Theory Comput Syst (2010) 47: 250–258
DOI 10.1007/s00224-008-9153-7

The Solvability Problem for Quadratic Equations over Free Groups is NP-Complete

O. Kharlampovich · I.G. Lysënok · A.G. Myasnikov · N.W.M. Touikan

Published online: 25 October 2008

Abstract We prove that the problems of deciding whether a quadratic equation over a free group has a solution is NP-complete.

Keywords Equations over free groups · NP-completeness

1 Introduction

The study of quadratic equations over free groups started with the work of Malc′ev [6] and has been deepened extensively ever since. One of the reasons research in this topic has been so fruitful is the deep connection between quadratic equations and the topology of surfaces.

In this paper we will show that the problem of deciding if a quadratic equation over a free group is satisfiable is NP-complete. This problem was shown to be decidable in [1]. In addition it was shown in [4, 8], and [5] that if n, the number of variables, is fixed, then deciding if a quadratic equation has a solution can be done in time polynomial in the sum of the lengths of the coefficients. These results imply that the problem is solvable in at most exponential time. We will improve on this by proving (see Theorem 2.2) that deciding if a quadratic equation over a free group has a solution is in NP.

In [2] it is shown that deciding if a quadratic word equation has a solution is NP-hard. We will prove (see Theorem 3.11) that deciding if a quadratic equations over a free group has a solution is also NP-hard. Our proofs are geometric, relying on the topological results of [8] and disc diagram techniques.

O. Kharlampovich · A.G. Myasnikov · N.W.M. Touikan (✉)
McGill University, Montreal, QC, Canada
e-mail: touikan@math.mcgill.ca

I.G. Lysënok
Steklov Institute, Moscow, Russia

2 The Solvability Problem for Quadratic Equations over Free Groups Is in NP

Let A be a finite alphabet and let A^{-1} be a set of formal inverses of elements of A. We denote by $(A \cup A^{-1})^*$ the free monoid with involution with basis A and for $w \in (A \cup A^{-1})^*$, we denote by w^{-1} its involution. We denote by $F(A)$ the free group on A.

2.1 Standard Form

A quadratic equation E with variables $\{x_i, y_i, z_j\}$ and non-trivial coefficients $\{w_i, d\} \in F(A)$ is said to be in *standard form* if its coefficients are expressed as freely and cyclically reduced words in A^* and E has either the form:

$$\left(\prod_{i=1}^{g}[x_i, y_i]\right)\left(\prod_{j=1}^{m-1} z_j^{-1} w_j z_j\right) d = 1 \quad \text{or} \quad \left(\prod_{i=1}^{g}[x_i, y_i]\right) d = 1 \tag{1}$$

where $[x, y] = x^{-1}y^{-1}xy$, in which case we say it is *orientable* or it has the form

$$\left(\prod_{i=1}^{g} x_i^2\right)\left(\prod_{j=1}^{m-1} z_j^{-1} w_j z_j\right) d = 1 \quad \text{or} \quad \left(\prod_{i=1}^{g} x_i^2\right) d = 1 \tag{2}$$

in which case we say it is non-orientable. The *genus* of a quadratic equation is the number g in (1) and (2) and m is the number of coefficients. If $g = 0$ then we will define E to be orientable. If E is a quadratic equation we define its *reduced Euler characteristic*, $\overline{\chi}$ as follows:

$$\overline{\chi}(E) = \begin{cases} 2 - 2g & \text{if } E \text{ is orientable} \\ 2 - g & \text{if } E \text{ is not orientable} \end{cases}$$

We finally define the *length* of a quadratic equation E to be

$$\text{length}(E) = |w_1| + \cdots + |w_{n-1}| + d + 2 \quad \text{(number of variables)}$$

It is a well known fact that an arbitrary quadratic equation over a free group can be brought to a standard form in time polynomial in its length.

2.2 Ol'shanskii's Result

The following is proved in [8].

