An Optimal Bound on the Solution Sets of One-Variable Word Equations and its Consequences

Dirk Nowotka[1]
Department of Computer Science, Kiel University, 24098 Kiel, Germany
dn@informatik.uni-kiel.de

Aleksi Saarela
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland
amsaar@utu.fi
https://orcid.org/0000-0002-6636-2317

Abstract

We solve two long-standing open problems on word equations. Firstly, we prove that a one-variable word equation with constants has either at most three or an infinite number of solutions. The existence of such a bound had been conjectured, and the bound three is optimal. Secondly, we consider independent systems of three-variable word equations without constants. If such a system has a nonperiodic solution, then this system of equations is at most of size 17. Although probably not optimal, this is the first finite bound found. However, the conjecture of that bound being actually two still remains open.

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words

Keywords and phrases combinatorics on words, word equations, systems of equations

Digital Object Identifier 10.4230/LIPIcs.ICALP.2018.136

Related Version A full version of the paper is available at `http://arxiv.org/abs/1805.09535`.

1 Introduction

If n words satisfy a nontrivial relation, they can be written as products of $n-1$ words. This folklore result is known as the defect theorem, and it can be seen as analogous to the simple fact of linear algebra that the dimension of the solution space of a homogeneous n-variable linear equation is $n-1$. If an independent equation is added to a system of linear equations, the dimension of the solution space decreases, which gives an upper bound n for the size of independent systems of linear equations, but no such results are known for word equations. In fact, the maximal size of independent systems of constant-free word equations has been one of the biggest open questions in combinatorics on words for many decades. In 1983, Culik and Karhumäki [4] pointed out that a conjecture of Ehrenfeucht about test sets of formal languages can be equivalently formulated as claiming that every infinite system of word equations is equivalent to a finite subsystem. Ehrenfeucht's conjecture was proved by Albert and Lawrence [1] and independently by Guba [9], and it follows that independent systems cannot be infinite, but no finite upper bounds depending only on the number of variables have been found. Independent systems of size $\Theta(n^4)$ on n variables were constructed by

[1] This work was partially supported by the DFG research project 181615770

45th International Colloquium on Automata, Languages, and Programming (ICALP 2018).
Editors: Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella;
Article No. 136; pp. 136:1–136:13

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Karhumäki and Plandowski [13], and the hidden constant in $\Theta(n^4)$ was improved in [14]. This is the best known lower bound.

The case of three variables is particularly interesting. In this case, it is easy to find systems of size two that are independent and have a nonperiodic solution, or systems of size three that are independent but have no nonperiodic solution, and Culik and Karhumäki conjectured that there are no larger such systems, but no finite upper bounds have been found even in this case. In fact, despite Ehrenfeucht's conjecture, even the existence of a bound is not guaranteed, because in principle it might be possible that there are unboundedly large finite independent systems. This case of three variables is very striking because it is the simplest nontrivial case, but the gap between the almost trivial lower bound and the infinite upper bound has remained huge despite the considerable attention the problem has received. Some results about systems of specific forms are known [10, 5, 6], and some upper bounds that depend on the sizes of the equations have been proved [17, 11, 16]. The best current bound is logarithmic with respect to the size of the smallest equation in the system [16].

In the above, we have considered constant-free word equations. If we add constants, the equations become more complicated. For constant-free equations, the three-variable case is the first nontrivial one, but for equations with constants, already the one-variable case is interesting. One-variable equations have been studied in many articles [8, 7, 15], and the main open question about them is the maximal number of solutions such an equation can have if we exclude equations with infinitely many solutions (if the solution set is infinite, it is known to be of a very specific form). Even finding an example with exactly two solutions is not entirely trivial, but a simple example was given by Laine and Plandowski [15]. An example with exactly three solutions was recently found [16]. No fixed upper bound, or even the existence of an upper bound, has been proved. The best known result is a bound that depends logarithmically on the number of occurrences of the variable in the equation [15]. It can be noted that the solutions of a one-variable equation can be found in linear time in the RAM model, as proved by Jeż [12].

In this article, we solve the open problem about sizes of solution sets of one-variable equations by proving that a one-variable equation has either infinitely many solutions or at most three, which is an optimal result. As a consequence, we prove the first upper bound for the sizes of independent systems of constant-free three-variable equations, thus settling the old open question about the existence of such a bound. More specifically, we prove that if an independent system of constant-free three-variable equations is independent and has a nonperiodic solution, then the system is of size at most 17 (if the system is not required to have a nonperiodic solution, then the size can be at most one larger). This bound is probably not optimal and the conjecture of Culik and Karhumäki remains open, as does the more general question about n-variable equations.

Two previous articles provide crucial tools for our proofs. The first article is [18], where new methods were introduced to solve a certain open problem on word equations. We use and further develop these methods to analyze one-variable equations. The second article is [16], where a surprising connection between the two topics we have discussed above was found: It was proved that a bound for the maximal size of a finite solution set of a one-variable equation implies a (larger) bound for the maximal size of independent systems of constant-free three-variable equations.

Some of the proofs have been omitted from this conference version to save space.

2 Preliminaries

We begin this section by considering constant-free word equations. Let Ξ be an alphabet of variables and Γ an alphabet of constants. A *constant-free word equation* is a pair $(U, V) \in \Xi^* \times \Xi^*$, and the solutions of this equation are the morphisms $h : \Xi^* \to \Gamma^*$ such that $h(U) = h(V)$. A solution h is *periodic* if there exists $p \in \Gamma^*$ such that $h(X) \in p^*$ for all $X \in \Xi$. Otherwise, h is *nonperiodic*. It is well-known that h is periodic if and only if $h(PQ) = h(QP)$ for all words $P, Q \in \Xi^*$.

