# THE MARKOFF SPECTRUM AND GEODESICS ON THE PUNCTURED TORUS

DAVID J. CRISP

Thesis submitted for the degree of
Doctor of Philosophy
in
The University of Adelaide
(Department of Pure Mathematics)

DECEMBER 1993

## CONTENTS

## ABSTRACT

This thesis is a contribution to the area of diophantine approximation concerning the Markoff spectrum. It is based on the connection Harvey Cohn discovered between the structure of the Markoff spectrum and the behaviour of geodesics on a hyperbolic punctured torus. The torus $\mathbf{T}$ involved is the quotient of the upper half-plane $\mathbf{H}$ by the commutator subgroup $\Gamma'$ of the modular group. The connection is made by associating a form $f$ with the geodesic $\gamma$ in $\mathbf{H}$ whose endpoints are the roots of $f$ and then projecting $\gamma$ to $\mathbf{T}$. Cohn, [8], found that under this map the Markoff forms, that is, the forms whose Markoff values lie below 3, correspond exactly to the simple closed geodesics. Here, we study the Markoff values arising from geodesics with low self-intersection number.

Closed geodesics can be studied via their free homotopy classes. In this manner, we show that the closed geodesics with one self-intersection fall into two classes; the proper closed 1-intersectors and the improper closed 1-intersectors. The Markoff values of the improper closed 1-intersectors lie in Hall's ray and are not considered further. For open geodesics we use cutting sequences. It is known that the geodesics which correspond to forms with Markoff value 3 have aperiodic linear cutting sequences and therefore are simple and open. We show that the only other simple open geodesics have half-linear cutting sequences. This is an improvement on Haas' topological characterisation, [21], because it provides a means of calculating the associated Markoff values. We show that they all lie in Hall's ray.

By developing an understanding of how the subgroup of Aut $\Gamma'$ which fixes Markoff values lies in Aut $\Gamma'$ we are able to convert our characterisation of the proper closed 1-intersectors into a characterisation of the associated forms. This leads to an expression for the forms in terms of the solutions to Markoff's equation.

We demonstrate an intriguing symmetry between the Markoff values of the proper closed 1-intersectors and those of the simple closed geodesics. An examination of the spectrum near the first few values of the proper closed 1-intersectors reveals that they are isolated points of the spectrum. We conjecture that they are all isolated. Whilst we cannot prove our conjecture we do obtain some useful partial results. The techniques introduced for this purpose also allow us to describe two new families of isolated points.

## STATEMENT OF ORIGINALITY

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

SIGNED: DATE: 17/12/93

## ACKNOWLEDGEMENTS

The research for this thesis was begun at the University of Adelaide in 1988 under the supervision of Professor William Moran. In 1991 Professor Moran transferred to the Flinders University of South Australia and Associate Professor Charles Pearce became my internal supervisor. Professor Moran agreed to be my external supervisor and continued his role as my chief mentor.

That the connection between the Markoff spectrum and geodesics on the punctured torus would make an interesting and fruitful topic for research was originally suggested by Professor Moran. It was his conjecture that the Markoff values associated with the closed geodesics on $\mathbf{T}$ with one self-intersection are isolated in the Markoff spectrum.

I am grateful to Professor Moran for his guidance throughout my studies. I also thank the Flinders University for resources made available to me.

This thesis was typeset by $\mathcal{AMS}$-TEX, the TEX macro system of the American Mathematical Society.

CHAPTER 1

# INTRODUCTION AND PRELIMINARIES

This thesis is a contribution to the area of diophantine approximation concerning the Markoff spectrum. The Markoff spectrum is the set of values which arise when one considers the appropriately normalised minima of real indefinite binary quadratic forms. While it is *in toto* a complicated subset of the interval $[\sqrt{5}, \infty]$, the portion of it lying below 3 has a simple structure. It is a discrete set with 3 as its only limit point. The specific values in this portion and the associated forms, called Markoff forms, may be calculated using Markoff's theory, [29] and [30]. The Markoff spectrum also contains a maximal interval of the form $[\nu, \infty)$. It is referred to as Hall's ray.

Our work is based on the surprising connection Harvey Cohn discovered between the properties of the Markoff spectrum and the behaviour of geodesics on a hyperbolic once punctured torus. The torus $\mathbf{T}$ involved is the quotient of the upper half-plane $\mathbf{H}$ by the commutator subgroup $\Gamma'$ of the modular group. The connection is made by associating a form $f$ with the geodesic $\gamma$ in $\mathbf{H}$ whose endpoints are the roots of $f$ and then projecting $\gamma$ to $\mathbf{T}$. Cohn, [8], found that under this map the Markoff forms correspond precisely to the simple closed geodesics. Haas, [21], has extended this result by showing that geodesics which correspond to forms with Markoff value equal to 3 are simple and open. In this thesis we examine the Markoff values which arise from geodesics on $\mathbf{T}$ with low self-intersection numbers. Some of our results shed light on the poorly understood part of the Markoff spectrum between 3 and $\nu$. A review of these results has already been published, [12]. Our approach also provides insight into Harvey Cohn's work.

The topology of closed geodesics can be studied via their free homotopy classes. Free homotopy classes on $\mathbf{T}$ correspond to conjugacy classes in the fundamental

group $\pi_1(\mathbf{T})$. The fundamental group of the punctured torus is well-understood, it is course the free group of rank two. Open geodesics are more awkward. For them we use cutting sequences. Cutting sequences were first discussed in this context by Series, [36]. We provide the relevant background material on cutting sequences in Chapter 2. Although, Chapter 2 is mainly a review of established notions and known facts, we do present some new results there. In particular, we introduce *half-linear* cutting sequences and establish some of their properties.

In Chapter 3, we study geodesics on $\mathbf{T}$ with low self-intersection numbers. Specifically, we characterise the closed geodesics with one self-intersection and the simple open geodesics which do not correspond to forms with Markoff value equal to 3. It is convenient to call the closed geodesics with one self-intersection *closed 1-intersectors*. We show that there are two types of closed 1-intersector. One type includes a loop which bounds a disc containing the puncture, the other type does not. We refer to geodesics of the latter type as *proper closed 1-intersectors* and to the others as *improper closed 1-intersectors*. The presence of this loop around the puncture forces the Markoff values of the improper closed 1-intersectors to lie in Hall's ray.

We continue Chapter 3 by studying the simple open geodesics on $\mathbf{T}$. They can be characterised in terms of their cutting sequences. It is already known, [21] and [36], that the simple open geodesics which correspond to forms with Markoff value equal to 3 have linear cutting sequences. We show that the remaining simple open geodesics have half-linear cutting sequences. This is an improvement on Haas' topological characterisation, [21], since it provides the means of calculating the associated Markoff values. We show that those values lie in Hall's ray.

Closed geodesics are naturally associated with conjugacy classes in $\Gamma'$. Hence we can assign Markoff values to the conjugacy classes in $\Gamma'$. We denote the subgroup of Aut $\Gamma'$ which fixes the Markoff values of such classes by $\Psi$. In order to convert our characterisation of the proper closed 1-intersectors into a characterisation of the associated classes of forms we need to develop an understanding of how $\Psi$ lies in Aut $\Gamma'$. We do this in Chapter 4. Our main result is a description of a right

transversal for $\Psi$ in $\Gamma'$.

In Chapter 5 we show how the Markoff values of the proper closed 1-intersectors may be calculated. We achieve this by describing both the associated doubly infinite sequences of positive integers and representatives of the corresponding classes of forms. We do not include the improper closed 1-intersectors because, as we have mentioned, their Markoff values lie in Hall's ray. Our description of the classes of forms is given in terms of the solutions to Markoff's equation. We find that corresponding to each Markoff number $m$ there is exactly one proper closed 1-intersector. It has Markoff value

$$\sqrt{9+\frac{4}{m^2}}.$$

Of course, the Markoff forms can also be expressed in terms of the solutions to Markoff's equation. As is well-known, the Markoff value of the Markoff form corresponding to $m$ is

$$\sqrt{9-\frac{4}{m^2}}.$$

On the basis of this intriguing symmetry between the two sets of values we make the following conjecture.

**Conjecture.** *The Markoff values of the proper closed 1-intersectors on* $\mathbf{T}$ *are isolated points of the Markoff spectrum.*

Our primary aim in Chapter 6 is to provide evidence for our conjecture. However, having established the means of doing so, we are also be able to prove the existence of two new families of values which are isolated in the spectrum. Our calculations are based on the description of the spectrum in terms of doubly infinite sequences of positive integers. While we cannot prove in general that the Markoff value of a proper closed 1-intersector is isolated we can, in effect, describe a large class of integer sequences whose Markoff values are bounded away from the given one. By estimating the possible Markoff values of the remaining integer sequences we verify that the first few proper closed 1-intersectors do indeed have isolated Markoff values.

Markoff's original work, [29] and [30], has been refined by several authors. In the

appendix we provide a further improvement to part of that theory. As a natural extension of the results discussed there, we also present a new characterisation of the linear cutting sequences.

In the remainder of this chapter we define the entities we shall be dealing with and introduce the notation we shall be using. In the process of doing this we shall also briefly review some of the known facts about the Markoff spectrum and especially the connection between its properties and the behaviour of geodesics on the hyperbolic once punctured torus.

### The Markoff spectrum

For the real indefinite binary quadratic form

$$f(x,y) = \alpha x^2 + \beta xy + \gamma y^2, \qquad \alpha, \beta, \gamma \in \mathbf{R}$$

with discriminant

$$d(f) = \beta^2 - 4\alpha\gamma > 0$$

define

$$m(f) = \inf\{|f(x,y)| \ : \ (x,y) \in \mathbf{Z} \times \mathbf{Z}, (x,y) \neq (0,0)\}$$

and

$$M(f) = \sqrt{d(f)}/m(f) \tag{1.1}$$

with the convention that $M(f) = \infty$ when $m(f) = 0$. The quantity $M(f)$ is called the *Markoff value* of the form $f$. The set of Markoff values taken over all possible real indefinite binary quadratic forms is called the *Markoff spectrum*. (Sometimes this term refers to the set of reciprocals of $M(f)$.) As mentioned earlier, the forms whose Markoff values lie below 3 are called *Markoff forms* and they may be calculated using Markoff's theory, [29] and [30]. For other treatments see [5],[14],[17] and [26]. We remind the reader that the Markoff spectrum contains a maximal interval of the form $[\nu, \infty]$. The value of $\nu$ is less than $\sqrt{21} = 4.582\ldots$, see [14].

For the remainder of this thesis by *form* we shall mean real indefinite binary quadratic form. The set of all forms can be partitioned into equivalence classes. We require a more general definition of equivalence of forms than is usual. Given two forms $f$ and $g$, we say that $g$ is *equivalent* to $f$ if there is some $\mu \in \mathbf{R}$ and some $T = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL(2, \mathbf{Z})$ such that

$$g(x, y) = \mu f\left(T\begin{pmatrix} x \\ y \end{pmatrix}\right) = \mu f(ax + by, cx + dy)$$

and

$$\mu \det(T) > 0.$$

This defines an equivalence relation for forms. The reason forms are partitioned into equivalence classes is that equivalent forms have equal Markoff values. Thus we can refer unambiguously to the Markoff value of an equivalence class of forms. While it is also evident that $M(f) = M(-f)$, it is convenient to treat their equivalence classes separately.

A form $f$ can also be expressed in terms of its roots. The *roots* of $f$ are the zeros of $f(x, 1) = 0$ with one of the roots being taken as $\infty$ when the leading coefficient of $f(x, 1)$ is zero. Forms which represent zero, that is, which take the value zero at some integer point $(x, y) \in \mathbf{Z} \times \mathbf{Z}$ other than $(0, 0)$, are exactly those with at least one rational root. They have Markoff value $\infty$ and are of no interest here. We exclude them from consideration. Forms which do not represent zero then, are exactly those which can be written as

$$f = \mu(x - \eta y)(x - \xi y)$$

where $\mu \neq 0$ and $\eta$ and $\xi$ are distinct irrationals. The roots $\eta$ and $\xi$ of $f$ may be ordered. We order them so that

$$\mu(\xi - \eta) > 0.$$

With this ordering, we call $\eta$ the *first root* of $f$ and $\xi$ the *second*. Note that Dickson, [16] and [17], uses the reverse of this ordering.

