On the expressive power of univariate equations over sets of natural numbers[*]

Alexander Okhotin[1,2], and Panos Rondogiannis[3]

[1] Academy of Finland
[2] Dept. of Mathematics, University of Turku, Finland
[3] Dept. of Informatics and Telecommunications, University of Athens, Greece

Abstract. Equations of the form $X = \varphi(X)$ are considered, where the unknown X is a set of natural numbers. The expression $\varphi(X)$ may contain the operations of set addition, defined as $S + T = \{m + n \mid m \in S, n \in T\}$, union and intersection, as well as ultimately periodic constants. An equation with a non-periodic solution of exponential growth is constructed. At the same time it is demonstrated that no sets with super-exponential growth can be represented. It is also shown that a restricted class of these equations cannot represent sets with super-linearly growing complements. The results have direct implications on the power of conjunctive grammars with one nonterminal symbol.

1 Introduction

Language equations, in which the unknowns are formal languages, have recently become an active topic of study [5]. Formal languages are typically considered over an alphabet containing at least two letters. For a unary alphabet $\Sigma = \{a\}$, they can be regarded as sets of natural numbers. Then the operation of concatenating such languages turns into pairwise addition of sets: $S + T = \{\, m + n \mid m \in S,\ n \in T \,\}$. Language equations accordingly become equations over sets of numbers. Even in this seemingly simple case they already have quite surprising properties.

Consider systems of equations of the form

$$X_i = \varphi_i(X_1, \ldots, X_n) \quad (1 \leqslant i \leqslant n), \tag{*}$$

where the unknowns X_i are subsets of $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, while the right-hand sides φ_i contain union, addition and singleton constants. These systems are equivalent to language equations of the same form (*) over a unary alphabet using the operations of union and concatenation, and accordingly represent context-free grammars. As it is well-known that all unary context-free languages

[*] Research supported by the Academy of Finland under grant 118540, and by the Greek General Secretariat for Research and Technology under the program $\Pi ENE\Delta$ (grant number $03E\Delta$ 330).

Please use the following format when citing this chapter:

Okhotin, A. and Rondogiannis, P., 2008, in IFIP International Federation for Information Processing, Volume 273; Fifth IFIP International Conference on Theoretical Computer Science; Giorgio Ausiello, Juhani Karhumäki, Giancarlo Mauri, Luke Ong; (Boston: Springer), pp. 215–227.

are regular, least solutions of systems (*) over sets of numbers are vectors of ultimately periodic sets.

Another kind of equations are systems of the form (*) with *addition and complementation.* An example of such an equation with a non-periodic solution was given by Leiss [6]. Later Okhotin and Yakimova [8] established the main properties of systems of such equations (in the more general case of language equations) and gave a direct proof that a certain rather simple non-periodic set is not representable.

Consider systems of the same general form (*), in which the allowed operations are *union, intersection and addition.* These systems correspond to an extension of the context-free grammars, the *conjunctive grammars* [7], which are again considered over a unary alphabet. The question of whether conjunctive grammars can generate any non-regular unary languages has been an open problem for some years [7], until recently solved by Jeż [3], who constructed a grammar for the language $\{ a^{4^n} \mid n \geqslant 0 \}$. This grammar can be regarded as a system (*) of four equations over sets of numbers using union, intersection and addition, such that one of the four components of its least solution is $\{ 4^n \mid n \geqslant 0 \}$.

The set $\{ 4^n \mid n \geqslant 0 \}$ grows exponentially, so this example left a question of whether any super-exponentially growing sets are representable. A strong answer was given by Jeż and Okhotin [4], who showed that for every given recursive function it is possible to represent a set that grows faster.

Despite these extensive positive results (and maybe to some extent *due* to these positive results), no results saying that some particular set *cannot* be represented by such equations could so far be obtained. The $DTIME(n^2) \cap DSPACE(n)$ complexity upper bound for conjunctive grammars over a unary alphabet is the only known restriction. Otherwise, no techniques of proving non-representability of sets by equations with union, intersection and addition are known.

This paper considers a particular case of systems (*) with $n = 1$: these are equations of the form $X = \varphi(X)$, where X is a unique variable and φ is an expression containing arbitrarily nested union, intersection, sum and ultimately periodic constants. Every such equation has a least solution given by $\bigcup_{n=0}^{\infty} \varphi^n(\varnothing)$. It is shown that these equations can represent a certain non-periodic set of an exponential growth rate: namely, the example of Jeż [3] is reconstructed using one variable instead of four. At the same time it is proved that no sets that grow asymptotically faster than exponential can be represented. Another class of sets is shown to be non-representable by a restricted class of such equations: these are *dense sets*, that is, sets with super-linearly growing complements. In overall, it is demonstrated that one-variable equations are weaker in power than systems of multiple equations. This also demonstrates that conjunctive grammars with a single nonterminal cannot generate all conjunctive languages.