▶ **Example 1.** Let $\Xi = \{X, Y, Z\}$ and consider the equation (XYZ, ZYX). For all $p, q \in \Gamma^*$ and $i, j, k \geq 0$, the morphism h defined by $h(X) = (pq)^i p$, $h(Y) = (qp)^j q$, $h(Z) = (pq)^k p$ is a solution of this equation because

$$h(XYZ) = (pq)^i p \cdot (qp)^j q \cdot (pq)^k p = (pq)^{i+j+k+1} p = (pq)^k p \cdot (qp)^j q \cdot (pq)^i p = h(ZYX).$$

Every nonperiodic solution of the equation is of this form.

A set of equations is a *system of equations*. A morphism is a solution of a system if it is a solution of every equation in the system. Two equations or systems are *equivalent* if they have exactly the same solutions. A system of equations is *independent* if it is not equivalent to any of its proper subsets.

▶ **Example 2.** Let $\Xi = \{X, Y, Z\}$ and $\Gamma = \{a, b\}$. The system of equations $S = \{(XYZ, ZYX), (XYYZ, ZYYX)\}$ is independent and has a nonperiodic solution h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = a$. To see independence, note that S is not equivalent to (XYZ, ZYX), because the morphism h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = aba$ is a solution of (XYZ, ZYX) but not of S, and S is not equivalent to $(XYYZ, ZYYX)$, because the morphism h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = abba$ is a solution of $(XYYZ, ZYYX)$ but not of S.

The following question is a big open problem on word equations: If a system of constant-free three-variable equations is independent and has a nonperiodic solution, then how large can the system be? The largest known examples are of size two, see Example 2, and it has been conjectured that these examples are optimal. Even the following weaker conjecture is open.

▶ **Conjecture 3.** *There exists a number c such that every independent system of constant-free three-variable equations with a nonperiodic solution is of size c or less.*

Currently, the best known result is the following.

▶ **Theorem 4** ([16])**.** *Every independent system of constant-free three-variable equations is of size $O(\log n)$, where n is the length of the shortest equation.*

Next, we will consider word equations with constants. As before, let Ξ be an alphabet of variables and Γ an alphabet of constants. A *word equation with constants* is a pair $(U, V) \in (\Xi \cup \Gamma)^* \times (\Xi \cup \Gamma)^*$, and the solutions of this equation are the constant-preserving morphisms $h : (\Xi \cup \Gamma)^* \to \Gamma^*$ such that $h(U) = h(V)$. If $U = V$, then the equation is *trivial*.

In this article, we are interested in the one-variable case $\Xi = \{X\}$. We use the notation $[u]$ for the constant-preserving morphism $h : (\{X\} \cup \Gamma)^* \to \Gamma^*$ defined by $h(X) = u$. If S is a set of words, we use the notation $[S] = \{[u] \mid u \in S\}$. If $[u]$ is a solution of a one-variable equation E, then u is called a *solution word* of E. The set of all solutions of E is denoted by $\mathrm{Sol}(E)$.

▶ **Example 5.** Let $\Gamma = \{a, b\}$. The equation (Xab, abX) has infinitely many solutions $[(ab)^i]$, where $i \geq 0$. The equation $(XaXbab, abaXbX)$ has exactly two solutions $[\varepsilon]$ and $[ab]$. The equation $(XXbaaba, aabaXbX)$ has exactly two solutions $[a]$ and $[aaba]$. The equation

$$(XaXbXaabbabaXbabaabbab, abaabbabaXbabaabbXaXbX)$$

has exactly three solutions $[\varepsilon], [ab], [abaabbab]$.

The following is a well-known open problem: If a one-variable equation has only finitely many solutions, then what is the maximal number of solutions it can have? Example 5 shows that the answer is at least three, but no upper bound is known. Currently, the best known result is the following.

▶ **Theorem 6** ([15, Theorems 23, 26, 29]). *If the solution set of a one-variable equation is finite, then it has size at most* $8 \log n + O(1)$*, where* n *is the number of occurrences of the variable.*

If the solution set is infinite and the equation is not trivial, then there are words p, q *such that* pq *is primitive and the solution set is* $[(pq)^* p]$*.*

We will need the following lemma.

▶ **Lemma 7** ([7, Lemma 1]). *Let* E *be a one-variable equation and let* pq *be primitive. The set*

$$\mathrm{Sol}(E) \cap [(pq)^+ p]$$

is either $[(pq)^+ p]$ *or has at most one element.*

A connection between constant-free three-variable equations and one-variable equations with constants was recently found [16]. Here we give the relevant special case of one of the results.

▶ **Theorem 8** ([16]). *If every one-variable word equation has either infinitely many solutions or at most three, then Conjecture 3 is true for* $c = 17$*.*

In this article, we will prove that every one-variable word equation has either infinitely many solutions or at most three, and thus Conjecture 3 is true for $c = 17$.

3 Sums of words

In this section, we will give some definitions and ideas that will be used in our proofs. Most of these were introduced in [18].

We can assume that the alphabet Γ is a subset of $\mathbb{R}$. Then we can define $\Sigma(w)$ to be the sum of the letters of a word $w \in \Gamma^*$, that is, if $w = a_1 \cdots a_n$ and $a_1, \ldots, a_n \in \Gamma$, then $\Sigma(w) = a_1 + \cdots + a_n$. Words w such that $\Sigma(w) = 0$ are called *zero-sum words*. If w is zero-sum, then the morphism $[w]$ can also be called zero-sum. The largest and smallest letters in a word w can be denoted by $\max(w)$ and $\min(w)$, respectively.

The *prefix sum word* of $w = a_1 \cdots a_n$ is the word $\mathrm{psw}(w) = b_1 \cdots b_n$, where $b_i = \Sigma(a_1 \cdots a_i)$ for all i. Of course, $\mathrm{psw}(w)$ is usually not a word over Γ, but over some other alphabet. The mapping psw is injective and length-preserving. We also use the notation $\mathrm{psw}_r(w) = c_1 \cdots c_n$, where $r \in \mathbb{R}$ and $c_i = b_i + r$ for all i.