### Markoff forms and the Markoff equation

Markoff's theory, [29] and [30], led to a formulae for representatives of the equivalence classes of Markoff forms and their associated Markoff values in terms of the solutions $(m, m_1, m_2)$ in positive integers to the diophantine equation

$$m^2 + m_1^2 + m_2^2 = 3mm_1m_2. \tag{1.2}$$

We refer to this equation as *Markoff's equation.* A good exposition of its solutions may be found in §3 of Chapter II of Cassels book [5]. We take our notation from there. Thus we order the triple $(m, m_1, m_2)$ so that

$$m \geq \max(m_1, m_2).$$

The solutions $(1,1,1)$ and $(2,1,1)$ are the only ones in which $m$, $m_1$ and $m_2$ are not distinct and for that reason are called *singular.* For each non-singular solution $(m, m_1, m_2)$, there is exactly one integer $k$ and one ordering of $m_1$ and $m_2$ such that

$$m_1k \equiv m_2 \pmod{m}, \qquad 0 < k < m/2.$$

Further, there is an integer $l$ satisfying

$$k^2 + 1 = lm.$$

This last condition is also true for the singular solutions $(1,1,1)$ and $(2,1,1)$ if we choose $k = 0$ and $k = 1$, respectively. Markoff showed that the form $f_m$ defined by

$$f_m(x, y) = mx^2 + (3m - 2k)xy + (l - 3k)y^2$$

is a Markoff form and that it has Markoff value

$$M(f_m) = \sqrt{9 - \frac{4}{m^2}}. \tag{1.3}$$

He also showed that any form $f$ with $M(f) < 3$ is equivalent to some $f_m$ and thereby characterised the Markoff forms. While the notation $f_m$ is ambiguous in that it is possible there are two distinct solutions to the equation with the same

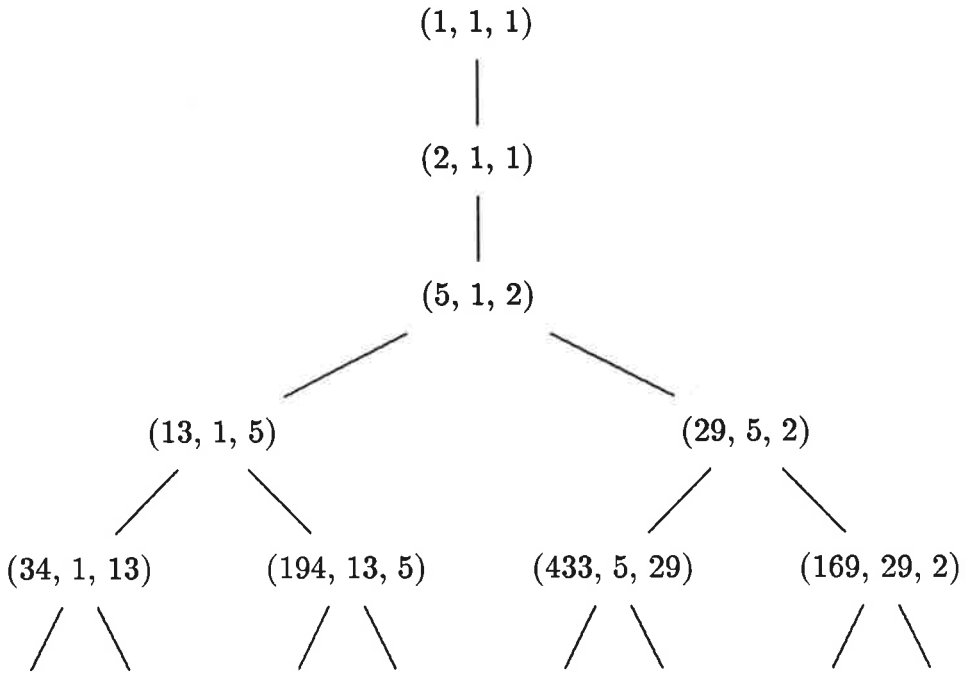


FIGURE 1.1. The set of solutions $(m, m_1, m_2)$ in positive integers to Markoff's equation $m^2 + m_1^2 + m_2^2 = mm_1m_2$ arranged as a tree.

maximum member $m$ there will be no problem with its use within the context of this monograph. In connection with this we mention that the conjecture that the solution $(m, m_1, m_2)$ is completely determined by $m$ is well-known and is referred to as the conjecture on the uniqueness of Markoff numbers.

The complete set of solutions to the equation is usually presented in the form of a tree. See Figure 1.1. At the top are the singular solutions $(1, 1, 1)$ and $(2, 1, 1)$. Immediately below these is the solution $(5, 1, 2)$. The tree is continued from this node and each of its successors by branching to both the left and the right. At the typical node $(m, m_1, m_2)$ one forms

$$(m_2', m_1, m), \qquad m_2' = 3mm_1 - m_2 \tag{1.4}$$

to branch left and

$$(m_1', m, m_2), \qquad m_1' = 3mm_2 - m_1 \tag{1.5}$$

to branch right. It is not hard to verify that these new triples are indeed solutions to the given equation. Conversely, it is possible to show that by applying this process

in reverse to any given non-singular solution one eventually arrives at the solution $(5,1,2)$. It follows that the tree contains every solution (in positive integers) to Markoff's equation. The same type of argument also shows that each solution occurs exactly once in the tree.

### Doubly infinite sequences of positive integers

To calculate the Markoff value of a form we shall also make use of the associated doubly infinite sequence of positive integers. Before we describe the association we need some definitions. As usual, $[x_0, x_1, x_2, \dots]$ denotes the simple continued fraction with partial quotients $x_0, x_1, x_2, \dots$. For a doubly infinite sequence of positive integers $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ we define

$$\eta_i(\mathcal{A}) = -[0, a_{i-1}, a_{i-2}, a_{i-3}, \dots], \qquad \xi_i(\mathcal{A}) = [a_i, a_{i+1}, a_{i+2}, \dots] \tag{1.6}$$

and

$$\lambda_i(\mathcal{A}) = \xi_i(\mathcal{A}) - \eta_i(\mathcal{A}).$$

The Markoff value of $\mathcal{A}$ is the quantity

$$M(\mathcal{A}) = \sup_{i \in \mathbf{Z}} \lambda_i(\mathcal{A}). \tag{1.7}$$

It is well-known that the set of values taken by $M(\mathcal{A})$ as $\mathcal{A}$ runs through all possible doubly infinite sequences of positive integers is exactly the Markoff spectrum.

We also need a notion of equivalence for integer sequences. We say that a doubly infinite sequence of positive integers $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is *equivalent* to the sequence $\mathcal{A}$ if there is some $k \in \mathbf{Z}$ such that $b_i = a_{i+k}$ for all $i \in \mathbf{Z}$. This defines an equivalence relation and clearly $M(\mathcal{B}) = M(\mathcal{A})$. An equivalence class for this relation should be thought of as an un-indexed doubly infinite sequence of positive integers. In general, we shall not differentiate between a particular integer sequence and its equivalence class.

The connection between the sequences of integers and the forms is given by the map

$$\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} \quad \longmapsto \quad f(x,y) = (x - \eta y)(x - \xi y) \tag{1.8}$$

where $\eta = \eta_0(\mathcal{A})$ is the first root of $f$ and $\xi = \xi_0(\mathcal{A})$ the second root. The significance of this map is that it preserves Markoff values, that is, $M(\mathcal{A}) = M(f)$. This non-trivial fact can be deduced from the exposition in Dickson's book, [16]. However, it is not hard to verify that

$$\{a_{i+1}\}_{i=-\infty}^{+\infty} \quad \longmapsto \quad \tfrac{1}{(a_0-\eta)(a_0-\xi)} f\left(\begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

and hence that equivalent sequences map to equivalent forms. We therefore have an induced map between equivalence classes of sequences and equivalence classes of forms not representing zero. Again, while not explicit, it can be deduced from Dickson's exposition [16] that the induced map is a bijection. We also note here, firstly that,

$$\overline{\mathcal{A}} = \{a_{-i}\}_{i=-\infty}^{+\infty} \quad \longmapsto \quad f\left(\begin{pmatrix} -1 & a_0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right),$$

so that the class of sequences containing the reverse of $\mathcal{A}$ is mapped to the class of forms containing $-f$, and secondly that, the periodic sequences of integers map exactly to the classes of forms which contain a representative with integral coefficients. We shall discuss such forms again later in this section.

To be specific, the results of [16] referred to are contained in sections §61, §62 and §67 where it is shown that every *proper equivalence* (our definition of equivalence but with $\mu = 1$) class of forms contains exactly one *chain* of *reduced* forms. Also section §65 from which it can be deduced that such a chain $\{\Phi_i\}_{i=-\infty}^{+\infty}$ where $\Phi_i(x,y) = a_i x^2 + b_i xy + a_{i+1} y^2$ and the chain $\{\Phi_i'\}_{i=-\infty}^{+\infty}$ defined by $\Phi_i'(x,y) = -a_i x^2 + b_i xy - a_{i+1} y^2$, as a pair, uniquely determine and are uniquely determined by a sequence $\{g_i\}_{i=-\infty}^{+\infty}$ of positive integers and the discriminant $d > 0$ of the chains. The correspondence is described by the fact that one of $\{\Phi_i\}_{i=-\infty}^{+\infty}$ or $\{\Phi_i'\}_{i=-\infty}^{+\infty}$ (depending on the parity of the location of the index $i = 0$) is the sequence

$$\left\{\frac{\sqrt{d}}{f_i - s_i}(x - f_i y)(x - s_i y)\right\}_{i=-\infty}^{+\infty}$$

where the first root and second roots are

$$f_i = (-1)^i[0, g_i, g_{i+1}, g_{i+2}, \ldots] \qquad \text{and} \qquad s_i = -(-1)^i[g_{i-1}, g_{i-2}, g_{i-3}, \ldots],$$

respectively. (Here we are using Dickson's ordering of the roots.) While this correspondence is not exactly the map above, it is easily seen that it induces a bijection like the one above except that $f$ must be replaced with $-f$. Finally Theorem 86 states that if the proper equivalence class of a form $f$ contains the chain $\{\Phi_i\}_{i=-\infty}^{+\infty}$ then $m(f)$ is the infimum of the $|a_i| = \sqrt{d(f)}/|f_i - s_i|$, that is, $M(f) = \sup_{i\in\mathbf{Z}} |f_i - s_i| = M(\{g_i\}_{i=-\infty}^{+\infty})$.

### The punctured torus T

The usual action of $SL(2,\mathbf{Z})$ on the upper half plane $\mathbf{H}$ is given by the homomorphism from $SL(2,\mathbf{Z})$ to $\Gamma = PSL(2,\mathbf{Z})$ defined by

$$T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \longmapsto \quad T(z) = \frac{az+b}{cz+d}. \tag{1.9}$$

This homomorphism is onto and its kernel is $\{\pm I\}$ thus allowing the standard abuse of notation whereby the same symbol denotes the matrix and the linear fractional transformation. The generators $A$ and $B$ of the commutator subgroup $\Gamma'$ of $\Gamma$ are the transformations associated with the matrices

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

We remark that the group generated by the matrices is the commutator subgroup of $SL(2,\mathbf{Z})$; it does not contain $-I$, and so is isomorphic to $\Gamma'$. The hyperbolic once punctured torus $\mathbf{T}$ which interests us is the quotient of $\mathbf{H}$ by $\Gamma'$. The projection map is

$$\sigma : \quad \mathbf{H} \quad \longrightarrow \quad \mathbf{H}/\Gamma' = \mathbf{T}. \tag{1.10}$$

We know $\mathbf{T}$ is a once punctured torus because the signature of $\Gamma'$ is $(1;\infty)$. Thus $\Gamma'$ has no elliptic elements and only one conjugacy class of maximal parabolic cyclic subgroups. Note in particular that every parabolic element of $\Gamma'$ is a conjugate of

$$ABA^{-1}B^{-1}(z) = z + 6$$

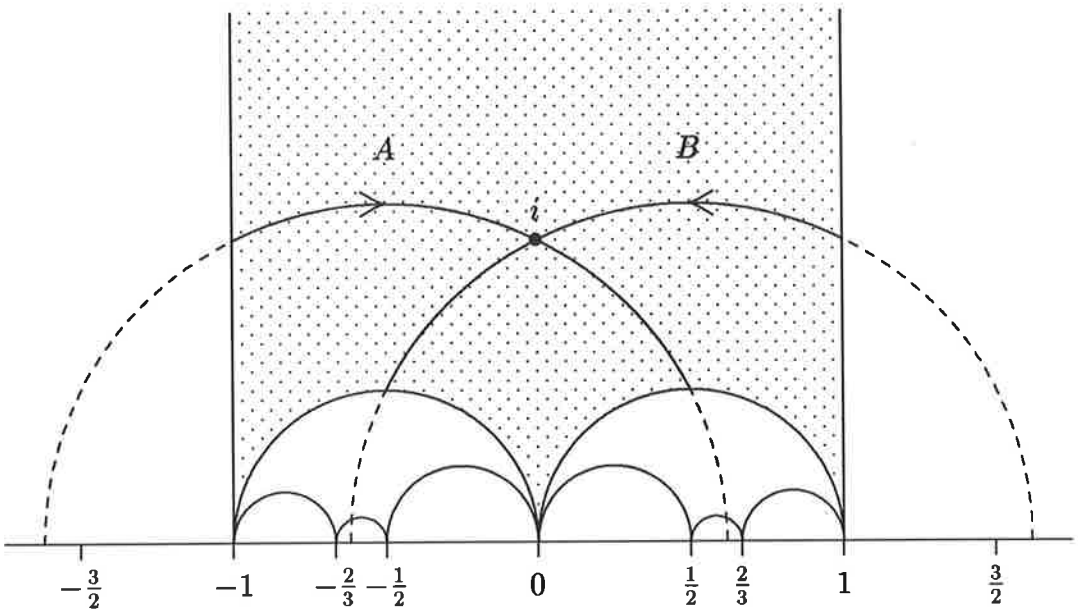


FIGURE 1.2. The fundamental domain $\mathcal{D}$ for $\Gamma'$ and its images under $A$ and $B$. Also shown are the axes (dashed) of $A$ and $B$ and fundamental segments (solid) thereof.

or one of its powers. A fundamental domain for $\Gamma'$ and its images under the transformations $A$ and $B$ are shown in Figure 1.2. The side pairings of the domain are indicated by the fundamental segments.

