Math. Ann. 196, 8—22 (1972)

Markoff Forms and Primitive Words

HARVEY COHN*

1. Introduction

Almost one hundred years ago, in 1879, Markoff's paper «Sur les formes binaires indéfinies» appeared in this journal [10]. The paper was concerned with the quadratic form (in $\boldsymbol{Z}$)

$$\Phi(x_1, x_2) = a x_1^2 + a' x_1 x_2 + a'' x_2^2$$

with minimum $m(\Phi)$, of $|\Phi|$ over its non-zero values, and discriminant $d(\Phi) = (a')^2 - 4aa''$ (> 0), not a perfect square. The problem was to consider when the following relation holds:

$$\mu(\Phi) = m(\Phi)/d(\Phi)^{1/2} > 1/3 .$$

Markoff showed that the inequality was satisfied only for a special sequence of forms Φ_a and forms (properly or improperly) equivalent to $g\Phi_a$, $(0 < 2g \in \boldsymbol{Z})$, (all called "Markoff forms"). Here the integers a are "Markoff numbers" defined as one of a "Markoff triple" (a, b, c) of positive integers satisfying "Markoff's equation"

$$a^2 + b^2 + c^2 = 3abc ,$$

$a = 1, 2, 5, 13, 29, 34, 89, \ldots$. For each such value, the form Φ_a was constructed such that

$$m(\Phi_a) = a, \;\; d(\Phi_a) = 9a^2 - 4, \;\; \mu(\Phi_a) = a/(9a^2 - 4)^{1/2} .$$

Most of Markoff's attention went to the construction of the purely periodic part of the continued fraction[1] for the roots ξ and η of the factored form:

$$\Phi_a(x, y) = a(x_1 - \xi x_2)(x_1 - \eta x_2) .$$

* This paper was supported in part by (U.S.) National Science Foundation Grant P-13967 and in part by Air Force Office of Scientific Research Grant AF-AFOSR-69-1712.

[1] Markoff's rule of formation of periods will not be repeated here because of a more suitable version in § 6 (below).

The first three such forms correspond to the triple (1, 2, 5):

$$\Phi_1(x_1, x_2) = x_1^2 + x_1 x_2 - x_2^2, \quad \text{period} \quad [1,1], \; \mu(\Phi_1) = 1/5^{1/2},$$

$$\Phi_2(x_1, x_2) = 2x_1^2 + 4x_1 x_2 - 2x_2^2, \quad \text{period} \quad [2,2], \; \mu(\Phi_2) = 1/8^{1/2},$$

$$\Phi_5(x_1, x_2) = 5x_1^2 + 11 x_1 x_2 - 5x_2^2, \quad \text{period} \quad [2,2,1,1], \; \mu(\Phi_5) = 5/221^{1/2}.$$

It is of great historical interest that $\mu(\Phi) > 1/3$ only for Markoff forms, with the discrete[2] sequence $a/(9a^2-4)^{1/2}$, and it is of further interest that even with real coefficients in Φ (and $x_1, x_2 \in \boldsymbol{Z}$) no other forms Φ arise with $\mu(\Phi) > 1/3$, (see [4]), (assuming "min" is modified to "non-zero inf" in $m(\Phi)$).

In the earlier papers [2] (in 1955) and [3] (in 1971), the author developed Markoff forms in triples, like Markoff numbers, by using triples of "Markoff matrices". This required a combination of techniques of Fricke, who in 1896 (see [5]) had used triples of matrices in $\boldsymbol{SL}_2(\boldsymbol{R})$ to generate Fuchsian groups, and Frobenius, who in 1913, (see [6]), had used single matrices in $\boldsymbol{SL}_2(\boldsymbol{Z})$ to generate Markoff forms. Thus the relationship of Markoff forms to the modular group and to geodesics in the Poincaré-Klein fashion (on a perforated torus) was developed.

The purpose of this present paper is to show how the idea of Markoff matrix-triples leads to theorems on the word problem for the free group $\boldsymbol{Z} * \boldsymbol{Z}$, specifically a normalized form for the primitive words and a "Nielsen-type" partition of some primitive words into a basis pair. The inevitable interpretation in terms of homotopy is also made.

Incidentally, the role of matrix-triples rather than single forms may be of some further interest in light of a connection recently made by Hirzebruch between Markoff triples and algebraic manifolds [7], using identities of Rademacher on Dedekind sums [14].

2. Main Theorem

Let $\boldsymbol{F}_2$ be the free group $\boldsymbol{Z} * \boldsymbol{Z}$ on the generators (A, B) and let $\boldsymbol{\Gamma}$ be the automorphism group on $\boldsymbol{F}_2$. Thus we write

$$G(A,B) = (w_1(A,B), w_2(A,B)), \qquad G \in \boldsymbol{\Gamma},$$

where w_1 or w_2 are the so-called primitive words of $\boldsymbol{F}_2$ (shown as a finite product):

$$w_i(A,B) = \prod_j A^{t_i^{(j)}} B^{u_i^{(j)}}; \qquad t_i^{(j)}, u_i^{(j)} \in \boldsymbol{Z}.$$

[2] This is unlike the case of definite forms, where the corresponding $\mu(\Phi)$ lies dense in the interval between the (Gaussian) maximum $1/3^{1/2}$ and 0.