Theorem 2.1 *Let E be a quadratic equation over $F(A)$ in standard form. If $g = 0$, $m = 2$, or E is not orientable and $g = 1, m = 1$ then we set $N = 1$. Otherwise we set $N = 3(m - \overline{\chi}(E))$. E has a solution if and only if for some $n \leq N$;*

(i) *there is a set $P = \{p_1, \ldots p_n\}$ of variables and a collection of m discs $D_1, \ldots, D_m$ such that;*

(ii) *the boundaries of these discs are circular* 1*-complexes with directed and labeled edges such that each edge has a label in* P *and each* $p_j \in P$ *occurs exactly twice in the union of boundaries*;

(iii) *if we glue the discs together by edges with the same label, respecting the edge orientations, then we will have a collection* $\Sigma_0, \ldots, \Sigma_l$ *of closed surfaces and the following inequalities*: *if* E *is orientable then each* Σ_i *is orientable and*

$$\left(\sum_{i=0}^{l} \chi(\Sigma_i)\right) - 2l \geq \overline{\chi}(E)$$

if E *is non-orientable either at least one* Σ_i *is non-orientable and*

$$\left(\sum_{i=0}^{l} \chi(\Sigma_i)\right) - 2l \geq \overline{\chi}(E)$$

or, each Σ_i *is orientable and*

$$\left(\sum_{i=0}^{l} \chi(\Sigma_i)\right) - 2l \geq \overline{\chi}(E) + 2$$

and

(iv) *there is a mapping* $\overline{\psi} : P \to (A \cup A^{-1})^*$ *such that upon substitution, the coefficients* $w_1, \ldots, w_{m-1}$ *and* d *can be read without cancellations around the boundaries of* $D_1, \ldots, D_{m-1}$ *and* D_m, *respectively*; *and finally that*

(v) *if* E *is orientable the discs* $D_1, \ldots, D_m$ *can be oriented so that* w_i *is read clockwise around* ∂D_i *and* d *is read clockwise around* ∂D_m, *moreover all these orientations must be compatible with the gluings*.

Proof It is shown in Sect. 2.4 [8] that the solvability of a quadratic equation over $F(A)$ coincides with the existence of a *diagram* Δ *over* $F(A)$ on the appropriate surface Σ with boundary. This diagram may not be *simple*, so via surgeries we produce from Σ a finite collection of surfaces $\Sigma_1, \ldots, \Sigma_l$ with induced simple diagrams $\Delta_1, \ldots, \Delta_l$ which we can recombine to get back Σ and Δ. So existence of a diagram Δ on Σ is equivalent to existence of a collection of simple diagrams Δ_i on surfaces Σ_i such that the inequalities involving Euler characteristics given in the statement of the Theorem are satisfied.

In Sect. 2.3 of [8] the bounds on n are proved. It is also shown in that section that if one can glue discs together as described in the statement of the Theorem with the condition on the boundaries, then there exist simple diagrams Δ_i on surfaces Σ_i. □

2.3 The Certificate

Theorem 2.1 enables us to construct a good certificate.

Theorem 2.2 *There exists a polynomial time algorithm $\mathcal{A}$ such that a quadratic equation E over $F(A)$ in standard form has a solution if and only if there is a certificate c of size bounded by*

$$2(|w_1| + \cdots + |w_m - 1| + |d| + 3(2g + m)) \leq 8 * \text{length}(E)$$

such that $\mathcal{A}$ answers "yes" on the input (E, c).

Proof The certificate will consist of the following:

1. A collection of variables $P = \{p_1, \ldots, p_n\}$ where $n \leq \max\{3(2g + m), 1\}$.
2. A collection of substitutions $\overline{\psi} = \{p_i \mapsto a_i,\ i = 1, \ldots, n\}$ where $a_i \in (A \cup A^{-1})^*$.
3. A collection of words in P^*

$$\mathcal{C} = \begin{cases} C_1 = p_{11}^{\epsilon_{11}} \cdots p_{1l}^{\epsilon_{1j(l)}} \\ \cdots \\ C_m = p_{m1}^{\epsilon_{m1}} \cdots p_{mj(m)}^{\epsilon_{mj(m)}} \end{cases}$$

with $p_{ij} \in P, \epsilon_{ij} \in \{-1, 1\}$ and each $p_i \in P$ occurring exactly twice.

