Gleichungen mit regulären Randbedingungen über freien Gruppen

Von der Fakultät Informatik der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Naturwissenschaften (Dr.rer.nat.) genehmigte Abhandlung

Vorgelegt von

Christian Hagenah

aus Rendsburg

Hauptberichter :	Prof. Dr. V. Diekert
Mitberichter :	Prof. Dr. C. Claus
	Prof. Dr. A. Muscholl

Tag der mündlichen Prüfung : 15. Juni 2000

Institut für Informatik der Universität Stuttgart
2000

Inhaltsverzeichnis

Zusammenfassung

Wir beweisen, daß das Erfüllbarkeitsproblem für Gleichungen mit regulären Randbedingungen über freien Gruppen PSPACE-vollständig ist. Wir zeigen auch, daß eine minimale Lösung einer solchen Gleichung höchstens eine doppelt exponentielle Länge hat und in 2-DEXPTIME berechnet werden kann.

Wir reduzieren zuerst das Problem Gleichungen mit regulären Randbedingungen über einer freien Gruppen zu lösen auf das Problem Gleichungen mit regulären Randbedingungen über freien Monoiden mit einer Anti-Involution zu lösen. Anschließend stellen wir einen Algorithmus vor, der in PSPACE entscheidet, ob diese Gleichungen lösbar sind und einen Algorithmus, der in 2-DEXPTIME eine Lösung berechnet, wenn die Gleichung lösbar ist.

Abstract

We prove that the satisfiability problem for equations with regular constraints over free groups is PSPACE-complete. We also show that a minimal solution of such an equation has at most a double exponential length and can be computed in 2-DEXPTIME.

We first reduce the problem to solve equations with regular constraints over free groups to the problem to solve equations with regular constraints over monoids with an anti-involution. Then we present an algorithm that decides in PSPACE whether these equations are solvable and an algorithm that computes a solution in 2-DEXPTIME if the equation in solvable.

KAPITEL 1

Extended Abstract

1.1. Introduction

Makanin proved in 1977 [**M77**] that the solvability of equations with constants over free monoids is decidable. The complexity of Makanin's algorithm was improved several times: 4-NEXPTIME [**S92**] (composition of four exponential functions), 3-NEXPTIME [**KP96**], 2-EXPSPACE [**D98**] and EXPSPACE [**G98**]. Makanins algorithm was implemented [**A87**] and a current version can be found in [**D00**]. Plandowski proved that the problem is in NEXPTIME [**P99A**] and PSPACE [**P99B**] decidable.

Schulz generalized Makanin's algorithm and proved that the solvability of equations with regular constraints over free monoids is also decidable [**S92**]. The complexity improvements for Makanins algorithm cited above applied to the generalized version, too. In [**P99B**] Plandowski announced a PSPACE algorithm of Rytter that decides, whether an equation with regular constraints over a free monoid has a solution.

Makanin proved in 1982 [**M82**] and in 1984 [**M84**] that the solvability of equations with constants over free groups is decidable. The running time of the algorithm was shown to be not primitive recursive [**KP98**]. The inherent complexity of the problem remained however unclear and Gutiérrez published 2000 [**G00B**] a PSPACE algorithm that used the new ideas of Plandowski.

In the present dissertation we will consider equations over free monoids with an anti-involution (FMA-equations). This approach is more general since we can reduce equations over free groups (FG-equations) to FMA-equations effectively. We generalize the results in [**PR98**], [**P99A**], and [**P99B**] and extend them to the case where we allow regular constraints. The overall strategy is to follow the method of Plandowski and Gutiérrez.

But each step has to be generalized and modified in order to cope with the presence of regular constraints.

In this extended abstract we list the essential lemmas and theorems which lead to the main results. All details are omitted for lack of space.

1.2. Definitions

Let $\Sigma = \{a, b, \dots\}$ be a finite alphabet of constants and $\overline{\ } : \Sigma^* \to \Sigma^*$ be an anti-involution with fixed points. It is $\overline{\overline{a}} = a$ for all $a \in \Sigma$, $I := \{a \in \Sigma \mid a = \overline{a}\}$ the set of fixed points, and $\overline{A} := \{\overline{a} \mid a \in A\}$ is the set of inverse constants for a set $A \subseteq \Sigma$. For a word $w = a_1 \cdots a_\ell$ with $a_i \in \Sigma$ we denote by $|w| := \ell$ the length of the word, by $\overline{w} := \overline{a_\ell} \cdots \overline{a_2}\,\overline{a_1}$ the inverse word, and by $w[\alpha, \beta]$ the subword $a_{\alpha+1}a_{\alpha+2} \cdots a_\beta$, if $\alpha \leq \beta$, and $\overline{w[\beta, \alpha]}$, if $\alpha > \beta$. The positions α and β are the borders of the subword. Let ϵ be the empty word.

Let $\Omega = \{X, Y, \dots\}$ be a finite set of variables. For each variable X exists an inverse variable $\overline{X}$ with $\overline{\overline{X}} = X$ and $\overline{X} \neq X$. By $\Omega_{\frac{1}{2}}$ we denote a subset of Ω with $\Omega_{\frac{1}{2}} \cup \overline{\Omega_{\frac{1}{2}}} = \Omega$ and $\Omega_{\frac{1}{2}} \cap \overline{\Omega_{\frac{1}{2}}} = \emptyset$.

An equation with regular constraints over a free monoid with an anti-involution with fixed points (FMA-equation with regular constraints) is an equation $E : L = R$ with $L, R \in (\Sigma \cup \Omega)^*$ together with the regular constraints $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$ for the variables $X \in \Omega$. A_X is a non-deterministic finite automata that recognizes the language $L(A_X)$. Let $d := |L| + |R|$ be the denotational length of the equation E. A solution of the equation E is a homomorphism $\sigma : (\Sigma \cup \Omega)^* \to \Sigma^*$ with $\sigma(a) = a$ for all $a \in \Sigma$, $\sigma(\overline{X}) = \overline{\sigma(X)}$ for all $X \in \Omega$, $\sigma(L) = \sigma(R)$ and $\sigma(X) \in L(A_X)$ for all $X \in \Omega$. A solution σ is minimal if $|\sigma(L)| \leq |\sigma'(L)|$ for all solutions σ'.

A formula with regular constraints over FMA-equations is a boolean formula with FMA-equations as atomic formulae.

LEMMA 1.2.1. *For each formula with regular constraints over FMA-equations of the size n exists an equivalent single FMA-equation with regular constraints of the size $\mathcal{O}(n^3)$.*

We calculate a monoid M and a homomorphism $h : \Sigma^* \to M$ that recognizes the regular constraints A_X, i.e. for each variable X exists a subset $M_X \subseteq M$ such that $h^{-1}(M_X) = L(A_X)$.

LEMMA 1.2.2. *The monoid M can be chosen such that*

$$|M| \leq 2^{\sum_{X \in \Omega} |Q_X|^2} \leq 2^{n^2}$$

and a constant

$$1 \leq c(M) \leq max\{|Q_X| \mid X \in \Omega\}!$$

exists such that $\forall m \in M : m^{c(M)} = m^{2c(M)}$.

For all pairs of monoid elements $m_1, m_2 \in M$ we fix a shortest word $\min_{m_1,m_2} \in \Sigma^*$ such that $h(\min_{m_1,m_2}) = m_1$ and $h(\overline{\min_{m_1,m_2}}) = m_2$. If no such word exists $\min_{m_1,m_2}$ is undefined. We set

$$\max_M := \max\{|\min_{m_1,m_2}| \mid m_1, m_2 \in M \wedge \min_{m_1,m_2} \text{ defined}\}.$$

LEMMA 1.2.3. *Is is $max_M < |M|^2$.*

1.3. Equations Over Free Groups

Let G be a free group that is generated by the generators $S = \{a, b, \dots\}$. Let the constants $\Sigma := S \cup \overline{S} = \{a, \overline{a}, b, \overline{b}, \dots\}$ be the union of the generators S and their inverses $\overline{S} = \{\overline{a}, \overline{b}, \dots\}$. It is $\overline{\overline{a}} = a$ and $\overline{a} \neq a$ for all $a \in \Sigma$. We write $v \to w$ if $v = ra\overline{a}s$ and $w = rs$ with $a \in \Sigma$ and $r, s \in \Sigma^*$. Let $\approx$ be the equivalence relation over $\to$. Each equivalence class corresponds to an element of the free group G.

For the FG-equations the regular constraints are rational languages. We specify them through non-deterministic finite automata $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$. We can see the symbols $\Sigma \cup \{\epsilon\}$ of the transitions as elements of the free group or as elements of a free monoid. By $L(A_X)$ we denote the language that is recognized by A_X where the symbols are interpreted as elements of a free monoid.

Let $\Omega = \{X, Y, \dots\}$ be the finite set of variables. For each variable X exists an inverse variable $\overline{X}$ with $\overline{\overline{X}} = X$ and $\overline{X} \neq X$. By $\Omega_{\frac{1}{2}}$ we denote a subset of Ω with $\Omega_{\frac{1}{2}} \cup \overline{\Omega_{\frac{1}{2}}} = \Omega$ and $\Omega_{\frac{1}{2}} \cap \overline{\Omega_{\frac{1}{2}}} = \emptyset$.

An equation with regular constraints over a free group (FG-equation with regular constraints) is an equation $E : L \approx R$ with $L, R \in (\Sigma \cup \Omega)^*$ together with the regular constraints $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$. Let $d := |L| + |R|$ be the denotational length of the equation. A solution of the equation E is a homomorphism $\sigma : (\Sigma \cup \Omega)^* \to \Sigma^*$ with $\sigma(a) = a$ for all $a \in \Sigma$, $\sigma(\overline{X}) = \overline{\sigma(X)}$ for all $X \in \Omega$, $\sigma(L) \approx \sigma(R)$, and $\sigma(X) \in L(A_X)$ for all $X \in \Omega$.

A formula with regular constraints over FG-equations is a boolean formula with FG-equations as atomic formulae.

LEMMA 1.3.1. *For each formula with regular constraints over FG-equations of the size n exists an equivalent single FG-equation with regular constraints of the size $\mathcal{O}(n^3)$.*

The following theorem allows us to calculate in non-deterministic polynomial time for a FG-equation E with regular constraints a FMA-equation E_i with regular constraints such that E_i has a solution σ_i if and only if E has a soluton σ.

THEOREM 1.3.2. *Let $E : L \approx \epsilon$ be a FG-equation with $L \in \Omega^*$, the regular constraints $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$ for the variables $X \in \Omega_{\frac{1}{2}}$, and $d \geq 2$. Then FMA-equations $E_1 : L_1 = R_1, E_2 : L_2 = R_2, \dots, E_k : L_k = R_k$ with regular constraints $A'_X = (Q_X, \Sigma, \delta'_X, I_X, F_X)$ exist such that*

1. $|L_i| + |R_i| \leq 14d - 16$, $k \in 2^{\mathcal{O}(d \log(d))}$ *and* $|\delta'_X| \leq |\delta_X| + |Q_X|^2$,
2. *if a solution σ exists for E then a solution σ_i exists for one of the equations E_i, and*
3. *if a solution σ_i exists for one of the equations E_i then a solution σ exists for E with $|\sigma(L)| \leq 2\, 2^{n^2}(|\sigma_i(L_i)| + n)$.*

1.4. The Exponent of Periodicity

DEFINITION 1.4.1. Let $w \in \Sigma^*$. The exponent of periodicity $\exp(w)$ of the word w is

$$\exp(w) := \max\{i \mid \exists r, s \in \Sigma^*, p \in \Sigma^+ : w = rp^i s\}.$$

We can bound the exponent of periodicity for minimal solutions.

THEOREM 1.4.2. *Let $E : L = R$ be a FMA-equation with regular constraints $M_X \subseteq M$ and σ be a minimal solution for E. Then $exp(\sigma(L)) \in \mathcal{O}(c(M)2^{1,6d})$.*

1.5. Free Intervals

DEFINITION 1.5.1. Let $E : L = R$ be an equation with $L = X_{i_1}X_{i_2}\cdots X_{i_c}$, $R = X_{i_{c+1}}X_{i_{c+2}}\cdots X_{i_d}$, and $\Omega = \{X_1, X_2, \ldots, X_{|\Omega|}\}$ and let $\delta_j := |\sigma(X_{i_1}X_{i_2}\cdots X_{i_{j-1}})|$ for $1 \leq j \leq c$ and $\delta_j := |\sigma(X_{i_{c+1}}X_{i_{c+2}}\cdots X_{i_{j-1}})|$ for $c+1 \leq j \leq d$. Then

$$[\delta_j + \alpha, \delta_j + \beta] \sim [\delta_k + \mu, \delta_k + \nu],$$

if $1 \leq j, k \leq d$, $0 \leq \alpha, \beta \leq |\sigma(X_{i_j})|$, $\alpha \neq \beta$, and

- $X_{i_j} = X_{i_k}$, $\alpha = \mu$, and $\beta = \nu$ or
- $X_{i_j} = \overline{X_{i_k}}$, $\alpha = |\sigma(X_{i_j})| - \mu$, and $\beta = |\sigma(X_{i_j})| - \nu$

By $\approx$ we denote the equivalence relation over $\sim$.

DEFINITION 1.5.2. An interval $[\alpha, \beta]$ is free if $[\mu, \nu] \approx [\alpha, \beta]$ implies that there are not cuts proper in the interval $[\mu, \nu]$. A free intervall $[\alpha, \beta]$ is maximal, if no free intervall $[\mu, \nu]$ exists such that $[\alpha, \beta]$ is a proper sub interval of $[\mu, \nu]$.

LEMMA 1.5.3. *There are at most $2d - 2$ different equivalence classes of maximal free intervals.*

Because all intervals in the same equivalence class correspond with the same word or its inverse and the word $\sigma(L)$ consist of maximal free intervals, the

word $\sigma(L)$ is a concatenation of at most $2d-2$ different words and each of these words corresponds to an equivalence class of a maximal free interval.

LEMMA 1.5.4. *Let σ be a minimal solution and $[\alpha, \beta]$ be a free interval. Then $|\beta - \alpha| \leq 2max_M + 1 < 2|M|^2$.*

1.6. ℓ-Factorisations

If a variable X starts at the position α in $\sigma(L)$ or $\sigma(R)$ then α and $\alpha+|\sigma(X)|$ are the borders of the variable. The borders $0 = \gamma_1 < \gamma_2 < \ldots < \gamma_k = |\sigma(L)|$ of all variables in $\sigma(L)$ and $\sigma(R)$ are the cuts.

The words with the length 2ℓ with $\ell > 0$ that appear over the cuts and their inverses are the characteristic words C_ℓ:

$$C_\ell := \{v \mid \exists 1 \leq i \leq k : \sigma(L)[\gamma_i - \ell, \gamma_i + \ell] = v \vee \sigma(L)[\gamma_i - \ell, \gamma_i + \ell] = \overline{v}\}.$$

A factorisation is a function

$$F : \Sigma^* \to (D \times \Sigma^+ \times D)^*$$

where D is an arbitrary set. The ℓ-factorisation F_ℓ of a word w is defined by the characteristic words C_ℓ. Let w be a word and $0 < \mu_1 < \mu_2 < \ldots < \mu_k < |w|$ be all positions over which a word from C_ℓ appears in w. It is

$$\{\mu_i \mid 1 \leq i \leq k\} = \{\mu \mid w[\mu - \ell, \mu + \ell] \in C_\ell\}.$$

Let $v_i := w[\mu_i - \ell, \mu_i + \ell]$ be the word from C_ℓ that appears over μ_i. By

$$F_\ell(w) := (\$, w[0, \mu_1], v_1)(v_1, w[\mu_1, \mu_2], v_2) \cdots (v_k, w[\mu_k, |w|], \$)$$

we denote the ℓ-factorisation of the word w where \$ is a new symbol with $\overline{\$} = \$$.

1.7. ℓ-Transformations

The ℓ-Transformation T_ℓ assigns each tripel $(E : L = R, M_\Omega, \sigma)$ a quadrupel $(E_\ell : L_\ell = R_\ell, M_{\Omega_\ell}, \sigma_\ell, f_\ell)$ that consist of an equation $E_\ell : L_\ell = R_\ell$, a solution σ_ℓ for this equation, the regular constraints M_{Ω_ℓ} and a function $f_\ell : ((\Sigma_\ell \cup \Omega_\ell)^* \to \Sigma_\ell^*) \to ((\Sigma \cup \Omega)^* \to \Sigma^*)$. The constants of the equation

E_ℓ are a finite subset of $D_\ell \times \Sigma^+ \times D_\ell$. The set of constants Σ_ℓ is the union of the constants that appear in $E_\ell : L_\ell = R_\ell$, their inverses, and the set $D_\ell \times \Sigma^{\geq 1 \wedge \leq 2\max_M + 1} \times D_\ell$.

Because there are at most exponential many non isomorphic equations E_ℓ and all isomorphic equations are near to each other the following theorems hold.

THEOREM 1.7.1. *Let E be a FMA-equation with regular constraints and σ be a minimal solution. Then $|\sigma(L)| \leq b^{2p} \in 2^{2^{\mathcal{O}(n^5)}}$ with $b := 4d(3 + exp(\sigma(L))) + 2max_M$ and $p := (32d^3 + |M| + exp(\sigma(L)))^{128d^3}$, i.e. the length of a minimal solution of a FMA-equation with regular constraints is double exponential bounded.*

THEOREM 1.7.2. *Let E be a FG-equation with regular constraints and σ be a minimal solution for E. Then $|\sigma(L)| \in 2^{2^{\mathcal{O}(n^{10})}}$, i.e. the length of a minimal solution of a FG-equation with regular constraints is double exponential bounded.*

1.8. The PSPACE Algorithm

To decide whether or not the equation $E : L = R$ has eine solution, we construct an equation $E_M : L = R$. The only difference of the equations E and E_M is the alphabet of constants

$$\Sigma_M := \{(h(w), h(\overline{w}), s) \mid w \in \Sigma^* \wedge (w = \overline{w} \rightarrow s = 0) \wedge (w \neq \overline{w} \rightarrow s = \pm 1)\}$$

with $\overline{(m_1, m_2, s)} := (m_2, m_1, -s)$. Note that each pair of monoid elements can be represented by a single constant in E_M, i.e. $\max_M = 1$.

LEMMA 1.8.1. *The equation E_M ist solvable if and only if the equation E is solvable.*

In the following we consider E_M instead of E. Intuitively the PSPACE algorithm searches all over a graph that contains only solvable equations. It starts with the equation $E_\ell : (\$, \sigma(L), \$) = (\$, \sigma(L), \$)$ and searches for a special equation E_0 which is similar to the equation E_M. If E_M is solvable then the algorithm will find E_0, because the path $E_{|\sigma(L)|} \rightarrow E_{|\sigma(L)|-1} \rightarrow$

$\cdots \to E_1 \to E_0$ exists. The critical property of $E \to E'$ is that if E is solvable then E' is solvable, too.

THEOREM 1.8.2. *The problem whether a formula with regular constraints over FMA-equations has a solution is PSPACE complete.*

THEOREM 1.8.3. *The problem whether a formula with regular constraints over FG-equations has a solution is PSPACE complete.*

1.9. Compression of Minimal Solutions

DEFINITION 1.9.1. A grammar is definite if the grammar is acyclic and for each nonterminal exists exactly one production with the nonterminal on the left side.

LEMMA 1.9.2. *Let σ be a minimal solution. For all $X \in \Omega$ exists a definite grammar in Chomsky normalform of the size*

$$\mathcal{O}(d^2(max_M + log(|\sigma(L)|))^2)$$

which generates exactly the word $\sigma(X)$.

THEOREM 1.9.3. *In NEXPTIME can be decided whether a formula with regular constraints over FMA-equations is solvable.*

Since the length of the solution can be double exponential it cannot be outputed in NEXPTIME. But this is possible in 2-DEXPTIME.

THEOREM 1.9.4. *A solution for a formula with regular constraints over FMA-equations can be computed in 2-DEXPTIME if a solution exists.*

The same holds for free groups.

THEOREM 1.9.5. *A solution for a formula with regular constraints over FG-equations can be computed in 2-DEXPTIME is a solution exists.*

THEOREM 1.9.6. *If the length of a minimal solution is exponential bounded and the monoid M is small, i.e. $|M| \in n^{\mathcal{O}(1)}$, then it can be decided in NP whether a formula with regular constraints over FMA-equations has a solution.*

Thus it is sufficient to prove that the length of a minimal solution is exponential bounded to show that the problem is NP-complete.

1.10. Solutions With a Given Length N

In this section N will be a part of the input (binary coded). Assume that a solution exists with the length N. We show that if σ is a solution with $|\sigma(L)| = N$ then a solution σ' exists with $|\sigma'(L)| = N$ such that σ' is highly compressible. The idea is that all maximal free intervals are replaced by highly compressible words.

LEMMA 1.10.1. *Let w be a word, M a monoid, and $h : \Sigma^* \to M$ a homomorphism. Then a word v exists with $|v| = |w|$, $h(v) = h(w)$, $h(\overline{v}) = h(\overline{w})$, $w = \overline{w} \Rightarrow v = \overline{v}$ and v is generated by a definite grammar in Chomsky normalform of the size $\mathcal{O}(|M|^4 log(|w|) + |M|^6)$.*

LEMMA 1.10.2. *Let $E : L = R$ be a FMA-equation with regular constraints. If σ is a solution with $|\sigma(L)| = N$ then a solution σ' exists with $|\sigma'(L)| = N$ such that $\sigma'(L)$ is generated by a grammer in Chomsky normalform of the size $\mathcal{O}(d^2|M|^8(log(N) + |M|^2)^2)$.*

THEOREM 1.10.3. *Whether a FMA-equation with regular constraints has a solution σ with $|\sigma(L)| = N$ can be decided in NEXPTIME. If the monoid M is small, i.e. $|M| \in n^{\mathcal{O}(1)}$, then it can be decided in NP.*

THEOREM 1.10.4. *Let $E : L = R$ be a FMA-equation with regular constraints. A solution σ with $\sigma(L) = N$ can be computed in 2-DEXPTIME if such a solution exists. If the monoid M is small, i.e. $|M| \in n^{\mathcal{O}(1)}$, then the solution can be computed in DEXPTIME.*

1.11. Conclusions

Consider the problem whether a FMA-equation without regular constraints has a solution. It is open whether the problem is NP- or PSPACE-complete. On the one side it is NP hard and on the other side we have a PSPACE algorithm. If we consider the length of a minimal solution we come to the same question. The length is only double exponential bounded but it is no

example known with an over exponential length. If it is possible to bound the length exponential then the problem is NP-complete.

KAPITEL 2

Einleitung

Makanin zeigte 1977 [**M77**], daß die Lösbarkeit von Gleichungen ohne reguläre Randbedingungen über freien Monoiden entscheidbar ist. Die Komplexität von Makanins Algorithmus wurde immer wieder verbessert: 4-NEXPTIME [**S92**] (eine Verkettung von 4 exponentiellen Funktionen), 3-NEXPTIME [**KP96**], 2-EXPSPACE [**D98**] und EXPSPACE [**G98**]. Makanins Algorithmus wurde auch implementiert [**A87**] und eine aktuelle Version findet man in [**D00**]. Mit einem neuen Ansatz zeigte Plandowski, daß das Problem in NEXPTIME [**P99A**] und PSPACE [**P99B**] entscheidbar ist.

Schulz verallgemeinerte Makanins Algorithmus und bewies, daß auch die Lösbarkeit von Gleichungen mit regulären Randbedingungen über freien Monoiden entscheidbar ist [**S92**]. Die oben zitierten Komplexitätsverbesserungen an Makanins Algorithmus ließen sich auch jeweils auf den verallgemeinerten Algorithmus übertragen. In [**P99B**] kündigt Plandowski einen PSPACE Algorithmus an, der entscheidet, ob eine Gleichungen mit regulären Randbedingungen über einem freien Monoid eine Lösung hat.

Makanin zeigte auch als erster 1982 [**M82**] (eine Ergänzung erfolgte 1984 [**M84**]), daß die Lösbarkeit von Gleichungen ohne reguläre Randbedingungen über freien Gruppen entscheidbar ist. Die Laufzeit des Algorithmus ist jedoch nicht primitiv rekursiv [**KP98**]. Gutiérrez veröffentlichte 2000 [**G00B**] einen PSPACE Algorithmus, bei dem er die neuen Ideen von Plandowski weiterentwickelte.

Für einige Spezialfälle ohne reguläre Randbedingungen über einem freien Monoid gibt es effiziente Algorithmen. Ob eine Gleichung mit nur zwei Variablen lösbar ist, kann in $\mathcal{O}(n^6)$ entschieden werden [**IP99**]. Ob eine Gleichung, in der jede Variable höchstens zweimal vorkommt und bei der die

Längen der Variablen vorgegeben sind, lösbar ist, kann in $\mathcal{O}(n)$ entschieden werden [**RD99**].

Außer der Erweiterung von Gleichungen um reguläre Randbedingungen, werden häufig auch Ausdrücke über Gleichungen betrachtet. Die bekannten Transformation eines Ausdrucks ohne reguläre Randbedingungen über Gleichungen über einem freien Monoid in eine einzelne Gleichung ohne reguläre Randbedingungen führen jedoch zu einer exponentiell größeren Gleichung [**KPM97**].

Wortgleichungen werden häufig benutzt, um Unifikationsprobleme auf sie zu reduzieren [**BS96**], [**GV97**], [**F88**]. Man kann auch Wortgleichungen auf Unifikationsprobleme reduzieren [**DMV96**]. In Prolog III können Variablen über dem Anfang und dem Ende eines Wortes liegen. Dies entspricht wieder dem Lösen von Wortgleichungen.

Anstelle von Gleichungen über freien Monoiden (FM-Gleichungen) werden wir Gleichungen über freien Monoiden mit einer Anti-Involution mit Fixpunkten betrachten (FMA-Gleichungen). Wir können nämlich die Gleichungen über freien Gruppen (FG-Gleichungen) auf die FMA-Gleichungen zurückführen. Im weiteren werden wir u.a. die folgenden Theoreme beweisen:

1. Die Länge einer minimalen Lösung einer FG-Gleichung mit regulären Randbedingungen ist doppelt exponentiell beschränkt.
2. Die Länge einer minimalen Lösung einer FMA-Gleichung mit regulären Randbedingungen ist doppelt exponentiell beschränkt.
3. Das Problem, ob ein Ausdruck mit regulären Randbedingungen über FG-Gleichungen eine Lösung hat, ist PSPACE-vollständig.
4. Das Problem, ob ein Ausdruck mit regulären Randbedingungen über FMA-Gleichungen eine Lösung hat, ist PSPACE-vollständig.
5. In 2-DEXPTIME kann eine Lösung für einen Ausdruck mit regulären Randbedingungen über FG-Gleichungen berechnet werden, wenn es eine Lösung gibt.
6. In 2-DEXPTIME kann eine Lösung für einen Ausdruck mit regulären Randbedingungen über FMA-Gleichungen berechnet werden, wenn es eine Lösung gibt.

7. Es ist in NEXPTIME entscheidbar, ob eine FMA-Gleichung mit regulären Randbedingungen eine Lösung der Länge N hat, wobei N ein Teil der Eingabe ist.

Hierzu werden wir vor allem die Ideen aus [**PR98**], [**P99A**] und [**P99B**] weiterentwickeln.

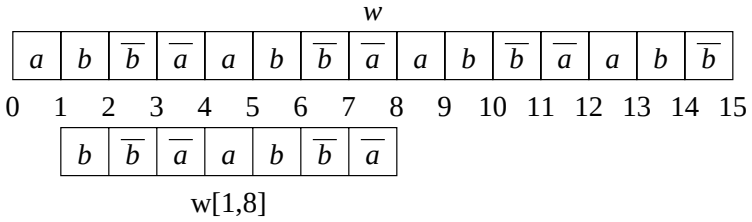
KAPITEL 3

FMA-Gleichungen

3.1. Definitionen

Mit $\Sigma = \{a, b, \dots\}$ bezeichnen wir das Alphabet der Konstanten. Es sei $\overline{\ } : \Sigma^* \to \Sigma^*$ eine Anti-Involution mit Fixpunkten. Es ist $\overline{\overline{a}} = a$ für alle $a \in \Sigma$ und $I := \{a \in \Sigma \mid a = \overline{a}\}$ die Menge der Fixpunkte. Für eine Menge A ist $\overline{A} := \{\overline{a} \mid a \in A\}$ die Menge der inversen Elemente.

Für ein Wort $w = a_1 \cdots a_\ell$ mit $a_i \in \Sigma$ bezeichnen wir mit $|w| := \ell$ die Länge des Wortes, mit $\overline{w} := \overline{a_\ell} \cdots \overline{a_2}\,\overline{a_1}$ das inverse Wort und mit $w[\alpha, \beta]$ das Teilwort $a_{\alpha+1}a_{\alpha+2} \cdots a_\beta$, wenn $\alpha \leq \beta$ ist, und $\overline{w[\beta, \alpha]}$, wenn $\alpha > \beta$ ist. Die Positionen α und β sind die Ränder des Teilwortes. Wir benutzen den Begriff Teilwort und nicht Faktor, da wir später, wenn wir Faktorisierungen einführen, noch von Faktoren sprechen werden. Die griechischen Buchstaben werden wir im weiteren für Positionen innerhalb von Wörtern benutzen. Das leere Wort sei ϵ.



Mit $\Omega = \{X, Y, \dots\}$ bezeichnen wir die Menge der Variablen. Zu jeder Variablen X muß es eine inverse Variable $\overline{X}$ mit $\overline{\overline{X}} = X$ und $\overline{X} \neq X$ geben. Mit $\Omega_{\frac{1}{2}}$ bezeichnen wir eine Teilmenge von Ω, für die $\Omega_{\frac{1}{2}} \cup \overline{\Omega_{\frac{1}{2}}} = \Omega$ und $\Omega_{\frac{1}{2}} \cap \overline{\Omega_{\frac{1}{2}}} = \emptyset$ gilt. Eine Gleichung mit regulären Randbedingungen über einem freien Monoid mit einer Anti-Involution mit Fixpunkten (kurz FMA-Gleichung mit regulären Randbedingungen) ist eine Gleichung

$E : L = R$ mit $L, R \in (\Sigma \cup \Omega)^*$ zusammen mit den regulären Randbedingungen $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$ für die Variablen $X \in \Omega$. A_X ist ein nichtdeterministischer endlicher Automat mit den Zuständen Q_X, dem Alphabet Σ, den Übergängen $\delta_X \subseteq Q_X \times (\Sigma \cup \epsilon) \times Q_X$, den Startzuständen $I_X \subseteq Q_X$ und den Endzuständen $F_X \subseteq Q_X$, der die reguläre Randbedingung für die Variable X beschreibt. Der Automat A_X muß nicht vollständig sein. Es muß jedoch für jede Variable ein Automat angegeben werden. Mit $L(A_X)$ bezeichnen wir die von A_X erkannte Sprache. Wir benutzen nichtdeterministische Automaten und nicht reguläre Ausdrücke zur Beschreibung der regulären Randbedingungen, da sie im allgemeinen eine kompaktere Darstellung ermöglichen. Die denotationelle Länge der Gleichung E bezeichnen wir mit $d := |L| + |R|$.

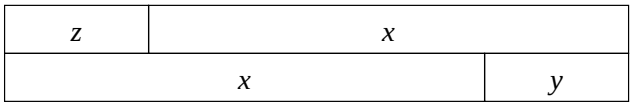
Die Lösung der Gleichung E ist ein Homomorphismus $\sigma : (\Sigma \cup \Omega)^* \to \Sigma^*$ mit $\sigma(a) = a$ für alle $a \in \Sigma$, $\sigma(\overline{X}) = \overline{\sigma(X)}$ für alle $X \in \Omega$, $\sigma(L) = \sigma(R)$ und $\sigma(X) \in L(A_X)$ für alle $X \in \Omega$. Eine Lösung ist bereits durch die Abbildung $\Omega_{\frac{1}{2}} \to \Sigma^*$ vollständig bestimmt. Eine Lösung σ ist minimal, wenn $|\sigma(L)| \leq |\sigma'(L)|$ ist für jede Lösung σ'.

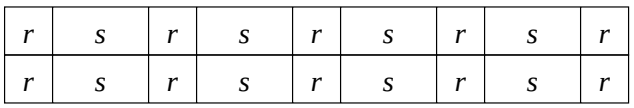
3.2. Wörter

Wir zeigen zunächst einige grundlegende Eigenschaften von Wörtern.

LEMMA 3.2.1. *Es sei $x \in \Sigma^*$ und $y, z \in \Sigma^+$. Dann sind die beiden folgenden Aussagen äquivalent:*

1. $xy = zx$,
2. $\exists r \in \Sigma^*, s \in \Sigma^+, i \geq 0 : x = (rs)^i r, y = sr, z = rs$.

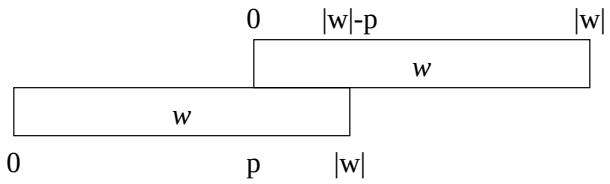




BEWEIS. s. Satz 1.3.4 in [**L83**]. □

DEFINITION 3.2.2. Eine Zahl p, $0 < p < \ell$ ist genau dann eine Periode des Wortes $w = a_1 \cdots a_\ell$, wenn $a_i = a_{p+i}$ für $1 \leq i \leq \ell - p$ ist.

LEMMA 3.2.3. *Eine Zahl p, $0 < p < |w|$ ist genau dann eine Periode des Wortes w, wenn $w[0, |w| - p] = w[p, |w|]$ ist.*

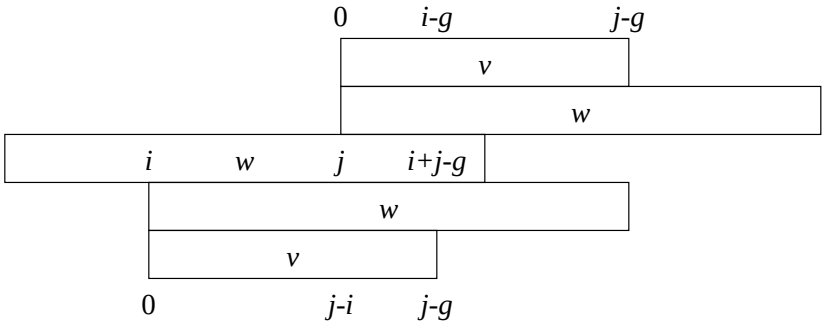


BEWEIS. Trivial. □

Wenn p eine Periode des Wortes w ist, sind aufgrund von Lemma 3.2.1 auch alle Vielfachen v von p mit $v < |w|$ Perioden von w. Um kleinere Perioden zu berechnen, kann man den Satz von Fine und Wilf [**FW65**], [**HHI98**] benutzen.

LEMMA 3.2.4. *Es seien i und j Perioden des Wortes $w \in \Sigma^*$. Wenn $|w| \geq i + j - ggt(i, j)$ ist, dann ist auch $ggt(i, j)$ eine Periode von w.*

BEWEIS. Es sei $g := \text{ggt}(i, j)$. O.B.d.A. sei $i \leq j$. Induktion über $|w|$. Wenn $|w| = 0$ oder $|w| = 1$ ist, hat w keine Perioden. Wenn $|w| = 2$ ist, können i und j nur 1 sein und es ist auch $g = 1$. Für $|w| > 2$ betrachten wir das Wort $v = w[0, j - g]$. Es ist g eine Periode von w, wenn $g = i$ oder $i = j$ ist. Also sei $g < i < j$.



Es ist $v[0, i-g] = w[0, i-g] = w[j, i+j-g] = w[j-i, j-g] = v[j-i, j-g]$. Also hat nach Lemma 3.2.3 das Wort v die Periode $j-i$. Wenn $g = j-i$ ist, hat v die Periode g. Wenn $g < j-i$ ist, hat v die Perioden i und $j-i$. Da $\mathrm{ggt}(j-i, i) = \mathrm{ggt}(i, j)$ ist, hat v auch die Periode g. Da v die Periode g hat, w die Periode i hat, $|v| \geq i$ ist und i ein Vielfaches von g ist, hat mit Lemma 3.2.1 auch w die Periode g. □

Von besonderer Bedeutung werden auch die primitiven Wörter sein.

DEFINITION 3.2.5. Ein Wort $p \in \Sigma^*$ ist primitiv, wenn sich p nicht von der Form w^i ist mit $w \in \Sigma^*$ und $i > 1$, d.h. es ist

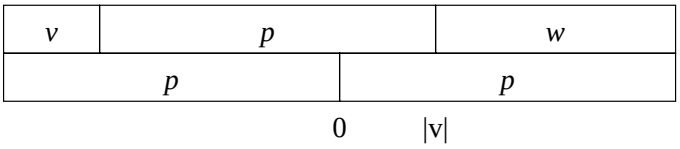
$$p \text{ primitiv} \Longleftrightarrow \neg\exists w \in \Sigma^*, i > 1 : p = w^i.$$

Man beachte, daß $\overline{p}$ genau dann primitiv ist, wenn p primitiv ist. Man könnte die primitiven Wörter auch anders definieren.

LEMMA 3.2.6. *Es ist $p \in \Sigma^+$ genau dann primitiv, wenn keine Wörter $v, w \in \Sigma^+$ existieren, so daß $p^2 = vpw$ ist, d.h. p kommt nicht echt in p^2 vor.*

BEWEIS. Es sei p nicht primitiv. Dann ist $p = u^i$ mit $i > 1$ und es ist $p^2 = u^{2i} = uu^iu^{i-1} = vpw$ mit $v = u$ und $w = u^{i-1}$.

