Word Equations With Two Variables

Witold Charatonik,
Institute of Computer Science
University of Wrocław
Wrocław, Poland

Leszek Pacholski
Institute of Mathematics
Polish Academy of Sciences
Wrocław, Poland

1 Introduction

The problem whether the set of all equations that are satisfiable in some free semigroup - or, equivalently, in an algebra of words with concatenation - is recursive (usually called the satisfiability problem for semigroup equations) was first formulated by A.A. Markov in early sixties (see [3]). Special cases of the problem were solved affirmatively by A.A. Markov (see [3]), Yu.I. Khmelevskiĭ [8], [7], G. Plotkin, [14] and A. Lentin [11]. The full positive solution, was given by G.S. Makanin in a paper [12], which is long and very technical.

Makanin's decision procedure for equational satisfiability in semigroups has received a lot of attention in the literature. Undoubtedly, this is because the notion of an algebra of words (or strings) with the operation of concatenation - is of fundamental importance in computer science: many algorithms and data structures refer to words. Thus, several improvements of Makanin's algorithm have been given by H. Abdulrab, J.-P. Pecuchet, K. Schulz, A. Kościelski and L. Pacholski (see [2], [13], [15], [9], and [10]), and attempts have even been made to implement the algorithm (see [1]). Moreover, related unification problems have been studied. In particular, J. Jaffar, in [6], basing on the Makanin's decision procedure, described an algorithm which, when an equation has a solution, generates all its solutions and halts if the set of solutions is finite.

An important fact used in the Makanin's algorithm and in the unification algorithms based on it, is that the periodicity exponent of a minimal solution of a word equation can be bounded by a recursive function of the length of the equation. In fact, V.K.Bulitko, in [4], proved that if d is the length of an equation, then the index of periodicity of its minimal solution (see below) does not exceed $(6d)^{2^{2d^4}} + 2$. Kościelski and Pacholski ([9], [10]) forced this bound down to $2^{1.07d}$. They also prove a lower bound of $2^{0.29d}$ for the exponent of periodicity of minimal solutions of a word equation of length d.

Although the bound on the exponent of periodicity given by Kościelski and Pacholski gave an over-exponential improvement of the algorithm its complexity is still so high that it prohibits any applications in practice. Moreover, Kościelski and Pacholski [10] proved that the problem of the solvability of word equations is NP-hard, even if a linear bound is put on the length of possible solutions. Thus, for a given constant $c > 2$ the problem of the existence of a solution of length cd for an equation of length d is NP-complete. This implies, that there does not exist any fast algorithm, which decides solvability of all word equations and suggests that good algorithms can be found only for restricted classes of equations.

This paper contains the first report on our research project with the aim to describe classes of word equations for which fast algorithms, deciding solvability or giving actual solutions, exist. In this paper by "fast" we mean "deterministic polynomial time". Of course for many actual applications it would be better to consider more restricted classes like linear time or DTIME(nlog(n)). This problem will be considered in subsequent papers.

We consider equations which have at most two distinct variables. For such equations we give a deterministic polynomial time algorithm deciding their solvability. Our technique and algorithm is based on the notion of an "equation in exponent" which has been introduced by Yu.I. Khmelevskiĭ [7].

2 Preliminaria

The set of nonnegative integers is denoted by IN. For a finite set Σ, Σ^* is the set of words over Σ (the free semigroup generated by Σ), Σ^+ is the set of nonempty words over Σ, and Σ^c is the set of words of length c over Σ. ε denotes is the empty word, and $|W|$ denotes the length of a word W.

Let $\Sigma = \{a_1, \ldots, a_n\}$ and $\Xi = \{x, y\}$ be two disjoint alphabets, called respectively the alphabet of coefficients and the alphabet of variables. A word equation $\mathcal{E}$ over (Σ, Ξ) is a pair of words (W_1, W_2) (also denoted by $W_1 = W_2$), where $W_1, W_2 \in (\Sigma \cup \Xi)^+$. $|W_1 W_2|$ is the length of $\mathcal{E}$. A solution of $\mathcal{E}$ is a function $\upsilon : \Xi \to \Sigma^+$ such that $W_1(\upsilon(x)/x), \upsilon(y)/y)) = W_2(\upsilon(x)/x), \upsilon(y)/y))$, where $W(\upsilon(x)/x))$ denotes the word obtained from W by replacing each occurrence of x by $\upsilon(x)$. Given any function $\upsilon : \Xi \to \Sigma^*$, slightly abusing the notation, by the same letter υ we denote the extension of υ to the homomorphism $\upsilon : (\Sigma \cup \Xi)^* \to \Sigma^*$, which is the identity on Σ. Sometimes we identify the function υ with the pair of words $(\upsilon(x), \upsilon(y))$. For words A_1, A_2 we write $A_1 < A_2$ if A_1 is a prefix of A_2. If $A_1, \ldots, A_n$ are words then $[A_i]_{i=1}^n$ denotes the concatenation of $A_1, \ldots, A_n$. A word A is primitive if $A \neq S^n$, for any word S and any integer $n > 1$. The length of a solution υ is $|\upsilon(x)| + |\upsilon(y)|$. A solution is minimal if it has minimal length.

Below we shall give some preliminary results. Most of them can be found in [8].