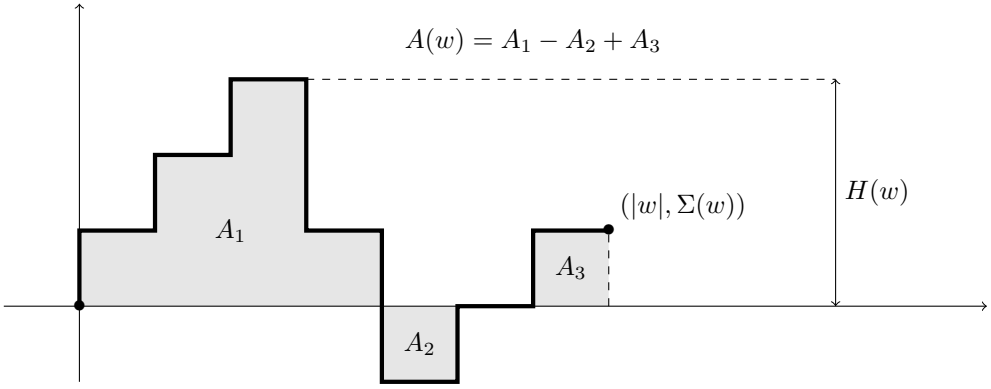

Figure 1 Representation of the word $w = aaabbaa$, where $a = 1$ and $b = -2$. We have $|w| = 7$, $\Sigma(w) = 1$, $H(w) = 3$, and $A(w) = 7$.

▶ **Example 9.** Let $w = bbcaac$, where $a = 1$, $b = 2$, and $c = -3$. We have $|w| = 6$, $\max(w) = 2$, and $\min(w) = -3$. Because $\Sigma(w) = 2 + 2 - 3 + 1 + 1 - 3 = 0$, w is a zero-sum word. The prefix sum word of w is $\mathrm{psw}(w) = 241230$, and $\max(\mathrm{psw}(w)) = 4$ and $\min(\mathrm{psw}(w)) = 0$.

For a word w, we define its *height* $H(w)$ and *area* $A(w)$:

$$H(w) = \max(\mathrm{psw}(w)) = \max\{\Sigma(u) \mid \varepsilon \neq u \sqsubseteq w\},$$
$$A(w) = \Sigma(\mathrm{psw}(w)) = \sum_{u \sqsubseteq w} \Sigma(u),$$

where $u \sqsubseteq w$ means that u is a prefix of w. For the empty word, $H(\varepsilon) = -\infty$ and $A(\varepsilon) = 0$.

These definitions have the following graphical interpretation: A word $w = a_1 \cdots a_n$ can be represented by a polygonal chain by starting at the origin, moving a_1 steps up, one step to the right, a_2 steps up, one step to the right, and so on. The end point of this curve is then $(|w|, \Sigma(w))$. The biggest y-coordinate (after the initial line segment starting at the origin) is $H(w)$. The number $A(w)$ is the area under the curve, defined in the same way as a definite integral, that is, parts below the x-axis count as negative areas. See Figure 1 for an example.

▶ **Lemma 10.** *For words $w_1, \ldots, w_n$, we have*

$$\Sigma(w_1 \cdots w_n) = \Sigma(w_1) + \cdots + \Sigma(w_n),$$
$$\mathrm{psw}(w_1 \cdots w_n) = \prod_{i=1}^{n} \mathrm{psw}_{\Sigma(w_1 \cdots w_{i-1})}(w_i),$$
$$H(w_1 \cdots w_n) = \max\{\Sigma(w_1 \cdots w_{i-1}) + H(w_i) \mid 1 \leq i \leq n\},$$
$$A(w_1 \cdots w_n) = \sum_{i=1}^{n} (A(w_i) + \Sigma(w_1 \cdots w_{i-1})|w_i|).$$

Proof. Follows easily from the definitions. ◀

When studying words from a combinatorial point of view, the choice of the alphabet is arbitrary (except for the size of the alphabet), so we can assign numerical values to the letters in any way we like, as long as no two letters get the same value. The next two lemmas show that, given any word w, the alphabet can be normalized so that w becomes a zero-sum word, and every zero-sum word can be written as a product of minimal zero-sum words in a unique way.

▶ **Lemma 11** ([18, Lemma 3]). *Let $w \in \Gamma^*$. There exists an alphabet Δ and an isomorphism $h : \Gamma^* \to \Delta^*$ such that $h(w)$ is zero-sum.*

▶ **Lemma 12** ([18, Lemma 4]). *The set of zero-sum words over Γ is a free monoid.*

4 Equations in normal form

If a one-variable equation has more occurrences of the variable on the left-hand side than on the right-hand side, or vice versa, then it is easy to see by a length argument that it can have at most one solution. Therefore every one-variable equation with more than one solution can be written in the form

$$(u_0Xu_1 \cdots Xu_n, v_0Xv_1 \cdots Xv_n), \tag{1}$$

where X is the variable, $n \geq 1$, and $u_0, \ldots, u_n, v_0, \ldots, v_n$ are constant words. Clearly, it must be $|u_0 \cdots u_n| = |v_0 \cdots v_n|$. If the equation is nontrivial, x_1, x_2 are solution words, and $|x_1| \leq |x_2|$, then it is quite easy to see that x_1 is a prefix and a suffix of x_2.

We say that the equation (1) is in *normal form* if the following conditions are satisfied:

(N1) It has the empty solution and at least one other zero-sum solution,
(N2) $|u_0 \cdots u_i| < |v_0 \cdots v_i|$ for all $i \in \{0, \ldots, n-1\}$,
(N3) $|u_0 \cdots u_i| \leq |v_0 \cdots v_{i-1}|$ for all $i \in \{0, \ldots, n\}$.