Let F_2^A be the abelian free group $\boldsymbol{Z} \oplus \boldsymbol{Z}$ on the generators (A, B) and let $\Gamma^A (= GL_2(\boldsymbol{Z}))$ be the automorphism group on F_2^A. The canonical projection, π, of course, operates as follows:

$$\pi w_i(A, B) = \sum_j t_i^{(j)} A + \sum_j u_i^{(j)} B = t_i A + u_i B, \qquad (i = 1, 2),$$

and

$$\pi G = \begin{pmatrix} t_1 & u_1 \\ t_2 & u_2 \end{pmatrix}, \qquad (t_1 u_2 - t_2 u_1 = \pm 1).$$

Let Γ_R be the subgroup of Γ defined as $\pi^{-1}(G_{12})$ where

$$G_{12} = \left\{ \pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \pm \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \pm \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \right.$$
$$\left. \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \pm \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix}, \pm \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix} \right\}.$$

Clearly, $G_{12} \approx Z_2 \times S_6$ (the symmetric group); i.e., G_{12} is the diedral group of the hexagon, and, moreover, Γ_R includes the inner automorphisms (Γ_I^+) of F_2. We shall later demonstrate the structure $\Gamma_R / \Gamma_I^+ = G_{12}$ (see §4 below).

We shall use symbols of type Γ_R^+ to mean the subgroup of Γ_R for which $\pi(\Gamma_R^+) \in SL_2(\boldsymbol{Z})$, the "oriented" or "proper" automorphisms and Γ_R^- to mean the coset, $\pi(\Gamma_R^-) \notin SL_2(\boldsymbol{Z})$.

Theorem 2.1 *(Main Theorem). The primitive word $w = w(A, B)$ has the unique normalized form*

$$M_{r,s}(A, B) = \prod_{t=1}^{r} A B^{a_t + 2}, M_{0,1}(A, B) = B,$$

$a_t = [ts/r] - [(t-1)s/r]$, $(r > 0, s \geqq 0, (r, s) = 1)$, *to within (several possible) automorphisms of F_2 lying in Γ_R. Specifically, for every primitive word there exists an $H_R \in \Gamma_R$ and a unique integer pair (r, s), $(r \geqq 0, s \geqq 0, (r, s) = 1)$, such that*

$$w(H_R(A, B)) = M_{r,s}(A, B).$$

Finally, the automorphism H_R is uniquely determined to within an inner automorphism except when $(r, s) = (0, 1)$ or $(1, 0)$ in which cases there are two cosets, one in Γ_R^+, the other in Γ_R^-.

Corollary 2.2. *Any primitive word expressible as positive powers of B and AB^2 must be of the form $M_{r,s}(A, B)$ or a cyclic rearrangement thereof. Thus the assignment of integers r and $\left(\sum_1^r a_t = \right) s$, with $r \geqq 0$, $s \geqq 0$, $(r, s) = 1$ completely (cyclically) determines the individual addends a_t.*

To take some simple illustrations:

$M_{0,1}(A,B)=B, M_{1,0}(A,B)=AB^2, M_{1,1}(A,B)=AB^3, M_{1,2}(A,B)=AB^4,$ $M_{2,1}(A,B)=AB^2AB^3$, etc.

$$w(A,B)=B^{-1}\Rightarrow w(A^{-1},B^{-1})=w(AB,B^{-1})=B,$$
$$w(A,B)=A\Rightarrow w(B,A)=w(B,B^{-1}A^{-1})=B,$$
$$w(A,B)=AB^{-1}\Rightarrow w(ABA^{-1},B^{-1}A^{-1})=w(AB,B^{-1})=AB^2,$$
$$w(A,B)=BA^2\Rightarrow w(B,A)=w(B^{-1}AB^2,B^{-1}A^{-1}B)=AB^2,$$
$$w(A,B)=BA^3\Rightarrow w(B,A)=AB^3.$$

In each case the automorphisms in Γ_R shown here are the only possible ones to within inner automorphisms. The computational procedure appears in Lemma 4.2 (below), and the cases are of later interest for remarks in § 7. On a more difficult level, $M_{3,4}(A,B) = AB^3AB^3AB^4$ is primitive but not $AB^2AB^3AB^5$, by the corollary. Also $M_{r,s}(A^{-1},B^{-1})^{-1}$, the "reversal", must be a cyclic rearrangement of $M_{r,s}(A,B)$ (also see Lemma 6.1 below).

3. Some Automorphism Subgroups

Here we shall list some group presentations which shall be required below. In all the following cases, the classical techniques (see [12, 13, 9]) suffice and details may be omitted:

$$P(A,B)=(B,A),\quad R(A,B)=(B^{-1},A),\quad S(A,B)=(B,B^{-1}A^{-1}).$$
$$\Gamma=\{P,R,S;P^2,R^4,S^3,(PR^2)^2,(PSR^2)^2,RS^2R^2S=SR^2S^2R\}.$$
$$\Gamma^+=\{R,S;R^4,S^3,RS^2R^2S=SR^2S^2R\}.$$
$$\Gamma_R=\{P,R^2,S;P^2,(R^2)^2,S^3,(PR^2)^2,(PSR^2)^2\}.$$
$$\Gamma_R^+=\{R^2,S;(R^2)^2,S^3\}.$$
$$\Gamma_I=\{P,R^2SR^2S^2,S^2R^2SR^2;P^2,P(R^2SR^2S^2)P=S^2R^2SR^2\}.$$
$$\Gamma_I^+=\{R^2SR^2S^2,S^2R^2SR^2\}.$$
$$\Gamma_F=\{P,SR,S^2R^2;P^2,P(SR)P=(S^2R^2)^{-1}(SR),(SRS^2R^2)^2=(S^2R^2)^3\}.$$
$$\Gamma_F^+=\{SR,S^2R^2;(SRS^2R^2)^2=(S^2R^2)^3\}.$$
$$\Gamma_N^+=\{RS,SR;(RS)(SR)(RS)=(SR)(RS)(SR)\}.$$

For simplicity, we shall take Γ_R to be defined primarily by the above presentation and verify (in § 4 below) it is the same as in § 2 (above).