Proposition 2.1 (Proposition 1.16 in [8]) *Let $A \in \Sigma^+$, $V, \Phi, \Psi \in \Sigma^*$, and assume that $V\Phi = A^a V\Psi$, for some $a > 0$. Then $V = A^t A_1$, for some $t \geq 0$ and $A_1 < A$.*

Notice that every equation with one variable is equivalent to one in the form

$$A[xA_i]_{i=1}^n = [xB_j]_{j=1}^m. \tag{1}$$

Lemma 2.2 (Proposition 1.19 in [8]) *If the equation (1) is solvable then it has a solution of length smaller then $M^2 + 3M$, where $M = \max_{i,j}\{n, m, |A_i|, |B_j|, |A|\}$.*

Corollary 2.3 *It can be decided in time $O(d^5)$ if a word equation $\mathcal{E}$ of length d with one variable has a solution.*

Proof. Without any loss of generality we can assume that $\mathcal{E}$ has the form $x\Phi = Px\Psi$, where $\Phi, \Psi \in (\Sigma \cup \{x\})^*$, $P \in \Sigma^+$. By Proposition 2.1 for any solution $\upsilon : \{x\} \to \Sigma^+$ of $\mathcal{E}$ we have $\upsilon(x) = P^t P_1$ for some integer t and a word $P_1 < P$. Moreover, by Proposition 2.2 $\mathcal{E}$ has a solution of length smaller than $d^2 + 3d$, so to decide if $\mathcal{E}$ is solvable it suffices to check if it is satisfied by one of $d^2 + 3d$ words of the form $P^t P_1$. This can be done in time $O(d^5)$ since there are $O(d^2)$ possibilities and in each case the word $\upsilon(x)\Phi Px\Psi(\upsilon(x)/x)$ has the length $O(d^3)$. ■

Definition 2.4 *An exponential equation is an expression of the form $P_0[S_i^{\lambda_i} P_i]_{i=1}^n = Q_0[T_j^{\mu_j} Q_j]_{j=1}^m$, where λ_i, μ_j are integer variables, $P_i, Q_j \in \Sigma^*$, $S_i, T_j \in \Sigma^+$. A solution of such an equation assigns integer values to variables in such a way, that both sides of the equation become graphically identical.*

Lemma 2.5 (Proposition 2.4′ in [8]) *If an exponential equation $P_0[S^{\lambda_i} P_i]_{i=1}^n = Q_0[S^{\mu_j} Q_j]_{j=1}^m$ with two variables, (i.e. $\lambda_i, \mu_j \in \{\lambda, \mu\}$) is solvable, then it has a solution such that $\lambda \leq 4h^2 H$ or $\mu \leq 4h^2 H$, where $h = \max\{n, m, 8\}$, and $H = \frac{\max\{|S|, |P_i|, |Q_j|\}}{|S|}$*

Lemma 2.6 (Proposition 2.7 in [8]) *If an exponential equation $P_0[S^{\lambda} P_i]_{i=1}^n = Q_0[S^{\lambda} Q_j]_{j=1}^m$ with one variable λ is solvable, then it has a solution such that $\lambda \leq 2hH$, where $h = \max\{n, m, 8\}$, and $H = \frac{\max\{|S|, |P_i|, |Q_j|\}}{|S|}$*

Lemma 2.7 (Implicite in the proof of Proposition 2.8 in [8]) *If an exponential equation $[(S^\sigma C)^{\lambda_i} S^\sigma A_i]_{i=1}^n = [(S^\sigma C)^{\mu_j} S^\sigma B_j]_{j=1}^m$ with variables λ_i, μ_j, σ has a solution such that $\lambda_i, \mu_j \geq 3$, then it has a solution such that $\lambda_i, \mu_j \geq 3$ and $\sigma|S| + |C| \leq \max_{i,j}\{|A_i|, |B_j|\} + 2|SC|$.*

Corollary 2.8 *If an exponential equation $P_0[S^{\lambda_i} P_i]_{i=1}^n = Q_0[S^{\mu_j} Q_j]_{j=1}^m$, where $\lambda_i, \mu_j \in \{\lambda, \mu\}$ are integer variables, is solvable, then it has a solution such that $\lambda \leq 4d^3$, $\mu \leq 8d^5$ or $\mu \leq 4d^3$, $\lambda \leq 8d^5$, where $d = |P_0[SP_i]_{i=1}^n Q_0[SQ_j]_{j=1}^m|$, so it can be solved in time $O(d^{14})$.*

Proof. It is an easy consequence of Lemma 2.5 and Lemma 2.6. ■

Definition 2.9 *A directed equation is an expression of the form $x\Phi \to y\Psi$ or $x\Phi \leftarrow y\Psi$, where $\Phi, \Psi \in (\Sigma \cup \Xi)^*$. A solution of $x\Phi \to y\Psi$ is any solution υ of the equation $x\Phi = y\Psi$ such that $|\upsilon(x)| > |\upsilon(y)|$.*

Lemma 2.10 (Proposition 3.1 in [8]) *Given words $\Phi, \Psi \in (\Sigma \cup \Xi)^*$, and $B_j \in \Sigma^*$ for $j \leq b$, where b is an integer ≥ 1, let*

$$x\Phi \to [yB_j]_{j=1}^{b} x\Psi \tag{2}$$

be a directed equation with two variables x, y. For integers t, k and a word B such that $0 \leq t, 0 \leq k < b$, and $B < B_{k+1}$, we put

$$\sigma_1(x) = ([yB_j]_{j=1}^{b})^t [yB_j]_{j=1}^{k} yB,\ \sigma_2(x) = ([yB_j]_{j=1}^{b})^t [yB_j]_{j=1}^{k} x,$$

and $\sigma_1(y) = \sigma_2(y) = y$.