It follows from these conditions that $u_0 = v_n = \varepsilon$. By the next two lemmas, it is usually sufficient to consider equations in normal form.

▶ **Lemma 13.** *Let E be a one-variable equation,* $\mathrm{Sol}(E) = \{[x_0], \ldots, [x_m]\}$, *and $|x_0| \leq |x_i|$ for all i. There exists a one-variable equation E' such that* $\mathrm{Sol}(E') = \{[\varepsilon], [x_0^{-1}x_1], \ldots, [x_0^{-1}x_m]\}$.

Proof. If $m = 0$, the claim is clear. Otherwise, we can assume that E is of the form (1). Let E' be the equation we get from E by replacing X by x_0X:

$$E' : (u_0x_0Xu_1 \cdots x_0Xu_n, v_0x_0Xv_1 \cdots x_0Xv_n).$$

Because E is nontrivial, x_0 is a prefix of every x_i. Clearly, the word $x_0^{-1}x_i$ is a solution word of E'. On the other hand, if x is a solution word of E', then x_0x is a solution word of E. This proves the claim. ◀

Next we will give an example of how to transform an equation that satisfies Condition N1 into an equation in normal form.

▶ **Example 14.** Consider the equation

$$(XabXababXaabaXbX, abXXXababaXaXbab).$$

By a length argument, it is equivalent to the system of equations

$$(Xab, abX), (X, X), (ababX, Xabab), (a, a), (abaXbX, XaXbab).$$

We can drop the trivial equations (X, X) and (a, a), and then switch the left-hand and right-hand sides of the equations $(ababX, Xabab)$ and $(abaXbX, XaXbab)$ to get the system

$$(Xab, abX), (Xabab, ababX), (XaXbab, abaXbX).$$

Then we can combine these equations into the equation

$$(XabXababXaXbab, abXababXabaXbX),$$

which satisfies Conditions N2 and N3. (Actually, this equation is equivalent to the equation $(XaXbab, abaXbX)$.)

▶ **Lemma 15.** *Let E be a nontrivial one-variable equation with the empty solution and at least one other solution. There exists an equation in normal form that is equivalent to E up to a renaming of the letters and not longer than E.*

Proof. Omitted. ◀

5 Sums and heights of solutions

In this section, we prove lemmas about the sums and heights of solution words of one-variable equations in normal form.

▶ **Lemma 16.** *All solutions of an equation in normal form are zero-sum.*

Proof. Let the equation be (1). Let $u'_i = u_0 \cdots u_{i-1}$ and $v'_i = v_0 \cdots v_{i-1}$ for all i. After applying a solution $[x]$ on the left-hand side and taking the area we get

$$\begin{aligned}
&A(u_0 x u_1 \cdots x u_n)\\
=&\sum_{i=0}^{n}(A(u_i) + \Sigma(u_0 x u_1 \cdots u_{i-1}x)|u_i|) + \sum_{i=1}^{n}(A(x) + \Sigma(u_0 x u_1 \cdots x u_{i-1})|x|)\\
=&\sum_{i=0}^{n}(A(u_i) + \Sigma(u'_i)|u_i| + i\Sigma(x)|u_i|) + \sum_{i=1}^{n}(A(x) + \Sigma(u'_i)|x| + (i-1)\Sigma(x)|x|)\\
=&A(u_0 \cdots u_n) + \Sigma(x)\sum_{i=0}^{n} i|u_i| + nA(x) + |x|\sum_{i=1}^{n}\Sigma(u'_i) + \frac{(n-1)n}{2}\cdot\Sigma(x)|x|.
\end{aligned}$$

We get a similar formula for $A(v_0 x v_1 \cdots x v_n)$. Because $u_0 x u_1 \cdots x u_n = v_0 x v_1 \cdots x v_n$, we get

$$\begin{aligned}
0 &= A(u_0 x u_1 \cdots x u_n) - A(v_0 x v_1 \cdots x v_n)\\
&= A(u_0 \cdots u_n) - A(v_0 \cdots v_n) + \Sigma(x)\sum_{i=0}^{n} i(|u_i| - |v_i|) + |x|\sum_{i=1}^{n}(\Sigma(u'_i) - \Sigma(v'_i))\\
&= \Sigma(x)\sum_{i=0}^{n} i(|u_i| - |v_i|) + |x|\sum_{i=1}^{n}(\Sigma(u'_i) - \Sigma(v'_i)). \qquad (2)
\end{aligned}$$

By the definition of normal form, the equation has a nonempty zero-sum solution $[x_1]$. Replacing x by x_1 in (2) gives

$$0 = |x_1|\sum_{i=1}^{n}(\Sigma(u'_i) - \Sigma(v'_i)).$$

Because $|x_1| > 0$, $\sum_{i=1}^{n}(\Sigma(u'_i) - \Sigma(v'_i)) = 0$. Then (2) takes the form

$$0 = \Sigma(x)\sum_{i=0}^{n} i(|u_i| - |v_i|),$$

so either $\Sigma(x) = 0$ or $\sum_{i=0}^{n} i(|u_i| - |v_i|) = 0$. The latter is not possible, because

$$\begin{aligned}
&\sum_{i=0}^{n} i(|u_i| - |v_i|) = \sum_{i=1}^{n}(|u_i \cdots u_n| - |v_i \cdots v_n|)\\
=&\sum_{i=1}^{n}(|u_0 \cdots u_n| - |u'_i| - (|v_0 \cdots v_n| - |v'_i|)) = \sum_{i=1}^{n}(-|u'_i| + |v'_i|) > 0,
\end{aligned}$$

by Condition N2 in the definition of normal form. Thus every solution $[x]$ is zero-sum. ◀