By a *geodesic* in $\mathbf{H}$ we mean a semi-circle (or half-line) which is orthogonal to the real axis. Each geodesic is determined up to orientation by its endpoints, that is, the points at which it meets the real axis. By ordering the endpoints we can also specify an *orientation*. We write

$$\gamma = [\eta, \xi], \qquad \eta \neq \xi$$

to mean that $\gamma$ is the geodesic with endpoints $\eta$ and $\xi$ and that $\gamma$ is directed from $\eta$ to $\xi$. For a transformation $T \in \Gamma'$ we define $T(\gamma)$ to be the oriented geodesic $T(\gamma) = [T(\eta), T(\xi)]$. Two oriented geodesics $\gamma_1$ and $\gamma_2$ are called $\Gamma'$-*equivalent* if there is some $T \in \Gamma'$ such that $\gamma_2 = T(\gamma_1)$.

It is natural to parameterise the points an oriented geodesic $\gamma$ in $\mathbf{H}$ using hyperbolic arc length from some fixed initial point. Thus we associate with $\gamma$ all the

open curves

$$c: \quad \mathbf{R} \quad \longrightarrow \quad \gamma$$

whose orientations agree with that of $\gamma$ and for which the hyperbolic distance between $c(x)$ and $c(y)$ is equal to the Euclidean distance between $x$ and $y$ for all $x, y \in \mathbf{R}$. In other words, a curve $c$ is associated with $\gamma$ if it corresponds to traversing $\gamma$ at unit speed. Note that, if $c$ is associated with $\gamma$ then every other open curve associated with $\gamma$ is of the form $c(x+t)$ where $t \in \mathbf{R}$. Further, the curve $c(-x)$ is associated with the geodesic $\gamma' = [\xi, \eta]$. We treat segments of geodesics similarly. Thus $[z_0, z_1]$, where $z_0 \neq z_1$ and $z_0, z_1 \in \mathbf{H}$, denotes the *oriented geodesic segment* which begins at $z_0$ and ends at $z_1$. Of course such a segment inherits its parameterisations as a curve from the oriented geodesic which contains it.

A *geodesic* $\gamma$ on $\mathbf{T}$ is the projection of a geodesic $\tilde{\gamma}$ in $\mathbf{H}$. It inherits its orientation and parameterisations from $\tilde{\gamma}$. That is, we associate with $\gamma$ any curve which is the projection of a curve associated with $\tilde{\gamma}$. Note that $\gamma$ and its associated curves are also the projection of any geodesic in $\mathbf{H}$ which is $\Gamma'$-equivalent to $\tilde{\gamma}$.

Since $\Gamma'$ contains no elliptic transformations the projection $\sigma : \mathbf{H} \to \mathbf{T}$ is a universal covering. It follows that, $\sigma$ together with a fixed lift $\tilde{p}_0 \in \mathbf{H}$ of the base point $p_0$ of the fundamental group $\pi_1(\mathbf{T})$ determine an isomorphism between $\Gamma'$ and $\pi_1(\mathbf{T})$. For $W \in \Gamma'$, the projection of all curves in $\mathbf{H}$ from $\tilde{p}_0$ to $W(\tilde{p}_0)$ is a homotopy class $w$ of loops on $\pi_1(\mathbf{T})$. The isomorphism is the map $W \mapsto w$. A point $\tilde{p}_1 \in \mathbf{H}$ is another lift of $p_0$ if and only if $\tilde{p}_1 = V(\tilde{p}_0)$ for some $V \in \Gamma'$. The isomorphism for $\tilde{p}_1$ is $VWV^{-1} \mapsto w$. It differs from the first only by an inner automorphism of $\Gamma'$ or equivalently of $\pi_1(\mathbf{T})$. Thus, if we do not insist on a fixed lift of $p_0$, the projection $\sigma$ determines a bijection between the conjugacy classes $[W]$ of $\Gamma'$ and the conjugacy classes $[w]$ of $\pi_1(\mathbf{T})$.

There is a natural choice for the base point $p_0$ of $\pi_1(\mathbf{T})$ and its lift $\tilde{p}_0$, namely, $\sigma(i)$ and $i$, respectively. We denote the isomorphism that this choice determines by $\theta$ and label the images of $A$ and $B$ by $a$ and $b$, respectively. Thus, $\theta$ is the isomorphism

$$\theta: \quad \Gamma' = F(A, B) \quad \longrightarrow \quad F(a, b) = \pi_1(\mathbf{T}) \tag{1.11}$$

defined by $\theta(A) = a$ and $\theta(B) = b$. A topological picture of loops representing the generators $a$ and $b$ of $\pi_1(\mathbf{T})$ may be obtained by identifying the sides of the fundamental domain $\mathcal{D}$ shown in Figure 1.2. The loops we refer to are the fundamental segments of $A$ and $B$, respectively, with their ends identified, see Figure 2.1.

There is also a natural bijection between the conjugacy classes $[w]$ of $\pi_1(\mathbf{T})$ and the free homotopy classes of loops on $\mathbf{T}$. As is well-known, for $w_1, w_2 \in \pi_1(\mathbf{T})$ with representative loops $l_1, l_2$ respectively; $w_1$ and $w_2$ are conjugate in $\pi_1(\mathbf{T})$ if and only if $l_1$ and $l_2$ are freely homotopic. The bijection is the map which identifies the conjugacy class of $w$ with the free homotopy class of the loops $l$ representing $w$. To see that this map is onto, note that, if $l$ is a loop on $\mathbf{T}$ and $c$ a curve joining the base point of $\pi_1(\mathbf{T})$ to $l$ then $l$ is freely homotopic to $clc^{-1}$ and such a loop represents some $w \in \pi_1(\mathbf{T})$. By combining this bijection with that mentioned above, we can identify the conjugacy classes of $\Gamma'$ with the free homotopy classes of loops on $\mathbf{T}$. Thus the free homotopy class corresponding to the conjugacy class $[W]$ is that which contains the loops representing $[w]$ where $w = \theta(W)$. It is not hard to verify that under this correspondence the projection of any curve in $\mathbf{H}$ from some $z$ to $W(z)$ is a loop $l$ in the free homotopy class $[w]$.

### Closed geodesics on T

Of particular interest are those geodesics on $\mathbf{T}$ which are the projection of the axes of the hyperbolic transformations in $\Gamma'$. The *axis* of $W \in \Gamma'$ is the unique geodesic $\tilde{\gamma}$ in $\mathbf{H}$ which is fixed by $W$. The effect of $W$ is to translate $\tilde{\gamma}$ along itself by a fixed hyperbolic distance. We orient $\tilde{\gamma}$ according to this effect. Specifically, *we direct $\tilde{\gamma}$ from the repulsive fixed point of $W$ to the attractive one.* Observe that $\tilde{\gamma}$ can be partitioned into a fundamental segment of the form $[z_0, W(z_0)]$ where $z_0 \in \tilde{\gamma}$ and its images under $W$. Recall that according to our notation the segment $[z_0, W(z_0)]$ is directed from $z_0$ to $W(z_0)$ and hence its orientation agrees with that of $\tilde{\gamma}$. Now let $\gamma$ be the projection of $\tilde{\gamma}$ to $\mathbf{T}$. Evidently, $\gamma$ is covered by and has the same orientation as the projection of the segment $[z_0, W(z_0)]$. Since $z_0$ and $W(z_0)$ project to the same point on $\mathbf{T}$, we can view $\gamma$ as a closed curve. We call $\gamma$ together

with the parameterisations it inherits from all the fundamental segments of $\tilde{\gamma}$ with respect to $W$ a *closed geodesic.* We emphasis that our definition is more specific than is usual in that we have associated with $\gamma$ a particular family of closed curves. For $T \in \Gamma'$, the geodesic $T(\tilde{\gamma})$ is the axis of the hyperbolic transformation $TWT^{-1}$ and therefore $TWT^{-1}$ defines the same closed geodesic as $W$. We say that $\gamma$ is the closed geodesic on $\mathbf{T}$ *defined by the conjugacy class* $[W]$ in $\Gamma'$.

We call an element of $\Gamma'$ *primitive* if it is not a non-trivial power of some other element of $\Gamma'$. Every conjugate of a primitive element is also primitive. Thus we can unambiguously refer to primitive conjugacy classes in $\Gamma'$. We call a closed geodesic on $\mathbf{T}$ *primitive* if it is defined by a primitive conjugacy class. We shall consider the situation where $\gamma$ is a non-primitive closed geodesic in some detail. Thus we let $[W]$ be the conjugacy class defining $\gamma$ and we suppose $W = V^n$ where $V \in \Gamma'$ and $n$ is not 0 or $\pm 1$. By replacing $V$ by its inverse if necessary, we may assume $n \geq 2$. It is not hard to verify that $V$ like $W$ is hyperbolic and that it has the same axis as $W$. Clearly each fundamental segment $[z_0, W(z_0)]$ of the axis of $W$ can be partitioned into $n$ fundamental segments for $V$, namely, $[V^{i-1}(z_0), V^i(z_0)]$ where $i = 1, 2, \ldots, n$. It follows that each closed curve associated with $\gamma$ is the $n$-th power of some closed curve associated with the closed geodesic defined by $[V]$. In particular, each closed curve associated with $\gamma$ has a continuum of self-intersections. Since we are only interested in geodesics with a low number of self-intersections we shall restrict our attention to primitive closed geodesics.

In the preceding section we described a bijection between the conjugacy classes of $\Gamma'$ and the free homotopy classes of loops on $\pi_1(\mathbf{T})$. In particular, given $W \in \Gamma'$, we saw that the projection of any curve in $\mathbf{H}$ from some $z$ to $W(z)$ is a loop in the free homotopy class $[w]$ where $w = \theta(W)$. It follows that if $\gamma$ is the closed geodesic defined by $[W]$ then the closed curves we have associated with it all lie in the free homotopy class $[w]$. Clearly the closed curves associated with distinct closed geodesics lie in distinct free homotopy classes. As is common, we shall say that $\gamma$ is the *unique closed geodesic in the free homotopy class* $[w]$. Not all free homotopy classes contain closed geodesics. Of course those which do not, correspond exactly

to the parabolic conjugacy classes in $\Gamma'$. We remind the reader that the latter are of the form $[(ABA^{-1}B^{-1})^n]$ where $n$ is a non-zero integer. We shall make much use of the well-known fact that *the number of self-intersections of a primitive closed geodesic is minimal amongst all the curves in its free homotopy class.*

### Self-intersection numbers of geodesics on T

We define the *number of self-intersections* of an open geodesic $\gamma$ on $\mathbf{T}$ to be the number of self-intersections of any of the open curves associated with it. By the number of self-intersections of an open curve

$$c: \quad \mathbf{R} \quad \longrightarrow \quad \gamma$$

we mean the number of unordered pairs $x, y$ such that $c(x) = c(y)$. To see that the self-intersection number of $\gamma$ is well-defined suppose $c : \mathbf{R} \to \gamma$ is a curve associated with $\gamma$. By definition, $c$ is the projection of a curve

$$\tilde{c}: \quad \mathbf{R} \quad \longrightarrow \quad \tilde{\gamma}$$

which is associated with a lift $\tilde{\gamma}$ of $\gamma$ to $\mathbf{H}$. Since $\tilde{c}(x) = \tilde{c}(y)$ if and only if $x = y$ we can deduce that the number of self-intersections of $c$ is the number of unordered pairs $z_1, z_2$ with $z_1, z_2 \in \tilde{\gamma}$ and $z_1 \neq z_2$ such that $z_1$ and $z_2$ project to the same point on $\mathbf{T}$. Obviously this number does not depend on the choice of $c$ or $\tilde{\gamma}$. It follows that the self-intersection number of $\gamma$ is well-defined.

We emphasis that our conventions are perhaps slightly unusual in that *we allow an open geodesic $\gamma$ on $\mathbf{T}$ to cover a closed geodesic on $\mathbf{T}$*. This happens precisely when one and hence all of the lifts of $\gamma$ to $\mathbf{H}$ are the axes of hyperbolic elements of $\Gamma'$. In this case it is clear that $\gamma$ has a continuum of self-intersections.