Es sei p primitiv. Nehmen wir an, daß p echt in p^2 vorkommt. Dann gibt es ein kürzestes primitives Wort p, so daß $p^2 = vpw$ ist mit $v, w \in \Sigma^+$. Nach Lemma 3.2.1 ist $p = (rs)^i r$, $v = rs$ und $p[0, |v|] = sr$ mit $r \in \Sigma^*$, $s \in \Sigma^+$, $i \geq 0$.



Wenn $r = \epsilon$ ist, muß $i = 1$ sein, da sonst p nicht primitiv ist. Es kann nicht $r = \epsilon$ und $i = 1$ sein, da dann $w = \epsilon$ ist.

Wenn $r \neq \epsilon$ ist, muß $i \geq 1$ sein, da sonst $|v| > |p|$ ist. Also ist $rs = v = p[0, |v|] = sr$. Das Wort $p' := rs$ ist nicht primitiv, da sonst $p'^2 = rsrs =$

$rp's$ und $|p'| < |p|$ ist, was ein Widerspruch ist. Da p' nicht primitiv ist, gilt $p' = p''^j$ mit $j > 1$. Wir wählen p'' und j, so daß j maximal ist. Dann ist p'' ein primitives Wort. Es ist $|r|$ kein Vielfaches von $|p''|$, da sonst $r = p''^k$ und $s = p''^\ell$ ist und damit p nicht primitiv ist. Es ist

$$p''^2 = (sr)[0, 2|p''|] = (rs)[0, 2|p''|] = v'p''w'$$

mit $v' = r[0, |r| \bmod |p''|]$ und $w' = s[0, |s| \bmod |p''|]$. Dann ist aber p nicht minimal, da $|p''| < |p|$. Dies ist ein Widerspruch. Also kommt p nicht echt in p^2 vor. □

Aufgrund dieser Eigenschaft der primitiven Wörter, liegt es nahe die stabil-primitiven Wörter einzuführen.

DEFINITION 3.2.7. Ein Wort $p \in \Sigma^*$ ist stabil-primitiv, wenn weder p noch $\overline{p}$ echt in p^2 vorkommt.

LEMMA 3.2.8. *Es sei $p \in \Sigma^*$ ein stabil-primitives Wort. Dann ist auch $\overline{p}$ stabil-primitiv.*

BEWEIS. Trivial.

□

3.3. FMA-Gleichungssysteme

Ein FMA-Gleichungssystem S mit regulären Randbedingungen ist eine Menge von FMA-Gleichungen, d.h.

$$S = \{E_1 : L_1 = R_1, E_2 : L_2 = R_2, \ldots, E_k : L_k = R_k\}.$$

Alle Gleichungen sind über denselben Konstanten Σ und denselben Variablen Ω definiert. Zusätzlich sind für die Variablen Ω die regulären Randbedingungen A_X gegeben. Eine Lösung σ für das Gleichungssystem muß alle Gleichungen erfüllen.

Wir lösen ein Gleichungssystem, indem wir es zuerst in eine einzelne FMA-Gleichung mit regulären Randbedingungen transformieren und dann diese lösen. Diese Gleichung wird genau dann eine Lösung haben, wenn das

Gleichungssystem eine Lösung hat. In der Gleichung werden dieselben Variablen wie in dem Gleichungssystem vorkommen und eine Lösung der Gleichung wird auch eine Lösung des Gleichungssystems sein.

LEMMA 3.3.1. *Es sei* $S = \{E_1 : L_1 = R_1, E_2 : L_2 = R_2, \ldots, E_k : L_k = R_k\}$ *ein FMA-Gleichungssystem mit den regulären Randbedingungen* A_X. *Dieses Gleichungssystem ist genau dann lösbar, wenn die Gleichung* E

$$L_1 a L_2 a \cdots a L_k L_1 b L_2 b \cdots b L_k = R_1 a R_2 a \cdots a R_k R_1 b R_2 b \cdots b R_k$$

lösbar ist, wobei $a, b \in \Sigma$ *zwei beliebige unterschiedliche Konstanten sind. Es ist* $d = 4k - 4 + 2\sum_{i=1}^{k} d_i$.

BEWEIS. O.B.d.A. können wir annehmen, daß $|\Sigma| \geq 2$ ist, da wir sonst einfach eine weitere Konstante hinzufügen. Diese kann in $\sigma(X)$ für alle $X \in \Omega$ aufgrund der regulären Randbedingungen A_X nicht vorkommen. Offensichtlich ist eine Lösung für das Gleichungssystem S auch eine Lösung für die Gleichung E.

Es sei σ eine Lösung für die Gleichung E. Es ist

$$\begin{aligned} |\sigma(L_1 a L_2 a \cdots a L_k)| &= |\sigma(L_1 b L_2 b \cdots b L_k)| \text{ und} \\ |\sigma(R_1 a R_2 a \cdots a R_k)| &= |\sigma(R_1 b R_2 b \cdots b R_k)|. \end{aligned}$$

Also muß

$$\begin{aligned} \sigma(L_1 a L_2 a \cdots a L_k) &= \sigma(R_1 a R_2 a \cdots a R_k) \text{ und} \\ \sigma(L_1 b L_2 b \cdots b L_k) &= \sigma(R_1 b R_2 b \cdots b R_k) \end{aligned}$$

sein. Dann muß $\sigma(L_1) = \sigma(R_1)$ sein, da sonst

$$\begin{aligned} \sigma(L_1 a L_2 a \cdots a L_k) &\neq \sigma(R_1 a R_2 a \cdots a R_k) \text{ oder} \\ \sigma(L_1 b L_2 b \cdots b L_k) &\neq \sigma(R_1 b R_2 b \cdots b R_k) \end{aligned}$$

ist. Wenn $\sigma(L_1) = \sigma(R_1)$ ist, dann muß

$$\begin{aligned} \sigma(L_2 a L_3 a \cdots a L_k) &= \sigma(R_2 a R_3 a \cdots a R_k) \text{ und} \\ \sigma(L_2 b L_3 b \cdots b L_k) &= \sigma(R_2 b R_3 b \cdots b R_k) \end{aligned}$$

sein. Aufgrund derselben Argumentation ist $\sigma(L_2) = \sigma(R_2)$ und wir erhalten induktiv $\sigma(L_i) = \sigma(R_i)$ für $1 \leq i \leq k$. □

3.4. Ausdrücke über FMA-Gleichungen

Ein Ausdruck mit regulären Randbedingungen über FMA-Gleichungen ist eine boolesche Formel mit FMA-Gleichungen als atomare Formeln. Alle Gleichungen sind über denselben Konstanten Σ und denselben Variablen Ω definiert. Zusätzlich sind für die Variablen $X \in \Omega$ die regulären Randbedingungen A_X gegeben. Eine Lösung σ muß den Ausdruck wahr werden lassen.

Wir lösen einen Ausdruck über FMA-Gleichungen mit regulären Randbedingungen, indem wir ihn zuerst in eine einzelne FMA-Gleichung mit regulären Randbedingungen transformieren. Diese Gleichung wird genau dann eine Lösung haben, wenn der Ausdruck eine Lösung hat. In der Gleichung werden alle Variablen aus dem Ausdruck vorkommen und eine Lösung der Gleichung (eingeschränkt auf die im Ausdruck verwendeten Variablen) wird auch eine Lösung des Ausdrucks sein.

Karhumäki, Plandowski und Mignosi geben in [**KPM97**] eine Transformation für Ausdrücke ohne reguläre Randbedingungen über FM-Gleichungen an, die aber eine bis zu exponentiell größere FM-Gleichung erzeugt. Wir werden die regulären Randbedingungen benutzen, um eine Gleichung mit nur polynomieller Größe zu erzeugen. Wir erstellen die Gleichung, indem wir die Konjunktionen der Form $L_1 = R_1 \wedge L_2 = R_2$, die Disjunktionen der Form $L_1 = R_1 \vee L_2 = R_2$ und die Negationen der Form $L \neq R$ schrittweise durch einzelne Gleichungen ersetzen. Hierzu benutzen wir die drei folgenden Lemmata.

LEMMA 3.4.1. *Der Ausdruck $L_1 = R_1 \wedge L_2 = R_2$ ist äquivalent mit der Gleichung $L_1 a L_2 = R_1 a R_2$ mit der neuen Konstante a mit $a \in I$.*

BEWEIS. Die regulären Randbedingungen der Variablen wurden nicht verändert. Die Konstante a kann also nicht in der Lösung für eine Variable $X \in \Omega$ vorkommen und sie kommt auch nicht in L_1, R_1, L_2 oder R_2 vor,

da es sich um eine neue Konstante handelt. Daher ist genau dann $L_1 a L_2 = R_1 a R_2$, wenn $L_1 = R_1 \wedge L_2 = R_2$ ist. □

Man beachte, daß die entstehende Gleichung nicht größer ist als der Ausdruck.

LEMMA 3.4.2. *Der Ausdruck $L_1 = R_1 \vee L_2 = R_2$ ist äquivalent mit der Gleichung*

$$W a X a b L_1 X b W L_2 b = R_1 a R_2 a Y b W X b Z$$

mit den neuen Konstanten a und b mit $a, b \in I$ und den neuen Variablen W, $\overline{W}$, X, $\overline{X}$, Y, $\overline{Y}$, Z und $\overline{Z}$ mit den regulären Randbedingungen $L(A_W), L(A_X) = (\Sigma \setminus \{a, b\})^*$ und $L(A_Y), L(A_Z) = (\Sigma \setminus \{a\})^*$.

BEWEIS. Die regulären Randbedingungen der Variablen wurden nicht verändert. Die Konstante a kann also nicht in der Lösung für eine Variable $X \in \Omega$ vorkommen und sie kommt auch nicht in L_1, R_1, L_2 oder R_2 vor, da es sich um eine neue Konstante handelt. Da die Konstante a genau zweimal auf der linken und rechten Seite auftritt, ist die Gleichung also äquivalent mit dem Ausdruck

$$W = R_1 \wedge X = R_2 \wedge b L_1 X b W L_2 b = Y b W X b Z.$$

Durch Einsetzen von R_1 bzw. R_2 für W bzw. X ergibt sich die äquivalente Gleichung

$$b L_1 R_2 b R_1 L_2 b = Y b R_1 R_2 b Z.$$

Die Konstante b kann nur in der Lösung für die Variablen Y, $\overline{Y}$, Z und $\overline{Z}$ vorkommen. Daher ist diese Gleichung äquivalent mit dem Ausdruck

$$L_1 R_2 = R_1 R_2 \vee R_1 L_2 = R_1 R_2$$

und durch Kürzen von R_2 bzw. R_1 erhalten wir den äquivalenten Ausdruck

$$L_1 = R_1 \vee L_2 = R_2.$$

□

Die Variablen W und X wurden eingeführt, damit L_1, R_1, L_2 und R_2 in der Gleichung nur jeweils einmal vorkommen. Daher ist die Gleichung selbst nur konstant viel größer als der Ausdruck $L_1 = R_1 \vee L_2 = R_2$. Hinzu kommen die regulären Randbedingungen für die neuen Variablen.

LEMMA 3.4.3. *Die Ungleichung $L \neq R$ ist äquivalent mit der Gleichung*

$$LbaRbaUV = XUYaXVZaW$$

mit den neuen Konstanten a und b mit $a, b \in I$ und den neuen Variablen U, $\overline{U}$, V, $\overline{V}$, W, $\overline{W}$, X, $\overline{X}$, Y, $\overline{Y}$, Z und $\overline{Z}$ mit den regulären Randbedingungen $L(A_U), L(A_V) = \Sigma \setminus \{a\}$, $L(A_W) = \{uv \mid u, v \in \Sigma \setminus \{a\} \wedge u \neq v\}$ und $L(A_X), L(A_Y), L(A_Z) = (\Sigma \setminus \{a\})^*$.

BEWEIS. Die regulären Randbedingungen der Variablen wurden nicht verändert. Die Konstante a kann also nicht in der Lösung für eine Variable $X \in \Omega$ vorkommen und sie kommt auch nicht in L oder R vor, da es sich um eine neue Konstante handelt. Da die Konstante a genau zweimal auf der linken und rechten Seite auftritt, ist die Gleichung also äquivalent mit dem Ausdruck

$$Lb = XUY \wedge Rb = XVZ \wedge UV = W.$$

Es sei σ eine Lösung für die Ungleichung. Wenn $\sigma(L) \neq \sigma(R)$ ist, muß eine der folgenden drei Aussagen wahr sein:

- $\sigma(R)$ ist ein echtes Präfix von $\sigma(L)$. Der Ausdruck wird von $\sigma(U) = \sigma(L)[|\sigma(R)|, |\sigma(R)|+1]$, $\sigma(V) = b$, $\sigma(W) = \sigma(L)[|\sigma(R)|, |\sigma(R)|+1]b$, $\sigma(X) = \sigma(R)$, $\sigma(Y) = \sigma(L)[|\sigma(R)|+1, |\sigma(L)|]b$ und $\sigma(Z) = \epsilon$ erfüllt.
- $\sigma(L)$ ist ein echtes Präfix von $\sigma(R)$. Der Ausdruck wird von $\sigma(U) = b$, $\sigma(V) = \sigma(R)[|\sigma(L)|, |\sigma(L)|+1]$, $\sigma(W) = b\,\sigma(R)[|\sigma(L)|, |\sigma(L)|+1]$, $\sigma(X) = \sigma(L)$, $\sigma(Y) = \epsilon$ und $\sigma(Z) = \sigma(R)[|\sigma(L)|+1, |\sigma(R)|]b$ erfüllt.
- $\sigma(L)$ und $\sigma(R)$ unterscheiden sich ab der Position i. Der Ausdruck wird von $\sigma(U) = \sigma(L)[i, i+1]$, $\sigma(V) = \sigma(R)[i, i+1]$, $\sigma(W) = \sigma(L)[i, i+1]\sigma(R)[i, i+1]$, $\sigma(X) = \sigma(L)[0, i]$, $\sigma(Y) = \sigma(L)[i+1, |\sigma(L)|]b$ und $\sigma(Z) = \sigma(R)[i+1, |\sigma(R)|]b$ erfüllt.

Wenn die Ungleichung nicht lösbar ist, gilt $L = R$. Dann ist auch $Lb = Rb$, $XUY = XVZ$, $UY = VZ$ und $U = V$. Also ist die reguläre Randbedingung für $W = UV$ nicht erfüllt und der Ausdruck nicht lösbar. □

Die Gleichung selbst ist nur nur konstant viel größer als die Ungleichung. Die regulären Randbedingungen haben die Größe $\mathcal{O}(\Sigma^2) \subseteq \mathcal{O}(n^2)$.

KOROLLAR 3.4.4. *Für jeden Ausdruck mit regulären Randbedingungen über FMA-Gleichungen der Größe n gibt es eine äquivalente einzelne FMA-Gleichung mit regulären Randbedingungen der Größe $\mathcal{O}(n^3)$.*

BEWEIS. Die entstehende Gleichung selbst hat die Größe $\mathcal{O}(n)$. Es gibt maximal $\mathcal{O}(n)$ neue Variablen und $\mathcal{O}(n)$ Konstanten. Die regulären Randbedingung für eine neue Variable hat maximal die Größe $\mathcal{O}(n^2)$. Also hat die Gleichung insgesamt höchstens die Größe $\mathcal{O}(n^3)$. □

3.5. Konstanten

Um die Darstellung im weiteren zu vereinfachen, eliminieren wir zuerst alle Konstanten in der Gleichung. Wir ersetzen alle Konstanten $a \in \Sigma$ in der Gleichung $E : L = R$ durch neue Variablen X_a mit den regulären Randbedingungen $L(A_{X_a}) = \{a\}$. Die Eingabe wird dadurch maximal um einen konstanten Faktor größer.

3.6. Reguläre Randbedingungen

Wir bestimmen ein Monoid M und einen Homomorphismus $h : \Sigma^* \to M$, der die regulären Randbedingungen A_X erkennt, d.h. für jede Variable X existiert eine Menge $M_X \subseteq M$, so daß $h^{-1}(M_X) = L(A_X)$ ist. Wir brauchen die beiden folgenden Lemmata, um später den Periodizitätsexponenten zu beschränken.

LEMMA 3.6.1. *Es sei B eine boolesche $m \times m$ Matrix. Dann ist $B^{m!} = B^{2m!}$.*

BEWEIS. Wir interpretieren die boolesche Matrix B als gerichteten Graphen mit m Knoten. Es ist genau dann $B_{i,j} = 1$, wenn es eine Kante vom Knoten i zum Knoten j gibt, und genau dann $B^{\ell}_{i,j} = 1$, wenn es einen Weg $i = i_0, i_1, i_2, \ldots, i_\ell = j$ der Länge ℓ vom Knoten i zum Knoten j gibt. Wir werden zeigen, daß $B^{m!}_{i,j} = 1 \Leftrightarrow B^{2m!}_{i,j} = 1$ ist.

Wenn $B^{m!}_{i,j} = 1$ ist, gibt es einen Weg der Länge $m!$ von i nach j. Betrachten wir die ersten $m+1$ Knoten dieses Weges. Da es nur m verschiedene Knoten gibt, kommt ein Knoten mehrmals vor und es existieren $0 \leq a < b \leq m$, so daß $i_a = i_b$ ist. Also gibt es auch Wege der Länge $m! + k(b-a)$ mit $k \geq 0$ von i nach j. Es ist $b - a \leq m$ und damit $b - a$ ein Teiler von $m!$. Also gibt es einen Weg der Länge $m! + \frac{m!}{b-a}(b-a) = 2m!$ von i nach j und es ist $B^{2m!}_{i,j} = 1$.

Wenn $B^{2m!}_{i,j} = 1$ ist, gibt es einen Weg der Länge $2m!$ von i nach j. Wir zerlegen diesen Weg in einen Basisweg der Länge $\ell \leq m(m-1)$ und Zyklen, die höchstens m lang sind. Es sei $s_1 := 0$ und $k = 1$. Wir betrachten die Knoten $i_{s_k}, i_{s_k+1}, \ldots, i_{s_k+m}$ des Weges. Da es nur m verschiedene Knoten gibt, kommt ein Knoten mehrmals vor und es existieren $s_k \leq a < b \leq s_k + m$, so daß $i_a = i_b$ ist. Wir verkürzen den Weg um die Knoten $i_{a+1}, i_{a+2}, \ldots, i_b$. Der entstehende Weg ist um $b - a \leq m$ kürzer. Da wir diesen Zyklus später in den Basisweg einsetzen wollen, müssen wir darauf achten, daß der Knoten i_a im Basisweg enthalten ist. Wenn der Knoten i_a nicht unter den Knoten $i_0, i_1, \ldots, i_{s_k}$ vorkommt, ist deshalb $s_{k+1} = a$ und sonst ist $s_{k+1} = s_k$. Wir wiederholen dieses Vorgehen für $k+1$, bis die Länge des verbleibenden Weges $\ell < s_{k+1} + m$ ist. Es wird höchstens $m-1$ mal $s_{k+1} = a$ gesetzt, da es nur $m-1$ Knoten gibt, die ungleich i_0 sind, und jedesmal ist $s_{k+1} = a \leq s_k + m - 1$. Also ist $\ell \leq (m-1)(m-1) + (m-1) = m(m-1)$. Wir können die Länge des ursprünglichen Wegs also schreiben als $\ell + \sum_{i=1}^{m} t_i i = 2m!$, wobei t_i die Anzahl ist, wie oft ein Zyklus der Länge i aus dem Weg ausgeschnitten wurde. Es sei j der kleinste Wert für den $t_j > 0$ ist. Für $i > j$ ersetzen wir jeweils j Zyklen der Länge i durch i Zyklen der Länge j, d.h. $t_j := t_j + i \left\lfloor \frac{t_i}{j} \right\rfloor$ und

$t_i := t_i \bmod j$. Dann ist $t_i < j$ für $i > j$ und

$$\ell + jt_j + \sum_{i=j+1}^{m} it_i = 2m!.$$

Wir unterscheiden drei Fälle, um zu zeigen, daß $jt_j \geq m!$ ist.

1. Für $m > 3$ ist $jt_j \geq m!$, da $\ell \leq m(m-1)$ ist und
$$\sum_{i=j+1}^{m} it_i \leq \frac{(j-1)(m-j)(m+j+1)}{2} \leq m(m-2)(m-3)$$
ist.
2. Für $m = 3$ ist $jt_j \geq m!$, wenn $j \neq 2$ ist, da $\ell \leq m(m-1)$ und $\sum_{i=j+1}^{m} it_i = 0$ ist. Es ist auch $jt_j \geq m!$, wenn $j = 2$ und $t_j \geq 3$ ist. Da $\ell + jt_j + \sum_{i=j+1}^{m} it_i = 2m!$ sein muß, bleibt nur noch der Fall $j = 2$, $t_2 = 2$ und $t_3 = 1$, bei dem es von jedem Knoten zu jedem Knoten einen Weg der Länge $m! = 6$ gibt. Also ist $B_{i,j}^{m!} = 1$.
3. Für $m < 3$ ist $jt_j \geq m!$, da $\ell \leq m(m-1)$ und $\sum_{i=j+1}^{m} it_i = 0$ ist.

Da $jt_j \geq m!$ ist, erhalten wir durch das Einfügen aller Zyklen bis auf $\frac{m!}{j}$ Zyklen der Länge j einen Weg der Länge $m!$ und es ist $B_{i,j}^{m!} = 1$. □

LEMMA 3.6.2. *Das Monoid M kann so gewählt werden, daß*

$$|M| \leq 2^{\sum_{X\in\Omega} |Q_X|^2} \leq 2^{n^2}$$

ist und es eine Konstante

$$1 \leq c(M) \leq max\{|Q_X| \mid X \in \Omega\}! \leq n!$$

gibt, so daß $\forall m \in M : m^{c(M)} = m^{2c(M)}$ ist.

BEWEIS. O.B.d.A. seien bei den Automaten A_X die Zustände $Q_X = \{q_{X,1}, q_{X,2}, \ldots q_{X,|Q_X|}\}$. Wir stellen ein Monoidelement durch $|\Omega|$ quadratische boolesche Matrizen B_X der Größen $|Q_X| \times |Q_X|$ mit $X \in \Omega$ dar. Wir verknüpfen zwei Monoidelemente, indem wir für alle $X \in \Omega$ die Matrizen B_X miteinander multiplizieren.

Wir definieren den Homomorphismus h wie folgt: Für ein Wort $w \in \Sigma^*$ hat das Matrixelement $B_{X,i,j}$ in Zeile i und Spalte j der Matrix B_X genau

dann der Wert 1, wenn der Automat A_X durch das Lesen des Wortes w vom Zustand $q_{X,i}$ in den Zustand $q_{X,j}$ übergehen kann, d.h.

$$B_{X,i,j} = 1 \Leftrightarrow q_{X,i} \xrightarrow{w} q_{X,j}.$$

Wir setzen $M := h(\Sigma^*)$ und $M_X := h(L(A_X))$, um die regulären Randbedingungen zu erkennen. Man kann leicht überprüfen, daß M ein Monoid, h ein Homomorphismus und $h^{-1}(M_X) = L(A_X)$ ist. Es ist $|M| \leq 2^{\sum_{X\in\Omega}|Q_X|^2} \leq 2^{n^2}$.

Wir setzen $c(M) := \max\{|Q_X| \mid X \in \Omega\}!$. Es sei m ein beliebiges Monoidelement. Nach Lemma 3.6.1 ist $B_X^{c(M)} = B_X^{2c(M)}$ für alle $X \in \Omega$ und damit $m^{c(M)} = m^{2c(M)}$. □

3.7. Minimale Wörter und reguläre Randbedingungen

Für alle Paare von Monoidelementen $m_1, m_2 \in M$ bestimmen wir ein kürzestes Wort $\min_{m_1,m_2} \in \Sigma^*$, so daß

$$h(\min_{m_1,m_2}) = m_1 \text{ und } h(\overline{\min_{m_1,m_2}}) = m_2$$

ist. Wenn kein solches Wort existiert ist $\min_{m_1,m_2}$ undefiniert. Die Menge der kürzesten Wörter sei

$$\min_M := \{\min_{m_1,m_2} \mid m_1, m_2 \in M \wedge \min_{m_1,m_2} \text{ definiert}\}$$

und die Länge des längsten Wortes in $\min_M$ sei $\max_M = \max\{|w| \mid w \in \min_M\}$.

LEMMA 3.7.1. *Es ist* $max_M < |M|^2$.

BEWEIS. Nehmen wir an, daß Monoidelemente m_1, m_2 existieren mit $|\min_{m_1,m_2}| \geq |M|^2$. Es sei $\ell = |\min_{m_1,m_2}|$. Dann gibt es $0 \leq \alpha < \beta \leq \ell$, so daß $h(\min_{m_1,m_2}[0,\alpha]) = h(\min_{m_1,m_2}[0,\beta])$ und $h(\min_{m_1,m_2}[\ell,\ell-\alpha]) =$

$h(\min_{m_1,m_2}[\ell, \ell-\beta])$ ist, da es nur $|M|^2$ verschiedene Paare von Monoidelementen gibt. Es sei $w := \min_{m_1,m_2}[0,\alpha]\min_{m_1,m_2}[\beta,\ell]$. Es ist

$$\begin{aligned} h(\min_{m_1,m_2}) &= h(\min_{m_1,m_2}[0,\beta]\min_{m_1,m_2}[\beta,\ell]) \\ &= h(\min_{m_1,m_2}[0,\beta])h(\min_{m_1,m_2}[\beta,\ell]) \\ &= h(\min_{m_1,m_2}[0,\alpha])h(\min_{m_1,m_2}[\beta,\ell]) \\ &= h(w), \end{aligned}$$

$$\begin{aligned} h(\overline{\min_{m_1,m_2}}) &= h(\min_{m_1,m_2}[\ell,\alpha]\min_{m_1,m_2}[\alpha,0]) \\ &= h(\min_{m_1,m_2}[\ell,\alpha])h(\min_{m_1,m_2}[\alpha,0]) \\ &= h(\min_{m_1,m_2}[\ell,\beta])h(\min_{m_1,m_2}[\alpha,0]) \\ &= h(\overline{w}) \end{aligned}$$

und $|w| < |\min_{m_1,m_2}|$. Dies ist ein Widerspruch zur Minimalität von $\min_{m_1,m_2}$. Also ist $|\min_{m_1,m_2}| < |M|^2$. □

LEMMA 3.7.2. *Es sei $w \in \Sigma^*$. Dann existiert ein Wort v, so daß $h(v) = h(w)$, $h(\overline{v}) = h(\overline{w})$ und $|v| \leq 2max_M + 1 < 2|M|^2$ ist und $v = \overline{v}$ ist, wenn $w = \overline{w}$ ist.*

BEWEIS. Wenn $w \neq \overline{w}$ ist, erfüllt $v = \min_{h(w),h(\overline{w})}$ die Voraussetzungen. Wenn $w = \overline{w}$ ist, sei $u = w[0, \lfloor |w|/2 \rfloor]$ und $a = \epsilon$, wenn $|w|$ gerade ist, und $a = w[(|w|-1)/2, (|w|+1)/2]$ sonst. Dann ist $w = ua\overline{u}$ und $a \in I \cup \{\epsilon\}$. Es sei $v := \min_{h(u),h(\overline{u})}a\overline{\min_{h(u),h(\overline{u})}}$. Es ist $v = \min_{h(u),h(\overline{u})}a\overline{\min_{h(u),h(\overline{u})}} = \overline{v}$,

$$\begin{aligned} h(v) &= h(\min_{h(u),h(\overline{u})})h(a)h(\overline{\min_{h(u),h(\overline{u})}}) \\ &= h(u)h(a)h(\overline{u}) \\ &= h(w), \end{aligned}$$

$h(\overline{v}) = h(\overline{w})$, da $v = \overline{v}$, $w = \overline{w}$ und $h(v) = h(w)$ ist, und

$$|v| = |\min_{h(u),h(\overline{u})}| + |a| + |\overline{\min_{h(u),h(\overline{u})}}| \leq 2\max_M + 1 < 2|M|^2.$$

□

KAPITEL 4

Reduktionen

Bevor wir die Theoreme für FMA-Gleichungen mit regulären Randbedingungen beweisen, werden wir zeigen, wie man FM-Gleichungen mit regulären Randbedingungen und FG-Gleichungen mit regulären Randbedingungen auf FMA-Gleichungen mit regulären Randbedingungen reduzieren kann.

4.1. Gleichungen über freien Monoiden

Diese Reduktion ist sehr einfach:

- Wir machen alle Konstanten zu Fixpunkten der Anti-Involution, d.h. wir setzen $I = \Sigma$.
- Wir erweitern die Menge der Variablen Ω um die Inversen $\overline{\Omega}$.

Die entstehende Gleichung ist eine FMA-Gleichung mit regulären Randbedingungen und hat genau dieselben Lösungen wie die ursprüngliche Gleichung.

4.2. Gleichungen über freien Gruppen

Gutiérrez hat in [**G00A**] gezeigt, wie man Gleichungen über freien Gruppen mit Hilfe von FMA-Gleichungen lösen kann. Er geht dabei jedoch den Umweg über FMA-Gleichungen mit irreduziblen Lösungen. Wir werden zeigen, wie man direkt von der freien Gruppe zum freien Monoid mit Anti-Involution übergehen kann und daß auch reguläre Randbedingungen berücksichtigt werden können.

Es sei G eine freie Gruppe, die von den Erzeugenden $S = \{a, b, \dots\}$ erzeugt wird. Die Menge der Erzeugenden S und ihrer Inversen $\overline{S} = \{\overline{a}, \overline{b}, \dots\}$ bezeichnen wir als die Konstanten $\Sigma := S \cup \overline{S} = \{a, \overline{a}, b, \overline{b}, \dots\}$. Es ist

$\overline{\overline{a}} = a$ und $\overline{a} \neq a$ für alle $a \in \Sigma$. Entsprechend sagen wir, daß G die freie Gruppe ist, die von den Konstanten Σ erzeugt wird.

Wir schreiben $v \rightarrow w$, wenn $v = ra\overline{a}s$ und $w = rs$ mit $a \in \Sigma$ und $r, s \in \Sigma^*$ ist. Mit $\approx$ bezeichnen wir die Äquivalenzrelation über $\rightarrow$. Jede Äquivalenzklasse entspricht einem Element in der freien Gruppe G. Es gelten die üblichen Gruppenaxiome, wobei ϵ das Einselement in der freien Gruppe ist:

$$\begin{aligned} xy &\approx z, \\ (xy)z &\approx x(yz), \\ \epsilon x &\approx x \approx x\epsilon, \\ \overline{x}x &\approx \epsilon \approx x\overline{x}. \end{aligned}$$

Für ein Wort $w = a_1 a_2 \cdots a_{|w|}$ ist $\overline{w} = \overline{a_{|w|}} \cdots \overline{a_2}\,\overline{a_1}$. Ein Wort $w \in \Sigma^*$ ist irreduzibel, wenn es kein Teilwort $a\overline{a}$ mit $a \in \Sigma$ enthält.

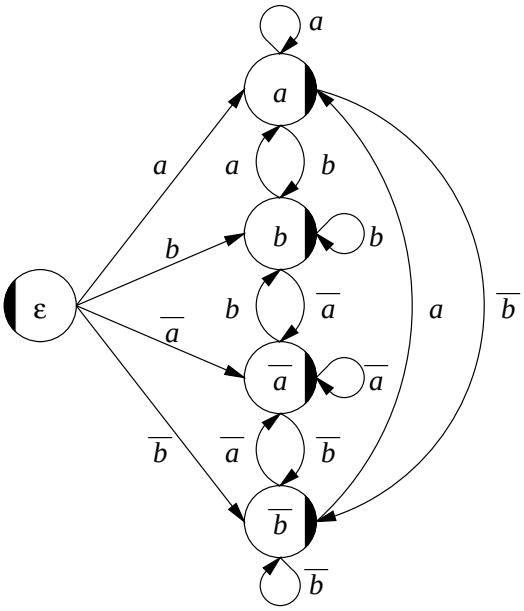
SATZ 4.2.1. *In jeder Äquivalenzklasse von $\approx$ gibt es genau ein irreduzibles Wort.*

BEWEIS. Siehe Lemma 7.1.1 in [**H59**]. □

Wir können also jedes Element x der freien Gruppe G durch ein irreduzibles Wort w repräsentieren.

Bei den Gleichungen über freien Gruppen werden die regulären Randbedingungen rationale Sprachen sein. Wir geben diese durch nichtdeterministische endliche Automaten $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$ mit den Zuständen Q_X, dem Alphabet Σ, den Startzuständen $I_X \subseteq Q_X$, den Übergängen $\delta_X \subseteq Q_X \times (\Sigma \cup \epsilon) \times Q_X$ und den Endzuständen $F_X \subseteq Q_X$. Wir benutzen diese Darstellung, da sie im allgemeinen kompakter ist als rationale Ausdrücke. Man kann die Symbole $\Sigma \cup \epsilon$ an den Übergängen entweder als Elemente der freien Gruppe oder als Elemente eines freien Monoids auffassen. Mit $L_G(A_X)$ bzw. $L(A_X)$ bezeichnen wir die von A_X erkannte Sprache, wobei die Symbole als Elemente der freien Gruppe bzw. des freien Monoids aufgefaßt werden.

BEISPIEL 4.2.2. Es sei $\Sigma = \{a, \overline{a}, b, \overline{b}\}$. Für den folgenden Automaten ist $L(A_X)$ die Menge aller irreduziblen Wörter bis auf ϵ und $L_G(A_X)$ die Menge aller Elemente der freien Gruppe bis auf ϵ.



Es sei $\Omega = \{X, Y, \dots\}$ die Menge der Variablen. Zu jeder Variablen X muß es eine inverse Variable $\overline{X}$ mit $\overline{\overline{X}} = X$ und $\overline{X} \neq X$ geben. Mit $\Omega_{\frac{1}{2}}$ bezeichnen wir eine Teilmenge von Ω, für die $\Omega_{\frac{1}{2}} \cup \overline{\Omega_{\frac{1}{2}}} = \Omega$ und $\Omega_{\frac{1}{2}} \cap \overline{\Omega_{\frac{1}{2}}} = \emptyset$ gilt. Eine Gleichung mit regulären Randbedingungen über einer freien Gruppe (kurz FG-Gleichung mit regulären Randbedingungen) ist eine Gleichung $E : L \approx R$ mit $L, R \in (\Sigma \cup \Omega)^*$ zusammen mit den regulären Randbedingungen $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$. Die denotationelle Länge der Gleichung ist $d := |L| + |R|$.

Normalerweise wäre die Lösung der Gleichung E ein Homomorphismus, der die Variablen auf Elemente der freien Gruppe abbildet und die Elemente der freien Gruppe müßten in $L_G(A_X)$ enthalten sein. Dies hat aber den Nachteil, daß es schwer ist die Länge der Lösung zu definieren und wir der Lösung nicht direkt ansehen können, wie das Wort aussieht, das von A_X akzeptiert wird. Wir werden daher fordern, daß die Lösung für eine Variable ein Element des freien Monoids ist. Dieses Wort muß in der

Äquivalenzklasse sein, die dem Element der freien Gruppe entspricht, und es muß in $L(A_X)$ sein. Entsprechend werden wir im weiteren nicht mehr Elemente der freien Gruppe betrachten sondern nur noch Elemente des freien Monoid. Die Äquivalenz bzgl. der freien Gruppe können wir weiterhin mit $\approx$ ausdrücken.

Die Lösung der Gleichung E ist ein Homomorphismus $\sigma : (\Sigma \cup \Omega)^* \to \Sigma^*$ mit $\sigma(a) = a$ für alle $a \in \Sigma$, $\sigma(\overline{X}) = \overline{\sigma(X)}$ für alle $X \in \Omega$, $\sigma(L) \approx \sigma(R)$, und $\sigma(X) \in L(A_X)$ für alle $X \in \Omega$. Eine Lösung ist bereits durch die Abbildung $\Omega_{\frac{1}{2}} \to \Sigma^*$ vollständig bestimmt. Eine Lösung σ ist minimal, wenn $|\sigma(LR)| \leq |\sigma'(LR)|$ ist für jede Lösung σ'.

O.B.d.A. können wir im weiteren davon ausgehen, daß

1. die Gleichung E die Form $L \approx \epsilon$ hat, da $L \approx R$ und $L\overline{R} \approx \epsilon$ äquivalent sind,
2. $d = |L| \geq 2$ ist, da $L \approx \epsilon$ und $La\overline{a} \approx \epsilon$ für ein beliebiges $a \in \Sigma$ äquivalent sind,
3. $L \in \Omega^*$ ist, da wir die Konstanten $a \in \Sigma$ durch neue Variablen X_a mit den regulären Randbedingungen $L(X_a) = a$ ersetzen können, und
4. die regulären Randbedingungen für alle $X \in \overline{\Omega_{\frac{1}{2}}}$ gleich Σ^* sind. Der Automat $A'_{\overline{X}} := (Q_{\overline{X}}, \Sigma, \delta'_{\overline{X}}, I'_{\overline{X}}, F'_{\overline{X}})$ mit $\delta'_{\overline{X}} := \{(q_1, a, q_2) \mid (q_2, \overline{a}, q_1) \in \delta_{\overline{X}}\}$, $I'_{\overline{X}} := F_{\overline{X}}$ und $F'_{\overline{X}} := I_{\overline{X}}$ erkennt die reguläre Randbedingung für $\overline{X}$ auf dem Wort $\sigma(X)$. Der Produktautomat von A_X und $A'_{\overline{X}}$ erkennt also gleichzeitig die regulären Randbedingungen A_X und $A_{\overline{X}}$. Wenn wir diesen anstelle von A_X verwenden, können wir $A_{\overline{X}}$ so wählen, daß $L(A_{\overline{X}}) = \Sigma^*$ ist. Man beachte, daß die Größe des Produktautomaten quadratisch bzgl. der Größe der Automaten A_X und $A_{\overline{X}}$ sein kann.