▶ **Lemma 17.** *Consider the nontrivial equation* (1)*. Let* $s_i = \Sigma(u_0 \cdots u_{i-1})$ *and* $t_i = \Sigma(v_0 \cdots v_{i-1})$ *for all* i*. If the equation has at least two zero-sum solutions, then* $(s_1, \ldots, s_n)$ *is a permutation of* $(t_1, \ldots, t_n)$*.*

Proof. Omitted. ◀

▶ **Lemma 18.** *Let* (1) *be an equation in normal form. Let*

$$h = H(u_0 \cdots u_n) - \max\{\Sigma(u_0 \cdots u_i) \mid i \in \{0, \ldots, n-1\}\}. \tag{3}$$

If the equation has at least three nonempty solutions, then every nonempty solution is of height h*. If the equation has two nonempty solutions, then the shorter one is of height* h *and the longer one of height at least* h*.*

Proof. The idea of the proof is to look at the first occurrences of the highest points on the curves of the left-hand side and the right-hand side of the equation; these must match. If the length of the solution changes, these first occurrences often move with respect to each other so that they no longer match; this puts a limit on the number of solutions under certain conditions. A first occurrence can be either inside a constant part or inside a variable. We will see that if the first occurrences are inside constant parts on both sides, then the solution is empty, if they are inside variables on both sides, then the solution is of height at least h and there can be at most one solution of height more than h, and if the first occurrence is inside a constant part on one side and inside a variable on the other side, then the solution is of height h, and if there is a solution of height more than h, then there can be at most one solution of height h.

For any word w, let $\phi(w)$ be its shortest prefix such that $H(\phi(w)) = H(w)$. For any solution $[x]$, we have

$$\phi(u_0 x u_1 \cdots x u_n) = \phi(v_0 x v_1 \cdots x v_n). \tag{4}$$

Let $s_i = \Sigma(u_0 \cdots u_{i-1})$ and $t_i = \Sigma(v_0 \cdots v_{i-1})$ for all i. Let i and j be such that $\phi(u_0 \cdots u_n) = u_0 \cdots u_{i-1}\phi(u_i)$ and $\phi(v_0 \cdots v_n) = v_0 \cdots v_{j-1}\phi(v_j)$. Because $[\varepsilon]$ is a solution, $\phi(u_0 \cdots u_n) = \phi(v_0 \cdots v_n)$ and thus

$$|u_0 \cdots u_{i-1}| + |\phi(u_i)| = |v_0 \cdots v_{j-1}| + |\phi(v_j)|. \tag{5}$$

By (5) and Condition N3 in the definition of normal form, $i > j$.

Because $[\varepsilon]$ is a solution, $H(u_0 \cdots u_n) = H(v_0 \cdots v_n)$, and by Lemma 17,

$$\max\{\Sigma(u_0 \cdots u_i) \mid i \in \{0, \ldots, n-1\}\} = \max\{\Sigma(v_0 \cdots v_i) \mid i \in \{0, \ldots, n-1\}\},$$

so

$$h = H(v_0 \cdots v_n) - \max\{\Sigma(v_0 \cdots v_i) \mid i \in \{0, \ldots, n-1\}\}.$$

Let k and l be the smallest indices such that $s_k = \max\{s_1, \ldots, s_n\}$ and $t_l = \max\{t_1, \ldots, t_n\}$. Then

$$\phi(u_0 x u_1 \cdots x u_n) = \begin{cases} u_0 x u_1 \cdots u_{i-1} x \phi(u_i) & \text{if } H(x) < h \text{ or if } H(x) = h \text{ and } i < k, \\ u_0 x u_1 \cdots x u_{k-1} \phi(x) & \text{if } H(x) > h \text{ or if } H(x) = h \text{ and } i \geq k, \end{cases}$$

$$\phi(v_0 x v_1 \cdots x v_n) = \begin{cases} v_0 x v_1 \cdots v_{j-1} x \phi(v_j) & \text{if } H(x) < h \text{ or if } H(x) = h \text{ and } j < l, \\ v_0 x v_1 \cdots x v_{l-1} \phi(x) & \text{if } H(x) > h \text{ or if } H(x) = h \text{ and } j \geq l, \end{cases}$$

This means that, for a given x, (4) can take one of four possible forms:

(i) If $H(x) < h$ or if $H(x) = h$, $i < k$ and $j < l$, then

$$u_0xu_1\cdots u_{i-1}x\phi(u_i) = v_0xv_1\cdots v_{j-1}x\phi(v_j)$$

and thus

$$|u_0\cdots u_{i-1}| + |\phi(u_i)| + (i-j)|x| = |v_0\cdots v_{j-1}| + |\phi(v_j)|.$$

Because $i > j$, it follows that this equality can hold for at most one $|x|$, so there is only one possible x in this case, namely, the empty word.

(ii) If $H(x) = h$, $i < k$ and $j \geq l$, then

$$u_0xu_1\cdots u_{i-1}x\phi(u_i) = v_0xv_1\cdots xv_{l-1}\phi(x),$$

but

$$\begin{aligned}&|u_0xu_1\cdots u_{i-1}x\phi(u_i)| = |u_0\cdots u_{i-1}| + |\phi(u_i)| + i|x| = |v_0\cdots v_{j-1}| + |\phi(v_j)| + i|x|\\ &> |v_0\cdots v_{l-1}| + l|x| \geq |v_0xv_1\cdots xv_{l-1}\phi(x)|\end{aligned}$$

by (5) and $i > j \geq l$, a contradiction.

(iii) If $H(x) > h$ or if $H(x) = h$, $i \geq k$ and $j \geq l$, then

$$u_0xu_1\cdots xu_{k-1}\phi(x) = v_0xv_1\cdots xv_{l-1}\phi(x)$$

and thus

$$|u_0\cdots u_{k-1}| + (k-l)|x| = |v_0\cdots v_{l-1}|.$$

By Condition N2 in the definition of normal form, $k > l$. It follows that this equality can hold for at most one $|x|$, so there is only one possible x in this case.