Then if $(\upsilon(x), \upsilon(y))$ is a solution of (2) then exactly one of the two conditions below holds:

1. *For some k, t such that $0 \leq k < b, 0 \leq t$ and a prefix B of B_{k+1} we have*

$$\upsilon(x) = ([\upsilon(y)B_j]_{j=1}^{b})^t [\upsilon(y)B_j]_{j=1}^{k} \upsilon(y)B \tag{3}$$

   and $\upsilon(y)$ is a solution of the equation

$$\sigma_1(x\Phi) = \sigma_1([yB_j]_{j=1}^{b} x\Psi) \tag{4}$$

   obtained by applying σ_1 to (2).

2. *For some integers k, t such that $0 \leq k < b, 0 \leq k$ and a word X' we have*

$$\upsilon(x) = ([\upsilon(y)B_j]_{j=1}^{b})^t [\upsilon(y)B_j]_{j=1}^{k} X' \tag{5}$$

   and $(X', \upsilon(y))$ is a solution of the equation

$$\sigma_2(x\Phi) = \sigma_2([yB_j]_{j=1}^{b} x\Psi) \tag{6}$$

   obtained by applying σ_2 to (2).

Moreover, for any solution $\upsilon(y)$ of (4) the pair $(\upsilon(x), \upsilon(y))$, where $\upsilon(x)$ is defined by (3) is a solution of (2).

Finaly, for any solution $(X', \upsilon(y))$ of (6), the pair $(\upsilon(x), \upsilon(y))$, where $\upsilon(x)$ is defined by (5) is a solution of (2).

Lemma 2.11 (Proposition 4.4 in [8])*Let $\nu(\Phi)$ denote the pair (t_1, t_2), where t_1 (respectively t_2) is the number of occurrences of the variable x (respectively y) in the word Φ. Let $\bar{\Phi}$ denote the projection of the word Φ onto the alphabet Σ. Let $\Phi_1\Phi_2 = \Psi_1\Psi_2$ be an equation with two variables such that $\nu(\Phi_1) = \nu(\Psi_1)$ and $c = |\bar{\Phi}_1| - |\bar{\Psi}_1| \geq 0$. Then the following equivalence holds:*

$$\Phi_1\Phi_2 = \Psi_1\Psi_2 \iff \exists R \in \Sigma^c\ (\Phi_1 = \Psi_1 R\ \&\ R\Phi_2 = \Psi_2)$$

Lemma 2.12 (Proposition 5.10 in [8]) *The directed equation $xAy \to yBx$ is solvable if and only if there exist words $P, S \in \Sigma^*$, $Q \in \Sigma^+$ such that $A = PQS$, $B = SQP$.*

Definition 2.13 *For an equation*

$$xAy\Phi \to yBx\Psi \tag{7}$$

and a substitution σ such that $\sigma(x) = (xAyB)^t x,\ \sigma(y) = xAy$. equations $xAy\sigma(\Phi) \to yBx\sigma(\Psi)$ and $xAy\sigma(\Phi) \leftarrow yBx\sigma(\Psi)$ are called t-images of (7).

We say that equation (7) has the property α if it is equivalent to an equation of the form $xAyP\xi\Phi' \to yBxQ\eta\Psi'$, for some $P, Q \in \Sigma^*,\ P \neq Q,\ \Phi, \Psi \in (\Sigma \cup \Xi)^*,\ \xi, \eta \in \Xi$.

Lemma 2.14 (Proposition 7.4 in [8]) *If $A \neq B$, then either equation (7) is equivalent to the equation $xAy \to yBx$, or each t-image of (7) has the property α.*

Definition 2.15 *The exponent of periodicity of a word W is the maximal positive integer p such that $W = U_1U^pU_2$ for some words U_1, U_2 and a nonempty word U. The exponent of periodicity of a minimal solution of equation $\mathcal{E} = (W_1, W_2)$ is the greatest of the exponents of periodicity of the words $\upsilon(x), \upsilon(y)$, where υ is a minimal (with respect to length) solution of $\mathcal{E}$.*

Lemma 2.16 *If p is the exponent of periodicity of a minimal solution of an equation of length $d \geq 6$ over $(\Sigma, \{x, y\})$, then $p \leq 4d^5$.*

A proof of this lemma is given in [5] following the idea of the proof of the main theorem of Chapter 3 in [9].

Lemma 2.17 *An equation $x\Phi = Px\Psi$, of length $d \geq 6$ with two variables x, y, and such that $P \in \Sigma^*$, $\Phi, \Psi \in (\Sigma \cup \{x, y\})^*$ is solvable if and only if it has solution υ such that $\upsilon(x) = P^tP_1$ for an integer $t \leq 4d^5$ and a word $P_1 < P$.*

Proof. It is easy to see that any solution υ of this equation is of the form $\upsilon(x) = P^tP_1$ for some integer $t \geq 0$ and $P_1 < P$. The bound $t \leq 4d^5$ follows from lemma 2.16. ■

Definition 2.18 *For a word S and an integer M we write $\tau(S, M)$ if the following condition holds: there exist words A, A_1 and integer t such that $A_1 < A$, $|A| < M$ and $S = A^tA_1$.*

A reduction of an equation (WW_1, WW_2) consists of transformation of it into an equivalent equation (W_1, W_2), where words W_1, W_2 begin with different symbols. An equation is trivially unsolvable if after reduction its sides begin with different symbols from the alphabet of coefficients or exactly one of its sides is the empty word.

3 Basic equations

In this section we consider equations of the form $xAy\Phi = yBx\Psi$ of length d, with $A, B \in \Sigma^*$, $\Phi, \Psi \in (\Sigma \cup \Xi)^*$, and $|A| = |B|$. We distinguish two cases, $A = B$ and $A \neq B$.