4.3. Ausdrücke über FG-Gleichungen

Ein Ausdruck mit regulären Randbedingungen über FG-Gleichungen ist eine boolesche Formel mit FG-Gleichungen als atomare Formeln. Alle Gleichungen sind über denselben Konstanten Σ und denselben Variablen Ω definiert. Zusätzlich sind für die Variablen Ω die regulären Randbedingungen A_X gegeben. Eine Lösung σ muß den Ausdruck wahr werden lassen.

Wir lösen einen Ausdruck mit regulären Randbedingungen über FG-Gleichungen, indem wir ihn zuerst in eine einzelne FG-Gleichung mit regulären Randbedingungen transformieren. Diese Gleichung wird genau dann eine Lösung haben, wenn der Ausdruck eine Lösung hat. In der Gleichung werden alle Variablen aus dem Ausdruck vorkommen und eine Lösung der Gleichung (eingeschränkt auf die im Ausdruck verwendeten Variablen) wird auch eine Lösung des Ausdrucks sein. Wir erstellen die Gleichung wie bei den FMA-Gleichungen, indem wir die Konjunktionen der Form $L_1 \approx \epsilon \wedge L_2 \approx \epsilon$, die Disjunktionen der Form $L_1 \approx \epsilon \vee L_2 \approx \epsilon$ und die Negationen der Form $L \not\approx \epsilon$ schrittweise durch einzelne Gleichungen ersetzen. Hierzu benutzen wir die drei folgenden Lemmata.

LEMMA 4.3.1. *Der Ausdruck $L_1 \approx \epsilon \wedge L_2 \approx \epsilon$ ist äquivalent mit der Gleichung $L_1 a L_2 \overline{a} \approx \epsilon$ mit den neuen Konstanten a und $\overline{a}$.*

BEWEIS. Die regulären Randbedingungen der Variablen wurden nicht verändert. Die Konstanten a und $\overline{a}$ können also nicht in der Lösung für eine Variable $X \in \Omega$ vorkommen und sie kommen auch nicht in L_1 oder L_2 vor, da es sich um neue Konstanten handelt. Daher ist genau dann $L_1 a L_2 \overline{a} \approx \epsilon$, wenn $L_1 \approx \epsilon \wedge L_2 \approx \epsilon$ ist. □

Man beachte, daß die entstehende Gleichung nicht größer ist als der Ausdruck.

LEMMA 4.3.2. *Der Ausdruck $L_1 \approx \epsilon \vee L_2 \approx \epsilon$ ist äquivalent mit der Gleichung*

$$a L_1 a L_2 a \overline{Y} \overline{a}\, \overline{a} \overline{X} \approx \epsilon$$

mit den neuen Konstanten a und $\overline{a}$ und den neuen Variablen X, $\overline{X}$, Y und $\overline{Y}$ mit den regulären Randbedingungen $L(A_X), L(A_Y) = (\Sigma \setminus \{\overline{a}\})^*$.

BEWEIS. Die Gleichung ist äquivalent mit

$$aL_1aL_2a \approx XaaY.$$

Die Konstante a kann nur in der Lösung für die Variablen X und Y vorkommen. Die Konstante $\overline{a}$ kann weder in der Lösung für L_1 oder L_2 noch in der Lösung für X oder Y vorkommen. Daher ist diese Gleichung äquivalent mit dem Ausdruck

$$L_1 \approx \epsilon \vee L_2 \approx \epsilon.$$

□

Die Gleichung selbst ist nur konstant viel größer als der Ausdruck $L_1 \approx \epsilon \vee L_2 \approx \epsilon$. Hinzu kommen die regulären Randbedingungen für die neuen Variablen. Da die Menge der rationalen Sprachen eine effektive boolesche Algebra ist, können wir das Komplement von ϵ durch eine reguläre Randbedingung darstellen.

LEMMA 4.3.3. *Die Ungleichung $L \not\approx \epsilon$ ist äquivalent mit dem Ausdruck $L\overline{X} \approx \epsilon$ mit den neuen Variablen X und $\overline{X}$ mit den regulären Randbedingungen $A_X := (Q_X, \Sigma, \delta_X, I_X, F_X)$ mit*

$$\begin{aligned} Q_X &:= \{q_\epsilon\} \cup \{q_a \mid a \in \Sigma\}, \\ \delta_X &:= \{(q_\epsilon, a, q_a) \mid a \in \Sigma\} \cup \{(q_a, b, q_b) \mid a, b \in \Sigma \wedge a \neq \overline{b}\}, \\ I_X &:= \{q_\epsilon\}, \\ F_X &:= \{q_a \mid a \in \Sigma\}. \end{aligned}$$

BEWEIS. Die Ungleichung $L \not\approx \epsilon$ ist äquivalent mit dem Ausdruck $L \approx X \wedge X \not\approx \epsilon$, wobei es noch keine reguläre Randbedingung für X gibt. Dieser Ausdruck ist äquivalent mit der Gleichung $L \approx X$ mit der regulären Randbedingung A_X für X, da A_X alle irreduziblen Wörter außer ϵ erkennt. Die Gleichung $L \approx X$ ist äquivalent mit $L\overline{X} \approx \epsilon$. □

Der entstehende Ausdruck ist nur konstant viel länger und die reguläre Randbedingung A_X hat nur die Größe $\mathcal{O}(\Sigma^2) \subseteq \mathcal{O}(n^2)$.

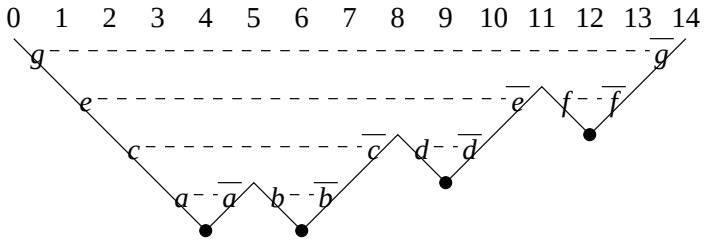
KOROLLAR 4.3.4. *Für jeden Ausdruck mit regulären Randbedingungen über FG-Gleichungen der Größe n gibt es eine äquivalente einzelne FG-Gleichung mit regulären Randbedingungen der Größe $\mathcal{O}(n^3)$.*

BEWEIS. Die entstehende Gleichung selbst hat die Größe $\mathcal{O}(n)$. Es gibt maximal $\mathcal{O}(n)$ neue Variablen und $\mathcal{O}(n)$ Konstanten. Die regulären Randbedingung für eine neue Variable hat maximal die Größe $\mathcal{O}(n^2)$. Also hat die Gleichung insgesamt höchstens die Größe $\mathcal{O}(n^3)$. □

4.4. Reduzierungen

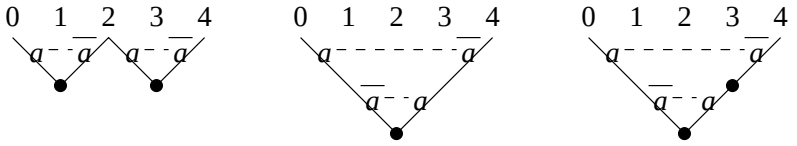
Es sei $w \approx \epsilon$. Wir betrachten Reduzierungen $w \xrightarrow{*} \epsilon$. Bei jeder Reduzierung können wir jedem Zeichen $w[\alpha, \alpha+1]$ genau ein anderes Zeichen $w[\beta, \beta+1] = \overline{w[\alpha, \alpha+1]}$ zuordnen, so daß sich die Zeichen gegenseitig eliminieren. Wir können dies graphisch darstellen, indem wir das Wort w als ein Gebirge darstellen, in dem die beiden Zeichen auf gleicher Höhe auf gegenüberliegenden Seiten eines Tals stehen.

BEISPIEL 4.4.1. Die graphische Darstellung für die Reduzierung des Wortes $geca\overline{a}b\overline{b}c\overline{c}d\overline{d}e\overline{e}f\overline{f}\overline{g}$:



Wenn zwei Teilwörter $w[\alpha, \alpha+\ell]$ und $w[\beta, \beta+\ell] = \overline{w[\alpha, \alpha+\ell]}$ sich bei der Reduzierung gegenseitig eliminieren, nennen wir diese die linke und rechte Seite eines Tales der Größe 2ℓ. Wenn die beiden Teilwörter in w aneinander angrenzen, d.h. $\alpha + \ell = \beta$ ist, sprechen wir von einem Basistal. Die Mitte der Basistäler kennzeichnen wir mit einem Punkt.

BEISPIEL 4.4.2. Für das Wort $w = a\overline{a}a\overline{a}$ gibt es z.B. folgende Reduzierungen:



In der linken Reduzierung eliminieren sich die Wörter $w[0,1]$ und $w[1,2]$ und die Wörter $w[2,3]$ und $w[3,4]$, in der mittleren Reduzierung eliminieren sich die Wörter $w[0,2]$ und $w[2,4]$ und in der rechten Reduzierung eliminieren sich die Wörter $w[1,2]$ und $w[2,3]$, die Wörter $w[3,3]$ und $w[3,3]$ und die Wörter $w[0,1]$ und $w[3,4]$.

LEMMA 4.4.3. *Es seien $w_1, w_2, \ldots, w_k \in \Sigma^*$ irreduzible Wörter und es sei $w_1 w_2 \cdots w_k \approx \epsilon$. Dann existieren Wörter $v_1, v_2, \ldots, v_\ell \in \Sigma^*$, so daß*

1. *$w_j = v_{i_j} v_{i_j+1} \cdots v_{i_{j+1}-1}$ mit $i_1 = 1$ und $i_{k+1} - 1 = \ell$,*
2. *in der freien von den Konstanten v_i erzeugten Gruppe $v_1 v_2 \cdots v_\ell \approx_{v_i} \epsilon$ ist und*
3. *$\ell \leq 4k - 6$ ist.*

BEWEIS. Wir betrachten eine Reduzierung von $w := w_1 w_2 \cdots w_k$ zu ϵ. Da die Wörter w_j irreduzibel sind, kann die Reduzierung nur an Positionen zwischen den Wörtern w_j starten. Wir betrachten alle Positionen zwischen den Wörtern w_j als Basistäler. Es gibt also $k-1$ Basistäler, wobei manche davon die Größe 0 haben können. Sobald zwei Täler zusammenstoßen verschmelzen sie zu einem neuen Tal. Da danach jeweils ein Tal weniger existiert, entstehen genau $k-2$ weitere Täler. Also gibt es insgesamt $(k-1) + (k-2) = 2k-3$ Täler. Es seien $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_\ell = |w|$ alle Positionen in w, an denen eine Talseite startet oder endet. Da es nur $2k-3$ Täler mit jeweils zwei Seiten gibt, ist $\ell \leq 4k-6$. Wir setzen $v_i := w[\alpha_{i-1}, \alpha_i]$. Da die linke Talseite immer genau das Inverse der rechten Seite ist, gilt in der freien über den Konstanten v_i erzeugten Gruppe $v_1 v_2 \cdots v_\ell \approx_{v_i} \epsilon$. □

Die Grenze $4k-6$ ist optimal, wie das Beispiel $w_1 = a_{2k-3}a_{2k-5}\cdots a_1$, $w_2 = \overline{a_1}a_2$, $w_3 = \overline{a_2}\,\overline{a_3}a_4$, $w_4 = \overline{a_4}\,\overline{a_5}a_6$, ..., $w_{k-1} = \overline{a_{2k-6}}\,\overline{a_{2k-5}}a_{2k-4}$ und $w_k = \overline{a_{2k-4}}\,\overline{a_{2k-3}}$ zeigt. Der Fall $k = 5$ ist im Beispiel 4.4.1 dargestellt.

4.5. Von der FG-Gleichung zur FMA-Gleichung

Beim Übergang von der freien Gruppe zum freien Monoid mit Anti-Involution werden wir die regulären Randbedingungen verändern. Das folgende Lemma beschreibt die Automaten für die regulären Randbedingungen im freien Monoid mit Anti-Involution.

LEMMA 4.5.1. *Es sei $A = (Q, \Sigma, \delta, I, F)$ ein nichtdeterministischer endlicher Automat. Dann existiert ein nichtdeterministischer endlicher Automat $A' = (Q, \Sigma, \delta', I, F)$, der*

1. *die Sprache $\{w \mid \exists v \in L(A) : v \xrightarrow{*} w\}$ erkennt,*
2. *maximal $|Q|^2$ mehr Übergänge hat und*
3. *in polynomieller Zeit konstruiert werden kann.*

BEWEIS. O.B.d.A sei $Q = \{q_1, q_2, q_3, \ldots, q_{|Q|}\}$. Wir erstellen die kontextfreie Grammatik $G = (V, \Sigma, P, S_{1,1})$ mit

$$\begin{aligned} V &= \{S_{i,j} \mid 1 \le i, j \le |Q|\}, \\ P &= \{S_{i,k} \to S_{i,j}S_{j,k} \mid 1 \le i, j, k \le |Q|\} \\ &\quad \cup \{S_{i,\ell} \to aS_{j,k}\overline{a} \mid a \in \Sigma \wedge q_i \xrightarrow{a} q_j \wedge q_k \xrightarrow{\overline{a}} q_\ell\} \\ &\quad \cup \{S_{i,j} \to \epsilon \mid q_i \xrightarrow{\epsilon} q_j\}. \end{aligned}$$

Der Startzustand $S_{1,1}$ wurde beliebig gewählt, da uns die von der Grammatik selbst erzeugte Sprache nicht interessiert. Die Grammatik hat maximal die Größe $\mathcal{O}(|\Sigma||Q|^4)$. Die Nichtterminale $S_{i,j}$ erzeugen genau die Wörter, die sich zu ϵ reduzieren lassen und mit denen man vom Zustand q_i zum Zustand q_j kommt, d.h. $L(S_{i,j}) = \{w \mid w \approx \epsilon \wedge q_i \xrightarrow{w} q_j\}$. Wir erzeugen δ', indem wir den Übergang $q_i \xrightarrow{\epsilon} q_j$ zu δ hinzufügen, wenn $L(S_{i,j}) \neq \emptyset$ ist, d.h. $\delta' := \delta \cup \{(q_i, \epsilon, q_j) \mid L(S_{i,j}) \neq \emptyset\}$. Wir fügen maximal $|Q|^2$ Übergänge hinzu. Ob ein Nichtterminal produktiv ist, kann man in polynomieller Zeit feststellen.

Es ist $\{w \mid \exists v \in L(A) : v \xrightarrow{*} w\} \subseteq L(A')$, da

$$v = u_o w[0,1]u_1 w[1,2]u_2 \cdots u_{|w|-1} w[|w|-1,|w|]u_{|w|}$$

mit $u_i \approx \epsilon$ sein muß und für jedes der Wörter u_i ein ϵ-Übergang in A' existiert. Es ist $\{w \mid \exists v \in L(A) : v \xrightarrow{*} w\} \supseteq L(A')$, da, wenn $w \in L(A')$ ist, es auch ein Wort

$$v = u_o w[0,1]u_1 w[1,2]u_2 \cdots u_{|w|-1} w[|w|-1,|w|]u_{|w|}$$

mit $u_i \approx \epsilon$ in $L(A)$ geben muß und damit $v \xrightarrow{*} w$ gilt. □

Man beachte, daß es für jedes produktive Nichtterminal $S_{i,j}$ ein Wort $w_{i,j} \in \Sigma^*$ mit $S_{i,j} \xrightarrow{*} w_{i,j}$ und $|w_{i,j}| \leq 2^{|Q|^2} + 2^{|Q|^2-1} - 2 < 2\,2^{|Q|^2}$ gibt, da es nur $|Q|^2$ Nichtterminale gibt. Damit haben wir jetzt die Voraussetzungen, um das folgende Theorem zu beweisen.

THEOREM 4.5.2. *Es sei $E : L \approx \epsilon$ eine FG-Gleichung mit $L \in \Omega^*$, den regulären Randbedingungen $A_X = (Q_X, \Sigma, \delta_X, I_X, F_X)$ für die Variablen $X \in \Omega_{\frac{1}{2}}$ und $d \geq 2$. Dann existieren FMA-Gleichungen $E_1 : L_1 = R_1, E_2 : L_2 = R_2, \ldots, E_k : L_k = R_k$ mit den regulären Randbedingungen $A'_X = (Q_X, \Sigma, \delta'_X, I_X, F_X)$, so daß*

1. *$|L_i| + |R_i| \leq 14d - 16$, $k \in 2^{\mathcal{O}(d \log(d))}$ und $|\delta'_X| \leq |\delta_X| + |Q_X|^2$ ist,*
2. *wenn es eine Lösung σ für E gibt, es auch eine Lösung σ_i für eine der Gleichungen E_i gibt und*
3. *wenn es eine Lösung σ_i für eine der Gleichungen E_i gibt, es auch eine Lösung σ für E gibt mit $|\sigma(L)| \leq 2\,2^{n^2}(|\sigma_i(L_i)| + n)$.*

BEWEIS. Wir erstellen die Gleichungen E_i wie folgt:

1. Wir führen $4d-6$ neue Variablen $\Omega' = \{Y_1, \overline{Y_1}, \ldots, Y_{2d-3}, \overline{Y_{2d-3}}\}$ ein.
2. Wir erstellen Gleichungssysteme

   $$S_i = \{X_{p_1} = T_1, X_{p_2} = T_2, \ldots, X_{p_d} = T_d\}.$$

   Für jede Variable X_{p_j} in $L = X_{p_1} X_{p_2} \cdots X_{p_d}$ erstellen wir eine Gleichung $X_{p_j} = T_j$ mit $T_j \in \Omega'^*$, wobei $\sum_{j=1}^{d} |T_j| \leq 4d - 6$ sein muß und in der freien über den Konstanten Ω' erzeugten Gruppe $T_1 T_2 \cdots T_d \approx_{\Omega'} \epsilon$ sein muß.

3. Wir fassen die Gleichungssysteme S_i mit Hilfe von Lemma 3.3.1 zu den Gleichungen E_i zusammen.
4. Die regulären Randbedingungen A'_X sind die Automaten aus Lemma 4.5.1 für die Automaten A_X.

Es ist $|L_i| + |R_i| = 4d - 4 + 2\sum_{j=1}^{d}(|X_{k_j}| + |T_j|) \leq 4d - 4 + 2(d + 4d - 6) = 14d - 16$. Da die Gleichungssysteme S_i sich nur auf der rechten Seite unterscheiden, diese rechten Seiten insgesamt nur eine Länge von maximal $4d-6$ haben und nur aus maximal $4d-6$ verschiedenen Variablen bestehen, ist $k \in 2^{\mathcal{O}(d\log(d))}$. Aufgrund von Lemma 4.5.1 ist $|\delta'_X| \leq |\delta_X| + |Q_X|^2$.

Es sei σ eine Lösung für E. Für alle $X_p \in \Omega$ bezeichnen wir mit $\hat{w}_p$ die irreduziblen Wörter die äquivalent mit $\sigma(X_p)$ sind, d.h. $\hat{w}_p \approx \sigma(X_p)$. Nach Lemma 4.4.3 existieren Wörter $v_1, v_2, \ldots, v_\ell$ mit $\ell \leq 4d - 6$, so daß $\hat{w}_{p_j} = v_{q_j} v_{q_j+1} \cdots v_{q_{j+1}-1}$ für $1 \leq j \leq d$, $q_1 = 1$ und $q_{d+1}-1 = \ell$ ist und in der freien von den Konstanten v_q erzeugten Gruppe $v_1 v_2 \cdots v_\ell \approx_{v_q} \epsilon$ ist. Wenn wir in den Gleichungen $\hat{w}_{p_j} = v_{q_j} v_{q_j+1} \cdots v_{q_{j+1}-1}$ die Symbole $\hat{w}_p$ durch X_p und die Symbole v_q durch T_q ersetzen erhalten wir ein Gleichungssystem S_i, das wir in Schritt 2 erzeugt haben. Dieses Gleichungssystem hat die Lösung σ_i mit $\sigma_i(X_p) = \hat{w}_p$ und $\sigma_i(Y_q) = v_q$. Die Lösung σ_i erfüllt auch die regulären Randbedingungen A'_X, da $\sigma(X_p) \in L(A_{X_p})$ ist und damit $\sigma_i(X_p) = \hat{w}_p \approx \sigma(X_p)$ in $L(A'_{X_p}) = \{w \mid \exists v \in L(A_{X_p}) : v \xrightarrow{*} w\}$ ist. Mit dem Gleichungssystem S_i hat auch die Gleichung E_i die Lösung σ_i nach Lemma 3.3.1.

Es sei σ_i eine Lösung für E_i. Wir wählen ein Wort mit minimaler Länge aus $\{v \in L(A_{X_p}) \mid v \xrightarrow{*} \sigma_i(X_p)\}$ als Lösungswort $\sigma(X_p)$ für alle $X_p \in \Omega_{\frac{1}{2}}$. Die Mengen $\{v \in L(A_{X_p}) \mid v \xrightarrow{*} \sigma_i(X_p)\}$ sind nicht leer, da die Wörter $\sigma_i(X_p) \in L(A'_{X_p}) = \{w \mid \exists v \in L(A_{X_p}) : v \xrightarrow{*} w\}$ sind. Es ist σ eine Lösung für die Gleichung $E : L \approx \epsilon$ mit $L = X_{p_1} X_{p_2} \cdots X_{p_d}$, da $\sigma(X_{p_j}) \approx \sigma_i(X_{p_j}) = \sigma_i(T_j)$ und $T_1 T_2 \cdots T_d \approx_{\Omega'} \epsilon$ ist. Die regulären Randbedingungen werden von σ erfüllt, da nur Wörter aus $L(A_{X_p})$ als $\sigma(X_p)$ in Frage kamen. Aufgrund der Anmerkung nach Lemma 4.5.1 existiert für zwei Zustände $q_{X,i}$ und $q_{X,j}$ im Automaten A_X, wenn es ein Wort $v_{X,i,j}$ mit $v_{X,i,j} \approx \epsilon$ und $q_{X,i} \xrightarrow{v_{X,i,j}} q_{X,j}$ gibt, auch ein Wort $w_{X,i,j}$ mit $|w_{X,i,j}| < 2\,2^{|Q_X|^2}$, $w_{X,i,j} \approx \epsilon$

und $q_{X,i} \overset{w_{X,i,j}}{\to} q_{X,j}$. Also ist $\sigma(X_p) \leq 2\,2^{|Q_{X_p}|^2}(|\sigma_i(X_p)|+1)$ und $|\sigma(L)| \leq 2\,2^{\max\{|Q_X|^2|X\in\Omega\}}(|\sigma_i(L_i)|+d) \leq 2\,2^{n^2}(|\sigma_i(L_i)|+n)$. □

Mit diesem Theorem können wir für eine FG-Gleichung E mit regulären Randbedingungen nichtdeterministisch in polynomieller Zeit eine FMA-Gleichung E_i mit regulären Randbedingungen bestimmen, die genau dann eine Lösung σ_i hat, wenn die Gleichung E eine Lösung σ hat. Die Gleichung E_i ist nur um einen konstanten Faktor länger und die Automaten A'_X haben gleich viel Zustände wie die Automaten A_X. Lediglich die Anzahl der Übergänge kann quadratisch größer sein. Dies hat jedoch keine Auswirkungen auf die unsere Komplexitätsabschätzung, da unsere obere Schranke für die Monoidgröße $|M| \leq 2^{\sum_{X\in\Omega}|Q_X|^2}$ nur von der Anzahl der Zustände abhängt. Durch den Übergang von der freien Gruppe zum freien Monoid mit Anti-Involution wird die Eingabe also eventuell quadratisch größer, sie verhält sich aber bzgl. unserer Komplexitätsabschätzungen immer wie eine nur um einen konstanten Faktor größere Eingabe.

Hinzu kommt noch, daß die Eingabe quadratisch größer wurde durch die Eliminierung der regulären Randbedingungen für alle $X \in \overline{\Omega_{\frac{1}{2}}}$. Das dies für unseren Beweis des Theorems 4.5.2 notwendig ist, sieht man am folgenden Beispiel.

Beispiel 4.5.3. Es sei die Gleichung $E : X\overline{X} = \epsilon$ mit $L(A_X) = a\overline{a}$ und $L(A_{\overline{X}}) = b\overline{b}$ gegeben, die offensichtlich keine Lösung hat. Es würde u.a. das Gleichungssystem $S = \{X = Y_1, \overline{X} = \overline{Y_1}\}$ mit $L(A'_X) = \epsilon|a\overline{a}$ und $L(A'_{\overline{X}}) = \epsilon|b\overline{b}$ erzeugt. Dieses hat aber die Lösung σ mit $\sigma(X) = \epsilon$ und $\sigma(Y_1) = \epsilon$.

KAPITEL 5

Der Periodizitätsexponent

DEFINITION 5.0.4. Es sei $w \in \Sigma^*$. Der Periodizitätsexponent $\exp(w)$ eines Wortes w ist definiert als

$$\exp(w) := \max\{i \mid \exists r, s \in \Sigma^*, p \in \Sigma^+ : w = rp^i s\}.$$

Schon seit dem ursprünglichen Algorithmus von Makanin ist die Beschränkung des Periodizitätsexponenten das Fundament für alle weiteren Abschätzungen, d.h. sie hängen in irgendeiner Form vom Periodizitätsexponenten ab.

Als nächstes werden wir die p-stabile Normalform definieren. Die Definition weicht von der üblichen in der Literatur verwendeten [**D00**] ab, da wir zusätzlich die Anti-Involution berücksichtigen. Wenn $p = \overline{p}$ ist, stimmen die Normalformen jedoch überein.

Wie werden anstelle von $\overline{w}^i$ zum Teil auch w^{-i} schreiben. Es gilt $(w^{-1})^i = w^{-i} = (w^i)^{-1}$, da $\overline{w}^i = \overline{w^i}$ ist.

DEFINITION 5.0.5. Es sei $p \in \Sigma^+$ ein stabil-primitives Wort. Die p-stabile Normalform eines Wortes $w \in \Sigma^*$ ist eine Sequenz

$$(u_0, r_1, u_1, r_2, \ldots, r_k, u_k)$$

mit $k \geq 0$, $u_i \in \Sigma^*$ für $0 \leq i \leq k$ und $r_i \in \mathbb{Z}$ für $1 \leq i \leq k$, so daß

1. $w = u_0 p^{r_1} u_1 p^{r_2} \cdots p^{r_k} u_k$ ist,
2. $r_i \geq 0$ für $1 \leq i \leq k$ ist, wenn $p = \overline{p}$ ist,
3. k minimal ist, d.h. es gibt keine kürzere Sequenz,
4. genau dann $k = 0$ ist, wenn weder p^2 noch p^{-2} Teilwörter von w sind, und

5. wenn $k > 0$ ist,

$$\begin{aligned} u_0 &\in \Sigma^* p^{s_1} \setminus \Sigma^* p^{\pm 2} \Sigma^*, \\ u_i &\in (p^{s_i}\Sigma^* \cap \Sigma^* p^{s_{i+1}}) \setminus \Sigma^* p^{\pm 2}\Sigma^*, \\ u_k &\in p^{s_k}\Sigma^* \setminus \Sigma^* p^{\pm 2}\Sigma^* \end{aligned}$$

für $0 < i < k$ ist mit

$$s_i = \begin{cases} \operatorname{sign}(r_i) & \text{wenn } r_i \neq 0, \\ \pm 1 & \text{wenn } r_i = 0 \wedge p \neq \overline{p}, \\ 1 & \text{wenn } r_i = 0 \wedge p = \overline{p} \end{cases}$$

für $1 \leq i \leq k$.

Als nächstes werden wir beweisen, daß es sich wirklich um eine Normalform handelt, d.h. sie ist eindeutig.

LEMMA 5.0.6. *Es sei $p \in \Sigma^+$ stabil-primitiv. Dann ist die p-stabile Normalform eines Wortes $w \in \Sigma^*$ eindeutig bestimmt.*

BEWEIS. Nehmen wir an, daß es zwei unterschiedliche p-stabile Normalformen $(u_0, r_1, u_1, r_2, \ldots, r_k, u_k)$ und $(u'_0, r'_1, u'_1, r'_2, \ldots, r'_{k'}, u'_{k'})$ für w gibt. Es ist $k = k'$, da sowohl k als auch k' minimal sind. Es seien s_i und s'_i entsprechend der Definition 5.0.5 gesetzt. Wir zeigen, daß

$$(u_0, r_1, u_1, r_2, \ldots, u_{i-1}, r_i) = (u'_0, r'_1, u'_1, r'_2, \ldots, u'_{i-1}, r'_i)$$

ist, mit Induktion für $0 \leq i \leq k$.

Induktionsanfang: Für $i = 0$ ist $() = ()$.
Induktionsvoraussetzung: Es gilt

$$(u_0, r_1, u_1, r_2, \ldots, u_{i-2}, r_{i-1}) = (u'_0, r'_1, u'_1, r'_2, \ldots, u'_{i-2}, r'_{i-1}).$$

Induktionsschritt: Nehmen wir an, daß $u_{i-1} \neq u'_{i-1}$ ist. Aus Symmetriegründen können wir o.B.d.A. davon ausgehen, daß $|u_{i-1}| < |u'_{i-1}|$ ist. Es sei $u_{i-1} = vp^{s_i}$ und $u'_{i-1} = v'p^{s'_i}$. In w folgt auf $u_0 \cdots p^{r_{i-1}}u_{i-1}$ auf jeden Fall ein weiteres p^{s_i}. Es ist $|u'_{i-1}| < |u_{i-1}| + |p|$, da sonst p^{2s_i} in u'_{i-1} enthalten ist. Also kommt $p^{s'_i}$ in p^{2s_i} vor. Dies ist ein Widerspruch, da p stabil-primitiv ist.

Es ist $s_i = s'_i$, da u_{i-1} und u'_{i-1} mit demselben Wort p bzw. $\overline{p}$ enden und, wenn $p = \overline{p}$ ist, $s_i, s'_i = 1$ sein müssen.
Nehmen wir an, daß $|r_i| \neq |r'_i|$ ist. Aus Symmetriegründen können wir o.B.d.A. davon ausgehen, daß $|r_i| < |r'_i|$ ist. Wir unterscheiden drei Fälle.

1. $|u_i| = |p|$:
   Es ist $u_i = p^{s'_i} = p^{s_i}$. Es muß $i < k$ sein, da sonst $|u'_i| < |p|$ ist. Es ist k nicht minimal, wenn $p = \overline{p}$ oder $s_i = s_{i+1}$ ist, da die Sequenz

   $$(u_0, r_1, \ldots, r_{i-1}, u_{i-1}, r_i + s_i + r_{i+1}, u_{i+1}, r_{i+2}, \ldots, r_k, u_k)$$

   kürzer ist und auch w repräsentiert. Wenn $p \neq \overline{p}$ und $s_i \neq s_{i+1}$ ist, gilt nicht $u_i \in p^{s_i}\Sigma^* \cap \Sigma^* p^{s_{i+1}}$.
2. $|p| < |u_i| < 2|p|$:
   Es muß $i < k$ sein, da sonst $|u'_i| < |p|$ ist. Es kommt p^{s_i} in $p^{2s'_i}$ vor. Dies ist ein Widerspruch, da p stabil-primitiv ist.
3. $|u_i| \geq 2|p|$:
   Es kommt $p^{2s'_i}$ in u_i vor. Dies ist ein Widerspruch zu $u_i \in \ldots \setminus \Sigma^* p^{\pm 2} \Sigma^*$

Also ist

$$(u_0, r_1, u_1, r_2, \ldots, u_{k-1}, r_k) = (u'_0, r'_1, u'_1, r'_2, \ldots, u'_{k-1}, r'_k)$$

für $i = k$. Dann muß auch $u_k = u'_k$ sein, da beide Sequenzen w repräsentieren. □

Wenn wir zeigen, daß der Periodizitätsexponent begrenzt ist, werden wir davon ausgehen, daß es keine Fixpunkte gibt. Dies ermöglicht uns das folgende Lemma.

LEMMA 5.0.7. *Es sei $E : L = R$ eine FMA-Gleichung über dem Alphabet Σ mit den regulären Randbedingungen $M_X \subseteq M$. Dann existiert eine FMA-Gleichung $E' : L = R$ über dem Alphabet Σ' mit den regulären Randbedingungen $M'_X \subseteq M'$, so daß Σ' keine Fixpunkte hat, $c(M') \leq 2c(M)$ ist und, wenn σ eine minimale Lösung für E ist, gibt es eine minimale Lösung σ' für E' mit $exp(\sigma(L)) \leq exp(\sigma'(L))$.*

BEWEIS. Wir setzen

$$\begin{aligned}\Sigma' &:= (\Sigma \setminus I) \cup \{a', \overline{a'} \mid a \in I\},\\ M' &:= ((\{\epsilon\} \cup I) \times M \times (\{\epsilon\} \cup I)) \cup \{m_0\},\\ M'_X &:= \{\epsilon\} \times M_X \times \{\epsilon\}.\end{aligned}$$

Es sei $m'_1, m'_2 \in M'$. Wenn $m'_1 = m_0$ oder $m'_2 = m_0$ ist, gilt $m'_1 m'_2 := m_0$ (m_0 ist das Nullelement von M'). Für $m'_1 = (a_1, m_1, b_1)$ und $m'_2 = (a_2, m_2, b_2)$ ist

$$m'_1 m'_2 := \begin{cases} (a_1, m_1 h(b_1) m_2, b_2) & \text{wenn } b_1 = a_2,\\ m_0 & \text{wenn } b_1 \neq a_2.\end{cases}$$

Es sei $\Psi : \Sigma^* \to \Sigma'^*$ ein Homomorphismus mit $\Psi(a) = a$ für $a \in \Sigma \setminus I$ und $\Psi(a) = a'\overline{a'}$ für $a \in I$. Es ist $\Psi(\overline{w}) = \overline{\Psi(w)}$. Es ist $h' : \Sigma'^* \to M'$ der Homomorphismus mit

$$h'(a) := \begin{cases} (\epsilon, h(a), \epsilon) & \text{wenn } a \in \Sigma \setminus I,\\ (\epsilon, 1, b) & \text{wenn } a = b' \text{ mit } b \in I,\\ (b, 1, \epsilon) & \text{wenn } a = \overline{b'} \text{ mit } b \in I.\end{cases}$$

Σ' hat keine Fixpunkte. Es ist $m_0^{2c(M)} = m_0 = m_0^{4c(M)}$, $(a, m, b)^2 = m_0 = (a, m, b)^{2c(M)} = (a, m, b)^{4c(M)}$ für $a \neq b$ und

$$\begin{aligned}(a, m, a)^{2c(M)} &= (a, (mh(a))^{2c(M)-1} m, a)\\ &= (a, (mh(a))^{c(M)} (mh(a))^{c(M)-1} m, a)\\ &= (a, (mh(a))^{3c(M)} (mh(a))^{c(M)-1} m, a)\\ &= (a, (mh(a))^{4c(M)-1} m, a)\\ &= (a, m, a)^{4c(M)}.\end{aligned}$$

Also ist $c(M') \leq 2c(M)$.

Es sei σ eine Lösung für E. Dann ist σ' mit $\sigma'(X) := \Psi(\sigma(X))$ eine Lösung für E', da $\sigma'(\overline{X}) = \Psi(\sigma(\overline{X})) = \Psi(\overline{\sigma(X)}) = \overline{\Psi(\sigma(X))} = \overline{\sigma'(X)}$ für alle $X \in \Omega$, $\sigma'(L) = \Psi(\sigma(L)) = \Psi(\sigma(R)) = \sigma'(R)$ und $h'(\sigma'(X)) = h'((\epsilon, h(\sigma(X)), \epsilon)) \in M'_X$ für alle $X \in \Omega$. Es ist $\exp(\sigma(L) \leq \exp(\sigma'(L))$, da, wenn das Wort $w = p^i$ mit $p \in \Sigma^*$ in $\sigma(L)$ vorkommt, auch das Wort $\Psi(p^i)$ in $\sigma'(L)$ vorkommt. □

Um den Periodizitätsexponenten zu beschränken, werden wir ein diophantisches Gleichungssystem erstellen, so daß jede Lösung des Gleichungssystems zu einer Lösung der FMA-Gleichung führen wird. Kościelski und Pacholski haben den folgenden Satz bewiesen.

SATZ 5.0.8. *Es sei* $N := \{\vec{n}_i \cdot \vec{x} + n_i = 0 \mid 1 \leq i \leq n\}$ *mit* $\vec{n}_i = (n_{i,1}, n_{i,2}, \ldots, n_{i,r}) \in \mathbb{Z}^r$ *und* $n_i \in \mathbb{Z}$ *für* $1 \leq i \leq n$ *ein System von linearen diophantischen Gleichungen. Es sei* $w = \sum_{i=1}^{n} |n_i|$ *und* $c = \sum_{i=1}^{n} \sum_{j=1}^{r} |n_{i,j}|$. *Wenn* $\vec{x} = (x_1, x_2, \ldots, x_r) \in \mathbb{N}_0^r$ *eine minimale Lösung bzgl. der natürlichen partiellen Ordnung für* $\mathbb{N}_0^r$ *ist, dann ist* $x_i \leq (w + r)e^{c/e}$ *für* $1 \leq i \leq r$.

BEWEIS. s. Korollar 4.4 in [**KP96**]. □

Der obige Satz basiert auf den Arbeiten von Zur Gathen und Sieveking [**GS78**] bzw. Lambert [**L87a**], [**L87b**]. Damit können wir jetzt den Periodizitätsexponenten beschränken.

