AN OPTIMAL BOUND ON THE SOLUTION SETS OF ONE-VARIABLE WORD EQUATIONS AND ITS CONSEQUENCES*

DIRK NOWOTKA† AND ALEKSI SAARELA‡

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 has at most 17 equations. Although probably not optimal, this is the first finite bound found. However, the conjecture of that bound being actually two still remains open.

Key words. combinatorics on words, word equations, systems of equations

AMS subject classifications. 68R15

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. This gives an upper bound n for the size of independent systems of linear equations. 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 II 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 Karhumäki and Plandowski [15], and the hidden constant in $\Theta(n^4)$ was improved in [16]. 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 II 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

*An extended version of an article published in the proceedings of the 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018).

Funding: This work was partially supported by the DFG research project 181615770.

†Department of Computer Science, Kiel University, 24098 Kiel, Germany (dn@informatik.uni-kiel.de).

‡Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland (amsaar@utu.fi).

systems of specific forms are known [10, 5, 6], and some upper bounds that depend on the sizes of the equations have been proved [20, 11, 18]. The best current bound is logarithmic with respect to the size of the smallest equation in the system [18].

Moving from constant-free equations to the more general family of word equations with constants. The algorithmic problem of determining whether a given word equation has a solution, also known as the satisfiability problem, can be easily proved to be NP-hard. Proving upper bounds for the complexity is much more challenging. The best known result is that the problem is in NSPACE(n), as proved by Jeż [14]. For constant-free equations, there are variations of the satisfiability problem that are as difficult as the general satisfiability problem [23]. However, if the number of variables is small, equations with constants can be more complex than constant-free equations. In particular, 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, 17], 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 [17]. An example with exactly three solutions was recently found [18]. 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 [17]. It can be noted that the solutions of a one-variable equation can be found in linear time in the RAM model [13].

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 optimal. 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. In addition to independent systems, we also study decreasing chains of equations and prove a similar result for them.

Two previous articles provide crucial tools for our proofs. The first article is [22] (or its earlier conference version [21]), 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 [18], 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.

This article is an extended version of the conference article [19]. One difference between the two versions is that many of the proofs, for example those in Section 6, were omitted from the conference version to save space. Another difference is that in this article we consider decreasing chains in addition to independent systems.

2. Preliminaries. We begin this section by giving some standard definitions. A word x is a *prefix* of a word w if $w = xy$ for some word y. Similarly, a word x is

a *suffix* of a word w if $w = yx$ for some word y. The empty word is denoted by ε. The length of a word w is denoted by $|w|$ and the number of occurrences of a letter a in w is denoted by $|w|_a$. A nonempty word is *primitive* if it is not a power of a shorter word. If Γ_1 and Γ_2 are alphabets, then a mapping $h : \Gamma_1^* \to \Gamma_2^*$ is a *morphism* if $h(xy) = h(x)h(y)$ for all $x, y \in \Gamma_1^*$.

Next, we consider 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 $(U, V) \in \Xi^* \times \Xi^*$. A *solution* of this equation is a morphism $h : \Xi^* \to \Gamma^*$ 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 2.1. 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 2.2. 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 renowned 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 2.2, and it has been conjectured that these examples are optimal. Even the following weaker conjecture is open.

CONJECTURE 2.3. *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 2.4 ([18]). *Every independent system of constant-free three-variable equations is of size $O(\log n)$, where n is the length of the shortest equation.*

A sequence of equations $E_1, \ldots, E_N$ is a *decreasing chain* if the systems of equations $E_1, \ldots, E_{i-1}$ and $E_1, \ldots, E_i$ are nonequivalent for all $i \in \{1, \ldots, N\}$ (the case $i = 1$ means that E_1 cannot be equivalent to the empty system, that is, E_1 cannot be a trivial equation (U, U)). A morphism is a solution of a decreasing chain if it is a solution of every equation in the chain. The equations of an independent system, ordered in any way, form a decreasing chain. Decreasing chains have been studied in, for example, [12], [16] and [18].

If a decreasing chain of constant-free three-variable equations has a nonperiodic solution, then how long can the chain be? The largest known examples are of length

four, see Example 2.7. We can state a conjecture and a theorem that are analogous to Conjecture 2.3 and Theorem 2.4.