The symbols $\Gamma^+, \Gamma_R^+, \Gamma_I^+, \Gamma_F^+$ denote the corresponding subgroup formed by suppressing the automorphism P, the only generator which reverses "orientation". These are invariant subgroups of index 2, as is better seen from the "conjugation" relations:

$$PRP = R^3,\ PR^2P = R^2,\ PSP = R^2S^2R^2,\ PS^2P = R^2SR^2.$$

(The second conjugation relation of Γ_F, $P(S^2R^2)P = (S^2R^2)^{-1}$, of course, follows from the first.)

Here Γ_I^+ corresponds to the inner automorphisms of F_2,

$$R^2SR^2S^2(A,B) = (A, ABA^{-1}),\ S^2R^2SR^2(A,B) = (BAB^{-1}, B).$$

Also, Γ_F^+ is the automorphism subgroup[3] which preserves the commutator $K = B^{-1}A^{-1}BA$, i.e.,

$$SR(A,B) = (A, A^{-1}B),\ S^2R^2(A,B) = (BA, A^{-1}).$$

The (full) Γ_F at worst interchanges K and K^{-1}. Fricke used Γ_F^+, (see [5, 2]), for real Fuchsian groups of genus 1, (compare § 5 below).

Finally $\Gamma_N^+ = R\Gamma_F^+R^{-1}$ is an involutory transformation of Γ_F^+ preserving $AB^{-1}A^{-1}B$. It has the extension property by $RPR^{-1} = PR^2$, (not by P), so we do not write "Γ_N". The group Γ_N^+ is related to partitions of primitive words of Nielsen's type (see § 8 below).

4. Relation to Modular Group

We now imbed F_2 in the modular group G_M so that the inner automorphisms of G_M will generate the subgroup Γ_R^+ in Γ^+.

For convenience, we take G_M as $PSL_2(Z)$, but we represent elements of G_M as matrix pairs $\pm T$ for $T \in SL_2(Z)$. Thus if $T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$ then $\pm T$ determines an element of G_M,

$$T(z) = (t_{11}z + t_{12})/(t_{21}z + t_{22}).$$

(As usual, I is the identity matrix and $\pm I$ leads to $I(z) = z$.) Then, (see [9, 8]),

$$G_M = Z_2 * Z_3 = \{X(z), Y(z); X^2(z), Y^3(z)\}$$

where $X = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $Y = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$. Note $X^2 = -I$, but $X^2(z) = I(z)$. (The

[3] The author is indebted to B. H. Neumann for advice on the word structure problem.

group Γ_R^+ is isomorphic to G_M by the interpretation following below.)

We next imbed in G_M its commutator $F_2 = [G_M, G_M]$, the free group generated by

$$A_0 = X^{-1}Y^{-1}XY = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad B_0 = YXY^{-1}X^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

It will be useful to use the same symbol for the free group $F_2 = \{A, B\}$ and $\{A_0, B_0\}$, the imbedded subgroup of G_M. Actually, F_2 is also the commutator subgroup of $SL_2(Z)$.

Now we see Γ_R^+ as the subgroup of those automorphisms of F_2 arising as inner automorphisms of $PSL_2(Z)\,(=G_M)$:

$$(A_0(z), B_0(z)) \mapsto (TA_0T^{-1}(z), TB_0T^{-1}(z)), \; T \in SL_2(Z).$$

It suffices to note

$$T = X \Rightarrow (A_0, B_0) \mapsto (A_0^{-1}, B_0^{-1}) = R^2(A_0, B_0)$$

$$T = Y \Rightarrow (A_0, B_0) \mapsto (B_0, A_0^{-1}B_0^{-1}) = R^2S^2R^2(A_0, B_0).$$

These automorphisms generate Γ_R^+.

We must consider, however, the "extended" inner automorphism subgroup of F_2, with $J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ extending $SL_2(Z)$ to $GL_2(Z)$, corresponding to $J(z) = -z$. Since $A_0(-z) = B_0(z)$ then the automorphism generated by J, i.e.

$$(A_0(z), B_0(z)) \mapsto (JA_0J^{-1}(z), JB_0J^{-1}(z))$$

is merely $P(A_0, B_0) = (B_0, A_0)$. Then the automorphism group Γ_R^+ becomes extended to Γ_R.