(iv) If $H(x) = h$, $i \geq k$ and $j < l$, then

$$u_0xu_1\cdots xu_{k-1}\phi(x) = v_0xv_1\cdots v_{j-1}x\phi(v_j)$$

and thus

$$|u_0\cdots u_{k-1}| + |\phi(x)| + (k-1-j)|x| = |v_0\cdots v_{j-1}| + |\phi(v_j)|. \tag{6}$$

If x and x' are solution words, then one of them is a prefix of the other, so if they have the same height, then $\phi(x) = \phi(x')$. Therefore, (6) can hold for more than one solution word x of height h only if $k-1-j = 0$. In general, this can happen (for example, if the equation has infinitely many solutions). However, if there exists a solution word of height more than h, then it follows from Case (iii) that $k > l$. Then $j < l < k$, so $k-1 > j$ and there is at most one solution word x of height h. ◀

▶ **Example 19.** Consider the equation

$$(XaXbXaabbabaXbabaabbab, abaabbabaXbabaabbXaXbX)$$

that was mentioned in Example 5. Let $a = 1$ and $b = -1$. The equation has exactly three solutions $[\varepsilon], [ab], [abaabbab]$. All of them are zero-sum, and their heights are $-\infty$, 1, 2, respectively. If we use the notation of the proof of Lemma 18, then $i = 3$, $j = 0$, $k = 2$, $l = 1$, and $h = 1$. We have $\phi(u_i) = \phi(aabbaba) = aa$, $\phi(v_j) = \phi(abaabbaba) = abaa$, $\phi(ab) = a$, and $\phi(abaabbab) = abaa$. Then

$$\phi(xaxbxaabbabaxbabaabbab) = \begin{cases} xaxbxaa & \text{if } x = \varepsilon,\\ xa\phi(x) & \text{if } x = abaabbab \text{ or if } x = ab,\end{cases}$$

$$\phi(abaabbabaxbabaabbxaxbx) = \begin{cases} abaa & \text{if } x = \varepsilon \text{ or if } x = ab,\\ abaabbaba\phi(x) & \text{if } x = abaabbab.\end{cases}$$

6 Some Lemmas

In this section, we state many lemmas about one-variable equations that will be used in the proof of the main result. The proofs are omitted.

A subset Z of Γ^* is called a *code* if the elements of Z do not satisfy any nontrivial relations. In other words, Z is a code if and only if for all $x_1, \ldots, x_m, y_1, \ldots, y_n \in Z$, $x_1 \cdots x_m = y_1 \cdots y_n$ implies $m = n$ and $x_i = y_i$ for all $i \in \{1, \ldots, m\}$. If Z is a code, then Z^* is a free monoid, and if Δ is an alphabet of the same size as Z, then the free monoids Z^* and Δ^* are isomorphic. More information about codes can be found in the book of Berstel, Perrin and Reutenauer [2].

The next lemma can be used to compress an equation into a shorter one. We will use it with two codes Z: The set of all minimal zero-sum words (those zero-sum words which cannot be written as a product of two shorter zero-sum words), and the set of words of a specific length.

▶ **Lemma 20.** *Let E be the equation (1) and let Z be a code. If $u_i, v_i \in Z^*$ for all i, then there exists an alphabet Δ and an isomorphism $h : Z^* \to \Delta^*$, and the equation*

$$(h(u_0)Xh(u_1)\cdots Xh(u_n), h(v_0)Xh(v_1)\cdots Xh(v_n)) \tag{7}$$

has the solution set $\{[h(x)] \mid [x] \in \mathrm{Sol}(E),\ x \in Z^*\}$.

Note that the equation E in Lemma 20 can have solution words that are not in Z^*, so (7) can have less solutions than E.

The next lemma can be used to cut off part of an equation so that all solutions are preserved, except possibly the empty solution (and maybe some additional solutions are added).

▶ **Lemma 21.** *Consider the equation (1). Let $k \in \{0, \ldots, n\}$ and let*

$$d = |v_0 \cdots v_{k-1}| - |u_0 \cdots u_k| \geq 0.$$

If all nonempty solutions of the equation are of length at least d, and if y is the common prefix of length d of all nonempty solution words, then each one of the nonempty solutions is a solution of the equation

$$(u_0 X u_1 \cdots X u_k y, v_0 X v_1 \cdots v_{k-1} X). \tag{8}$$

Using Lemma 21 requires the existence of a suitable index k. The next two lemmas can sometimes be used to find such an index. The proof of Lemma 22 is somewhat similar to the proof of Lemma 18, but simpler.

▶ **Lemma 22.** *Let (1) be an equation in normal form. If it has at least three nonempty solutions, and if there exists $k \in \{1, \ldots, n-1\}$ such that*

$$\Sigma(u_0) = \cdots = \Sigma(u_{k-1}) = 0 \neq \Sigma(u_k),$$

then every nonempty solution is of length more than $|v_0 \cdots v_{k-1}| - |u_0 \cdots u_k|$.

▶ **Lemma 23.** *Let the equation (1) have the solution set $[p^*]$ for some primitive word p. Let $u_0 = v_n = \varepsilon$. Let $j \in \{0, \ldots, n\}$ be the largest index such that the lengths of $u_0, \ldots, u_{j-1}$ and $v_0, \ldots, v_{j-1}$ are divisible by $|p|$. Then $j > 0$ and $|v_0 \cdots v_{j-1}| - |u_0 \cdots u_j| \leq |p|$.*

Lemma 21 does not guarantee that the new, shorter equation would have the empty solution. Sometimes the next lemma can be used to get around this problem.

