, j'. As discussed in the second paragraph of this proof, since X_s and Y_s have the same suffix of length γ, then $j \equiv j' \bmod P$ which means that $q(T + j) + i \in X \Leftrightarrow q(T + j') + i \in X$, or $X_s \in L \Leftrightarrow Y_s \in L$.

Following the same lines as in the proof of Lemma 3.2.12 we conclude that the minimal automaton of $L \cap L'$ has at least $\gamma + 1$ states.

For any DFA $\mathcal{A}$ having Q as set of states and δ as transition function, we set

$$\mathcal{T}_\eta(\mathcal{A}) = \{ f_w : Q \to Q, q \mapsto \delta(q, w) \mid w \in \Sigma_k^*, |w| \geq \eta \}$$

as the set of actions on Q derived from words of length at least η. Consider the product automaton $\mathcal{P} = \mathcal{A}_L \times \mathcal{A}_{L'}$ recognizing $L \cap L'$. We clearly have $\#\mathcal{T}_\eta(\mathcal{P}) \leq \#\mathcal{T}_\eta(\mathcal{A}_L).\#\mathcal{T}_\eta(\mathcal{A}_{L'})$. Since there is a canonical morphism of automata from $\mathcal{P}$ onto $\mathcal{A}_{L \cap L'}$, we get $\#\mathcal{T}_\eta(\mathcal{A}_{L \cap L'}) \leq \#\mathcal{T}_\eta(\mathcal{P})$.

Using Proposition 3.2.3 and Theorem 3.3.1, we get $\#\mathcal{T}_\eta(\mathcal{A}_{L'}) = q.s$. With the above discussion about the number of states of $\mathcal{A}_{L \cap L'}$, we have

$$\#\mathcal{T}_\eta(\mathcal{A}_{L \cap L'}) \geq \gamma + 1.$$

Putting all of this together, we obtain

$$\gamma + 1 \leq \#\mathcal{T}_\eta(\mathcal{A}_{L \cap L'}) \leq \#\mathcal{T}_\eta(\mathcal{P}) \leq \#\mathcal{T}_\eta(\mathcal{A}_L).q.s.$$

Therefore, $\#\mathcal{T}_\eta(\mathcal{A}_L) \geq \frac{\gamma+1}{q \cdot s}$. Since $\mathcal{T}_\eta(\mathcal{A}_L)$ is isomorphic to a subset of the syntactic monoid of L, this monoid has at least $\frac{\gamma+1}{q \cdot s}$ classes, each one containing a word of length at least η. Since for all words w of length at least η, we can use Lemma 3.4.2 to get $w \leftrightarrow_{0^* \mathrm{rep}_k(X)} (0^{s-1}1)^m w$, integers corresponding to elements of the different equivalence classes can be chosen arbitrarily large. $\qquad\square$

3.5 Application to a decision procedure

Let $X \subseteq \mathbb{N}$ be a k-recognizable set of integers such that $L = 0^* \mathrm{rep}_k(X)$ is accepted by some DFA $\mathcal{A}_L$. As explained in Section 3.1, a typical technique to decide whether or not X is ultimately periodic is to prove that whenever X is ultimately periodic, then its period and its preperiod must be bounded by some quantities depending only on the size of the DFA $\mathcal{A}$. Therefore, one has a finite number of admissible periods and preperiods to test, leading to a decision procedure. In particular, the following result [BCFR09, Prop. 44] stated in full generality for any abstract numeration system shows that we only have to obtain an upper bound on the admissible periods.

Proposition 3.5.1. *Let $S = (L, \Sigma, <)$ be an abstract numeration system. If $X \subseteq \mathbb{N}$ is an ultimately periodic set of period p_X such that $\mathrm{rep}_S(X)$ is accepted by a DFA with d states, then the preperiod of X is bounded by an effectively computable constant C depending only on d and on p_X.*

The following result is a consequence of Proposition 3.4.8.

Theorem 3.5.2. *Let $k \geq 2$. If $X \subseteq \mathbb{N}$ is an ultimately periodic set of period $p_X = dk^n q$ given as in (3.7), then the syntactic complexity of $0^* \mathrm{rep}_k(X)$ is greater than or equal to $\max\{q.s, \frac{\gamma+1}{q.s}\}$ where γ is defined as in (3.11).*

Proof. Let I be the preperiod of X. Even though Proposition 3.4.8 is about (purely) periodic sets of integers, if we consider instead an ultimately periodic set, since we can choose words belonging to different equivalence classes in such a way that their numerical value is greater than I, then the lower bound on the number of classes is still valid for the ultimately periodic case. $\square$

Assume that $k \geq 2$. Therefore, given a DFA $\mathcal{A}$ accepting $0^* \operatorname{rep}_k(X)$ and so the corresponding syntactic monoid, if X is ultimately periodic, then we get an upper bound on its period. Indeed, if m written $dk^n q$ as in (3.7) tends to infinity, then at least one of the three quantities d, n or q tends to infinity and thanks to Remark 3.4.6, the corresponding function $\max\{q.s, \frac{\gamma+1}{q.s}\}$ tends to infinity. As soon as a bound on the possible period is obtained, one can use Proposition 3.5.1 to get a bound on the possible preperiod. It suffices then to test a finite number of admissible preperiod/period pairs or, equivalently, a finite number of ultimately periodic sets Y and compare those with the given set X. In order to test if $X = Y$ we construct $L' = 0^* \operatorname{rep}_k(Y)$ and we have

$$X = Y \quad \Leftrightarrow \quad L = L'.$$

So we only have to test whether $L \backslash L' = \emptyset$ and $L' \backslash L = \emptyset$. Then our problem is decidable since the emptiness problem of the languages $L \backslash L'$ and $L' \backslash L$ is decidable.

Chapter 4

Automaticity

4.1 Introduction

In previous chapters, regular sets play a somehow crucial role since a set is recognizable for a numeration system if the language of its representations in this system is regular. But what happens if we deal with non-regular languages? Recall that some properties that can be tested algorithmically for regular sets become undecidable for context-free languages. Then it is quite natural to introduce the following notion. Automaticity is a measure of how close a non-regular language is to being regular. We can approximate a non-regular language L by considering a regular language L' such that the words of length at most n in L are exactly the words of length at most n in L'. The automaticity $A_L(n)$ of L is the number of states of a smallest deterministic finite automaton accepting some n-th order approximation L' to L, i.e., the words of length at most n of L' are exactly those of L. Non-deterministic automaticity can be defined similarly. In this chapter, we study the automaticity of the language of primitive words, the language of unbordered words and the set of representations of irreducible polynomials over a finite field. The material of this chapter can be found in [LR13].

Automaticity was first introduced by Trakhtenbrot [Tra64] who gives some upper and lower bounds on the automaticity with a slightly different definition than the one considered in this chapter. His results were improved by Grinberg and Korshunov [GK66]. Karp [Kar67] is the first person who studied the concept of automaticity exactly as we have defined it; he proved Theorem 4.2.9 given page 88. Breitbart [Bre71] studied the automaticity of the set of k-th powers in prime base p and proved that $A_L(n) = \Omega(p^{n/k})$. Dwork and Stockmeyer [DS89] introduced what they call a "measure of non-regularity". Their measure, as a function of n, is defined to be the maximum

number of distinct words that are pairwise n-dissimilar (see Definition 4.2.4
page 86). As proved by Kaneps and Freivalds [KF90], this measure coincides
with $A_L(n)$ (see Theorem 4.2.6). In 1996, Shallit and Breitbart [SB96] wrote
a survey of the basic results concerning automaticity known at that time and
gave some bounds on the deterministic and non-deterministic automaticity
of non-regular languages on alphabets Σ such that $|\Sigma| \geq 2$. The unary
case was studied in [PRS97] by Pomerance, Robson and Shallit. In [GS98],
Glaister and Shallit studied the closure properties of languages with polyno-
mial automaticity and showed how to construct a context-free language of
automaticity arbitrarily close to the maximum possible. In [Sha96], Shallit
introduced the k-automaticity $A_s^k(n)$ of a sequence s over a finite alphabet,
which is the smallest possible number of states in any DFA that for all $i \in \mathbb{N}$,
takes $\text{rep}_k(i)$ as input and computes $s(i)$.

For instance, consider the self-generating set

$$X = F^\omega(I) = \{0, 1, 3, 7, 9, 15, 19, 21, 27, 31, \ldots\}$$

generated by $I = \{0\}$ and $F = \{\varphi_0 : x \mapsto x, \varphi_1 : x \mapsto 2x + 1, \varphi_2 : x \mapsto 3x\}$
(see Definition 2.1.7 page 26). By Corollary 2.5.8 page 50, we know that X
is not k-recognizable for any integer base $k \geq 2$. In particular, the language
$L = \text{rep}_2(X)$ is not regular and the first words of L in genealogical order are
given by

$$\text{rep}_2(X) = \{\varepsilon, 1, 11, 111, 1001, 1111, 10011, 10101, 11011, 11111, \ldots\}.$$

The automata represented in Figure 4.1 show that $A_L(3) \leq 2$ and $A_L(4) \leq 5$.
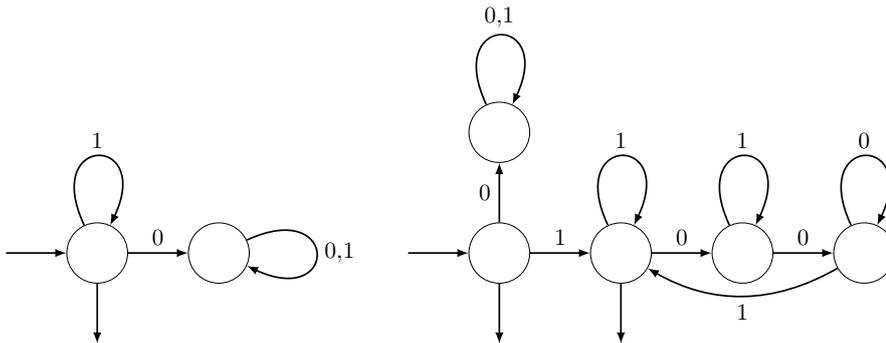
Figure 4.1: Automata accepting some 3rd and 4th approximations to
$\text{rep}_2(X)$.

In the first part of this chapter, we give bounds for the non-deterministic
automaticity of the language of primitive words and the language of unbor-
dered words. A word is *primitive* if it is not a power of a smaller word. A

word is *unbordered* if it has no non-trivial period. The first words of the languages Q_2 of primitive words over Σ_2 and UB_2 of unbordered words over Σ_2 are given in Table 4.1.

Q_2	UB_2
0	
1	
01	01
10	10
001	001
010	
011	011
100	100
101	
110	110
$\vdots$	$\vdots$

Table 4.1: First primitive and unbordered words over Σ_2.

The language of primitive words has been well studied (see the survey by Lischke [Lis11], for example). It is not difficult to show that the language of primitive words is not regular, but it is a long-standing open problem to show that this language is not context-free. It is also not difficult to show that the language of unbordered words is not regular. For a proof that this language is not context-free see [RSW11].

The set of base k representations of the prime numbers is not a regular language for any base k. Shallit [Sha96] gave a lower bound on the automaticity of the set of prime numbers in any base.

Theorem 4.1.1 ([Sha96]). *The set P of base k representations of the prime numbers has automaticity*

$$A_P(n) = \Omega(n^{1/43})$$

for all integers $k \geq 2$.

In the second part of the chapter, we consider a similar problem in the setting of polynomials over a finite field and, using the same kind of arguments as Shallit, we give a bound on the automaticity of the set of irreducible polynomials over a finite field. Given a fixed non-constant polynomial b, one

can also define the base b representation for such polynomials (see for example [Rig08]). Rigo and Waxweiler [RW11] proved that the set of base b representations of the irreducible polynomials is again a non-regular language for any base b. We obtain our bound for the automaticity using arguments similar to those of [Sha96].

There is an interesting connection between primitive words and irreducible polynomials over a finite field. The number of primitive words of length n over an alphabet of size q is

$$\sum_{d|n} \mu(d) q^{n/d}, \tag{4.1}$$

where μ is the Möbius function defined as follows (see [Lot83, Section 1.3])

- $\mu(n) = 1$ if n is a square-free positive integer with an even number of prime factors;

- $\mu(n) = -1$ if n is a square-free positive integer with an odd number of prime factors;

- $\mu(n) = 0$ if n is not square-free.

Similarly, the number of monic irreducible polynomials of degree n over the finite field with q elements is (see [LN83], p.91)

$$\frac{1}{n} \sum_{d|n} \mu(d) q^{n/d}.$$

This is equal to the number of equivalence classes of primitive words of length n under the conjugacy relation $x \sim y$ if x is a cyclic shift of y. For an explicit bijection between the set of irreducible polynomials and the set of primitive necklaces, see [Reu93, Section 7.6.2].

4.2   Definitions and basic properties

Definition 4.2.1. Let L be a language over a finite alphabet Σ. A language L' is an *n-th order approximation* to L if

$$L' \cap \Sigma^{\leq n} = L \cap \Sigma^{\leq n},$$

i.e., the words of L' of length at most n are exactly those of L. We define the *automaticity* $A_L(n)$ of a language L to be the number of states of a smallest DFA accepting some n-th order approximation to L. Similarly, the *nondeterministic automaticity* $N_L(n)$ of a language L is the number of states of a smallest NDFA accepting some n-th order approximation to L.

Example 4.2.2. Let $L = \{0^n 1^n : n \geq 0\}$. The automaton in Figure 4.2 shows that $A_L(6) \leq 8$ (In fact, one can show that $A_L(6) = 7$.).
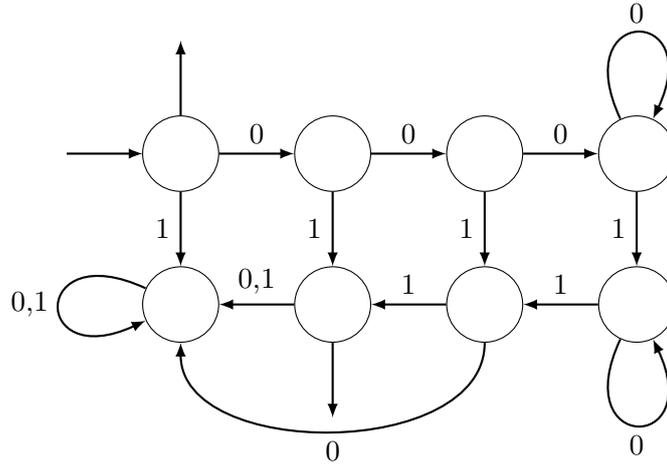

Figure 4.2: DFA accepting a 6-th order approximation of $L = \{0^n 1^n : n \geq 0\}$.

The proofs of the following properties come easily from the previous definition.

Theorem 4.2.3. *Let L be a language over a finite alphabet Σ.*

1. *For all $n \geq 0$, we have $A_L(n) \leq A_L(n+1)$ and $N_L(n) \leq N_L(n+1)$.*

2. *[DS89, KF90] The language L is regular if and only if $A_L(n) = O(1)$. The same statement holds for $N_L(n)$.*

3. *For all $n \geq 0$, we have $N_L(n) \leq A_L(n) \leq 2^{N_L(n)}$.*

4. *For all $n \geq 0$, we have $A_L(n) = A_{\bar{L}}(n)$ where $\bar{L}$ denotes the complement of L.*

5. *For all $n \geq 0$, we have*

$$A_L(n) \leq 2 + \sum_{w \in L \cap \Sigma^{\leq n}} |w| \leq 2 + n \#(L \cap \Sigma^{\leq n})$$

and

$$N_L(n) \leq 1 + \sum_{w \in L \cap \Sigma^{\leq n}} |w| \leq 1 + n \#(L \cap \Sigma^{\leq n})$$

Definition 4.2.4. Let $x, y \in \Sigma^*$. We say that x and y are *n-similar for L* if for all $z \in \Sigma^*$ with $|xz|, |yz| \leq n$, we have $xz \in L$ if and only if $yz \in L$. If x and y are *n*-similar for L, we write $x \sim_{L,n} y$, or $x \sim_n y$ if the context is clear. If x and y are not *n*-similar, then they are *n-dissimilar for L*.