The $C_i's$ represent the labels of the boundaries of the discs $D_1, \ldots, D_l$. It follows that checking conditions (i) and (ii) of Theorem 2.1 can be done quickly, moreover we see that the size of $\mathcal{C}$ is at most $2n \leq 6(2g + m)$.

$\overline{\psi}$ extends to a monoid homomorphism $\psi : (P \cup P^{-1})^* \to (A \cup A^{-1})^*$. (iv) can also be verified quickly since for $i = 1, \ldots, m - 1$ we just need to check that some cyclic permutation of $\psi(C_i)$ is equal to w_i and some cyclic permutation of $\psi(C_m)$ is equal do d. Moreover, since the equality is graphical we have that

$$|a_1| + \cdots + |a_n| \leq |w_1| + \cdots + |w_m| + |d|$$

Therefore the size of the certificate is bounded as advertised. All that is left is to determine the topology of the glued together discs. We describe the algorithm without too much detail.

Step 1: *Built a forest of discs*: We make a graph Γ such that each vertex $v_i \in V(\Gamma)$ corresponds the disc D_i and each edge $e_j \in E(\Gamma)$ corresponds to the variable $p_j \in P$. The edge e_k goes from v_i to v_j if and only if the variable p_k occurs in the boundary of D_i and in the boundary of D_j or if $i = j$ then there are two different occurrences of the variable p_k. We construct a spanning forest $\mathcal{F}$. This enables us to count the number of connected components $\Sigma_0, \ldots, \Sigma_l$.

Step 2: *Determine orientability*: For each maximal tree $T_r \subset \mathcal{F}$ we get a "tree of discs" by gluing together only the pairs of edges whose labels correspond to elements of $E(T_r)$. The resulting tree of discs is a simply connected topological space that can be embedded in the plane and we can read a cyclic word $c(T_r)$ in P^* along its boundary. The surface Σ_r obtained by gluing together the remaining paired edges of the tree of discs will be orientable only if whenever $p_j^{\pm 1}$ occurs in $c(T_r)$ then $p_j^{\mp 1}$ also occurs. We can also check (v) at this point.

Step 3: *Compute Euler characteristic*: The identification of the boundary of the discs with graphs, enables us to think of the discs as polygons. If a disc D_i has N_i

sides then we give each corner of D_i an angle of $\pi(N_i - 2)/N_i$. Then for each tree of discs produced in the previous step, we identify the remaining pairs of edges to get the surfaces $\Sigma_0, \ldots, \Sigma_l$, which now have an extra angular structure. To each Σ_i, we can apply the Combinatorial Gauss-Bonnet Theorem (see Sect. 4 of [7]) which states that for an angled two-complex X,

$$2\pi\chi(X) = \sum_{f \in X^{(2)}} \kappa(f) + \sum_{v \in X^{(0)}} \kappa(v)$$

where $X^{(2)}$ is the set of faces and $X^{(0)}$ is the set of vertices. This angle assignment gives each face f a curvature $\kappa(f) = 0$ and each vertex has curvature

$$\kappa(v) = 2\pi - \left(\sum_{c \in link(v)} \measuredangle(c) \right)$$

i.e. $\kappa(v)$ is 2π minus the sum of the angles that meet at v.

With an appropriate data structure one can perform steps 1–3 (not necessarily in sequential order) in at most quadratic time in the size of $\mathcal{C}$. Once all that is done, verifying the inequalities of (iii) is easy and we are finished. □

3 The Solvability Problem for Quadratic Equations over Free Groups Is NP-Hard

We will present the bin packing problem which is known to be NP-complete and show that it is equivalent to deciding if a certain type of quadratic equation has a solution.