2 Conjunctive grammars and systems of equations

Conjunctive grammars form a natural extension of the context-free grammars that supports intersection in the right-hand sides of rules:

Definition 1 ([7]). A conjunctive grammar is a quadruple $G = (\Sigma, N, P, S)$, where Σ and N are disjoint finite nonempty sets of terminal and nonterminal symbols respectively, P is a finite set of rules, each of the form

$$A \to \alpha_1 \& \ldots \& \alpha_n \qquad (n \geqslant 1,\ A \in N,\ \alpha_i \in (\Sigma \cup N)^*) \tag{1}$$

and $S \in N$ is the start symbol. A grammar is said to be linear conjunctive if furthermore each α_i in each rule (1) is in $\Sigma^* N \Sigma^*$ or in Σ^*.

One way to define the semantics of conjunctive grammars is by term rewriting. Consider terms over concatenation and conjunction. Then a subterm A can be rewritten with $(\alpha_1 \& \ldots \& \alpha_n)$ for every rule (1), and any subterm of the form $(w \& \ldots \& w)$, with $w \in \Sigma^*$, can be rewritten with w. Then $L(G)$ is defined as the set of all strings $w \in \Sigma^*$ that are derivable from the term S.

An equivalent definition can be given using language equations.

Definition 2. For every conjunctive grammar $G = (\Sigma, N, P, S)$, the associated system of language equations is a system of equations in variables N, in which each variable assumes a value of a language over Σ, and which contains the following equation for every variable A:

$$A = \bigcup_{A \to \alpha_1 \& \ldots \& \alpha_m \in P} \ \bigcap_{i=1}^{m} \alpha_i \quad (\text{for all } A \in N)\ . \tag{2}$$

Each instance of a symbol $a \in \Sigma$ in such a system defines a constant language $\{a\}$, while each empty string denotes a constant language $\{\varepsilon\}$. A solution of such a system is a vector of languages $(\ldots, L_C, \ldots)_{C \in N}$, such that the substitution of L_C for C, for all $C \in N$, turns each equation (2) into an equality.

Let $(\ldots, L_C, \ldots)$ be the least solution of the system and define $L_G(C) = L_C$ for all $C \in N$ and $L(G) = L_G(S)$.

Consider conjunctive grammars over a one-symbol alphabet, with $\Sigma = \{a\}$. A formal language $L \subseteq a^*$ can be regarded as a set of numbers $\{\, n \mid a^n \in L \,\}$. The operation of concatenation of languages is replaced with pairwise addition of sets: for all $S, T \subseteq \mathbb{N}$, define

$$S + T = \{\, m + n \mid m \in S,\ \text{and } n \in T \,\}$$

Thus a system of language equations (2) corresponding to a conjunctive grammar over $\{a\}$ can be regarded as a system of equations over sets of natural numbers.

For unary languages, being regular means to be ultimately periodic as a set of numbers. A set S is *ultimately periodic* if there exist numbers $d, p \geqslant 0$, such that for any $n \geqslant d$, the number n is in S if and only if $n+p$ is in S. Such a set is also said to be periodic starting from d with period p.

The first example of a system of equations with union, intersection and addition representing a non-periodic set (originally presented in the form of a conjunctive grammar) is as follows:

Example 1 (Jeż [3]). The system of equations

$$\begin{cases} X_1 = \big((X_1 + X_3) \cap (X_2 + X_2)\big) \ \cup \ \{1\} \\ X_2 = \big((X_1 + X_1) \cap (X_6 + X_2)\big) \ \cup \ \{2\} \\ X_3 = \big((X_1 + X_2) \cap (X_6 + X_6)\big) \ \cup \ \{3\} \\ X_6 = \big((X_1 + X_2) \cap (X_3 + X_3)\big) \end{cases}$$

has the least solution $X_k = \{\, k \cdot 4^n \mid n \geqslant 0 \,\}$, for $k = 1, 2, 3, 6$.

The idea of this construction is best understood in terms of *positional notation* of numbers. Let $\Sigma_k = \{\mathtt{0}, \mathtt{1}, \ldots, k-1\}$ be digits in base-k notation. For every $w \in \Sigma_k^*$, let $(w)_k$ be the number defined by this string of digits. Define $(L)_k = \{\, (w)_k \mid w \in L \,\}$. Now the solution of the above system can be represented in base-4 notation as the vector $\big((\mathtt{10}^*)_4, (\mathtt{20}^*)_4, (\mathtt{30}^*)_4, (\mathtt{120}^*)_4\big)$. Let us substitute this vector into the right-hand side of the first equation:

$$\begin{aligned} &\big((\mathtt{10}^*)_4 + (\mathtt{30}^*)_4\big) \cap \big((\mathtt{20}^*)_4 + (\mathtt{20}^*)_4\big) = \\ &\quad = \big((\mathtt{10}^*\mathtt{30}^*)_4 \cup (\mathtt{10}^+)_4 \cup (\mathtt{30}^*\mathtt{10}^*)_4\big) \cap \big((\mathtt{20}^*\mathtt{20}^*)_4 \cup (\mathtt{10}^+)_4\big) = (\mathtt{10}^+)_4 \end{aligned}$$

Taking the singleton $\{1\}$ into account, the set $(\mathtt{10}^*)_4$ is obtained.