CONJECTURE 2.5. *There exists a number c such that every decreasing chain of constant-free three-variable equations with a nonperiodic solution is of length c or less.*

THEOREM 2.6 ([18]). *Every decreasing chain of constant-free three-variable equations is of length $O(\log n)$, where n is the length of the first equation.*

Example 2.7. Let $\Xi = \{X, Y, Z\}$ and $\Gamma = \{a, b\}$. The sequence

$$(XYZ, ZXY), (XYXZYZ, ZXZYXY), (XZ, ZX), (Z, \varepsilon)$$

is a decreasing chain and has a nonperiodic solution. This can be seen as follows:

1. The morphism h defined by $h(X) = a$, $h(Y) = \varepsilon$, $h(Z) = b$ is not a solution of (XYZ, ZXY).
2. The morphism h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = abab$ is a solution of (XYZ, ZXY) but not of $(XYXZYZ, ZXZYXY)$.
3. The morphism h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = ab$ is a solution of (XYZ, ZXY) and $(XYXZYZ, ZXZYXY)$ but not of (XZ, ZX).
4. The morphism h defined by $h(X) = \varepsilon$, $h(Y) = \varepsilon$, $h(Z) = a$ is a solution of (XYZ, ZXY), $(XYXZYZ, ZXZYXY)$ and (XZ, ZX) but not of (Z, ε).
5. The morphism h defined by $h(X) = a$, $h(Y) = b$, $h(Z) = \varepsilon$ is nonperiodic and a solution of all the equations.

Next, we 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 $(U, V) \in (\Xi \cup \Gamma)^* \times (\Xi \cup \Gamma)^*$. A *solution* of this equation is a 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 2.8. 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 2.8 shows that the answer is at least three, but no upper bound is known. Currently, the best known result is the following.

THEOREM 2.9 ([17, 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, or rather its corollary.

LEMMA 2.10 ([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.

COROLLARY 2.11. *Let E be a one-variable equation with only finitely many solutions and let pq be primitive. The set*

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

has at most two elements.

A connection between constant-free three-variable equations and one-variable equations with constants was recently found [18]. Here we give the relevant special cases of one of the results.

THEOREM 2.12 ([18]). *If every one-variable word equation has either infinitely many solutions or at most three, then Conjecture 2.3 is true for $c = 17$.*

THEOREM 2.13 ([18]). *If every one-variable word equation has either infinitely many solutions or at most three, then Conjecture 2.5 is true for $c = 20$.*

In this article, we will prove that every one-variable word equation has either infinitely many solutions or at most three, and thus Conjecture 2.3 is true for $c = 17$ and Conjecture 2.5 is true for $c = 20$.

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 [21].

We can assume that the alphabet Γ is a subset of $\mathbb{R}$. Then we 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 $[w]$ is also called zero-sum. The largest and smallest letters in a word w are 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.

Example 3.1. 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 3.1 for an example.

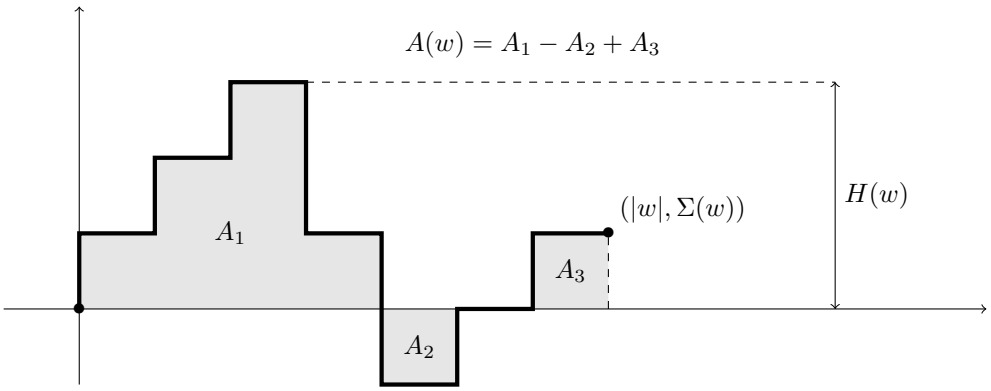

FIG. 3.1. *Representation of the word $w = aaabbaa$, where $a = 1$ and $b = -2$. We have $|w| = 7$, $\Sigma(w) = 1$, $H(w) = 3$, and $A(w) = 7$.*

LEMMA 3.2. *For words $w_1, \ldots, w_n$, we have*

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

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 lemma shows that, given any word w, the alphabet can be normalized so that w becomes a zero-sum word.

LEMMA 3.3 ([21, 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.*

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

$$(u_0 X u_1 \cdots X u_n, v_0 X v_1 \cdots X v_n), \tag{4.1}$$

where X is the variable, $n \ge 1$, 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 (4.1) 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 4.1. *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 (4.1). 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. After the example, we will prove that this can always be done. The example is preceded by a lemma that is often useful when working with word equations.