3.1 Bin Packing

Problem 3.1 (Bin Packing)

- *INPUT*: *A k-tuple of positive integers $(r_1, \ldots, r_k)$ and positive integers B, N.*
- *QUESTION*: *Is there a partition of $\{1, \ldots, k\}$ into N subsets*

$$\{1, \ldots, k\} = S_1 \sqcup \cdots \sqcup S_N$$

such that for each $i = 1, \ldots, N$ we have

$$\sum_{j \in S_i} r_j \leq B \tag{3}$$

This problem is NP-hard in the strong sense (see [3, p. 226]), i.e. there are NP-hard instances of this problem when both B and the r_j are bounded by a polynomial function of k.

Let $t = NB - \sum_{i=1}^{k} r_i$. Then by replacing $(r_1, \ldots, r_k)$ by the $k + t$-tuple $(r_1, \ldots, r_k, \ldots, 1, \ldots, 1)$ we can assume that the inequalities (3) are actually equalities. This modified version is still NP hard in the strong sense. We state it explicitly:

Problem 3.2 (Exact Bin Packing)

- *INPUT*: *A k-tuple of positive integers $(r_1, \ldots, r_k)$ and positive integers B, N.*
- *QUESTION*: *Is there a partition of $\{1, \ldots, k\}$ into N subsets*

$$\{1, \ldots, k\} = S_1 \sqcup \cdots \sqcup S_N$$

such that for each $i = 1, \ldots, N$ we have

$$\sum_{j \in S_i} r_j = B \tag{4}$$

The authors warmly thank Laszlo Babai for drawing their attention to this problem in connection to tiling problems.

3.2 Tiling Discs

Throughout this section we will consider the discs to be embedded in the Euclidean plane $\mathbb{E}^2$ and will always read clockwise around closed curves.

Definition 3.3 An $[a, b^n]$-*disc* is a disc as in Sect. 2.2 equipped with an orientation along whose boundary one can read the cyclic word $[a, b^n]$ in the clockwise direction. *We will always assume that $n \geq 1$.*

Definition 3.4 An $[a, b^n]$-*ribbon* is a rectangular cell complex embedded in $\mathbb{E}^2$ obtained by attaching $[a, b^j]$-discs by their a-labeled edges, while respecting orientation, such that we can read $[a, b^n]$ along its boundary. The *top* of an $[a, b^n]$ ribbon is the boundary subpath along which we can read the word b^{-n}, the *bottom* is the boundary subpath along which we can read the word b^n.

Definition 3.5 Let D be a disc embedded in $\mathbb{E}^2$ tiled by coherently oriented $[a, b^n]$-discs. We define the *a-pattern* of D to be the graph defined as follows:

1. In the middle of each a-labeled edge put a vertex.
2. Between any two vertices contained in the same $[a, b^n]$-disc draw an edge.

Connected components of a-patterns are called *a-tracks*.

Lemma 3.6 *A disc D embedded in $\mathbb{E}^2$ tiled by finitely many coherently oriented $[a, b^n]$-discs cannot have any circular a-tracks.*

Proof It is clear that every a-track is a graph whose vertices have valence at most 2. If an a-track t has vertices of valence 1 then they must lie on ∂D.