In order to minimize the number of brackets, the subsequent examples will assume the following default precedence of operations: addition has the highest precedence, intersection has intermediate precedence, and the precedence of union is the lowest. Also, singleton constants $\{n\}$ will sometimes be written as n.

Let us define the notion of a growth rate of a set. Every infinite set of numbers $L = \{i_1, i_2, \ldots, i_n, \ldots\}$, with $0 \leqslant i_1 < i_2 < \ldots < i_n < \ldots$, can be regarded as an increasing integer sequence. The *growth rate* of such sequences is represented by a function $g(n) = i_n$. The set from Example 1 has exponential growth rate.

The method of manipulating positional notations of numbers using addition of sets has been further extended in the following way. Consider a linear conjunctive grammar generating base-k positional notations of some numbers. Then the set of these numbers can be specified by a system of equations over sets of numbers.

Theorem 1 (Jeż, Okhotin [4]). *For every $k \geqslant 2$ and for every linear conjunctive grammar G over Σ_k there exists a system of equations $X = \varphi_i(X_1, \ldots, X_n)$ over sets of natural numbers with the least solution $X_i = S_i$, in which $S_1 = (L(G))_k$.*

This theorem has several important implications. One of them is that the growth rate of representable sets is not bounded by any fixed recursive function.

Theorem 2 (Jeż, Okhotin [4]). *For every recursively enumerable set of natural numbers S there exists a system $X_i = \varphi_i(X_1, \dots, X_n)$ over sets of natural numbers with the least solution $X_i = S_i$, such that the growth function of S_1 is greater than that of S at any point.*

There are four variables in the system in Example 1, while Theorems 1–2 use quite many variables. The purpose of this paper is to investigate the expressibility of univariate equations.

3 Equations with one variable

Consider an equation

$$X = \varphi(X),$$

where the unknown X is a set of natural numbers, while φ uses union, intersection and addition, as well as ultimately periodic constants. These operations can, in general, be arbitrarily nested. It is known from the fixed point theory that $\bigcup_{i \geqslant 0} \varphi^i(\varnothing)$ is the least (wrt set inclusion) among all the solutions of the equation.

A particular case of such equations are those corresponding to one-nonterminal conjunctive grammars, where φ must be a union of intersections of sums, and it is interesting to note that already in this case every ultimately periodic set can be represented using singleton constants.

Lemma 1 (Alhazov [1]). *Every unary regular language is generated by a one-nonterminal conjunctive grammar.*

Proof. Let $K \cup (a^p)^+ L$ be the given language, where $K, L \subseteq \{\varepsilon, a, \dots, a^{p-1}\}$. Then the required grammar is

$$\begin{aligned} &S \to a^i \quad (a^i \in K \cup a^p L \cup a^{2p} L) \\ &S \to a^p S \& a^{2p} S \end{aligned} \qquad \square$$

The question is, whether any non-periodic sets can be represented using univariate equations. As the following lemma demonstrates, this is indeed the case:

Lemma 2. *The following one-variable equation has the unique solution* $\{\, 4^n - 8 \mid n \geqslant 3 \,\} \cup \{\, 2 \cdot 4^n - 15 \mid n \geqslant 3 \,\} \cup \{\, 3 \cdot 4^n - 11 \mid n \geqslant 3 \,\} \cup \{\, 6 \cdot 4^n - 9 \mid n \geqslant 3 \,\}$*:*

$$\begin{aligned} X = {} & \big(11{+}X{+}X \,\cap\, 22{+}X{+}X\big) \cup \big(1{+}X{+}X \,\cap\, 9{+}X{+}X\big) \cup \\ & \cup \big(7{+}X{+}X \,\cap\, 12{+}X{+}X\big) \cup \big(13{+}X{+}X \,\cap\, 14{+}X{+}X\big) \cup \{56, 113, 181\} \end{aligned}$$

Here addition is assumed to have higher precedence than intersection.

The idea behind this construction is to encode four variables from Example 1 into a single variable. The unique solution of the constructed equation is a union of four disjoint sets:

$$\begin{aligned} L_1 &= \{\, 4^n - 8 \mid n \geqslant 3 \,\} \\ L_2 &= \{\, 2 \cdot 4^n - 15 \mid n \geqslant 3 \,\} \\ L_3 &= \{\, 3 \cdot 4^n - 11 \mid n \geqslant 3 \,\} \\ L_6 &= \{\, 6 \cdot 4^n - 9 \mid n \geqslant 3 \,\} \end{aligned}$$

Each of them represents the corresponding component of the solution of the system from Example 1. These components are represented with an *offset*: the numbers in L_1, L_2, L_3 and L_6 are smaller by $d_1 = 8$, $d_2 = 15$, $d_3 = 11$ and $d_6 = 9$, respectively.