We finally equate the definitions of Γ_R in § 2 and § 3 by considering the coset structure (using Nielsen's result [11, 9] $\pi^{-1}(I) = \Gamma_I^+$)

$$\Gamma_R/\Gamma_I^+ = G_{12} = \{H_{t,u,v}\},$$

$$H_{t,u,v} = P^t R^{2u} S^{2v}, \qquad t, u (\mathrm{mod}\, 2), v (\mathrm{mod}\, 3):$$

$$I(A,B) = (A,B),\; S(A,B) = (B, B^{-1}A^{-1}),\; S^2(A,B) = (B^{-1}A^{-1}, A),$$

$$R^2(A,B) = (A^{-1}, B^{-1}),\; R^2S(A,B) = (B^{-1}, AB),\; R^2S^2(A,B) = (AB, A^{-1}),$$

$$P(A,B) = (B,A),\; PS(A,B) = (B^{-1}A^{-1}, B),\; PS^2(A,B) = (A, B^{-1}A^{-1}),$$

$$PR^2(A,B) = (B^{-1}, A^{-1}),\; PR^2S(A,B) = (AB, B^{-1}),\; PR^2S^2(A,B) = (A^{-1}, AB).$$

The projections $\pi(H_{t,u,v})$ into Γ^A, (the automorphisms of the corresponding free group $\mathbf{Z}\oplus\mathbf{Z}$), give as before

$$G_{12}=\left\{\pm\begin{pmatrix}1&0\\0&1\end{pmatrix},\ \pm\begin{pmatrix}0&1\\-1&-1\end{pmatrix},\ \pm\begin{pmatrix}-1&-1\\1&0\end{pmatrix};\right.$$
$$\left.\pm\begin{pmatrix}0&1\\1&0\end{pmatrix},\ \pm\begin{pmatrix}-1&-1\\0&1\end{pmatrix},\ \pm\begin{pmatrix}1&0\\-1&-1\end{pmatrix}\right\}.$$

Lemma 4.1. *Consider again the canonical projection of $F_2=\mathbf{Z}*\mathbf{Z}$ onto $F_2^A=\mathbf{Z}\oplus\mathbf{Z}$ so that*

$$\pi\prod_j A^{t(j)}B^{u(j)}=\left(\sum_j t^{(j)}\right)A+\left(\sum_j u^{(j)}\right)B$$

and let $pA+qB$ be any primitive vector of F^A, i.e. $p,q\in\mathbf{Z}$, $(p,q)=1$. Then there exists a unique normalized (primitive word) $M_{r,s}(A,B)$ and an automorphism $G\in\Gamma_R$ such that

$$\pi M_{r,s}(G(A,B))=pA+qB\,.$$

Here G is uniquely determined to within inner automorphisms Γ_I^+ (as a member of $H_{t,u,v}$) with two exceptions:

$$\pi M_{0,1}(G(A,B))=B \qquad \text{for}\quad G=\Gamma_I^+ \quad \text{or}\quad \Gamma_I^+(PS)\,,$$
$$\pi M_{1,0}(G(A,B))=A+2B \quad \text{for}\quad G=\Gamma_I^+ \quad \text{or}\quad \Gamma_I^+(PR^2S^2)\,.$$

For proof, note that the primitive words $M_{r,s}(A,B)$ of the main theorem cover exactly those $tA+uB$ where $(t,u)=1$, $u\geqq 2t\geqq 0$. Now take the general primitive vector $pA+qB$ and apply the matrices of G_{12} to (A,B). Then (p,q) will transform (by the "dual" matrices) into the following 12 pairs (t_i,u_i):

$$\pm(p,q),\ \pm(-q,p-q),\ \pm(-p+q,-p);$$
$$\pm(q,p),\ \pm(-p,-p+q),\ \pm(p-q,-q)\,.$$

We assert that exactly one of these 12 pairs satisfies $u_i\geqq 2t_i\geqq 0$ unless $(p,q)=\pm(1,0)$, $\pm(0,1)$, $\pm(1,1)$ in one case, or $\pm(1,2)$, $\pm(2,1)$, $\pm(1,-1)$ in the other. This assertion, in homogeneous form, $(\xi=p/q)$, states that

$$\xi,\ 1/(1-\xi),\ (\xi-1)/\xi;\ 1/\xi,\ (1-\xi),\ \xi/(\xi-1)$$

are six functions (forming the group S_6) which transform the closed interval $[2,\infty]$ so that it covers $[-\infty,\infty]$ with only end-point duplication (in the two exceptional cases).

In terms of the diedral group, G_{12}, the exceptional vectors correspond to the 60° and the 30° lines of symmetry of the regular hexagon.

Lemma 4.2. *To reduce a primitive word $w(A, B)$ to normal form, substitute the set of 12 automorphisms $H_{t,u,v}(A, B)$ for (A, B) and test each word $w(H_{t,u,v}(A, B))$ to see if negative powers can be cancelled by automorphisms of Γ_I^+, i.e., look for a $w(H_{t,u,v}(H_I(A, B)))$ with $H_I \in \Gamma_I^+$ which has only positive powers of AB^2 and B. Then rotate cyclically to form $M_{r,s}$. (If $w(A, B)$ is not a primitive word, the process will reveal this "finitistically".)*

Criteria of this type have many variations; (see Nielsen [11] and Lemma 7.1 below).

We note, for now, that Lemma 4.1 provides the uniqueness part (of the selection of H_R) in Theorem 2.1 using Lemma 4.2 as the algorithm.