If $\gamma$ is an open geodesic on $\mathbf{T}$ which does not cover a closed geodesic then we can express the self-intersection number of $\gamma$ in terms of its lifts to $\mathbf{H}$. In particular, we claim that if $\tilde{\gamma}$ is any lift of $\gamma$ to $\mathbf{H}$ then the self-intersection number of $\gamma$ is half the number of geodesics in $\mathbf{H}$ which are $\Gamma'$-equivalent to $\tilde{\gamma}$ and intersect it. To see this, recall that the self-intersection number of $\gamma$ is the number of unordered pairs

$z_1, z_2$ with $z_1, z_2 \in \tilde{\gamma}$ and $z_1 \neq z_2$ such that $\sigma(z_1) = \sigma(z_2)$. For each such pair of points $z_1, z_2$ there is a unique non-trivial pair of inverse elements $T^{\pm 1}$ in $\Gamma'$ such that $z_2 = T(z_1)$ and $z_1 = T^{-1}(z_2)$. Since $\gamma$ does not cover a closed geodesic $T(\tilde{\gamma})$ and $T^{-1}(\tilde{\gamma})$ are not equal to $\tilde{\gamma}$ and hence $\{z_2\} = \tilde{\gamma} \cap T(\tilde{\gamma})$ and $\{z_1\} = \tilde{\gamma} \cap T^{-1}(\tilde{\gamma})$. Conversely, if $T^{\pm 1}$ is a pair of non-trivial elements of $\Gamma'$ such that the intersections of $\tilde{\gamma}$ with $T(\tilde{\gamma})$ and $T^{-1}(\tilde{\gamma})$ are non-empty then there is pair of points $z_1, z_2$ satisfying $\{z_2\} = \tilde{\gamma} \cap T(\tilde{\gamma})$ and $\{z_1\} = \tilde{\gamma} \cap T^{-1}(\tilde{\gamma})$. Hence $z_2 = T(z_1)$ and $z_1 = T^{-1}(z_2)$ and so $z_1 \neq z_2$ and $\sigma(z_1) = \sigma(z_2)$. Thus each self-intersection of $\gamma$ corresponds to a pair of inverse elements $T^{\pm 1}$ in $\Gamma'$ with $\tilde{\gamma} \cap T(\tilde{\gamma})$ and $\tilde{\gamma} \cap T^{-1}(\tilde{\gamma})$ non-empty. We conclude that the number of self-intersections of $\gamma$ is half the number of transformations $T \in \Gamma'$ such that $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$. The truth of our claim is now evident.

As with open geodesics, we define the *number of self-intersections* of a closed geodesic $\gamma$ on $\mathbf{T}$ to be the number of self-intersections of any of the closed curves associated with it. By the number of self-intersections of a closed curve

$$l: \quad [x_0, x_1] \subset \mathbf{R} \quad \longrightarrow \quad \gamma$$

we mean the number of unordered pairs $x, y$ with $x_0 \leq x, y < x_1$ such that $l(x) = l(y)$. To see that this number is the same for all closed curves associated with $\gamma$, let $[W]$ be the conjugacy class defining $\gamma$. By definition a closed curve $l$ associated with $\gamma$ is the projection of a suitably parameterised fundamental segment of the axis, say $\tilde{\gamma}$, of some transformation in $[W]$. Since the axes of all the members of $[W]$ are $\Gamma'$-equivalent we may assume $\tilde{\gamma}$ is the axis of $W$. Thus $l$ is the projection of a curve of the form

$$\tilde{l}: \quad [x_0, x_1] \subset \mathbf{R} \quad \longrightarrow \quad [z_0, W(z_0)] \subset \tilde{\gamma}$$

which parameterises $[z_0, W(z_0)]$ by hyperbolic arc length. Because $\tilde{l}(x) = \tilde{l}(y)$ if and only if $x = y$ we find that the number of self-intersections of $l$ is the number of unordered pairs $z_1, z_2$ with $z_1, z_2 \in [z_0, W(z_0)]$ and $z_1, z_2 \neq W(z_0)$ such that $z_1$ and $z_2$ project to the same point of $\mathbf{T}$. It is not hard to verify that this number depends only $[W]$ and not the choice of $z_0$.

We saw in the previous section that if $\gamma$ is a non-primitive closed geodesic then each closed curve associated with $\gamma$ is a non-trivial power of some other closed curve. It follows that non-primitive closed geodesics have a continuum of self-intersections. Note that the underlying reason for this is also the reason open geodesics on $\mathbf{T}$ which cover closed geodesics have a continuum of self-intersections. Of course, we can express the self-intersection number of a primitive closed geodesic $\gamma$ on $\mathbf{T}$ in terms of its lifts to $\mathbf{H}$. In particular, if $W$ is a representative of the conjugacy class defining $\gamma$ and if $\tilde{\gamma}$ is the axis of $W$ then the self-intersection number of $\gamma$ is half the number of geodesics in $\mathbf{H}$ which are $\Gamma'$-equivalent to $\tilde{\gamma}$ and intersect a fixed fundamental segment $[z_0, W(z_0)]$ of $\tilde{\gamma}$. (In this context the point $W(z_0)$ is not included in the set $[z_0, W(z_0)]$.) The verification of this is similar to that given above for the open geodesics. We shall not be using this result and we omit the details.

### The connection between the Markoff spectrum and T

We now elucidate the connection that Harvey Cohn found between the structure of the Markoff spectrum and the behaviour of geodesics on $\mathbf{T}$. Cohn's original work appears in a sequence of papers [6], [8], [9] and [10]. He has reviewed them and related matters in [11]. Haas, [20] and [21], and Series, [36] and [38], have reinterpreted and extended Cohn's work. The connection is made by relating forms to geodesics of the upper half-plane $\mathbf{H}$. As noted earlier, a geodesic in $\mathbf{H}$ as a point set is uniquely determined by its endpoints. Cohn associates with a form $f$ the geodesic $\gamma$ whose endpoints are the roots of $f$. We also orient $\gamma$ by directing it from the first root of $f$ to the second. We write this association as the map

$$f(x,y) = \mu(x - \eta y)(x - \xi y) \quad \longmapsto \quad \gamma = [\eta, \xi] \tag{1.12}$$

where $\mu(\xi - \eta) > 0$. Note that the geodesic associated with $-f$ covers the same point set as $\gamma$ but has the opposite orientation. Clearly forms map to the same geodesic if and only if they are positive multiples of one another. Also, if geodesics with a rational endpoint are excluded, the map is onto.

In order to best describe the relationship between geodesics which correspond to equivalent forms we extend the group $\Gamma$. We extend it to the group $\Gamma^*$ generated by $\Gamma$ and the transformation

$$T_1 : \quad z \quad \longmapsto \quad -\bar{z}.$$

Note that $T_1$ is reflection in the imaginary axis. The index of $\Gamma$ in $\Gamma^*$ is 2. By mapping the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ to the transformation $T_1$ the homomorphism from $SL(2,\mathbf{Z})$ to $\Gamma$ defined by (1.9) is extended to one from $GL(2,\mathbf{Z})$ to $\Gamma^*$.

It can now be deduced from the results in [11] or [20] that equivalent forms map to $\Gamma^*$-equivalent geodesics. (Here we have extended the notion of $\Gamma'$-equivalence for geodesics to $\Gamma^*$-equivalence in the obvious manner.) In fact for $\mu \in \mathbf{R}$ and $T \in GL(2,\mathbf{Z})$ with $\mu \det(T) > 0$ we have

$$\mu f\left(T\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)\right) \quad \longmapsto \quad T^{-1}(\gamma) \tag{1.13}$$

under the correspondence (1.12). Thus (1.12) induces a bijection between equivalence classes of forms and $\Gamma^*$-equivalence classes of oriented geodesics. We remark here that the Markoff value of a class of forms is exactly the supremum of the diameters of the geodesics (as semi-circles) in the associated $\Gamma^*$-equivalence class of geodesics on $\mathbf{H}$, see [11] and [20].

At this stage Cohn had the insight to consider the effect of projecting the geodesics to $\mathbf{T}$. He obtained the now well-known result that the simple closed geodesics on $\mathbf{T}$ correspond exactly to the Markoff forms.

The group $\Gamma'$ is a normal subgroup of $\Gamma^*$. Its index in $\Gamma^*$ is 12 and therefore there are twelve geodesics (not necessarily distinct) on $\mathbf{T}$ corresponding to each $\Gamma^*$-equivalence class of geodesics in $\mathbf{H}$. Hence there are twelve geodesics on $\mathbf{T}$ corresponding to each class of forms. The relationship between these geodesics is easy to discover. For $T \in \Gamma^*$, consider the map from $\mathbf{T}$ to itself defined by

$$z \quad \longmapsto \quad \sigma(T(\tilde{z})) \tag{1.14}$$

where $\tilde{z}$ is any lift of $z$ to $\mathbf{H}$. If $\tilde{z}'$ is another lift of $z$ to $\mathbf{H}$ then there is $W \in \Gamma'$ such that $\tilde{z}' = W(\tilde{z})$. Hence

$$\sigma(T(\tilde{z}')) = \sigma(TW(\tilde{z})) = \sigma(VT(\tilde{z})) = \sigma(T(\tilde{z})),$$

where $V \in \Gamma'$ satisfies $V = TWT^{-1}$. It follows that the map (1.14) is well-defined. It is one-to-one and onto since the map corresponding to $T^{-1}$ is its inverse. We know $T$ is an isometry of $\mathbf{H}$ and therefore (1.14) is an isometry of $\mathbf{T}$. It is not hard to verify that this correspondence between $\Gamma^*$ and the isometries of $\mathbf{T}$ is a homomorphism and that its kernel is $\Gamma'$. In other words, $\Gamma^*/\Gamma'$ is isomorphic to a group of isometries of $\mathbf{T}$. It is clear from the construction that the twelve geodesics on $\mathbf{T}$ corresponding to each class of forms are images of each other under this group. We remark that $\Gamma^*/\Gamma'$ accounts for all the isometries of $\mathbf{T}$.

A description of $\Gamma^*/\Gamma'$ may be found in [31]. Its generators are $U_1\,\Gamma'$ and $T_1\,\Gamma'$ where

$$U_1 : \quad z \quad \longmapsto \quad z+1$$

The relations follow from $U_1^6(z) = B^{-1}A^{-1}BA(z) \in \Gamma'$ and $T_1^2(z) = z$ and $U_1T_1(z) = T_1U_1^{-1}(z)$. Thus a set of coset representatives is

$$\{U_1^n, T_1U_1^n \ : \ n = 0, 1, \ldots, 5\}.$$

We emphasis that the transformations $U_1$ and $T_1$ leave the diameters of geodesics (as semi-circles) in $\mathbf{H}$ unchanged and therefore the Markoff value of a geodesic on $\mathbf{T}$ can be defined, in terms of $\Gamma'$ only, as the supremum of the diameters of all its lifts to $\mathbf{H}$.

### Forms mapping to closed geodesics on T

To be rigorous we shall say that *the form $f$ maps to the closed geodesic $\gamma$* if under the correspondence (1.12) the form $f$ maps to a geodesic in $\mathbf{H}$ which covers and has the same orientation as $\gamma$. Obviously the geodesics $\tilde{\gamma}$ in $\mathbf{H}$ with this property are exactly the axes (with the appropriate orientation) of the transformations in the conjugacy class $[W]$ defining $\gamma$. Forms which map to $\gamma$ can be obtained directly from the elements of $[W]$.

For $W = \pm \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma'$, we define

$$f_W(x,y) = \begin{cases} cx^2 + (d-a)xy - by^2 & \text{if } a+d > 0 \\ -cx^2 - (d-a)xy + by^2 & \text{if } a+d < 0. \end{cases} \tag{1.15}$$

Note that $f_W = f_{-W}$ and so there is no loss of generality here in identifying a transformation with the associated matrices. We claim that $f_W$ maps to the closed geodesic $\gamma$ defined by the conjugacy class $[W]$. To see this let $\tilde{\gamma}$ be the axis of $W$. It is not hard to verify that the roots of $f_W$ are the fixed points of $W$ and hence one of $f_W$ or $-f_W$ maps to $\tilde{\gamma}$ under the correspondence (1.12). Now recall that $\tilde{\gamma}$ is directed from the repulsive fixed point of $W$ to the attractive one. The dependency of $f_W$ on the sign of $a+d$ ensures that the ordering of its roots agrees with the ordering of the endpoints of $\tilde{\gamma}$ due to its orientation and hence $f_W$ maps to $\tilde{\gamma}$. The truth of our claim is now evident. We remark here that since $W$ is hyperbolic, $f_W$ does not represent zero. Also $f_{W^{-1}} = -f_W$ and therefore $f_{W^{-1}}$ maps to the inverse of $\gamma$, as expected.

For $T \in \Gamma'$, the transformation $TWT^{-1}$ lies in $[W]$ and so the form $f_{TWT^{-1}}$ also maps to $\gamma$. Suppose more generally that $T \in \Gamma^*$. Since $\Gamma'$ is normal in $\Gamma^*$ the transformation $TWT^{-1}$ belongs to $\Gamma'$. Moreover, $TWT^{-1}$ is hyperbolic and its axis is the image of $\tilde{\gamma}$ under $T$. It follows that the form $f_{TWT^{-1}}$ maps to the closed geodesic on $\mathbf{T}$ which is the image of $\gamma$ under the isometry of $\mathbf{T}$ induced by $T$. It also follows that $f_{TWT^{-1}}$ is equivalent to $f_W$. A straightforward calculation reveals that

$$f_{TWT^{-1}}(x,y) = \det(T) f_W\left(T^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right). \tag{1.16}$$

From this we can deduce that $W$ is an automorph of $f_W$, that is,

$$f_W\left(W\begin{pmatrix} x \\ y \end{pmatrix}\right) = f_W(x,y).$$

We have just seen that for each closed geodesic on $\mathbf{T}$ there is a form with integral coefficients and not representing zero which maps to it. The converse is also true, that is, each form $f$ with integral coefficients which does not represent zero maps to a closed geodesic on $\mathbf{T}$. To prove this, we use the fact, shown in sections §69 and §70 of Dickson's book [16], that such a form has a non-trivial automorph $\pm W \in \Gamma$. Again, since

$$f\left(W\begin{pmatrix} x \\ y \end{pmatrix}\right) = f\left(-W\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

there really is no loss of generality here in identifying a matrix with its negative and hence with the associated transformation of $\mathbf{H}$. Now let $\tilde{\gamma}$ be the geodesic in $\mathbf{H}$ corresponding to $f$. We are assuming $f\left(W\begin{pmatrix}x\\y\end{pmatrix}\right) = f(x,y)$ and so (1.13) implies $W^{-1}(\tilde{\gamma}) = \tilde{\gamma}$. Hence $W$ fixes $\tilde{\gamma}$. Because $f$ is indefinite, $W$ is not elliptic of order 2. It follows that $W$ is hyperbolic and by replacing $W$ by its inverse if necessary we may assume $\tilde{\gamma}$ is its axis. We know $\Gamma'$ is of finite index in $\Gamma$ and therefore there exists $n \geq 1$ such that $W^n \in \Gamma'$. Clearly $W^n$ defines a closed geodesic on $\mathbf{T}$ which is covered by the projection of $\tilde{\gamma}$ to $\mathbf{T}$. We conclude that $f$ maps to a closed geodesic on $\mathbf{T}$.