Consider first the following system:

$$\begin{cases} Y_1 = \big(11{+}Y_1{+}Y_3 \ \cap \ 22{+}Y_2{+}Y_2\big) \ \cup \ \{56\} \\ Y_2 = \big(1{+}Y_1{+}Y_1 \ \cap \ 9{+}Y_6{+}Y_2\big) \ \cup \ \{113\} \\ Y_3 = \big(7{+}Y_6{+}Y_6 \ \cap \ 12{+}Y_1{+}Y_2\big) \ \cup \ \{181\} \\ Y_6 = 13{+}Y_3{+}Y_3 \cap 14{+}Y_1{+}Y_2 \end{cases} \tag{3}$$

This system is obtained from the system in Example 1 as follows. First, the constant sets $\{1\}$, $\{2\}$ and $\{3\}$ are replaced with $\{64\}$, $\{128\}$ and $\{192\}$, so that the values of n in the solution start from 3. Then the substitution $X_1 = Y_1 + 8$, $X_2 = Y_2 + 15$, $X_3 = Y_3 + 11$, $X_6 = Y_6 + 9$ is applied. It is easy to see that the solution of system (3) is the vector (L_1, L_2, L_3, L_6).

Note that each set L_i is a subset of a periodic set $\{\, 64m - d_i \mid m \geqslant 1 \,\}$. Let us call every such periodic superset *a track*. The sum of any two of these sets, $L_i + L_j$, is a subset of $\{\, 64m - d_i - d_j \mid m \geqslant 2 \,\}$, which is a track as well. The numbers 8, 15, 11 and 9 have been chosen so that the sums of all pairs of these numbers are pairwise distinct: $d_i + d_j = d_k + d_\ell$ with $i \leqslant j$ and $k \leqslant \ell$ implies $i = k$ and $j = \ell$. In other words, the tracks are pairwise disjoint, and the calculations in the right-hand sides of different equations occur in different tracks.

This property is used to ensure that if the same set $L_1 \cup L_2 \cup L_3 \cup L_6$ is substituted for *every variable* in the right-hand sides of (3), then every right-hand side still evaluates to L_1, L_2, L_3 and L_6, respectively. Now the equation in Lemma 2 is obtained from the system (3) by identifying all four variables into one.

It must be admitted that these ideas do not work in general, and Lemma 2 is not proved by a formal transformation. However, they happen to work for the given example and with the given assignment of offsets to variables. The lemma can actually be proved by substituting the given set into the equation and verifying that it is indeed a solution. The proof is omitted in this extended abstract due to its pure technicality.

The equation in Lemma 2 has a simple form corresponding to a conjunctive grammar. The result can thus be restated in the following form.

Example 2. The following one-nonterminal conjunctive grammar generates the language $\{\, a^{4^n-8} \mid n \geqslant 3 \,\} \cup \{\, a^{2 \cdot 4^n-15} \mid n \geqslant 3 \,\} \cup \{\, a^{3 \cdot 4^n-11} \mid n \geqslant 3 \,\} \cup \{\, a^{6 \cdot 4^n-9} \mid n \geqslant 3 \,\}$:

$$S \to a^{22}SS \& a^{11}SS \mid a^{9}SS \& aSS \mid a^{7}SS \& a^{12}SS \mid a^{13}SS \& a^{14}SS \mid a^{56} \mid a^{113} \mid a^{181}$$

This example answers the question raised by Jeż [3] about the least number of nonterminals in a conjunctive grammar necessary to generate non-regular languages over $\{a\}$: *one is enough.*

4 Non-representability of fast growing sets

The set represented in Lemma 2 has exponential growth. It will now be shown that sets with asymptotically super-exponential growth cannot be represented by univariate equations. The following statement is also applicable to some sets that do not formally fit this description.

Theorem 3. *Let $L = \{n_1, n_2, \ldots, n_i, \ldots\}$ with $0 \leqslant n_1 < n_2 < \ldots < n_i < \ldots$ be an infinite set of natural numbers, for which $\liminf_{i \to \infty} \frac{n_i}{n_{i+1}} = 0$. Then L is not the least solution of any univariate equation $X = \varphi(X)$.*

In particular, the theorem asserts non-representability of sets like $\{\, 2^{2^n} \mid n \geqslant 0 \,\}$ and $\{\, n! \mid n \geqslant 1 \,\}$, as well as sets like $\{\, n!, n!+1 \mid n \geqslant 1 \,\}$.

The assumption that limit inferior of $\frac{n_i}{n_{i+1}}$ as n approaches infinity is zero means that the size of gaps between consecutive numbers (measured relatively to the smaller number) is not bounded. That is, for every k there is $n \in L$ so that L does not contain any numbers between $n+1$ and kn.