5. Matrix Version of Markoff's Theorem

We speak of a "Markoff matrix" as a matrix $T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} \in \boldsymbol{SL}_2(\boldsymbol{Z})$ whose fixed points ($T(z) = z$ for $z = \xi, \eta$) are roots of a Markoff form,

$$\Phi(x_1, x_2) = t_{21}x_1^2 + (t_{22} - t_{11})x_1x_2 - t_{12}x_2^2 .$$

We shall generate triples of such matrices according to two theorems in [2]:

Theorem 5.1. *Let A, B, C be matrices of $\boldsymbol{SL}_2(\boldsymbol{Z})$ satisfying*

$$ABC = I, \ CBA = K = \begin{pmatrix} -1 & -6 \\ 0 & -1 \end{pmatrix}.$$

Then A, B, C are a Markoff matrix-triple. One third of the traces yields the corresponding Markoff numbers

$$a = \boldsymbol{S}(A)/3, \qquad b = \boldsymbol{S}(B)/3, \qquad c = \boldsymbol{S}(C)/3$$

which satisfy the equation of the Markoff-triple:

$$a^2 + b^2 + c^2 = 3abc .$$

Theorem 5.2. *Matrices giving all classes of Markoff forms can be found by starting with*

$$A_0 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \quad C_0 = \begin{pmatrix} 3 & -1 \\ 1 & 0 \end{pmatrix}$$

and $(a_0, b_0, c_0) = (1, 1, 1)$ *and then applying* Γ *to* (A, B) *with* C *always defined as* $B^{-1}A^{-1}$. *Thus*[4]

$$P(A, B) = (B, A), \qquad P(a, b, c) = (b, a, c)$$
$$R(A, B) = (B^{-1}, A), \qquad R(a, b, c) = (b, a, c')$$
$$S(A, B) = (B, B^{-1}A^{-1}), \qquad S(a, b, c) = (b, c, a).$$

Here $c' = 3ab - c$, *the "second root" of Markoff's equation (in* c*).*

Remarks. In the earlier work [2], the operations in Theorem 5.2 were restricted to the subgroup Γ_F^+ which preserves the commutator $B^{-1}A^{-1}BA = K$, (for the benefit of Theorem 5.1). This is not really necessary since

$$P(B^{-1}A^{-1}BA) = K^{-1}, \qquad R(B^{-1}A^{-1}BA) = BKB^{-1},$$
$$S(B^{-1}A^{-1}BA) = AKA^{-1}.$$

Thus by "extended" automorphisms of F_2, $H \in \Gamma_I$, we obtain new matrices, $H(A, B)$, which satisfy Theorem 5.1. Thus the "Markoff" property (and traces) of the new matrices remains the same. Specifically, under Γ_F^+, (compare § 3 above),

$$SR(a, b, c) = (a, c, b), \quad S^2R^2(a, b, c) = (c, b', a),$$

$(b' = 3ac - b)$, thus generating all Markoff triples, (see [2]). In this paper Γ_N^+ will be used instead (in § 8 below).

6. Nature of the Period

The roots of a Markoff form were shown in [10] to have a purely periodic expansion of "double ones" [1, 1] and "double twos" [2, 2] with special arrangements, which we explain now according to [3]. Let

$$V_1(z) = 1 + 1/(1 + 1/z) = B_0^{-1}(z)$$
$$V_2(z) = 2 + 1/(2 + 1/z) = B_0^{-1}A_0^{-1}B_0^{-1}(z).$$

Then an arbitrary Markoff period is uniquely determined by a pair of integers (r, s), with $r \geqq 0$, $s \geqq 0$, $(r, s) = 1$ by a pattern of "double ones" and "double twos" corresponding to the matrix

$$M_{r,s}^*(A_0, B_0) = V_2 V_1^{a_1} V_2 V_1^{a_2} \dots V_2 V_1^{a_r}; M_{0,1}^* = V_1,$$

[4] The terminology $H(a, b, c) = (a^*, b^*, c^*)$ means that $H(A, B, C) = (A^*, B^*, C^*)$ with $3a = S(A)$, $3a^* = S(A^*)$, etc., for any Markoff matrix-triple. The independence of the trace relation (for arbitrary Markoff matrix-triple) is the very essence of Fricke's identities (see [5, 2]).

and $a_t = [ts/r] - [(t-1)s/r]$, (as in Theorem 2.1). This becomes our normal form (in terms of $A = A_0$, $B = B_0$) if we note the words

$$V_1 = v_1(A, B) = B^{-1}, \quad V_2 = v_2(A, B) = B^{-1}A^{-1}B^{-1},$$
$$S^2R^2S(A, B) = (B^{-1}A^{-1}B, B^{-1}),$$
$$v_1(S^2R^2S(A, B)) = B, \quad v_2(S^2R^2S(A, B)) = AB^2,$$
$$M^*_{r,s}(S^2R^2S(A, B)) = M_{r,s}(A, B).$$

Lemma 6.1. *If the sequence* $a_t = [ts/r] - [(t-1)s/r]$, $(r, s) = 1$, $r > 0$, $s > 0$ *is written in reverse order for* $1 \leqq t \leqq r$

$$a_r, a_{r-1}, \ldots, a_1$$

then the sequence is equivalent to a cyclic shift (modulo r) of its direct order $a_1, \ldots, a_r$. *In terms of automorphisms, for* $H_I^+ \in \Gamma_I^+$ *and* $H_R^+ = H_I^+ R^2 \in \Gamma_R^+$,

$$M_{r,s}(A^{-1}, B^{-1})^{-1} = M_{r,s}(H_I^+(A, B)),$$
$$M_{r,s}(A, B)^{-1} = M_{r,s}(H_R^+(A, B)).$$

For proof, we use induction. First we can reduce s to its least residue (modulo r), and next we can replace r by s since the count of the "zeros" ($a_t = 0$) between the "ones" ($a_t = 1$) is what must be shifted cyclically by the reversal of order. In this way the problem becomes reduced to small r and s for which it is trivial.