LEMMA 4.2. *Let (UU', VV') be a word equation. Assume that $|U| = |V|$ and $|U|_X = |V|_X$ for every variable X. Then the equation (UU', VV') is equivalent to the system $(U, V), (U', V')$.*

Proof. It is clear that every solution of the system is also a solution of the equation (UU', VV'). If h is a solution of (UU', VV'), then $h(U)$ and $h(V)$ are both prefixes of $h(UU') = h(VV')$, and by the assumption about U and V, they have the same length. Thus they are equal, and then also $h(U') = h(V')$, so h is a solution of the system $(U, V), (U', V')$. This shows the equivalence. ∎

Example 4.3. Consider the equation

$$(XabXababXaabaXbX, abXXXababaXaXbab).$$

By using Lemma 4.2 repeatedly, we see that the equation 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 4.4. *Let E be a nontrivial one-variable equation with the empty solution and at least one other solution. There exists an equation not longer than E and in normal form that is equivalent to E up to a renaming of the letters.*

Proof. We can assume that E has a nonempty zero-sum solution by Lemma 3.3. We can also assume that E is a shortest equation among all the equivalent equations, and E is written as (4.1). Finally, we can let $j \in \{0, \dots, n\}$ be the smallest index such that $|u_0 \cdots u_j| \geq |v_0 \cdots v_j|$ (the inequality holds for $j = n$, so j exists), and assume that there does not exists an equivalent equally long equation for which the index j would be larger.

We are going to prove that E is in normal form. We already know that Condition N1 holds.

If it were $j < n$ and $|u_0 \cdots u_j| = |v_0 \cdots v_j|$, then for any word x we would have the sequence of equivalences

$$\begin{aligned} & u_0xu_1 \cdots xu_n = v_0xv_1 \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_j = v_0xv_1 \cdots xv_j \wedge u_{j+1}xu_{j+2} \cdots xu_n = v_{j+1}xv_{j+2} \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_ju_{j+1}xu_{j+2} \cdots xu_n = v_0xv_1 \cdots xv_jv_{j+1}xv_{j+2} \cdots xv_n. \end{aligned}$$

(here we have essentially used Lemma 4.2). Thus E would be equivalent to the shorter equation

$$(u_0Xu_1 \cdots Xu_ju_{j+1}Xu_{j+2} \cdots Xu_n, v_0Xv_1 \cdots Xv_jv_{j+1}Xv_{j+2} \cdots Xv_n),$$

which would contradict the minimality of E. On the other hand, if it were $j < n$ and $|u_0 \cdots u_j| > |v_0 \cdots v_j|$, then there would exist words p, q such that $u_j = pq$ and $|u_0 \cdots u_{j-1}p| = |v_0 \cdots v_j|$, and for any word x we would have the sequence of equivalences

$$\begin{aligned} & u_0xu_1 \cdots xu_n = v_0xv_1 \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_{j-1}xp = v_0xv_1 \cdots xv_j \wedge qxu_{j+1} \cdots xu_n = xv_{j+1} \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_{j-1}xpxv_{j+1} \cdots xv_n = v_0xv_1 \cdots xv_jqxu_{j+1} \cdots xu_n, \end{aligned}$$

so E would be equivalent to the equation

$$((u_0Xu_1 \cdots Xu_{j-1}XpXv_{j+1} \cdots Xv_n, v_0Xv_1 \cdots Xv_jqXu_{j+1} \cdots Xu_n).$$

But this contradicts the assumption that there does not exists an equivalent equally long equation for which the index j would be larger. We conclude that the only possibility is that $j = n$, so Condition N2 holds.

If there were an index $i \in \{0, \dots, n\}$ such that $|u_0 \cdots u_i| > |v_0 \cdots v_{i-1}|$, then there would exist words p, q, r such that $u_i = pq$, $v_i = qr$, and $|u_0 \cdots u_{i-1}p| = |v_0 \cdots v_{i-1}|$, and for any word x we would have the sequence of equivalences

$$\begin{aligned} & u_0xu_1 \cdots xu_n = v_0xv_1 \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_{i-1}xp = v_0xv_1 \cdots xv_{i-1}x \wedge xu_{i+1} \cdots xu_n = rxv_{i+1} \cdots xv_n \\ \iff & u_0xu_1 \cdots xu_{i-1}xpxu_{i+1} \cdots xu_n = v_0xv_1 \cdots xv_{i-1}xrxv_{i+1} \cdots xv_n, \end{aligned}$$

so E would be equivalent to the shorter equation

$$(u_0Xu_1\cdots Xu_{i-1}XpXu_{i+1}\cdots Xu_n = v_0Xv_1\cdots Xv_{i-1}XrXv_{i+1}\cdots Xv_n),$$

which would contradict the minimality of E. This shows that also Condition N3 holds, so E is in normal form. ∎

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 5.1. *All solutions of an equation in the normal form are zero-sum.*

Proof. Let the equation be (4.1). 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

$$\begin{aligned}
&A(u_0xu_1\cdots xu_n)\\
=&\sum_{i=0}^{n}(A(u_i)+\Sigma(u_0xu_1\cdots u_{i-1}x)|u_i|)+\sum_{i=1}^{n}(A(x)+\Sigma(u_0xu_1\cdots xu_{i-1})|x|)\\
=&\sum_{i=0}^{n}(A(u_i)+\Sigma(u'_i)|u_i|+i\Sigma(x)|u_i|)+\sum_{i=1}^{n}(A(x)+\Sigma(u'_i)|x|+(i-1)\Sigma(x)|x|)\\
=&A(u_0\cdots u_n)+\Sigma(x)\sum_{i=0}^{n}i|u_i|+nA(x)+|x|\sum_{i=1}^{n}\Sigma(u'_i)+\frac{(n-1)n}{2}\cdot\Sigma(x)|x|.
\end{aligned}$$