Example 4.2.5. Consider the alphabet $\Sigma = \{a, b\}$ and the language $L = a^* b^*$. The words b and a are 2-dissimilar for L since $ba \notin L$ and $aa \in L$ but ab and abb are *n*-similar for L for all n.

Theorem 4.2.6 ([KF90]). *Let $L \subseteq \Sigma^*$. For all $n \geq 0$, $A_L(n)$ is the maximum possible cardinality of a set of pairwise n-dissimilar words for L.*

Proof. Let $d_{L,n}$ denote the maximum possible cardinality of a set of pairwise *n*-dissimilar words for L.

We consider the case $n = 0$ separately. Evidently, $d_{L,0} = 1$. On the other hand, the minimal automaton recognizing the 0-th order approximation to L (either containing ε or not containing it) needs no more than one state. Hence, Theorem 4.2.6 holds for $n = 0$.

Let $n > 0$. A word $w \in \Sigma^{\leq n}$ is called *short* if there exists no word $w' \in \Sigma^{\leq n}$ such that $|w'| < |w|$ and $w' \sim_n w$. Since L and n are fixed for the rest of the proof, we denote $d_{L,n}$ by d. We take d pairwise *n*-dissimilar words $w_1, w_2, \ldots, w_d \in \Sigma^{\leq n}$ (Their existence is implied by the definition of d). For every $i \in [\![1, d]\!]$ we fix a word w'_i as the shortest word with the property $w_i \sim_n w'_i$. (If several words of the same minimum length exist, we take for instance the lexicographically smallest). Clearly $|w'_i| < |w_i|$.

Now we prove that w'_i is short. Assume the contrary. Then there exists $w''_i \in \Sigma^{\leq n}$ such that $|w''_i| < |w'_i|$ and $w''_i \sim_n w'_i$. Since $|w'_i| \leq \max\{|w_i|, |w''_i|\}$, $w'_i \sim_n w_i$, $w'_i \sim_n w''_i$, we have that $w_i \sim_n w''_i$ and we get a contradiction, since w'_i is the shortest word *n*-similar to w_i. Hence, all the words w'_i ($i = 1, 2, \ldots, d$) are short.

Now we prove that for distinct i, j the words w'_i and w'_j are *n*-dissimilar. Assume the contrary, namely, assume $w'_i \sim_n w'_j$. From this and $|w'_i| \leq \max\{|w_i|, |w'_j|\}$ it follows that $w'_j \sim_n w_i$. Now we obtain the same type of conclusion from $w'_j \sim_n w_j$ and $|w'_j| \leq \max\{|w_i|, |w_j|\}$. We get $w_i \sim_n w_j$ and then a contradiction. Thus all the words $w'_1, w'_2, \ldots w'_d$ are short and pairwise *n*-dissimilar with respect to L.

Assume to the contrary that there is a deterministic finite automaton with less than d states recognizing an *n*-th order approximation to L. Then there are distinct i, j such that w'_i and w'_j move the automaton from the initial state to the same current state. But then the automaton cannot

distinguish between w_i' and w_j'. Hence for arbitrary $w \in \Sigma^*$ either both $w_i'w$ and $w_j'w$ are accepted or the two words are both rejected. We have

$$(\forall w \in \Sigma^*) \quad (w_i'w \in \Sigma^{\leq n} \text{ and } w_j'w \in \Sigma^{\leq n}) \Rightarrow (w_i'w \in L \Leftrightarrow w_j'w \in L).$$

Hence $w_i' \sim_n w_j'$ and we get a contradiction.

To conclude the proof it remains to construct a deterministic finite automaton with d states recognizing an n-th order approximation to L. Since $\{w_1', w_2', \ldots, w_d'\}$ is a maximal set of pairwise n-dissimilar words, for arbitrary $w \in \Sigma^{\leq n}$ there is an i such that $w \sim_n w_i'$. Hence, for some i, we have $\varepsilon \sim_n w_i'$. From the shortness of w_i' if follows that $w_i' = \varepsilon$. For the sake of brevity, we assume $w_1' = \varepsilon$. We define the function $\delta : \{q_1, q_2, \ldots, q_d\} \times \Sigma \to \{q_1, q_2, \ldots, q_d\}$ as follows. If $i \in [\![1, d]\!]$, $|w_i'| < n - 1$, $v \in \Sigma$ then $\delta(q_i, v)$ is defined to be q_j such that $w_i'v \sim_n w_j'$ (Such a value j exists because $w'v \in \Sigma^{\leq n}$). If several distinct j with this property are found, we take one of them. For those pairs (i, v) where $|w_i'| = n$, the value $\delta(q_i, v)$ is not really needed and we define, for example, $\delta(q_i, v) = q_1$. We consider the automaton $\mathcal{A}$ with the set of states $Q = \{q_1, \ldots, q_d\}$, the initial state q_1, the set of accepting states $F = \{q_i : w_i' \in L\}$ and the transition function δ. We extend the function δ to the domain $Q \times \Sigma^*$ in the standard way. Let $\delta(q_i, w)$ be used to denote the state of $\mathcal{A}$ into which $\mathcal{A}$ is moved after reading w, provided it has been in q_i initially.

By induction on the length of w we prove

$$(\forall w \in \Sigma^{\leq n}) \quad \delta(q_1, w) = q_i \;\Rightarrow\; w \sim_n w_i'. \tag{4.2}$$

Since $\delta(q_1, \varepsilon) = q_1$ and $\varepsilon = w_1' \sim_n \varepsilon$ then for $w = \varepsilon$ the property holds. Let (4.2) hold for the words of length k ($k < n$). We will prove (4.2) for an arbitrary word $wv \in \Sigma^{\leq n}$ where $w \in \Sigma^k$ and $v \in \Sigma$. Let $\delta(q_1, w) = q_i$. By the induction hypothesis, $w \sim_n w_i'$. Since the word w_i' is short, $|w_i'| < |w| = k < n$. Hence, $w_i'v \in \Sigma^{\leq n}$. It follows that $wv \sim_n w_i'v$. Let $\delta(q_i, w) = q_j$, i.e., $w_i'v \sim_n w_j'$. Since $w_i'v \sim_n wv$, $w_i'v \sim_n w_j'$ and $|w_i'v| \leq \max\{|wv|, |w|\}$, it follows that $wv \sim_n w_j'$. Taking into consideration $\delta(q_1, wv) = \delta(\delta(q_1, w), v) = \delta(q_i, v) = q_j$ we have the property (4.2) for the word wv.

Let $w \in \Sigma^{\leq n}$ and $\delta(q_1, w) = q_j$. It follows from (4.2) that $w \sim_n w_j'$ and this implies $w \in L \Leftrightarrow w_j' \in L$. Hence, $q_j \in F \Leftrightarrow w_j' \in L \Leftrightarrow w \in L$. Hence, the automaton $\mathcal{A}$ recognizes a n-th order approximation of the language L.
$\square$

Example 4.2.7. Let $L = \{0^n 1^n : n \geq 0\}$. Then $A_L(n) \geq m + 1$ for $n = 2m, 2m + 1$, since $\{\varepsilon, 0, 00, \ldots, 0^m\}$ is a set of pairwise $2m$-dissimilar

words for L. To see this, consider 0^j and 0^k for $0 \le j < k \le n$. Then $0^j 1^j \in L$ and $0^k 1^j \notin L$.

Let U be a finite set of words. We say that U is a set of *uniformly n-dissimilar words for L* if for each $x \in U$ there exists z such that

  (i) $|xz| \le n$ and $xz \in L$ and

  (ii) for each $y \in U$ such that $x \ne y$, we have $|yz| \le n$ and $yz \notin L$.

Theorem 4.2.8 ([GS98]). *Let $L \subseteq \Sigma^*$ and let U be a set of uniformly n-dissimilar words for L. Then $N_L(n) \ge |U|$.*

Proof. Consider a word $u \in U$. By definition, there exists a witness word w satisfying conditions (i) and (ii). Let $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$ be any non-deterministic finite automaton that accepts an n-th order approximation to L. Now $uw \in L$, and since $\mathcal{A}$ accepts all words in L of length at most n, we have $\delta(q_0, uw) \cap F \ne \emptyset$. Hence, there exists at least one state $q \in \delta(q_0, u)$ such that $p \in \delta(q, w)$, where $p \in F$.

However, for every other word $v \in U$, with $v \ne u$, we must have $q \notin \delta(q_0, v)$, since if $q \in \delta(q_0, v)$, we would have $p \in \delta(q_0, vw)$ and so $vw \in L$, a contradiction (since $|vw| \le n$). Hence, every set $\delta(q_0, u)$ contains a state q which does not appear in any other set $\delta(q_0, v)$ for $u \ne v$. It follows that there must be at least $|U|$ different states in Q. $\square$

The following theorem from Karp gives a lower bound on the deterministic automaticity of non-regular languages.

Theorem 4.2.9 ([Kar67]). *Let L be a non-regular language over a finite alphabet Σ. Then $A_L(n) \ge \frac{n+3}{2}$ for infinitely many n.*

Breitbart and Shallit proved that the bound given by Karp is tight.

Theorem 4.2.10 ([SB96]). *The bound in Theorem 4.2.9 is best possible in the sense that the result is not true if the "2" in the denominator is replaced by any smaller positive real number, nor if the "3" is replaced by any larger real number.*

4.3 Primitive and unbordered words

In this section, we study the non-deterministic automaticity of the languages of primitive words and unbordered words over a finite alphabet.

Definition 4.3.1. Let $k \geq 2$ be an integer. A word y is a *k-power* if y can be written as $y = x^k$ for some non-empty word x. If y cannot be so written for any $k \geq 2$, then y is *primitive*.

It is easy to prove the following property.

Proposition 4.3.2. *The language of primitive words over a finite alphabet is not regular.*

Definition 4.3.3. Bordered words are generalizations of powers. We say that a word x is *bordered* if there exist words $u, v, w \in \Sigma^+$ such that $x = uv = wu$. In this case, the word u is said to be a *border* for x. Otherwise, x is *unbordered*.

Let $w = w_0 \cdots w_{\ell-1}$ and let $p < \ell$. The word w has a *period* p if $w_i = w_{i+p}$ for all $0 \leq i \leq \ell - p - 1$. Note that a word is unbordered if it has no period.

Example 4.3.4. The word $w = church$ is primitive and bordered. Its period is 4 since $w_i = w_{i+4}$ for all $0 \leq i \leq 1$.

The following property is well known.

Proposition 4.3.5. *The language of unbordered words over a finite alphabet is not regular.*

Indeed this language is not context-free (see [RSW11] or [Sha09, Exercise 16,p.134]), and thus not regular.

We recall the notation $O(\cdot)$ and $\Omega(\cdot)$. Let f and g be functions from $\mathbb{N}$ to $\mathbb{R}$. The function f is $O(g)$ if there exist $C > 0$ and n_0 such that, for all $n > n_0$, we have $f(n) \leq C \cdot g(n)$. The function f is $\Omega(g)$ if there exist $C > 0$ and n_0 such that for all $n > n_0$ we have $f(n) \geq C \cdot g(n)$.

The following theorem gives a lower bound on the non-deterministic automaticity of the set of primitive words over Σ_k.

Theorem 4.3.6. *Let $\varepsilon > 0$ be a real number. The non-deterministic auto-*

maticity of the set Q_k of primitive words over the alphabet $\Sigma_k = \{0, \ldots, k-1\}$ is such that

$$N_{Q_k}(n) \geq \beta_k k^{\lfloor n/2 \rfloor} + O((k+\varepsilon)^{n/4}),$$

where β_k is a constant that depends only on k.

Proof. For $n \geq 0$, we will define a set of uniformly n-dissimilar words as follows. Let

$$D_n = \{w \in \Sigma_k^{\lfloor n/2 \rfloor} \mid w \text{ is unbordered }\}.$$

To show that the words in D_n are uniformly n-dissimilar, let $x, y \in D_n$, $x \neq y$. Observe that xx is not primitive, but yx is, for if yx were not primitive, then $yx = z^\ell$ for some word z and some $\ell \geq 3$. In this case $|yx|/\ell$ is a period of y, and hence y is bordered, which contradicts the definition of D_n.

Guibas and Odlyzko [GO81, Theorem 7.2] gave the following formula for the size of D_n (see also [Nie73]): there exists a constant β_k such that

$$\#D_n = \beta_k k^{\lfloor n/2 \rfloor} + O((k+\varepsilon)^{n/4}).$$

By Theorem 4.2.8 we have $N_{Q_k}(n) \geq \#D_n$, which is the desired result.   □

Note that Guibas and Odlyzko gave an explicit formula for the β_k, which permits one to calculate β_k to any desired degree of accuracy. For example, if $k = 2$, we have $\beta_2 = 0.26771654 \cdots$.

Next we give an upper bound on the non-deterministic automaticity of Q_k.

Theorem 4.3.7. *The non-deterministic automaticity of the set Q_k of primitive words over the alphabet $\Sigma_k = \{0, \ldots, k-1\}$ is such that*

$$N_{Q_k}(n) \leq \left(\frac{2k^{3/2}}{(k-1)(\sqrt{k}-1)} \right) k^{n/2} + n^3 k^{n/3}$$

Proof. For each $n \geq 0$ we construct a deterministic automaton that accepts all words of length at most n in the complement of Q_k. The automaton is constructed as follows. First consider the language of square words (2-powers) of length i. We can construct an automaton accepting this language by first constructing the complete k-ary tree with $k^{i/2}$ leaves so that each path from the root to a leaf is labeled by a different word of length $i/2$. We then make a copy of this tree, but reflected, so that the arrows are directed away from the leaves towards the root of the tree. The leaves of the
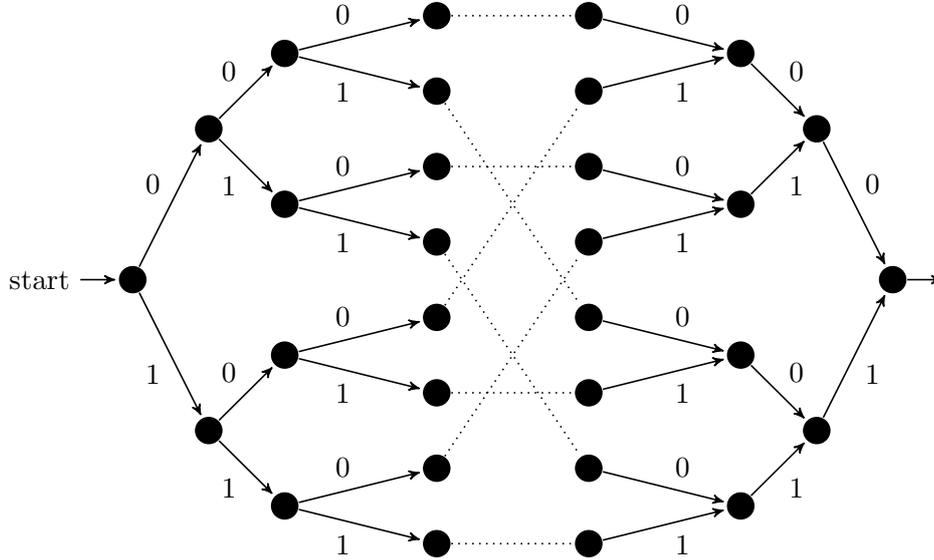
Figure 4.3: An automaton accepting binary squares of length 6.

first tree are identified with the leaves of the second tree. This construction is illustrated in Figure 4.3, which shows the automaton accepting all binary squares of length 6. In the figure, dotted lines connect states to be identified, and transitions not shown go to a sink state.