We are especially interested in those classes of forms which map to closed geodesics on $\mathbf{T}$ with low self-intersection numbers. For that reason we have restricted our attention, amongst the closed geodesics, to those which are primitive. We observe that as far as the associated classes of forms are concerned this is a natural restriction anyway. We have seen that if a closed geodesic $\gamma$ is not primitive then there is some integer $n > 1$ and a primitive closed geodesic $\delta$ such that each closed curve associated with $\gamma$ is the $n$-th power of some closed curve associated with $\delta$. In this case, we can choose representatives $W$ and $V$ of the conjugacy classes defining $\gamma$ and $\delta$, respectively, such that $W = V^n$. Obviously $W$ and $V$ share the same axis. It follows that $f_V$ is not only equivalent to $f_W$ but is a positive multiple of $f_W$.

### The group $\Psi$ of automorphisms of $\Gamma'$

An automorphism $G$ of $\Gamma'$ is described by its effect on the generators $A$ and $B$ of $\Gamma'$. Thus we write $G(A,B) = (G(A),\ G(B))$. The composition $GH$ of $G, H \in$ Aut $\Gamma'$ is defined by $GH(A,B) = (G(H(A)),\ G(H(B)))$. The group Aut $\Gamma'$ of all automorphisms of $\Gamma'$ has presentation, see Cohn [9],

$$\text{(1.17)}\quad \text{Aut } \Gamma' = \langle\ P,R,S \mid P^2, R^4, S^3, (RP)^2, (R^2SP)^2, SR^2S^2R = RS^2R^2S\ \rangle$$

where

$$\text{(1.18)}\quad P(A,B) = (B,\ A),\quad R(A,B) = (B^{-1},\ A),\quad S(A,B) = (B,\ B^{-1}A^{-1}).$$

Note that our convention for composition differs from that of Cohn.

Since $\Gamma'$ is normal in $\Gamma^*$ it is preserved by the inner automorphisms of $\Gamma^*$. Consequently, the map

$$T \longmapsto G_T(A,B) = (TAT^{-1}, TBT^{-1}) \tag{1.19}$$

is a homomorphism from $\Gamma^*$ to Aut $\Gamma'$. Because the axes of $TAT^{-1}$ and $TBT^{-1}$ are the images of those of $A$ and $B$, respectively, under $T$ we can deduce that $G_T$ is the identity automorphism only when $T$ is trivial. It follows that the map (1.19) is an isomorphism. We denote the image of $\Gamma^*$ by $\Psi$. Direct calculation shows that under (1.19) we have

$$\begin{aligned} T_1(z) = -\bar{z} &\longmapsto P(A,B) = (B,A) \\ U_1(z) = z+1 &\longmapsto SR^2(A,B) = (B^{-1}, AB) \\ U_2(z) = -1/z &\longmapsto R^2(A,B) = (A^{-1}, B^{-1}). \end{aligned}$$

The transformations $T_1$, $U_1$ and $U_2$ generate $\Gamma^*$ and thus the automorphisms $P$, $R^2$ and $S$ generate $\Psi$. Note for example that $A = U_2U_1^{-1}U_2U_1$ and so $G_A = S^2R^2SR^2$ and similarly $B = U_2U_1U_2U_1^{-1}$ and so $G_B = R^2SR^2S^2$. Of course, $G_A$ and $G_B$ generate Inn $\Gamma'$, the inner automorphisms of $\Gamma'$.

The significance of the automorphisms in $\Psi$ is that they preserve the classes of forms and hence Markoff values associated with the conjugacy classes in $\Gamma'$. To explain what we mean by this, let $[W]$ be a conjugacy class in $\Gamma'$. We know from (1.16) that the class of forms represented by $f_W$ does not depend on the particular representative $W$ of $[W]$. We say that the class of forms containing $f_W$ is *associated* with the conjugacy class $[W]$. Now let $G \in \Psi$. Obviously $G$ permutes the conjugacy classes of $\Gamma'$. We know $G = G_T$ for some $T \in \Gamma^*$. Therefore $G(W) = TWT^{-1}$ and (1.16) implies that the form $f_{G(W)}$ is equivalent to $f_W$. It follows that the same class of forms is associated with both $[W]$ and $[G(W)]$. In terms of geodesics, the closed geodesic defined by $[G(W)]$ is the image of that defined by $[W]$ under the isometry of $\mathbf{T}$ induced by $T$ and hence they give rise to the same class of forms. It will become apparent that $\Psi$ is the largest subgroup of Aut $\Gamma'$ with this property when we consider the effect of automorphisms on the conjugacy class $[AB]$.

As we have noted, the image of $\Gamma'$ under (1.19) is Inn $\Gamma'$. Hence

$$\Gamma^*/\Gamma' \quad \cong \quad \Psi/\text{Inn } \Gamma'.$$

The factor group $\Psi/\text{Inn } \Gamma'$ corresponds to the group $G_{12}$ in [9]. The isomorphism between $\Gamma^*/\Gamma'$ and $\Psi/\text{Inn } \Gamma'$ has an interpretation on $\mathbf{T}$. Each isometry of $\mathbf{T}$ permutes the free homotopy classes of loops on $\mathbf{T}$ and hence the corresponding conjugacy classes of $\pi_1(\mathbf{T})$. It follows that each isometry induces an outer isomorphism of $\pi_1(\mathbf{T})$. We have identified the isometries of $\mathbf{T}$ with the elements of $\Gamma^*/\Gamma'$ and hence we can also associate the cosets of $\Gamma'$ in $\Gamma^*$ with the outer automorphisms of $\pi_1(\mathbf{T})$. It should come as no surprise that under this correspondence the coset $T\ \Gamma'$ maps to the image of the outer isomorphism $G_T$ Inn $\Gamma'$ under the isomorphism $\theta$. We leave the verification to the reader.

### Cohn's commutator map

Cohn's original work was based on first projecting $\mathbf{H}$ to the complex plane $\mathbf{C}$ with a certain lattice of points deleted. We denote the latter by $\mathbf{P}$. His projection $\sigma_1 : z \mapsto w$ satisfies the equation

$$1 - \mathcal{J}(z) = 4\mathcal{P}(w) + 1. \tag{1.20}$$

where $\mathcal{J}$ is the elliptic modular function and $\mathcal{P}$ is the Weierstrass elliptic function defined by

$$(\mathcal{P}'(w))^2 = 4\mathcal{P}(w) + 1.$$

We shall need a better account of this map than Cohn has provided. The elliptic modular function is well-known. A description of it may be found in Cohn's book, [7]. Weierstrass elliptic functions in general are also discussed there although the particular function $\mathcal{P}$ is not. The properties of $\mathcal{P}$ may be deduced from those of $\mathcal{P}_0$ where

$$(\mathcal{P}_0'(w))^2 = 4\mathcal{P}_0(w) - 1.$$

This function, referred as the "equianharmonic case", is described in [41], for instance. Its relationship to $\mathcal{P}$ is given by $\mathcal{P}(w) = -\mathcal{P}_0(iw)$. Note also that

$(\mathcal{P}'(w))^2 = -(\mathcal{P}_0'(iw))^2$. The associated tessellations of $\mathbf{H}$ and $\mathbf{P}$ (to be introduced shortly) and their groups are fully reviewed in Magnus's book, [28].

The elliptic modular function is an analytic map from the upper half-plane $\mathbf{H}$ to the complex plane $\mathbf{C}$ which is automorphic with respect to the modular group $\Gamma$. That is, $\mathcal{J}(T(z)) = \mathcal{J}(z)$ for all $T \in \Gamma$. The group $\Gamma$ is discontinuous in $\mathbf{H}$ and a fundamental domain for it consists of the hyperbolic triangle with vertices at $\rho$, $\rho + 1$ and $\infty$ where $\rho = e^{i2\pi/3}$. The boundary points to the left of the imaginary axis are included but those to the right are not. The function $\mathcal{J}$ is a bijection from this triangle to $\mathbf{C}$.

In order to make better use of Cohn's projection and its properties we consider $\mathcal{J}$ in relation to the extended modular group $\Gamma^*$. If $T \in \Gamma^*$ and $T \notin \Gamma$ then $\mathcal{J}(T(z))$ is the complex conjugate of $\mathcal{J}(z)$. A fundamental domain for $\Gamma^*$ is the hyperbolic triangle with vertices $\rho$, $i$ and $\infty$. The function $\mathcal{J}$ maps this triangle to the half-plane in $\mathbf{C}$ lying above the real axis. Further, $\mathcal{J}$ maps its boundary to the real axis and in particular $\mathcal{J}(\rho) = 0$ and $\mathcal{J}(i) = 1$. Observe that the fundamental domain for $\Gamma$ together with its boundary consists of the fundamental domain for $\Gamma^*$ and its reflection in the imaginary axis. Since reflection in the imaginary axis is an element of $\Gamma^*$ but not $\Gamma$ it is clear that $\mathcal{J}$ maps the hyperbolic triangle with vertices $i$, $\rho + 1$ and $\infty$ to the half-plane in $\mathbf{C}$ lying below the real axis.

The situation for $\mathcal{P}$ is analogous. However, before we describe it, it is convenient to introduce some notation. We take as generators for $\Gamma^*$ the hyperbolic reflections in each of the three geodesics containing a side of its fundamental domain. That is, we take as generators for $\Gamma^*$ the transformations

$$T_1(z) = -\overline{z} \qquad T_2(z) = 1/\overline{z} \qquad T_3(z) = -\overline{z} - 1.$$

A presentation for $\Gamma^*$ is

$$\Gamma^* = \langle\, T_1, T_2, T_3 \mid T_1^2 = T_2^2 = T_3^2 = (T_1T_2)^2 = (T_2T_3)^3 = \mathrm{Id} \,\rangle.$$

The modular group $\Gamma$ is the subgroup of index 2 consisting of all words of even length. It is generated by $U_2 = T_1T_2$ and $U_3 = T_2T_3$ and has presentation

$$\Gamma = \langle\, U_2, U_3 \mid U_2^2 = U_3^3 = \mathrm{Id} \,\rangle.$$

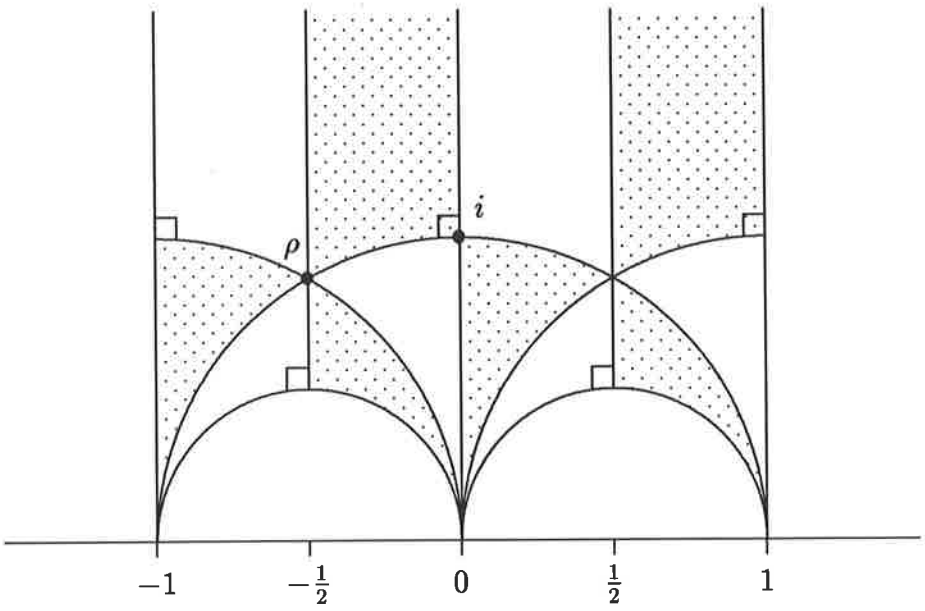


FIGURE 1.3. A tiling of the fundamental domain for $\Gamma'$ induced by $\Gamma^*$. Tiles which are congruent under $\Gamma$ are either both shaded or both unshaded and *vica-versa*.