We get a similar formula for $A(v_0xv_1\cdots xv_n)$. Because $u_0xu_1\cdots xu_n = v_0xv_1\cdots xv_n$, we get

(5.1)
$$\begin{aligned}
0 &= A(u_0xu_1\cdots xu_n)-A(v_0xv_1\cdots xv_n)\\
&= A(u_0\cdots u_n)-A(v_0\cdots v_n)+\Sigma(x)\sum_{i=0}^{n}i(|u_i|-|v_i|)+|x|\sum_{i=1}^{n}(\Sigma(u'_i)-\Sigma(v'_i))\\
&= \Sigma(x)\sum_{i=0}^{n}i(|u_i|-|v_i|)+|x|\sum_{i=1}^{n}(\Sigma(u'_i)-\Sigma(v'_i)).
\end{aligned}$$

By the definition of normal form, the equation has a nonempty zero-sum solution $[x_1]$. Replacing x by x_1 in (5.1) gives

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

Because $|x_1|>0$, we have $\sum_{i=1}^{n}(\Sigma(u'_i)-\Sigma(v'_i))=0$. Then (5.1) 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 5.2. *Consider a nontrivial equation (4.1). 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. Let $[x]$ and $[y]$ be two zero-sum solutions and let $|x| > |y|$. Because y is a prefix and a suffix of x, also $\mathrm{psw}_r(y)$ is a prefix and a suffix of $\mathrm{psw}_r(x)$ for every r. If a is any letter that appears in $\mathrm{psw}_r(y)$, and if its last occurrence in $\mathrm{psw}_r(y)$ is at position k, then its last occurrence in $\mathrm{psw}_r(x)$ is at a later position $k + |x| - |y|$. Consequently, every letter that appears in $\mathrm{psw}_r(y)$ appears more often in $\mathrm{psw}_r(x)$. Let $(s'_1, \ldots, s'_n)$ be the permutation of $(s_1, \ldots, s_n)$ such that $s'_i \le s'_{i+1}$ for all i, and let $(t'_1, \ldots, t'_n)$ be the permutation of $(t_1, \ldots, t_n)$ such that $t'_i \le t'_{i+1}$ for all i. Let j be the largest index such that $s'_j \neq t'_j$ (if there is no such index, then we have proved the lemma). Without loss of generality, let $s'_j > t'_j$. Let $a = H(x) + s'_j$. Then

$$
\begin{aligned}
0 =& |\mathrm{psw}(u_0 x u_1 \cdots x u_n)|_a - |\mathrm{psw}(v_0 x v_1 \cdots x v_n)|_a \\
& - |\mathrm{psw}(u_0 y u_1 \cdots y u_n)|_a + |\mathrm{psw}(v_0 y v_1 \cdots y v_n)|_a && (5.2)\\
=& \mathrm{psw}(u_0) + \sum_{i=1}^{n} (|\mathrm{psw}_{s_i}(x)|_a + |\mathrm{psw}_{s_i}(u_i)|_a) \\
& - \mathrm{psw}(v_0) - \sum_{i=1}^{n} (|\mathrm{psw}_{t_i}(x)|_a + |\mathrm{psw}_{t_i}(v_i)|_a) \\
& - \mathrm{psw}(u_0) - \sum_{i=1}^{n} (|\mathrm{psw}_{s_i}(y)|_a + |\mathrm{psw}_{s_i}(u_i)|_a) \\
& + \mathrm{psw}(v_0) + \sum_{i=1}^{n} (|\mathrm{psw}_{t_i}(y)|_a + |\mathrm{psw}_{t_i}(v_i)|_a) && (5.3)\\
=& \sum_{i=1}^{n} (|\mathrm{psw}_{s_i}(x)|_a - |\mathrm{psw}_{t_i}(x)|_a - |\mathrm{psw}_{s_i}(y)|_a + |\mathrm{psw}_{t_i}(y)|_a) && (5.4)\\
=& \sum_{i=1}^{n} (|\mathrm{psw}_{s'_i}(x)|_a - |\mathrm{psw}_{t'_i}(x)|_a - |\mathrm{psw}_{s'_i}(y)|_a + |\mathrm{psw}_{t'_i}(y)|_a) && (5.5)\\
=& \sum_{i=1}^{j} (|\mathrm{psw}_{s'_i}(x)|_a - |\mathrm{psw}_{t'_i}(x)|_a - |\mathrm{psw}_{s'_i}(y)|_a + |\mathrm{psw}_{t'_i}(y)|_a) && (5.6)\\
=& \sum_{i=1}^{j} (|\mathrm{psw}_{s'_i}(x)|_a - |\mathrm{psw}_{s'_i}(y)|_a) && (5.7)\\
\ge& |\mathrm{psw}_{s'_j}(x)|_a - |\mathrm{psw}_{s'_j}(y)|_a > 0, && (5.8)
\end{aligned}
$$