The left tree has

$$\frac{k^{(i/2)+1} - 1}{k - 1}$$

states, so the automaton for the squares of length i has at most

$$2 \left(\frac{k^{(i/2)+1} - 1}{k - 1} \right)$$

states.

For each $d > 2$, to accept d-powers of length i we simply construct a tree with $k^{i/d}$ leaves so that each path from the root to a leaf is labeled by a different d-power of length i. This tree has at most $ik^{i/d}$ states.

To create the automaton accepting all non-primitive words of length i, we can combine all of these automata, sharing edges and transitions whenever

possible. The resulting automaton has at most

$$2\left(\frac{k^{(i/2)+1}-1}{k-1}\right) + \sum_{\substack{d|i \\ d>2}} ik^{i/d}$$

$$\leq \quad \left(\frac{2k}{k-1}\right)k^{i/2} + \sum_{\substack{d|i \\ d>2}} ik^{i/d}$$

$$\leq \quad \left(\frac{2k}{k-1}\right)k^{i/2} + \sum_{\substack{d|i \\ d>2}} ik^{i/3}$$

$$\leq \quad \left(\frac{2k}{k-1}\right)k^{i/2} + i^2 k^{i/3}$$

states.

We can therefore construct an automaton accepting all non-primitive words of length at most n using at most

$$\sum_{i=1}^{n}\left(\left(\frac{2k}{k-1}\right)k^{i/2} + i^2 k^{i/3}\right)$$

$$\leq \quad \frac{2k}{k-1}\sum_{i=1}^{n}k^{i/2} + \sum_{i=1}^{n}i^2 k^{i/3}$$

$$\leq \quad \frac{2k}{k-1}\left(\frac{k^{(n+1)/2}-\sqrt{k}}{\sqrt{k}-1}\right) + \sum_{i=1}^{n}i^2 k^{i/3}$$

$$\leq \quad \frac{2k}{k-1}\left(\frac{k^{(n+1)/2}-\sqrt{k}}{\sqrt{k}-1}\right) + n^3 k^{n/3}$$

$$\leq \quad \left(\frac{2k^{3/2}}{(k-1)(\sqrt{k}-1)}\right)k^{n/2} + n^3 k^{n/3}$$

states. Since this automaton is deterministic, the automaton accepting Q_k has at most this many states as well. $\square$

Next we consider the language of unbordered words and we first need the following lemma.

Lemma 4.3.8 ([FS09]). *For each $\varepsilon > 0$ there exists j such that the number of words of length m over a k-letter alphabet that avoid 1^j is $\Omega((k-\varepsilon)^m)$.*

We can now prove the following theorem about the non-deterministic automaticity of the set of unbordered words over Σ_k.

Theorem 4.3.9. *Let $\varepsilon > 0$. The non-deterministic automaticity of the set UB_k of unbordered words over the alphabet Σ_k is*

$$N_{UB_k}(n) = \Omega((k - \varepsilon)^{n/2}).$$

Proof. Fix a j as in Lemma 4.3.8. For $n \geq 2(j + 2)$ we define

$$D_n = \{0w01^j : |w| = \lfloor n/2 \rfloor - (j + 2) \text{ and } w \text{ does not contain } 1^j\}.$$

To show that the words in D_n are uniformly n-dissimilar, let $x, y \in D_n$, $x \neq y$. Since $x, y \in D_n$, there exist w_1 and w_2 such that

$$x = 0w_1 01^j \quad \text{and} \quad y = 0w_2 01^j.$$

Clearly xx is bordered; however, xy is not bordered. If xy were bordered, then either w_2 contains 1^j or $y = x$, both of which are contradictions. By Lemma 4.3.8, we have

$$\begin{aligned} \#D_n &= \Omega((k - \varepsilon)^{n/2 - (j+2)}) \\ &= \Omega((k - \varepsilon)^{n/2}). \end{aligned}$$

Then we have the result by Theorem 4.2.8. $\qquad \square$

4.4 Irreducible polynomials

In this section we consider the automaticity of the language of representations of irreducible polynomials over a finite field with respect to some base b.

Let $\mathbb{F}_q$ be a field with q elements. Let $\mathbb{F}_q[X]$ be the polynomial ring over $\mathbb{F}_q$. If $f \in \mathbb{F}_q[X]$ we denote its degree by $\deg f$. Let B be an integer and let $\mathbb{F}_q[X]_{<B}$ denote the set of polynomials over $\mathbb{F}_q$ of degree strictly less than B. If b is a fixed non-constant polynomial, then any polynomial f can be uniquely written as

$$f = \sum_{i=0}^{\ell} c_i b^{\ell - i}, \quad c_0 \neq 0,$$

where each c_i has degree less than $\deg b$.

We define a function $\Psi : \mathbb{F}_q[X]_{<B} \to \mathbb{F}_q^B$ by

$$\Psi(f) := (\underbrace{0, \ldots, 0}_{B - N - 1}, F_0, \ldots, F_N)$$

if $f = F_0 X^N + \cdots + F_N$, i.e., $\Psi(f)$ is the vector containing coefficients of f with leading zeros if necessary. The word $\mathrm{rep}_b(f) := \Psi(c_\ell)\Psi(c_{\ell-1})\cdots\Psi(c_0)$ over the alphabet $\mathbb{F}_q^B$ is the b-representation of f. By convention, the representation of the zero polynomial is ε. Given a b-representation $w \in (\mathbb{F}_q^B)^*$, we denote its value in $\mathbb{F}_q[X]$ by $\mathrm{val}_b(w)$. Note that we have chosen to write f starting with the least significant "digit" and ending with the most significant "digit".

A set $\mathcal{T} \subseteq \mathbb{F}_q[X]$ is *b-recognizable* if the language

$$\mathrm{rep}_b(\mathcal{T}) = \{\mathrm{rep}_b(f) : f \in \mathcal{T}\} \subseteq (\mathbb{F}_q^B)^*$$

is regular. Rigo and Waxweiler [RW11] proved that for any base b, the set of irreducible polynomials over $\mathbb{F}_q$ is not b-recognizable.

Let b be a non-constant polynomial and $\mathcal{T} \subset \mathbb{F}_q[X]$. The *b-automaticity* of $\mathcal{T}$ is denoted by $A_{\mathcal{T}}^b(n)$ and is defined as the automaticity $A_L^b(n)$ of the language $L = \{\mathrm{rep}_b(f) \mid f \in \mathcal{T}\}$.

If f is a polynomial over $\mathbb{F}_q$, let $\Phi(f)$ denote the number of polynomials of degree less than $\deg f$ that are relatively prime with f.

Let m be a non-constant polynomial over $\mathbb{F}_q$. We recall the notion of a *Dirichlet character modulo m*, which is defined in the same manner as in the classical case of the integers. A Dirichlet character modulo m is a function $\chi : \mathbb{F}_q[X] \to \mathbb{C}$ such that

(a) $\chi(a + bm) = \chi(a)$ for all $a, b \in \mathbb{F}_q[X]$,

(b) $\chi(a)\chi(b) = \chi(ab)$ for all $a, b \in \mathbb{F}_q[X]$,

(c) $\chi(a) \neq 0$ if and only if $(a, m) = 1$.

A Dirichlet character modulo m induces a homomorphism from the set[1] $(\mathbb{F}_q[X]/m\,\mathbb{F}_q[X])^\star$ to $\mathbb{C}^\star$, and hence for all a, the value of $\chi(a)$ is either 0 or a root of unity. The *trivial Dirichlet character* χ_0 is defined by $\chi_0(a) = 1$ if $(a, m) = 1$ and $\chi_0(a) = 0$ otherwise. There are $\Phi(m)$ Dirichlet characters modulo m.

Let $\bar{\chi}$ be defined by $\bar{\chi}(a) = \overline{\chi(a)}$, where $\overline{z}$ denotes the complex conjugate of z. Let a and b be elements of $\mathbb{F}_q[X]$ relatively prime with m. We have the following *orthogonality relation*:

$$\sum_\chi \chi(a)\bar{\chi}(b) = \begin{cases} \Phi(m) & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases}$$

---

[1] Here, $X^\star$ denotes the set of invertible elements of the set X.

Let χ be a Dirichlet character modulo m. Let $L(s, \chi)$ be the Dirichlet L-series corresponding to χ, that is,

$$L(s, \chi) = \sum_{f \text{ monic}} \frac{\chi(f)}{q^{s \deg f}}, \qquad s \in \mathbb{C},$$

where the sum converges absolutely for $\Re(s) > 1$. We also have the Euler product representation

$$L(s, \chi) = \prod_{P} (1 - \chi(P) q^{-s \deg P})^{-1},$$

where the product is over all monic irreducible polynomials P. We also have

$$L(s, \chi_0) = \prod_{P|m} (1 - q^{-s \deg P}) \frac{1}{1 - q^{1-s}}. \qquad (4.3)$$

Let a and m be polynomials over $\mathbb{F}$ such that $(a, m) = 1$. Let $\#S_N(a, m)$ denote the number of monic irreducible polynomials of degree N congruent to a modulo m and let $M = \deg m$. Rosen [Ros02] proved that

$$\#S_N(a, m) \geq \frac{q^N}{N\Phi(m)} + O\left(\frac{q^{N/2}}{N}\right).$$

This result is a refinement of the analogue of Dirichlet's theorem on primes in arithmetic progressions due to Kornblum and Landau. However, this result is not sufficiently precise for our purposes. In particular, we need to know explicitly the constants hidden in the big-Oh term. We therefore prove a slightly more precise version of this result below. We follow Rosen's proof, but we replace all big-Oh terms with explicit bounds.

Theorem 4.4.1. *Let a and m be polynomials over $\mathbb{F}$ such that $(a, m) = 1$. The number of monic irreducible polynomials of degree N congruent to a modulo m is such that*

$$\#S_N(a, m) \geq \frac{q^N}{N\Phi(m)} - \frac{(M+1)q^{N/2}}{N} - \frac{Mq^M}{N\Phi(m)}.$$

Proof. We can rewrite (4.3) in terms of the variable $u = q^{-s}$ as

$$L^*(u, \chi_0) = \prod_{P|m} (1 - u^{\deg P}) \frac{1}{1 - qu}. \qquad (4.4)$$

For $\chi \neq \chi_0$, the series $L(s, \chi)$ is a polynomial in q^{-s} of degree at most $M - 1$, and hence we have

$$L^*(u, \chi) = \prod_{i=1}^{M-1} (1 - \alpha_i(\chi)u), \tag{4.5}$$

where the $\alpha_i(\chi)$ are complex numbers.

We can also rewrite the Euler product representation of L as

$$L(s, \chi) = \prod_{d=1}^{\infty} \prod_{\substack{P \nmid m \\ \deg(P) = d}} (1 - \chi(P)q^{-ds})^{-1},$$

and make the substitution $u = q^{-s}$ to obtain

$$L^*(u, \chi) = \prod_{d=1}^{\infty} \prod_{\substack{P \nmid m \\ \deg(P) = d}} (1 - \chi(P)u^d)^{-1}. \tag{4.6}$$

We will take the logarithmic derivatives of (4.4)–(4.6), express the results as series in u,

$$u\frac{d}{du} \log(L^*(u, \chi)) = \sum_{k=1}^{\infty} C_k(\chi)u^k,$$

and examine the coefficients $C_N(\chi)$ of u^N in these series. To perform these calculations we will make frequent use of the identity

$$u\frac{d}{du}(\log(1 - \alpha u)^{-1}) = \sum_{k=1}^{\infty} \alpha^k u^k,$$

where α is a complex number. This identity is valid for all u such that $|u| < |\alpha|^{-1}$.

Taking the logarithmic derivative of (4.4), we obtain

$$
\begin{aligned}
u\frac{d}{du} \log L^*(u, \chi_0) &= u\sum_{P|m} -\deg P \frac{u^{\deg P - 1}}{1 - u^{\deg P}} + u\frac{d}{du} \log \frac{1}{1 - qu} \\
&= -\sum_{P|m} \deg P \frac{u^{\deg P}}{1 - u^{\deg P}} + \sum_{k=1}^{\infty} q^k u^k \\
&= -\sum_{P|m} \deg P \left(\sum_{j=0}^{\infty} u^{j \deg P + 1} \right) + \sum_{k=1}^{\infty} q^k u^k
\end{aligned}
$$

The coefficient of u^N in the first sum is at most Mq^M and is exactly q^N in the second sum. Hence

$$C_N(\chi_0) \geq q^N - Mq^M. \tag{4.7}$$

Taking the logarithmic derivative of (4.5), we obtain

$$
\begin{aligned}
u\frac{d}{du}\log L^*(u, \chi_0) &= \sum_{i=1}^{M-1} u\frac{d}{du}\log(1 - \alpha_i(\chi)u) \\
&= -\sum_{i=1}^{M-1}\sum_{k=1}^{\infty} \alpha_i(\chi)^k u^k.
\end{aligned}
$$

By Weil's analogue of the Riemann Hypothesis for function fields, we have that $|\alpha_i(\chi)|$ is either 1 or $\sqrt{q}$. Hence, the coefficient of u^N in the above sum is

$$C_N(\chi) \geq -(M-1)q^{N/2}. \tag{4.8}$$

Since there are $\Phi(m)$ Dirichlet characters modulo m, applying (4.7) and (4.8) gives

$$\sum_{\chi} \bar{\chi}(a)C_N(\chi) \geq q^N - Mq^M - (\Phi(m) - 1)(M-1)q^{N/2}. \tag{4.9}$$

Taking the logarithmic derivative of (4.6), we obtain (after some calculation)

$$
\begin{aligned}
C_N(\chi) &= N \sum_{\substack{k,P \\ k\deg P=N}} \deg P\,\chi(P)^k \\
&= N \sum_{\deg P=N} \chi(P) + \sum_{\substack{d|N \\ d\leq N/2}} d \sum_{\deg P=d} \chi(P)^{N/d}.
\end{aligned}
$$

The absolute value of the sum

$$\sum_{\deg P=d} \chi(P)^{N/d}$$

is at most the number of monic irreducible polynomials P of degree d, which is at most q^d/d. Hence,

$$\sum_{\substack{d|N \\ d\leq N/2}} d \sum_{\deg P=d} \chi(P)^{N/d} \leq \sum_{d=1}^{\lfloor N/2 \rfloor} q^d \leq 2q^{N/2},$$

and

$$C_N(\chi) \leq N \sum_{\deg P = N} \chi(P) + 2q^{N/2}.$$

This gives the inequality

$$
\begin{aligned}
\sum_\chi \bar{\chi}(a) C_N(\chi) &\leq \sum_\chi \bar{\chi}(a) \left(N \sum_{\deg P = N} \chi(P) + 2q^{N/2} \right) \\
&= N \sum_\chi \bar{\chi}(a) \sum_{\deg P = N} \chi(P) + 2q^{N/2} \sum_\chi \bar{\chi}(a) \\
&= N \Phi(m) \# S_N(a, m) + 2q^{N/2} \Phi(m), \qquad (4.10)
\end{aligned}
$$

where we have applied the orthogonality relations for Dirichlet characters in the last step.