▶ **Lemma 24.** *If the equation* (1) *has a nonempty solution,* $u_n = ua^m$ *for some* $u \in \Gamma^*$, $a \in \Gamma$ *and* $m \geq 0$, *and* $u_0 \cdots u_{n-1}u$ *is a prefix of* $v_0 \cdots v_n$, *then the equation has the empty solution.*

7 Main results

Now we are ready to prove our main results.

▶ **Theorem 25.** *If a one-variable equation has only finitely many solutions, it has at most three solutions.*

Proof. Assume that there is a counterexample. Then there is one with an empty solution by Lemma 13. Of all equations with the empty solution, at least three nonempty solutions, and only finitely many solutions, let E_1 be a shortest one. We are going to prove a contradiction by showing that there exists a shorter equation with these properties. By Lemma 15, we can assume that E_1 is the equation (1) and it is in normal form. By Lemma 16, each one of its solutions is zero-sum.

The idea of the proof is to cut off part of the equation to get a shorter equation E_2 that has at least three nonempty solutions but only finitely many. Unfortunately, E_2 does not necessarily have the empty solution. We map E_2 with a length-preserving mapping to get an equation E_3 that has at least three nonempty solution and also the empty solution. Unfortunately, E_3 might have infinitely many solutions. We analyze E_3 to find another way to cut off part of E_1 to get an equation E_4, which is then modified to an equation E_5. For E_5, we can finally prove that it has the empty solution and at least three but only finitely many nonempty solutions.

If $\Sigma(u_i) = 0$ for all $i < n$, then $\Sigma(v_i) = 0$ for all $i < n$ by Lemma 17, and then also $\Sigma(u_n) = 0$, because $\Sigma(u_0 \cdots u_n) = \Sigma(v_0 \cdots v_n)$ and $v_n = \varepsilon$. Thus all u_i, v_i are zero-sum, and we can use Lemma 20 with Z the set of all minimal zero-sum words to get a shorter equation with the same number of solutions, one of them empty.

For the rest of the proof, we assume that there exists a minimal $k < n$ such that $\Sigma(u_k) \neq 0$. By symmetry, we can assume that $\Sigma(u_k) > 0$. By Lemmas 22 and 21, we get a shorter equation

$$E_2 : (u_0 X u_1 \cdots X u_k y, v_0 X v_1 \cdots v_{k-1} X)$$

that has at least all the same nonempty solutions as E_1. It might have some other solutions as well, but it cannot have infinitely many solutions, because the intersection of an infinite solution set of a nontrivial one-variable equation and a finite solution set of a one-variable equation is of size at most two by Theorem 6 and Lemma 7. If it has also the empty solution, then we are done, but we do not know yet whether this is the case. We can use Lemma 17 for E_2 to see that $(\Sigma(u_0), \ldots, \Sigma(u_0 \cdots u_{k-1}))$ and $(\Sigma(v_0), \ldots, \Sigma(v_0 \cdots v_{k-1}))$ are permutations of each other. We know that $u_0, \ldots, u_{k-1}$ are zero-sum, so also $v_0, \ldots, v_{k-1}$ are zero-sum.

Let $[x_1]$ be the shortest nonempty solution of E_1. Let $\{a, b\}$ be an alphabet and let g be the morphism that maps the letter $\min(\mathrm{psw}(x_1))$ to b and every other letter to a. Let $f = g \circ \mathrm{psw}$. Then f is length-preserving, and if w is zero-sum, then $f(ww') = f(w)f(w')$. If $[x]$ is a nonempty solution of E_1, then $[f(x)]$ is a solution of the equation

$$E_3 : (f(u_0) X f(u_1) \cdots X f(u_k y), f(v_0) X f(v_1) \cdots f(v_{k-1}) X).$$

We have $f(u_k y) = f(u_k)g(\mathrm{psw}_{\Sigma(u_k)}(y))$. Because $\Sigma(u_k) > 0$ and y is a prefix of x_1, $\min(\mathrm{psw}_{\Sigma(u_k)}(y)) > \min(\mathrm{psw}(x_1))$. Thus $g(\mathrm{psw}_{\Sigma(u_k)}(y)) \in a^*$. Because $u_0 \cdots u_k$ is a prefix of $v_0 \cdots v_{k-1}$, also $f(u_0 \cdots u_k) = f(u_0) \cdots f(u_k)$ is a prefix of $f(v_0 \cdots v_{k-1}) = f(v_0) \cdots f(v_{k-1})$. We can use Lemma 24 with $g(\mathrm{psw}_{\Sigma(u_k)}(y))$ as a^m, so E_3 has the empty solution. If it has only finitely many solutions, then we are done. For the rest of the proof, we assume that it has infinitely many solutions. Then its solution set is $[p^*]$ for some primitive word p. Consequently, the length of every solution word of E_1 is divisible by $|p|$. Because the solution word $f(x_1)$ of E_3 contains the letter b, also p must contain b. This means that p cannot be a suffix of $g(\mathrm{psw}_{\Sigma(u_k)}(y)) \in a^*$, so $|p| > |y|$.

We can use Lemma 23 for E_3 to find an index j such that the lengths of $u_0, \ldots, u_{j-1}$ and $v_0, \ldots, v_{j-1}$ are divisible by $|p|$ and, if $j < k$, $|v_0 \cdots v_{j-1}| - |u_0 \cdots u_j| \leq |p|$ (remember that f is length-preserving). By letting $z = y$ if $j = k$, or by using Lemma 21 with j as k for E_1 otherwise, we get an equation

$$E_4 : (u_0 X u_1 \cdots X u_j z, v_0 X v_1 \cdots v_{j-1} X)$$

that has at least all the same nonempty solutions as E_1. In both cases, $|z| \leq |p|$. Like in the case of E_2, we see that E_4 cannot have infinitely many solutions. The lengths of all the constant words in E_4 are divisible by $|p|$, and so are the lengths of at least three nonempty solutions (the solutions of E_1). We can use Lemma 20 with $Z = \Gamma^{|p|}$ for E_4. If h is the morphism of Lemma 20, then we get the equation

$$E_5 : (h(u_0) X h(u_1) \cdots X h(u_j z), h(v_0) X h(v_1) \cdots h(v_{j-1}) X).$$