7. Proof of Main Theorem

Let $w(A, B)$ be a primitive word in F_2. Then the matrix $T = w(A_0, B_0)$ is a Markoff matrix by Theorem 5.2. The problem is that the matrix contains "more information" than the corresponding form.

Lemma 7.1. *The Markoff matrix T is equivalent to an exact power of some* $M_{r,s}(A_0, B_0)$, *i.e., for some* $N \in GL_2(\mathbf{Z})$, *and integer* $u(\neq 0)$

$$\pm NTN^{-1} = M_{r,s}(A_0, B_0)^u.$$

To sketch the proof of this lemma briefly, let N transform[5] the fixed points of T into those of some $M_{r,s}(A_0, B_0)$, (by § 6). Now the matrices NTN^{-1} and $M_{r,s}(A_0, B_0)$ have the same fixed points, and the eigenvalues of each are units (of norm 1) lying in a quadratic field. The existence of a fundamental unit assures us that both of these matrices are (positive

[5] Since the continued fractions are generated by $N_1(z) = 1 + z$ and $N_2(z) = 1/z$, it is necessary to permit matrices of determinant -1 for the transform N (See remark at end of § 7.)

or negative) powers of a common matrix. Since $M_{r,s}$ is not a higher power (by construction), Lemma 7.1 follows.

To conclude the proof of the main theorem, note that if T is a primitive word, it is not an exact power (of $N^{-1}M_{r,s}N$) unless $u = \pm 1$. If $u = +1$ the Main Theorem follows from the automorphism properties of $\boldsymbol{\Gamma}_R$. If $u = -1$, we have only to apply Lemma 6.1 to obtain exponent $u = +1$ by an inner automorphism ($\in \boldsymbol{\Gamma}_I^+ \subset \boldsymbol{\Gamma}_R$). The "uniqueness" was noted in § 4 above.

Now Corollary 2.2 to the main theorem follows from Lemma 4.2. If $w_0(A, B)$ is a positive word in both B and AB^2, then the lemma shows none of the twelve words $w_0(H_{t,u,v}(A, B))$ except $H_{t,u,v} = I$ can be made to have only positive powers of B and AB^2 through inner automorphisms of $\boldsymbol{F}_2$ (in $\boldsymbol{\Gamma}_I^+$). Now by the uniqueness of the period of the Markoff form, $w_0(A, B)$ must be equivalent (in $\boldsymbol{\Gamma}_I^+$) to $M_{r,s}(A, B)$.

There is a "computational" test in Lemma 7.1.

Lemma 7.2. *To test whether or not the word $w(A, B)$ in $\boldsymbol{F}_2$ is a power of a primitive word, test to see if the matrix $T = w(A_0, B_0)$ is a Markoff matrix, (presumably by the continued fraction for its fixed-points).*

Remark. Since it takes only a "small" subgroup of $\boldsymbol{\Gamma}^+$ (e.g., $\boldsymbol{\Gamma}_F^+$) to generate all classes of Markoff forms in Theorem 5.1, we might wonder if the smaller group $\boldsymbol{\Gamma}_R^+$ could replace $\boldsymbol{\Gamma}_R$ in the main theorem. This involves asking if N could be restricted to $\boldsymbol{SL}_2(\boldsymbol{Z})$ in Lemma 7.1.

This is possible only for the first two Markoff forms Φ_1 and Φ_2 of periods [1, 1] and [2, 2] respectively, (see § 1), which naturally are the exceptional cases of the main theorem and Lemma 4.1. Since the roots permit the improper transformations $z \mapsto n + 1/z$, $(n = 1, 2)$, any numbers equivalent to them in $\boldsymbol{PGL}_2(\boldsymbol{Z})$ are equivalent in $\boldsymbol{PSL}_2(\boldsymbol{Z})$. For Φ_a, $(a > 2)$, this situation does not prevail because the period is "genuinely" even. (Compare the illustrations in § 2.)

8. Partitions of Primitive Words on Markoff Periods

We consider analogues of theorems of Nielsen [11] (also see Whitehead [16] and Weinbaum [15]), on the partitioning[6] of primitive words into the product of basis elements.

The construction of $M_{r,s}(A, B)$ (in § 2 above) lends itself to the identity

$$M_{r,s}M_{t,u} = M_{r+s,t+u}$$

where

$$r/s > (r+t)/(s+u) > t/u, \qquad ru - st = 1 .$$

[6] These results were called to the author's attention by W. Magnus.