a contradiction. Here, (5.2) follows from x and y being solution words, (5.3) from them being zero-sum, (5.4) from cancelling the terms related to the u_i and v_i words, (5.5) from permuting the order of the terms in the sums, (5.6) from the definition of j, (5.7) from $a > H(x) + t'_j \ge H(x) + t'_i \ge H(y) + t'_i$ for all $i \in \{1, \ldots, j\}$, and (5.8) from $|\mathrm{psw}_{s'_j}(x)|_a > 0$ and the fact that for all r, every letter that appears in $\mathrm{psw}_r(y)$ appears more often in $\mathrm{psw}_r(x)$. □

LEMMA 5.3. *Let (4.1) be an equation in normal form. Let*

$$h = H(u_0 \cdots u_n) - \max\{\Sigma(u_0 \cdots u_i) \mid i \in \{0, \ldots, n-1\}\}. \tag{5.9}$$

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 $\Sigma(\phi(w)) = H(w)$. For any solution $[x]$, we have

$$\phi(u_0xu_1\cdots xu_n) = \phi(v_0xv_1\cdots xv_n). \tag{5.10}$$

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, we have $\phi(u_0\cdots u_n) = \phi(v_0\cdots v_n)$ and thus

$$|u_0\cdots u_{i-1}| + |\phi(u_i)| = |v_0\cdots v_{j-1}| + |\phi(v_j)|. \tag{5.11}$$

By (5.11) and Condition N3 in the definition of normal form, we have $i > j$.

Because $[\varepsilon]$ is a solution, we have $H(u_0\cdots u_n) = H(v_0\cdots v_n)$, and by Lemma 5.2,

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

From this and (5.9) it follows that

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

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

$$h = H(u_0\cdots u_n) - s_k = H(v_0\cdots v_n) - t_k.$$

To determine $\phi(u_0xu_1\cdots xu_n)$, let us look at sums of prefixes of $u_0xu_1\cdots xu_n$. If u'_r is a prefix of u_r, we have

$$\Sigma(u_0xu_1\cdots u_{r-1}xu'_r) = \Sigma(u_0\cdots u_{r-1}u'_r) \le H(u_0\cdots u_n) = s_k + h,$$

and equality is first reached for $r = i$ and $u'_r = \phi(u_i)$. If x' is a prefix of x, we have

$$\Sigma(u_0xu_1\cdots xu_{r-1}x') = \Sigma(u_0\cdots u_{r-1}x') = s_r + \Sigma(x') \le s_k + H(x),$$

and equality is first reached for $r = k$ and $x' = \phi(x)$. Thus

$$\phi(u_0xu_1\cdots xu_n) = \begin{cases} u_0xu_1\cdots u_{i-1}x\phi(u_i) & \text{if } H(x) < h \text{ or if } H(x) = h \text{ and } i < k, \\ u_0xu_1\cdots xu_{k-1}\phi(x) & \text{if } H(x) > h \text{ or if } H(x) = h \text{ and } i \ge k. \end{cases}$$

Similarly, we see that

$$\phi(v_0xv_1\cdots xv_n)=\begin{cases}v_0xv_1\cdots v_{j-1}x\phi(v_j) & \text{if } H(x)<h \text{ or if } H(x)=h \text{ and } j<l,\\ v_0xv_1\cdots xv_{l-1}\phi(x) & \text{if } H(x)>h \text{ or if } H(x)=h \text{ and } j\geq l,\end{cases}$$

This means that, for a given x, (5.10) 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 (5.11) 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

$$|u_0\cdots u_{k-1}|+|\phi(x)|+(k-1-j)|x|=|v_0\cdots v_{j-1}|+|\phi(v_j)|. \tag{5.12}$$

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, (5.12) 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.

This proves that the equation cannot have nonempty solution words of height less than h, and if the equation has a solution word of height more than h, then there is at most one other nonempty solution word, and it has height h. ∎

Example 5.4. Consider the equation

$$(XaXbXaabbabaXbabaabbab, abaabbabaXbabaabbXaXbX)$$

that was mentioned in Example 2.8. 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 5.3, 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.

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 $x_1, \dots, x_m, y_1, \dots, y_n \in Z$, $x_1 \cdots x_m = y_1 \cdots y_n$ implies $m = n$ and $x_i = y_i$ for all $i \in \{1, \dots, m\}$. If Z is a code, then Z^* is a free monoid. Conversely, the minimal generating set of a free monoid is a code. If Δ is an alphabet of the same size as Z, then the free monoids Z^* and Δ^* are isomorphic. More information about codes and free monoids 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. The fact that the set of all minimal zero-sum words is a code follows from [21, Lemma 4].