THEOREM 5.0.9. *Es sei* $E : L = R$ *einer FMA-Gleichung mit den regulären Randbedingungen* $M_X \subseteq M$ *und* σ *eine minimale Lösung für* E. *Dann existiert eine Konstante* g, *so daß* $exp(\sigma(L)) \leq gc(M)2^{1,6d}$ *ist.*

BEWEIS. Wenn $I \neq \emptyset$ ist, bestimmen wir den Periodizitätsexponenten für die Gleichung E' aus Lemma 5.0.7. Es ist

$$\exp(\sigma(L)) \leq \exp(\sigma'(L)) \leq g'c(M')2^{1,6d} \leq 2g'c(M)2^{1,6d} \leq gc(M)2^{1,6d}$$

mit $g := 2g'$. Wir können also o.B.d.A. davon ausgehen, daß es keine Fixpunkte gibt. Auch können wir annehmen, daß $c(M) \geq 2$ ist, da $2c(M)$ auch die Bedingungen für $c(M)$ erfüllt.

Wir wählen ein primitives Wort $p' \in \Sigma^+$, so daß $p'^{\exp(\sigma(L))}$ in $\sigma(L)$ enthalten ist. Wenn p' stabil-primitiv ist, sei $p := p'$. Sonst ist $\overline{p'}$ in p'^2 enthalten und es ist $p' = u'v'$ mit $u' = \overline{u'}$ und $v' = \overline{v'}$. Dann ist $u' = u\overline{u}$ und $v' = v\overline{v}$, denn es ist $I = \emptyset$. Das konjugierte Wort $p := \overline{u}v\overline{v}u$ ist stabil-primitiv, da mit p' auch p primitiv ist und $p = \overline{p}$ ist, und es kommt $p^{\exp(\sigma(L))-1}$ in $\sigma(L)$ vor.

Es sei $L = X_{i_1}X_{i_2}\cdots X_{i_c}$ und $R = X_{i_{c+1}}X_{i_{c+2}}\cdots X_{i_d}$. Das Gleichungssystem

$$S := \left\{ \begin{array}{rclcrcl} X_{i_1} & = & Y_1, & & X_{i_{c+1}} & = & Y_{c+1}, \\ Y_1X_{i_2} & = & Y_2, & & Y_{c+1}X_{i_{c+2}} & = & Y_{c+2,} \\ & \vdots & & & & \vdots & \\ Y_{c-1}X_{i_c} & = & Y_c, & & Y_{d-1}X_{i_d} & = & Y_d, \\ & & & Y_c = Y_d & & & \end{array} \right\}$$

mit den neuen Variablen $\{Y_1, \overline{Y_1}, Y_2, \overline{Y_2}, \ldots, Y_d, \overline{Y_d}\}$ ist äquivalent zu E. Es ist $\exp(\sigma(L)) = \exp(\sigma(Y_d))$. Nach einer offensichtlichen Elimination von Variablen, ist das Gleichungssystem äquivalent zu einem System mit $d-2$ Gleichungen vom Typ $XY = Z$ mit $X, Y, Z \in \Omega$.

Wir betrachten eine Gleichung $XY = Z$. Wir bestimmen die p-stabilen Normalformen von $\sigma(X)$, $\sigma(Y)$ und $\sigma(Z)$. Es sei

$$\begin{array}{rcl} \sigma(X) & : & (u_0, r_1c(M) + \alpha_1, u_1, r_2c(M) + \alpha_2, \ldots, r_kc(M) + \alpha_k, u_k), \\ \sigma(Y) & : & (v_0, s_1c(M) + \beta_1, v_1, s_2c(M) + \beta_2, \ldots, s_\ell c(M) + \beta_\ell, v_\ell), \\ \sigma(Z) & : & (w_0, t_1c(M) + \gamma_1, w_1, t_2c(M) + \gamma_2, \ldots, t_mc(M) + \gamma_m, v_m) \end{array}$$

mit $-c(M) < \alpha_i, \beta_i, \gamma_i < c(M)$ und

$$\begin{array}{rcl} \alpha_i \neq 0 & \Rightarrow & \mathrm{sign}(r_i) \in \{0, \mathrm{sign}(\alpha_i)\}, \\ \beta_i \neq 0 & \Rightarrow & \mathrm{sign}(s_i) \in \{0, \mathrm{sign}(\beta_i)\}, \\ \gamma_i \neq 0 & \Rightarrow & \mathrm{sign}(t_i) \in \{0, \mathrm{sign}(\gamma_i)\}. \end{array}$$

Man beachte, daß, wenn

$$(u_0, r_1c(M) + \alpha_1, u_1, r_2c(M) + \alpha_2, \ldots, r_kc(M) + \alpha_k, u_k)$$

die p-stabile Normalform für $\sigma(X)$ ist, die p-stabile Normalform für $\sigma(\overline{X})$

$$(\overline{u_k}, (-r_k)c(M) - \alpha_k, \overline{u_{k-1}}, (-r_{k-1})c(M) - \alpha_{k-1}, \ldots, (-r_1)c(M) - \alpha_1, \overline{u_0})$$

ist, wenn $p \neq \overline{p}$ ist, und

$$(\overline{u_k}, r_kc(M) + \alpha_k, \overline{u_{k-1}}, r_{k-1}c(M) + \alpha_{k-1}, \ldots, r_1c(M) + \alpha_1, \overline{u_0})$$

ist, wenn $p = \overline{p}$ ist. Wir können daher für die p-stabilen Normalformen von $\sigma(X)$ und $\sigma(\overline{X})$ dieselben Variablen benutzen.

Da σ eine Lösung ist, gibt es viele Gleichungen zwischen den Wörtern und den Zahlen. Z.B. $u_0 = w_0$, $v_\ell = w_m$, $r_1 = t_1$, $\alpha_1 = \gamma_1$, usw. für $k, \ell \geq 2$. Wir werden nur die Gleichungen mit den Variablen r_i, s_i und t_i benutzen. Dies sind u.a. die Gleichungen

$$\begin{array}{rclrclcrcl} r_1 & = & t_1, & r_2 & = & t_2, & \ldots, & r_{k-1} & = & t_{k-1} \\ s_2 & = & t_{m-l+2}, & s_3 & = & t_{m-l+3}, & \ldots, & s_\ell & = & t_m. \end{array}$$

Es sind k, ℓ und m nur beschränkt durch $|k+\ell-m| \leq 2$. Welche Gleichungen noch hinzukommen hängt von der p-stabilen Normalform von $u_k v_0$ ab. Es ist

$$u_k, v_0 \in \Sigma^* \setminus \Sigma^* p^{\pm 2} \Sigma^*.$$

Daher können nur die folgenden neun Fälle auftreten

$$\begin{array}{rcl} (u_k, v_0) & : & t_k = r_k,\ t_{k+1} = s_1 \\ (p^s, t', p^s) & : & t_{k'} = r_k + c_1 + s_1 + c_1' \\ (u_k' p^s, t', p^s) & : & t_k = r_k,\ t_{k+1} = s_1 + c_1' \\ (p^s, t', p^s v_0') & : & t_{k'} = r_k + c_1,\ t_{k'+1} = s_1 \\ (u_k' p^s, t', p^s v_0') & : & t_k = r_k,\ t_{k+1} = 0,\ t_{k+2} = s_1 \\ (p^s, 0, w', 0, p^{s'}) & : & t_{k'} = r_k + c_1,\ t_{k'+1} = s_1 + c_1' \\ (u_k' p^s, 0, w', 0, p^{s'}) & : & t_k = r_k,\ t_{k+1} = 0,\ t_{k+2} = s_1 + c_1' \\ (p^s, 0, w', 0, p^{s'} v_0') & : & t_{k'} = r_k + c_1,\ t_{k'+1} = 0,\ t_{k'+2} = s_1 \\ (u_k' p^s, 0, w', 0, p^{s'} v_0') & : & t_k = r_k,\ t_{k+1} = 0,\ t_{k+2} = 0,\ t_{k+3} = s_1 \end{array}$$

mit $s, s' = \pm 1$, $t' \in \{0, s\}$, $u_k', v_0' \in \Sigma^+$, $w' \in (p^s \Sigma^* \cap \Sigma^* p^{s'}) \setminus \Sigma^* p^{\pm 2} \Sigma^*$, $c_1, c_1' \in \{-1, 0, 1\}$, $r_0 = -c_1$, $s_{\ell+1} = -c_1'$ und $k' = \max\{1, k\}$, wobei alle Gleichungen, in denen t_0 oder t_{m+1} vorkommen, weggelassen werden.

Z.B. erhalten wir für

$$\begin{array}{rcl} \sigma(X) & : & (acb\overline{b}cb) \\ \sigma(Y) & : & (\overline{b}\overline{c}b\overline{b}\overline{c}, -3, b\overline{b}\overline{c}b) \\ \sigma(Z) & : & (acb\overline{b}, 0, cb\overline{b}\overline{c}, -4, b\overline{b}\overline{c}b) \end{array}$$

mit $\Sigma = \{a, \overline{a}, b, \overline{b}, c, \overline{c}\}$, $I = \emptyset$, $c(M) = 2$ und $p = cb\overline{b}$ die Gleichungen

$$t_1 = 0, t_2 = s_1 - 1.$$

Wir eliminieren alle Gleichungen der Form $x = 0$, $x = y$ und $x = -y$ durch Substitution. Die Werte r_i, s_i und t_i, die wir aus den p-stabilen Normalformen gewonnen haben, geben uns eine Lösung für das Gleichungssystem. Wir ersetzen alle Variablen x, bei denen wir als Lösung $x = 0$ erhalten, durch 0, da, wenn man sie verändert, sich die Monoidelemente der Lösung ändern können. Es ist zwar

$$m^{xc(M)+\alpha} = m^{yc(M)+\alpha}$$

für $x, y > 0$ und $0 \leq \alpha < c(M)$, aber nicht unbedingt $m^{\alpha} = m^{c(M)+\alpha}$ für $0 \leq \alpha < c(M)$. Die übrigen Variablen x ersetzen wir durch $1 + x'$, wenn die Lösung für x größer 0 ist, und durch $-1 - x'$, wenn die Lösung für x kleiner 0 ist, wobei x' jeweils in $\mathbb{N}_0$ sein muß.

Wir erhalten in allen Fällen für jede Gleichung $XY = Z$ entweder höchstens zwei Gleichungen der Form $x - y = c$ oder $x = c$ mit $c \in \{0, 1\}$ oder eine Gleichung der Form $x - y - z = c$ oder $x - y = c$ mit $c \in \{0, 1, 2, 3\}$. Wenn noch zwei Gleichungen übrig bleiben, eliminieren wir eine der beiden durch Substitution. Nachdem wir dies für alle $d - 2$ Gleichungen $XY = Z$ gemacht haben, erhalten wir höchstens $d - 2$ nicht triviale Gleichung der Form

$$c_1 x - c_2 y - c_3 z = c_4$$

mit $c_1, c_2, c_3 \in \{0, 1\}$ und $|c_4| \leq 2d - 1$.

Wir haben jetzt ein System $N := \{\overrightarrow{n}_i \cdot \overrightarrow{x} + n_i = 0 \mid 1 \leq i \leq q\}$ mit $\overrightarrow{n}_i = (n_{i,1}, n_{i,2}, \ldots, n_{i,r})$ für $1 \leq i \leq q$ von linearen diophantischen Gleichungen. Es ist $q \leq d - 2$, $r \leq 3d - 6$, $\sum_{i=1}^{n} |n_i| \leq (2d - 1)(d - 2)$ und $\sum_{i=1}^{n} \sum_{j=1}^{r} |n_{i,j}| \leq 3d - 6$. Die Lösung $\overrightarrow{x} = (x_1, x_2, \ldots, x_r)$, die wir für dieses System haben, ist minimal bzgl. der natürlichen partiellen Ordnung für $\mathbb{N}_0^r$, da σ minimal ist und aus jeder kleineren Lösung des Gleichungssystems sich direkt eine kleinere Lösung für die Gleichung E ergibt. Damit

gilt nach Satz 5.0.8 also

$$x_i \leq ((2d-1)(d-2)+3d-6)e^{(3d-6)/e} \leq (2d^2-2d-4)2^{1,593d}$$

für $1 \leq i \leq r$. Durch die Substitutionen kann sich der absolute Wert der größten vorkommenden Variable maximal um $d-1$ verkleinert haben. Also ist

$$\begin{aligned} \exp(\sigma(L)) &\leq 1+2+(c(M)-1)+((2d^2-2d-4)2^{1,593d}+d-1)c(M) \\ &\leq 2c(M)d^2 2^{1,593d} \\ &\in \mathcal{O}(c(M)2^{1,6d}). \end{aligned}$$

□

Man beachte, daß der maximale Periodizitätsexponent nur von der denotationellen Länge d und den Konstanten g und $c(M)$ abhängt. Die Größe des Alphabets Σ hat dagegen keinen Einfluß.

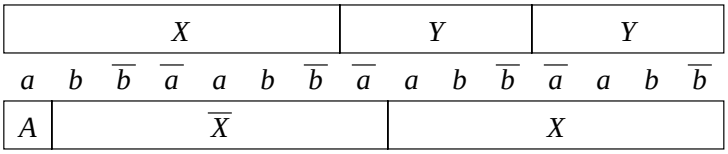
KAPITEL 6

Die Länge einer minimalen Lösung

Im weiteren sei eine FMA-Gleichung $E : L = R$ mit $L, R \in \Omega^*$, ein Monoid M, ein Homomorphismus $h : \Sigma^* \to M$, die regulären Randbedingungen $M_X \subseteq M$ für alle $X \in \Omega$ und eine minimale Lösung σ fest gegeben. Wir können o.B.d.A. voraussetzen, daß $|\sigma(L)| > 0$ ist, da wir die Länge der Lösung abschätzen wollen. Im weiteren können wir daher davon ausgehen, daß $|L| > 0$, $|R| > 0$ und $d \geq 2$ ist.

Als Beispiel wird in diesem Kapitel die Gleichung $E : XYY = A\overline{X}X$ mit $\Sigma = \{a, \overline{a}, b, \overline{b}\}$, $I = \emptyset$, $\Omega = \{A, \overline{A}, X, \overline{X}, Y, \overline{Y}\}$, das Monoid $M = \{1, m_a, m_1, m_2\}$, der Homomorphismus h mit $h(a) = m_a$, $h(\overline{a}) = m_1$, $h(b) = m_2$ und $h(\overline{b}) = m_2$, die regulären Randbedingungen $M_A = \{m_a\}$, $M_X = M$, $M_Y = \{m_2\}$ und die minimale Lösung $\sigma(A) = a$, $\sigma(X) = ab\overline{b}\overline{a}ab\overline{b}$ und $\sigma(Y) = \overline{a}ab\overline{b}$ benutzt. Es ist $d = 6$, $h^{-1}(1) = \epsilon$, $h^{-1}(m_a) = a$, $h^{-1}(m_1) = \overline{a}|(a|\overline{a})^{\geq 2}$ und $h^{-1}(m_2) = \Sigma^*(b|\overline{b})\Sigma^*$.

$\cdot$	1	m_a	m_1	m_2
1	1	m_a	m_1	m_2
m_a	m_a	m_1	m_1	m_2
m_1	m_1	m_1	m_1	m_2
m_2	m_2	m_2	m_2	m_2

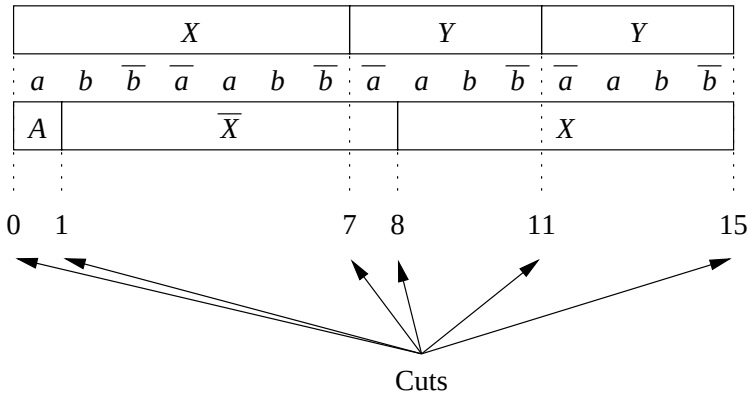


6.1. Charakteristische Wörter

Um die Länge der minimalen Lösung abzuschätzen, werden wir ℓ-Transformationen der minimalen Lösung betrachten. Die ℓ-Transformationen werden mit Hilfe von ℓ-Faktorisierungen definiert, die auf den charakteristischen Wörter der Länge ℓ basieren.

Wenn eine Variable X an der Position α in $\sigma(L)$ bzw. $\sigma(R)$ beginnt, sind α und $\alpha + |\sigma(X)|$ die Ränder der Variablen. Wir bezeichnen die Ränder $0 = \gamma_1 < \gamma_2 < \ldots < \gamma_k = |\sigma(L)|$ aller Variablen in $\sigma(L)$ bzw. $\sigma(R)$ als Cuts. Es ist $\gamma_1 = 0$, da der linke Rand der linkesten Variable bei 0 liegt, und es ist $\gamma_k = |\sigma(L)|$, da der rechte Rand der rechtesten Variable bei $|\sigma(L)|$ liegt. Es gibt maximal d Cuts, da jede Variable zwei Ränder hat und die Ränder paarweise identisch sind.

Beispiel 6.1.1. Für unser Beispiel ergeben sich die folgenden Cuts:



Es sei v ein Wort mit gerader Länge. Wir sagen, daß das Wort v in einem Wort w über einer Position α vorkommt, wenn v mittig über α in w liegt, d.h. $w[\alpha - |v|/2, \alpha + |v|/2] = v$. Die Wörter der Länge 2ℓ mit $\ell > 0$, die über den Cuts vorkommen, und ihre Inversen bezeichnen wir als die charakteristischen Wörter C_ℓ:

$$C_\ell := \{v \mid \exists 1 \leq i \leq k : \sigma(L)[\gamma_i - \ell, \gamma_i + \ell] = v \vee \sigma(L)[\gamma_i - \ell, \gamma_i + \ell] = \overline{v}\}.$$

Man beachte, daß immer $|C_\ell| \leq 2(d-2)$ ist, da es maximal d Cuts gibt und über dem linkesten Cut 0 und dem rechtesten Cut $|\sigma(L)|$ kein charakteristisches Wort vorkommen kann. Es ist immer $C_\ell = \overline{C_\ell}$.

Beispiel 6.1.2. Die charakteristischen Wörter für unser Beispiel sind:

$$\begin{aligned} C_1 &= \{ab, \overline{b}\overline{a}, \overline{a}a\} \\ C_2 &= \{b\overline{b}\overline{a}a, \overline{a}ab\overline{b}, \overline{b}\overline{a}ab\} \\ C_3 &= \{ab\overline{b}\overline{a}ab, \overline{b}\overline{a}ab\overline{b}\overline{a}, b\overline{b}\overline{a}ab\overline{b}\} \\ &\vdots \end{aligned}$$

6.2. Die Äquivalenzrelation ≈

Wir werden die Äquivalenzrelation ≈ mit Hilfe der Relation ∼ definieren.

Definition 6.2.1. Für die Gleichung $E : L = R$ sei $L = X_{i_1}X_{i_2}\cdots X_{i_c}$ und $R = X_{i_{c+1}}X_{i_{c+2}}\cdots X_{i_d}$ mit $\Omega = \{X_1, X_2, \ldots, X_{|\Omega|}\}$. Es sei $\delta_j := |\sigma(X_{i_1}X_{i_2}\cdots X_{i_{j-1}})|$ für $1 \leq j \leq c$ und $\delta_j := |\sigma(X_{i_{c+1}}X_{i_{c+2}}\cdots X_{i_{j-1}})|$ für $c+1 \leq j \leq d$. Dann gilt

$$[\delta_j + \alpha, \delta_j + \beta] \sim [\delta_k + \mu, \delta_k + \nu],$$

wenn $1 \leq j, k \leq d$, $0 \leq \alpha, \beta \leq |\sigma(X_{i_j})|$, $\alpha \neq \beta$ und

- $X_{i_j} = X_{i_k}$, $\alpha = \mu$ und $\beta = \nu$ ist oder
- $X_{i_j} = \overline{X_{i_k}}$, $\alpha = |\sigma(X_{i_j})| - \mu$ und $\beta = |\sigma(X_{i_j})| - \nu$

ist.

Man beachte, daß

1. $[\alpha, \beta] \sim [\mu, \nu]$ genau dann ist, wenn $[\beta, \alpha] \sim [\nu, \mu]$ ist.
2. aus $[\alpha, \beta] \sim [\mu, \nu]$ folgt, daß $\sigma(L)[\alpha, \beta] = \sigma(L)[\mu, \nu]$ ist, da die Intervalle dem gleichen Teilwort aus derselben Variable oder der inversen Variable entsprechen.
3. wenn $[\alpha_1, \beta_1] \sim [\alpha_2, \beta_2] \sim [\alpha_3, \beta_3]$ ist, $[\alpha_1, \beta_1]$ und $[\alpha_3, \beta_3]$ nicht einem Teilwort derselben Variablen oder der inversen Variable entsprechen müssen, da $[\alpha_2, \beta_2]$ sowohl einem Teilwort einer Variablen in L als

auch einem Teilwort einer anderen Variablen in R entsprechen kann. Es ist aber natürlich $\sigma(L)[\alpha_1, \beta_1] = \sigma(L)[\alpha_2, \beta_2] = \sigma(L)[\alpha_3, \beta_3]$.

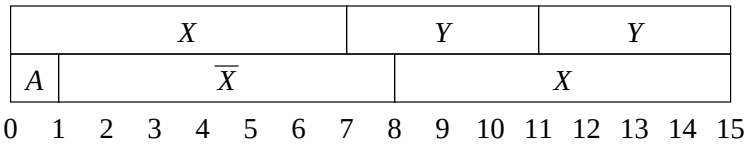
Die Relation $\sim$ ist bereits reflexiv und symmetrisch. Wir benötigen im weiteren die transitive Hülle von $\sim$.

DEFINITION 6.2.2. Mit $\approx$ bezeichnen wir die von $\sim$ erzeugte Äquivalenzrelation.

BEISPIEL 6.2.3. Für unser Beispiel ergeben sich folgende Äquivalenzklassen mit mehr als einem Element:

$$[0,1] \approx [8,9] \approx [12,13] \approx [4,5] \approx [4,3] \approx [12,11] \approx [8,7]$$
$$[1,0] \approx [9,8] \approx [13,12] \approx [5,4] \approx [3,4] \approx [11,12] \approx [7,8]$$
$$[1,2] \approx [9,10] \approx [13,14] \approx [5,6] \approx [3,2] \approx [11,10] \approx [15,14] \approx [7,6]$$
$$[2,1] \approx [10,9] \approx [14,13] \approx [6,5] \approx [2,3] \approx [10,11] \approx [14,15] \approx [6,7]$$

$$[0,2] \approx [8,10] \approx [12,14] \approx [4,6] \approx [4,2] \approx [12,10] \approx [8,6]$$
$$[2,0] \approx [10,8] \approx [14,12] \approx [6,4] \approx [2,4] \approx [10,12] \approx [6,8]$$
$$[1,3] \approx [9,11] \approx [13,15] \approx [5,7] \approx [3,1] \approx [11,9] \approx [15,13] \approx [7,5]$$
$$[3,5] \approx [11,13] \approx [7,9] \approx [5,3] \approx [13,11] \approx [9,7]$$
$$\vdots$$
$$[0,7] \approx [8,15] \approx [8,1]$$
$$[7,0] \approx [15,8] \approx [1,8]$$



6.3. Freie Intervalle

Wir sagen, daß ein Intervall $[\alpha, \beta]$ in einem Intervall $[\mu, \nu]$ enthalten ist bzw. $[\alpha, \beta] \subseteq [\mu, \nu]$ ist, wenn $\min\{\mu, \nu\} \leq \alpha, \beta \leq \max\{\mu, \nu\}$ ist. Das Intervall

ist echt enthalten bzw. $[\alpha, \beta] \subset [\mu, \nu]$, wenn zusätzlich $|\nu - \mu| > |\beta - \alpha|$ ist. Wir sagen, daß eine Position γ in einem Intervall $[\alpha, \beta]$ enthalten ist bzw. $\gamma \in [\alpha, \beta]$ ist, wenn $\min\{\alpha, \beta\} < \gamma < \max\{\alpha, \beta\}$ ist.

DEFINITION 6.3.1. Ein Intervall $[\alpha, \beta]$ heißt frei, wenn aus $[\mu, \nu] \approx [\alpha, \beta]$ folgt, daß keine Cuts im Intervall $[\mu, \nu]$ enthalten sind. Ein freies Intervall $[\alpha, \beta]$ ist maximal, wenn kein freies Intervall $[\mu, \nu]$ existiert, so daß $[\alpha, \beta]$ in $[\mu, \nu]$ echt enthalten ist.

Man beachte, daß die folgenden Lemmata 6.3.2, 6.3.3, 6.3.4 und 6.3.5 nicht verwenden, daß σ eine minimale Lösung ist. Sie gelten also auch für nicht minimale Lösungen.

LEMMA 6.3.2. *Es sei $[\alpha, \beta]$ ein freies Intervall. Dann ist auch das Intervall $[\beta, \alpha]$ frei.*

BEWEIS. Nehmen wir an, daß das Intervall $[\beta, \alpha]$ nicht frei ist. Dann gibt es einen Cut γ und Intervalle $[\beta_i, \alpha_i]$, so daß

$$[\beta, \alpha] = [\beta_1, \alpha_1] \sim [\beta_2, \alpha_2] \sim \cdots \sim [\beta_k, \alpha_k]$$

und $\gamma \in [\beta_k, \alpha_k]$ ist. Aufgrund der Definition von $\sim$, ist

$$[\alpha, \beta] = [\alpha_1, \beta_1] \sim [\alpha_2, \beta_2] \sim \cdots \sim [\alpha_k, \beta_k].$$

Also ist auch $[\alpha, \beta]$ nicht frei. Dies ist ein Widerspruch. Also ist $[\beta, \alpha]$ frei. □

Man beachte, daß, wenn $[\alpha, \beta]$ ein maximales freies Intervall ist, auch $[\beta, \alpha]$ ein maximales freies Intervall ist.

LEMMA 6.3.3. *Es seien $[\alpha, \beta]$ und $[\mu, \nu]$ freie Intervalle mit $\alpha < \mu < \beta < \nu$. Dann ist das Intervall $[\alpha, \nu]$ frei.*

BEWEIS. Nehmen wir an, daß das Intervall $[\alpha, \nu]$ nicht frei ist. Dann gibt es einen Cut γ und Intervalle $[\alpha_i, \nu_i]$, so daß $[\alpha, \nu] = [\alpha_1, \nu_1] \sim [\alpha_2, \nu_2] \sim \cdots \sim [\alpha_k, \nu_k]$ und $\gamma \in [\alpha_k, \nu_k]$ ist. Da $[\alpha, \beta] \subset [\alpha, \nu]$ und $[\mu, \nu] \subset [\alpha, \nu]$ ist

und aufgrund der Definition von $\sim$, ist

$$\begin{aligned}[\alpha,\beta] &= [\alpha_1, \alpha_1 + s_1(\beta-\alpha)] \sim \cdots \sim [\alpha_k, \alpha_k + s_k(\beta-\alpha)],\\ [\mu,\nu] &= [\nu_1 - s_1(\nu-\mu), \nu_1] \sim \cdots \sim [\nu_k - s_k(\nu-\mu), \nu_k]\end{aligned}$$

mit $s_i := \text{sign}(\nu_i - \alpha_i)$. Da sowohl $[\alpha,\beta]$ als auch $[\mu,\nu]$ frei sind, gibt es keinen Cut γ, so daß $\gamma \in [\alpha_k, \alpha_k + s_k(\beta-\alpha)]$ oder $\gamma \in [\nu_k - s_k(\nu-\mu), \nu_k]$ ist. Da $(\beta-\alpha)+(\nu-\mu) > \nu-\alpha = |\nu_k - \alpha_k|$ ist, gibt es also keinen Cut γ, so daß $\gamma \in [\alpha_k, \nu_k]$ ist. Dies ist ein Widerspruch. Also ist $[\alpha,\nu]$ frei. □

LEMMA 6.3.4. *Es seien $[\alpha,\beta]$ und $[\mu,\nu]$ maximale freie Intervalle mit $\alpha < \beta$, $\mu < \nu$ und $\alpha < \mu$. Dann ist $\beta \leq \mu$, d.h. maximale freie Intervalle überlappen sich nicht.*

BEWEIS. Wenn $\nu \leq \beta$ ist, dann ist $\alpha < \mu < \nu \leq \beta$ und damit das freie Intervall $[\mu,\nu]$ nicht maximal. Dies ist ein Widerspruch. Also ist $\beta < \nu$.

Nehmen wir an, daß $\mu < \beta$ ist. Dann ist $\alpha < \mu < \beta < \nu$. Nach Lemma 6.3.3 ist auch das Intervall $[\alpha,\nu]$ frei. Das Intervall $[\mu,\nu]$ ist also nicht maximal, da $\alpha < \mu < \nu = \nu$ ist. Dies ist ein Widerspruch. Also ist $\beta \leq \mu$. □

Man beachte, daß, wenn $[\alpha,\beta]$ ein maximales freies Intervall ist, das maximale freie Intervall $[\beta,\alpha]$ in einer anderen oder auch derselben Äquivalenzklasse wie $[\alpha,\beta]$ liegen kann.

LEMMA 6.3.5. *Es gibt maximal $2d-2$ verschiedene Äquivalenzklassen von maximalen freie Intervallen.*

BEWEIS. Es seien $[\alpha,\beta]$ und $[\mu,\nu]$ maximale freie Intervalle, die nicht in derselben Äquivalenzklasse sind. Als erstes werden wird zeigen, daß es Intervalle $[\alpha',\beta'] \approx [\alpha,\beta]$ und $[\mu',\nu'] \approx [\mu,\nu]$ gibt, so daß β' und ν' Cuts sind.

Wenn $\beta = 0$ oder $\beta = |\sigma(L)|$ ist, sind wir fertig, da dann $\beta' = \beta$ ein Cut ist. Das Intervall $[\alpha, \beta + s]$ mit $s := \text{sign}(\beta - \alpha)$ ist nicht frei, da $[\alpha,\beta]$ maximal ist. Es gibt also ein Intervall $[\alpha', \beta' + s']$ mit $s' := \text{sign}(\beta' - \alpha')$, so daß $[\alpha, \beta + s] \approx [\alpha', \beta' + s']$ ist und es einen Cut $\gamma \in [\alpha', \beta' + s']$ gibt. Aufgrund der Definition von $\sim$ ist auch $[\alpha,\beta] \approx [\alpha',\beta']$. Da das Intervall $[\alpha,\beta]$ frei ist, gibt es keinen Cut $\gamma' \in [\alpha',\beta']$. Also ist $\gamma = \beta'$ und es gibt

ein Intervall $[\alpha', \beta'] \approx [\alpha, \beta]$, so daß β' ein Cut ist. Entsprechend können wir zeigen, daß es ein Intervall $[\mu', \nu']$ gibt, so daß ν' ein Cut ist.

Da die Intervalle $[\alpha, \beta]$ und $[\mu, \nu]$ maximal und nicht in derselben Äquivalenzklasse sind, muß $\beta' \neq \nu'$ oder $\text{sign}(\beta' - \alpha') \neq \text{sign}(\nu' - \mu')$ sein. Für jeden Cut γ kann es nur zwei maximale freie Intervalle $[\lambda, \gamma]$ geben: Eines links vom Cut ($\lambda < \gamma$) und eines rechts vom Cut ($\gamma < \lambda$). Da links vom Cut 0 und rechts vom Cut $|\sigma(L)|$ kein maximales freies Intervall liegen kann und es maximal d Cuts gibt, kann es maximal $2d - 2$ Äquivalenzklassen von maximalen freien Intervallen geben. □

Betrachten wir wieviele verschiedene Wörter diesen Äquivalenzklassen entsprechen. Die Äquivalenzklasse eines maximalen freien Intervalls $[\alpha, \beta]$ kann den zwei Wörtern $\sigma(L)[\alpha, \beta]$ und $\overline{\sigma(L)[\alpha, \beta]}$ entsprechen. Doch, wenn $[\alpha, \beta] \approx [\beta, \alpha]$ ist, gilt $\sigma(L)[\alpha, \beta] = \overline{\sigma(L)[\alpha, \beta]}$ und, wenn $[\alpha, \beta] \not\approx [\beta, \alpha]$ ist, gilt für die Äquivalenzklassen der maximalen freien Intervalle $[\alpha, \beta]$ und $[\beta, \alpha]$, daß $\sigma(L)[\alpha, \beta] = \overline{\sigma(L)[\beta, \alpha]}$ und $\overline{\sigma(L)[\alpha, \beta]} = \sigma(L)[\beta, \alpha]$ ist. Es gibt also maximal soviele verschiedene Wörter, wie es Äquivalenzklassen gibt.

Da alle Intervalle in derselben Äquivalenzklasse demselben Wort oder seinem Inversen entsprechen und das Wort $\sigma(L)$ sich aus maximalen freien Intervallen zusammensetzt, ist das Wort $\sigma(L)$ eine Konkatenation aus maximal $2d-2$ verschiedenen Wörtern und jedes dieser Wörter entspricht einer Äquivalenzklasse eines maximalen freien Intervalls. In der Konkatenation kann jedes der Wörter natürlich beliebig oft vorkommen. Diese Wörter sind sozusagen die Bausteine, aus denen die Lösung zusammengesetzt ist. Man kann weitere Lösungen konstruieren, indem man die Wörter w und $\overline{w}$, die einem maximalen freien Intervall $[\alpha, \beta]$ entsprechen, durch andere Wörter v und $\overline{v}$ ersetzt. Dabei muß man darauf achten, daß sich sowohl die Monoidelemente nicht ändern, d.h. $h(v) = h(w)$ und $h(\overline{v}) = h(\overline{w})$ ist, als auch $v = \overline{v}$ ist, wenn $[\alpha, \beta] \approx [\beta, \alpha]$ ist.

Für das folgende Lemma ist es wieder notwendig, das σ eine minimale Lösung ist.

LEMMA 6.3.6. *Es sei $[\alpha, \beta]$ ein freies Intervall. Dann ist $|\beta - \alpha| \leq 2max_M + 1 < 2|M|^2$.*

BEWEIS. Das freie Intervall $[\alpha, \beta]$ ist in einem maximalen freien Intervall $[\mu, \nu]$ enthalten. Aufgrund von Lemma 6.3.4 überlappen sich alle maximalen Intervalle nicht. Alle Intervalle in der Äquivalenzklasse von $[\mu, \nu]$ enthalten keinen Cut, da $[\mu, \nu]$ frei ist, und sie entsprechen denselben Wörtern w und $\overline{w}$. Wenn wir alle diese Intervalle durch die Wörter v und $\overline{v}$ aus Lemma 3.7.2 ersetzen, erhalten wir auch eine Lösung, da die regulären Randbedingungen auch von v erfüllt werden und auch $v = \overline{v}$ gilt, wenn $[\mu, \nu] \approx [\nu, \mu]$ ist. Da die Lösung σ minimal ist, gilt

$$|\nu - \mu| = |\sigma(L)[\mu, \nu]| \leq |v| \leq 2\text{max}_M + 1$$

und damit auch $|\beta - \alpha| \leq 2\text{max}_M + 1 < 2|M|^2$. □

Von besonderer Bedeutung ist die Negation von Lemma 6.3.6. Jedes Intervall $[\alpha, \beta]$ mit $|\beta - \alpha| > 2\text{max}_M + 1$ ist nicht frei, d.h. es gibt ein Intervall $[\mu, \nu] \approx [\alpha, \beta]$, so daß das Intervall $[\mu, \nu]$ einen Cut γ enthält. Da $\sigma(L)[\mu, \nu] = \sigma(L)[\alpha, \beta]$ ist, kommt also für jedes Teilwort $v = \sigma(L)[\alpha, \beta]$ von $\sigma(L)$ mit $|v| > 2\text{max}_M + 1$ entweder das Wort v selbst oder sein Inverses $\overline{v}$ über einem Cut γ in $\sigma(L)$ vor.

6.4. ℓ-Faktorisierungen

Eine Faktorisierung ist eine Funktion

$$F : \Sigma^* \to (D \times \Sigma^+ \times D)^*$$

wobei D eine beliebige Menge ist. Diese Definition weicht von der ursprünglichen Definition in [**KPM97**] ab. Eine ähnliche Definition wird in [**P99B**] als Quasi-Faktorisierung bezeichnet. Da wir jedoch nur diese Quasi-Faktorisierungen betrachten werden, nennen wir sie einfach Faktorisierungen.