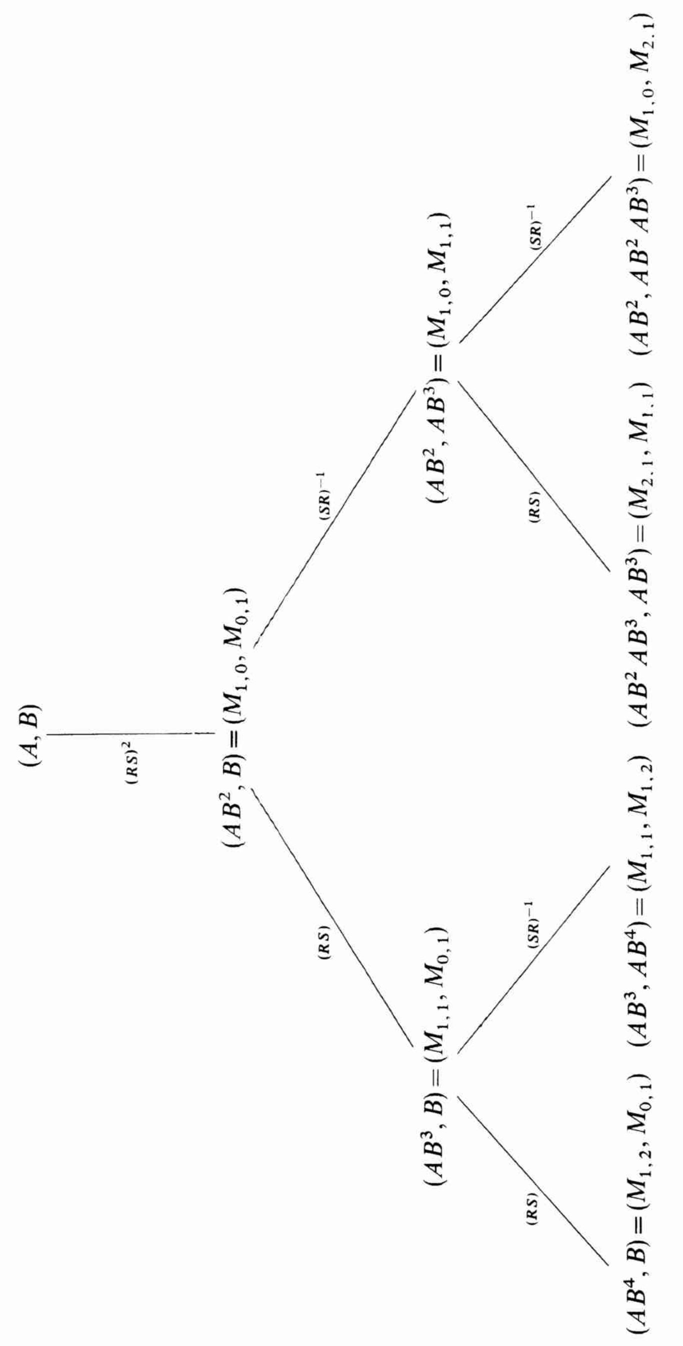

Fig. 1. Generation of normalized words by Γ_N^+

(The definition here requires that $r \geqq 0$, $s \geqq 0$, $t \geqq 0$, $u \geqq 0$.) This identity is essentially Smith's "geometric interpretation of the continued fraction", (compare [3]).

Thus the totality of normalized words $M_{r,s}(A, B)$ can be constructed by the following operations of Γ_N^+ (see § 3 above).

$$(RS)^2(A, B) = (AB^2, B) = (M_{1,0}, M_{0,1})$$
$$RS(M_{r,s}, M_{t,u}) = (M_{r,s}M_{t,u}, M_{t,u})$$
$$(SR)^{-1}(M_{r,s}, M_{t,u}) = (M_{r,s}, M_{r,s}M_{t,u}).$$

Indeed by the geometric method mentioned above, we can find some $H_N \in \Gamma_N^+$, given any $M_{r,s}$, $M_{t,u}$, such that

$$H_N(A, B) = (M_{r,s}(A, B), M_{t,u}(A, B)).$$

Theorem 8.1. *Let (a, b, c) be a Markoff triple. Then non-negative integers r, s, t, u, $(ru - st = 1)$, can be found so that the traces yield the values:*

$$3a = SM_{r,s}(A_0, B_0), \quad 3b = SM_{t,u}(A_0, B_0), \quad 3c = SM_{r+t,s+u}(A_0, B_0).$$

This triple can be constructed by using the Γ_N^+ group in the manner of Theorem 2.5, (indeed by using the semi-group generated by RS and $(SR)^{-1}$). In terms of Markoff triples,

$$RS(a, b, c) = (c, b, a'), \quad (SR)^{-1}(a, b, c) = (a, c, b'),$$

(with $a' = 3bc - a$, $b' = 3ac - b$, as usual).

This construction is illustrated in Fig. 1. It is analogous to the construction of all relatively prime non-negative pairs (r, s) by the euclidean algorithm. Markoff numbers for $M_{r,s}$ are as follows, (for $a < 100$):

$(r, s) =$	(0, 1)	(1, 0)	(1, 1)	(1, 2)	(2, 1)	(1, 3)	(1, 4)
$a =$	1	2	5	13	29	34	89

Remark. From Theorem 8.1, and the uniqueness of the class of the Markoff form for a given Markoff number, the continued fraction rules (see $M_{r,s}^*$ of § 5 above) automatically follow independently of [3].

9. Remarks on the Perforated Torus and its Geodesics

Looking back, we see that the "minimal property" of Markoff forms was not used directly. We used only the fact that the Markoff matrices were "self-perpetuating" and identifiable by the (invariant) trace with the corresponding form, (see Theorem 5.2 above).

To use the minimal property, we repeat the construction of [3]. For each Markoff form Φ we draw a "Poincaré-Klein" geodesic semicircle in the upper half z-plane, $(y > 0)$, orthogonal to the (real) x-axis where it connects the roots ξ, η of the Markoff form Φ. Then under the equivalence class of the form Φ, the geodesics will all lie below $y=3/2$. Indeed the maximum y, (= radius = $(\xi-\eta)/2$), is exactly $(2\mu(\Phi))^{-1}$, in the terminology of § 1. Only the class of a Markoff form would have the property that all geodesics lie (a positive amount) below $y=3/2$.