3.1 The algorithm

Case 1 $A = B$

Make in parallel steps 1.1, 1.2, 1.3.

Step 1.1 For each suffix Z of A and prefixes Z_1, Z_2 of Z such that $Z_1A = Z^k$, and $Z_2A = Z^l$ for some $k, l \in \mathbb{N}$ substitute $Z^\lambda Z_1$ for x, $Z^\mu Z_2$ for y and solve the exponential equation with variables λ, μ obtained by this substitution.

Step 1.2 For each λ, μ such that either $(\lambda < 3, \mu \leq 4d^2)$ or $(\mu < 3, \lambda \leq 4d^2)$ or $(3 \leq \lambda \leq 4(12d^3)^3, 3 \leq \mu \leq (8(12d^3)^5)$ or $(3 \leq \mu \leq 4(12d^3)^3, 3 \leq \lambda \leq (8(12d^3)^5)$ substitute $(\upsilon A)^\lambda\upsilon$ for x, $(\upsilon A)^\mu\upsilon$ for y and solve the word equation with one variable υ obtained by this substitution.

Step 1.3 Substitute $(\upsilon A)^\lambda\upsilon$ for x, $(\upsilon A)^\mu\upsilon$ for y and solve the exponential equation obtained in this way. This is an equation with two variables λ, μ, over alphabet $\Sigma \cup \{\upsilon\}$ of coefficients.

Case 2 $A \neq B$

Step 2.1 substitute y for x and solve the equation with one variable obtained by the substitution

Step 2.2 solve two systems of equations $(xAy \to yBx, \Phi = \Psi)$ and $(xAy \leftarrow yBx, \Phi = \Psi)$

Step 2.2.1 input: a system $xAy \to yBx$, $\Phi = \Psi$ of equations
output: at most one system of equations of the form $xAy \to yBx$, $P\xi\Phi' = \eta\Psi'$ ($\xi, \eta \in \Xi$).

Use the following procedure to transform the equation $\Phi = \Psi$ to an equation of the form $P\xi\Phi' = \eta\Psi'$ with $P \neq \varepsilon$ (or to the empty or a trivially unsolvable equation):

REPEAT

- reduce the equation
- if the equation has the form $xA_1[\xi_i A_i]_{i=2}^n = (yB)^q xB_{q+1}[\eta_i B_i]_{i=q+2}^m$, then replace it with the equation $xA_1[\xi_i A_i]_{i=2}^n = x(Ay)^q B_{q+1}[\eta_i B_i]_{i=q+2}^m$
- if the equation has the form $xAyA_2[\xi_i A_i]_{i=3}^n = yB_1[\eta_i B_i]_{i=2}^m$, then replace it with the equation $yBxA_2[\xi_i A_i]_{i=3}^n = yB_1[\eta_i B_i]_{i=2}^m$

UNTIL none of the rules above applies.

If the equation obtained untill now is not of the form $P\xi\Phi' = \eta\Psi'$ then for each $t \leq d$ and $C < B$

- substitute $(yB)^t yC$ for x and solve the equation with one variable obtained in this way
- substitute $(xAyB)^t x$ for x, xAy for y, reduce the equation and if now it has the form $xAyBx\overline{\Phi} = yCxAy\overline{\Psi}$ then replace it with equation $yBxBx\overline{\Phi} = yCxAy\overline{\Psi}$ and reduce the last one

If this procedure gives the empty equation, then solve the equation $xAy \to yBx$ by finding decomposition $A = PQS$, $B = SQP$ with $Q \neq \varepsilon$. $x = QSQ$, $y = Q$ is a solution of this equation. If the procedure gives a trivially unsolvable equation, then give up this branch of algorithm.

Step 2.2.2 input: system of equations $xAy \to yBx$, $P\xi\Phi = \eta\Psi$ with $|P\xi\Phi\eta\Psi| = d$
For each $t \leq 4d^5$ and $P_1 < P$ substitute yx for x, and then $P^t P_1$ for y in the equation $xAyP\xi\Phi \to yBx\eta\Psi$ and solve equation with one variable obtained by this substitution.

3.2 Correctness of the algorithm

Theorem 3.1 (correctness of algorithm 3.1)

- *in case 1 an input equation is solvable if and only if one of the equation created in steps 1.1–1.3 has a solution*
- *in case 2 an input equation is solvable if and only if one of the equation (or systems of equations) created in steps 2.1–2.2 has a solution*
- *if output in step 2.2.1 is nonempty, then the input system is equivalent to the output system; otherwise the input system is solvable if and only if there exist a decomposition described in this step*
- *in step 2.2.2 the input system of equations is solvable if and only if one of the equations created in this step has a solution*

Proof. **Case 1**. If $A = B$, then $xAyA = yAxA$, so there exists such word $Z \in \Sigma^+$ that $xA = Z^k$ and $yA = Z^l$ for some $k, l \in \mathbb{N}$.

Step 1.1 corresponds to the case $|Z| \leq |A|$. If $|Z| > |A|$, then $Z = \upsilon A$ for some word υ, so $x = (\upsilon A)^\lambda \upsilon$, and $y = (\upsilon A)^\mu \upsilon$ for some $\lambda, \mu \in \mathbb{N}$.