Die ℓ-Faktorisierung F_ℓ für ein Wort w ist durch die Menge C_ℓ der charakteristischen Wörter von E der Länge ℓ definiert. Für ein Wort w seien $0 < \mu_1 < \mu_2 < \ldots < \mu_k < |w|$ alle Positionen über denen ein Wort aus C_ℓ in w vorkommt. Es ist

$$\{\mu_i \mid 1 \leq i \leq k\} = \{\mu \mid w[\mu - \ell, \mu + \ell] \in C_\ell\}.$$

Das Wort aus C_ℓ, das über der Position μ_i vorkommt, sei v_i, d.h. $v_i := w[\mu_i - \ell, \mu_i + \ell]$. Dann ist

$$F_\ell(w) := (\$, w[0, \mu_1], v_1)(v_1, w[\mu_1, \mu_2], v_2) \cdots (v_k, w[\mu_k, |w|], \$)$$

die ℓ-Faktorisierung des Wortes w, wobei \$ ein neues Symbol ist mit $\overline{\$} = \$$. Entsprechend sagen wir, daß \$ über 0 und $|w|$ vorkommt. Die ℓ-Faktorisierung ist also eine Sequenz

$$W = (v_0, w_1, v_1)(v_1, w_2, v_2) \cdots (v_k, w_{k+1}, v_{k+1})$$

von Tripeln $(v_{i-1}, w_i, v_i) \in D_\ell \times \Sigma^+ \times D_\ell$ mit $D_\ell := C_\ell \cup \{\$\}$. Die Tripel (v_{i-1}, w_i, v_i) sind die Faktoren des Wortes w und repräsentieren die Teilwörter w_i. Das Inverse für einen Faktor (v_{i-1}, w_i, v_i) sei $\overline{(v_{i-1}, w_i, v_i)} := (\overline{v_i}, \overline{w_i}, \overline{v_{i-1}})$. Wir bezeichnen die Anzahl der Faktoren mit $|W|$, die Konkatenation der repräsentierten Teilwörter aller Faktoren mit $\mathrm{concat}(W)$, den ersten Faktor mit $\mathrm{kopf}(W)$, die Sequenz aus allen Faktoren bis auf den ersten und letzten mit $\mathrm{rumpf}(W)$, den letzten Faktor mit $\mathrm{rest}(W)$, die inverse Sequenz mit $\overline{W}$ und die leere Sequenz ohne Faktoren mit ϵ.

$$\begin{aligned}
|W| &:= k+1 \\
\mathrm{concat}(W) &:= w_1 w_2 \ldots w_{k+1} \\
\mathrm{kopf}(W) &:= (v_0, w_1, v_1) \\
\mathrm{rumpf}(W) &:= (v_1, w_2, v_2)(v_2, w_3, v_3) \cdots (v_{k-1}, w_k, v_k) \\
\mathrm{rest}(W) &:= (v_k, w_{k+1}, v_{k+1}) \\
\overline{W} &:= (\overline{v_{k+1}}, \overline{w_{k+1}}, \overline{v_k})(\overline{v_k}, \overline{w_k}, \overline{v_{k-1}}) \cdots (\overline{v_1}, \overline{w_1}, \overline{v_0})
\end{aligned}$$

Wenn $|W| = 0$ ist, sei $\mathrm{kopf}(W) = \epsilon$, $\mathrm{rumpf}(W) = \epsilon$ und $\mathrm{rest}(W) = \epsilon$. Wenn $|W| = 1$ ist, sei $\mathrm{kopf}(W) = (v_0, w_1, v_1)$, $\mathrm{rumpf}(W) = \epsilon$ und $\mathrm{rest}(W) = \epsilon$. Für zwei Sequenzen W_1 und W_2 sei $W_1 W_2$ die Konkatenation der beiden Sequenzen. Mit W^r bezeichnen wir die r-fache Konkatenation der Sequenz W. Die Sonderfälle für $|W| = 0$ und $|W| = 1$ haben wir eingeführt, damit immer $W = \mathrm{kopf}(W)\mathrm{rumpf}(W)\mathrm{rest}(W)$ ist.

BEISPIEL 6.4.1. Mit $C_2 = \{b\overline{b}\overline{a}a, \overline{a}ab\overline{b}, \overline{b}\overline{a}ab\}$ ist

$$\begin{aligned} F_2(\epsilon) &= \epsilon \\ F_2(a) &= (\$, a, \$) \\ F_2(ab\overline{b}\overline{a}a) &= (\$, ab\overline{b}, b\overline{b}\overline{a}a)(b\overline{b}\overline{a}a, \overline{a}a, \$) \\ F_2(ab\overline{b}\overline{a}ab\overline{b}) &= (\$, ab\overline{b}, b\overline{b}\overline{a}a)(b\overline{b}\overline{a}a, \overline{a}, \overline{b}\overline{a}ab)(\overline{b}\overline{a}ab, a, \overline{a}ab\overline{b})(\overline{a}ab\overline{b}, b\overline{b}, \$) \end{aligned}$$

und für $W = F_2(ab\overline{b}\overline{a}ab\overline{b})$ ist

$$\begin{aligned} |W| &= 4 \\ \text{concat}(W) &= ab\overline{b}\overline{a}ab\overline{b} \\ \text{kopf}(W) &= (\$, ab\overline{b}, b\overline{b}\overline{a}a) \\ \text{rumpf}(W) &= (b\overline{b}\overline{a}a, \overline{a}, \overline{b}\overline{a}ab)(\overline{b}\overline{a}ab, a, \overline{a}ab\overline{b}) \\ \text{rest}(W) &= (\overline{a}ab\overline{b}, b\overline{b}, \$) \\ \overline{W} &= (\$, b\overline{b}, b\overline{b}\overline{a}a)(b\overline{b}\overline{a}a, \overline{a}, \overline{b}\overline{a}ab)(\overline{b}\overline{a}ab, a, \overline{a}ab\overline{b})(\overline{a}ab\overline{b}, b\overline{b}\overline{a}, \$) \end{aligned}$$

Für $1 \leq i \leq |W| + 1$ sei $\delta_i := |w_1 w_2 \cdots w_{i-1}|$. Für $1 \leq i \leq |W|$ ist δ_i die Startposition des Faktors (v_{i-1}, w_i, v_i) in W. Mit $W[\alpha, \beta]$ bezeichnen wir die Teilsequenz von W, die $w[\alpha, \beta]$ entspricht, wobei $\alpha, \beta \in \{\delta_i \mid 1 \leq i \leq |W| + 1\}$ und $\alpha \leq \beta$ sein muß. Wenn $\alpha = \delta_i$ und $\beta = \delta_j$ ist, dann ist

$$W[\alpha, \beta] := (v_{i-1}, w_i, v_i)(v_i, w_{i+1}, v_{i+1}) \cdots (v_{j-2}, w_{j-1}, v_{j-1})$$

Wenn wir im weiteren $W[\alpha, \beta]$ schreiben, gilt implizit $\alpha, \beta \in \{\delta_i \mid 1 \leq i \leq |W| + 1\}$, da sonst $W[\alpha, \beta]$ nicht definiert ist.

Aufgrund des \$ Symbols am Anfang und Ende von $F_\ell(w[\alpha, \beta])$ ist meistens $F_\ell(w)[\alpha, \beta] \neq F_\ell(w[\alpha, \beta])$. Es ist aber auch im allgemeinen nicht $|F_\ell(w)[\alpha, \beta]| = |F_\ell(w[\alpha, \beta])|$. So ist für $w = ab\overline{b}\overline{a}ab\overline{b}$:

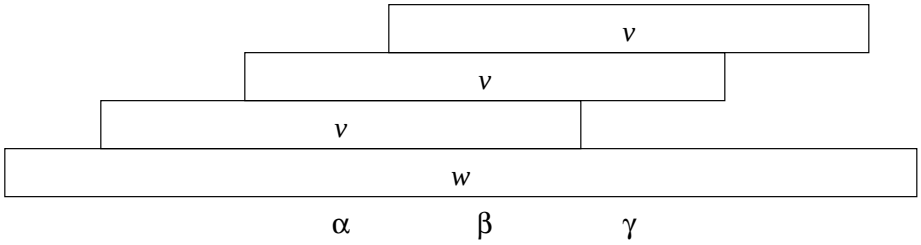
$$\begin{aligned} F_2(w)[0, 5] &= (\$, ab\overline{b}, b\overline{b}\overline{a}a)(b\overline{b}\overline{a}a, \overline{a}, \overline{b}\overline{a}ab)(\overline{b}\overline{a}ab, a, \overline{a}ab\overline{b}), \\ F_2(w[0, 5]) &= (\$, ab\overline{b}, b\overline{b}\overline{a}a)(b\overline{b}\overline{a}a, \overline{a}a, \$). \end{aligned}$$

Der Unterschied entsteht, weil das Wort $\overline{b}\overline{a}ab$ über der Position 4 in w vorkommt, aber nicht über der Position 4 in $w[0, 5] = ab\overline{b}\overline{a}a$ vorkommt, da das b bereits vorher abgeschnitten wurde.

Anstelle von $\mathrm{kopf}(F_\ell(w))$, $\mathrm{rumpf}(F_\ell(w))$, $\mathrm{rest}(F_\ell(w))$ und $h(\mathrm{concat}(W))$ werden wir im weiteren $\mathrm{kopf}_\ell(w)$, $\mathrm{rumpf}_\ell(w)$, $\mathrm{rest}_\ell(w)$ und $h(W)$ schreiben. Es ist $F_\ell(\overline{w}) = \overline{F_\ell(w)}$ und $\mathrm{rumpf}_\ell(\overline{w}) = \overline{\mathrm{rumpf}_\ell(w)}$, da $C_\ell = \overline{C_\ell}$ ist. Es ist auch $\mathrm{kopf}_\ell(\overline{w}) = \overline{\mathrm{rest}_\ell(w)}$ und $\mathrm{rest}_\ell(\overline{w}) = \overline{\mathrm{kopf}_\ell(w)}$, wenn $\mathrm{rumpf}_\ell(w) \neq ()$ ist.

6.5. Exponentielle Ausdrücke

LEMMA 6.5.1. *Es seien $\alpha < \beta < \gamma$ drei aufeinanderfolgende Positionen über denen ein Wort $v \in C_\ell$ in w vorkommt, d.h. über den Positionen zwischen α und γ kommt v außer bei β nicht vor. Wenn $\gamma \leq \alpha + 2\ell$ ist, dann ist $F_\ell(w)[\alpha, \beta] = F_\ell(w)[\beta, \gamma]$.*



BEWEIS. Sowohl $\beta - \alpha$ als auch $\gamma - \beta$ sind Perioden von v. Nach Lemma 3.2.4 ist $\mathrm{ggt}(\beta - \alpha, \gamma - \beta)$ eine Periode von v, da $(\beta - \alpha) + (\gamma - \beta) = \gamma - \alpha \leq 2\ell = |v|$ ist. Da $\mathrm{ggt}(\beta - \alpha, \gamma - \beta) \leq \beta - \alpha$ ist und über den Positionen zwischen α und β das Wort v nicht vorkommt, ist $\mathrm{ggt}(\beta - \alpha, \gamma - \beta) = \beta - \alpha$. Da $\mathrm{ggt}(\beta - \alpha, \gamma - \beta) \leq \gamma - \beta$ ist und über den Positionen zwischen β und γ das Wort v nicht vorkommt, ist $\mathrm{ggt}(\beta - \alpha, \gamma - \beta) = \gamma - \beta$. Also ist $\beta - \alpha = \gamma - \beta$ und $w[\alpha, \beta] = w[\beta, \gamma]$. Es ist $F_\ell(w)[\alpha, \beta] = F_\ell(w)[\beta, \gamma]$, da

$$w[\alpha - \ell, \beta + \ell] = v[0, \ell]w[\alpha, \beta]v[\ell, 2\ell] = w[\beta - \ell, \gamma + \ell]$$

ist. □

Als nächstes werden wir zeigen, daß sich Teilsequenzen einer ℓ-Faktorisierung effizient mit exponentiellen Ausdrücken darstellen lassen.

LEMMA 6.5.2. *Gegeben sei eine Sequenz $F_\ell(w)[\alpha,\beta]$ mit $\beta-\alpha \leq 2\ell$. Dann existieren Sequenzen W_i, so daß*

$$F_\ell(w)[\alpha,\beta] = W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}$$

ist mit $1 \leq s \leq 4d-7$, $|W_i| \leq 2d-3$ und $r_1, r_s = 1$.

BEWEIS. Es seien $\alpha = \mu_0 < \mu_1 < \cdots < \mu_{m+1} = \beta$ alle Positionen zwischen α und β über denen ein Wort aus D_ℓ in w vorkommt. Es sei v_i das Wort aus D_ℓ, das über μ_i vorkommt. Wir bestimmen die Sequenzen W_i nacheinander, wobei die folgenden Invarianten gelten:

1. Es existieren Sequenzen W_i, so daß $F_\ell(w)[\alpha,\mu_{j_k}] = W_1^{r_1} W_2^{r_2} \cdots W_{2k}^{r_{2k}}$ ist.
2. Es gibt höchstens $2d-4-k$ unterschiedliche Wörter, die über den Position $\mu_{j_k+1}, \mu_{j_k+2}, \ldots, \mu_m$ vorkommen, d.h. $|\{v_i \mid j_k < i < m+1\}| \leq 2d-4-k$.

Induktion über k bis $j_k = m$ ist:

Induktionsanfang: Für $k = 0$ und $j_0 := 0$ sind die Invarianten erfüllt, da nur Wörter aus C_ℓ mit $|C_\ell| \leq 2d-4$ über den Positionen $\mu_1, \mu_2, \ldots, \mu_m$ vorkommen.

Induktionsvoraussetzung: Die Invarianten gelten für k.

Induktionsschritt: Wir betrachten die Wörter v_i, die über den Positionen $\mu_{j_k+1} < \mu_{j_k+2} < \cdots < \mu_m$ vorkommen, von links nach rechts. Wenn jedes Wort nur einmal vorkommt, setzen wir $j_{k+1} := m$, $W_{2k+1} := F_\ell(w)[\mu_{j_k}, \mu_{j_{k+1}}]$, $r_{2k+1} := 1$, $W_{2k+2} := \epsilon$ und $r_{2k+2} := 1$. Sonst sei v das erste Wort, das von links aus gesehen mehrfach vorkommt. Die Startpositionen von v seien $\mu_{i_1} < \mu_{i_2} < \cdots < \mu_{i_r}$. Wir setzen $j_{k+1} := i_r$, $W_{2k+1} := F_\ell(w)[\mu_{j_k}, \mu_{i_1}]$, $r_{2k+1} := 1$, $W_{2k+2} := F_\ell(w)[\mu_{i_1}, \mu_{i_2}]$ und $r_{2k+2} := r-1$. Da v das erste Wort war, das mehrfach vorkam, kommen alle Wörter über den Positionen $\mu_{j_k+1} < \mu_{j_k+2} < \cdots < \mu_{i_2-1}$ nur einmal vor und es ist $|W_{2k+1}|, |W_{2k+2}| \leq |C_\ell| \leq 2d-4$. Es ist $\mu_{i_r} - \mu_{i_1} \leq \beta - \alpha \leq 2\ell$ und aufgrund von Lemma 6.5.1 wiederholt sich die Sequenz W_{2k+2} immer wieder. Nach μ_{i_r} kommt das Wort v nicht mehr vor.

Damit haben wir jetzt den exponentiellen Ausdruck für $F_\ell(w)[\alpha, \mu_m]$ bestimmt. Wir setzen noch $s := 2k + 1$ und $W_s := (v_m, w[\mu_m, \mu_{m+1}], v_{m+1})$. Da die Anzahl der unterschiedlichen Wörter jedesmal um eins abnimmt, ist $k \leq 2d - 4$ und $s \leq 4d - 7$. □

Wir können auch Teilsequenzen, die ein Wort repräsentieren, das länger als 2ℓ ist, mit exponentiellen Ausdrücken darstellen.

LEMMA 6.5.3. *Gegeben sei eine Sequenz $F_\ell(w)[\alpha, \beta]$. Dann existieren Sequenzen W_i, so daß*

$$F_\ell(w)[\alpha, \beta] = W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}$$

ist mit $1 \leq s \leq \left\lceil \frac{\beta-\alpha}{2\ell} \right\rceil (4d - 7)$, $|W_i| \leq 2d - 2$ und $r_1 = r_s = 1$.

BEWEIS. Induktion über $\beta - \alpha$:

Induktionsanfang: Für $\beta - \alpha = 0$ ist die Aussage trivial und für $1 \leq \beta - \alpha \leq 2\ell$ ergibt sie sich direkt aus Lemma 6.5.2.

Induktionsvoraussetzung: Es sei $\beta - \alpha > 2\ell$ und für $\beta' - \alpha' \leq \beta - \alpha - 2\ell$ sei die Aussage wahr.

Induktionsschritt: Es seien $\alpha = \mu_0 < \mu_1 < \cdots < \mu_{m+1} = \beta$ alle Positionen zwischen α und β über denen ein Wort aus D_ℓ in w vorkommt. Es sei v_i das Wort aus D_ℓ, das über μ_i vorkommt. Wir wählen j, so daß μ_j die größte Position ist mit $\mu_j \leq 2\ell$. Wir bestimmen den exponentiellen Ausdruck für $F_\ell(w)[\alpha, \mu_j]$ mit Lemma 6.5.2. An die letzte Sequenz des Ausdrucks hängen wir noch den Faktor $(v_j, w[\mu_j, \mu_{j+1}], v_{j+1})$ an. Als nächstes bestimmen wir induktiv den exponentiellen Ausdruck für $F_\ell(w)[\mu_{j+1}, \beta]$. Wir konkatenieren die beiden Ausdrücke und erhalten einen Ausdruck mit

$$s \leq 4d - 7 + \left\lceil \frac{\beta - \alpha - 2\ell}{2\ell} \right\rceil (4d - 7) \leq \left\lceil \frac{\beta - \alpha}{2\ell} \right\rceil (4d - 7).$$

□

LEMMA 6.5.4. *Gegeben sei eine Sequenz $F_\ell(w)[\alpha, \beta]$. Dann existieren Sequenzen W_i, so daß*

$$F_\ell(w)[\alpha, \beta] = W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s} \mathit{rumpf}_\ell(w[\alpha, \beta]) W_{s+1}^{r_{s+1}} W_{s+2}^{r_{s+2}} \cdots W_t^{r_t}$$

ist mit $1 \leq s \leq t \leq 8d-14$, $|W_i| \leq 2d-2$,

$$\begin{aligned} \mathit{concat}(W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}) &= \mathit{concat}(\mathit{kopf}_\ell(w[\alpha,\beta])), \\ \mathit{concat}(W_{s+1}^{r_{s+1}} W_{s+2}^{r_{s+2}} \cdots W_t^{r_t}) &= \mathit{concat}(\mathit{rest}_\ell(w[\alpha,\beta])). \end{aligned}$$

Beweis. Es seien $\alpha = \mu_0 < \mu_1 < \cdots < \mu_{m+1} = \beta$ alle Positionen zwischen α und β über denen ein Wort aus D_ℓ in w vorkommt. Es sei v_i das Wort aus D_ℓ, das über μ_i vorkommt.

Wenn ein i existiert, so daß $\alpha+\ell \leq \mu_i \leq \beta+\ell$ ist, wählen wir j und k, so daß μ_j die kleinste und μ_k die größte Position ist, für die $\alpha+\ell \leq \mu_j, \mu_k \leq \beta-\ell$ gilt. Dann ist

$$\begin{aligned} F_\ell(w[\alpha,\beta]) &= \mathrm{kopf}_\ell(w[\alpha,\beta])\mathrm{rumpf}_\ell(w[\alpha,\beta])\mathrm{rest}_\ell(w[\alpha,\beta]) \\ &= (\$, w[\alpha,\mu_j], v_j)(v_j, [\mu_j,\mu_{j+1}], v_{j+1}) \cdots (v_k, w[\mu_k,\beta], \$), \\ F_\ell(w)[\alpha,\beta] &= F_\ell(w)[\alpha,\mu_j]\mathrm{rumpf}_\ell(w[\alpha,\beta])F_\ell(w)[\mu_k,\beta] \end{aligned}$$

mit

$$\begin{aligned} \mathrm{concat}(F_\ell(w)[\alpha,\mu_j]) &= \mathrm{kopf}_\ell(w[\alpha,\beta]), \\ \mathrm{concat}(F_\ell(w)[\mu_k,\beta]) &= \mathrm{rest}_\ell(w[\alpha,\beta]). \end{aligned}$$

Da $\mu_{j-1}-\alpha \leq \ell$ und $\beta-\mu_{k+1} \leq \ell$ sind, können wir exponentielle Ausdrücke für die Sequenzen $F_\ell(w)[\alpha,\mu_{j-1}]$ und $F_\ell(w)[\mu_{k+1},\beta]$ mit Lemma 6.5.2 bestimmen. Wir hängen noch den Faktor $(v_{j-1}, w[\mu_{j-1},\mu_j], v_j)$ an die letzte Sequenz des ersten Ausdrucks und den Faktor $(v_k, w[\mu_k,\mu_{k+1}], v_{k+1})$ vor die erste Sequenz des zweiten Ausdrucks und erhalten somit $W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}$ und $W_{s+1}^{r_{s+1}} W_{s+2}^{r_{s+2}} \cdots W_t^{r_t}$.

Wenn kein i existiert, so daß $\alpha+\ell \leq \mu_i \leq \beta+\ell$ ist, muß $\mathrm{rumpf}_\ell(w[\alpha,\beta]) = \epsilon$ und $\mathrm{rest}_\ell(w[\alpha,\beta]) = \epsilon$ sein. Wenn $\beta-\alpha \leq 2\ell$ ist, bestimmen wir den exponentiellen Ausdruck $W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}$ direkt mit Lemma 6.5.2. Sonst wählen wir j und k, so daß μ_j die größte Position ist mit $\mu_j < \alpha+\ell$ und μ_k die kleinste Position ist mit $\beta-\ell < \mu_k$. Es ist $j+1 = k$. Wir bestimmen die Ausdrücke für $F_\ell(w)[\alpha,\mu_j]$ und $F_\ell(w)[\mu_k,\beta]$ mit Lemma 6.5.2. Wir erhalten $W_1^{r_1} W_2^{r_2} \cdots W_s^{r_s}$, indem wir den Faktor $(v_j, w[\mu_j,\mu_k], v_k)$ an die letzte Sequenz des ersten Ausdrucks anhängen und dann die beiden Ausdrücke konkatenieren. □

6.6. ℓ-Transformationen

Die ℓ-Transformation T_ℓ weist dem Tripel $(E : L = R, M_\Omega, \sigma)$ ein Quadrupel $(E_\ell : L_\ell = R_\ell, M_{\Omega_\ell}, \sigma_\ell, f_\ell)$ zu, das aus einer Gleichung $E_\ell : L_\ell = R_\ell$, einer Lösung σ_ℓ dieser Gleichung, den regulären Randbedingungen M_{Ω_ℓ} und einer Funktion f_ℓ besteht. In der Gleichung E_ℓ sind die Konstanten eine Teilmenge von $D_\ell \times \Sigma^+ \times D_\ell$. Die Menge der Konstanten Σ_ℓ ist die Vereinigung der in der Gleichung $E_\ell : L_\ell = R_\ell$ vorkommenden Konstanten Σ'_ℓ, ihrer Inversen und der Menge der Tripel $D_\ell \times \Sigma^{\geq 1 \wedge \leq 2\max_M + 1} \times D_\ell$, d.h. $\Sigma_\ell = \Sigma'_\ell \cup \overline{\Sigma'_\ell} \cup (D_\ell \times \Sigma^{\geq 1 \wedge \leq 2\max_M + 1} \times D_\ell)$. Die Menge der Variablen ist $\Omega_\ell := \{X \mid X \in \Omega \wedge \text{rumpf}_\ell(\sigma(X)) \neq \epsilon\}$. Die Gleichung E_ℓ wird berechnet, indem für alle Variablen mit $\text{rumpf}_\ell(\sigma(X)) \neq \epsilon$ folgende Ersetzung in $F_\ell(\sigma(L))$ bzw. $F_\ell(\sigma(R))$ gleichzeitig durchgeführt wird:

Es sei X eine Variable in $L = VXW$ mit $V, W \in \Omega^*$ und $\text{rumpf}_\ell(\sigma(X)) \neq \epsilon$. Die Startposition von $\sigma(X)$ in $\sigma(L)$ sei $\alpha := |\sigma(V)|$. Wir ersetzen in $F_\ell(\sigma(L))$ die Sequenz

$$F_\ell(\sigma(L))[\alpha + |\text{concat}(\text{kopf}_\ell(\sigma(X)))|, \alpha + |\sigma(X)| - |\text{concat}(\text{rest}_\ell(\sigma(X)))|]$$

durch X und setzen $\sigma_\ell(X) := \text{rumpf}_\ell(\sigma(X))$ und $M_{X,\ell} = \{h(\text{rumpf}_\ell(\sigma(X)))\}$. Man beachte, daß alle Faktoren von $\text{rumpf}_\ell(\sigma(X))$ in Σ_ℓ enthalten sind. Für alle Variablen X in R_ℓ verfahren wir entsprechend.

Die Funktion $f_\ell : ((\Sigma_\ell \cup \Omega_\ell)^* \to \Sigma_\ell^*) \to ((\Sigma \cup \Omega)^* \to \Sigma^*)$ weist jeder Lösung σ'_ℓ für E_ℓ eine Lösung σ' für E zu, so daß $|\text{concat}(\sigma'_\ell(L_\ell))| = |\sigma'(L)|$ ist. Die Lösung $\sigma' = f_\ell(\sigma'_\ell)$ ist definiert durch

$$\sigma'(X) := \text{concat}(\text{kopf}_\ell(\sigma(X))\sigma'_\ell(X)\text{rest}_\ell(\sigma(X))),$$

wenn $X \in \Omega_\ell$ ist, und $\sigma'(X) := \sigma(X)$ sonst.

Beispiel 6.6.1. Für unser Beispiel ist $T_2(E : L = R, M_\Omega, \sigma) = (E_2 : L_2 = R_2, M_{\Omega_2}, \sigma_2, f_2)$. Es sei $a_2 := (\overline{b}\overline{a}ab, a, \overline{a}ab\overline{b})$, $b_2 := (\overline{a}ab\overline{b}, b\overline{b}, b\overline{b}\overline{a}a)$, $c_2 := (\$, ab\overline{b}, b\overline{b}\overline{a}a)$, $d_2 := (\overline{a}ab\overline{b}, b\overline{b}, \$)$. Dann ist

$$
\begin{aligned}
L_2 &= c_2 X b_2 \overline{a_2} a_2 b_2 \overline{a_2} a_2 d_2,\\
R_2 &= c_2 \overline{X} b_2 \overline{a_2} a_2 b_2 X d_2,\\
\Sigma_2' &= \{a_2, \overline{a_2}, b_2, c_2, d_2\},\\
\Sigma_2 &= \{a_2, \overline{a_2}, b_2, c_2, d_2\} \cup \{\overline{a_2}, a_2, b_2, \overline{c_2}, \overline{d_2}\} \cup D_2 \times \Sigma^{\geq 1 \wedge \leq 3} \times D_2\\
&= \{b\overline{b}\overline{a}a, \overline{a}ab\overline{b}, \overline{b}\overline{a}ab, \$\} \times \{a, \overline{a}, b, \overline{b}\}^{\geq 1 \wedge \leq 3} \times \{b\overline{b}\overline{a}a, \overline{a}ab\overline{b}, \overline{b}\overline{a}ab, \$\}\\
I_2 &= \{(v, w, \overline{v}) \mid v \in \{b\overline{b}\overline{a}a, \overline{a}ab\overline{b}, \overline{b}\overline{a}ab, \$\} \wedge w \in \{a\overline{a}, \overline{a}a, b\overline{b}, \overline{b}b\}\}\\
\Omega_2 &= \{X, \overline{X}\},\\
M_X &= \{m_1\},\\
M_{\overline{X}} &= \{m_1\},\\
\sigma_2(X) &= \overline{a_2} a_2,\\
f_2(\sigma_2')(A) &= a,\\
f_2(\sigma_2')(X) &= ab\overline{b}\,\mathrm{concat}(\sigma_2'(X))\,b\overline{b},\\
f_2(\sigma_2')(Y) &= \overline{a}ab\overline{b}.
\end{aligned}
$$

Die Lösung σ_2 ist in diesem Fall die einzige Lösung für die FMA-Gleichung E_2 mit den regulären Randbedingungen M_{Ω_2}.

LEMMA 6.6.2. *Wenn in zwei Gleichungen E_i und E_j, die durch die ℓ-Transformationen $T_i(E : L = R, M_\Omega, \sigma)$ und $T_j(E : L = R, M_\Omega, \sigma)$ entstanden sind, gleich viele Variablen vorkommen, kommen in beiden Gleichungen dieselben Variablen in derselben Reihenfolge vor.*

BEWEIS. O.B.d.A. sei $i \geq j$. Eine Variable X kommt in der Gleichung E_i genau dann vor, wenn $\mathrm{rumpf}_i(X) \neq \epsilon$ ist. Dann ist aber auch $\mathrm{rumpf}_j(X) \neq \epsilon$, da die Wörter in C_j Teilwörter sind von den Wörtern in C_i und immer, wenn ein Wort aus C_i über einer Position α einen Faktor in $F_i(\sigma(L))$ startet, es auch ein Wort in C_j gibt, das über der Position α vorkommt und einen Faktor in $F_j(\sigma(L))$ startet. Also sind in E_j alle Variablen vorhanden, die in E_i vorkommen. Da in den Gleichungen E_i und E_j gleich viele Variablen vorkommen, kommen also in beiden Gleichungen dieselben Variablen in derselben Reihenfolge vor. □

6.7. Isomorphe Gleichungen

Von besonderer Bedeutung werden im weiteren die isomorphen Gleichungen sein, da wir die Lösung einer Gleichung so abändern können, daß sie eine isomorphe Gleichung löst.

DEFINITION 6.7.1. Es seien E_i und E_j zwei Gleichungen über den Konstanten $\Sigma_i \subset D_i \times \Sigma^+ \times D_i$ bzw. $\Sigma_j \subset D_j \times \Sigma^+ \times D_j$, mit denselben Variablen und den regulären Randbedingungen M_{Ω_i} bzw. M_{Ω_j} über demselben Monoid M. Die in der Gleichung E_i bzw. E_j vorkommenden Konstanten seien Σ_i' bzw. Σ_j'. Wir nennen die Gleichungen E_i und E_j isomorph, wenn

1. es eine bijektive Abbildung $\psi : D_j \to D_i$ mit $\overline{\psi(v)} = \psi(\overline{v})$ und eine bijektive Abbildung $\phi : \Sigma^+ \to \Sigma^+$ mit $\overline{\phi(w)} = \phi(\overline{w})$ gibt, so daß, wenn die Konstanten Σ_j' in der Gleichung E_j durch die bijektive Abbildung $\Phi : \Sigma_j' \to \Sigma_i'$ mit $\Phi((v_1, w, v_2)) := (\psi(v_1), \phi(w), \psi(v_2))$ umbenannt werden, die Gleichungen $E_i : L_i = R_i$ und $E_j : L_j = R_j$ identisch sind,
2. die regulären Randbedingungen M_{Ω_i} und M_{Ω_j} der Variablen in den Gleichungen gleich sind und
3. sich die Monoidelemente der Konstanten und ihrer Inversen durch die Umbenennung nicht verändern, d.h. $h(w) = h(\phi(w))$ für alle $(v_1, w, v_2) \in \Sigma_j' \cup \overline{\Sigma_j'}$.

LEMMA 6.7.2. *Es seien E_i und E_j zwei isomorphe Gleichungen und σ_j eine Lösung für E_j. Dann existiert eine Lösung $\sigma_i' : (\Sigma_i \cup \Omega_i)^* \to \Sigma_i^*$ für E_i, so daß $|\sigma_i'(L_i)| = |\sigma_j(L_j)|$ ist.*

BEWEIS. Da die Gleichungen E_i und E_j isomorph sind, gibt es die bijektiven Abbildungen $\psi : D_j \to D_i$ und $\phi : \Sigma^+ \to \Sigma^+$ aus Definition 6.7.1. Es sei $\Psi : \Sigma_j^* \to \Sigma_i^*$ ein Homomorphismus mit

$$\Psi((v_1, w, v_2)) := \begin{cases} (\psi(v_1), \phi(w), \psi(v_2)) & \text{wenn } (v_1, w, v_2) \in \Sigma_j' \cup \overline{\Sigma_j'} \\ (\psi(v_1), w, \psi(v_2)) & \text{sonst} \end{cases}$$

für alle Konstanten $(v_1, w, v_2) \in \Sigma_j = \Sigma'_j \cup \overline{\Sigma'_j} \cup (D_j \times \Sigma^{\leq 2\max_M+1} \times D_j)$. Es ist

$$\begin{aligned}\Psi(\overline{(v_1, w, v_2)}) &= \Psi((\overline{v_2}, \overline{w}, \overline{v_1})) = (\psi(\overline{v_2}), \phi(\overline{w}), \psi(\overline{v_1})) = (\overline{\psi(v_2)}, \overline{\phi(w)}, \overline{\psi(v_1)}) \\ &= \overline{(\psi(v_2), \phi(w), \psi(v_1))} = \overline{\Psi((v_1, w, v_2))}\end{aligned}$$

für $(v_1, w, v_2) \in \Sigma'_j \cup \overline{\Sigma'_j}$ und auch

$$\begin{aligned}\Psi(\overline{(v_1, w, v_2)}) &= \Psi((\overline{v_2}, \overline{w}, \overline{v_1})) = (\psi(\overline{v_2}), \overline{w}, \psi(\overline{v_1})) = (\overline{\psi(v_2)}, \overline{w}, \overline{\psi(v_1)}) \\ &= \overline{(\psi(v_2), w, \psi(v_1))} = \overline{\Psi((v_1, w, v_2))}\end{aligned}$$

sonst. Also ist $\Psi(\overline{w}) = \overline{\Psi(w)}$ für $w \in \Sigma_j^*$.

Wir setzen $\sigma'_i(X) := \Psi(\sigma_j(X))$ für alle Variablen $X \in \Omega_i = \Omega_j$. Es ist $|\sigma'_i(L_i)| = |\sigma_j(L_j)|$ und σ'_i eine Lösung für E_i, da

$$\begin{aligned}\sigma'_i(\overline{X}) &= \Psi(\sigma_j(\overline{X})) = \Psi(\overline{\sigma_j(X)}) = \overline{\Psi(\sigma_j(X))} = \overline{\sigma'_i(X)}, \\ \sigma'_i(L_i) &= \Psi(\sigma_j(L_j)) = \Psi(\sigma_j(R_j)) = \sigma'_i(R_i), \\ h(\sigma'_i(X)) &= h(\Psi(\sigma_j(X))) = h(\sigma_j(X)) \in M_{X,j} = M_{X,i}\end{aligned}$$

ist. □

Als nächstes werden wir abschätzen, wieviele verschiedene Äquivalenzklassen von isomorphen Gleichungen es gibt.

LEMMA 6.7.3. *Es existieren maximal* $(32d^3 + |M| + exp(\sigma(L)))^{128d^3}$ *Gleichungen, die paarweise nicht isomorph sind und durch* ℓ*-Transformationen einer Gleichung entstanden sind.*

BEWEIS. Wir komprimieren die Gleichung E_ℓ mit exponentiellen Ausdrücken entsprechend Lemma 6.5.4. In der komprimierten Gleichung kommen maximal d Variablen und $(8d-14)(2d-2)d$ Konstanten vor, da für jede der Variablen maximal $8d-14$ Sequenzen mit jeweils maximal $2d-2$ Konstanten entstehen. Der größte mögliche Exponent ist $\exp(\sigma(L))$. Wir benutzen für jede Variable, jedes charakteristische Wort, jedes von einer Konstanten repräsentierte Wort, jeden Exponenten und jedes Monoidelement jeweils ein eigenes Symbol. Hinzu kommen noch die Symbole (,) und

=. Wir brauchen für die Gleichung mit regulären Randbedingungen also

$$d + 2d - 3 + 2(8d - 14)(2d - 2)d + \exp(\sigma(L)) + |M| + 3$$
$$\leq 32d^3 + |M| + \exp(\sigma(L))$$

verschiedene Symbole. Es ist zu beachten, daß ein Symbol für ein Wort in zwei unterschiedlichen Gleichungen für ein unterschiedliches Wort stehen kann. Ein Symbol für eine Variablen, einen Exponenten oder ein Monoidelement beschreibt dagegen immer dieselbe Variable, dieselbe Zahl oder dasselbe Monoidelement.

Wir benötigen maximal für jede Variable vier Symbole (Variable, Monoidelement, inverse Variable und Monoidelement der inversen Variable), für jede Konstante acht Symbole (vier Symbole für das Tripel und das Monoidelement und vier Symbole für das inverse Tripel und sein Monoidelement), für die Klammern und Exponenten aller Sequenzen $3(8d - 14)d$ Symbole und ein Symbol für das = Zeichen. Die Komprimierung der Gleichung besteht also aus maximal

$$4d + 8(8d - 14)(2d - 2)d + 3(8d - 14)d + 1 \leq 128d^3$$

Symbolen.

Wenn zwei Gleichungen durch dieselbe Symbolfolge dargestellt werden, sind sie isomorph. Es kann also maximal soviele paarweise nicht isomorphe Gleichungen geben wie es unterschiedliche Zeichenfolgen geben kann. Davon gibt es aber maximal $(32d^3 + |M| + \exp(\sigma(L)))^{128d^3}$. □

6.8. Die Länge einer minimalen Lösung

Wenn wir zwei isomorphe Gleichungen E_i und E_j haben, können wir die Lösung von E_j benutzen, um E_i zu lösen, wobei sich die Anzahl der Faktoren in der Lösung nicht ändert. In den nächsten Lemmata werden wir zeigen, daß, wenn j viel größer als i ist, es eine Lösung $f_i(\sigma'_i(L_i))$ für E gibt, die kürzer ist als die minimale Lösung σ. Da dies ein Widerspruch ist, müssen die isomorphen Gleichungen nahe beieinander liegen und damit gibt es nur eine beschränkte Anzahl von zueinander isomorphen Gleichungen. Da auch die Anzahl der nicht isomorphen Gleichungen beschränkt ist,

können wir die durch ℓ-Transformationen entstehenden Gleichungen und damit auch ℓ und die Länge beschränken.

LEMMA 6.8.1. *Das von einem Faktor von $F_\ell(\sigma(L))$ repräsentierte Teilwort ist von der Form $v_1v_2^kv_3$ mit $|v_1|, |v_2| \leq 4d\ell$, $|v_3| \leq 8d\ell + 2max_M$ und $k \leq exp(\sigma(L))$. Es hat also höchstens eine Länge von $4d(3 + exp(\sigma(L)))\ell + 2max_M$.*

BEWEIS. Es sei $F_\ell[\alpha, \beta]$ ein beliebiger Faktor. Die Aussage ist trivial, wenn $\beta - \alpha \leq 16d\ell + 2\max_M$ ist. Es sei $\beta - \alpha > 16d\ell + 2\max_M$. Wir definieren zunächst ein Folge von Wörtern $w_0, w_1, \ldots w_{4d-3}$ mit $w_i = \sigma(L)[\alpha_i, \beta_i]$ für $0 \leq i \leq 4d-3$ und $|w_i| - \ell < |w_{i+1}| < |w_i|$ für $0 \leq i < 4d-3$.