The transformation $U_2$ is hyperbolic rotation through $\pi$ about $i$ and $U_3$ is hyperbolic rotation through $2\pi/3$ clockwise about $\rho$. The generators $A$ and $B$ of $\Gamma'$ satisfy

$$A = U_2U_3^{-1}U_2U_3 \qquad \text{and} \qquad B = U_3U_2U_3^{-1}U_2.$$

A tessellation of the fundamental domain for $\Gamma'$ by $\Gamma^*$ is shown in Figure 1.3.

The Weierstrass elliptic function $\mathcal{P}$ is an analytic map from $\mathbf{P}$ to the complex plane $\mathbf{C}$ which is automorphic with respect to a group $\Sigma$ of transformations generated by

$$u_2(w) = -w + 2w_0 \qquad \text{and} \qquad u_3(w) = w/\epsilon^2 + 2w_0. \tag{1.21}$$

Here $w_0 = -i1.52995\ldots$ is a known constant and $\epsilon = e^{i\pi/3}$. The domain $\mathbf{P}$ is obtained from $\mathbf{C}$ by removing all the images of the origin O under $\Sigma$. The transformation $u_2$ is rotation through $\pi$ about $w_0$ and $u_3$ is clockwise rotation through $2\pi/3$ about $w_1 = w_0 - iw_0/\sqrt{3}$. Again we consider an extension $\Sigma^*$ of $\Sigma$ with index 2. The coset of transformations lying in $\Sigma^*$ but not $\Sigma$ is represented

by reflection in the imaginary axis. If $T$ is any transformation in this coset then $\mathcal{P}(T(w))$ is the complex conjugate of $\mathcal{P}(w)$. A fundamental domain for $\Sigma^*$ is the triangle with vertices $w_1$, $w_0$ and O and hence a fundamental domain for $\Sigma$ is this triangle together with its reflection in the imaginary axis. (To be rigorous, the fundamental domain for $\Sigma$ includes the boundary points which lie to the left of the imaginary axis and excludes those to the right.) As with $\mathcal{J}$, the function $\mathcal{P}$ is a bijection from this fundamental domain to $\mathbf{C}$. It maps the left half to those points lying below the real axis and the right half to those points lying above it. Further, the boundary of each half is mapped to the real axis and in particular $\mathcal{P}(w_0) = -1/4$ and $\mathcal{P}(w_1) = 0$.

As with $\Gamma^*$, we take as generators for $\Sigma^*$ the reflections in each of the sides of its fundamental domain, namely,

$$t_1(w) = -\overline{w} \qquad t_2(w) = \overline{w} + 2w_0 \qquad t_3(w) = \epsilon^2\overline{w}.$$

A presentation for $\Sigma^*$ is

$$\Sigma^* = \langle\, t_1, t_2, t_3 \mid t_1^2 = t_2^2 = t_3^2 = (t_1t_2)^2 = (t_2t_3)^3 = (t_3t_1)^6 = \mathrm{Id} \,\rangle.$$

Likewise, $\Sigma$ is generated by $u_2 = t_1t_2$ and $u_3 = t_2t_3$ and has presentation

$$\Sigma = \langle\, u_2, u_3 \mid u_2^2 = u_3^3 = (u_2u_3)^6 = \mathrm{Id} \,\rangle.$$

Naturally, we shall also be interested in the subgroup $\Sigma'$ generated by

$$a = u_2u_3^{-1}u_2u_3 \qquad \text{and} \qquad b = u_3u_2u_3^{-1}u_2.$$

Whilst this notation is in conflict with our earlier usage of the symbols $a$ and $b$ to denote the generators of $\pi_1(\mathbf{T})$ no ambiguity will arise as a result of this. The meaning of $a$ and $b$ will always be clear from the context. An elementary calculation shows

$$a(w) = w + 2\epsilon w_0 \qquad \text{and} \qquad b(w) = w + 2w_0/\epsilon. \tag{1.22}$$

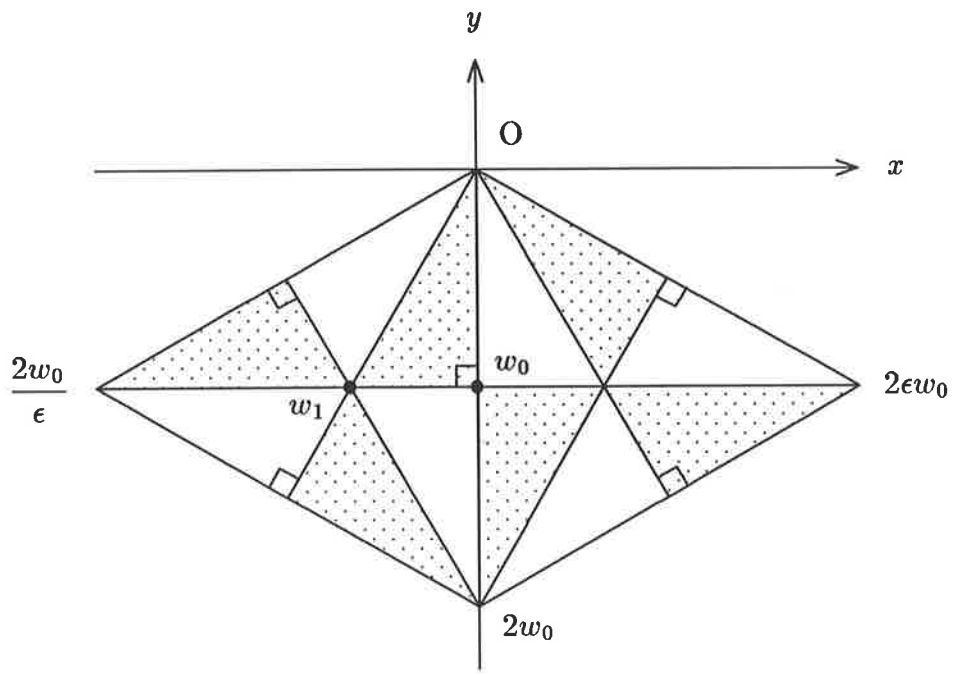


FIGURE 1.4. A tiling of the fundamental domain for $\Sigma'$ induced by $\Sigma^*$. Here $w_0 = -i1.52995\ldots$, $w_1 = w_0 - iw_0/\sqrt{3}$ and $\epsilon = e^{i\pi/3}$. Tiles congruent under $\Sigma$ are either both shaded or both unshaded and *vica-versa*.

Thus $a$ and $b$ are translation by $2\epsilon w_0$ and $2w_0/\epsilon$, respectively, and $\Sigma'$ is the free abelian group of rank two. A tessellation of the fundamental domain for $\Sigma'$ by $\Sigma^*$ is shown in Figure 1.4.

We are ready to describe Cohn's map. It is clear from the discussion above that equation (1.20) defines a bijection $\sigma_1$ from the fundamental domain of $\Gamma$ to that of $\Sigma$. Moreover, $\sigma_1$ and its inverse are analytic on the interiors of these domains. We can use the groups $\Gamma$ and $\Sigma$ to extend $\sigma_1$ to all of $\mathbf{H}$ and $\mathbf{P}$. We know from the presentations for $\Gamma$ and $\Sigma$ that there is a homomorphism $\pi$ from $\Gamma$ to $\Sigma$ defined by

$$\pi(U_2) = u_2 \qquad \pi(U_3) = u_3. \tag{1.23}$$

We extend $\sigma_1$ by defining $\sigma_1(z) = t(\sigma_1(T^{-1}(z)))$ where $t = \pi(T)$ and $T$ is the element of $\Gamma$ for which $T^{-1}(z) \in \mathcal{D}$. The only points in $\mathbf{H}$ at which there may be a problem with this definition are the points which are fixed by elliptic elements of $\Gamma$. By noting that $\sigma_1(i) = w_0$ and $\sigma_1(\rho) = w_1$ it is not hard to verify that $\sigma_1$ is well-defined at such points. An immediate consequence of the definition of $\sigma_1$ is

that for all $z \in \mathbf{H}$ and $T \in \Gamma$ we have

$$\sigma_1(T(z)) = t(\sigma_1(z)), \qquad \text{where } t = \pi(T). \tag{1.24}$$

Clearly $\sigma_1$ maps $\mathbf{H}$ onto $\mathbf{P}$. With care at the elliptic fixed points it is possible to verify that $\sigma_1$ is analytic. Note that $\sigma_1$ maps the fundamental domain for $\Gamma'$ and its tiling by $\Gamma^*$ shown in Figure 1.3 to the fundamental domain for $\Sigma'$ and its tiling by $\Sigma^*$ shown in Figure 1.4.

It is evident from the presentations of $\Gamma$ and $\Sigma$ that the kernel of the homomorphism $\pi$ is the smallest normal subgroup of $\Gamma$ containing the transformation

$$B^{-1}A^{-1}BA = (U_2U_3)^6.$$

That is, Ker $\pi$ is generated by $B^{-1}A^{-1}BA$ and its conjugates in $\Gamma$. Every such conjugate is a conjugate of $B^{-1}A^{-1}BA$ by an element of $\Gamma'$ and so Ker $\pi$ is in fact the smallest normal subgroup of $\Gamma'$ containing $B^{-1}A^{-1}BA$. We know $\Gamma'$ is the free group generated by $A$ and $B$ and it follows that Ker $\pi$ is the commutator subgroup $\Gamma''$ of $\Gamma'$. Now observe that since $\pi$ is defined by (1.23) we have $\pi(A) = a$ and $\pi(B) = b$ and therefore the image of $\Gamma'$ under $\pi$ is the group $\Sigma'$. In other words the map

$$\pi : \quad \Gamma' \quad \longrightarrow \quad \Sigma' \tag{1.25}$$

is a homomorphism defined by $\pi(A) = a$ and $\pi(B) = b$. Obviously its kernel is $\Gamma''$. This is of course a realisation of the well-known fact that when the free group of rank two is factored by its commutator subgroup the result is the free abelian group of rank two.

Given that Ker $\pi = \Gamma''$ and (1.24) holds, it is clear that $\sigma_1$ identifies points in $\mathbf{H}$ which are $\Gamma''$-equivalent. We claim that $\sigma_1$ only identifies points which are $\Gamma''$-equivalent. To see this, suppose $\sigma_1(z) = \sigma_1(z')$ for some $z, z' \in \mathbf{H}$. Choose $T$ and $T'$ in $\Gamma'$ so that $T(z)$ and $T'(z')$ both lie in $\mathcal{D}$. Obviously $\sigma_1(T(z))$ and $\sigma_1(T'(z'))$ both lie in the fundamental domain for $\Sigma'$ shown in Figure 1.4. Moreover, if $t$ and $t'$ are the images of $T$ and $T'$ under $\pi$, respectively, then

$$t't^{-1}(\sigma_1(T(z))) = t'(\sigma_1(z)) = t'(\sigma_1(z')) = \sigma_1(T'(z')).$$

We conclude that $\sigma_1(T(z))$ and $\sigma_1(T'(z'))$ are equivalent under the transformation $t't^{-1} \in \Sigma'$. This is only possible if $\sigma_1(T(z)) = \sigma_1(T'(z'))$ and $t't^{-1} = \mathrm{Id}$. Therefore $T(z) = T'(z')$ and $T'T^{-1}$ lies in $\Gamma''$. It follows that $z$ and $z'$ are $\Gamma''$-equivalent.

We have shown that the effect of $\sigma$ is precisely to identify points which are $\Gamma''$-equivalent. Since $\sigma_1$ is analytic we conclude that $\mathbf{P}$ is quotient of $\mathbf{H}$ by $\Gamma''$. That is, $\sigma_1$ is the projection

$$\sigma_1 : \quad \mathbf{H} \quad \longrightarrow \quad \mathbf{H}/\Gamma'' = \mathbf{P}. \tag{1.26}$$

As with $\Gamma$, the group $\Gamma''$ contains no elliptic elements and so $\sigma_1$ defines a universal covering of $\mathbf{P}$. We use $\sigma_1$ to transfer the hyperbolic metric to $\mathbf{P}$ and hence can speak of geodesics on $\mathbf{P}$. Of course, if we then form the quotient of $\mathbf{P}$ with respect to the group $\Sigma'$ the combined projection map is (1.10) and we obtain the punctured torus $\mathbf{T}$. We point out here that a standard method for constructing the torus is to form a quotient space like $\mathbf{C}/\Sigma'$. Since we have obtained $\mathbf{P}$ from $\mathbf{C}$ by removing $O$ and its images under $\Sigma'$ from $\mathbf{C}$ it is clear that $\mathbf{P}/\Sigma'$ is a punctured torus.

We shall refer to the map $\sigma_1$ as Cohn's *commutator map*. As Cohn discovered, [8], the significance of $\sigma_1$ is that it maps the geodesics in $\mathbf{H}$ which correspond to the Markoff forms precisely to the geodesics in $\mathbf{P}$ which pass between the points of the lattice $\Sigma'(O)$ like those straight lines in $\mathbf{P}$ which are parallel to the vectors joining $O$ to the points of $\Sigma'(O)$. (We shall formulate this statement more precisely in the section of Chapter 2 dealing with cutting sequences.)