If such a set is a least solution of an equation, then L can be expressed from itself and from ultimately periodic constants using union, intersection and addition. Then the gaps between elements of the set have to be bridged either by summing up several smaller elements of this set in an expression $X + \ldots + X$, or by adding an ultimately periodic constant to X. The expression φ contains only finitely many additions, and hence only a bounded number of smaller elements can be added up. Larger gaps can only be bridged by adding an ultimately periodic constant. However, this addition would make the sum ultimately periodic as well.

This reasoning is formalized in the following statement:

Lemma 3. *Let $\varphi(X)$ be an expression that contains instances of a unique variable X ultimately periodic constants with a common period p starting from d,* *and the operations of union, intersection and addition. Let h be the greatest number of nested additions in φ. Let a number n and a set of numbers L be such that $n \in \varphi(L)$, $L \cap \{\lceil \frac{n}{2^h} \rceil, \lceil \frac{n}{2^h} \rceil + 1, \ldots, n-1\} = \varnothing$ and $\frac{n}{2^h} \geqslant d + p$. Then $n \in L$ or $n - p \in \varphi(L)$.*

Proof. Induction on the structure of φ.

Basis I: $\varphi(X) = X$. Then $n \in \varphi(L)$ means $n \in L$.

Basis II: $\varphi(X) = C$, where C is an ultimately periodic set of natural numbers. Then $h = 0$ and hence $n \geqslant d + p$ by assumption. Since C has period p starting from d, $n \in \varphi(L) = C$ is equivalent to $n - p \in C = \varphi(L)$.

Induction step I: $\varphi(X) = \varphi_1(X) \cup \varphi_2(X)$. Then $n \in \varphi(L)$ implies that $n \in \varphi_i(L)$ for some $i \in \{1, 2\}$. Assume without loss of generality that $n \in \varphi_1(L)$. Let h_1 be the greatest number of nested additions in φ_1; obviously, $h_1 \leqslant h$. Then $\frac{n}{2^{h_1}} \geqslant \frac{n}{2^h}$ and therefore $L \cap \{\lceil \frac{n}{2^{h_1}} \rceil, \lceil \frac{n}{2^{h_1}} \rceil + 1, \ldots, n-1\} = \varnothing$ and $\frac{n}{2^{h_1}} \geqslant d + p$. Thus the induction hypothesis is applicable to φ_1 and n, giving that $n \in L$ or $n - p \in \varphi_1(L) \subseteq \varphi(L)$.

Induction step II: $\varphi(X) = \varphi_1(X) \cap \varphi_2(X)$. In this case $n \in \varphi(L)$ implies both $n \in \varphi_1(L)$ and $n \in \varphi_2(L)$. Let h_1 and h_2 be the greatest numbers of nested additions in φ_1 and φ_2, respectively, for which it is known that $h_1 \leqslant h$ and $h_2 \leqslant h$. As in the case of union, the induction hypothesis is applicable to φ_1 and n, as well as to φ_2 and n, which gives $n \in L$ or $n - p \in \varphi_1(L)$, and at the same time $n \in L$ or $n - p \in \varphi_2(L)$. If either subexpression yields $n \in L$, this immediately proves the claim for φ and n. Otherwise the number $n - p$ is known to be both in $\varphi_1(L)$ and in $\varphi_2(L)$, which means $n - p \in \varphi(L)$.

Induction step III: $\varphi(X) = \varphi_1(X) + \varphi_2(X)$. Then it follows from $n \in \varphi(L)$ that there are two numbers $n_1, n_2 \geqslant 0$ with $n_1 + n_2 = n$ and $n_i \in \varphi_i(L)$ for $i \in \{1, 2\}$. Assume without loss of generality that $n_1 \geqslant n_2$. Let h_1 be the greatest number of nested additions in φ_1, which is known to be at most $h - 1$. Then $\frac{n_1}{2^{h_1}} \geqslant \frac{n_1}{2^{h-1}} \geqslant \frac{n}{2} \cdot \frac{1}{2^{h-1}} = \frac{n}{2^h}$, and therefore $L \cap \{\lceil \frac{n_1}{2^{h_1}} \rceil, \lceil \frac{n_1}{2^{h_1}} \rceil + 1, \ldots, n-1\} = \varnothing$ and $\frac{n_1}{2^{h_1}} \geqslant d + p$. By the induction hypothesis for φ_1 and n_1 it follows that $n_1 \in L$ or $n_1 - p \in \varphi_1(L)$. Consider each of these cases:

- In the former case, note that $\frac{n}{2} \leqslant n_1 \leqslant n$. Since $h \geqslant 1$ and $L \cap \{\lceil \frac{n}{2^h} \rceil, \lceil \frac{n}{2^h} \rceil + 1, \ldots, n-1\} = \varnothing$ by assumption, $n_1 \in L$ implies that n_1 must be equal to n, while n_2 must be zero. This proves that $n \in L$.
- If $n_1 - p \in \varphi_1(L)$, then $n - p = (n_1 - p) + n_2 \in \varphi(L)$.