Wir setzen $\alpha_0 = \alpha$ und $\beta_0 = \beta$. Es wird also

$$|w_i| \geq \beta - \alpha - i\ell \geq \beta - \alpha - (4d-3)\ell > 12d\ell + 2\max_M$$

sein. Nach Lemma 6.3.6 gibt es ein Intervall $[\mu_i', \nu_i'] \approx [\alpha_i, \beta_i]$ und einen Cut $\gamma_i \in [\mu_i', \nu_i']$. Es sei $j := |\gamma_i - \mu_i'|$. Wenn $\mu_i' + \ell \leq \gamma_i \leq \nu_i' - \ell$ bzw. $\nu_i' + \ell \leq \gamma_i \leq \mu_i' - \ell$ ist, kommt über der Position $\alpha_i + j$ das Wort $\sigma(L)[\gamma_i - \ell, \gamma_i + \ell] \in C_\ell$ bzw. $\overline{\sigma(L)[\gamma_i - \ell, \gamma_i + \ell]} \in C_\ell$ vor und es startet damit ein neuer Faktor an dieser Position. Da dies ein Widerspruch ist, muß $0 < j < \ell$ oder $|w_i| - \ell < j < |w_i|$ sein. Wenn $j < \ell$ ist, setzen wir $\alpha_{i+1} := \alpha_i + j$, $\beta_{i+1} := \beta_i$, $\mu_{i+1} := \gamma_i$ und $\nu_{i+1} := \nu_i'$. Sonst setzen wir $\alpha_{i+1} := \alpha_i$, $\beta_{i+1} := \alpha_i + j$, $\mu_{i+1} := \mu_i'$ und $\nu_{i+1} := \gamma_i$. Dann ist $[\alpha_{i+1}, \beta_{i+1}] \approx [\mu_{i+1}, \nu_{i+1}]$ und μ_{i+1} ist ein Cut, wenn $\alpha_{i+1} \neq \alpha_i$ ist, und ν_{i+1} ist ein Cut, wenn $\beta_{i+1} \neq \beta_i$ ist.

Unter den Wörter $w_1, w_2, \ldots, w_{4d-3}$ gibt es entweder $2d-1$ Wörter w_i mit $\alpha_{i-1} \neq \alpha_i$ oder es gibt $2d-1$ Wörter w_i mit $\beta_{i-1} \neq \beta_i$. Nehmen wir an, daß es $2d-1$ Wörter gibt mit $\alpha_{i-1} \neq \alpha_i$ (der Beweis für den anderen Fall verläuft analog). Da es nur d Cuts gibt und das Intervall $[\mu_i, \nu_i]$ nur links ($\nu_i < \mu_i$) oder rechts ($\mu_i < \nu_i$) von dem Cut μ_i liegen kann, aber nicht links vom Cut 0 oder rechts vom Cut $|\sigma(L)|$, gibt es zwei Intervalle $[\mu_p, \nu_p]$ und $[\mu_q, \nu_q]$ mit $1 \leq p < q \leq 4d-3$, $\alpha_{p-1} \neq \alpha_p$ und $\alpha_{q-1} \neq \alpha_q$, die auf derselben Seite desselben Cuts liegen. Dann ist

$$\sigma(L)[\alpha_p, \alpha_p + |w_q|] = \sigma(L)[\mu_q, \nu_q] = \sigma(L)[\alpha_q, \beta_q].$$

Das Wort $w_q = \sigma(L)[\alpha_q, \beta_q]$ hat also die Periode $\alpha_q - \alpha_p \leq \alpha_{4d-3} - \alpha_1 \leq (4d-4)\ell$. Aufgrund von Lemma 3.2.1 ist $w_q = (rs)^k r$ mit $|rs| \leq (4d-4)\ell$ und $k \leq \exp(\sigma(L))$. Wir setzen $v_1 := \sigma(L)[\alpha, \alpha_q]$, $v_2 := rs$ und $v_3 := \sigma(L)[\beta_q - |r|, \beta]$. □

LEMMA 6.8.2. *Es ist $|\sigma(L)| > \ell |F_{b\ell}(\sigma(L))|$, wenn $b \geq d(exp(\sigma(L))+1)$ und $|\sigma(L)| > \ell$ ist.*

BEWEIS. Es ist $|\sigma(L)| > \ell |F_{b\ell}(\sigma(L))|$, wenn $|F_{b\ell}(\sigma(L))| = 1$ ist. Sonst seien $b\ell \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq |\sigma(L)| - b\ell$ alle Positionen über denen ein Wort aus $C_{b\ell}$ in $\sigma(L)$ vorkommt. Es fängt also genau bei 0 und jedem μ_i ein Faktor an und es ist $k+1 = |F_{b\ell}(\sigma(L))|$.

Es sei $1 \leq i \leq k-2d$. Dann gibt es unter den Positionen $\mu_i < \mu_{i+1} < \ldots < \mu_{i+2d}$ zwei Positionen $\alpha < \beta$, so daß dasselbe Wort $v \in C_{b\ell}$ über α und β vorkommt, da $|C_{b\ell}| \leq 2(d-2)$ ist. Wir unterscheiden zwei Fälle, um zu zeigen, daß $\beta - \alpha \geq 2d\ell$ ist.

1. Wenn $\beta - \alpha < 2b\ell = |v|$ ist, überlappt sich v mit sich selbst und $\beta - \alpha$ ist eine Periode von v. Es existieren also $r, s \in \Sigma^*$, so daß $v = (rs)^p r$ und $|rs| = \beta - \alpha$ ist. Es ist $p \leq \exp(\sigma(L))$ und damit $(\beta - \alpha)(\exp(\sigma(L)) + 1) \geq |v| = 2b\ell \geq 2d\ell(\exp(\sigma(L)) + 1)$. Also ist $\beta - \alpha \geq 2d\ell$.
2. Wenn $\beta - \alpha \geq 2b\ell$ ist, gilt $\beta - \alpha \geq 2d\ell(\exp(\sigma(L)) + 1) \geq 2d\ell$.

Es ist auch $\mu_{i+2d} - \mu_i \geq 2d\ell$, da $\mu_i \leq \alpha$ und $\beta \leq \mu_{i+2d}$ ist. Wir wählen j, so daß $k - 2d < 1 + 2dj \leq k$ ist. Für $i = 1, 1+2d, \ldots, 1+2d(j-1)$ ist jeweils $\mu_{i+2d} - \mu_i \geq 2d\ell$. Also ist $\mu_{1+2dj} - \mu_1 \geq 2dj\ell$ und

$$\begin{aligned} |\sigma(L)| &\geq \mu_1 + (\mu_{1+2dj} - \mu_1) + (|\sigma(L)| - \mu_k) \\ &\geq b\ell + 2dj\ell + b\ell \\ &> 2dj\ell + (2d+1)\ell \\ &\geq \ell(k+1) \\ &= \ell |F_{b\ell}(\sigma(L))|. \end{aligned}$$

□

Jetzt können wir zeigen, daß die isomorphen Gleichungen nahe beieinander liegen müssen.

LEMMA 6.8.3. *Es seien E_i und E_j isomorphe Gleichungen, die durch die ℓ-Transformationen $T_i(E : L = R, M_\Omega, \sigma)$ und $T_j(E : L = R, M_\Omega, \sigma)$ mit $1 \leq i < j < |\sigma(L)|$ entstanden sind. Dann ist $j < ib^2$ mit $b := 4d(3 + exp(\sigma(L))) + 2max_M$.*

BEWEIS. Nehmen wir an, daß $j \geq ib^2$ ist. Aufgrund von Lemma 6.7.2 existiert eine Lösung $\sigma'_i : (\Sigma_i \cup \Omega_i)^* \to \Sigma_i^*$ für E_i, so daß $|\sigma'_i(L_i)| = |\sigma_j(L_j)|$ ist und $f_i(\sigma'_i)$ eine Lösung von E ist mit $|\text{concat}(\sigma'_i(L_i))| = |f_i(\sigma'_i)(L)|$. Wir werden zeigen, daß $|\sigma(L)| > |f_i(\sigma'_i)(L)|$ ist, was ein Widerspruch zur Minimalität von σ ist. Mit Lemma 6.8.2 ist

$$\begin{aligned}
|\sigma(L)| &> \left\lfloor \frac{j}{b} \right\rfloor |F_{b\lfloor \frac{j}{b} \rfloor}(\sigma(L))| \\
&\geq ib|F_{b\lfloor \frac{j}{b} \rfloor}(\sigma(L))| \\
&\geq ib|F_j(\sigma(L))| \\
&= ib|\sigma_j(L_j)| \\
&= ib|\sigma'_i(L_i)| \\
&\geq (4d(3 + \exp(\sigma(L)))i + 2\text{max}_M)|\sigma'_i(L_i)| \\
&\geq |\text{concat}(\sigma'_i(L_i))| \\
&= |f_i(\sigma'_i)(L)|.
\end{aligned}$$

□

Damit können wir die Länge der minimalen Lösung abschätzen.

THEOREM 6.8.4. *Es sei $b := 4d(3 + exp(\sigma(L))) + 2max_M$ und $p := (32d^3 + |M| + exp(\sigma(L)))^{128d^3}$. Dann ist $|\sigma(L)| \leq b^{2p} \in 2^{2^{\mathcal{O}(n^5)}}$, d.h. die Länge einer minimalen Lösung einer FMA-Gleichung mit regulären Randbedingungen ist doppelt exponentiell beschränkt.*

BEWEIS. Nach Lemma 6.7.3 gibt maximal p nicht isomorphe Gleichungen und nach Lemma 6.8.3 ist $j < ib^2$ für zwei isomorphe durch ℓ-Transformationen entstandene Gleichungen E_i und E_j, wenn $1 \leq i < j < |\sigma(L)|$

ist. Für $j \geq |\sigma(L)| > |\sigma(L)|/2$ sind alle Gleichungen E_j isomorph zueinander, da $C_j = \emptyset$ und $F_j(\sigma(L)) = (\$, \sigma(L), \$)$ ist. Nach Lemma 3.7.1 ist $\max_M < |M|^2$, nach Lemma 3.6.2 ist $c(M) \leq n!$ und $|M| \leq 2^{n^2}$ und nach Theorem 5.0.9 ist $\exp(\sigma(L)) \leq gc(M)2^{1,6d}$. Also muß

$$\begin{aligned} |\sigma(L)| &\leq (b^2)^p = (4d(3+\exp(\sigma(L))) + 2\max_M)^{2(32d^3+|M|+\exp(\sigma(L)))^{128d^3}} \\ &\leq (4n(3+gn!2^{1,6n}) + 2\,2^{2n^2})^{2(32n^3+2^{n^2}+gn!2^{1,6n})^{128n^3}} \in 2^{2^{\mathcal{O}(n^5)}} \end{aligned}$$

sein. □

THEOREM 6.8.5. *Es sei E eine FG-Gleichung mit regulären Randbedingungen und σ eine minimale Lösung für E. Dann ist $|\sigma(L)| \in 2^{2^{\mathcal{O}(n^{10})}}$, d.h. die Länge einer minimalen Lösung einer FG-Gleichung mit regulären Randbedingungen ist doppelt exponentiell beschränkt.*

BEWEIS. Nach Theorem 4.5.2 gibt es eine FMA-Gleichung $E_i : L_i = R_i$ mit regulären Randbedingungen und eine minimale Lösung σ_i für E_i, so daß es eine Lösung σ' für E gibt mit $|\sigma'(L)| \leq 2\,2^{n^2}(|\sigma_i(L_i)| + n)$. Da die Gleichung E_i höchstens quadratisch größer ist und aufgrund von Theorem 6.8.4, ist $|\sigma_i(L_i)| \in 2^{2^{\mathcal{O}(n^{10})}}$. Es ist $|\sigma(L)| \leq |\sigma'(L)| \leq 2\,2^{n^2}(|\sigma_i(L_i)| + n) \in 2^{2^{\mathcal{O}(n^{10})}}$.

Aus Theorem 6.8.5 folgt direkt die Entscheidbarkeit in 2-NEXPTIME. Dies ist bereits besser als der Algorithmus von Makanin, dessen Laufzeit nicht primitiv rekursiv ist [**KP98**]. Aufgrund von Korollar 4.3.4 ist also die existentielle Theorie der Gleichungen über einer freien Gruppe in 2-NEXPTIME entscheidbar. Im nächsten Kapitel werden wir zeigen, daß es sogar in PSPACE möglich ist. □

KAPITEL 7

Der PSPACE Algorithmus

In [**P99B**] hat Plandowski einen PSPACE Algorithmus vorgestellt, um zu entscheiden, ob eine Gleichung ohne reguläre Randbedingungen über einem freien Monoid eine Lösung hat. Er zitiert auch eine Mitteilung von Rytter, daß mit den vorgestellten Ideen in PSPACE entschieden werden kann, ob eine Gleichung mit regulären Randbedingungen über einem freien Monoid eine Lösung hat. Dieser Algorithmus ist jedoch nicht offensichtlich. Gutiérrez hat Plandowskis Algorithmus in [**G00B**] erweitert, um in PSPACE zu entscheiden, ob eine Gleichung über einer freien Gruppe eine Lösung hat.

Wir stellen jetzt einen PSPACE Algorithmus vor, der entscheidet, ob eine FMA-Gleichung mit regulären Randbedingungen lösbar ist. Es handelt sich eigentlich um einen NPSPACE Algorithmus. Nach dem Satz von Savitch ist jedoch NPSPACE=PSPACE.

Im weiteren seien eine FMA-Gleichung $E : L = R$ mit $L, R \in \Omega^*$ und die regulären Randbedingungen A_X für die Variablen $X \in \Omega$ fest gegeben.

7.1. Die Gleichung E_M

Wir können das Monoid nicht vollständig vorweg berechnen, da es exponentiell groß sein kann. Wir benutzen daher grundsätzlich die quadratischen booleschen Matrizen aus Lemma 3.6.2, um die Monoidelemente darzustellen. Da es Matrizen und damit auch Monoidelemente m geben kann, für die kein Wort $w \in \Sigma^*$ existiert, so daß $h(w) = m$ ist, beschränken wir das Monoid auf die Elemente zu denen es auch ein Wort aus Σ^* gibt, d.h. $M = h(\Sigma^*)$. Dies ist möglich, da wir in NPSPACE testen können, ob es ein Wort zu einem Monoidelement gibt. Wir können zwei Monoidelemente verknüpfen, indem wir die Matrizen multiplizieren. Wir können auch

bestimmen, ob ein Monoidelement m in M_X enthalten ist, d.h. ob das Monoidelement m die reguläre Randbedingung A_X erfüllt, indem wir in den Matrizen die Einträge $B_{X,i,j}$ mit $i \in I_X$ und $j \in F_X$ betrachten. Genau dann, wenn einer oder mehrere dieser Einträge 1 ist, akzeptiert der Automat A_X die Wörter $h^{-1}(m)$ und es ist $m \in M_X$.

Um zu entscheiden, ob die Gleichung $E : L = R$ eine Lösung hat, konstruieren wir eine Gleichung $E_M : L = R$. Der einzige Unterschied zwischen den Gleichungen E und E_M ist das Alphabet der Konstanten

$$\Sigma_M := \{(h(w), h(\overline{w}), s) \mid w \in \Sigma^* \wedge (w = \overline{w} \rightarrow s = 0) \wedge (w \neq \overline{w} \rightarrow s = \pm 1)\}$$

mit $\overline{(m_1, m_2, s)} := (m_2, m_1, -s)$. Es ist $I_M := \{(m, m, 0) \in \Sigma_M\}$. Der Homomorphismus $h_M : \Sigma_M^* \rightarrow M$ bildet die Konstanten Σ_M auf das Monoidelement des entsprechenden Wortes ab, d.h. $h_M((m_1, m_2, s)) := m_1$. Sowohl die Gleichung als auch die Variablen und ihre regulären Randbedingungen M_X werden nicht verändert. Man beachte, daß bei der Gleichung E_M jedes Paar von Monoidelementen durch eine einzelne Konstante dargestellt werden kann und damit die Länge des kürzesten Wortes $\min_{m_1,m_2} \leq 1$ ist. Also ist $\max_M = 1$.

LEMMA 7.1.1. *Die Gleichung E_M ist genau dann lösbar, wenn die Gleichung E lösbar ist.*

BEWEIS. Es sei σ_M eine Lösung für die Gleichung E_M. Es sei $\min_m \in \Sigma^*$ das ein kürzestes Wort mit $h(\min_m) = m$ und $\min_m = \overline{\min_m}$. Wenn kein solches Wort existiert, ist $\min_m$ undefiniert. Wir definieren den Homomorphismus $\Psi : \Sigma_M^* \rightarrow \Sigma^*$ mit

$$\Psi((m_1, m_2, s)) := \begin{cases} \min_{m_1,m_2} & \text{wenn } s = 1, \\ \overline{\min_{m_2,m_1}} & \text{wenn } s = -1, \\ \min_{m_1} & \text{wenn } s = 0. \end{cases}$$

Es ist

$$\begin{aligned} \Psi(\overline{(m_1, m_2, s)}) &= \overline{\Psi((m_1, m_2, s))}, \\ h(\Psi((m_1, m_2, s))) &= h_M((m_1, m_2, s)) \end{aligned}$$

und damit $\Psi(\overline{w}) = \overline{\Psi(w)}$ und $h(\Psi(w)) = h_M(w)$ für $w \in \Sigma_M^*$. Also ist σ mit $\sigma(X) := \Psi(\sigma_M(X))$ eine Lösung für die Gleichung E.

Es sei σ eine Lösung für die Gleichung E. Nach Lemma 6.3.5 setzt sich die Lösung σ aus maximal $2d-2$ verschiedenen Wörtern w_i zusammen, die den Äquivalenzklassen von maximalen freien Intervallen entsprechen. Wenn wir in der Lösung σ die Wörter w_i bzw. $\overline{w_i}$ ersetzen durch die Konstanten $(h(w_i), h(\overline{w_i}), 1)$ bzw. $(h(\overline{w_i}), h(w_i), -1)$, wenn $w_i \neq \overline{w_i}$ ist, und durch $(h(w_i), h(\overline{w_i}), 0)$, wenn $w_i = \overline{w_i}$ ist, erhalten wir eine Lösung σ_M für E_M. □

Beispiel 7.1.2. Für unsere Beispielgleichung aus Kapitel 6 ergibt sich die Gleichung $E_M : XYY = A\overline{X}X$ mit den Konstanten Σ_M, den Fixpunkten I_M, der Lösung σ_M und den charakteristischen Wörtern C_ℓ mit

$$\begin{aligned}
\Sigma_M &= \{(1,1,0),(m_1,m_1,0),(m_2,m_2,0),\\
&\quad (m_a,m_1,1),(m_1,m_1,1),(m_1,m_a,1),(m_2,m_2,1),\\
&\quad (m_a,m_1,-1),(m_1,m_1,-1),(m_1,m_a,-1),(m_2,m_2,-1)\},\\
I_M &= \{(1,1,0),(m_1,m_1,0),(m_2,m_2,0)\},\\
\sigma_M(A) &= (m_a,m_1,1),\\
\sigma_M(X) &= (m_a,m_1,1)(m_2,m_2,0)(m_1,m_a,-1)(m_a,m_1,1)(m_2,m_2,0),\\
\sigma_M(Y) &= (m_1,m_a,-1)(m_a,m_1,1)(m_2,m_2,0),\\
C_1 &= \{(m_a,m_1,1)(m_2,m_2,0),(m_2,m_2,0)(m_1,m_a,-1),\\
&\quad (m_1,m_a,-1)(m_a,m_1,1)\},\\
C_2 &= \{(m_a,m_1,1)(m_2,m_2,0)(m_1,m_a,-1)(m_a,m_1,1),\\
&\quad (m_1,m_a,-1)(m_a,m_1,1)(m_2,m_2,0)(m_1,m_a,-1),\\
&\quad (m_2,m_2,0)(m_1,m_a,-1)(m_a,m_1,1)(m_2,m_2,0)\},\\
&\vdots
\end{aligned}$$

Wenn wir in einer minimalen Lösung für E_M die Konstanten $a_M \in \Sigma_M$ durch die Wörter $\Psi(a_M)$ ersetzen, erhalten wir zwar eine Lösung für die Gleichung E, doch ist diese normalerweise keine minimale Lösung für E. Dies liegt daran, daß der Maßstab für eine minimale Lösung der Gleichung

E_M nicht die Länge der den Konstanten entsprechenden Wörtern ist, sondern die Anzahl der Konstanten selbst.

Im weiteren betrachten wir nur noch die FMA-Gleichung E_M und nicht mehr die Gleichung E. Es ist zu beachten, daß $|\Sigma_M| \leq 3|M|^2 \leq 3\,2^{2n^2}$ ist. Mit σ bezeichnen wir im weiteren eine minimale Lösung für E_M, wenn E_M lösbar ist. Wir können o.B.d.A. voraussetzen, daß $|L| > 0$, $|R| > 0$ und $|\sigma(L)| > 0$ ist, da sonst $\sigma(X) = \epsilon$ für alle in der Gleichung vorkommenden Variablen gelten muß und wir diese Lösungen vorweg überprüfen können. Damit ist auch $d \geq 2$.

7.2. Die Relation $\rightarrow$

Die Anzahl der Symbole (s. auch Lemma 6.7.3), die wir benötigen um einen exponentiellen Ausdruck A darzustellen, sei die Länge von A. Im weiteren dürfen exponentielle Ausdrücke auch verschachtelt sein, d.h. anstelle von einer Konstanten darf auch wieder exponentieller Ausdruck stehen.

Die folgenden beiden Lemmata können nur benutzt werden, wenn E_M lösbar ist, da ℓ-Faktorisierungen der minimalen Lösung σ betrachtet werden.

LEMMA 7.2.1. *Es sei $F_{\ell+1}(\sigma(L))[\alpha,\beta] = (v,w,v')$ ein Faktor der Faktorisierung $F_{\ell+1}(\sigma(L))$. Dann entspricht dieser Faktor in der Faktorisierung $F_\ell(\sigma(L))$ einer Sequenz*

$$F_\ell(\sigma(L))[\alpha,\beta] = (v_0,w_1,v_1)(v_1,w_2,v_2)\cdots(v_m,w_{m+1},v_{m+1})$$

und diese Sequenz kann durch einen exponentiellen Ausdruck der Länge $\mathcal{O}(n^3)$ dargestellt werden.

BEWEIS. Es seien $\alpha = \mu_0 < \mu_1 < \cdots < \mu_{m+1} = \beta$ alle Positionen zwischen α und β über denen ein Wort aus D_ℓ in $\sigma(L)$ vorkommt. Es ist $\mu_0 = \alpha$ und $\mu_{m+1} = \beta$, da über den Position α und β in $\sigma(L)$ ein Wort aus $D_{\ell+1}$ vorkommt und damit auch ein Wort aus D_ℓ über den Position α und β in $\sigma(L)$ vorkommt. Es sei v_i das Wort aus D_ℓ, das über μ_i vorkommt.

Nach Lemma 6.8.1 existieren Wörter u_i, so daß $w = u_1 u_2^p u_3$ ist mit $|u_1|, |u_2| \leq 4d(\ell+1)$, $|u_3| \leq 8d(\ell+1) + 2\max_M$ und $p \leq \exp(\sigma(L))$. Wir unterscheiden zwei Fälle:

1. $\exists i : \alpha + |u_1| + \ell \leq \mu_i \leq \beta - |u_3| - \ell$:
   Wir wählen j und k, so daß μ_j die kleinste Position und μ_k die größte Position ist mit $\alpha + |u_1| + \ell \leq \mu_j, \mu_k \leq \beta - |u_3| - \ell$. Da $\mu_{j-1} - \alpha \leq 4d(\ell+1) + \ell$ und $\beta - \mu_{k+1} \leq 8d(\ell+1) + 2\max_M + \ell$ ist können wir nach Lemma 6.5.3 die Sequenzen $W_1' := F_\ell(\sigma(L))[\alpha, \mu_{j-1}]$ und $W_4' := F_\ell(\sigma(L))[\mu_{k+1}, \beta]$ durch exponentielle Ausdrücke darstellen, die aus maximal
   $$\left\lceil \frac{4d(\ell+1)+\ell}{2\ell} \right\rceil (4d-7) \;\leq\; (4d+1)(4d-7) \;\leq\; 16d^2,$$
   $$\left\lceil \frac{8d(\ell+1)+2\max_M+\ell}{2\ell} \right\rceil (4d-7) \;\leq\; (8d+2)(4d-7) \;\leq\; 32d^2$$
   Sequenzen mit jeweils maximal $2d-2$ Konstanten bestehen. Die Gleichung E_M ist nur notwendig, damit wir in der obigen Ungleichung und noch zweimal weiter unten $\max_M$ durch 1 abschätzen können. Wir erhalten die Sequenzen W_1 bzw. W_4, indem wir an W_1' den Faktor $F_\ell(\sigma(L))[\mu_{j-1}, \mu_j]$ anhängen bzw. vor W_4' den Faktor $F_\ell(\sigma(L))[\mu_k, \mu_{k+1}]$ hängen.
   Es ist
   $$F_\ell(\sigma(L))[\mu_j, \mu_j + |u_2|] = F_\ell(\sigma(L))[\mu_j + i|u_2|, \mu_j + (i+1)|u_2|]$$
   für $0 < i < \left\lfloor \frac{\mu_k - \mu_j}{|u_2|} \right\rfloor$, da mit $\delta := \mu_j - (\alpha + |u_1|)$
   $$\begin{aligned}
   \sigma(L)[\mu_j - \ell, \mu_j + |u_2| + \ell] &= u_2^p[\delta - \ell, \delta + |u_2| + \ell] \\
   &= u_2^p[\delta + i|u_2| - \ell, \delta + (i+1)|u_2| + \ell] \\
   &= \sigma(L)[\mu_j + i|u_2| - \ell, \mu_j + (i+1)|u_2| + \ell]
   \end{aligned}$$
   ist. Also ist $F_\ell(\sigma(L))[\mu_j, \mu_k] = W_2^{\left\lfloor \frac{\mu_k - \mu_j}{|u_2|} \right\rfloor} W_3$ mit
   $$\begin{aligned}
   W_2 &:= F_\ell(\sigma(L))[\mu_j, \mu_j + |u_2|], \\
   W_3 &:= F_\ell(\sigma(L))[\mu_j + \left\lfloor \frac{\mu_k - \mu_j}{|u_2|} \right\rfloor |u_2|, \mu_k].
   \end{aligned}$$

Da $(\mu_j+|u_2|)-\mu_j \leq 4d(\ell+1)$ und $\mu_k-(\mu_j+\left\lfloor\frac{\mu_k-\mu_j}{|u_2|}\right\rfloor|u_2|) < 4d(\ell+1)$ ist, können wir nach Lemma 6.5.3 die Sequenzen W_2 und W_3 durch exponentielle Ausdrücke darstellen, die aus maximal

$$\left\lceil\frac{4d(\ell+1)}{2\ell}\right\rceil(4d-7) \leq 4d(4d-7) \leq 16d^2$$

Sequenzen mit jeweils maximal $2d-2$ Konstanten bestehen.
Es ist $F_\ell(\sigma(L))[\alpha,\beta] = W_1W_2^{\left\lfloor\frac{\mu_k-\mu_j}{|u_2|}\right\rfloor}W_3W_4$ und der exponentielle Ausdruck hat die Länge $\mathcal{O}(n^3)$.

2. $\neg\exists i : \alpha+|u_1|+\ell \leq \mu_i \leq \beta-|u_3|-\ell$:
Wir wählen j und k, so daß μ_j die größte Position mit $\mu_j < \alpha+|u_1|+\ell$ und μ_k die kleinste Position ist mit $\beta-|u_3|-\ell < \mu_k$. Es ist $j+1=k$. Da $\mu_j-\alpha < 4d(\ell+1)+\ell$ und $\beta-\mu_k < 8d(\ell+1)+2\text{max}_M+\ell$ ist können wir nach Lemma 6.5.3 die Sequenzen $W_1' := F_\ell(\sigma(L))[\alpha,\mu_j]$ und $W_2 := F_\ell(\sigma(L))[\mu_k,\beta]$ durch exponentielle Ausdrücke darstellen, die aus maximal

$$\begin{aligned}\left\lceil\frac{4d(\ell+1)+\ell}{2\ell}\right\rceil(4d-7) &\leq (4d+1)(4d-7) \leq 16d^2,\\ \left\lceil\frac{8d(\ell+1)+2\text{max}_M+\ell}{2\ell}\right\rceil(4d-7) &\leq (8d+2)(4d-7) \leq 32d^2\end{aligned}$$

Sequenzen mit jeweils maximal $2d-2$ Konstanten bestehen. Wir erhalten die Sequenzen W_1, indem wir an W_1' den Faktor $F_\ell(\sigma(L))[\mu_j,\mu_k]$ anhängen. Es ist $F_\ell(\sigma(L))[\alpha,\beta] = W_1W_2$ und der exponentielle Ausdruck hat die Länge $\mathcal{O}(n^3)$.

□

Bisher wurden die Wörter v_1 und v_2 der Faktoren (v_1,w,v_2) eigentlich nicht benötigt. Der Grund, warum wir sie überhaupt eingeführt haben, ist, daß sie im nächsten Lemma zwingend gebraucht werden.

LEMMA 7.2.2. *Es seien $F_{\ell+1}(\sigma(L))[\alpha,\beta] = (v_1,w,v_2)$ und $F_{\ell+1}(\sigma(L))[\mu,\nu] = (v_1',w',v_2')$ zwei Faktoren der Faktorisierung $F_{\ell+1}(\sigma(L))$. Dann ist*

$$\begin{aligned}F_\ell(\sigma(L))[\alpha,\beta] &= F_\ell(\sigma(L))[\mu,\nu] \quad, \text{wenn } (v_1,w,v_2) = (v_1',w',v_2') \text{ ist, und}\\ F_\ell(\sigma(L))[\alpha,\beta] &= \overline{F_\ell(\sigma(L))[\mu,\nu]} \quad, \text{wenn } (v_1,w,v_2) = \overline{(v_1',w',v_2')} \text{ ist.}\end{aligned}$$

BEWEIS. Es ist $w = \sigma(L)[\alpha, \beta] = \text{concat}(F_\ell(\sigma(L))[\alpha, \beta])$ und $w' = \sigma(L)[\mu, \nu] = \text{concat}(F_\ell(\sigma(L))[\mu, \nu])$.

Wenn $(v_1, w, v_2) = (v_1', w', v_2')$ ist, gilt entweder

$$\sigma(L)[\alpha - \ell, \alpha] = v_1[1, \ell + 1] = v_1'[1, \ell + 1] = \sigma(L)[\mu - \ell, \mu]$$
$$\sigma(L)[\beta, \beta + \ell] = v_2[\ell + 1, 2\ell + 1] = v_2'[\ell + 1, 2\ell + 1] = \sigma(L)[\nu, \nu + \ell],$$

oder v_1 und v_1' bzw. v_2 und v_2' sind gleich \$. Da auch $w = w'$ ist, kommt ein Wort aus D_ℓ genau dann über einer Position $\alpha \leq \gamma \leq \beta$ vor, wenn es über der Position $\mu + (\gamma - \alpha)$ vorkommt. Also ist $F_\ell(\sigma(L))[\alpha, \beta] = F_\ell(\sigma(L))[\mu, \nu]$.

Wenn $(v_1, w, v_2) = \overline{(v_1', w', v_2')}$ ist, gilt entweder

$$\sigma(L)[\alpha - \ell, \alpha] = v_1[1, \ell + 1] = \overline{v_2'[\ell + 1, 2\ell + 1]} = \overline{\sigma(L)[\nu, \nu + \ell]},$$
$$\sigma(L)[\beta, \beta + \ell] = v_2[\ell + 1, 2\ell + 1] = \overline{v_1'[1, \ell + 1]} = \overline{\sigma(L)[\mu - \ell, \mu]}$$

oder v_1 und v_2' bzw. v_2 und v_1' sind gleich \$. Da auch $w = \overline{w}$ ist kommt ein Wort aus D_ℓ genau dann über einer Position $\alpha \leq \gamma \leq \beta$ vor, wenn sein Inverses über der Position $\nu - (\gamma - \alpha)$ vorkommt. Also ist $F_\ell(\sigma(L))[\alpha, \beta] = \overline{F_\ell(\sigma(L))[\mu, \nu]}$. □

Anschaulich gesehen sucht der Algorithmus einen Graph ab, der lösbare Gleichungen als Knoten hat. Er beginnt bei der Gleichung $E_\ell : (\$, \sigma(L), \$) = (\$, \sigma(L), \$)$, die immer lösbar ist, und sucht nach einer speziellen Gleichung E_0, die sehr ähnlich zu der Gleichung E_M ist. Wenn E_M lösbar ist, wird er E_0 finden, da der Weg $E_{|\sigma(L)|} \to E_{|\sigma(L)|-1} \to \cdots \to E_1 \to E_0$ existiert.

DEFINITION 7.2.3. Es seien $E_2 : L_2 = R_2$ und $E_1 : L_1 = R_1$ FMA-Gleichungen mit regulären Randbedingungen über demselben Monoid M. Es seien E_2' und E_1' Gleichungen, die durch die folgenden Ersetzungen erstellt werden.

- Es werden alle Konstanten $a \in \Sigma_2$ in E_2 durch Wörter $w_a \in \Sigma_1^*$ ersetzt. Es muß $h_M(a) = h_M(w_a)$, $w_{\overline{a}} = \overline{w_a}$ sein und w_a durch einen exponentiellen Ausdruck A_a der Länge $\mathcal{O}(n^3)$ dargestellt werden können. Damit ist auch $h_M(\overline{a}) = h_M(w_{\overline{a}}) = h_M(\overline{w_a})$ und $w_a = \overline{w_a}$, wenn $a = \overline{a}$ ist.

- Für alle Variablen $X \in \Omega_{\frac{1}{2}}$ werden X und $\overline{X}$ in E_1 durch $v_X X w_X$ und $\overline{w_X}\overline{X}\overline{v_X}$ oder durch w_X und $\overline{w_X}$ ersetzt mit $v_X, w_X \in \Sigma_1^*$. Es muß

$$\begin{aligned} M_{X,1} &= \{h_M(v_X) m h_M(w_X) \mid m \in M_{X,2}\}, \\ M_{\overline{X},1} &= \{h_M(\overline{w_X}) m h_M(\overline{v_X}) \mid m \in M_{\overline{X},1}\} \end{aligned}$$

  bzw. $M_{1,X} = \{h_M(w_X)\}$ und $M_{1,\overline{X}} = \{h_M(\overline{w_X})\}$ sein. Die Wörter v_X und w_X müssen durch exponentielle Ausdrücke A_X und B_X der Länge $\mathcal{O}(n^3)$ dargestellt werden können.

Es gilt $E_2 \to E_1$ genau dann, wenn Gleichungen E_2' und E_1' existieren, die identisch sind.

Die erste wichtige Eigenschaft von $E_2 \to E_1$ ist, daß, wenn E_2 lösbar ist, auch E_1 lösbar ist. Dadurch wird der PSPACE Algorithmus immer nur mit lösbaren Gleichungen arbeiten.

LEMMA 7.2.4. *Es seien E_2 und E_1 FMA-Gleichungen mit regulären Randbedingungen, σ_2 sei eine Lösung für E_2 und es gelte $E_2 \to E_1$. Dann gibt es auch eine Lösung σ_1 für E_1.*

BEWEIS. Es seien E_2' und E_1' die identischen Gleichungen aus Definition 7.2.3. Es sei $\Psi : \Sigma_2^* \to \Sigma_1^*$ ein Homomorphismus mit $\Psi(a) := w_a$. Dann ist σ' mit $\sigma'(X) := \Psi(\sigma_2(X))$ eine Lösung für E_2'. Da E_1' und E_2' identisch sind, ist σ' auch eine Lösung für E_1'. Dann ist σ_1 mit $\sigma_1(X) := v_X \sigma'(X) w_X$ bzw. $\sigma_1(X) := w_X$ für alle $X \in \Omega_{\frac{1}{2}}$ eine Lösung für E_1. □

Aufgrund des nächsten Lemma findet der PSPACE Algorithmus einen Weg von der Gleichung E_ℓ zu der Gleichung E_1.

LEMMA 7.2.5. *Es seien $E_{\ell+1}$ und E_ℓ die Gleichungen, die durch die ℓ-Transformationen $T_{\ell+1}(E_M : L = R, M_\Omega, \sigma)$ und $T_\ell(E_M : L = R, M_\Omega, \sigma)$ entstanden sind. Dann gilt $E_{\ell+1} \to E_\ell$.*

BEWEIS. Wir führen folgende Ersetzungen durch, um $E_{\ell+1}'$ und E_ℓ' zu erzeugen.