We conclude this chapter with the following interesting aside. We know $\sigma_1$ determines a family of isomorphisms between $\Gamma''$ and $\pi_1(\mathbf{P})$. We take the base point of $\pi_1(\mathbf{P})$ to be $w_0$ and we consider the isomorphism determined by lifting $w_0$ to $i$. Recall that this isomorphism maps a transformation $W$ to the homotopy class which contains the projection of the geodesic segment $[i, W(i)]$. It is easy to see that if $C = B^{-1}A^{-1}BA$ then loop $\sigma_1([i, C(i)])$ is based at $w_0$ and bounds a disc containing the point $O$. Now observe that $\Gamma''$ is generated by $C$ and all its conjugates in $\Gamma'$ and consider $W = VCV^{-1}$ where $V \in \Gamma'$. In this case, the geodesic segment $[i, W(i)]$ is homotopic to the composition of $[i, V(i)]$ with $[V(i), VC(i)]$ and $[VC(i), W(i)]$. The projection of $[i, V(i)]$ is a curve from $w_0$ to $v(w_0)$ where $v$ is the

image of $V$ under the homomorphism (1.25). Since $[V(i), VC(i)] = V([i, C(i)])$, its projection is the image of $\sigma_1([i, C(i)])$ under $v$ and so forms a loop based at $v(w_0)$ which bounds a disc containing $v(O)$. Finally, $[VC(i), W(i)] = W([V(i), i])$ and since $W \in \Gamma''$ the projection of $W([V(i), i])$ is the reverse of $\sigma_1([i, V(i)])$. It follows that $\sigma_1([i, W(i)])$ is homotopic to a loop based at $w_0$ which bounds a disc containing the point $v(O)$. We conclude, as expected, that $\pi_1(\mathbf{P})$ is generated by all the homotopy classes containing such loops.

CHAPTER 2

# CUTTING SEQUENCES

Our main tool for studying the topology of open geodesics on $\mathbf{T}$ is cutting sequences. Descriptions of cutting sequences and their properties may be found in [27], [35], [36] and [39]. In this chapter we shall mostly be reviewing the circle of ideas presented by Caroline Series in [36], however we do produce some new material. Theorem 2.2 and Remark 2.1 contain new results about the cutting sequences of closed geodesics. More importantly, the sections on reducibility for doubly infinite sequences and automorphisms of $\Gamma'$ applied to reduced sequences provide a rigourous basis for the application of the automorphisms of $\Gamma'$ to cutting sequences. Also, the final section which deals with $O$-radial and half-linear cutting sequences is completely original.

Throughout this chapter unless specifically stated otherwise, all geodesics in $\mathbf{H}$ will have irrational endpoints and all geodesics on $\mathbf{T}$ will be the projection of such geodesics. Recall from Chapter 1 that this restriction ensures that the corresponding forms do not represent zero. It also ensures that the associated cutting sequences are doubly infinite. The complete definition of the cutting sequence $\mathbf{S}(\gamma)$ of a geodesic $\gamma$ is given in the first section. For the moment, we merely note that they are doubly infinite sequences composed of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$. It is conventional to omit the commas when listing such sequences. Thus we write

$$\mathbf{S}(\gamma) = \ldots\ldots X_{-1}X_0X_1\ldots\ldots$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. We shall also use the abbreviation

$$W^n = \overbrace{W \ldots W}^{n}$$

where $n$ is a positive integer and $W$ is any word in the symbols $A$, $B$, $A^{-1}$ and $B^{-1}\}$. Likewise, we interpret $W^{\infty}$ as $\ldots WWW$ or $WWW\ldots$ or $\ldots WWW\ldots$ depending on the context. Although we shall eventually identify the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ with the corresponding elements of $\Gamma'$ we shall in general refrain from using the notation $W^{-n}$ except when $n = 1$. Of course $W^{-1}$ denotes the word obtained by reversing $W$ and replacing each symbol by its inverse.

### Cutting sequences

To define cutting sequences we must first describe a labelled grid of geodesics in $\mathbf{H}$. The grid is that which partitions $\mathbf{H}$ into the standard tessellation by $\Gamma'$. Specifically, the grid is obtained by taking all images under $\Gamma'$ of (the sides of) the fundamental domain $\mathcal{D}$ shown in Figure 1.2. It is labelled by first labelling $\mathcal{D}$. The label $A$ is placed next to the side joining $-1$ to $\infty$ and then, proceeding anti-clockwise around $\mathcal{D}$, the labels $B$, $A^{-1}$ and $B^{-1}$, in that order, are placed next to the remaining three sides. Since $\Gamma'$ is freely generated by $A$ and $B$ we know that each of its elements produces a distinct image of $\mathcal{D}$. Thus we can use $\Gamma'$ to copy the labelling in $\mathcal{D}$ to all its images in the associated tessellation of $\mathbf{H}$. We refer to the resulting grid as the *labelled grid induced by* $\Gamma$ and we denote it by $\Lambda$. Recall that we are interested only in geodesics whose endpoints are irrational. Each such oriented geodesic $\gamma$ cuts the lines of $\Lambda$ infinitely often in each direction. The *cutting sequence* $\mathbf{S}(\gamma)$ of $\gamma$ is the doubly infinite sequence of labels which records its intersections with $\Lambda$, with the conventions that reading $\mathbf{S}(\gamma)$ from left to right corresponds to traversing $\gamma$ according to its orientation and only the label immediately after each grid line is listed in $\mathbf{S}(\gamma)$. See Figure 2.1. Although we shall use the notation

$$\mathbf{S}(\gamma) = \ldots\ldots X_{-1}X_0X_1\ldots\ldots,$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, we shall usually consider $\mathbf{S}(\gamma)$ to be un-indexed. That is, we identify the different possible indexings of $\mathbf{S}(\gamma)$ in exactly the same manner as we have for the doubly infinite sequences of positive integers. This is consistent with our decision to not differentiate between the different possible

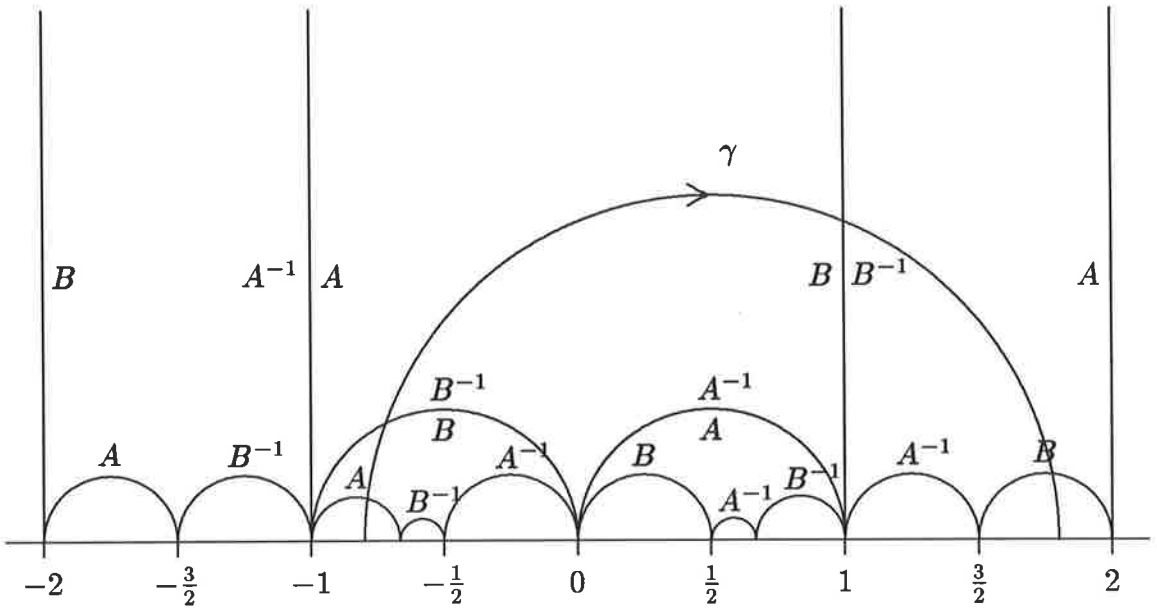


FIGURE 2.1. The labelled grid $\Lambda$ induced by $\Gamma'$. The cutting sequence of a geodesic $\gamma$ is the sequence of labels $\mathbf{S}(\gamma) = \ldots AB^{-1}B^{-1}B^{-1}\ldots$ which records its intersections with $\Lambda$, with the convention that only the label which occurs immediately after each grid line is listed.

parameterisations of $\gamma$. Note that there is no ambiguity in omitting the commas from $\mathbf{S}(\gamma)$ since it is composed of only the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$.

By using $\sigma$ to project $\Lambda$ to $\mathbf{T}$ we can likewise define cutting sequences for geodesics on $\mathbf{T}$. It is clear that the grid on $\mathbf{T}$ is covered by the projection of the sides of $\mathcal{D}$. Since $\sigma$ identifies the opposite sides of $\mathcal{D}$, the projected grid consists of only two geodesics; each begins and ends at the puncture. Also, since we labelled $\Lambda$ by using $\Gamma'$ to copy the labels in $\mathcal{D}$, the labels on its projection to $\mathbf{T}$ all agree and are simply the projection of those in $\mathcal{D}$. Obviously the cutting sequences of geodesics are preserved under this projection, that is,

$$\mathbf{S}(\sigma(\gamma)) = \mathbf{S}(\gamma) \tag{2.1}$$

where $\gamma$ is any geodesic in $\mathbf{H}$.

In order to use cutting sequences to study the topology of geodesics on $\mathbf{T}$ we need to identify the labels $A$, $B$, $A^{-1}$ and $B^{-1}$ with the corresponding elements of $\Gamma'$. This will allow us, for instance, to establish a direct relationship between

the cutting sequence of a closed geodesic on $\mathbf{T}$ and the conjugacy class in $\Gamma'$ which defines $\gamma$. Before we discuss such things we describe some of the properties of cutting sequences.

It is not hard to deduce from the fact that cutting sequences are preserved by $\sigma$ that they are also preserved under the action of $\Gamma'$ on $\mathbf{H}$. In other words, if $\gamma$ is a geodesic in $\mathbf{H}$ and $T \in \Gamma'$ then $\mathbf{S}(T(\gamma)) = \mathbf{S}(\gamma)$. Series points out that the converse is also true, that is, if $\gamma$ and $\gamma'$ are both geodesics in $\mathbf{H}$ and $\mathbf{S}(\gamma') = \mathbf{S}(\gamma)$ then $\gamma' = T(\gamma)$ for some $T \in \Gamma'$. Briefly, the way to prove this is to choose $T$ so that $\gamma'$ and $T(\gamma)$ traverse the same sequence of tiles in $\mathbf{H}$. In this case, $\gamma'$ and $T(\gamma)$ have the same endpoints and hence are identical. To summarise,

$$\mathbf{S}(\gamma') = \mathbf{S}(\gamma) \qquad \Longleftrightarrow \qquad \gamma' = T(\gamma) \text{ for some } T \in \Gamma'. \tag{2.2}$$

Note in particular that a geodesic on $\mathbf{T}$ is uniquely determined by its cutting sequence.

By considering our fundamental domain for $\Gamma'$ and its neighbouring tiles, see Figure 2.1, and by noting that $\Lambda$ and its labels are preserved under the action of $\Gamma'$, it is not hard to deduce that if the label $X$ appears on one side of a grid line then the label $X^{-1}$ appears on the other. By $X^{-1}$ we mean of course the label for which $XX^{-1} = \mathrm{Id}$ in $\Gamma'$. One consequence of this is that the string $XX^{-1}$ never occurs in a cutting sequence. The reason being that a geodesic $\gamma$ in $\mathbf{H}$ must enter and leave each tile it crosses by different sides. If $XX^{-1}$ occurred in $\mathbf{S}(\gamma)$ then the label $X$ would appear twice within the same tile which is impossible. A sequence which has no occurrence of $XX^{-1}$ is called *reduced*. Evidently cutting sequences are reduced. We also note that if $\gamma'$ is the geodesic $\gamma$ with the opposite orientation then $\mathbf{S}(\gamma')$ can be obtained from $\mathbf{S}(\gamma)$ by reversing it and interchanging $A$ with $A^{-1}$ and $B$ with $B^{-1}$.

Not only are cutting sequences reduced but they also do not contain strings of the form

$$(ABA^{-1}B^{-1})^{\infty} \qquad \text{or} \qquad (BAB^{-1}A^{-1})^{\infty}. \tag{2.3}$$

This can be deduced from the properties of the tessellation. The set of images of the vertices of $\mathcal{D}$ under $\Gamma$ is precisely the set of rationals together with $\infty$. Moreover,

for any given rational $r$, the images of $\mathcal{D}$ which have a vertex at $r$ form a fan of neighbouring tiles. Associated with this fan is a sequence of grid lines emanating from $r$ and hence a sequence of labels. By considering the canonical fan at $\infty$ it is not hard to see that, depending upon the direction in which the fan is traversed, the sequence is one those listed in (2.3). Now suppose that some geodesic $\gamma$ has a cutting sequence $\mathbf{S}(\gamma)$ which ends with such a sequence. Since the corresponding sequence of grid lines crossed by $\gamma$ is completely determined by the initial one and that initial one belongs to some fan, it follows that at some point $\gamma$ enters a fan of tiles and never leaves again. This can only happen if one endpoint of $\gamma$ is the rational $r$ upon which the fan is based. However, no pair of geodesics with endpoint $r$ can intersect and therefore the cutting sequence of $\gamma$ terminates when it enters the fan. This contradiction shows that $\mathbf{S}(\gamma)$ cannot end with one of the sequences in (2.3). A similar argument shows that $\mathbf{S}(\gamma)$ cannot begin with such a sequence.