From (4.9) and (4.10), we obtain

$$q^N - Mq^M - (\Phi(m) - 1)(M - 1)q^{N/2} \leq N\Phi(m) \# S_N(a, m) + 2\Phi(m)q^{N/2},$$

whence

$$\# S_N(a, m) \geq \frac{q^N}{N\Phi(m)} - \frac{(M+1)q^{N/2}}{N} - \frac{Mq^M}{N\Phi(m)}.$$

<div style="text-align:right">□</div>

The following corollary is the main tool for the proof of Theorem 4.4.5. A similar result has been proved by Hsu [Hsu96, Corollary 3.4].

Corollary 4.4.2. *There exists a constant C such that for all $N \geq CM$, the quantity $\# S_N(a, m)$ is positive.*

Proof. Since $\Phi(m) \leq q^M$, we have

$$\# S_N(a, m) \geq \frac{1}{N}(q^{N-M} - (M+1)q^{N/2} - Mq^M),$$

and the result follows.   □

Remark 4.4.3. Since there exist irreducible polynomials of all degree, there exist monic irreducible polynomials $f \in \mathbb{F}_q$ such that $\deg c_0 = 0$ with the notation introduced above. In this case, note that we have

$$\deg f = \deg(b^{|\operatorname{rep}_b(f)| - 1}) = (|\operatorname{rep}_b(f)| - 1) \deg b.$$

The proof of the following lemma is similar to that of [Sha96, Lemma 6], which is in turn based on an idea found in [HS68] and [All68].

Lemma 4.4.4. *Let $d \in \mathbb{F}_q[X]$ such that $\deg d > 0$ and let $f, g \in \mathbb{F}_q[X]_{<\deg d}$ such that $f \neq g$ and $(d, f) = (d, g) = 1$. Then there exists a constant C_q and a polynomial h such that $hd + f$ is irreducible and $hd + g$ is not irreducible, where $\deg h \leq C_q \deg d$.*

Proof. By Corollary 4.4.2 there exists a polynomial h_0 and a constant C such that $r = h_0 d + f$ is irreducible, with $\deg r > C \deg d$.

Furthermore, we have

$$\deg r > C \deg d \quad \Leftrightarrow \quad \deg h_0 + \deg d > C \deg d$$
$$\Leftrightarrow \quad \deg h_0 > (C - 1) \deg d.$$

Then there exists h_0 such that r is irreducible, where $\deg h_0 \leq C \deg d$. If $s = h_0 d + g$ is reducible, we are done.

Otherwise, since $(sd, sd + r) = 1$, by Theorem 4.4.2 again, there exists h_1 such that $h_1(sd) + (sd + r) = ((h_1 + 1)s + h_0)d + f$ is irreducible, where $\deg h_1 \leq C(\deg sd) = C(\deg s + \deg d)$. However, $h_1(sd) + (sd + s) = ((h_1 + 1)s + h_0)d + g$ is a multiple of s and hence is reducible. Therefore we set $h = s(h_1 + 1) + h_0$. Furthermore, we have

$$\begin{aligned}
\deg h \quad &\leq \quad \max\{\deg h_1 + \deg s, \deg h_0\} \\
&\leq \quad \max\{C(\deg s + \deg d) + \deg s, C \deg d\} \\
&\leq \quad (C + 1)(\deg h_0 + \deg d) + C \deg d \\
&\leq \quad (2C + 1) \deg d + (C + 1)C \deg d \\
&\leq \quad (C^2 + 3C + 1) \deg d.
\end{aligned}$$

Then we set $C_q = C^2 + 3C + 1$ and this concludes the proof. $\square$

Theorem 4.4.5. *Let b be a fixed non-constant polynomial and let*

$$B(n) := \left(\frac{n - 1}{1 + C_q} - 1 \right) \deg b$$

where C_q is as in Lemma 4.4.4. Then the set $\mathcal{S}$ of monic irreducible polynomials over $\mathbb{F}_q$ has b-automaticity

$$A_{\mathcal{S}}^b(n) \geq q^{B(n)}/B(n) + O(q^{B(n)/2}/B(n)).$$

Proof. To prove Theorem 4.4.5 we will contruct a set D_n of n-dissimilar words for $\mathrm{rep}_b(\mathcal{S})$. Let C_q be as in Lemma 4.4.4. Let

$$D_n = \{[f]_b \mid f \in \mathcal{S}, (f, b) = 1, \deg f = \left(\frac{n-1}{1+C_q} - 1\right) \deg b\}.$$

By Remark 4.4.3, note that all words in D_n are of length $\frac{n-1}{1+C_q}$. Consider two distinct elements $x, y \in D_n$. Let $f = \mathrm{val}_b(x)$, $g = \mathrm{val}_b(y)$. We have $(f, b) = 1 = (g, b)$ thus $(f, b^{|x|}) = 1 = (g, b^{|x|})$ and $\deg f = \deg g < \deg b^{|x|}$. Then, by Lemma 4.4.4, there exists h such that $hb^{|x|} + f$ is irreducible and $hb^{|x|} + g$ is not, where $\deg h \leq C_q \deg b^{|x|}$. Let $z = \mathrm{val}_b(h)$. Then $xz = \mathrm{rep}_b(hb^{|x|} + f) \in \mathrm{rep}_b(\mathcal{S})$ and $yz = \mathrm{rep}_b(hb^{|x|} + g) \notin \mathrm{rep}_b(\mathcal{S})$. Since $\deg b^{|z|-1} \leq \deg h \leq C_q \deg b^{|x|}$, we have

$$(|z| - 1) \deg b \leq C_q |x| \deg b \quad \Leftrightarrow \quad |z| - 1 \leq C_q |x|$$
$$\Leftrightarrow \quad |z| \leq C_q |x| + 1$$

Hence $|xz| \leq \frac{n-1}{1+C_q} + C_q \frac{n-1}{1+C_q} + 1 = n$.

We now estimate the size of D_n. Let $B(n) = \left(\frac{n-1}{1+C_q} - 1\right) \deg b$. Note that there are $q^{B(n)}/B(n) + O(q^{B(n)/2}/B(n))$ monic irreducible polynomials in $\mathbb{F}_q[X]$ of degree $B(n)$. Since $\deg b$ is a constant, there are at most a constant number of polynomials f that divide b. Hence $\#D_n = q^{B(n)}/B(n) + O(q^{B(n)/2}/B(n))$. $\qquad \square$

Chapter 5

Multidimensional sets

5.1 Introduction

In this chapter we characterize the subsets of $\mathbb{N}^d$ that are simultaneously recognizable in all abstract numeration systems. Lecomte and Rigo [LR01] provided such a characterization for the case $d = 1$ based on the well-known correspondence between unary regular languages and ultimately periodic subsets of $\mathbb{N}$. When $d > 1$ we no longer have such a nice correspondence and the situation becomes somewhat more complicated. To obtain our characterization we instead use a classical decomposition theorem due to Eilenberg, Elgot, and Shepherson [EES69]. The motivation for studying such sets comes from the well-known result of Cobham (see Theorem 1.3.16) (and its multidimensional generalization due to Semenov) concerning the sets recognizable in integer bases. The material of this chapter can be found in [CLR12].

5.1.1 Recognizability in $\mathbb{N}$

Let $k \geq 2$ be an integer. Recall that a set $X \subseteq \mathbb{N}$ is *k-recognizable* (or *k-automatic*) if the language $\mathrm{rep}_k(X) = \{\mathrm{rep}_k(x) \colon x \in X\}$ is regular or, equivalently, if the characteristic sequence of X is k-automatic (see Definition 1.5.1). Cobham proved that a sequence is k-automatic if and only if it is the image under a coding of a fixed point of a k-uniform morphism and that the sets that are k-recognizable for all integer bases $k \geq 2$ are exactly the ultimately periodic sets.

In particular, recall that we say that a set $X \subseteq \mathbb{N}$ is *1-recognizable* if the language $\{a^n \colon n \in X\}$ consisting of the unary representations of the elements of X is accepted by a finite automaton. It is well known that a set of integers is 1-recognizable if and only if it is ultimately periodic.

Let $S = (L, \Sigma, <)$ be an abstract numeration system. Lecomte and Rigo [LR01] introduced the concept of S-recognizable sets, that is a generalization of k-recognizable sets (see Definition 1.4.1 and Definition 1.4.3) and they proved that any ultimately periodic set is S-recognizable for any abstract numeration system S. Suppose on the other hand that $X \subseteq \mathbb{N}$ is S-recognizable for every abstract numeration system S. Then, in particular, the set X must be 1-recognizable, and hence must be ultimately periodic. We therefore have that the sets of integers that are S-recognizable for all abstract numeration systems S are exactly the ultimately periodic sets.

The notion of automatic sequence can also be extended to the abstract numeration systems.

Definition 5.1.1. Let $u = (u_n)_{n \in \mathbb{N}}$ be an infinite sequence over a finite alphabet Σ' and $S = (L, \Sigma, <)$ be an abstract numeration system. The sequence u is *S-automatic* if there exists a DFAO $\mathcal{M} = (Q, \Sigma, \delta, q_0, \Sigma', \tau)$ such that $u_n = \tau(\delta(q_0, w))$ for all $n \geq 0$ and where $\mathrm{val}_S(w) = n$.

Rigo and Maes [RM02] extended Theorem 1.5.8 to the case of abstract numeration systems by proving that a sequence is S-automatic if and only if it is morphic.

5.1.2   Recognizability in $\mathbb{N}^d$

For the formal definitions we need to introduce the following "padding" function.

Definition 5.1.2. If $w_1, \ldots, w_d$ are finite words over the alphabet Σ, the *padding map*

$$(\cdot)^{\#} : (\Sigma^*)^d \to ((\Sigma \cup \{\#\})^d)^*$$

is defined by

$$(w_1, \ldots, w_d)^{\#} := (w_1 \#^{m - |w_1|}, \ldots, w_d \#^{m - |w_d|})$$

where $m = \max\{|w_1|, \ldots, |w_d|\}$. Here we write (ac, bd) to denote the concatenation $(a, b)(c, d)$.

If $R \subseteq (\Sigma^*)^d$, then

$$R^{\#} = \{(w_1, \ldots, w_d)^{\#} : (w_1, \ldots, w_d) \in R\}.$$

Note that R is not necessarily a language, whereas $R^{\#}$ is; that is, the set R consists of d-tuples of words over Σ, whereas $R^{\#}$ consists of words over the alphabet $(\Sigma \cup \{\#\})^d$.

Definition 5.1.3. Let $k \geq 2$ be an integer and $X \subseteq \mathbb{N}^d$. The set X is *k-recognizable* (or *k-automatic*) if the language $\mathrm{rep}_k(X)^{\#}$ is regular, with

$$\mathrm{rep}_k(X) = \{(\mathrm{rep}_k(n_1), \ldots, \mathrm{rep}_k(n_d)) \colon (n_1, \ldots, n_d) \in X\}$$

or, equivalently, if the characteristic sequence of X is k-automatic. Here the *characteristic sequence* $\mathbf{c}_X$ of X is defined by $\mathbf{c}_X(n_1, \ldots, n_d) = 1$, if $(n_1, \ldots, n_d) \in X$ and $\mathbf{c}_X(n_1, \ldots, n_d) = 0$, otherwise.

Definition 5.1.4. Let $k \geq 2$ be an integer and φ be a multidimensional morphism. This morphism is *k-uniform* if all images are squares with side k (see Definition 1.6.7 and Definition 1.6.1).

The following result extends Theorem 1.5.8 to the multidimensional case.

Theorem 5.1.5 ([Sal87])**.** *Let $k \geq 2$ and $d \geq 1$ be integers. A sequence u over $\mathbb{N}^d$ is k-automatic if and only if u is generated by a k-uniform morphism.*

Example 5.1.6. Consider the set $X \subset \mathbb{N}$ defined as

$$X = \{(n, m) \ : \ \forall k \geq 0, \text{ the same power } 2^k \text{ of 2 does not occur}$$
$$\text{in both binary expansions of } n \text{ and } m\}.$$

This set is 2-recognizable as $X^{\#}$ is accepted by the automaton depicted in Figure 5.1.
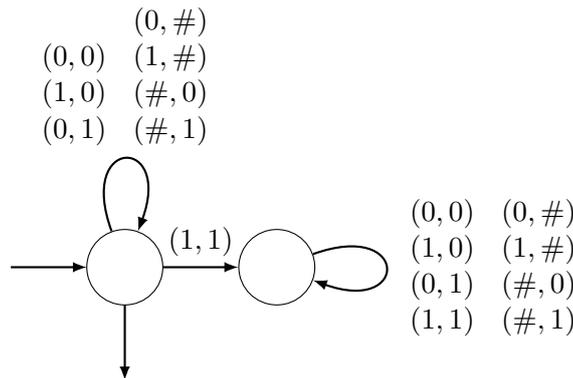
Figure 5.1: DFA accepting the set X of Example 5.1.6.

The characteristic sequence of the set X is defined by

$$\mathbf{c}_X(n, m) = \begin{cases} 0 \text{ if , for some } k \geq 0, \text{ the same power } 2^k \text{ of } 2 \text{ occurs in the} \\ \quad \text{binary expansions of } n \text{ and } m, \\ 1 \text{ otherwise.} \end{cases}$$

and $\mathbf{c}_X$ is generated by the 2-uniform morphism φ defined by

$$\varphi(1) = \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array} \quad , \quad \varphi(0) = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & 0 \\ \hline \end{array}.$$

Indeed, we have

$$\mathbf{c}_X = \varphi^{\omega}(1) = \begin{array}{|cc|cc|c} \vdots & & & & \iddots \\ \hline 1 & 0 & 0 & 0 & \\ 1 & 1 & 0 & 0 & \\ \hline 1 & 0 & 1 & 0 & \\ 1 & 1 & 1 & 1 & \cdots \\ \hline \end{array}$$

Definition 5.1.7. Let $S = (L, \Sigma, <)$ be an abstract numeration system. Let $X \subseteq \mathbb{N}^d$. The set X is *S-recognizable* (or *S-automatic*) if the language $\mathrm{rep}_S(X)^{\#}$ is regular, where

$$\mathrm{rep}_S(X) = \{(\mathrm{rep}_S(n_1), \ldots, \mathrm{rep}_S(n_d)) \colon (n_1, \ldots, n_d) \in X\}$$

or, equivalently, if the characteristic sequence of X is S-automatic. Let $k \geq 2$ be an integer. Observe that the notions of k-recognizability and 1-recognizability are special cases of S-recognizability. The set X is k-*recognizable* (or k-*automatic*) if it is S-recognizable for the abstract numeration system S built on the language consisting of the base k representations of the elements of X. The set X is 1-*recognizable* (or 1-*automatic*) if it is S-recognizable for the abstract numeration system S built on a^*.

Now we need the property of shape-symmetry that was first introduced by Maes and that is a natural generalization of uniform morphisms.

Definition 5.1.8. Let $\mu : \Sigma \to B_d(\Sigma)$ be a d-dimensional morphism having the d-dimensional infinite word x as a fixed point (See Definition 1.6.7). If the images $\mu(x(n, \ldots, n))$, for all $n \in \mathbb{N}$, of the letter on the diagonal of x are squares then x is said to be *shape-symmetric* (with respect to μ).