LEMMA 6.1. *Let E be the equation (4.1) 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*

$$(h(u_0)Xh(u_1)\cdots Xh(u_n), h(v_0)Xh(v_1)\cdots Xh(v_n)) \tag{6.1}$$

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

Proof. There exists an alphabet Δ and an isomorphism $h : Z^* \to \Delta^*$ by the definition of code. If $x \in Z^*$ is a solution word of E, then

$$\begin{aligned} h(u_0)h(x)h(u_1)\cdots h(x)h(u_n) &= h(u_0xu_1\cdots xu_n) \\ &= h(v_0xv_1\cdots xv_n) = h(v_0)h(x)h(v_1)\cdots h(x)h(v_n), \end{aligned}$$

so $[h(x)]$ is a solution of (6.1). On the other hand, if $[y]$ is a solution of (6.1), then there exists $x \in Z^*$ such that $h(x) = y$, and

$$\begin{aligned} h(u_0xu_1\cdots xu_n) &= h(u_0)yh(u_1)\cdots yh(u_n) \\ &= h(v_0)yh(v_1)\cdots yh(v_n) = h(v_0xv_1\cdots xv_n), \end{aligned}$$

so $u_0xu_1\cdots xu_n = v_0xv_1\cdots xv_n$ and $[x]$ is a solution of E. This completes the proof. ∎

Note that the equation E in Lemma 6.1 can have solution words that are not in Z^*, so (6.1) 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 6.2. *Consider the equation (4.1). 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

$$(u_0 X u_1 \cdots X u_k y, v_0 X v_1 \cdots v_{k-1} X). \tag{6.2}$$

Proof. If h is a nonempty solution of (4.1), then

$$h(u_0 X u_1 \cdots X u_n) = h(v_0 X v_1 \cdots X v_n).$$

Here the left-hand side has a prefix $h(u_0 X u_1 \cdots X u_k y)$ and the right-hand side has a prefix $h(v_0 X v_1 \cdots v_{k-1} X)$. These prefixes are of the same length, so they are equal. Thus h is a solution of (6.2). ∎

Using Lemma 6.2 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 6.3 is somewhat similar to the proof of Lemma 5.3, but simpler.

LEMMA 6.3. *Let (4.1) 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 $|v_0 \cdots v_{k-1}| - |u_0 \cdots u_k|$.

Proof. By symmetry, we can assume that $\Sigma(u_k) > 0$. By Lemma 5.3, the nonempty solutions have a common height h. For any word w of height at least $\Sigma(u_k) + h$, let $\psi(w)$ be its shortest prefix such that $H(\psi(w)) \geq \Sigma(u_k) + h$. If $[x]$ is a nonempty solution, then there exist indices i, j and words u, v such that u is a nonempty prefix of $u_i x$, v is a nonempty prefix of $v_j x$ and

$$\psi(u_0 x u_1 \cdots x u_n) = u_0 x u_1 \cdots u_{i-1} x u, \ \psi(v_0 x v_1 \cdots x v_n) = v_0 x v_1 \cdots v_{j-1} x v.$$

Here i, j, u, v are the same for all x, because every x has sum zero and height h, and the shortest x is a prefix of every other x. Clearly $i \leq k$, because

$$H(u_0 x u_1 \cdots u_k x) \geq \Sigma(u_0 x u_1 \cdots x u_k) + h = \Sigma(u_k) + h.$$

We know that $\psi(u_0 x u_1 \cdots x u_n) = \psi(v_0 x v_1 \cdots x v_n)$ (actually, we only need the fact that these words have the same length). Because

$$|u_0 x u_1 \cdots u_{i-1} x u| = |v_0 x v_1 \cdots v_{j-1} x v|$$

for more than one $|x|$, it must be $i = j$, and then $|u_0 \cdots u_{i-1} u| = |v_0 \cdots v_{i-1} v|$. Because $|u_0 \cdots u_i| \leq |v_0 \cdots v_{i-1}|$ by Condition N3 in the definition of normal form, u cannot

be a prefix of u_i. This means that $H(u_0xu_1\cdots xu_i) < \Sigma(u_k) + h$. If $i < k$, then u_i is zero-sum and thus adding x after xu_i does not increase the height, so also $H(u_0xu_1\cdots u_ix) < \Sigma(u_k) + h$, which is a contradiction. Therefore $i = k$. If there exists a nonempty solution $[x]$ of length at most $|v_0\cdots v_{k-1}| - |u_0\cdots u_k|$, then

$$|u_0\cdots u_{k-1}u| \le |u_0\cdots u_kx| \le |v_0\cdots v_{k-1}| < |v_0\cdots v_{k-1}v|,$$

a contradiction. ∎

LEMMA 6.4. *Let the equation (4.1) have the solution set $[p^*]$ for some primitive word p. Let $u_0 = v_n = \varepsilon$. Let $j \in \{0,\dots,n\}$ be the largest index such that the lengths of $u_0,\dots,u_{j-1}$ and $v_0,\dots,v_{j-1}$ are divisible by $|p|$. Then $j > 0$ and $|v_0\cdots v_{j-1}| - |u_0\cdots u_j| < |p|$.*

Proof. If $j = n$, the claim is clear. Otherwise, at least one of $|u_j|, |v_j|$ is not divisible by $|p|$. Let m be such that $|p^{m-1}| \ge |v_0\cdots v_j| - |u_0\cdots u_j|$. Let $d = |v_0\cdots v_{j-1}| - |u_0\cdots u_j|$.