Now $\boldsymbol{F}_2=\{A, B\}$ is the commutator group of $\boldsymbol{G}_M=\boldsymbol{PSL}_2(\boldsymbol{Z})$. As such, its fundamental domain is six replicas of the standard fundamental domain (cusp at ∞) of $\boldsymbol{G}_M$, (see Maass [8]), adjacent under translation $z \mapsto z+n$. The fundamental domain is mapped onto $\boldsymbol{U}$, a regular hexagon on the u-plane, for

$$du = dJ/(J^{2/3}(1-J)^{1/2})$$

and $J(z)$ the Klein invariant of $\boldsymbol{G}_M$, $(J(i)=1,\ J(\exp 2\pi i/3)=0,\ J(i\infty)=+\infty)$. We endow $\boldsymbol{U}$ with the metric $ds=|dz|/y$ as transferred from the z-plane.

This hexagon $\boldsymbol{U}$ is a torus under the identifications A_0, B_0, C_0 of opposite sides. If we excise the closed disk

$$y \geqq 3/2, \quad (\text{inc.}\, z=i\infty),$$

which surrounds $u=0$, $(J(z)=\infty)$, we get the perforated torus $\boldsymbol{U}^* \subset \boldsymbol{U}$; (actually, $\boldsymbol{U}$ has an "almost circular" disk cut out). Then $\boldsymbol{F}_2$ is the fundamental group of $\boldsymbol{U}^*$. It was shown in [3] that with a well-selected homotopy basis of $\boldsymbol{U}$, essentially $V_1=B_0^{-1}$, $V_2=B_0^{-1}A_0^{-1}B_0^{-1}$ (of § 5), the closed geodesics are precisely in the homotopy classes of the words $M^*_{r,s}$.

We use the same V_1, V_2 as generators of the (one-dimensional) homology group of $\boldsymbol{U}$ (not $\boldsymbol{U}^*$). Then each closed geodesic of $\boldsymbol{U}^*$ represents a primitive homology class of $\boldsymbol{U}$. Conversely, however, (by Lemma 4.1), each primitive homology class of $\boldsymbol{U}$ is represented by a unique closed $M^*_{r,s}$ geodesic of $\boldsymbol{U}^*$ under $\boldsymbol{\Gamma}_R$. Part of $\boldsymbol{\Gamma}_R$ is $\boldsymbol{\Gamma}_I^+$, the inner automorphisms of $\boldsymbol{F}_2$, which constitutes the change of homotopy basis; but the other factor represents $\boldsymbol{\Gamma}_R/\boldsymbol{\Gamma}_I^+=\boldsymbol{G}_{12}$ (in the terminology of § 4), realized by the diedral group of the hexagon.

Furthermore, in the spirit of § 8 (on partitions), every closed geodesic other than type V_1 or V_2 is homotopically equivalent to the sum of two geodesics which form a basis pair.

References

1. Cassels, J. W. S.: The Markoff chain. Ann. of Math. **50**, 676—685 (1949).
2. Cohn, H.: Approach to Markoff's minimal forms through modular functions. Ann. of Math. **61**, 1—12 (1955).
3. — Representation of Markoff's binary quadratic forms by geodesics on a perforated torus. Acta Arith. **18**, 125—136 (1971).

4. Dickson, L. E.: Studies in the theory of numbers, p. 79—107. University of Chicago, press: 1930.
5. Fricke, R.: Über die Theorie der automorphen Modulgruppen. Nach. Akad. Wiss. Göttingen 91—101 (1896).
6. Frobenius, G.: Über die Markoffschen Zahlen. Preuss. Akad. Wiss. Sitzungsber. 458—487 (1913).
7. Hirzebruch, F.: The signature theorem: Reminiscences and recreation, (manuscript, 1970).
8. Maass, H.: Lectures on modular functions of one complex variable. Tata Institute, (Bombay), 70 (1964).
9. Magnus, W., Karass, A., Solitar, D.: Combinatorial group theory. New York: Wiley 1966.
10. Markoff, A.: Sur les formes binaires indéfinies. Math. Ann. **15**, 381—409 (1879).
11. Nielsen, J.: Die Isomorphismen der allgemeinen unendlichen Gruppe mit zwei Erzeugenden. Math. Ann. **78**, 385—397 (1918).
12. — Die Isomorphismengruppe der freien Gruppen. Math. Ann. **91**, 169—209 (1924).
13. Neumann, B. H.: Die Automorphismengruppe der freien Gruppen. Math. Ann. **107**, 367—386 (1933).
14. Rademacher, H.: Lectures on analytic number theory. Tata Institute, (Bombay) 273 (1954—1955).
15. Weinbaum, C. M.: Partitioning a primitive word into a generating set. Math. Ann. **181**, 157—162 (1969).
16. Whitehead, J. H. C.: On certain sets of elements in a free group. Proc. London Math. Soc. **41** (2), 48—56 (1936).

Harvey Cohn
Institute for Advanced Study
Princeton, N.J. 08540 (USA)

present address:

Dept. of Mathematics
College of City of New York
New York 10031, USA

(Received May 6, 1971)