Example 5.1.9. Consider the morphism μ defined by

$$\mu(a) = \mu(f) = \begin{array}{|c|c|} \hline c & d \\ \hline a & b \\ \hline \end{array}, \quad \mu(b) = \begin{array}{|c|} \hline c \\ \hline e \\ \hline \end{array}, \quad \mu(c) = \begin{array}{|c|c|} \hline e & b \\ \hline \end{array}, \quad \mu(d) = \begin{array}{|c|} \hline f \\ \hline \end{array},$$

$$\mu(e) = \begin{array}{|c|c|} \hline g & d \\ \hline e & b \\ \hline \end{array} \;,\;\; \mu(g) = \begin{array}{|c|c|} \hline h & b \\ \hline \end{array} \;,\;\; \mu(h) = \begin{array}{|c|c|} \hline c & d \\ \hline h & b \\ \hline \end{array}.$$

This morphism has a fixed point $\mu^\omega(a)$ which is shape-symmetric.

$$\mu^\omega(a) =$$

$\vdots$										$\ddots$
h	b	f	e	b	e	b	f	h	b	
g	d	c	g	d	c	d	c	g	d	
e	b	e	e	b	a	b	e	e	b	
e	b	e	e	b	e	b	e	a	b	
g	d	c	g	d	g	d	c	c	d	
g	d	c	c	d	g	d	c	c	d	
e	b	e	a	b	e	b	e	h	b	
e	b	f	e	b	h	b	f	h	b	
c	d	c	g	d	g	d	c	g	d	
a	b	e	e	b	e	b	e	e	b	$\cdots$

Rigo and Maes [RM02] considered S-recognizability in a multidimensional setting. This concept was further studied by Charlier, Kärki, and Rigo [CKR10] who proved the following result, extending Theorem 1.4.4 to the multidimensional sequences.

Theorem 5.1.10. *Let $d \geq 1$ be an integer. The d-dimensional infinite word x is S-automatic for some abstract numeration system $S = (L, \Sigma, <)$ where $\varepsilon \in L$ if and only if x is the image under a coding of a shape symmetric infinite d-dimensional word.*

The multidimensional analogue of Theorem 1.3.16 is due to Semenov [Sem77] and requires an analogous notion of ultimate periodicity in the multidimensional setting.

Definition 5.1.11. A set $X \subseteq \mathbb{N}^d$ is *linear* if there exists $\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_t \in \mathbb{N}^d$ such that

$$X = \{\mathbf{v}_0 + n_1\mathbf{v}_1 + n_2\mathbf{v}_2 + \cdots + n_t\mathbf{v}_t : n_1, \ldots, n_t \in \mathbb{N}\}.$$

A set $X \subseteq \mathbb{N}^d$ is *semi-linear* if it is a finite union of linear sets.

For more on semi-linear sets, see [GS66]. We can now state the multidimensional version of Cobham's first theorem.

Theorem 5.1.12 ([Sem77])**.** *Let $k, \ell \geq 2$ be two multiplicatively independent integers and let $X \subseteq \mathbb{N}^d$. The set X is both k-recognizable and ℓ-recognizable if and only if it is semi-linear.*

In other words, the semi-linear sets are precisely the sets recognizable in all integer bases $k \geq 2$. One might therefore expect that, as in Theorem 1.4.4, the semi-linear sets are recognizable in all abstract numeration systems. However, this fails to be the case, as the following example shows.

Example 5.1.13. The semi-linear set $X = \{n(1, 2) : n \in \mathbb{N}\} = \{(n, 2n) : n \in \mathbb{N}\}$ is not 1-recognizable. Consider the language $\{(a^n \# ^n, a^{2n}) : n \in \mathbb{N}\}$, consisting of the unary representations of the elements of X. An easy application of the pumping lemma shows that this is not a regular language.

Observe that in the one-dimensional case, we have the following equivalences:

$$\text{semi-linear} \Leftrightarrow \text{ultimately periodic} \Leftrightarrow \text{1-recognizable.}$$

However, Example 5.1.13 shows that these equivalences no longer hold in a multidimensional setting. In order to get a multidimensional analogue of Theorem 1.4.4, we must consider the class of 1-recognizable sets, which forms a proper subclass of the class of semi-linear sets.

Another well-studied subclass of the class of semi-linear sets is the class of recognizable sets. A subset X of $\mathbb{N}^d$ is *recognizable* if there exists a finite monoid M, a monoid homomorphism $\varphi : \mathbb{N}^d \to M$, and a subset $B \subseteq M$ such that $X = \varphi^{-1}(B)$. When $d = 1$, we have again the following equivalences: recognizable $\Leftrightarrow$ ultimately periodic $\Leftrightarrow$ 1-recognizable. However, for $d > 1$ these equivalences no longer hold. An unpublished result of Mezei (see [Eil74, Proposition III.12.2]) demonstrates that the recognizable subsets of $\mathbb{N}^2$ are precisely finite unions of sets of the form $Y \times Z$, where Y and Z are ultimately periodic subsets of $\mathbb{N}$. In particular, the *diagonal set* $D = \{(n, n) : n \in \mathbb{N}\}$ is not recognizable [Eil74, Exercise III.12.7]. However, the set D is clearly a 1-recognizable subset of $\mathbb{N}^2$. So we see that for $d > 1$, the class of 1-recognizable sets corresponds neither to the class of semi-linear sets, nor to the class of recognizable sets. For further information on recognizable sets, their different characterizations and the classical Cobham–Semenov Theorem, see [BHMV94].

Our main result is the following, which generalizes the result of Lecomte and Rigo (Theorem 1.4.4).

Theorem 5.1.14. *Let $X \subseteq \mathbb{N}^d$. Then X is S-recognizable for all abstract numeration systems S if and only if X is 1-recognizable.*

To illustrate this theorem, we give the following example.

Example 5.1.15. Let

$$X = \{(2n, 3m + 1) : n, m \in \mathbb{N} \text{ and } 2n \geq 3m + 1\} \quad \cup$$
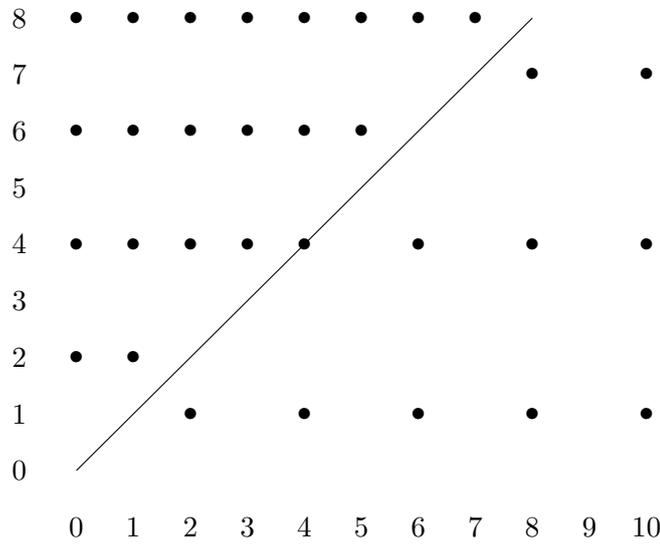$$\{(n, 2m) : n, m \in \mathbb{N} \text{ and } n < 2m\}.$$

Figure 5.2: The set X of Example 5.1.15

It is clear that X is 1-recognizable. Let $S = (L, \Sigma, <)$ be an abstract numeration system. By Theorem 1.4.4, the sets $\{2n : n \in \mathbb{N}\}$ and $\{3m + 1 : m \in \mathbb{N}\}$ are both S-recognizable, and so the set $\{(2n, 3m + 1) : n, m \in \mathbb{N}\}$ is also S-recognizable. In other words, the set $\{(\mathrm{rep}_S(2n), \mathrm{rep}_S(3m + 1))^\# : n, m \in \mathbb{N}\}$ is accepted by a finite automaton. Furthermore, the set $\{(x, y)^\# : x, y \in L \text{ and } x \geq y\}$ is also accepted by a finite automaton, and so by taking the product of these two automata we obtain an automaton accepting

$$\{(\mathrm{rep}_S(2n), \mathrm{rep}_S(3m + 1))^\# : n, m \in \mathbb{N} \text{ and } 2n \geq 3m + 1\}.$$

In the same way we can construct an automaton to accept the set

$$\{(\mathrm{rep}_S(n), \mathrm{rep}_S(2m))^\# : n, m \in \mathbb{N} \text{ and } n < 2m\}.$$

Since the union of two regular languages is regular, we see that X is S-recognizable.

5.2   Proof of our main result

In order to obtain our main result, we will need a classical result of Eilenberg, Elgot, and Shepherdson [EES69, Theorem 11.1] (see also [Rub04, Theorem C.1.1]). We first need the following definition.

Definition 5.2.1. Let A be a non-empty subset of $\{1, \ldots, d\}$. Define the subalphabet

$$\Sigma_A = \{x \in (\Sigma \cup \{\#\})^d : \text{the } i\text{-th component of } x \text{ is } \# \text{ exactly when } i \notin A\}.$$

Example 5.2.2. Let $\Sigma = \{a, b\}$ and $d = 4$. If $A = \{1, 2, 3, 4\}$, then $\Sigma_A = \{(\sigma_1, \sigma_2, \sigma_3, \sigma_4) : \sigma_i \in \Sigma \text{ for } i = 1, 2, 3, 4\}$. If $A = \{2, 3\}$, then $\Sigma_A = \{(\#, \sigma_2, \sigma_3, \#) : \sigma_i \in \Sigma \text{ for } i \in \{2, 3\}\}$. If $A = \{3\}$, then $\Sigma_A = \{(\#, \#, a, \#), (\#, \#, b, \#)\}$.

Theorem 5.2.3 (Decomposition [EES69]). *Let $R \subseteq (\Sigma^*)^d$. The language $R^\# \subseteq ((\Sigma \cup \{\#\})^d)^*$ is regular if and only if it is a finite union of languages of the form*

$$R_0 \cdots R_t, \quad t \in \mathbb{N},$$

where each factor $R_i \subseteq (\Sigma_{A_i})^*$ is regular and $A_t \subseteq \cdots \subseteq A_0 \subseteq \{1, \ldots, d\}$.

Remark 5.2.4. Theorem 5.2.3 does not hold if $R^\#$ is replaced by an arbitrary language over $(\Sigma \cup \{\#\})^d$. It is only valid due to the definition of the map $(\cdot)^\#$.

Example 5.2.5. Let $R = \{(a^{5n}, a^{6m}) : n, m \in \mathbb{N}\}$. Then $R^\#$ is regular, since one can easily construct an automaton that simultaneously checks that the length of the first component of its input is a multiple of 5 and that the length of the second component is a multiple of 6. Moreover, we have

$$R^\# = \bigcup_{\ell=0}^{5} (a^{30}, a^{30})^* (a^{5\ell} \#^\ell, a^{6\ell}) (\#^6, a^6)^* \quad \cup$$

$$\bigcup_{\ell=0}^{4} (a^{30}, a^{30})^* (a^{5(\ell+1)}, a^{6\ell} \#^{5-\ell}) (a^5, \#^5)^*.$$

Observe that each of the languages appearing in the unions above are products of the form described in Theorem 5.2.3.

Lemma 5.2.6. *Let $X \subseteq \mathbb{N}^d$. Then X is 1-recognizable if and only if X is a finite union of sets of the form*

$$
\left\{ \sum_{\ell=0}^{t} (c_\ell(n_{\ell,1}, \dots, n_{\ell,d}) + (b_{\ell,1}, \dots, b_{\ell,d})) \colon (\forall \ell)(\forall i)\ n_{\ell,i} \in \mathbb{N} \ and \right.
$$

$$
\left. (\forall \ell)(\forall i)\ (i \notin A_\ell \Rightarrow n_{\ell,i} = 0)\ and\ (\forall \ell)(\forall i)(\forall j)(i, j \in A_\ell \Rightarrow n_{\ell,i} = n_{\ell,j}) \right\}
$$

$$\tag{5.1}$$

where

- $t \in \mathbb{N}$,

- $A_t \subseteq \dots \subseteq A_0 \subseteq \{1, \dots, d\}$,

- $c_0, \dots, c_t \in \mathbb{N}$,

- $(\forall \ell)(\forall i)\ b_{\ell,i} \in \mathbb{N}$,

- $(\forall \ell)(\forall i)\ (i \notin A_\ell \Rightarrow b_{\ell,i} = 0)$, *and*

- $(\forall \ell)(\forall i)(\forall j)\ (i, j \in A_\ell \Rightarrow b_{\ell,i} = b_{\ell,j})$.

Proof. Let $\Sigma = \{a\}$ and let $S = (\Sigma^*, \Sigma, <)$. We define

$$
R := \mathrm{rep}_S(X) = \{(a^{n_1}, \dots, a^{n_d}) \colon (n_1, \dots, n_d) \in X\}.
$$

The set X is 1-recognizable if and only if the language $R^{\#}$ is regular. By Theorem 5.2.3, the language $R^{\#}$ is regular if and only if it is a finite union of languages of the form

$$
R_0 \cdots R_t, \quad t \in \mathbb{N},
$$

where each factor $R_\ell \subseteq (\Sigma_{A_\ell})^*$ is regular and $A_t \subseteq \dots \subseteq A_0 \subseteq \{1, \dots, d\}$. Since $|\Sigma| = 1$, we have $|\Sigma_{A_\ell}| = 1$. Let $\Sigma_{A_\ell} = \{x\}$. It is well known [Eil74, Proposition V.1.1] that R_ℓ is a finite union of languages of the form $\{x^{pi+q} : i \in \mathbb{N}\}$, where $p, q \in \mathbb{N}$. Without loss of generality we can assume that R_ℓ is exactly of this form. Hence, the language R_ℓ consists of the representations of a set of the form

$$
\{c_\ell(n_{\ell,1}, \dots, n_{\ell,d}) + (b_{\ell,1}, \dots, b_{\ell,d}) : (\forall i)(n_{\ell,i} \in \mathbb{N})\}.
$$

The conditions $A_t \subseteq \cdots \subseteq A_0 \subseteq \{1, \ldots, d\}$ impose the restrictions on the $n_{\ell,i}$'s and the constants $b_{\ell,i}$ in the statement of the lemma. The concatenation of the R_ℓ's gives the sum described above. $\qquad\qquad\square$

Remark 5.2.7. We can give an alternative description of the 1-recognizable sets. Let $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{N}^d$. We define $\mathrm{Supp}(\mathbf{v}) = \{i \in \{1, \ldots, d\} : v_i \neq 0\}$. Let $X \subseteq \mathbb{N}^d$. Then X is a finite union of sets of the form described in Lemma 5.2.6 if and only if X is a finite union of sets of the form

$$(b_0 + c_0\mathbb{N})\mathbf{v_0} + \cdots + (b_t + c_t\mathbb{N})\mathbf{v_t},$$

where

- $t \in \mathbb{N}$,

- $b_i, c_i \in \mathbb{N}$ for $i = 1, \ldots, t$,

- $\mathbf{v}_i \in \{0, 1\}^d$ for $i = 1, \ldots, t$,

- $\mathrm{Supp}(\mathbf{v}_t) \subseteq \cdots \subseteq \mathrm{Supp}(\mathbf{v}_0)$.

Example 5.2.8. Let $X = \{(5n, 5n + 4m + 6\ell + 1, 5n + 4m + 6\ell + 3, 5n) : n, m, \ell \in \mathbb{N}\}$. The unary representation of X is

$$R^\# = ((a, a, a, a)^5)^*((\#, a, a, \#)^4)^*((\#, a, a, \#)^6)^*(\#, a, a, \#)(\#, \#, a, \#)^2.$$

Since $R^\#$ is regular the set X is 1-recognizable. The set X can be written as

$$\begin{aligned}
X = \{5(n, n, n, n) + 4(0, m, m, 0) + 6(0, \ell, \ell, 0) \\
+ (0, 1, 1, 0) + (0, 0, 2, 0) \ : \ n, m, \ell \in \mathbb{N}\}, \quad (5.2)
\end{aligned}$$

which is an expression of the form (5.1) where $t = 3$; $A_0 = \{1, 2, 3, 4\}$, $A_1 = A_2 = \{2, 3\}$, $A_3 = \{3\}$; $c_0 = 5, c_1 = 4, c_2 = 6, c_3 = 0$; and $b_{0,i} = b_{1,i} = 0$ for all i, $(b_{2,1}, b_{2,2}, b_{2,3}, b_{2,4}) = (0, 1, 1, 0)$, $(b_{3,1}, b_{3,2}, b_{3,3}, b_{3,4}) = (0, 0, 2, 0)$.