Let r be the prefix of p^m of length $|p^m| - |v_0\cdots v_j| + |u_0\cdots u_j| \ge |p|$, and let p' be the suffix of r of length $|p|$. Because p is primitive, $p' = p$ if and only if $|r|$ is divisible by $|p|$. We have $u_0p^mu_1\cdots u_jp^m = v_0p^mv_1\cdots p^mv_jr$, and it follows that $p = p'$, so $|r|$ is divisible by $|p|$. This means that $|u_j|$ and $|v_j|$ are congruent modulo $|p|$, so neither of them is divisible by $|p|$. Consequently, $j \ne 0$ and d is not divisible by $|p|$.

Let s be the prefix of p^m of length d. If $d > |p|$, we can let p'' be the suffix of s of length $|p|$. Because p is primitive, $p'' = p$ if and only if $|s|$ is divisible by $|p|$. We have $u_0p^mu_1\cdots p^mu_js = v_0p^mv_1\cdots v_{j-1}p^m$, and it follows that $p = p''$, so $|s| = d$ is divisible by $|p|$. This is a contradiction, so $d < |p|$. ∎

Lemma 6.2 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 6.5. *If the equation (4.1) has a nonempty solution, $u_n = ua^m$ for some $u \in \Gamma^*$, $a \in \Gamma$ and $m \ge 0$, and $u_0\cdots u_{n-1}u$ is a prefix of $v_0\cdots v_n$, then the equation has the empty solution.*

Proof. Let y be a word such that $u_0\cdots u_{n-1}uy = v_0\cdots v_n$. We say that words p, q are *abelian equivalent* if $|p|_b = |q|_b$ for all letters b. Because (4.1) has a solution, $u_0\cdots u_{n-1}ua^m$ and $v_0\cdots v_n$ are abelian equivalent. Thus $u_0\cdots u_{n-1}uy$ and $u_0\cdots u_{n-1}ua^m$ are abelian equivalent, so y and a^m are abelian equivalent and $y = a^m$. The claim follows. ∎

7. Main results. Now we are ready to prove our main results.

THEOREM 7.1. *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 4.1. 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 4.4, we can assume that E_1 is the equation (4.1) and it is in normal form. By Lemma 5.1, 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 solutions 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 5.2, 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. If it were $u_i, v_i \in 0^*$ for all i, then the equation would have infinitely many solutions, which is not possible. We can use Lemma 6.1 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 6.3 and 6.2, 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 2.9 and Corollary 2.11. 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 5.2 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 6.5 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 $[(qp)^* q]$ for some primitive word qp by Theorem 2.9. From $\varepsilon \in [(qp)^* q]$ it follows that $q = \varepsilon$, so the solution set is $[p^*]$ and p is primitive. 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 6.4 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 $j < k$, we have $|v_0 \cdots v_{j-1}| - |u_0 \cdots u_j| < |p|$ (remember that f is length-preserving). By letting $z = y$ if $j = k$, or by using Lemma 6.2 with j as k for E_1 otherwise, we get an equation

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

where z is the common prefix of length $|v_0 \cdots v_{j-1}| - |u_0 \cdots u_j| < |p|$ of all nonempty

solution words of E_1, and E_4 has at least all the same nonempty solutions as E_1. 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 6.1 with $Z = \Gamma^{|p|}$ for E_4. If h is the morphism of Lemma 6.1, then we get the equation

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

It has at least three nonempty solutions, but only finitely many. Because $|z| \leq |p|$, we have $h(u_jz) = h(u)c$, 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 6.5 with c as a and $m = 1$, so E_5 has the empty solution. This contradicts the minimality of E_1. ∎

THEOREM 7.2. *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 7.1 and Theorem 2.12. ∎

THEOREM 7.3. *If a decreasing chain of constant-free three-variable equations has a nonperiodic solution, then it has at most 20 equations.*

Proof. Follows from Theorem 7.1 and Theorem 2.13. ∎

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 2 and 17.

Improving the bound 17 is an obvious open problem. A possible approach would be to improve the results in [18].

Another open problem is proving similar bounds for more than three variables. The result in [18] 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 [24, 25]. 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 2.9, 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*, Theoretical Computer Science, 41 (1985), pp. 121–123, https://doi.org/10.1016/0304-3975(85)90066-0.

[2] J. BERSTEL, D. PERRIN, AND C. REUTENAUER, *Codes and Automata*, Cambridge University Press, 2010.