This last case completes the proof of the lemma. □

Proof (Theorem 3). Suppose there exists an equation $X = \varphi(X)$ with the least solution L_0. Let $C_1, \ldots, C_m$ be all constants used in φ, and let each C_i have period p_i starting from d_i. Let $p = \operatorname{lcm}\{p_1, \ldots, p_m\}$ and $d = \max\{d_1, \ldots, d_m\}$; then all constants have period p starting from d. Denote the greatest number of nested additions in φ by h.

By the definition of limit inferior, there exist infinitely many numbers i with $\frac{n_i}{n_{i+1}} < \frac{1}{2^h}$. Then it is possible to choose a sufficiently large i so that $\frac{n_{i+1}}{2^h} \geqslant d + p$.

Now $n_i < \frac{n_{i+1}}{2^h} \leqslant \lceil \frac{n_{i+1}}{2^h} \rceil$, and since L_0 contains no elements between $n_i + 1$ and $n_{i+1} - 1$, it follows that $L_0 \cap \{\lceil \frac{n_{i+1}}{2^h} \rceil, \ldots, n_{i+1} - 1\} = \varnothing$.

Since L_0 is the least fixed point of φ, there exists a number of iterations ℓ, for which $n_{i+1} \notin \varphi^{\ell}(\varnothing)$ and $n_{i+1} \in \varphi^{\ell+1}(\varnothing)$. Denote $L = \varphi^{\ell}(\varnothing)$, that is, $n_{i+1} \notin L$ and $n_{i+1} \in \varphi(L)$. Since $L \subseteq L_0$, it is known that $L \cap \{\lceil \frac{n_{i+1}}{2^h} \rceil, \ldots, n_{i+1} - 1\} = \varnothing$. Therefore, Lemma 3 is applicable to φ, n and L, and it asserts that $n \in L$ or $n - p \in \varphi(L)$. The former contradicts the assumption, while the latter is not possible since $\lceil \frac{n}{2^h} \rceil \leqslant n - p \leqslant n - 1$. The contradiction obtained proves the theorem. □

Theorem 3 implies a separation of one-nonterminal conjunctive languages from conjunctive languages of the general form.

Theorem 4. *The following proper containments hold:*

$$\textsc{Reg}_{\{a\}} \subset \textsc{Conj}^1_{\{a\}} \subset \textsc{Conj}_{\{a\}}$$

Proof. In particular, $\textsc{Conj}^1_{\{a\}} \setminus \mathit{Reg}$ contains the language from Example 2, while $\textsc{Conj}_{\{a\}} \setminus \textsc{Conj}^1_{\{a\}}$ contains some languages growing faster than exponential (and as it will be demonstrated in the next section, also some languages with super-linearly growing complements). □

5 Non-representability of dense sets

In this section we derive non-representability results concerning a class of sets that are known as *additive bases*:

Definition 3. Let $S \subseteq \mathbb{N}$ be an infinite set of natural numbers, and let $k > 0$. For any $n \in \mathbb{N}$, define the number of its representations as a sum of k elements of S by

$$r_{k,S}(n) = |\{(a_1, \ldots, a_k) \in S^k : a_1 + \cdots + a_k = n\}|.$$

The set S is said to be a basis of order k if every sufficiently large natural number n can be represented as sum of k (not necessarily distinct) elements of S, or equivalently if $r_{k,S}(n) \geqslant 1$. In other words, S is a basis of order k if and only if $\underbrace{(S + \cdots + S)}_{k}$ is co-finite.

As an example, there is a well-known result, *Legendre's theorem*, that the set of squares of the natural numbers is a basis of order four.

Clearly, if a set S is a basis of order k then it is also a basis of every order $n > k$. The non-representability results we will obtain in this section, are for sets that are bases of order 2.

We start with a class of sets that are dense additive bases of order 2:

Definition 4. Given any $m, n \in \mathbb{N}$, let $[m, n]$ denote the discrete closed interval $[m, n] = \{i \in \mathbb{N} : m \leqslant i \leqslant n\}$.

A set $L \subseteq \mathbb{N}$ is said to be *dense* if $\lim_{n\to\infty} \frac{|L \cap [0,n]|}{n} = 1$.

For example, the set $\mathbb{N} \setminus \{ 2^n \mid n \geqslant 0 \}$ is obviously dense, and so is the set of composite numbers.

The following lemma is easy to establish using basic properties of limits:

Lemma 4. *Let L be a dense set. Then*

$$\lim_{n\to\infty} \frac{|(\mathbb{N} \setminus L) \cap [0,n]|}{|L \cap [0,n]|} = \lim_{n\to\infty} \frac{|(\mathbb{N} \setminus L) \cap [0,n]|}{n} = 0.$$

Similarly to Theorem 3, the following theorem states that sets of the above form cannot be represented using univariate equations that use finite or co-finite constants:

Theorem 5. *Let L be a dense non-ultimately periodic set. Then there is no univariate equation $X = \varphi(X)$ using finite and co-finite constants, which would have the least solution L.*

The proof of the theorem is based upon the following three lemmas.