Alternatively, by Remark 5.2.7 we can write

$$X = 5\mathbb{N}(1, 1, 1, 1) + 4\mathbb{N}(0, 1, 1, 0) + (1 + 6\mathbb{N})(0, 1, 1, 0) + (2 + 0\mathbb{N})(0, 0, 1, 0).$$

Furthermore, we have a factorization of $R^\#$ as given in Theorem 5.2.3: that is, $R^\# = R_0 R_1 R_2 R_3$, where $R_0 = ((a, a, a, a)^5)^*$, $R_1 = ((\#, a, a, \#)^4)^*$, $R_2 = ((\#, a, a, \#)^6)^*(\#, a, a, \#)$, and $R_3 = (\#, \#, a, \#)^2$, with the same A_ℓ's as those defined above. The term $5(n, n, n, n)$ corresponds to R_0, the term $4(0, m, m, 0)$ corresponds to R_1, the term $6(0, \ell, \ell, 0) + (0, 1, 1, 0)$ corresponds to R_2, and the term $(0, 0, 2, 0)$ corresponds to R_3.

We need the following classical number-theoretical result (see [RA05, Theorem 1.0.1]).

Theorem 5.2.9. *Let $a_1, \ldots, a_n$ be integers with $a_i \geq 2$ for $i = 1, \ldots, n$. If*

$$\gcd(a_1, \ldots, a_n) = 1,$$

then there exists a positive integer $F(a_1, \ldots, a_n)$ such that $F(a_1, \ldots, a_n)$ cannot be expressed as a non-negative linear combination of $a_1, \ldots, a_n$, but all integers greater than $F(a_1, \ldots, a_n)$ can be so expressed.

In the sequel we write $\mathbf{e}_i$ to denote the element of $\mathbb{N}^d$ that contains a 1 in its i-th component and 0's in all others.

Lemma 5.2.10. *A set $X \subseteq \mathbb{N}^d$ of the form (5.1) can be written as a union $A \cup B$, where A is made up of finite unions and intersections of sets having one of the forms (5.3)–(5.6) below and B is a finite intersection of sets of the form (5.3) or (5.4) below:*

$$\left\{ \sum_{\substack{i=1 \\ i \neq j}}^{d} n_i \mathbf{e}_i + (rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \geq N \right\} \tag{5.3}$$

where $1 \leq j \leq d$, and $r, s, N \in \mathbb{N}$;

$$\left\{ \sum_{\substack{i=1 \\ i \neq j}}^{d} n_i \mathbf{e}_i + (n_k + rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \geq N \right\} \tag{5.4}$$

where $1 \leq j, k \leq d$, $j \neq k$, and $r, s, N \in \mathbb{N}$;

$$\left\{ \sum_{\substack{i=1 \\ i \neq j}}^{d} n_i \mathbf{e}_i + (rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \in C \right\} \tag{5.5}$$

where $1 \leq j \leq d$, $r, s \in \mathbb{N}$, and $C \subseteq \mathbb{N}$ is a finite set; or

$$\left\{ \sum_{\substack{i=1 \\ i \neq j}}^{d} n_i \mathbf{e}_i + (n_k + rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \in C \right\} \tag{5.6}$$

where $1 \leq j, k \leq d$, $j \neq k$, and $r, s \in \mathbb{N}$, and $C \subseteq \mathbb{N}$ is a finite set.

Proof. Let X be a set of the form (5.1) where t, the A_ℓ's, the c_ℓ's, and the $b_{\ell,i}$'s are fixed and satisfy the conditions listed in Lemma 5.2.6. We will write $X = A \cup B$, where

$$B = \bigcap_{j=1}^{d} Y_j,$$

where each Y_j is either of the form (5.3) or (5.4), and A is made up of finite unions and intersections of sets of the forms (5.3)–(5.6).

First observe that if $j \in \{1, \ldots, d\} \setminus A_0$ the set X contains only vectors whose j-th component is always 0. For each such j, we define

$$Y_j = \left\{ \sum_{\substack{i=1 \\ i \neq j}}^{d} n_i \mathbf{e}_i + 0 \mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N} \right\},$$

which is of the form (5.3).

First consider the case where $A_0 = \cdots = A_t$. Define $j_1 < \cdots < j_{|A_0|}$ to be the elements of A_0. Define

$$Y_{j_1} = \left\{ \sum_{\substack{i=1 \\ i \neq j_1}}^{d} n_i \mathbf{e}_i + (r n_{j_1} + s) \mathbf{e}_{j_1} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_1} \geq N \right\},$$

where $r = \gcd(c_0, \ldots, c_t)$, $s = \sum_{\ell=0}^{t} b_{\ell, j_1}$, and $N - 1$ is the largest integer n such that rn cannot be written as a non-negative integer linear combination of $c_0, \ldots, c_t$ (note that by Theorem 5.2.9, N exists and is finite). Note that Y_{j_1} is of the form (5.3).

Define

$$Y'_{j_1} = \left\{ \sum_{\substack{i=1 \\ i \neq j_1}}^{d} n_i \mathbf{e}_i + (r n_{j_1} + s) \mathbf{e}_{j_1} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_1} \in C \right\},$$

where C is the set of all non-negative integers $n < N$ such that rn can be written as a non-negative integer linear combination of $c_0, \ldots, c_t$. Note that Y'_{j_1} is of the form (5.5).

For $k \in \{2, \ldots, |A_0|\}$, define

$$Y_{j_k} = \left\{ \sum_{\substack{i=1 \\ i \neq j_k}}^{d} n_i \mathbf{e}_i + n_{j_{k-1}} \mathbf{e}_{j_k} : n_1, \ldots, n_d \in \mathbb{N} \right\},$$

which is of the form (5.4).

The set X can be written as the union $A \cup B$ where

$$B = \bigcap_{j \in \{1,\ldots,d\} \setminus A_0} Y_j \quad \cap \quad \bigcap_{k \in \{1,\ldots,|A_0|\}} Y_{j_k}$$

and

$$A = \bigcap_{j \in \{1\ldots,d\} \setminus A_0} Y_j \quad \cap \quad \bigcap_{k \in \{2\ldots,|A_0|\}} Y_{j_k} \quad \cap \quad Y'_{j_1}.$$

Now consider the case where there is at least one index ℓ such that $A_\ell \setminus A_{\ell+1} \neq \emptyset$. Define $\ell_1 < \cdots < \ell_{t'}$ to be the indices of the sets A_ℓ satisfying $A_{\ell_k} \setminus A_{\ell_k+1} \neq \emptyset$ for each $k \in \{1, \ldots, t'\}$. We clearly have $1 \leq t' \leq t$ and $0 \leq \ell_{t'} < t$.

Define $d_1 = |A_{\ell_1} \setminus A_{\ell_1+1}|$ and $j_{1,1} < \cdots < j_{1,d_1}$ to be the elements of $A_{\ell_1} \setminus A_{\ell_1+1}$. Define

$$Y_{j_{1,1}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{1,1}}}^{d} n_i \mathbf{e}_i + (r_1 n_{j_{1,1}} + s_1) \mathbf{e}_{j_{1,1}} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_{1,1}} \geq N_1 \right\},$$

where $r_1 = \gcd(c_0, \ldots, c_{\ell_1})$, $s_1 = \sum_{\ell=0}^{\ell_1} b_{\ell,j_{1,1}}$, and $N_1 - 1$ is the largest integer n such that $r_1 n$ cannot be written as a non-negative integer linear combination of $c_0, \ldots, c_{\ell_1}$. Note that $Y_{j_{1,1}}$ is of the form (5.3).

Define

$$Y'_{j_{1,1}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{1,1}}}^{d} n_i \mathbf{e}_i + (r_1 n_{j_{1,1}} + s_1) \mathbf{e}_{j_{1,1}} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_{1,1}} \in C_1 \right\},$$

where C_1 is the set of all non-negative integers $n < N_1$ such that $r_1 n$ can be written as a non-negative integer linear combination of $c_0, \ldots, c_{\ell_1}$. Note that $Y'_{j_{1,1}}$ is of the form (5.5).

For $k \in \{2, \ldots, d_1\}$, define

$$Y_{j_{1,k}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{1,k}}}^{d} n_i \mathbf{e}_i + n_{j_{1,k-1}} \mathbf{e}_{j_{1,k}} : n_1, \ldots, n_d \in \mathbb{N} \right\},$$

which is of the form (5.4).

Define $d_2 = |A_{\ell_2} \setminus A_{\ell_2+1}|$ and $j_{2,1} < \cdots < j_{2,d_2}$ to be the elements of $A_{\ell_2} \setminus A_{\ell_2+1}$. Define

$$Y_{j_{2,1}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{2,1}}}^{d} n_i \mathbf{e}_i + (n_{j_{1,1}} + r_2 n_{j_{2,1}} + s_2) \mathbf{e}_{j_{2,1}} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_{2,1}} \geq N_2 \right\},$$

where $r_2 = \gcd(c_{\ell_1+1}, \ldots, c_{\ell_2})$, $s_2 = \sum_{\ell=\ell_1+1}^{\ell_2} b_{\ell,j_{2,1}}$, and $N_2 - 1$ is the largest integer n such that $r_2 n$ cannot be written as a non-negative integer linear combination of $c_{\ell_1+1}, \ldots, c_{\ell_2}$. Note that $Y_{j_{2,1}}$ is of the form (5.4).

Define

$$Y'_{j_{2,1}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{2,1}}}^{d} n_i \mathbf{e}_i + (n_{j_{1,1}} + r_2 n_{j_{2,1}} + s_2) \mathbf{e}_{j_{2,1}} : n_1, \ldots, n_d \in \mathbb{N}, n_{j_{2,1}} \in C_2 \right\},$$

where C_2 is the set of all non-negative integers $n < N_2$ such that $r_2 n$ can be written as a non-negative integer linear combination of $c_{\ell_1+1}, \ldots, c_{\ell_2}$. Note that $Y'_{j_{2,1}}$ is of the form (5.6).

For $k \in \{2, \ldots, d_2\}$, define

$$Y_{j_{2,k}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{2,k}}}^{d} n_i \mathbf{e}_i + n_{j_{2,k-1}} \mathbf{e}_{j_{2,k}} : n_1, \ldots, n_d \in \mathbb{N} \right\},$$

which is of the form (5.4).

We continue in this way to define d_p, $Y_{j_{p,k}}$, and $Y'_{j_{p,1}}$ for all $p \in \{1, \ldots, t'\}$ and $k \in \{1, \ldots, d_p\}$. Finally observe that we have $A_{\ell_{t'}} \setminus A_{\ell_{t'}+1} \neq \emptyset$ and $A_{\ell_{t'}+1} = \cdots = A_t$. Define $d_{t'+1} = |A_t|$ and $j_{t'+1,1} < \cdots < j_{t'+1,d_{t'+1}}$ to be the elements of A_t. Define

$$Y_{j_{t'+1,1}} = \Bigg\{ \sum_{\substack{i=1 \\ i \neq j_{t'+1,1}}}^{d} n_i \mathbf{e}_i + (n_{j_{t',1}} + r_{t'+1} n_{j_{t'+1,1}} + s_{t'+1}) \mathbf{e}_{j_{t'+1,1}} :$$

$$n_1, \ldots, n_d \in \mathbb{N}, n_{j_{t'+1,1}} \geq N_{t'+1} \Bigg\},$$

where $r_{t'+1} = \gcd(c_{\ell_{t'}+1}, \ldots, c_t)$, $s_{t'+1} = \sum_{\ell=\ell_{t'}+1}^{\ell_t} b_{\ell,j_{t'+1,1}}$, and $N_{t'+1} - 1$ is the largest integer n such that $r_{t'+1} n$ cannot be written as a non-negative

integer linear combination of $c_{\ell_{t'}+1}, \ldots, c_t$. Again note that $Y_{j_{t'+1,1}}$ is of the form (5.4).

Define

$$Y'_{j_{t'+1,1}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{t'+1,1}}}^{d} n_i \mathbf{e}_i + (n_{j_{t',1}} + r_{t'+1} n_{j_{t'+1,1}} + s_{t'+1}) \mathbf{e}_{j_{t'+1,1}} : \right.$$

$$\left. n_1, \ldots, n_d \in \mathbb{N}, n_{j_{t'+1,1}} \in C_{t'+1} \right\},$$

where $C_{t'+1}$ is the set of all non-negative integers $n < N_{t'+1}$ such that $r_{t'+1} n$ can be written as a non-negative integer linear combination of $c_{\ell_{t'}+1}, \ldots, c_t$. Note that $Y'_{j_{t'+1,1}}$ is of the form (5.6).

For $k \in \{2, \ldots, d_{t'+1}\}$, define

$$Y_{j_{t'+1,k}} = \left\{ \sum_{\substack{i=1 \\ i \neq j_{t'+1,k}}}^{d} n_i \mathbf{e}_i + n_{j_{t'+1,k-1}} \mathbf{e}_{j_{t'+1,k}} : n_1, \ldots, n_d \in \mathbb{N} \right\},$$

which is of the form (5.4).

The set X can be written as the union $A \cup B$ where

$$B = \bigcap_{j \in \{1,\ldots,d\} \setminus A_0} Y_j \quad \cap \quad \bigcap_{\substack{p \in \{1,\ldots,t'+1\} \\ k \in \{1,\ldots,d_p\}}} Y_{j_{p,k}}$$

and

$$A = \bigcap_{j \in \{1,\ldots,d\} \setminus A_0} Y_j \quad \cap$$

$$\bigcup_{p \in \{1,\ldots,t'+1\}} \left(Y'_{j_{p,1}} \quad \cap \quad \bigcap_{\substack{q \in \{1,\ldots,t'+1\} \setminus \{p\} \\ k \in \{1,\ldots,d_q\}}} Y_{j_{q,k}} \quad \cap \quad \bigcap_{k \in \{2,\ldots,d_p\}} Y_{j_{p,k}} \right).$$

$\square$

Example 5.2.11. We continue Example 5.2.8. We will write $X = A \cup B$ as in Lemma 5.2.10. The A_ℓ's are not all the same, so we can define $t' = 2$, $\ell_1 = 0 < \ell_2 = 2$ as in the proof of Lemma 5.2.10.