Step 1.2 corresponds to the case $\tau(\upsilon, M)$, where M is the maximal length of coefficient subword of Φ or Ψ. In this case $\upsilon = S^\sigma S_1$ for some $\sigma \in \mathbb{N}$, $S_1 < S$, and $|S| \leq M$, so $x = (S^\sigma S_1 A)^\lambda S^\sigma S_1$, and $y = (S^\sigma S_1 A)^\mu S^\sigma S_1$. If $\lambda < 3$ then (treating λ and σ as fixed) by proposition 2.7 of [8] we have $\mu \leq 4d^2$, and similarly for $\mu < 3$. If $\lambda, \mu \geq 3$ then by proposition 2.8 of [8] we have $\sigma \leq 6M$ and by substituting $(S^\sigma S_1 A)^\lambda S^\sigma S_1$ for x, and $(S^\sigma S_1 A)^\mu S^\sigma S_1$ for y we get an exponential equation of length less then $12d^3$ which can be solved using lemma 2.8.

Step 1.3 corresponds to the case $\neg\tau(\upsilon, M)$. After substituting $(\upsilon A)^\lambda\upsilon$ for x and $(\upsilon A)^\mu\upsilon$ for y (with fixed λ, μ) our equation gets the form $A_0[\upsilon A_i]_{i=1}^n = B_0[\upsilon B_j]_{j=1}^m$. Since we have $\neg\tau(\upsilon, M)$, proposition 1.16 of [8] gives us $n = m,\ A_i = B_i$ for $i \leq n$.

Case 2. Steps 2.1 and 2.2 correspond to the three possibilities: either $|x| = |y|$ (step 2.1), or $|x| < |y|$, or $|x| > |y|$ (step 2.2).

Step 2.2.1. Proof of correctness of this step can be found in the proof of propositions 7.4 and 5.10 of [8].

Step 2.2.2. From the definition of directed equation it follows that any solution of the system $xAy \to yBx$, $P\xi\Phi = \eta\Psi$ is such that $x = yx'$ for some nonempty word x'. After substitution yx for x the equation $P\xi\Phi = \eta\Psi$ gets the form $Py\Phi' = y\Psi'$ and the thesis follows from lemma 2.17. ■

4 Reduction to basic equations

Now we consider an arbitrary equation of length d with two distinct variables. Reducing the same symbols in the beginning of both sides of the equation we can assume the equation has the form $P\xi\Phi = \eta\Psi$, where $P \in \Sigma^*, \xi, \eta \in \Xi = \{x, y\}$, and $\Phi, \Psi \in (\Sigma \cup \Xi)^*$

4.1 The algorithm

Step 1 input: an equation $P\xi\Phi = \eta\Psi$ of length d
output: $O(d)$ equations of the form $x\Phi' = y\Psi'$, each of length $O(d^2)$.

Case 1.1 $\xi = \eta$, and the equation has the form $Px\Phi = x\Psi$.
For each $t \leq 4d^5$ and each $P_1 < P$ substitute P^tP_1 for x and solve the equation with one variable obtained by this substitution.

Case 1.2 $\xi \neq \eta$, and the equation has the form $Px\Phi = y\Psi$,

Step 1.2.1 For each $P_1 < P$ substitute P_1 for y and solve the equation with one variable obtained by this substitution.

Step 1.2.2 Substitute Py for y and solve the equation $x\Phi' = y\Psi'$ obtained by this substitution.

Step 2 input: an equation $x\Phi = y\Psi$
output: two equations of the form $x\Phi \to y\Psi$

Step 2.1 Substitute y for x and solve the equation with one variable obtained by this substitution.

Step 2.2 Solve the directed equation $x\Phi \to y\Psi$

Step 2.3 Solve the directed equation $x\Phi \leftarrow y\Psi$

Step 3 input: an equation $x\Phi \to y\Psi$ of length d
output: $O(d^2)$ equations of length $O(d^2)$ of the form $xAy\Phi' = yBx\Psi'$

We can assume that input is of the form $xA\xi\Phi'' \to [yB_j]_{j=1}^b x\Psi''$ where $A, B_j \in \Sigma^*$ (if not, we consider the equivalent equation $x\Phi x \to y\Psi x$)

Step 3.1 For each $t \leq d$, $k < b$ and $B'_{k+1} < B_{k+1}$ substitute $([yB_j]_{j=1}^b)^t[yB_j]_{j=1}^k yB'_{k+1}$ for x and solve the equation with one variable obtained by this substitution.

Step 3.2 For each $t \leq d$, $k \leq b$ substitute $([yB_j]_{j=1}^b)^t[yB_j]_{j=1}^k x$ for x. In this way $O(d^2)$ equations of length $O(d^2)$ of the form $xAy\bar{\Phi} \leftarrow [yB'_j]_{j=1}^b x\bar{\Psi}$ are obtained.

Step 3.3 Substitute xy for y and get $Axy\bar{\Phi}(xy/y) = [yB'_jx]_{j=1}^b\bar{\Psi}(xy/y)$

Step 3.4 For each $A_1 < A$ substitute A_1 for y and solve the equation with one variable obtained by this substitution.

Step 3.5 Substitute Ay for y and get $xAy\Phi' = yB_1'x\Psi'$

Step 4 input: an equation $xAy\Phi = yBx\Psi$ of length d
output: $O(d^2)$ equations of the form $xA'y\Phi' = yB'x\Psi'$ of length $O(d^2)$ with $|A'| = |B'|$
We can assume that $|A| - |B| = c > 0$ (if not then just replace A with B and x with y)

Step 4.1 For each $0 < i \leq c$ let A_i be the prefix of A of length $|B| + i$. For each $0 \leq j \leq |A_i|$ let A_j' be the suffix of A_i of length j. Let $R_{i,j}$ be the prefix of length c of the word $A_j'A_i$. Substitute $R_{i,j}$ for y and solve the equation with one variable obtained by this substitution.