Lemma 5. *Let $L_1 \subseteq \mathbb{N}$ and $L_2 \subseteq \mathbb{N}$ be dense sets. Then the set $L_1 + L_2$ is co-finite.*

Notice that this lemma implies that dense sets are additive bases of order 2 (just take $L_1 = L_2$).

Proof. The main idea of the proof is that every sufficiently large element of $\mathbb{N}$ can be written as the sum of two elements of $\mathbb{N}$ in *too many ways*. Now, since the sets $\mathbb{N} \setminus L_1$ and $\mathbb{N} \setminus L_2$ are "sparse", every sufficiently large element of $\mathbb{N}$ can also be written as the sum of at least two elements of L_1 and L_2. In other words, $\mathbb{N} \setminus (L_1 + L_2)$ is finite.

More formally now, it suffices to show that for every sufficiently large $n \in \mathbb{N}$ there exist $\ell_1 \in L_1$ and $\ell_2 \in L_2$ such that $n = \ell_1 + \ell_2$. Consider the number of ways in which a number n can be written as a sum of two numbers $n_1 \in L_1$ and $n_2 \in L_2$. More specifically, given $n \in \mathbb{N}$, define the functions:

$$\begin{aligned} p(n) &= |\{(n_1, n_2) : (n_1 \in \mathbb{N}) \text{ and } (n_2 \in \mathbb{N}) \text{ and } (n_1 + n_2 = n)\}| \\ r_1(n) &= |\{k : (k \in \mathbb{N} \setminus L_1) \text{ and } (k \leqslant n)\}| \\ r_2(n) &= |\{k : (k \in \mathbb{N} \setminus L_2) \text{ and } (k \leqslant n)\}| \end{aligned}$$

Now it is easy to see that every sufficiently large number n in $\mathbb{N}$ can be written as $n = \ell_1 + \ell_2$, with $\ell_1 \in L_1$ and $\ell_2 \in L_2$, in at least $p(n) - r_1(n) - r_2(n)$ ways. To prove that $p(n) - r_1(n) - r_2(n) > 0$ for large values of n, it suffices to show that $\lim_{n\to\infty} \frac{p(n)}{n} > 0$, while $\lim_{n\to\infty} \frac{r_1(n)}{n} = 0$ and $\lim_{n\to\infty} \frac{r_2(n)}{n} = 0$.

Notice now that $p(n) = n + 1$ since n can be written as the sum of two elements of $\mathbb{N}$ in the following ways: $(0, n), (1, n-1), \ldots, (n, 0)$. Therefore, $\lim_{n\to\infty} \frac{p(n)}{n} = 1$. Consider now the case of $r_1(n)$ (the case of $r_2(n)$ is identical). Since L_1 is a dense set, Lemma 4 asserts that $\lim_{n\to\infty} \frac{|(\mathbb{N}\setminus L_1)\cap[0,n]|}{n} = 0$, and therefore $\lim_{n\to\infty} \frac{r_1(n)}{n} = 0$. It follows that $p(n) - r_1(n) - r_2(n) > 0$ (that is, $n \in L_1 + L_2$) for all sufficiently large $n \in \mathbb{N}$. Therefore, $\mathbb{N} \setminus (L_1 + L_2)$ is a finite set. □

Lemma 6. *Let $S_1, S_2 \subseteq \mathbb{N}$ be dense sets, let $T \subseteq \mathbb{N}$ be any non-empty set. Then the sets $S_1 \cap S_2$, $S_1 \cup T$ and $S_1 + T$ are dense.*

The proof, which is omitted, proceeds by using simple set-theoretic arguments, and the basic properties of limits.

Lemma 7. *Let $\varphi(X)$ be an expression using the variable X, finite or co-finite constants, together with the operations of union, intersection and addition. Let L be a dense set and assume that $\varphi(L)$ is infinite. Then, $\varphi(L)$ is a dense set.*

Proof. Follows from Lemma 6 by a straightforward induction.

Proof (Proof of Theorem 5). Let $X = \varphi(X)$ be an equation. Let us prove that L cannot be its least solution. The proof is by an induction on the number of subexpressions of the form $\psi(X) + \xi(X)$ in φ, in which both ψ and ξ contain some instances of X.

Basis. If there are no such additions, then the least solution must be ultimately periodic by the known results on language equations with one-sided concatenation [2]. Since L is non-periodic, a contradiction is obtained.

Induction step. Consider any of the smallest such subexpressions of φ, that is, let $\varphi(X) = \widehat{\varphi}(X, \widetilde{\varphi}(X))$, where $\widetilde{\varphi} = \psi + \xi$.