We have $d_1 = |A_0 \setminus A_1| = 2$, $j_{1,1} = 1$ and $j_{1,2} = 4$. We also have $r_1 = \gcd(c_0) = \gcd(5) = 5$ and $s_1 = 0$, and hence $N_1 = 0$. Therefore,

$$Y_1 = \{n_2\mathbf{e_2} + n_3\mathbf{e_3} + n_4\mathbf{e_4} + (5n_1 + 0)\mathbf{e_1} : n_1, n_2, n_3, n_4 \in \mathbb{N}, n_1 \geq 0\},$$

$$Y_1' = \{n_2\mathbf{e_2} + n_3\mathbf{e_3} + n_4\mathbf{e_4} + (5n_1 + 0)\mathbf{e_1} : n_2, n_3, n_4 \in \mathbb{N}, n_1 \in C_1\} = \emptyset,$$

since $C_1 = \emptyset$, and

$$Y_4 = \{n_1\mathbf{e_1} + n_2\mathbf{e_2} + n_3\mathbf{e_3} + n_1\mathbf{e_4} : n_1, n_2, n_3 \in \mathbb{N}\}.$$

Next we have $d_2 = |A_2 \setminus A_3| = 1$ and $j_{2,1} = 2$. We also have $r_2 = \gcd(c_1, c_2) = \gcd(4, 6) = 2$ and $s_2 = b_{1,2} + b_{2,2} = 0 + 1 = 1$, and hence $N_2 = 2$. Therefore,

$$Y_2 = \{n_1\mathbf{e_1} + n_3\mathbf{e_3} + n_4\mathbf{e_4} + (n_1 + 2n_2 + 1)\mathbf{e_2} : n_1, n_2, n_3 \in \mathbb{N}, n_2 \geq 2\},$$

and

$$\begin{aligned} Y_2' &= \{n_1\mathbf{e_1} + n_3\mathbf{e_3} + n_4\mathbf{e_4} + (n_1 + 2n_2 + 1)\mathbf{e_2} : n_1, n_3 \in \mathbb{N}, n_2 \in C_2\} \\ &= \{n_1\mathbf{e_1} + n_3\mathbf{e_3} + n_4\mathbf{e_4} + (n_1 + 1)\mathbf{e_2} : n_1, n_3 \in \mathbb{N}\}, \end{aligned}$$

since $C_2 = \{0\}$.

Finally, we have $d_3 = |A_3| = 1$ and $j_{3,1} = 3$. We also have $r_3 = \gcd(c_3) = \gcd(0) = 0$ and $s_3 = b_{3,3} = 2$, and hence $N_3 = 0$. Therefore,

$$Y_3 = \{n_1\mathbf{e_1} + n_2\mathbf{e_2} + n_4\mathbf{e_4} + (n_2 + 0n_3 + 2)\mathbf{e_3} : n_1, n_2, n_3 \in \mathbb{N}, n_3 \geq 0\},$$

and

$$Y_3' = \{n_1\mathbf{e_1} + n_2\mathbf{e_2} + n_4\mathbf{e_4} + (n_2 + 0n_3 + 2)\mathbf{e_3} : n_1, n_2 \in \mathbb{N}, n_3 \in C_3\} = \emptyset,$$

since $C_3 = \emptyset$.

Hence $A = Y_1 \cap Y_2' \cap Y_3 \cap Y_4$ and $B = Y_1 \cap Y_2 \cap Y_3 \cap Y_4$.

We now need to introduce the notion of synchronous transducers and synchronous relations.

Definition 5.2.12. Let Σ and Δ be two finite alphabets. A *synchronous transducer* (from $(\Sigma \cup \{\#\})^*$ to $(\Delta \cup \{\#\})^*$) is a letter-to-letter transducer respecting the padding condition, that is no letter can appear after the padding symbol on the same component. Synchronous transducers realise synchronous relations obtained by the projection which erases the padding symbol $\#$.

The two following results will be useful for the proof of Lemma 5.2.16.

Definition 5.2.13. Let i be a non-negative integer. The $(i+1)$-th word of a language L in the genealogical order is denoted by $\langle i \rangle_L$. The successor of a word x in L is the unique word y of L such that

$$(x <_{gen} y) \wedge (\forall z \in L)((x <_{gen} z) \Rightarrow ((y = z) \vee (y <_{gen} z))).$$

The successor function[1] on L is the function $\mathrm{Succ}_L : \Sigma^* \to \Sigma^*$ that maps a word $x = \langle i \rangle_L$ of L onto its successor $y = \langle i+1 \rangle_L$ in L.

Proposition 5.2.14 ([BFRS07, BR10]). *Let L be a regular language. The successor function* Succ_L *is a synchronous relation.*

Theorem 5.2.15 ([FS93, BR10]). *The family of synchronous relation is closed under composition.*

We are ready to prove the following lemma.

Lemma 5.2.16. *Let $k \in \mathbb{N}$ and let S be an abstract numeration system. The set $X = \{(n, n+k) : n \in \mathbb{N}\}$ is S-recognizable.*

Proof. The proof follows easily from known results and so we only give a sketch of the proof. Let $R = \mathrm{rep}_S(X)$. To show that X is S-recognizable we must show that $R^\#$ is a regular language. Consider first the set $Y = \{(\mathrm{rep}_S(n), \mathrm{rep}_S(n+1)) : n \in \mathbb{N}\}$. If we interpret Y as the function mapping $\mathrm{rep}_S(n)$ to $\mathrm{rep}_S(n+1)$, then Y is the so-called successor function. From Proposition 5.2.14, we have that Y is a synchronous relation. Synchronous relations are defined in terms of letter-to-letter transducers, but this definition is equivalent to the fact that the language $Y^\#$ is accepted by a finite automaton. Moreover, from Theorem 5.2.15, we have that the composition of synchronous relations is again a synchronous relation. Hence R, which is the k-fold composition of Y with itself, is a synchronous relation. We conclude that $R^\#$ is a regular language, as required. □

Lemma 5.2.17. *A set $X \subseteq \mathbb{N}^d$ having one of the forms (5.3)–(5.6) defined in Lemma 5.2.10 is S-recognizable for any abstract numeration system S.*

Proof. We will give the proof for the cases where X is either of the form (5.3) or (5.4) (the other two cases are similar).

---

[1] See [AS10] or [LR01] for more on the successor function.

Let $S = (L, \Sigma, <)$ be an abstract numeration system and let $\mathcal{A}$ be a finite automaton accepting L. Let $R = \operatorname{rep}_S(X)$. We will show that $R^\#$ is regular. That is, we will define a (non-deterministic) finite automaton $\mathcal{M}$ that accepts $R^\#$. Let $(w_1, \ldots, w_d)^\#$ be an arbitrary input to the automaton $\mathcal{M}$.

Suppose that X is of the form (5.3). That is,

$$X = \left\{ \sum_{\substack{i=1 \\ i \neq j}}^d n_i \mathbf{e}_i + (rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \geq N \right\},$$

where $1 \leq j \leq d$, and $r, s, N \in \mathbb{N}$. Suppose first that $r = 0$. In this case, the automaton $\mathcal{M}$ simulates $\mathcal{A}$ on $w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d$. The automaton $\mathcal{M}$ accepts its input if and only if $\mathcal{A}$ accepts $w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d$ and $w_j = \operatorname{rep}_S(s)$.

Now suppose that $r > 0$. By increasing the value of N, we may, without loss of generality, assume that $s < r$. By Theorem 1.4.4 the language $\{\operatorname{rep}_S(rn_j + s) : n_j \geq N\}$ is regular. Let $\mathcal{A}'$ be an automaton accepting $\{\operatorname{rep}_S(rn_j + s) : n_j \geq N\}$. As before, the automaton $\mathcal{M}$ simulates $\mathcal{A}$ on $w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d$, but now also simulates $\mathcal{A}'$ on w_j. The automaton $\mathcal{M}$ accepts its input if and only if $\mathcal{A}$ accepts $w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d$ and $\mathcal{A}'$ accepts w_j.

Next suppose that X is of the form (5.4). That is,

$$\left\{ \sum_{\substack{i=1 \\ i \neq j}}^d n_i \mathbf{e}_i + (n_k + rn_j + s)\mathbf{e}_j : n_1, \ldots, n_d \in \mathbb{N}, n_j \geq N \right\},$$

where $1 \leq j, k \leq d$, $j \neq k$, and $r, s, N \in \mathbb{N}$. Again, suppose first that $r = 0$. By Lemma 5.2.16, the language $\{(\operatorname{rep}_S(n_k), \operatorname{rep}_S(n_k + s))^\# : n_k \in \mathbb{N}\}$ is regular. Let $\mathcal{A}''$ be a finite automaton accepting this language. The automaton $\mathcal{M}$ simulates $\mathcal{A}$ on each of the words in $\{w_1, \ldots, w_d\} \setminus \{w_j, w_k\}$. Simultaneously, the automaton $\mathcal{M}$ simulates $\mathcal{A}''$ on the pair $(w_k, w_j)^\#$. The automaton $\mathcal{M}$ accepts its input if and only if $\mathcal{A}$ accepts $\{w_1, \ldots, w_d\} \setminus \{w_j, w_k\}$ and $\mathcal{A}''$ accepts $(w_k, w_j)^\#$.

Now suppose that $r > 0$. Again, without loss of generality, we may assume that $s < r$. Using the same ideas as in the proof of [BR10, Theorem 3.3.1], it is not hard to see that the language

$$\{(\operatorname{rep}_S(m), \operatorname{rep}_S(n))^\# : m, n \in \mathbb{N} \text{ and } (n - m) \equiv s \pmod{r}\}$$

is regular. Let $\mathcal{Z}$ be an automaton accepting this language. Let $\mathcal{Z}'$ be an automaton accepting the language $\{(\operatorname{rep}_S(n_k), \operatorname{rep}_S(n_k + rN + s))^\# : n_k \in \mathbb{N}\}$ (since $rN + s$ is a constant, we may apply Lemma 5.2.16).

The automaton $\mathcal{M}$ simulates $\mathcal{A}$ on each of the words in $\{w_1, \ldots, w_d\} \setminus \{w_j, w_k\}$. Simultaneously, $\mathcal{M}$ simulates $\mathcal{Z}$ on the pair $(w_k, w_j)^{\#}$. The automaton $\mathcal{M}$ also non-deterministically "guesses" a word $v = b_1 \cdots b_{|v|}$ and simulates $\mathcal{Z}'$ on the pair $(w_k, v)^{\#}$. This "guess" works as follows. Let $w_k = a_1 \cdots a_{|w_k|}$, where each $a_i \in \Sigma$. For each $i = 1, \ldots, |w_k|$, we simulate $\mathcal{Z}'$ by non-deterministically choosing to follow one of the transitions of $\mathcal{Z}'$ labelled (a_i, b_i), where $b_i \in \Sigma$; and for $i > |w_k|$ (i.e., w_k has been completely read), the simulation may make a non-deterministic choice among transitions of the form $(\#, b_i)$, where $b_i \in \Sigma$. This non-deterministic choice of b_i at each step of the simulation is what defines the "guessed" word v. Note that if $\mathcal{Z}'$ accepts $(w_k, v)^{\#}$, then $\mathrm{val}_s(v) = \mathrm{val}_S(w_k) + rN + s$. As this non-deterministic simulation is performed, the automaton $\mathcal{M}$ also simultaneously verifies that w_j is greater than or equal to (in the radix order) the guessed word v.

The automaton $\mathcal{M}$ accepts its input if and only if

- $\mathcal{A}$ accepts each of the words in $\{w_1, \ldots, w_d\} \setminus \{w_j, w_k\}$,

- $\mathcal{Z}$ accepts $(w_k, w_j)^{\#}$ (and hence $\mathrm{val}_S(w_j) - \mathrm{val}_S(w_k) \equiv s \pmod{r}$),

- $\mathcal{Z}'$ accepts $(w_k, v)^{\#}$ for some guessed word v as described above (and hence $\mathrm{val}_S(v) = \mathrm{val}_S(w_k) + rN + s$), and

- w_j is greater than or equal to v in the radix order (and hence $\mathrm{val}_S(w_j) \geq \mathrm{val}_S(v)$).

The last three of these conditions guarantee that $\mathrm{val}_S(w_j) = \mathrm{val}_S(w_k) + rn_j + s$ for some $n_j \geq N$.

This completes the proof for the cases where X is either of the form (5.3) or (5.4). As previously stated, we omit the details for the other two cases since they are similar. $\qquad\square$

We are ready for the proof of Theorem 5.1.14.

Proof of Theorem 5.1.14. One direction is clear: if X is S-recognizable for all abstract numeration systems S, then it is certainly 1-recognizable.

To prove the other direction, suppose that X is 1-recognizable. The result now follows from Lemmas 5.2.6, 5.2.10, and 5.2.17. $\qquad\square$

Remark 5.2.18. Thanks to Theorem 5.1.14, the set $X = \{(n, m) \; : \; n \leq m, n, m \in \mathbb{N}\}$ is S-recognizable for all abstract numeration systems S. In-

deed, we have

$$\begin{aligned}
\{(n,m) \ : \ n \le m, n, m \in \mathbb{N}\} &= \{(n, n+k) \ : \ n, k \in \mathbb{N}\} \\
&= \mathbb{N}(1,1) + \mathbb{N}(0,1)
\end{aligned}$$

and by Lemma 5.2.6, X is 1-recognizable.

However, the set $Y = \{(n,m) \ : \ 2n \le m, n, m \in \mathbb{N}\}$ is not always S-recognizable. Otherwise, the set $\{(n, 2n) \ : \ n \in \mathbb{N}\}$ would be S-recognizable but, as shown by Theorem 1.4.7, multiplication by a constant does not preserve S-recognizability.

Bibliography

[AB08]   B. Adamczewski and J. Bell. Function fields in positive char-
          acteristic: expansions and Cobham's theorem. *J. Algebra*,
          319(6):2337–2350, 2008.

[AB11]   B. Adamczewski and J. Bell. An analogue of Cobham's theorem
          for fractals. *Trans. Amer. Math. Soc.*, 363(8):4421–4442, 2011.

[Ale04]   B. Alexeev. Minimal DFA for testing divisibility. *J. Comput.
          System Sci.*, 69(2):235–243, 2004.

[All68]   D. Allen, Jr. On a characterization of the nonregular sets of
          primes. *J. Comput. System Sci.*, 2:464–467, 1968.

[ARS09]   J.-P. Allouche, N. Rampersad, and J. Shallit. Periodicity, repe-
          titions, and orbits of an automatic sequence. *Theoret. Comput.
          Sci.*, 410(30-32):2795–2803, 2009.

[AS03]   J.-P. Allouche and J. Shallit. *Automatic sequences*. Cambridge
          University Press, Cambridge, 2003. Theory, applications, gener-
          alizations.

[AS10]   P.-Y. Angrand and J. Sakarovitch. Radix enumeration of rational
          languages. *RAIRO Theor. Inform. Appl.*, 44(1):19–36, 2010.

[ASS05]   J.-P. Allouche, J. Shallit, and G. Skordev. Self-generating sets,
          integers with missing blocks, and substitutions. *Discrete Math.*,
          292(1-3):1–15, 2005.

[AST00]   J.-P. Allouche, K. Scheicher, and R. F. Tichy. Regular maps in
          generalized number systems. *Math. Slovaca*, 50(1):41–58, 2000.

[BB07]   B. Boigelot and J. Brusten. A generalization of Cobham's the-
          orem to automata over real numbers. In *Automata, languages
          and programming*, volume 4596 of *Lecture Notes in Comput. Sci.*,
          pages 813–824. Springer, Berlin, 2007.

[BB09]   B. Boigelot and J. Brusten. A generalization of Cobham's the-
             orem to automata over real numbers. *Theoret. Comput. Sci.*,
             410(18):1694–1703, 2009.

[BBLT06]   G. Barat, V. Berthé, P. Liardet, and J. Thuswaldner. Dynam-
             ical directions in numeration. *Ann. Inst. Fourier (Grenoble)*,
             56(7):1987–2092, 2006. Numération, pavages, substitutions.

[BCFR09]   J. Bell, É. Charlier, A. S. Fraenkel, and M. Rigo. A decision
             problem for ultimately periodic sets in nonstandard numeration
             systems. *Internat. J. Algebra Comput.*, 19(6):809–839, 2009.

[Bès97]   A. Bès. Undecidable extensions of Büchi arithmetic and Cobham-
             Semënov theorem. *J. Symbolic Logic*, 62(4):1280–1296, 1997.

[Bès00]   A. Bès. An extension of the Cobham-Semënov theorem. *J. Sym-
             bolic Logic*, 65(1):201–211, 2000.