- Wir ersetzen alle Konstanten $(v_1, w, v_2) = F_{\ell+1}(\sigma(L))[\alpha, \beta]$ in $E_{\ell+1}$ durch $F_\ell(\sigma(L))[\alpha, \beta]$. Es ist

  $$h_M((v_1, w, v_2)) = h_M(\sigma(L)[\alpha, \beta]) = h_M(F_\ell(\sigma(L))[\alpha, \beta]),$$

  nach Lemma 7.2.2 $w_{\overline{(v_1,w,v_2)}} = \overline{w_{(v_1,w,v_2)}}$ und nach Lemma 7.2.1 haben die Ausdrücke für $F_\ell(\sigma(L))[\alpha, \beta]$ die Länge $\mathcal{O}(n^3)$.
- Es sei α der linke und β der rechte Rand eines beliebigen Vorkommens der Variable X in der Gleichung E_M in $\sigma(L)$ und es sei

  $$\begin{aligned} k_i &:= |\text{concat}(\text{kopf}_i(\sigma(X)))|, \\ r_i &:= |\text{concat}(\text{rest}_i(\sigma(X)))| \end{aligned}$$

  für $i = \ell, \ell + 1$. Wir unterscheiden drei Fälle für jede Variablen $X \in \Omega_{\frac{1}{2}}$.
  1. $\text{rumpf}_{\ell+1}(\sigma(X)) \neq \epsilon$:
     Dann ist auch $\text{rumpf}_\ell(\sigma(X)) \neq \epsilon$ und die Variablen X und $\overline{X}$ kommen in E_ℓ und $E_{\ell+1}$ vor. Wir ersetzen X durch $v_X X w_X$ und $\overline{X}$ durch $\overline{w_X}\,\overline{X}\,\overline{v_X}$ mit

     $$\begin{aligned} v_X &:= F_\ell(\sigma(X))[k_\ell, k_{\ell+1}], \\ w_X &:= F_\ell(\sigma(X))[|\sigma(X)| - r_{\ell+1}, |\sigma(X)| - r_\ell]. \end{aligned}$$

     Es ist

     $$\begin{aligned} F_\ell(\sigma(X))[k_\ell, k_{\ell+1}] &= F_\ell(\sigma(L))[\alpha + k_\ell, \alpha + k_{\ell+1}], \\ F_\ell(\sigma(X))[|\sigma(X)| - r_{\ell+1}, |\sigma(X)| - r_\ell] &= F_\ell(\sigma(L))[\beta - r_{\ell+1}, \beta - r_\ell]. \end{aligned}$$

     Es sei μ die Startposition des Faktors von $F_{\ell+1}(\sigma(L))$, der bei $\alpha + k_{\ell+1}$ endet, und ν die Endposition des Faktors, der bei $\beta - r_{\ell+1}$ startet. Es ist also $F_\ell(\sigma(L))[\mu, \alpha + k_{\ell+1}]$ und $F_\ell(\sigma(L))[\beta - r_{\ell+1}, \nu]$ jeweils genau ein Faktor. Da nach Lemma 7.2.1 ein Faktor von $F_{\ell+1}(\sigma(L))$ einem exponentiellen Ausdruck der Länge $\mathcal{O}(n^3)$ von Faktoren von $F_\ell(\sigma(L))$ entspricht, können $F_\ell(\sigma(L))[\mu, \alpha + k_{\ell+1}]$ und $F_\ell(\sigma(L))[\beta - r_{\ell+1}, \nu]$ durch exponentielle Ausdrücke der Länge $\mathcal{O}(n^3)$ dargestellt werden. Es ist $\mu \leq \alpha + k_\ell$ und $\nu \geq \beta - r_\ell$, da sonst $\mu > \alpha + k_\ell \geq \alpha + \ell$ oder $\nu < \beta - r_\ell \leq \beta - \ell$ ist und der entsprechende Faktor

zu rumpf $_{\ell+1}(\sigma(X))$ gehören würde. Also haben auch die Ausdrücke für v_X und w_X die Länge $\mathcal{O}(n^3)$. Es ist

$$\begin{aligned} M_{X,\ell+1} &= \{h_M(\text{rumpf}_{\ell+1}(\sigma(X)))\}, \\ M_{\overline{X},\ell+1} &= \{h_M(\text{rumpf}_{\ell+1}(\sigma(\overline{X})))\} \end{aligned}$$

und damit

$$\begin{aligned} M_{X,\ell} &= \{h_M(\text{rumpf}_\ell(\sigma(X)))\} \\ &= \{h_M(v_X)h_M(\text{rumpf}_{\ell+1}(\sigma(X)))h_M(w_X)\} \\ &= \{h_M(v_X)mh_M(w_X) \mid m \in M_{X,\ell+1}\}, \\ M_{\overline{X},\ell} &= \{h_M(\text{rumpf}_\ell(\sigma(\overline{X})))\} \\ &= \{h_M(\overline{\text{rumpf}_\ell(\sigma(X))})\} \\ &= \{h_M(\overline{w_X})h_M(\overline{\text{rumpf}_{\ell+1}(\sigma(X))})h_M(\overline{v_X})\} \\ &= \{h_M(\overline{w_X})mh_M(\overline{v_X}) \mid m \in M_{\overline{X},\ell+1}\}. \end{aligned}$$

2. rumpf$_{\ell+1}(\sigma(X)) = \epsilon \wedge$ rumpf$_\ell(\sigma(X)) \neq \epsilon$:
Die Variablen X und $\overline{X}$ kommen dann in E_ℓ aber nicht in $E_{\ell+1}$ vor. Wir ersetzen X durch $w_X :=$ rumpf$_\ell(\sigma(X))$ und $\overline{X}$ durch $\overline{w_X}$. Es ist

$$\text{rumpf}_\ell(\sigma(X)) = F_\ell(\sigma(L))[\alpha + k_\ell, \beta - r_\ell].$$

Es sei μ die größte Position mit $\mu < \alpha + \ell + 1$ und ν die kleinste Position mit $\nu > \beta - (\ell + 1)$ über der ein Wort aus $D_{\ell+1}$ in $\sigma(L)$ vorkommt. Dann ist $|F_{\ell+1}(\sigma(L))[\mu, \nu]| \leq 2$, da sonst rumpf$_{\ell+1}(\sigma(X)) \neq \epsilon$ wäre. Da nach Lemma 7.2.1 ein Faktor von $F_{\ell+1}(\sigma(L))$ einem exponentiellen Ausdruck der Länge $\mathcal{O}(n^3)$ von Faktoren von $F_\ell(\sigma(L))$ entspricht, kann $F_\ell(\sigma(L))[\mu, \nu]$ durch einen exponentiellen Ausdruck der Länge $\mathcal{O}(n^3)$ dargestellt werden. Es ist $\mu \leq \alpha + \ell \leq \alpha + k_\ell$ und $\nu \geq \beta - \ell \geq \beta - r_\ell$. Also hat auch der Ausdruck für w_X die Länge $\mathcal{O}(n^3)$. Es ist

$$\begin{aligned} M_{X,\ell} &= \{h_M(\text{rumpf}_\ell(\sigma(X)))\} = \{h_M(w_X)\}, \\ M_{\overline{X},\ell} &= \{h_M(\text{rumpf}_\ell(\sigma(\overline{X})))\} = \{h_M(\overline{\text{rumpf}_\ell(\sigma(X))})\} = \{h_M(\overline{w_X})\}. \end{aligned}$$

3. $\text{rumpf}_\ell(\sigma(X)) = \epsilon$:
   Die Variablen X und $\overline{X}$ kommen weder in $E_{\ell+1}$ noch in E_ℓ vor. Daher wird auch keine Ersetzung durchgeführt.

Die beiden Gleichungen E'_{l+1} und E'_ℓ sind identisch. Alle Nebenbedingungen aus der Definition 7.2.3 sind auch erfüllt. □

Mit dem nächsten Lemma kann der Algorithmus von der Gleichung E_1 zur Gleichung E_0 übergehen.

LEMMA 7.2.6. *Es existieren FMA-Gleichungen* $E_{0,1} : L = R, E_{0,2} : L = R, \ldots E_{0,k} : L = R$ *mit regulären Randbedingungen* $M_{\Omega_{0,1}}, M_{\Omega_{0,2}}, \ldots, M_{\Omega_{0,k}}$, *so daß*

1. $k \leq |M|^n$ und die Menge der Konstanten $\Sigma_0 \subseteq D \times \Sigma_M \times D$ mit $|D| = 1$ ist,
2. wenn $\sigma_{0,i}$ eine Lösung für die Gleichung $E_{0,i}$ ist, es auch eine Lösung σ_M für E_M gibt und
3. wenn E_M lösbar ist, es ein i mit $1 \leq i \leq k$ gibt, so daß $E_1 \to E_{0,i}$ gilt, wobei E_1 die durch die ℓ-Transformation $T_1(E_M : L = R, M_\Omega, \sigma)$ entstandene Gleichung ist.

BEWEIS. Die einzigen Unterschiede zwischen der Gleichung E_M und einer Gleichung $E_{0,i}$ sind die Menge der Konstanten und die regulären Randbedingungen. Für alle Gleichungen $E_{0,i}$ ist die Menge der Konstanten $\Sigma_0 := \{(\epsilon, a, \epsilon) \mid a \in \Sigma_M\}$. Für alle Variablen $X \in \Omega$ bestehen die Mengen $M_{X,0,i}$ der Gleichung $E_{0,i}$ genau aus einem Element. Dieses Element muß in der Menge M_X der Gleichung E_M enthalten sein, d.h. es muß gelten $|M_{X,0,i}| = 1$ und $M_{X,0,i} \subseteq M_X$. Damit ist $k \leq |M|^n$. Es sei $\Psi : \Sigma_0^* \to \Sigma_M^*$ der Isomorphismus mit $\Psi((\epsilon, a, \epsilon)) := a$.

Es sei $\sigma_{0,i}$ eine Lösung für $E_{0,i}$. Dann ist σ_M mit $\sigma_M(X) := \Psi(\sigma_{0,i}(X))$ auch eine Lösung für E_M, da die regulären Randbedingungen $M_{X,0,i}$ jeweils eine Teilmenge von M_X sind.

Es sei σ unsere minimale Lösung für E_M. Wir wählen i, so daß $M_{X,0,i} = \{h_M(\sigma(X))\}$ für alle Variablen $X \in \Omega$ ist. Dieses i existiert, da σ eine

Lösung für E_M ist und damit $h_M(\sigma(X)) \in M_X$ sein muß. Für führen folgende Ersetzungen durch, um E_1' und $E_{0,i}'$ zu erzeugen.

- Wir ersetzen alle Konstanten $(v_1, w, v_2) \in \Sigma_1$ in E_1 durch $\Psi^{-1}(w)$. Es ist $h_M((v_1, w, v_2)) = h_M(w) = h_M((\epsilon, w, \epsilon))$ und $w_{\overline{(v_1,w,v_2)}} = (\epsilon, \overline{w}, \epsilon) = \overline{w_{(v_1,w,v_2)}}$. Nach Lemma 6.8.1 existieren Wörter u_i, so daß $w = u_1 u_2^p u_3$ ist mit $|u_1|, |u_2| \leq 4d$, $|u_3| \leq 8d + 2\max_M$ und $p \leq \exp(\sigma(L))$. Die Ausdrücke haben also die Länge $\mathcal{O}(n)$.
- Wir ersetzen alle Variablen X mit $\text{rumpf}_1(\sigma(X)) \neq \epsilon$ in $E_{0,i}$ durch $v_X X w_X$ mit

$$\begin{aligned} v_X &:= \Psi^{-1}(\text{concat}(\text{kopf}_1(\sigma(X)))), \\ w_X &:= \Psi^{-1}(\text{concat}(\text{rest}_1(\sigma(X)))). \end{aligned}$$

  Es ist

$$\begin{aligned} M_{X,1} &= \{h_M(\text{rumpf}_1(\sigma(X)))\}, \\ M_{\overline{X},1} &= \{h_M(\text{rumpf}_1(\sigma(\overline{X})))\} \end{aligned}$$

  und damit

$$\begin{aligned} M_{X,0,i} &= \{h_M(\sigma(X))\} = \{h_M(v_X)h_M(\text{rumpf}_1(\sigma(X)))h_M(w_X)\} \\ &= \{h_M(v_X)mh_M(w_X) \mid m \in M_{X,1}\}, \\ M_{\overline{X},0,i} &= \{h_M(\sigma(\overline{X}))\} = \{h_M(\overline{w_X})h_M(\overline{\text{rumpf}_1(\sigma(X))})h_M(\overline{v_X})\} \\ &= \{h_M(\overline{w_X})mh_M(\overline{v_X}) \mid m \in M_{\overline{X},1}\}. \end{aligned}$$

  Alle Variablen X mit $\text{rumpf}_1(\sigma(X)) = \epsilon$ ersetzen wir durch $w_X := \Psi^{-1}(\sigma(X))$. Es ist

$$\begin{aligned} M_{X,0,i} &= \{h_M(\sigma(X))\} = \{h_M(w_X)\}, \\ M_{\overline{X},0,i} &= \{h_M(\sigma(\overline{X}))\} = \{h_M(\overline{w_X})\}. \end{aligned}$$

  Aufgrund von Lemma 6.5.4 haben die verwendeten Ausdrücke die Länge $\mathcal{O}(n^2)$.

Die beiden Gleichungen E_1' und $E_{0,i}'$ sind identisch. Alle Nebenbedingungen aus der Definition 7.2.3 sind auch erfüllt. □

7.3. Der Algorithmus

DEFINITION 7.3.1. Eine Grammatik heißt eins-eindeutig, wenn die Grammatik azyklisch ist und es für jedes Nichtterminal genau eine Produktion gibt, in der das Nichtterminal auf der linken Seite steht.

Offensichtlich erzeugt eine eins-eindeutige kontextfreie Grammatik genau ein Wort. Wir benötigen noch den folgenden Satz.

SATZ 7.3.2. *Es sei $G = (V, \Sigma, P, S)$ eine eins-eindeutige kontextfreie Grammatik in Chomsky-Normalform. Dann kann in $\mathcal{O}(|G|^5 log(|G|))$ überprüft werden, ob zwei Nichtterminale $X, Y \in V$ dasselbe Wort erzeugen.*

BEWEIS. s. Algorithmus Test und Theorem 12 in [**P94**]. □

Damit können wir jetzt den NPSPACE Algorithmus angeben.

ALGORITHMUS 7.3.3. *Ob eine FMA-Gleichung mit regulären Randbedingungen lösbar ist, entscheidet der Algorithmus GleichungLösen in NPSPACE.*

```
procedure GleichungLösen(E : L = R, A_X)
    ; σ sei eine minimale Lösung für E_M
    Rate E_0, m_1 = h_M(σ(L)), m_2 = h_M(σ̄(L)) und, ob σ(L) = σ̄(L) ist
    E := ($, σ(L), $) = ($, σ(L), $)
    while E ≠ E_0
        Rate E'
        Überprüfe E → E'
        E' := E
    wend
    return true
end GleichungLösen
```

BEWEIS. Der Algorithmus betrachtet Gleichungen über einem Alphabet $\Sigma \subseteq D \times \Sigma_M^+ \times D$ mit $|D| \leq 2d - 2$. Die Gleichungen werden wie bei Lemma 6.7.3 komprimiert und mit Symbolen gespeichert. Die komprimierte Darstellung der Gleichung darf höchstens eine Länge von $128d^3$ Symbolen haben.

Für jede verwendete Konstante a überprüft der Algorithmus, ob ein Wort $w \in \Sigma_M^+$ existiert, so daß $h_M(a) = h_M(w)$, $h_M(\overline{a}) = h_M(\overline{w})$ und, wenn $a = \overline{a}$ ist, auch $w = \overline{w}$ ist. Mit Satz 7.3.2 können wir feststellen, ob zwei komprimierte Gleichungen identisch sind.

Es gilt $E_{|\sigma(L)|} \to E_{|\sigma(L)|-1} \to \cdots \to E_1 \to E_0$ aufgrund der Lemmata 7.2.5 und 7.2.6. Da nach Lemma 7.2.4 alle Gleichungen E lösbar sind und, wenn E_0 lösbar ist, es auch eine Lösung für E_M gibt. Gibt der Algorithmus korrekt true aus, wenn die Gleichung lösbar ist.

Wenn die Gleichung E_M nicht lösbar ist, ist auch E_0 nicht lösbar und der Algorithmus bricht aufgrund der Platzbeschränkung ab, da E immer eine lösbare Gleichung ist und damit niemals $E = E_0$ ist. □

Damit können wir beweisen, daß die folgenden beiden Probleme PSPACE-vollständig sind.

THEOREM 7.3.4. *Das Problem, ob ein Ausdruck mit regulären Randbedingungen über FMA-Gleichungen eine Lösung hat, ist PSPACE-vollständig.*

BEWEIS. Betrachten wir den Ausdruck $X_1 = X_2 \wedge X_2 = X_3 \wedge \ldots \wedge X_{k-1} = X_k$ mit den regulären Randbedingungen A_{X_i} für $1 \leq i \leq k$. Es gibt genau dann eine Lösung, wenn $\bigcap_{1 \leq i \leq k} L(A_{X_i}) \neq \emptyset$ ist. Da das Problem, ob der Schnitt von mehreren regulären Sprachen leer ist, PSPACE-schwierig ist und der Algorithmus 8.3.1 in NPSPACE=PSPACE entscheidet, ob es eine Lösung gibt, ist das Problem PSPACE-vollständig. □

THEOREM 7.3.5. *Das Problem, ob ein Ausdruck mit regulären Randbedingungen über FG-Gleichungen eine Lösung hat, ist PSPACE-vollständig.*

BEWEIS. Aufgrund von Korollar 4.3.4 gibt es eine zum Ausdruck äquivalente Gleichung E, die nur polynomiell größer ist. Nach Theorem 4.5.2 gibt es eine FMA-Gleichung $E_i : L_i = R_i$ mit regulären Randbedingungen, die genau dann eine Lösung hat, wenn E lösbar ist. Der Algorithmus 8.3.1 entscheidet in NPSPACE=PSPACE, ob E_i und damit auch E und der Ausdruck lösbar sind.

Es seinen A_i mit $1 \leq i \leq k$ beliebige nichtdeterministische endliche Automaten über einem Alphabet Σ. Wir erweitern das Alphabet Σ um $\overline{\Sigma}$ mit

$\Sigma \cap \overline{\Sigma} = \emptyset$. Der Schnitt $\bigcap_{1 \leq i \leq k} L(A_i)$ ist genau dann nicht leer, wenn die Gleichung

$$X_1 \approx X_2 \wedge X_2 \approx X_3 \wedge \ldots \wedge X_{k-1} \approx X_k$$

mit $A_{X_i} = A_i$ eine Lösung σ hat, da in $\sigma(X_i)$ aufgrund der regulären Randbedingungen keine Konstante aus $\overline{\Sigma}$ vorkommen kann. Es ist das Problem, ob der Schnitt $\bigcap_{1 \leq i \leq k} L(A_i)$ leer ist, PSPACE-vollständig. □

KAPITEL 8

Komprimieren von minimalen Lösungen

Es sei $E : L = R$ eine FMA-Gleichung mit regulären Randbedingungen und σ eine minimale Lösung für E. In diesem Kapitel werden wir die Wörter $\sigma(X)$ mit Grammatiken beschreiben. Für eine Grammatik $G = (V, \Sigma, P, S)$ mit den Nichtterminalen V, den Terminalen T, den Produktionen P und dem Startsymbol S sei die Größe $|G| := |P|$ die Anzahl der Produktionen.

8.1. Intervall Grammatiken

Da wir mit einer Grammatik nur ein einziges Wort erzeugen wollen, sind für uns die eins-eindeutigen Grammatiken von besonderer Bedeutung. Für ein Nichtterminal A sei $|A|$ die Länge des von A erzeugten Wortes.

Definition 8.1.1. Es sei $G = (V, \Sigma, P, S)$ eine eins-eindeutige Grammatik und W die rechte Seite der Produktion für das Nichtterminal X, d.h. $X \to W \in P$. Dann ist die Höhe

$$H(X) := \begin{cases} \max\{H(Y) \mid Y \in V \text{ in } W\} + 1 & \text{wenn ein } Y \in V \text{ in } W \text{ vorkommt,} \\ 1 & \text{sonst.} \end{cases}$$

Definition 8.1.2. Eine Intervall Grammatik ist eine Grammatik $G = (V, \Sigma, P, S)$ bei der alle Produktionen von der Form

$$\begin{aligned} X &\to a \\ X &\to Y[\alpha, \beta]Z[\mu, \nu] \end{aligned}$$

sind. Dabei sind $X, Y, Z \in V$, $a \in \Sigma$, $0 \le \alpha \le \beta \le \min_Y$ und $0 \le \mu \le \nu \le \min_Z$, wobei $\min_X$ die Länge des kleinsten von X ableitbaren Wortes ist, d.h. $\min_X := \min\{|w| \mid w \in \Sigma^* \wedge X \xrightarrow{*} w\}$.

BEISPIEL 8.1.3. Die eins-eindeutige Intervall Grammatik $G = (V, \Sigma, P, S)$ mit

$$\begin{aligned} V &= \{A, B, C, D, S\} \\ \Sigma &= \{a, b\} \\ P &= \{A \to a, B \to b \\ &\quad C \to A[0,1]B[0,1], \\ &\quad D \to C[0,2]C[0,2], \\ &\quad S \to D[1,3], D[0,3]\} \end{aligned}$$

erzeugt genau das Wort *baaba*. Es ist $H(A) = H(B) = 1$, $H(C) = 2$, $H(D) = 3$ und $H(S) = 4$.

Die von einer Intervall Grammatik erzeugte Sprache ist immer endlich, da die Länge der Wörter beschränkt ist. Wenn eine Intervall Grammatik eins-eindeutig ist, erzeugt sie genau ein Wort. Wir werden solche Grammatiken benutzen, um die Lösung der Gleichung darzustellen. Um überprüfen zu können, ob es sich um eine Lösung handelt, transformieren wir die Intervall Grammatik in eine kontextfreie Grammatik in Chomsky-Normalform.

Der folgende Algorithmus 8.1.4 bekommt als Eingabe eine eins-eindeutige Intervall Grammatik und gibt eine eins-eindeutige Grammatik in Chomsky-Normalform aus, die dasselbe Wort erzeugt und nur quadratisch größer ist. Er konstruiert für die Nichtterminale nacheinander kontextfreie Produktionen.

ALGORITHMUS 8.1.4. *Es sei $G = (V, \Sigma, P, S)$ eine eins-eindeutige Intervall Grammatik. Der Algorithmus ChomskyErzeugen gibt eine eins-eindeutige kontextfreie Grammatik in Chomsky-Normalform der Größe $\mathcal{O}(|G|^2)$ aus, die dasselbe Wort erzeugt.*

; Erzeugt ein Nichtterminal $X_{\alpha,|X|}$, das $X[\alpha, |X|]$ entspricht
procedure *SuffixErzeugen*(X, α)
    $V' := V' \cup \{X_{\alpha,|X|}\}$
    if $X \to a \in P'$ then
        if $\alpha = 0$ then

$P' := P' \cup \{X_{0,1} \to a\}$
else
$P' := P' \cup \{X_{1,1} \to \epsilon\}$
end if
elseif $X \to YZ \in P'$ then
if $|Y| \leq \alpha$ then
SuffixErzeugen$(Z, \alpha - |Y|)$
$P' := P' \cup \{X_{\alpha,|X|} \to Z_{\alpha-|Y|,|Z|}\}$
else
SuffixErzeugen(Y, α)
$P' := P' \cup \{X_{\alpha,|X|} \to Y_{\alpha,|Y|}Z\}$
end if
end if
end *SuffixErzeugen*

; Erzeugt ein Nichtterminal $X_{0,\beta}$, das $X[0, \beta]$ entspricht
procedure *PräfixErzeugen*(X, β)
$V' := V' \cup \{X_{0,\beta}\}$
if $X \to a \in P'$ then
if $\beta = 1$ then
$P' := P' \cup \{X_{0,1} \to a\}$
else
$P' := P' \cup \{X_{0,0} \to \epsilon\}$
end if
elseif $X \to YZ \in P'$ then
if $\beta \leq |Y|$ then
PräfixErzeugen(Y, β)
$P' := P' \cup \{X_{0,\beta} \to Y_{0,\beta}\}$
else
PräfixErzeugen$(Z, \beta - |Y|)$
$P' := P' \cup \{X_{0,\beta} \to YZ_{0,\beta-|Y|}\}$
end if
end if
end *PräfixErzeugen*

; Erzeugt ein Nichtterminal $X_{\alpha,\beta}$, das $X[\alpha,\beta]$ entspricht
procedure *TeilwortErzeugen*(X, α, β)
    $V' := V' \cup \{X_{\alpha,\beta}\}$
    if $X \to a \in P'$ then
        if $\beta - \alpha = 1$ then
            $P' := P' \cup \{X_{0,1} \to a\}$
        else
            $P' := P' \cup \{X_{\alpha,\beta} \to \epsilon\}$
        end if
    elseif $X \to YZ \in P'$ then
        if $\beta \leq |Y|$ then
            TeilwortErzeugen(Y, α, β)
            $P' := P' \cup \{X_{\alpha,\beta} \to Y_{\alpha,\beta}\}$
        elseif $|Y| \leq \alpha$ then
            TeilwortErzeugen$(Z, \alpha - |Y|, \beta - |Y|)$
            $P' := P' \cup \{X_{\alpha,\beta} \to Z_{\alpha-|Y|,\beta-|Y|}\}$
        else
            SuffixErzeugen(Y, α)
            PräfixErzeugen$(Z, \beta - |Y|)$
            $P' := P' \cup \{X_{\alpha,\beta} \to Y_{\alpha,|Y|} Z_{0,\beta-|Y|}\}$
        end if
    end if
end *TeilwortErzeugen*

procedure *ChomskyErzeugen*(V, Σ, P, S)
    $V' := \emptyset$
    $P' := \emptyset$
    V *sortieren* nach $H(X)$
    for $i := 1$ to $|V|$
        $X := V[i]$
        $V' := V' \cup \{X\}$
        if $X \to a \in P$ then
            $P' := P' \cup \{X \to a\}$

elseif $X \to Y[\alpha,\beta]Z[\mu,\nu] \in P$ then
TeilwortErzeugen(Y,α,β)
TeilwortErzeugen(Z,μ,ν)
$P' := P' \cup \{X \to Y_{\alpha,\beta}Z_{\mu,\nu}\}$
end if
next i
(V',Σ,P',S) *in Chomsky-Normalform bringen*
Unerreichbare Nichtterminale aus (V',Σ,P',S) *entfernen*
return (V',Σ,P',S)
end *ChomskyErzeugen*

BEWEIS. Nachdem V nach $H(X)$ sortiert ist, gilt $Y < X$ und $Z < X$ für $X \to Y[\alpha,\beta]Z[\mu,\nu] \in P$. Wir konstruieren eine eins-eindeutige Grammatik, die dasselbe Wort erzeugt und bei der alle Produktionen von der Form $X \to a$, $X \to \epsilon$, $X \to Y$ oder $X \to YZ$ sind. Man kann leicht mit Induktion nachweisen, daß

1. die Prozedur *SuffixErzeugen*(X,α) bzw. *PräfixErzeugen*(X,β) maximal $H(X)$ Produktionen erzeugt, die Höhe des erstellten Nichtterminals $X_{\alpha,|X|}$ bzw. $X_{0,\beta}$ maximal $H(X)$ ist und das erstellte Nichtterminal $X[\alpha,|X|]$ bzw. $X[0,\beta]$ entspricht;
2. die Prozedur *TeilwortErzeugen*(X,α,β) maximal $2H(X)-1$ Produktionen erzeugt, die Höhe des erstellten Nichtterminals $X_{\alpha,\beta}$ maximal $H(X)$ ist und das erstellte Nichtterminal $X[\alpha,\beta]$ entspricht und
3. nach dem i-ten Durchlauf der for-Schleife
$$|P'| \leq 2 + \sum_{j=1}^{i}(2(2(j-1)-1)+1) = 2i^2 - 3i + 2$$
ist, die Höhe des erstellten Nichtterminals $X \in V'$ maximal i ist und das erstellte Nichtterminal dasselbe Wort wie $X \in V$ erzeugt.

Nach der for-Schleife erzeugt also $S \in V'$ genau dasselbe Wort wie $S \in V$ und es ist $|P'| \leq 2|V|^2 - 3|V| + 2 \in \mathcal{O}(|G|^2)$. Nachdem alle Produktionen der Form $X \to \epsilon$ bzw. $X \to Y$ ersetzt sind und die nicht erreichbaren Nichtterminale entfernt sind, haben wir eine äquivalente eins-eindeutige Grammatik in Chomsky-Normalform, die nicht größer ist. □

BEISPIEL 8.1.5. Für die im Beispiel 8.1.3 angegebene Grammatik $G = (V, \Sigma, P, S)$ wird die Grammatik $G' = (V', \Sigma, P', S)$ mit

$$\begin{aligned} V' &= \{A_{0,1_{0,1_{0,1}}}, B_{0,1_{0,1}}, C_{0,2_{1,2}}, C_{0,2_{0,1}}, C_{0,2_{0,2}}, D_{1,3}, D_{0,3}, S\} \\ \Sigma &= \{a, b\} \\ P' &= \{A_{0,1_{0,1_{0,1}}} \to a, B_{0,1_{0,1}} \to b, \\ &\quad C_{0,2_{1,2}} \to b, C_{0,2_{0,1}} \to a, C_{0,2_{0,2}} \to A_{0,1_{0,1_{0,1}}} B_{0,1_{0,1}}, \\ &\quad D_{1,3} \to C_{0,2_{1,2}} C_{0,2_{0,1}}, D_{0,3} \to C_{0,2_{0,2}} C_{0,2_{0,1}}, \\ &\quad S \to D_{1,3} D_{0,3}\} \end{aligned}$$

erstellt.

8.2. Eine Intervall Grammatik für die Lösung

Um die Darstellung zu vereinfachen, definieren wir

$$\lfloor \alpha, \beta \rfloor := \sigma(L)[\min\{\max\{\alpha, 0\}, |\sigma(L)|\}, \min\{\max\{\beta, 0\}, |\sigma(L)|\}].$$

Der Ausdruck $\lfloor \alpha, \beta \rfloor$ bezieht sich also immer auf das Lösungswort $\sigma(L)$ und falls α oder β kleiner 0 oder größer als $|\sigma(L)|$ sind, werden sie entsprechend angepaßt.

LEMMA 8.2.1. *Es gibt eine eins-eindeutige Intervall Grammatik der Größe*

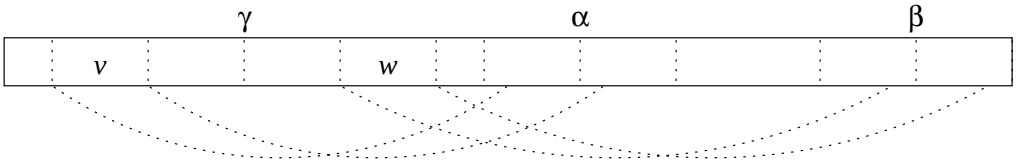
$$\mathcal{O}(d(max_M + log(|\sigma(L)|))),$$

die genau das Wort $\sigma(L)$ erzeugt.

BEWEIS. Es sei $m := 2\max_M + 2$. Das Lemma ist trivial, wenn $|\sigma(L)| < 2m$ ist. Es sei also $|\sigma(L)| \geq 2m$. Wir erstellen Nichtterminale $A_{\gamma,i,s}$, die genau die Wörter $\lfloor \gamma - sm2^i, \gamma + sm2^i \rfloor$ erzeugen. Für alle Cuts γ, $0 \leq i \leq$

$\lceil \log(|\sigma(L)|/m) \rceil$ und $s \in \{-1, 1\}$ erstellen wir die Produktionen

$$A_{\gamma,i,s} \to \begin{cases} \lfloor \gamma - sm, \gamma + sm \rfloor & \text{wenn } i = 0, \\ A_{\alpha,i-1,s_\alpha}[\mu_\alpha, \nu_\alpha] A_{\gamma,i-1,s}[0, |A_{\gamma,i-1,s}|] & \\ \quad A_{\beta,i-1,s_\beta}[\mu_\beta, \nu_\beta] & \text{wenn } i > 0 \wedge s = 1, \\ A_{\beta,i-1,-s_\beta}[|A_{\beta,i-1,s_\beta}| - \nu_\beta, |A_{\beta,i-1,s_\beta}| - \mu_\beta] & \\ \quad A_{\gamma,i-1,s}[0, |A_{\gamma,i-1,s}|] & \\ \quad A_{\alpha,i-1,-s_\alpha}[|A_{\alpha,i-1,s_\alpha}| - \nu_\alpha, |A_{\alpha,i-1,s_\alpha}| - \mu_\alpha] & \text{wenn } i > 0 \wedge s = -1. \end{cases}$$



Die Produktionen $A_{\gamma,i,-1}$ erzeugen genau die Inversen der Wörter, die von den Produktionen $A_{\gamma,i,1}$ erzeugt werden.

Es sei γ ein beliebiger Cut, $1 \leq i \leq \lceil \log(|\sigma(L)|/m) \rceil$ und $s = 1$. Wir bestimmen α, s_α, μ_α, ν_α, β, s_β, μ_β und ν_β wie folgt.

Es sei $v := \lfloor \gamma - m2^i, \max\{\gamma - m2^{i-1}, m\} \rfloor$. Da $|v| \geq m$ ist, kommt nach Lemma 6.3.6 v oder $\overline{v}$ über einem Cut α vor. Also kommt v und damit auch das Teilwort $\lfloor \gamma - m2^i, \gamma - m2^{i-1} \rfloor$ von v in dem von $A_{\alpha,i-1,1}$ oder $A_{\alpha,i-1,-1}$ erstellten Wort vor. Die Start- und Endposition von $\lfloor \gamma - m2^i, \gamma - m2^{i-1} \rfloor$ in dem von $A_{\alpha,i-1,s_\alpha}$ erstellten Wort sind μ_α und ν_α.

Es sei $w := \lfloor \min\{\gamma + m2^{i-1}, |\sigma(L)| - m\}, \gamma + m2^i \rfloor$. Da $|w| \geq m$ ist, kommt nach Lemma 6.3.6 w oder $\overline{w}$ über einem Cut β vor. Also kommt w und damit auch das Teilwort $\lfloor \gamma + m2^{i-1}, \gamma + m2^i \rfloor$ von w in dem von $A_{\beta,i-1,1}$ oder $A_{\beta,i-1,-1}$ erstellten Wort vor. Die Start- und Endposition von $\lfloor \gamma + m2^{i-1}, \gamma + m2^i \rfloor$ in dem von $A_{\beta,i-1,s_\beta}$ erstellten Wort sind μ_β und ν_β.

Für alle Cuts γ erzeugt das Nichtterminal $A_{\gamma,\lceil \log(|\sigma(L)|/m) \rceil,1}$ genau das Wort $\sigma(L)$, da

$$\lfloor \gamma - m2^{\lceil \log(|\sigma(L)|/m) \rceil}, \gamma + m2^{\lceil \log(|\sigma(L)|/m) \rceil} \rfloor = \sigma(L)[0, |\sigma(L)|] = \sigma(L)$$

ist. Die Produktionen für $A_{\gamma,i,s}$ sind keine Produktionen einer Intervall Grammatik, da Wörter bzw. drei Nichtterminale auf der rechten Seite vorkommen. Wir können sie jedoch einfach durch Intervall Grammatik Produktionen ersetzen. Die Wörter mit einer maximalen Länge von $2m$ können wir mit $\mathcal{O}(m)$ Produktionen darstellen. Die Produktionen mit drei Nichtterminalen können wir durch jeweils zwei Produktionen ersetzen. Da es maximal d Cuts gibt und $m = 2\max_M+2$ ist, hat die Intervall Grammatik also höchstens $\mathcal{O}(d(\max_M + \log(|\sigma(L)|)))$ Produktionen. □

Es sei $G = (V, \Sigma, P, S)$ eine eins-eindeutige Intervall Grammatik, die das Wort $\sigma(L)$ erzeugt. Dann gibt es natürlich auch für jede Variable $X \in \Omega$ eine eins-eindeutige Intervall Grammatik, die das Wort $\sigma(X)$ erzeugt. Man muß nur ein neues Startsymbol S_X und die Produktion $S_X \to S[\alpha, \beta]S[0, 0]$ hinzufügen, wobei α und β die Ränder eines Vorkommen der Variable X im Lösungswort $\sigma(L)$ sind.

8.3. Der Algorithmus

Damit können wir jetzt einen NEXPTIME Algorithmus zum Lösen einer FMA-Gleichung mit regulären Randbedingungen A_X angeben:

ALGORITHMUS 8.3.1. *Es sei $E : L = R$ eine FMA-Gleichung mit den regulären Randbedingungen A_X. Der Algorithmus GleichungLösen entscheidet in NEXPTIME, ob die Gleichung lösbar ist.*

procedure *GleichungLösen*$(E : L = R, A_X)$
; σ sei eine minimale Lösung
Berechne ein Monoid M und $\forall X : M_X \subseteq M$ mit $h^{-1}(M_X) = L(A_X)$
$\forall X \in \Omega_{\frac{1}{2}}$: *Rate eine* eins-*eindeutige Grammatik in Chomsky-Normalform, die $\sigma(X)$ erzeugt*
$\forall X \in \overline{\Omega_{\frac{1}{2}}}$: *Erstelle eine* eins-*eindeutige Grammatik in Chomsky-Normalf*orm*, die $\sigma(X)$ erzeugt*
if $\exists X \in \Omega : h(\sigma(X)) \notin M_X$ then return false
if $\sigma(L) \neq \sigma(R)$ then return false
return true
end *GleichungLösen*

BEWEIS. Wenn es eine Lösung gibt, existiert auch eine minimale Lösung σ. Aufgrund der Lemmata 3.7.1 und 3.6.2 ist $\max_M < |M|^2 \leq 2^{2n^2}$. Nach Lemma 8.2.1 existiert eine eins-eindeutige Intervall Grammatik der Größe $\mathcal{O}(d(\max_M + \log(|\sigma(L)|)))$, die das Wort $\sigma(L)$ erzeugt. Also gibt es auch für jede Variable $X \in \Omega_{\frac{1}{2}}$ eine eins-eindeutige Intervall Grammatik der Größe $\mathcal{O}(d(\max_M + \log(|\sigma(L)|)))$, die das Wort $\sigma(X)$ erzeugt. Diese Grammatiken können mit dem Algorithmus 8.1.4 in eins-eindeutige Grammatiken in Chomsky-Normalform der Größe $\mathcal{O}((d(\max_M + \log(|\sigma(L)|)))^2)$ umwandelt werden. Da die Länge einer minimalen Lösung nach Theorem 6.8.4 durch $2^{2^{\mathcal{O}(n^5)}}$ beschränkt ist, haben diese Grammatiken maximal die Größe $2^{\mathcal{O}(n^5)}$. Ob die regulären Randbedingungen erfüllt sind, läßt sich einfach überprüfen, indem man die Monoidelemente für alle Nichtterminale in den Grammatiken in Chomsky-Normalform berechnet. Ob $\sigma(L) = \sigma(R)$ ist, kann man testen, indem man aus den bestehenden Grammatiken eine Grammatik in Chomsky-Normalform erstellt mit zwei Nichtterminalen A_L und A_R, die genau die Wörter $\sigma(L)$ und $\sigma(R)$ erzeugen und dann den Algorithmus aus Satz 7.3.2 benutzt. Wenn eine minimale Lösung σ existiert, gibt der Algorithmus also korrekt true aus.