It has at least three nonempty solutions, but only finitely many. Because $|z| \leq |p|$, $h(u_j z) = h(u)c$, where u is a prefix of u_j and c is a letter. Because $u_0 \cdots u_j$ is a prefix of $v_0 \cdots v_{j-1}$, also $h(u_0 \cdots u_{j-1} u) = h(u_0) \cdots h(u_{j-1}) h(u)$ is a prefix of $h(v_0 \cdots v_{k-1}) = h(v_0) \cdots h(v_{k-1})$. We can use Lemma 24 with c as a and $m = 1$, so E_5 has the empty solution. This contradicts the minimality of E_1. ◀

▶ **Theorem 26.** *If a system of constant-free three-variable equations is independent and has a nonperiodic solution, then it has at most 17 equations.*

Proof. Follows from Theorem 25 and Theorem 8. ◀

8 Conclusion

We have proved that the maximal size of a finite solution set of a one-variable word equation is three, and that the maximal size of an independent system of constant-free three-variable equations with a nonperiodic solution is somewhere between two and 17.

Improving the bound 17 is an obvious open problem. A possible approach would be to improve the results in [16].

Another open problem is proving similar bounds for more than three variables. The result in [16] is based on a characterization of three-generator subsemigroups of a free semigroup by Budkina and Markov [3], or alternatively a similar result by Spehner [19, 20]. This means that it is very specific to the three-variable case, and analyzing the general case would require an entirely different approach.

Finally, characterizing possible solution sets of one-variable equations would be interesting. The possible infinite solution sets are given by Theorem 6, and every singleton set is possible, but for sets of size two or three the question is open.

References

1 M. H. Albert and J. Lawrence. A proof of Ehrenfeucht's conjecture. *Theoret. Comput. Sci.*, 41(1):121–123, 1985. doi:10.1016/0304-3975(85)90066-0.

2 Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2010.

3 L. G. Budkina and Al. A. Markov. F-semigroups with three generators. *Mat. Zametki*, 14:267–277, 1973.

4 Karel Culik, II and Juhani Karhumäki. Systems of equations over a free monoid and Ehrenfeucht's conjecture. *Discrete Math.*, 43(2–3):139–153, 1983. doi:10.1016/0012-365X(83)90152-8.

5 Elena Czeizler and Juhani Karhumäki. On non-periodic solutions of independent systems of word equations over three unknowns. *Internat. J. Found. Comput. Sci.*, 18(4):873–897, 2007. doi:10.1142/S0129054107005030.

6 Elena Czeizler and Wojciech Plandowski. On systems of word equations over three unknowns with at most six occurrences of one of the unknowns. *Theoret. Comput. Sci.*, 410(30–32):2889–2909, 2009. doi:10.1016/j.tcs.2009.01.023.

7 Robert Dąbrowski and Wojciech Plandowski. On word equations in one variable. *Algorithmica*, 60(4):819–828, 2011. doi:10.1007/s00453-009-9375-3.

8 S. Eyono Obono, P. Goralčík, and M. Maksimenko. Efficient solving of the word equations in one variable. In *Proceedings of the 19th MFCS*, volume 841 of *LNCS*, pages 336–341. Springer, 1994. doi:10.1007/3-540-58338-6_80.

9 V. S. Guba. Equivalence of infinite systems of equations in free groups and semigroups to finite subsystems. *Mat. Zametki*, 40(3):321–324, 1986. doi:10.1007/BF01142470.

10 Tero Harju and Dirk Nowotka. On the independence of equations in three variables. *Theoret. Comput. Sci.*, 307(1):139–172, 2003. doi:10.1016/S0304-3975(03)00098-7.

11 Štěpán Holub and Jan Žemlička. Algebraic properties of word equations. *J. Algebra*, 434:283–301, 2015. doi:10.1016/j.jalgebra.2015.03.021.

12 Artur Jeż. One-variable word equations in linear time. *Algorithmica*, 74(1):1–48, 2016. doi:10.1007/s00453-014-9931-3.

13 Juhani Karhumäki and Wojciech Plandowski. On the defect effect of many identities in free semigroups. In Gheorghe Paun, editor, *Mathematical aspects of natural and formal languages*, pages 225–232. World Scientific, 1994.

14 Juhani Karhumäki and Aleksi Saarela. On maximal chains of systems of word equations. *Proc. Steklov Inst. Math.*, 274:116–123, 2011. doi:10.1134/S0081543811060083.

15 Markku Laine and Wojciech Plandowski. Word equations with one unknown. *Internat. J. Found. Comput. Sci.*, 22(2):345–375, 2011. doi:10.1142/S0129054111008088.

16 Dirk Nowotka and Aleksi Saarela. One-variable word equations and three-variable constant-free word equations. *Internat. J. Found. Comput. Sci.*, To appear.

17 Aleksi Saarela. Systems of word equations, polynomials and linear algebra: A new approach. *European J. Combin.*, 47:1–14, 2015. doi:10.1016/j.ejc.2015.01.005.

18 Aleksi Saarela. Word equations where a power equals a product of powers. In *Proceedings of the 34th STACS*, volume 66 of *LIPIcs*, pages 55:1–55:9. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. doi:10.4230/LIPIcs.STACS.2017.55.

19 Jean-Claude Spehner. *Quelques problémes d'extension, de conjugaison et de présentation des sous-monoïdes d'un monoïde libre*. PhD thesis, Univ. Paris, 1976.

20 Jean-Claude Spehner. Les systemes entiers d'équations sur un alphabet de 3 variables. In *Semigroups*, pages 342–357, 1986.