[3] L. G. BUDKINA AND A. A. MARKOV, *F-semigroups with three generators*, Akademiya Nauk SSSR. Matematicheskie Zametki, 14 (1973), pp. 267–277.

[4] K. CULIK, II AND J. KARHUMÄKI, *Systems of equations over a free monoid and Ehrenfeucht's conjecture*, Discrete Mathematics, 43 (1983), pp. 139–153, https://doi.org/10.1016/0012-365X(83)90152-8.

[5] E. CZEIZLER AND J. KARHUMÄKI, *On non-periodic solutions of independent systems of word equations over three unknowns*, International Journal of Foundations of Computer Science, 18 (2007), pp. 873–897, https://doi.org/10.1142/S0129054107005030.

[6] E. CZEIZLER AND W. PLANDOWSKI, *On systems of word equations over three unknowns with at most six occurrences of one of the unknowns*, Theoretical Computer Science, 410 (2009), pp. 2889–2909, https://doi.org/10.1016/j.tcs.2009.01.023.

[7] R. Dąbrowski and W. Plandowski, *On word equations in one variable*, Algorithmica, 60 (2011), pp. 819–828, https://doi.org/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, vol. 841 of LNCS, Springer, 1994, pp. 336–341, https://doi.org/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*, Matematicheskie Zametki, 40 (1986), pp. 321–324, https://doi.org/10.1007/BF01142470.

[10] T. Harju and D. Nowotka, *On the independence of equations in three variables*, Theoretical Computer Science, 307 (2003), pp. 139–172, https://doi.org/10.1016/S0304-3975(03)00098-7.

[11] Š. Holub and J. Žemlička, *Algebraic properties of word equations*, Journal of Algebra, 434 (2015), pp. 283–301, https://doi.org/10.1016/j.jalgebra.2015.03.021.

[12] J. Honkala, *On chains of word equations and test sets*, Bulletin of the European Association for Theoretical Computer Science, 68 (1999), pp. 157–160.

[13] A. Jeż, *One-variable word equations in linear time*, Algorithmica, 74 (2016), pp. 1–48, https://doi.org/10.1007/s00453-014-9931-3.

[14] A. Jeż, *Word equations in nondeterministic linear space*, in Proceedings of the 44th ICALP, vol. 80 of LIPIcs, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 95:1–13, https://doi.org/10.4230/LIPIcs.ICALP.2017.95.

[15] J. Karhumäki and W. Plandowski, *On the defect effect of many identities in free semigroups*, in Mathematical aspects of natural and formal languages, G. Paun, ed., World Scientific, 1994, pp. 225–232, https://doi.org/10.1142/9789814447133_0012.

[16] J. Karhumäki and A. Saarela, *On maximal chains of systems of word equations*, Proceedings of the Steklov Institute of Mathematics, 274 (2011), pp. 116–123, https://doi.org/10.1134/S0081543811060083.

[17] M. Laine and W. Plandowski, *Word equations with one unknown*, International Journal of Foundations of Computer Science, 22 (2011), pp. 345–375, https://doi.org/10.1142/S0129054111008088.

[18] D. Nowotka and A. Saarela, *One-variable word equations and three-variable constant-free word equations*, International Journal of Foundations of Computer Science, 29 (2018), pp. 935–950, https://doi.org/10.1142/S0129054118420121.

[19] D. Nowotka and A. Saarela, *An optimal bound on the solution sets of one-variable word equations and its consequences*, in Proceedings of the 45th ICALP, vol. 107 of LIPIcs, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, pp. 136:1–136:13, https://doi.org/10.4230/LIPIcs.ICALP.2018.136.

[20] A. Saarela, *Systems of word equations, polynomials and linear algebra: A new approach*, European Journal of Combinatorics, 47 (2015), pp. 1–14, https://doi.org/10.1016/j.ejc.2015.01.005.

[21] A. Saarela, *Word equations where a power equals a product of powers*, in Proceedings of the 34th STACS, vol. 66 of LIPIcs, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 55:1–55:9, https://doi.org/10.4230/LIPIcs.STACS.2017.55.

[22] A. Saarela, *Word equations with kth powers of variables*, Journal of Combinatorial Theory. Series A, 165 (2019), pp. 15–31, https://doi.org/10.1016/j.jcta.2019.01.004.

[23] A. Saarela, *Hardness results for constant-free pattern languages and word equations*, in Proceedings of the 47th ICALP, vol. 168 of LIPIcs, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2020, pp. 140:1–140:15, https://doi.org/10.4230/LIPIcs.ICALP.2020.140.

[24] J.-C. Spehner, *Quelques problémes d'extension, de conjugaison et de présentation des sous-monoïdes d'un monoïde libre*, PhD thesis, Univ. Paris, 1976.

[25] J.-C. Spehner, *Les systemes entiers d'équations sur un alphabet de 3 variables*, in Semigroups Theory and Applications, Springer, 1988, pp. 342–357, https://doi.org/10.1007/BFb0083443.