Wenn es keine Lösung gibt, existieren auch keine Grammatiken, so daß $\forall X \in \Omega : h(\sigma(X)) \in M_X$ und $\sigma(L) = \sigma(R)$ ist. Der Algorithmus gibt also korrekt false aus.

Alle Schritte lassen sich in NEXPTIME ausführen. □

Damit können wir jetzt die folgenden vier Theoreme beweisen.

THEOREM 8.3.2. *Es ist in NEXPTIME entscheidbar, ob ein Ausdruck mit regulären Randbedingungen über FMA-Gleichungen lösbar ist.*

BEWEIS. Aufgrund von Korollar 3.4.4 gibt es eine zum Ausdruck äquivalente Gleichung, die nur polynomiell größer ist. Der Algorithmus 8.3.1 entscheidet in NEXPTIME, ob diese Gleichung und damit auch der Ausdruck lösbar ist. □

Da die Länge der Lösung doppelt exponentiell sein kann, kann in NEXPTIME die Lösung nicht ausgegeben werden. Dies ist jedoch in 2-DEXPTIME möglich.

THEOREM 8.3.3. *In 2-DEXPTIME kann eine Lösung für einen Ausdruck mit regulären Randbedingungen über FMA-Gleichungen berechnet werden, wenn es eine Lösung gibt.*

BEWEIS. Aufgrund von Korollar 3.4.4 gibt es eine zum Ausdruck äquivalente Gleichung, die nur polynomiell größer ist. Der Algorithmus 8.3.1 entscheidet in NEXPTIME, ob diese Gleichung und damit auch der Ausdruck lösbar ist. Der Algorithmus kann in 2-DEXPTIME simuliert werden. Wenn es eine Lösung gibt, werden vom Algorithmus eins-eindeutige Grammatiken in Chomsky-Normalform für $\sigma(X)$ für alle Variablem $X \in \Omega$ berechnet. Die Wörter die von diesen Grammatiken erzeugt werden, können in 2-DEXPTIME ausgegeben werden.

Dasselbe gilt für freie Gruppen. □

THEOREM 8.3.4. *In 2-DEXPTIME kann eine Lösung für einen Ausdruck mit regulären Randbedingungen über FG-Gleichungen berechnet werden, wenn es eine Lösung gibt.*

BEWEIS. Aufgrund von Korollar 4.3.4 gibt es eine zum Ausdruck äquivalente Gleichung E, die nur polynomiell größer ist. Nach Theorem 4.5.2 gibt es eine FMA-Gleichung $E_i : L_i = R_i$ mit den regulären Randbedingungen A'_X, die genau dann eine Lösung hat, wenn E lösbar ist. Nach Theorem 8.3.3 können wir in 2-DEXPTIME eine Lösung σ_i für E_i berechnen. Damit die Lösung die regulären Randbedingungen A_X erfüllt, fügen wir die Wörter $w_{X,i,j}$ mit $|w_{X,i,j}| \leq 2\,2^{|Q_X|^2}$, $w_{X,i,j} \approx \epsilon$ und $q_{X,i} \overset{w_{X,i,j}}{\to} q_{X,j}$ ein. Wir erhalten eine Lösung für die Gleichung E und damit auch für den Ausdruck. □

Es ist bisher kein Beispiel bekannt für eine FMA-Gleichung mit regulären Randbedingungen, bei der die minimale Lösung größer als exponentiell ist. Die Länge könnte also exponentiell beschränkt sein.

THEOREM 8.3.5. *Wenn die Länge einer minimalen Lösung exponentiell beschränkt ist und das Monoid nur polynomiell groß ist, kann man ist in NP entscheiden, ob ein Ausdruck mit regulären Randbedingungen über FMA-Gleichungen lösbar ist.*

BEWEIS. Aufgrund von Korollar 3.4.4 gibt es eine zum Ausdruck äquivalente Gleichung, die nur polynomiell größer ist. Der Algorithmus 8.3.1 entscheidet in NP, ob diese FMA-Gleichung und damit auch der Ausdruck lösbar ist. □

Um zu zeigen, daß das Problem NP-vollständig ist, muß man also nur die Länge einer minimalen Lösung exponentiell beschränken. Es ist NP-schwierig, da bereits das Problem, ob eine Gleichung ohne reguläre Randbedingungen über einem freien Monoid, die keine Variablen auf der rechten Seite hat, lösbar ist, NP-schwierig ist [**A79**]. Die Vorteile von diesem Algorithmus gegenüber dem in [**PR98**] sind, daß zusätzlich eine Anti-Involution mit Fixpunkten und reguläre Randbedingungen, die zu einem Monoid mit maximal polynomieller Größe führen, beachtet werden.

KAPITEL 9

Lösungen mit vorgegebener Länge N

In diesem Kapitel werden wir uns damit beschäftigen, in welcher Zeit es sich für eine FMA-Gleichung $E : L = R$ mit regulären Randbedingungen entscheiden läßt, ob es eine Lösung der Länge N gibt, wobei N Teil der Eingabe ist.

9.1. Eine alternative Lösung

Nehmen wir an, daß es eine Lösung der Länge N gibt. Da diese Lösung nicht minimal ist, kann es sein, daß diese Lösung nicht gut komprimierbar ist. Daher zeigen wir zuerst, daß, wenn es eine Lösung σ mit $|\sigma(L)| = N$ gibt, es auch eine Lösung σ' mit $|\sigma'(L)| = N$ gibt, so daß σ' gut komprimierbar ist. Die Idee ist, daß wir alle maximalen freien Intervalle durch gut komprimierbare Wörter ersetzen. Dazu werden wir den folgenden Algorithmus benutzen.

ALGORITHMUS 9.1.1. *Es sei w ein Wort, M ein Monoid und $h : \Sigma^* \to M$ ein Homomorphismus. Der Algorithmus WortKomprimieren gibt ein Word v aus, so daß $|v| = |w|$, $h(v) = h(w)$, $h(\overline{v}) = h(\overline{w})$ und $v = x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}$ ist mit $0 \leq n \leq |M|^4$, $|x_i| < |M|^2$ für $1 \leq i \leq n+1$, $1 \leq |y_i| \leq |M|^2$ und $1 \leq r_i \leq |w|$ für $1 \leq i \leq n$.*

procedure *WortKomprimieren*(w, M, h)
    $n := 0$
    $u := w$
    while $|u| \geq |M|^2$
        Bestimme $0 \leq \alpha < \beta \leq |M|^2 : h(u[0,\alpha]) = h(u[0,\beta])$
                $\wedge\, h(\overline{u[0,\alpha]}) = h(\overline{u[0,\beta]})$
        if $\neg\exists 1 \leq k \leq n : |y_k| = \beta - \alpha \wedge h(y_k) = h(u[\alpha,\beta])$

$\wedge\, h(\overline{y_k}) = h(\overline{u[\alpha,\beta]})$ then
$x_{n+1} := u[0,\alpha]$
$y_{n+1} := u[\alpha,\beta]$
$r_{n+1} := 1$
$u := u[\beta,|u|]$
$n := n+1$
else
$r_k := r_k + 1$
$u := u[0,\alpha]u[\beta,|u|]$
end if
wend
$x_{n+1} := u$
return $x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}$
end *WortKomprimieren*

BEWEIS. Es gelten die folgenden Invarianten für die while-Schleife:

1. $|x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} u| = |w|$,
2. $h(x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} u) = h(w)$ und
3. $h(\overline{x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} u}) = h(\overline{w})$.

Beweis mit Induktion über n:

Induktionsanfang: Es ist $n = 0$ und $u = w$. Die Invarianten sind erfüllt.

Induktionsvoraussetzung: Die Invarianten sind für n erfüllt.

Induktionsschritt: Es gibt nur $|M|^2$ verschiedene Paare von Monoidelementen. Also müssen unter den $|M|^2 + 1$ Wörtern

$$u[0,0], u[0,1], \ldots, u[0,|M|^2]$$

zwei sein, bei denen ihre Inversen und sie selbst auf jeweils dieselben Monoidelemente abgebildet werden. Damit existieren $0 \leq \alpha < \beta \leq |M|^2$, so daß $h(u[0,\alpha]) = h(u[0,\beta]) \wedge h(\overline{u[0,\alpha]}) = h(\overline{u[0,\beta]})$ ist.

1. $\neg\exists 1 \leq k \leq n : |y_k| = \beta - \alpha \wedge h(y_k) = h(u[\alpha,\beta]) \wedge h(\overline{y_k}) = h(\overline{u[\alpha,\beta]})$

Es ist

$$\begin{aligned} x_1y_1^{r_1}x_2y_2^{r_2}\cdots x_ny_n^{r_n}u &= x_1y_1^{r_1}x_2y_2^{r_2}\cdots x_ny_n^{r_n}u[0,\alpha]u[\alpha,\beta]u[\beta,|u|] \\ &= x_1y_1^{r_1}x_2y_2^{r_2}\cdots x_ny_n^{r_n}x_{n+1}y_{n+1}^{r_{n+1}}u[\beta,|u|]. \end{aligned}$$

Also sind nach den Zuweisungen $u := u[\beta,|u|]$ und $n := n+1$ die Invarianten erfüllt. Man beachte, daß

$$h(x_{n+1}) = h(u[0,\alpha]) = h(u[0,\beta]) = h(u[0,\alpha]u[\alpha,\beta]) = h(x_{n+1}y_{n+1}),$$
$$h(\overline{x_{n+1}}) = h(\overline{u[0,\alpha]}) = h(\overline{u[0,\beta]}) = h(\overline{u[0,\alpha]u[\alpha,\beta]}) = h(\overline{x_{n+1}y_{n+1}})$$

ist.

2. $\exists 1 \leq k \leq n : |y_k| = \beta - \alpha \wedge h(y_k) = h(u[\alpha,\beta]) \wedge h(\overline{y_k}) = h(\overline{u[\alpha,\beta]})$
   Die erste Invariante ist erfüllt, da

$$\begin{aligned} |u| &= |u[0,\alpha]| + |u[\alpha,\beta]| + |u[\beta,|u|]| \\ &= |u[0,\alpha]| + |y_k| + |u[\beta,|u|]| \\ &= |u[0,\alpha]u[\beta,|u|]| + |y_k| \end{aligned}$$

   ist. Es ist

$$\begin{aligned} &h(x_1y^{r_1}\cdots x_{k-1}y_{k-1}^{r_{k-1}}x_ky_k^{r_k}x_{k+1}y_{k+1}^{r_{k+1}}\cdots x_ny_n^{r_n}u) \\ &= h(x_1y_1^{r_1}\cdots x_{k-1}y_{k-1}^{r_{k-1}})h(x_ky_k^{r_k})h(x_{k+1}y_{k+1}^{r_{k+1}}\cdots x_ny_n^{r_n})h(u) \end{aligned}$$

   und

$$\begin{aligned} &h(\overline{x_1y_1^{r_1}\cdots x_{k-1}y_{k-1}^{r_{k-1}}x_ky_k^{r_k}x_{k+1}y_{k+1}^{r_{k+1}}\cdots x_ny_n^{r_n}u}) \\ &= h(\overline{u})h(\overline{x_{k+1}y_{k+1}^{r_{k+1}}\cdots x_ny_n^{r_n}})h(\overline{x_ky_k^{r_k}})h(\overline{x_1y_1^{r_1}\cdots x_{k-1}y_{k-1}^{r_{k-1}}}). \end{aligned}$$

Es sind auch die zweite und dritte Invariante erfüllt, da

$$\begin{aligned}
h(x_k y_k^{r_k}) &= h(x_k)h(y_k^{r_k}) = h(x_k y_k)h(y_k^{r_k}) = h(x_k y_k^{r_k+1}),\\
h(u) &= h(u[0,\beta]u[\beta,|u|]) = h(u[0,\beta])h(u[\beta,|u|])\\
&= h(u[0,\alpha])h(u[\beta,|u|]) = h(u[0,\alpha]u[\beta,|u|]),\\
h(\overline{x_k y_k^{r_k}}) &= h(\overline{y_k^{r_k}})h(\overline{x_k}) = h(\overline{y_k^{r_k}})h(\overline{x_k y_k}) = h(\overline{x_k y_k^{r_k+1}}),\\
h(\overline{u}) &= h(\overline{u[0,\beta]u[\beta,|u|]}) = h(\overline{u[\beta,|u|]})h(\overline{u[0,\beta]})\\
&\quad h(\overline{u[\beta,|u|]})h(\overline{u[0,\alpha]}) = h(\overline{u[0,\alpha]u[\beta,|u|]})
\end{aligned}$$

ist und $r_k := r_k + 1$ und $u := u[0,\alpha]u[\beta,|u|]$ gesetzt wird.

Nach der while-Schleife und der Zuweisung $x_{n+1} := u$ gilt also

$$\begin{aligned}
|x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}| &= |w|,\\
h(x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}) &= h(w),\\
h(\overline{x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}}) &= h(\overline{w}).
\end{aligned}$$

Es ist $|x_i| = \alpha < \beta \leq |M|^2$ und $1 \leq \beta - \alpha = |y_i| \leq \beta \leq |M|^2$ für $1 \leq i \leq n$. Aufgrund der while-Schleifen Bedingung $|u| \geq |M|^2$ ist $|x_{n+1}| = |u| < |M|^2$. Es werden maximal $|M|^4$ viele y_i erstellt, da es nur $|M|^2$ verschiedene Paare von Monoidelemente gibt und $1 \leq |y_i| \leq |M|^2$ ist. Damit gilt $0 \leq n \leq |M|^4$. Es ist $r_i \leq |w|$, da $|y_i| \geq 1$ ist. □

LEMMA 9.1.2. *Es sei w ein Wort, M ein Monoid und $h : \Sigma^* \to M$ ein Homomorphismus. Dann existiert ein Wort v, so daß $|v| = |w|$, $h(v) = h(w)$, $h(\overline{v}) = h(\overline{w})$, $w = \overline{w} \Rightarrow v = \overline{v}$ und v durch eine eins-eindeutige Grammatik in Chomsky-Normalform der Größe $\mathcal{O}(|M|^4 log(|w|) + |M|^6)$ erzeugt wird.*

BEWEIS. Wenn $w \neq \overline{w}$ ist, berechnen wir mit dem Algorithmus 9.1.1 das Wort v für w. Es ist $|v| = |w|$, $h(v) = h(w)$ und $h(\overline{v}) = h(\overline{w})$. Zusätzlich ist

$$v = x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}$$

mit $0 \leq n \leq |M|^4$, $|x_i| < |M|^2$ für $1 \leq i \leq n+1$, $1 \leq |y_i| \leq |M|^2$ und $1 \leq r_i \leq |v|$ für $1 \leq i \leq n$. Wir erstellen eins-eindeutige Grammatiken in Chomsky-Normalform für die Wörter $x_1, x_2, \ldots, x_{n+1}$ und die

Wörter $y_1, y_2, \ldots, y_n$. Diese Grammatiken haben zusammen die Größe $\mathcal{O}(n|M|^2) = \mathcal{O}(|M|^6)$. Jetzt erstellen wir eins-eindeutige Grammatiken in Chomsky-Normalform, die mit Hilfe der Grammatiken für $y_1, y_2, \ldots, y_n$ die Wörter $y_1^{r_1}, y_2^{r_2}, \ldots, y_n^{r_n}$ erzeugen. Diese Grammatiken haben zusätzlich die Größe $\mathcal{O}(|M|^4 \log(|w|))$. Abschließend erstellen wir die eins-eindeutige Grammatik in Chomsky-Normalform für $v = x_1 y_1^{r_1} x_2 y_2^{r_2} \cdots x_n y_n^{r_n} x_{n+1}$, die auf den Grammatiken für $x_1, x_2, \ldots, x_{n+1}$ und $y_1^{r_1}, y_2^{r_2}, \ldots, y_n^{r_n}$ aufbaut. Diese Grammatik hat zusätzlich die Größe $\mathcal{O}(|M|^4)$. Insgesamt hat die Grammatik für v also die Größe $\mathcal{O}(|M|^4 \log(|v|) + |M|^6)$.

Wenn $w = \overline{w}$ ist, muß $w = w' a \overline{w'}$ sein mit $w' \in \Sigma^*$ und $a \in I \cup \{\epsilon\}$. Für w' und $\overline{w'}$ erstellen wir wie im Fall $w \neq \overline{w}$ beschrieben, die eins-eindeutigen Grammatiken in Chomsky-Normalform für v' und $\overline{v'}$. Diese haben jeweils die Größe $\mathcal{O}(|M|^4 \log(|w'|) + |M|^6)$. Auf diese Grammatiken aufbauend können wir eine eins-eindeutige Grammatik in Chomsky-Normalform der Größe $\mathcal{O}(|M|^4 \log(|w|) + |M|^6)$ für $v = v' a \overline{v'}$ erzeugen. Es ist

$$\begin{aligned} |v| &= |v'| + |a| + |\overline{v'}| = |w'| + |a| + |\overline{w'}| = |w|, \\ v &= v' a \overline{v'} = \overline{v}, \\ h(v) &= h(v')h(a)h(\overline{v'}) = h(w')h(a)h(\overline{w'}) = h(w) \end{aligned}$$

und $h(\overline{v}) = h(\overline{w})$, da $h(v) = h(w)$, $v = \overline{v}$ und $w = \overline{w}$ ist. □

Da eine eins-eindeutige Grammatik in Chomsky-Normalform mit der Größe m einfach in eine eins-eindeutige Intervall Grammatik mit der Größe m transformiert werden kann, können wir jetzt das entscheidende Lemma beweisen.

LEMMA 9.1.3. *Wenn es eine Lösung σ mit $|\sigma(L)| = N$ für die FMA-Gleichung $E : L = R$ mit regulären Randbedingungen gibt, dann gibt es auch eine Lösung σ' mit $|\sigma'(L)| = N$, so daß das Wort $\sigma'(L)$ durch eine Intervall Grammatik der Größe $\mathcal{O}(d|M|^4(log(N)+|M|^2))$ erzeugt wird, d.h. das Wort $\sigma'(L)$ ist gut komprimierbar.*

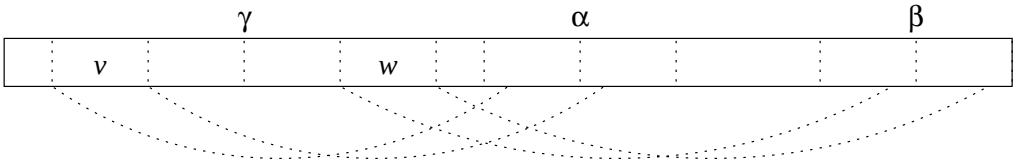
BEWEIS. Wir ändern den Homomorphismus σ so ab, daß dadurch in $\sigma(L)$ und $\sigma(R)$ alle maximalen freien Intervalle durch gut komprimierbare Wörter entsprechend Lemma 9.1.2 ersetzt werden. Den entstehenden Homomorphismus nennen wir σ'.

Dies ist möglich, da aufgrund von Lemma 6.3.4 sich maximale freie Intervalle nicht überlappen, freie Intervalle immer vollständig in einer Variablen enthalten sind, ein freies Intervall immer in allen Vorkommen einer Variable und an derselben Position vorkommt und das neue Wort auch identisch mit seinem Inversen ist, wenn dies beim maximalen freien Intervall der Fall war. Es ist σ' eine Lösung, da $\sigma'(L) = \sigma'(R)$ ist und die regulären Randbedingungen weiterhin erfüllt sind, weil $\forall X \in \Omega : h(\sigma(X)) = h(\sigma'(X))$ ist. Es ist $|\sigma'(L)| = N$, da die maximalen freien Intervalle durch Wörter derselben Länge ersetzt werden.

Aufgrund von Lemma 6.3.5 gibt es maximal $2d-2$ Wörter w_j, die einem maximalen freien Intervall entsprechen. Wir erstellen eins-eindeutige Intervall Grammatiken $G_j(V_j, \Sigma, P_j, S_j)$, die die Wörter w_j erzeugen. Nach Lemma 9.1.2 können wir Grammatiken erstellen mit $|G_j| \in \mathcal{O}(|M|^4 \log(N) + |M|^6)$.

Wir erstellen nun eine eins-eindeutige Intervall Grammatik für $\sigma'(L)$. Das prinzipielle Vorgehen ist dasselbe wie in Lemma 8.2.1. Wir erstellen Nichtterminale $A_{\gamma,i,s}$, die genau die Wörter $\lfloor \gamma - s2^i, \gamma + s2^i \rfloor$ erzeugen für $i \geq 0$ und ein Wort, das einem maximalen freien Intervall entspricht, für $i = -1$. Für alle Cuts γ, $-1 \leq i \leq \lceil \log(N) \rceil$ und $s \in \{-1, 1\}$ für $i \geq 0$ bzw. $s \in \{-1, 0, 1\}$ für $i = -1$ erstellen wir die Produktionen

$$A_{\gamma,i,s} \rightarrow \begin{cases} S_{j_{\gamma,s}} & \text{wenn } i = -1, \\ \lfloor \gamma - s, \gamma + s \rfloor & \text{wenn } i = 0, \\ A_{\alpha,i_\alpha,s_\alpha}[\mu_\alpha, \nu_\alpha] A_{\gamma,i-1,s}[0, |A_{\gamma,i-1,s}|] & \\ \quad A_{\beta,i_\beta,s_\beta}[\mu_\beta, \nu_\beta] & \text{wenn } i > 0 \wedge s = 1, \\ A_{\beta,i_\beta,-s_\beta}[|A_{\beta,i_\beta,s_\beta}| - \nu_\beta, |A_{\beta,i_\beta,s_\beta}| - \mu_\beta] & \\ \quad A_{\gamma,i-1,s}[0, |A_{\gamma,i-1,s}|] & \\ \quad A_{\alpha,i_\alpha,-s_\alpha}[|A_{\alpha,i_\alpha,s_\alpha}| - \nu_\alpha, |A_{\alpha,i_\alpha,s_\alpha}| - \mu_\alpha] & \text{wenn } i > 0 \wedge s = -1. \end{cases}$$



Im Beweis von Lemma 6.3.5 war d nur eine Abschätzung für die Anzahl der Cuts. Es gibt also höchstens doppelt soviele Wörter, die einem maximalen freien Intervall entsprechen, wie Cuts. Wir wählen die $j_{\gamma,s}$, so daß alle diese Wörter von den Produktionen $A_{\gamma,-1,s}$ erzeugt werden und $S_{j_{\gamma,-s}}$ das inverse Wort wie $S_{j_{\gamma,s}}$ erzeugt. Welche Produktion welches Wort erzeugt, ist egal.

Die Produktionen $A_{\gamma,i,-s}$ erzeugen genau die Inversen der Wörter, die von den Produktionen $A_{\gamma,i,s}$ erzeugt werden.

Es sei γ ein beliebiger Cut, $1 \leq i \leq \lceil \log(N) \rceil$ und $s \in \{-1, 1\}$. Wir bestimmen α, i_α, s_α, μ_α und ν_α,wie folgt. Es sei $v := \lfloor \gamma - 2^i, \gamma - 2^{i-1} \rfloor$. Wenn das v entsprechende Intervall nicht frei ist, kommt v über einem Cut α vor. Also kommt v in dem von $A_{\alpha,i_\alpha,s_\alpha}$ mit $i_\alpha := i-1$ erstellten Wort vor. Wenn das v entsprechende Intervall frei ist, kommt v in einem maximalen freien Intervall vor. Das Wort, das diesem Intervall entspricht, wird von einer Produktion $A_{\alpha,i_\alpha,s_\alpha}$ mit $i_\alpha := -1$ erstellt. In beiden Fällen sind μ_α und ν_α die Start- und Endposition von v in $A_{\alpha,i_\alpha,s_\alpha}$. Für $w = \lfloor \gamma + 2^i, \gamma + 2^{i-1} \rfloor$ können wir β, i_β, s_β, μ_β und ν_β entsprechend bestimmen.

Für alle Cuts γ erzeugt das Nichtterminal $A_{\gamma,\lceil \log(N) \rceil,1}$ genau das Wort $\sigma(L)$, da

$$\lfloor \gamma - 2^{\lceil \log(N) \rceil}, \gamma + 2^{\lceil \log(N) \rceil} \rfloor = \sigma(L)[0, |\sigma(L)|] = \sigma(L)$$

ist.

Die Produktionen für $A_{\gamma,i,s}$ sind keine Produktionen einer Intervall Grammatik, da einzelne Nichtterminale, Wörter und drei Nichtterminale auf der rechten Seite vorkommen. Wir können sie jedoch einfach durch Intervall Grammatik Produktionen ersetzen. Die Nichtterminale S_j ersetzen wir durch $S_j[0, |S_j|]S_j[0, 0]$. Die Wörter mit einer maximalen Länge von 2 können wir mit maximal drei Produktionen darstellen. Die Produktionen

mit drei Nichtterminalen können wir durch jeweils zwei Produktionen ersetzen. Da es maximal d Cuts gibt, hat die Intervall Grammatik insgesamt also höchstens

$$\mathcal{O}(d(|M|^4\log(N) + |M|^6 + \log(N))) = \mathcal{O}(d|M|^4(\log(N) + |M|^2))$$

Produktionen. □

9.2. Der Algorithmus

Der folgende Algorithmus ist sehr ähnlich zu dem Algorithmus 8.3.1.

ALGORITHMUS 9.2.1. *Es sei $E : L = R$ eine FMA-Gleichung mit den regulären Randbedingungen A_X. Der Algorithmus GleichungLösen entscheidet in NEXPTIME, ob die Gleichung eine Lösung der Länge N hat. Wenn das Monoid M klein ist, d.h. $|M| \in n^{\mathcal{O}(1)}$ ist, läuft der Algorithmus in NP.*

procedure *GleichungLösen*$(E : L = R, A_X, N)$
; σ sei eine gut komprimierbare Lösung der Länge N
Berechne ein Monoid M und $\forall X : M_X \subseteq M$ mit $h^{-1}(M_X) = L(A_X)$
$\forall X \in \Omega_{\frac{1}{2}}$: *Rate eine* eins-*eindeutige Grammatik in Chomsky-Normalform, die $\sigma(X)$ erzeugt*
$\forall X \in \overline{\Omega_{\frac{1}{2}}}$: *Erstelle eine* eins-*eindeutige Grammatik in Chomsky-Normalform die $\sigma(X)$ erzeugt*
if $\exists X \in \Omega : h(\sigma(X)) \notin M_X$ then return false
if $\sigma(L) \neq \sigma(R)$ then return false
if $|\sigma(L)| \neq N$ then return false
return true
end *GleichungLösen*

BEWEIS. Wenn es eine Lösung mit der Länge N gibt, existiert nach Lemma 9.1.3 auch eine Lösung σ, so daß es eine eins-eindeutige Intervall Grammatik der Größe $\mathcal{O}(d|M|^4(\log(N) + |M|^2))$ gibt, die das Wort $\sigma(L)$ erzeugt. Also existiert auch für jede Variable $X \in \Omega_{\frac{1}{2}}$ eine eins-eindeutige Intervall Grammatik der Größe $\mathcal{O}(d|M|^4(\log(N) + |M|^2))$, die das Wort $\sigma(X)$ erzeugt. Diese Grammatiken können mit dem Algorithmus

8.1.4 in eins-eindeutige Grammatiken in Chomsky-Normalform der Größe $\mathcal{O}((d|M|^4(\log(N) + |M|^2))^2)$ umwandelt werden.

Es ist $N \in 2^{\mathcal{O}(n)}$. Wenn $|M| \in n^{\mathcal{O}(1)}$ ist, haben die Grammatiken insgesamt maximal die Größe $n^{\mathcal{O}(1)}$ und der Algorithmus läuft in NP. Sonst haben die Grammatiken insgesamt maximal die Größe $2^{\mathcal{O}(n^2)}$ und der Algorithmus läuft in NEXPTIME.

Ob die regulären Randbedingungen erfüllt sind, läßt sich einfach überprüfen, indem man die Monoidelemente für alle Nichtterminale in den Grammatiken in Chomsky-Normalform berechnet. Ob $\sigma(L) = \sigma(R)$ ist, kann man testen, indem man aus den bestehenden Grammatiken eine Grammatik in Chomsky-Normalform erstellt mit zwei Nichtterminalen A_L und A_R, die genau die Wörter $\sigma(L)$ und $\sigma(R)$ erzeugen und dann den Algorithmus aus Satz 7.3.2 benutzt. Ob $|\sigma(L)| = N$ ist, läßt sich einfach überprüfen, indem die Längen der erzeugten Wörter für alle Nichtterminale in den Grammatiken in Chomsky-Normalform berechnet. □

Da die Länge der Lösung exponentiell länger als die Eingabe sein kann, kann in NEXPTIME bzw. NP die Lösung nicht ausgegeben werden. Dies ist jedoch in 2-DEXPTIME bzw. DEXPTIME möglich.

THEOREM 9.2.2. *In 2-DEXPTIME kann eine Lösung der Länge N für eine FMA-Gleichung mit regulären Randbedingungen berechnet werden, wenn es eine solche Lösung gibt. Wenn das Monoid M klein ist, d.h. $|M| \in n^{\mathcal{O}(1)}$ ist, kann die Lösung in DEXPTIME berechnet werden.*

BEWEIS. Der Algorithmus 9.2.1 entscheidet in NEXPTIME bzw. NP, ob die Gleichung eine Lösung der Länge N hat. Der Algorithmus kann in 2-DEXPTIME bzw. DEXPTIME simuliert werden. Wenn es eine Lösung gibt, werden vom Algorithmus eins-eindeutige Grammatiken in Chomsky-Normalform für $\sigma(X)$ berechnet. Die Wörter die von diesen Grammatiken erzeugt werden, können in 2-DEXPTIME bzw. DEXPTIME ausgegeben werden. □

KAPITEL 10

Schlußbemerkungen

Betrachten wir das Problem, ob eine FMA-Gleichung ohne reguläre Randbedingungen eine Lösung hat. Es stellt sich die Frage, ob es NP- oder PSPACE-vollständig ist. Auf der einen Seite ist das Problem NP-schwierig und auf der anderen Seite haben wir einen PSPACE Algorithmus. Diese Frage manifestiert sich auch in der Länge einer minimalen Lösung. Die Länge ist bisher nur doppelt exponentiell beschränkt. Doch es ist kein Beispiel bekannt mit einer mehr als exponentiellen Länge. Wenn wir die Länge exponentiell beschränken können, ist das Problem NP-vollständig.

Eine andere interessante Frage ist, ob sich Makanins Algorithmus weiter optimieren läßt, so daß er auch in PSPACE läuft. Dies wäre für eine Implementierung von Vorteil, weil Makanins Algorithmus gezielter nach einer Lösung sucht als die hier vorgestellten Algorithmen und man dadurch meistens schneller eine positive Antwort finden könnte.

Literaturverzeichnis

[A87] Habib Abdulrab, Résolution d'équations sur les mots: Etude et implémentation LISP de l'algorithme de Makanin, Ph. D. Thesis, Universite de Rouen, 1987.

[A79] Dana Angluin, Finding Patterns Common to a Set of Strings, Proceedings of the 11th ACM Symposium on the Theory of Computing (STOC), 130-141, ACM Press, 1979.

[BS96] Franz Baader und Klaus U. Schulz, Unification in the Union of Disjoint Equational Theories: Combining Decision Procedures, Journal of Symbolic Computation 21(2), 211-243, 1996.

[D98] Volker Diekert, Makanin's Algorithm for Solving Word Equations with Regular Constraints, Bericht 1998/02, Fakultät Informatik, Universität Stuttgart, 1998.

[D00] Volker Diekert, Makanin's Algorithm, Jean Berstel und Dominique Perrin, Editoren, Algebraic Combinatorics on Words, Cambridge University Press, 2000.

[DMV96] Anatoli Degtyarev, Yuri Matiyasevich und Andrei Voronkov, Simultaneous E-Unification and Related Algorithmic Problems, Proceedings of the 11th Annual IEEE Symposium on Logic in Computer Science (LICS), 494-502, IEEE Computer Society Press, 1996.

[F88] William M. Farmer, A Unification Algorithm for Second Order Monadic Terms, Annals of Pure and Applied Logic, Vol. 39, 131-174, 1988.

[FW65] Nathan J. Fine und Herbert S. Wilf, Uniqueness Theorems for Periodic Functions, Proceedings of the American Mathematical Society, Vol. 16, 109-114, 1965.

[G98] Claudio Gutiérrez, Satisfiability of Word Equations with Constants is in Exponential Space, 39th Annual Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society Press, 1998.

[G00A] Claudio Gutiérrez, Equations in free Semigroups with anti-involution and their relation to equations in free Groups, Proceedings Latin American Theoretical INformatics (LATIN), 2000.

[G00B] Claudio Gutiérrez, Satisfiability of Equations in Free Groups is in PSPACE, 32nd Annual ACM Symposium on Theory of Computing (STOC), 2000.

[GS78] Joachim von zur Gathen and Malte Sieveking, A Bound on Solutions of Linear Integer Equalities and Inequalities, Proceedings of the American Mathematical Society, 72(1), 155-158, 1978.

[GV97] Yuri Gurevich and Andrei Voronkov, Monadic Simultaneous Rigid E-Unification and Related Problems, 24th International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science, Vol. 1256, 154-165, Springer-Verlag, 1997.

[H59] Marshall Hall, The Theory of Groups, The Macmillan Company, 1959.

[HHI98] Vesa Halava, Tero Harju und Lucian Ilie, Periods and binary words technical report, Turku Centre for Computer Science (TUCS), Number TUCS-TR-213, 1998.

[IP99] Lucian Ilie und Wojciech Plandowski, Two-Variable Word Equations, Turku Centre for Computer Science (TUCS), Number TUCS-TR-306, 1999.

[KP96] Antoni Kościelski und Leszek Pacholski, Complexity of Makanin's Algorithm, Journal of the ACM 43(4), 670-684, 1996.

[KP98] Antoni Kościelski und Leszek Pacholski, Makanin's algorithm is not primitive recursive, Theoretical Computer Science 191(1-2), 145-156, 1998.

[KPM97] Juhani Karhumäki, Wojciech Plandowski und Filippo Mignosi, The Expressibility of Languages and Relations by Word Equations, 24th International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science, Vol. 1256, 98-109, Springer-Verlag, 1997.

[L83] M. Lothaire, Combinatorics on Words, Vol. 17 of Encyclopedia of Mathematics and its Applications. Addison-Wesley, Reading, MA.

[L87a] Jean-Luc Lambert, Le Problème de l'accessibilité dans les réseaux de Petri, Ph. D. Thesis, Universite de Orsay, 1987.

[L87b] Jean-Luc Lambert, Une borne pour les générateurs des solutions entières positives d'une équation diophantienne linéaire, Compte-rendu de L'Académie des Sciences de Paris, 305(1), 39-40, 1987.

[M77] Gennadií S. Makanin, The Problem of Solvability of Equations in a Free Semigroup, Mat. Sb., 103(2), 147-236, russisch; englische Übersetzung in Math. USSR Sbornik, 32, 129-198, 1977.

[M82] Gennadií S. Makanin, Equations in a Free Group, Izvestiya Akad. Nauk SSSR, Ser. Mat. 46, 1199-1273, 1982, russisch; englische Übersetzung in Math. USSR Izvestiya, 21, 483-546, 1983.

[M84] Gennadií S. Makanin, Decidability of the Universal and Positive Theories of a Free Group, Izvestiya Akad. Nauk SSSR, Ser. Mat. 48, 735-749, 1984, russisch; englische Übersetzung in Math. USSR Izvestiya, 25, 75-88, 1985.

[P94] Wojciech Plandowski, Testing Equivalence of Morphisms on Context-Free Languages, Second Annual European Symposium on Algorithms (ESA), Lecture Notes in Computer Science, Vol. 855, 460-470, Springer-Verlag, 1994.

[P99A] Wojciech Plandowski, Satisfiability of Word Equations with Constants is in NEXPTIME, Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC), 1999.

[P99B] Wojciech Plandowski, Satisfiability of Word Equations with Constants is in PSPACE, Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS), 495-500, IEEE Computer Society Press, 1999.

[PR98] Wojciech Plandowski und Wojciech Rytter, Application of Lempel-Ziv Encodings to the Solution of Word Equations, Annual International Colloquium on Automata Languages and Programming (ICALP), 1998.

[RD99] John M. Robson und Volker Diekert, On Quadratic Word Equations, 16th Symposium on Theoretical Aspects of Computer Science (STACS), Lecture Notes in Computer Science, Vol. 1563, 217-226, Springer-Verlag, 1999.

[S92] Klaus U. Schulz, Makanin's Algorithm for Word Equations: Two Improvements and a Generalization, Lecture Notes in Computer Science, Vol. 572, 85-150, Springer-Verlag, 1992.