Step 4.2 For each decomposition $A = PQS$ of the word A such that $P, S \in \Sigma^*$, $Q \in \Sigma^+$, let $R \in \Sigma^c$ be a word such that $RB = SQP$ (if such a word exists). Substitute yR for y and solve the equation $xAyR\Phi(yR/y) = yRBx\Psi(yR/y)$

4.2 The correctness

Theorem 4.1 (correctness of algorithm 4.1) *In each step of the algorithm 4.1 an input equation is solvable if and only if there exists a solvable equation created in this step.*

Proof. We will prove this theorem separately for each step of the algorithm.

Step 1. In the case 1.1 thesis of the theorem follows from lemma 2.17. In the case 1.2 there are two possibilities: either $|y| \leq |P|$ (this is step 1.2.1) or $|y| > |P|$ (this is step 1.2.2).

Step 2. It follows from the fact that there are three possibilities: either $|x| = |y|$ or $|x| < |y|$ or $|x| > |y|$.

Step 3. Steps 3.1 and 3.2 follow from proposition 3.1 of [8]. Step 3.3 follows from definition of a directed equation. The next two steps (3.4, 3.5) correspond to cases $|y| \leq |A|$ and $|y| > |A|$.

Step 4. By proposition 4.4 of [8] the equation $xAy\Phi = yBx\Psi$ is solvable if and only if there exists a word R of length c such that system of two equations $xAy = yBxR$, $R\Phi = \Psi$ is solvable.

Step 4.1 correspond to the case $|y| \leq c$. In this case we have $y = R_1$ for some suffix R_1 of R, and the first equation of our system gets the form $xAR_1 = R_1BxR$. Any solution of this equation must satisfy condition $x = (R_1B)^tR'$ for some $t \in I\!N$ and $R' < R_1B$.

Reading this equation from right to left we get $\overleftarrow{R}\,\overleftarrow{x}\,\overleftarrow{B}\,\overleftarrow{R_1} = \overleftarrow{R_1}\,\overleftarrow{A}\,\overleftarrow{x}$ where $\overleftarrow{P}$ is the reverse of the word P, i.e. the word P read from right to left. After reducing this equation we get $\overleftarrow{x}\,\overleftarrow{B}\,\overleftarrow{R_1} = \overleftarrow{A_i}\,\overleftarrow{x}$, where $i = |R_1|$. Any solution of this equation is of the form $\overleftarrow{x} = (\overleftarrow{A_i})^t\,\overleftarrow{A'}$ for some $t \in I\!N$ and $\overleftarrow{A'} < \overleftarrow{A_i}$ which means that $x = A'(A_i)^t$ where A' is some suffix of A_i. Since $|A_i|=|BR_1|$, we get $R_1BR' = A'A_i$, so $R_1 = R_{i,j}$ where $j = |A'|$.

Step 4.2 follows from proposition 5.10 of [8]. ■

Theorem 4.2 *The algorithm presented above works in polynomial time.*

Proof. This follows from the fact that in each step we create polynomial number of equations of polynomial length and from lemmas 2.3 and 2.8. ■

Remark. For the simplicity of the algorithm we did not take care of its complexity. In fact the complexity of the algorithm we have presented is of order $O(d^{100})$, what suggests that this algorithm is not of practical value. However, this complexity can be improved – namely in most places the number d can be replaced by a much smaller number called by Khmelevskiĭ the characteristics of an equation (see [8] for details).

5 An example

As an aplication of the algorithms given in sections 3 and 4, we shall solve the following problem.

Problem. Find all integers k, l such that the equation

$$(ax)^k = (yb)^l, \tag{8}$$

with coefficients $\Sigma = \{a, b\}$, and variables $\Xi = \{x, y\}$ is solvable. Describe the set of all solutions.

Without any loss of generality we can assume that $k \leq l$, since otherwise we can consider the equivalent equation $(by)^l = (xa)^k$, obtained by reversing the order in (8). Therefore, it suffices to consider only solutions υ, for which $|\upsilon(x)| \geq |\upsilon(y)|$. First, we shall follow the algorithm described in Section 4.

5.1 Reduction of the example to basic equations

Step 1 The equation has the form described by **Case 1.2**, so we follow instructions described in Step 1.2.1 and Step 1.2.2.
Step 1.2.1 We substitute a for y and get the equation

$$(ax)^k = (ab)^l$$

According to lemma 2.3 any solutions of this equation is of the form $\upsilon(x) = (ba)^t$ or $\upsilon(x) = (ba)^t b$. There is no solution of the first form since after substitution the left hand side of the equation ends with the symbol a while the right hand side ends with b. The second form provides a solution if $k(t+1) = l$, so we get the first solution of (8).

Lemma 5.1 *If $k|l$, then $\upsilon(y) = a,\ \upsilon(x) = b(ab)^{\frac{l}{k}-1}$ is a solution of $(ax)^k = (yb)^l$.*

Step 1.2.2 We substitute ay for y and get the equation

$$x(ax)^{k-1} = yb(ayb)^{l-1}$$

as an input for Step 2.
Step 2.1 We substitute y for x and get the

$$(ay)^{k-1} = b(ayb)^{l-1}$$

which is trivially unsolvable.

Since we are looking for solutions such that $|\upsilon(x)| \geq |\upsilon(y)|$, we do not follow Step 2.3. To execute Step 2.2 we go to Step 3.
Step 3 We are going to solve the directed equation $x(ax)^{k-1} \to yb(ayb)^{l-1}$.
Step 3.1 We substitute $(yb(ayb)^{l-1})^t Y$ for x where $Y \in \{\varepsilon, yb(ayb)^s, yb(ayb)^s a : s < l-1\}$. In this way we obtain the equation

$$(yb(ayb)^{l-1})^t Y[a(yb(ayb)^{l-1})^t Y]^{k-1} = yb(ayb)^{l-1}$$

By comparing lengths of both sides of the equation it is easy to check that if the equation is solvable, then $t = 0$. So, it remains to solve the equation $Y(aY)^{k-1} = yb(ayb)^{l-1}$ We consider three cases for different types of Y.