[BFRS07]   V. Berthé, C. Frougny, M. Rigo, and J. Sakarovitch. On the cost
             and complexity of the successor function. *Proceeding of WORDS
             2007*, 2007. (CIRM, Luminy, Marseille).

[BH97]   V. Bruyère and G. Hansel. Bertrand numeration systems and
             recognizability. *Theoret. Comput. Sci.*, 181(1):17–43, 1997. Latin
             American Theoretical INformatics (Valparaíso, 1995).

[BHMV94]   V. Bruyère, G. Hansel, C. Michaux, and R. Villemaire. Logic
             and p-recognizable sets of integers. *Bull. Belg. Math. Soc. Simon
             Stevin*, 1(2):191–238, 1994.

[BR10]   V. Berthé and M. Rigo. In *Combinatorics, automata and num-
             ber theory*, volume 135 of *Encyclopedia Math. Appl.*, pages 1–33.
             Cambridge Univ. Press, Cambridge, 2010.

[Bre71]   Ju. Ja. Breĭtbart. The automaton and "tape" complexity of the
             predicate "to be the kth power of a number". *Dokl. Akad. Nauk
             SSSR*, 196:16–19, 1971.

[Brz63]   J. A. Brzozowski. Canonical regular expressions and minimal
             state graphs for definite events. In *Proc. Sympos. Math. The-
             ory of Automata (New York, 1962)*, pages 529–561. Polytechnic
             Press of Polytechnic Inst. of Brooklyn, Brooklyn, N.Y., 1963.

[Büc60]   J. R. Büchi. Weak second-order arithmetic and finite automata.
             Z. Math. Logik Grundlagen Math., 6:66–92, 1960.

[BY11]   J. Brzozowski and Y. Ye. Syntactic complexity of ideal and closed
           languages. In *Developments in language theory*, volume 6795 of
           Lecture Notes in Comput. Sci., pages 117–128. Springer, Heidel-
           berg, 2011.

[CKR10]   É. Charlier, T. Kärki, and M. Rigo. Multidimensional general-
           ized automatic sequences and shape-symmetric morphic words.
           Discrete Math., 310(6-7):1238–1252, 2010.

[CLR12]   É. Charlier, A. Lacroix, and N. Rampersad. Multi-dimensional
           sets recognizable in all abstract numeration systems. *RAIRO
           Theor. Inform. Appl.*, 46(1):51–65, 2012.

[Cob68]   A. Cobham. On the hartmanis-stearns problem for a class of
           tag machines. In *IEEE Conference Record of 1968 Ninth Annual
           Symposium on Switching and Automata Theory*, pages 51–60.
           1968.

[Cob69]   A. Cobham. On the base-dependence of sets of numbers recog-
           nizable by finite automata. *Math. Systems Theory*, 3:186–192,
           1969.

[Cob72]   A. Cobham. Uniform tag sequences. *Math. Systems Theory*,
           6:164–192, 1972.

[DS89]   C. Dwork and L. Stockmeyer. On the power of 2-way probabilis-
           tic finite state automata. *IEEE Press, New York*, pages 480–485,
           1989. Proc. 30th Ann. Symp. Found. Comput. Sci.

[Dur98]   F. Durand. A generalization of Cobham's theorem. *Theory Com-
           put. Syst.*, 31(2):169–185, 1998.

[Dur02a]   F. Durand. Combinatorial and dynamical study of substitutions
           around the theorem of Cobham. In *Dynamics and randomness
           (Santiago, 2000)*, volume 7 of *Nonlinear Phenom. Complex Sys-
           tems*, pages 53–94. Kluwer Acad. Publ., Dordrecht, 2002.

[Dur02b]   F. Durand. A theorem of Cobham for non-primitive substitu-
           tions. *Acta Arith.*, 104(3):225–241, 2002.

[Dur08]   F. Durand. Cobham-Semenov theorem and $\mathbb{N}^d$-subshifts. *Theo-
           ret. Comput. Sci.*, 391(1-2):20–38, 2008.

[Dur11]   F. Durand. Decidability of the HD0L ultimate periodicity prob-
           lem. *preprint*, 2011. to appear in RAIRO, "arXiv"1111.3268.

[Dur12]   F. Durand. HD0L ω-equivalence and periodicity problems in the primitive case. *Unif. Distrib. Theory*, 7(1):199–215, 2012.

[EES69]   S. Eilenberg, C. C. Elgot, and J. C. Shepherdson. Sets recognized by n-tape automata. *J. Algebra*, 13:447–464, 1969.

[Eil74]   S. Eilenberg. *Automata, languages, and machines. Vol. A*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York, 1974. Pure and Applied Mathematics, Vol. 58.

[Fab94]   S. Fabre. Une généralisation du théorème de Cobham. *Acta Arith.*, 67(3):197–208, 1994.

[Fra85]   A. S. Fraenkel. Systems of numeration. *Amer. Math. Monthly*, 92(2):105–114, 1985.

[Fro92]   C. Frougny. Representations of numbers and finite automata. *Math. Systems Theory*, 25(1):37–60, 1992.

[FS93]   C. Frougny and J. Sakarovitch. Synchronized rational relations of finite and infinite words. *Theoret. Comput. Sci.*, 108(1):45–82, 1993. International Colloquium on Words, Languages and Combinatorics (Kyoto, 1990).

[FS09]   P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[GG07]   D. Garth and A. Gouge. Affinely self-generating sets and morphisms. *J. Integer Seq.*, 10(1):Article 07.1.5, 13, 2007.

[GK66]   V. S. Grinberg and A. D. Korshunov. Asymptotic behaviourof the maximum of the weight of a finite tree. *Problems Inform. Transmission*, 2:75–78, 1966.

[GO81]   L. J. Guibas and A. M. Odlyzko. Periods in strings. *J. Combin. Theory Ser. A*, 30(1):19–42, 1981.

[Gra08]   A. Granville. Smooth numbers: computational number theory and beyond. In *Algorithmic number theory: lattices, number fields, curves and cryptography*, volume 44 of *Math. Sci. Res. Inst. Publ.*, pages 267–323. Cambridge Univ. Press, Cambridge, 2008.

[GS66]   S. Ginsburg and E. H. Spanier. Semigroups, Presburger formulas, and languages. *Pacific J. Math.*, 16:285–296, 1966.

[GS98]   I. Glaister and J. Shallit. Automaticity. III. Polynomial automaticity and context-free languages. *Comput. Complexity*, 7(4):371–387, 1998.

[Han98] G. Hansel. Systèmes de numération indépendants et syndéticité. *Theoret. Comput. Sci.*, 204(1-2):119–130, 1998.

[HL86]   T. Harju and M. Linna. On the periodicity of morphisms on free monoids. *RAIRO Inform. Théor. Appl.*, 20(1):47–54, 1986.

[Hon86] J. Honkala. A decision method for the recognizability of sets defined by number systems. *RAIRO Inform. Théor. Appl.*, 20(4):395–403, 1986.

[HR04]   J. Honkala and M. Rigo. Decidability questions related to abstract numeration systems. *Discrete Math.*, 285(1-3):329–333, 2004.

[HS68]   J. Hartmanis and H. Shank. On the recognition of primes by automata. *J. Assoc. Comput. Mach.*, 15:382–389, 1968.

[Hsu96] C. Hsu. The distribution of irreducible polynomials in $\mathbf{F}_q[t]$. *J. Number Theory*, 61(1):85–96, 1996.

[HW79]   G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers.* The Clarendon Press Oxford University Press, New York, fifth edition, 1979.

[Kar67] R. M. Karp. Some bounds on the storage requirements of sequential machines and Turing machines. *J. Assoc. Comput. Mach.*, 14:478–489, 1967.

[KF90]   J. Kaņeps and R. Freivalds. Minimal nontrivial space complexity of probabilistic one-way Turing machines. In *Mathematical foundations of computer science (Banská Bystrica, 1990)*, volume 452 of *Lecture Notes in Comput. Sci.*, pages 355–361. Springer, Berlin, 1990.

[Kim00] C. Kimberling. A self-generating set and the golden mean. *J. Integer Seq.*, 3(2):Article 00.2.8, 1 HTML document, 2000.

[Kim04]   C. Kimberling. Affinely recursive sets and orderings of languages. *Discrete Math.*, 274(1-3):147–159, 2004.

[Kle56]   S. C. Kleene. Representation of events in nerve nets and finite automata. In *Automata studies*, Annals of mathematics studies, no. 34, pages 3–41. Princeton University Press, Princeton, N. J., 1956.

[KLR09]   T. Kärki, A. Lacroix, and M. Rigo. On the recognizability of self-generating sets. In *Mathematical foundations of computer science 2009*, volume 5734 of *Lecture Notes in Comput. Sci.*, pages 525–536. Springer, Berlin, 2009.

[KLR10]   T. Kärki, A. Lacroix, and M. Rigo. On the recognizability of self-generating sets. *J. Integer Seq.*, 13(2):Article 10.2.2, 18, 2010.

[Lis11]   G. Lischke. Primitive words and roots of words. *Acta Univ. Sapientiae, Informatica*, 3:5–34, 2011.

[LN83]   R. Lidl and H. Niederreiter. *Finite fields*, volume 20 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1983. With a foreword by P. M. Cohn.

[Lot83]   M. Lothaire. *Combinatorics on words*. Cambridge University Press, Cambridge, 1983.

[Lot02]   M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2002.

[LR01]   P. B. A. Lecomte and M. Rigo. Numeration systems on a regular language. *Theory Comput. Syst.*, 34(1):27–44, 2001.

[LR13]   A. Lacroix and N. Rampersad. Automaticity of primitive words and irreducible polynomials. *DMTCS*, 15(1):29–36, 2013.

[LRRV12]   A. Lacroix, N. Rampersad, M. Rigo, and É. Vandomme. Syntactic complexity of ultimately periodic sets of integers and application to a decision procedure. *Fund. Inform.*, 116(1-4):175–187, 2012.

[Mae99]   A. Maes. *Morphic Predicates and Applications to the Decidability of Arithmetic Theories*. PhD Thesis, Univ. Mons Hainaut. 1999.

[Mit11]   I. Mitrofanov. A proof for the decidability of HD0L ultimate periodicity. *preprint*, 2011. "arXiv"1110.4780.

[MS13]   V. Marsault and J. Sakarovitch. Ultimate periodicity of b-recognisable sets: a quasilinear procedure. *preprint*, 2013. "arXiv"1301.2691.

[MV96]   C. Michaux and R. Villemaire. Presburger arithmetic and recognizability of sets of natural numbers by automata: new proofs of Cobham's and Semenov's theorems. *Ann. Pure Appl. Logic*, 77(3):251–277, 1996.

[Ner58]   A. Nerode. Linear automaton transformations. *Proc. Amer. Math. Soc.*, 9:541–544, 1958.

[Nie73]   P. T. Nielsen. A note on bifix-free sequences. *IEEE Trans. Information Theory*, IT-19:704–706, 1973.

[Pan86]   J.-J. Pansiot. Decidability of periodicity for infinite words. *RAIRO Inform. Théor. Appl.*, 20(1):43–46, 1986.

[PB97]   F. Point and V. Bruyère. On the Cobham-Semenov theorem. *Theory Comput. Syst.*, 30(2):197–220, 1997.

[Pin97]   J.-E. Pin. Syntactic semigroups. In *Handbook of formal languages, Vol. 1*, pages 679–746. Springer, Berlin, 1997.

[PRS63]   M. Perles, M. O. Rabin, and E. Shamir. The theory of definite automata. *IEEE Trans. Electronic Computers*, EC-12:233–243, 1963.

[PRS97]   C. Pomerance, J. M. Robson, and J. Shallit. Automaticity. II. Descriptional complexity in the unary case. *Theoret. Comput. Sci.*, 180(1-2):181–201, 1997.

[RA05]   J. L. Ramírez Alfonsín. *The Diophantine Frobenius problem*, volume 30 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2005.

[Reu93]   C. Reutenauer. Free Lie algebras. Oxford, Oxford, 1993.

[Rig00]   M. Rigo. Generalization of automatic sequences for numeration systems on a regular language. *Theoret. Comput. Sci.*, 244(1-2):271–281, 2000.

[Rig01]   M. Rigo. Numeration systems on a regular language: arithmetic operations, recognizability and formal power series. *Theoret. Comput. Sci.*, 269(1-2):469–498, 2001.

[Rig08]   M. Rigo. Syntactical and automatic properties of sets of polynomials over finite fields. *Finite Fields Appl.*, 14(1):258–276, 2008.

[RM02]   M. Rigo and A. Maes. More on generalized automatic sequences. *J. Autom. Lang. Comb.*, 7(3):351–376, 2002.

[Ros02]   M. Rosen. *Number theory in function fields*, volume 210 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[RS59]   M. O. Rabin and D. Scott. Finite automata and their decision problems. *IBM J. Res. Develop.*, 3:114–125, 1959.

[RSW11]   N. Rampersad, J. Shallit, and M. Wang. Inverse star, borders, and palstars. *Inform. Process. Lett.*, 111(9):420–422, 2011.

[Rub04]   S. Rubin. Automatic structures. 2004. Ph.D. thesis. University of Auckland, New Zealand.

[RV11]   M. Rigo and É. Vandomme. Syntactic complexity of ultimately periodic sets of integers. *Lect. Notes in Comput. Sci.*, 6638:477–488, 2011. Proceedings of the fifth conference LATA (Languages and Automata, Theory and Applications).

[RW06]   M. Rigo and L. Waxweiler. A note on syndeticity, recognizable sets and Cobham's theorem. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, (88):169–173, 2006.

[RW11]   M. Rigo and L. Waxweiler. Logical characterization of recognizable sets of polynomials over a finite field. *Internat. J. Found. Comput. Sci.*, 22(7):1549–1563, 2011.

[Sak03]   J. Sakarovitch. *Éléments de théorie des automates*. Vuibert, Paris, 2003.

[Sal87]   O. Salon. Suites automatiques à multi-indices. *Séminaire de théorie des nombres de Bordeaux*, 4:4.01–4.27, 1986-1987.

[SB96]   J. Shallit and Y. Breitbart. Automaticity. I. Properties of a measure of descriptional complexity. *J. Comput. System Sci.*, 53(1):10–25, 1996.

[Sem77]   A. L. Semenov. The Presburger nature of predicates that are regular in two number systems. *Sibirsk. Mat. Ž.*, 18(2):403–418, 479, 1977.

[Sha88]   J. Shallit. A generalization of automatic sequences. *Theoret. Comput. Sci.*, 61(1):1–16, 1988.

[Sha96]   J. Shallit. Automaticity. IV. Sequences, sets, and diversity. *J. Théor. Nombres Bordeaux*, 8(2):347–367, 1996.

[Sha09]   J. Shallit. *A Second Course in Formal Languages and Automata Theory*. Cambridge University Press, 2009.

[Tra64]   B. Trakhtenbrot. On an estimate for the weight of a finite tree. *Sibirsk. Mat. Zh.*, 5:186–191, 1964.

[Vil92a] R. Villemaire. Joining k- and l-recognizable sets of natural numbers. In *STACS 92 (Cachan, 1992)*, volume 577 of *Lecture Notes in Comput. Sci.*, pages 83–94. Springer, Berlin, 1992.

[Vil92b] R. Villemaire. The theory of $\langle \mathbf{N}, +, V_k, V_l \rangle$ is undecidable. *Theoret. Comput. Sci.*, 106(2):337–349, 1992.

[Zec72]   E. Zeckendorf. Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas. *Bull. Soc. Roy. Sci. Liège*, 41:179–182, 1972.

Index