Suppose towards a contradiction that D has a circular a-track c. Then c divides D into two components: an interior and an exterior. If we examine the interior we see that it is a planar union of discs with only the letter b occurring on its boundary, it follows that the interior contains a disc D' with circular a-track. Repeating the argument we find that D must have infinitely many cells which is a contradiction. □

Corollary 3.7 *Let D be a disc embedded in $\mathbb{E}^2$ tiled by finitely many coherently oriented $[a, b^n]$-discs. Then it is impossible for an a-track t to start and end inside a segment $\alpha \subset \partial D$ labeled a^m for some $m \geq 1$.*

Proof Suppose towards a contradiction that this was not true. Then for some D and some a-track t in D, we have that t starts and ends in some arc $\alpha \subset \partial D$ labeled a^m. Without loss of generality α lies on the x-axis of $\mathbb{E}^2$ and consider the reflection about the x-axis, then we have a resulting disc D', and reversing the orientations of all the b-labeled edges, makes D' another disc tiled by finitely many $[a, b^n]$-discs. Attaching D to D' along α gives a new disc D'' that has a circular a-track, contradicting Lemma 3.6 □

Corollary 3.8 *We cannot tile a sphere S with finitely many coherently oriented $[a, b^n]$-discs.*

Proof Suppose towards a contradiction that this was possible. Then in particular all the a-tracks are closed and compact and therefore circles. If S contains only one a-track t, then S is obtained as some topological quotient of an annulus A such that A is obtained by gluing the edges labeled a in the boundary of some $[a, b^N]$-ribbon. Now ∂A consists of two circles c_1, c_2 with label b^N. Since t separates S into two discs, we see that the images of c_1, c_2 are disjoint via the quotient map $\pi : A \to S$. It therefore follows that π must continuously map c_1, which has label b^N, to something simply connected, i.e. a simplicial tree, while respecting the orientations of the edges, which is impossible.

Otherwise S has at least two a-tracks, if we remove from S some $[a.b^n]$-disc D not lying in some track t. Then $S - D$ embeds into E^2, has a circular a-track, and therefore contradicts Lemma 3.6. □

Lemma 3.9 *Let R be some $[a, b^N]$-ribbon. Suppose there is a continuous map $\psi : R \to D$ where D is a disc embedded in $\mathbb{E}^2$ tiled by finitely many coherently oriented $[a, b^n]$-discs, such that ψ is injective on the interior of R and sends edges to edges, labels to labels and preserves edge orientations. Then ψ is an embedding.*

Proof Let $t \subset R$ be the unique a-track and let t_R and b_R be the top and bottom of R respectively. By Lemma 3.6 the edges labeled a of ∂R have disjoint images. We can remove $[a, b^n]$-discs from D to get a smaller disc D' such that $\psi(t)$ separates D' into two pieces. From this it is clear that the images $\psi(t_R) \cap \psi(b_R)$ are disjoint. It follows that the only possible failures of injectivity are in restrictions to t_R or b_R. Suppose ψ is not injective on, say, t_R. Then if $\psi(t_R)$ bounds a sub-disc in $D'' \leq D$, then we see that D'' must have a circular a-track—contradiction. It follows that $\psi(t_R)$ maps onto a tree of edges labeled b, but this would contradict the fact that ψ preserved edge orientations. □

Proposition 3.10 *Suppose that D is a disc embedded in $\mathbb{E}^2$ with boundary label $[a^N, b^B]$ that is the result of gluings of $[a, b^n]$-discs respecting the orientation, then it is obtained from a collection of M $[a, b^B]$-ribbons $R_1, \ldots, R_M$ such that the bottom of R_{i+1} is glued to the top of R_i, $i = 1, \ldots, M$.*

Proof We divide ∂D into four arcs l_a, t_b, r_a, b_b that have labels a^{-N}, b^{-B}, a^N, b^B respectively, i.e. the left, top, right and bottom sides. By Lemma 3.6 and Corollary 3.7 each a-track starts in l_a and ends in r_a. By Lemma 3.9 each a-track lies in an embedded ribbon. Since each $[a, b^n]$ disc lies in one of these ribbons, it follows that D is obtained by gluing together N-ribbons as stated in the Proposition. Now b_b must lie in the bottom-most ribbon R_1 which means that R_1 is an $[a, b^B]$-ribbon. It follows that all the ribbons are $[a, b^B]$-ribbons. □

3.3 A Special Genus Zero Quadratic Equation

Equipped with Proposition 3.10 we shall deduce NP hardness of the following equation:

$$\prod_{j=1}^{k} z_j^{-1}[a, b^{n_j}]z_j = [a^N, b^B] \tag{5}$$

By Theorem 2.1, (5) has a solution if and only if there is a collection of discs D_j with boundary labels $[a, b^{n_j}]$ for $j = 1, \ldots, k$ respectively and a disc D_m with boundary label $[a^N, b^B]$ such that, glued together in a way that respect labels and orientation of edges, form a union of spheres (this is forced by the first inequality in (iii), Theorem 2.1.