- $Y = \varepsilon$

  We get the equation

  $$a^{k-1} = yb(ayb)^{l-1}$$

  whose reverse is trivially unsolvable

- $Y = yb(ayb)^s$

  We get the equation

  $$yb(ayb)^s[ayb(ayb)^s]^{k-1} = yb(ayb)^{l-1},$$

  equivalent to $[(ayb)^{s+1}]^{k-1} = (ayb)^{l-s-1}$, which is solvable if and only if $(s+1)k = l$, and we get a second solution to (8).

Lemma 5.2 *If $k|l$, then $\upsilon(y) = ay,\ \upsilon(x) = yb(ayb)^{\frac{l}{k}-1}$ is a solution of $(ax)^k = (yb)^l$.*

- $Y = yb(ayb)^s a$

  We get the equation

  $$yb(ayb)^s a[ayb(ayb)^s a]^{k-1} = yb(ayb)^{l-1},$$

  which is equivalent to $a[(ayb)^{s+1}a]^{k-1} = (ayb)^{l-s-1}$. The reverse of the last equation is trivially unsolvable

Step 3.2 We substitute $(yb(ayb)^{l-1})^t Y$ for x where $Y \in \{x, yb(ayb)^s ax : s < l-1\}$. Again we get the equation

$$(yb(ayb)^{l-1})^t Y[a(yb(ayb)^{l-1})^t Y]^{k-1} \leftarrow yb(ayb)^{l-1},$$

and again we conclude that $t = 0$. Therefore the last equation reduces to $Y(aY)^{k-1} \leftarrow yb(ayb)^{l-1}$. Since $t = 0$, the case $Y = x$ gives the degenerated substitution of x for x, and it suffices to consider the case $Y = yb(ayb)^s ax$. Now, our equation has the form

$$yb(ayb)^s ax[ayb(ayb)^s ax]^{k-1} \leftarrow yb(ayb)^{l-1}$$

equivalent to $ax[(ayb)^{s+1}ax]^{k-1} = (ayb)^{l-s-1}$.
Step 3.3 We substitute xy for y and get $[(axyb)^{s+1}ax]^{k-1} = yb(axyb)^{l-s-2}$
Step 3.4 We substitute a for y and get

$$[(axab)^{s+1}ax]^{k-1} = ab(axab)^{l-s-2}.$$

This is equation of the form $x\Phi = bax\Psi$. We have to consider two cases

- Substitution of $(ba)^t$ for x gives a trivially unsolvable equation
- Substitution of $(ba)^t b$ for x gives the equation

  $$(ab)^{[(t+2)(s+1)+t+1](k-1)} = (ab)^{1+(t+2)(l-s-2)},$$

  which is solvable if and only if $k[(t+2)(s+1)+t+1] = l(t+2)$. Composing all the substitutions we have used, we get the third answer to our problem:

Lemma 5.3 *If there exist integers t, s, such that $k[(t+2)(s+1)+t+1] = l(t+2)$ and $s < l$ then, $\upsilon(y) = (ab)^{t+1}a$, $\upsilon(x) = b(ab)^{(t+2)(s+1)+t}$ is solution of $(ax)^k = (yb)^l$.*

Step 3.5 We substitute ay for y and get $[(axayb)^{s+1}ax]^{k-1} = ayb(axayb)^{l-s-2}$ as input of Step 4
Step 4.1 We have to solve equation of the form $xay\Phi = ybax\Psi$, so $c = 1$, $A_1 = ba$, $R_{1,0} = b$, $R_{1,1} = a$. Again, we consider two cases:

- Substitution of a for x gives an equation whose reverse is trivially unsolvable
- Substitution of b for x gives

  $$[(abayb)^{s+1}ab]^{k-1} = ayb(abayb)^{l-s-2}$$

  which is of the form $bay\Phi = y\Psi$. There are two possibilities of choosing solution:

  - Substitution of $(ba)^t b$ for y gives a trivially unsolvable equation
  - Substitution of $(ba)^t$ for y gives

    $$(ab)^{[(t+2)(s+1)+1](k-1)} = (ab)^{t+1+(t+2)(l-s-2)}$$

    which is solvable if and only if $k[(t+2)(s+1)+1] = l(t+2)$. Composing all the substitutions we have used, we get the fourth answer to our problem:

Lemma 5.4 *If there exist numbers t, s, such that $k[(t+2)(s+1)+1] = l(t+2)$ and $s < l$, then $\upsilon(y) = (ab)^{t+1}a$, $\upsilon(x) = b(ab)^{(t+2)(s+1)}$ is a solution of $(ax)^k = (yb)^l$.*

Step 4.2 The only possible decomposition is $P = S = \varepsilon$, $Q = A = ba$, so $R = b$, and we substitute xb for x. We get the equation

$$[(axbayb)^{s+1}axb]^{k-1} = ayb(axbayb)^{l-s-2}$$

of the form $xbay\Phi = ybax\Psi$, so we follow Case 1 of Algorithm 3.1.