Consider first the case where both $\psi(L)$ and $\xi(L)$ are infinite. Let us show that $\widetilde{\varphi}(L)$ is co-finite. Indeed, by Lemma 7, $\psi(L)$ is a dense set and $\xi(L)$ is also a dense set. Then Lemma 5 states that $\mathbb{N} \setminus (\psi(L) + \xi(L))$ is a finite set. In other words, $\psi(L) + \xi(L) = \mathbb{N} \setminus F$ for some finite $F \subset \mathbb{N}$. Denote $\mathbb{N} \setminus F$ by R'.

Then $\varphi(L) = \widehat{\varphi}(L, \widetilde{\varphi}(L)) = \widehat{\varphi}(L, R')$. Let $\varphi'(X)$ be a new expression defined as $\widehat{\varphi}(X, R')$. Then L should be the least solution of the equation $X = \varphi'(X)$. Since $\varphi'(X)$ contains fewer subexpressions of the form $\psi(X) + \xi(X)$, by the induction hypothesis, L cannot be the least solution of this equation. A contradiction.

Now consider the remaining case of $\psi(L)$ being a finite set, say F. Then $\varphi(L) = \widehat{\varphi}(L, \widetilde{\varphi}(L)) = \widehat{\varphi}(L, F + \xi(L))$. Define a new expression $\varphi'(X)$ as $\widehat{\varphi}(X, F + \xi(X))$; the set L should be the least solution of the equation $X = \varphi'(X)$. However, φ' contains fewer subexpressions of the form $\psi(X) + \xi(X)$, and hence L is not its least solution. This last contradiction establishes the induction step and concludes the proof. □

An immediate consequence of this result is that the class of sets of natural numbers that can be defined using univariate equations containing only finite or co-finite constants is not closed under complementation. Indeed, the complement of the language in Lemma 2 is dense and falls under Theorem 5. In particular, the class of unary languages generated by conjunctive grammars with one nonterminal is not closed under complementation.

Note that the equations corresponding to conjunctive grammars have a particular form, in which union and intersection may not be nested within addition. Further non-representability results for one-nonterminal conjunctive grammars can be obtained by using this form:

Theorem 6. *Let L be an additive basis of order 2 that is not ultimately periodic. Then L is not the least solution of any univariate equation $X = \varphi(X)$ that uses ultimately periodic constants, together with the operations of union, intersection and addition and in which union and intersection can not be nested within addition.*

Proof (a sketch). Let $X = \varphi(X)$ be an equation. Let us prove that L cannot be its least solution. Consider any subexpression of φ of the form $X + \cdots + X$. Since L is a basis, the corresponding sum $L + \cdots + L$ is co-finite and therefore ultimately periodic. Replace every such expression in φ by a corresponding constant. If there are no such additions left, then the least solution of the resulting equation must be ultimately periodic by the known results on language equations with one-sided concatenation [2], which is a contradiction. □

It follows that the family of unary languages generated by one-nonterminal conjunctive grammars does not contain any non-periodic additive bases of order 2.

6 Conclusions

It was shown that univariate equations $X = \varphi(X)$ with union, intersection and addition are, on one hand, nontrivial in the sense that they can represent some non-periodic sets. On the other hand, counting arguments were used to show that they cannot represent some sets that are known to be representable using systems of equations.

These non-representability results become the first of their kind, since no methods of proving sets non-representable by systems of equations with union, intersection and addition are currently known. This task appears challenging, though it is the authors' hope that the results obtained in this paper may also shed some light on this more general case.

Acknowledgement

We are grateful to an anonymous referee for his thorough comments.

References

1. A. Alhazov, personal communication, September 2007.
2. F. Baader, A. Okhotin, “Complexity of language equations with one-sided concatenation and all Boolean operations”, *20th International Workshop on Unification* (UNIF 2006, Seattle, USA, August 11, 2006), 59–73.
3. A. Jeż, “Conjunctive grammars can generate non-regular unary languages”, *Developments in Language Theory* (DLT 2007, Turku, Finland, July 3–6, 2007), LNCS 4588, 242–253.
4. A. Jeż, A. Okhotin, “Conjunctive grammars over a unary alphabet: undecidability and unbounded growth”, *Computer Science in Russia* (CSR 2007, Ekaterinburg, Russia, September 3–7, 2007), LNCS 4649, 168–181.
5. M. Kunc, “What do we know about language equations?”, *Developments in Language Theory* (DLT 2007, Turku, Finland, July 3–6, 2007), LNCS 4588, 23–27.
6. E. L. Leiss, “Unrestricted complementation in language equations over a one-letter alphabet”, *Theoretical Computer Science*, 132 (1994), 71–93.
7. A. Okhotin, “Conjunctive grammars”, *Journal of Automata, Languages and Combinatorics*, 6:4 (2001), 519–535.
8. A. Okhotin, O. Yakimova, “On language equations with complementation”, *Developments in Language Theory* (DLT 2006, Santa Barbara, USA, June 26–29, 2006), LNCS 4036, 420–432.