Theorem 3.11 *Deciding if the quadratic equation* (5) *with coefficients*

$$[a, b^{n_1}], \ldots, [a, b^{n_k}] \quad \textit{and} \quad [a^N, b^B]$$

has a solution is equivalent to deciding if problem 3.2; *with input* $(n_1, \ldots, n_m)$ *and positive integers* B, N; *has a positive answer.*

Proof "Bin packing ⇒ solution." Suppose that Problem 3.2 has a positive answer on the specified inputs. For each subset S_i of the given partition of $\{1, \ldots, k\}$ we form a $[a, b^B]$-ribbon R_i by gluing together the $[a, b^{n_j}]$-discs for $j \in S_i$, this is possible by (iv) of Theorem 2.1 and equation (4). We then construct one hemisphere by gluing the ribbons $R_1, \ldots, R_N$. The other hemisphere is the remaining disc with boundary label $[a^N, b^B]^{-1}$, the resulting sphere proves the solvability of (5) with the given coefficients.

"Solution ⇒ bin packing." If (5) has a solution then there is a union of spheres tiled with $[a, b^{n_i}]$-discs and one $[a^N, b^B]^{-1}$-disc, moreover these discs are coherently oriented. By condition (v) and Corollary 3.8 there can only be one sphere: the sphere S_0 containing the unique $[a^N, b^B]^{-1}$-disc. If we remove this $[a^N, b^B]^{-1}$-disc from S_0 what remains will be a disc D with boundary label $[a^N, b^B]$ tiled with $[a, b^{n_i}]$-discs. Applying Proposition 3.10 divides D into ribbons $R_1, \ldots, R_N$ and we immediately see that these ribbons provide a partition of $\{n_1, \ldots, n_k\}$, showing that Problem 3.2 has a positive solution on the given input. □

References

1. Comerford, L.P., Jr., Edmunds, C.C.: Quadratic equations over free groups and free products. J. Algebra **68**(2), 276–297 (1981)
2. Diekert, V., Robson, J.M.: Quadratic word equations. In: Jewels Are Forever, pp. 314–326. Springer, Berlin (1999)
3. Garey, M.R., Johnson, D.S.: Computers and Intractability. A Guide to the Theory of NP-Completeness. A Series of Books in the Mathematical Sciences. Freeman, New York (1979)
4. Grigorchuk, R.I., Kurchanov, P.F.: On quadratic equations in free groups. In: Proceedings of the International Conference on Algebra, Part 1 (Novosibirsk, 1989). Contemp. Math., vol. 131, pp. 159–171. Amer. Math. Soc., Providence (1989)
5. Grigorchuk, R.I., Lysionok, I.G.: A description of solutions of quadratic equations in hyperbolic groups. Int. J. Algebra Comput. **2**(3), 237–274 (1992)
6. Mal'cev, A.I.: On the equation $zxyx^{-1}y^{-1}z^{-1} = aba^{-1}b^{-1}$ in a free group. Algebra Log. Sem. **1**(5), 45–50 (1962)
7. McCammond, J.P., Wise, D.T.: Fans and ladders in small cancellation theory. Proc. Lond. Math. Soc. (3) **84**(3), 599–644 (2002)
8. Ol'shanskiĭ, A.Yu.: Diagrams of homomorphisms of surface groups. Sib. Mat. Z. **30**(6), 150–171 (1989)