5.2 Applying the algorithm for basic equations

Step 1.1 The only possible choice of Z, Z_1, Z_2 is $Z = ba,\ Z_1 = Z_2 = \varepsilon$. We substitute $(ba)^n$ for x and $(ba)^m$ for y. This substitution gives the equation

$$[(a(ba)^n ba(ba)^m b)^{s+1} a(ba)^n b]^{k-1} = a(ba)^m b[a(ba)^n ba(ba)^m b]^{l-s-2}$$

equivalent to $[(ab)^{(n+m+2)(s+1)+n+1}]^{k-1} = (ab)^{m+1+(n+m+2)(l-s-2)}$ which is solvable if and only if $k((n+m+2)(s+1)+n+1) = l(n+m+2)$. Here we get the fifth answer to our problem:

Lemma 5.5 *If there exist integers n, m, s such that $k((n+m+2)(s+1)+n+1) = l(n+m+2)$ and $s < l,\ n, m > 0$, then $\upsilon(y) = (ab)^{n+m+1}a,\ \upsilon(x) = b(ab)^{(n+m+2)(s+1)+n}$ is a solution of $(ax)^k = (yb)^l$.*

Steps 1.2 and **1.3** We substitute $(zba)^n z$ for x and $(zba)^m z$ for y thus obtaining the equation

$$[(azb)^{(n+m+2)(s+1)+n+1}]^{k-1} = (azb)^{m+1+(n+m+2)(l-s-2)},$$

which is solvable if and only if $k((n+m+2)(s+1)+n+1) = l(n+m+2)$. The last answer to our problem is:

Lemma 5.6 *If there exist nonnegative integers n, m, s, such that $k((n+m+2)(s+1)+n+1) = l(n+m+2)$ and $s < l'$ then $\upsilon(y) = (azb)^{n+m+1}az,\ \upsilon(x) = zb(azb)^{(n+m+2)(s+1)+n}$ is solution of equation $(ax)^k = (yb)^l$.*

Now we can recapitulate our solution:

Theorem 5.7 *Equation $(ax)^k = (yb)^l$ is solvable if and only if one of the four conditions below is satisfied:*

- *$k|l$ (solution: $y = az, x = zb(azb)^{\frac{l}{k}-1}$ for any $z \in \Sigma^*$)*
- *$l|k$ (solution: $x = az, y = zb(azb)^{\frac{k}{l}-1}$ for any $z \in \Sigma^*$)*
- *there exist nonnegative numbers n, m, s such that $k((n+m+2)(s+1)+n+1) = l(n+m+2)$ and $s < l$ (solution: $y = (azb)^{n+m+1}az, x = zb(azb)^{(n+m+2)(s+1)+n}$ for any $z \in \Sigma^*$)*
- *there exist nonnegative numbers n, m, s such that $l((n+m+2)(s+1)+n+1) = k(n+m+2)$ and $s < k$ (solution: $x = (azb)^{n+m+1}az, y = zb(azb)^{(n+m+2)(s+1)+n}$ for any $z \in \Sigma^*$)*

References

[1] H. Abdulrab, Résolution d'équations sur les mots: étude et implémentation LISP de l'algorithme de Makanin. *Ph.D. Thesis, Université de Rouen*, 1987.

[2] H. Abdulrab, J.P. Pecuchet, Solving word equations. *Journal of Symbolic Computation* 8 (1989), pp. 499-521.

[3] S.I. Adyan, G.S. Makanin, Investigation on algorithmic questions of algebra. *Trudy Matem. Inst. Steklova* 168 (1984), English translation in *Proc. of Steklov Institute of Mathematics* 1986, issue 3, pp. 207-226.

[4] V.K. Bulitko, Equations and inequalities in a free group and a free semigroup. (in Russian) *Tul. Gos. Ped. Inst. Ucen. Zap. Mat. Kafedr Vyp 2, Geometr. i Algebra.* (1970), pp.242-252.

[5] W. Charatonik, Equations in Free Semigroup (in Polish),*Master Thesis, Wrocław University* 1991 (unpublished).

[6] J. Jaffar, Minimal and Complete Word Unification. *Journal of ACM* 37 (1990), pp.47-85.

[7] Yu.I. Khmelevskiĭ, Solution of word equations in three unknowns. *Dokl. Akad. Nauk SSSR* 177 (1967), 1023-1025; English transl. in *Soviet Math. Dokl.* 8 (1967).

[8] Yu.I. Khmelevskiĭ, Equations in a Free Semigroup (in Russian), *Trudy Matem. Inst. Steklova*, 107 (1971) pp. 1–284.

[9] A. Kościelski, L. Pacholski, Complexity of Unification in Free Groups and Free Semigroups, *Proceedings 31st Annual Symposium on Foundations of Computer Science*, Los Alamitos 1990, vol. II, pp.824–830.

[10] A. Kościelski, L. Pacholski, Complexity of Makanins' Algorithms *submitted*.

[11] A. Lentin, Equations in free monoids. in Nivat, editor, *Automata, Languages and Programing*, Amsterdam 1972, pp. 67-85.

[12] G.S. Makanin, The Problem of Solvability of Equations in a Free Semigroup. (in Russian), *Matematicheskiĭ Sbornik* 103 (1977), pp. 147-236; English translation in *Math. USSR Sbornik* 32 (1977), pp. 129-198.

[13] J.-P. Pecuchet, Solutions principales et rang d'un système d'équations avec constantes dans le monoïde libre. *Discrete Mathematics* 48 (1984), pp.253-274.

[14] G. Plotkin, Building in equational theories. *Machine Intelligence* 7 (1972), pp. 73-90.

[15] K.U. Schulz, Makanin's algorithm - two improvements and a generalization, *CIS-report* 91-39, Centrum für Informations- und Sprachverarbeitung, University of Munique, 1991.