We have seen that the cutting sequence of a geodesic with irrational endpoints is a reduced doubly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s in which neither of the sequences (2.3) occurs. The converse is also true, that is, if $\mathbf{S}$ is a reduced sequence of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ in which neither of the sequences (2.3) occurs then $\mathbf{S}$ is a cutting sequence of some geodesic $\gamma$ with irrational endpoints. The proof of this involves constructing from $\mathbf{S}$ the endpoints of $\gamma$. Series, [36], has given a brief indication of how this may be done. Her idea is to choose a polygonal path (a sequence of adjoining geodesic segments) with cutting sequence $\mathbf{S}$. Since $\mathbf{S}$ is reduced such a path converges to two points $\eta$ and $\xi$ on the real axis. Further, since neither of the sequences (2.3) occurs $\eta$ and $\xi$ are irrational. It is possible to prove from the construction that if $\gamma = [\eta, \xi]$ then $\mathbf{S} = \mathbf{S}(\gamma)$.

We shall conclude this section with a lemma. It may be found in [39].

**Lemma 2.1.** *If $z_0 \in \mathcal{D}$ and if $[z_0, z_1]$ is a geodesic segment with cutting sequence $W = X_1 X_2 \dots X_n$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, then $W^{-1}(z_1)$ lies in $\mathcal{D}$.*

*Proof.* Let the segment $[z_0, z_1]$ be as described. We prove the lemma by induction on $n$. It is easy to verify the lemma is true when $n = 1$. Now suppose $n > 1$. Let $z_2$ be a point on $[z_0, z_1]$ such that $[z_0, z_2]$ has cutting sequence

$V = X_1 X_2 \ldots X_{n-1}$ and let $[z_2', z_1']$ be the image of $[z_2, z_1]$ under $V^{-1}$. Our inductive hypothesis is that $z_2'$ lies in $\mathcal{D}$. Since $V$ preserves cutting sequences, $[z_2', z_1']$ has cutting sequence $X_n$. Using our inductive hypothesis again we conclude that $X_n^{-1}(z_1')$ also lies in $\mathcal{D}$. Obviously $W^{-1}(z_1) = X_n^{-1}V^{-1}(z_1) = X_n^{-1}(z_1')$ and hence the lemma is true. □

### Boundary expansions

The cutting sequence of a geodesic in **H** can be obtained from the boundary expansions of its endpoints. Boundary expansions are defined by Series in [34]. More details may be found in [3] and [39]. The definition Series gives applies to the limit points of a Fuchsian group acting on the unit disc. We reformulate Series work in terms of $\Gamma'$ acting on the upper half-plane **H**. The limit points of $\Gamma'$ in **H** are exactly the irrationals. Motivated by Series work, we define the *boundary expansion* $\mathbf{S}(\xi)$ of an irrational point $\xi$ to be the cutting sequence of any oriented geodesic ray which begins within $\mathcal{D}$ and ends at $\xi$ and we write

$$\mathbf{S}(\xi) = X_0 X_1 X_2 \ldots\ldots,$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. (We can assume if desired that the ray starts at $i$.) It is not hard to see from our comments on cutting sequences that each boundary expansion is an infinite reduced sequence of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ in which neither $(ABA^{-1}B^{-1})^{\infty}$ nor $(BAB^{-1}A^{-1})^{\infty}$ occurs. It is also not hard to verify that the operation of forming boundary expansions defines a bijection between the irrationals and the set of all such sequences.

We claim that if the irrationals $\eta$ and $\xi$ have boundary expansions

$$\mathbf{S}(\eta) = X_{-1} X_{-2} X_{-3} \ldots\ldots \qquad \text{and} \qquad \mathbf{S}(\xi) = X_0 X_1 X_2 \ldots\ldots,$$

respectively, then the geodesic $\gamma = [\eta, \xi]$ has cutting sequence

$$\mathbf{S}(\gamma) = \ldots\ldots X_{-k-3}^{-1} X_{-k-2}^{-1} X_{-k-1}^{-1} X_k X_{k+1} X_{k+2} \ldots\ldots$$

where $k \geq 0$ is the smallest integer such that $X_{-k-1} \neq X_k$. Moreover, the segment of $\gamma$ between the labels $X_{-k-1}^{-1}$ and $X_k$ lies inside the image of $\mathcal{D}$ under the

transformation $W = X_0X_1 \dots X_{k-1}$. To see this, let $k$ be as described. We can choose a point $z_1$ on the ray $[i, \xi]$ so that the cutting sequence of the segment $[i, z_1]$ is $X_0X_1 \dots X_{k-1}$. It follows from Lemma 2.1 that $W^{-1}(z_1)$ lies in $\mathcal{D}$. Similarly, we can choose a point $z_2$ on the ray $[i, \eta]$ so that the cutting sequence of $[i, z_2]$ is $X_{-1}X_{-2} \dots X_{-k}$. Since $W = X_{-1}X_{-2} \dots X_{-k}$ we know that $W^{-1}(z_2)$ also lies in $\mathcal{D}$. Thus $z_1$ and $z_2$ both lie in $W(\mathcal{D})$. Now let $\gamma$ be a geodesic with cutting sequence $\mathbf{S}(\gamma) = \dots X_{-k-3}^{-1}X_{-k-2}^{-1}X_{-k-1}^{-1}X_kX_{k+1}X_{k+2} \dots$ We can choose $\gamma$ so that the segment of it between the labels $X_{-k-1}^{-1}$ and $X_k$ also lies in $W(\mathcal{D})$. In this case, there is a ray contained in $\gamma$ which begins in $W(\mathcal{D})$ and has the same cutting sequence as the ray $[z_1, \xi]$. We conclude that $\xi$ is an endpoint of $\gamma$. Similarly, there is a ray contained $\gamma$ which ends in $W(\mathcal{D})$ and has the same cutting sequence as $[\eta, z_2]$ (note the change of orientation here). Hence $\eta$ is the other endpoint of $\gamma$ and the claim is proved.

It is evident from Theorem A of Birman and Series paper, [3], that there is an ordering of boundary expansions which reflects the natural ordering of the corresponding irrationals. It is based on the cyclic ordering of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ given by the ordering

$$A^{-1} < B < A < B^{-1} \tag{2.4}$$

and all its cyclic permutations. Let $X_0X_1X_2 \dots$ and $X_0'X_1'X_2' \dots$ be distinct boundary expansions and let $k \geq 0$ be the smallest integer such that $X_k \neq X_k'$. We write

$$X_0X_1X_2 \dots\dots \quad < \quad X_0'X_1'X_2' \dots\dots$$

if either $k = 0$ and $X_0 < X_0'$ in the ordering (2.4) or $k \geq 1$ and $X_k < X_k'$ in the ordering obtained from (2.4) by cyclically permuting it so that $X_{k-1}^{-1}$ is the smallest term. Birman and Series refer to this ordering as the cyclic *lexicographic ordering* of boundary expansions. We omit the word cyclic. While it is straightforward to deduce from Theorem A of Birman and Series' paper that the lexicographic ordering of boundary expansions agrees with that inherited from the natural ordering of the corresponding irrationals, we present the following proof to provide insight into the use of boundary expansions.

**Theorem 2.1.** *Let $\xi$ and $\xi'$ be distinct irrationals with boundary expansions*

$$\mathbf{S}(\xi) = X_0X_1X_2\ldots\ldots \qquad \text{and} \qquad \mathbf{S}(\xi') = X_0'X_1'X_2'\ldots\ldots,$$

*respectively. Then $\xi < \xi'$ if and only if $X_0X_1X_2\ldots < X_0'X_1'X_2'\ldots$.*

*Proof.* It is sufficient to prove only the forward implication. Thus we assume $\xi < \xi'$ and we shall prove that $X_0X_1X_2\ldots < X_0'X_1'X_2'\ldots$. Let $k$ be the smallest non-negative integer such that $X_k \neq X_k'$. We deal with the case $k = 0$ first. The sides of $\mathcal{D}$ partition the real axis into four intervals. Since the cutting sequence of $[i, \xi]$ begins with $X_0$ we know that $\xi$ lies in the interval bounded by the side with external label $X_0$. Similarly, $\xi'$ lies in the interval bounded by the side with external label $X_0'$. We are assuming $\xi < \xi'$ and $X_0 \neq X_0'$. Since the external labels of $\mathcal{D}$ appear in the order given by (2.4) as one traverses the real axis from left to right we conclude that $X_0 < X_0'$ in the ordering (2.4) and hence $X_0X_1X_2\ldots < X_0'X_1'X_2'\ldots$.

Now suppose $k \geq 1$ and let $W = X_0X_1\ldots X_{k-1}$. We know from Lemma 2.1 that any segment of the ray $[i, \xi]$ with cutting sequence $W$ ends in the tile $W(\mathcal{D})$. Hence the term $X_{k-1}$ in the boundary expansion of $\xi$ records the intersection of the ray $[i, \xi]$ with a side $\gamma$ of $W(\mathcal{D})$ as it enters that tile. Since the ray $[i, \infty]$ lies entirely inside $\mathcal{D}$ it cannot intersect $\gamma$ and hence the points $\xi$ and $\infty$ are separated by the endpoints of $\gamma$. It follows that $\xi$ lies in an interval of the real axis bounded by $\gamma$. The remaining sides of $W(\mathcal{D})$ partition this interval into three sub-intervals. Clearly, $\xi$ lies in the sub-interval bounded by the side of $W(\mathcal{D})$ with external label $X_k$. Since $X_i = X_i'$ for $0 \leq i \leq k-1$ we know $\xi'$ also lies in the interval of the real axis bounded by $\gamma$. Obviously $\xi$ lies in the sub-interval bounded by the side of $W(\mathcal{D})$ with external label $X_k'$. The transformation $W$ is orientation preserving and hence the order of the labels external to $W(\mathcal{D})$ is the same as that of the labels external to $\mathcal{D}$. In particular, if we start with the external label $X_{k-1}^{-1}$ of $W(\mathcal{D})$ and then proceed in an anti-clockwise direction around $W(\mathcal{D})$, the external labels of $W(\mathcal{D})$ appear in the same order as that obtained by cyclically permuting (2.4) until $X_{k-1}^{-1}$ is the smallest term. Since $\xi < \xi'$ we conclude that $X_k < X_k'$ in this ordering. Again, $X_0X_1X_2\ldots < X_0'X_1'X_2'\ldots$ and the theorem is proved. □

### Cutting sequences of closed geodesics

Given our convention that closed geodesics on $\mathbf{T}$ are closed curves without distinguished starting points, it is awkward to directly define cutting sequences for them. We overcome this difficulty by using the open geodesics which cover them. Thus we define the *cutting sequence* $\mathbf{S}(\gamma)$ *of a closed geodesic* $\gamma$ by $\mathbf{S}(\gamma) = \mathbf{S}(\gamma')$ where $\gamma'$ is the open geodesic on $\mathbf{T}$ which covers $\gamma$ and whose orientation agrees with that of $\gamma$. As Series points out, a cutting sequence $\mathbf{S}$ is periodic if and only if it is the cutting sequence of a closed geodesic $\gamma$. However, more can be proved. The period of the cutting sequence can be related to the conjugacy class defining the closed geodesic. We remind the reader that to make the connection we need to be able to view a word $W$ of the form $W = X_1X_2 \dots X_n$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, as both a sequence of labels and an element of $\Gamma'$. As mentioned earlier, we do this by identify each label with the corresponding element of $\Gamma'$. Before we state the theorem, recall that a word $W = X_1X_2 \dots X_n$ is *cyclically reduced* if it is reduced and $X_n \neq X_1^{-1}$.

**Theorem 2.2.** *Let $W$ be a word of the form $W = X_1X_2 \dots X_n$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. If $\gamma$ is a closed geodesic on $\mathbf{T}$ and if $W$ is a cyclically reduced representative of the conjugacy class defining $\gamma$ then $\mathbf{S}(\gamma)$ is periodic with period $W$. Conversely, if $\mathbf{S}$ is a periodic cutting sequence with period $W$ then $\mathbf{S} = \mathbf{S}(\gamma)$ where $\gamma$ is the closed geodesic on $\mathbf{T}$ defined by the conjugacy class $[W]$.*

*Proof.* Let $\gamma$ be a closed geodesic on $\mathbf{T}$ and suppose $W$ is a cyclically reduced representative of the conjugacy class defining $\gamma$. Since $W$ is hyperbolic and $X_1X_2 \dots X_n$ is cyclically reduced, the sequence

$$\dots\dots X_1X_2 \dots X_n\, X_1X_2 \dots X_n\, X_1X_2 \dots X_n \dots\dots \tag{2.5}$$

is a cutting sequence. Hence there is an oriented geodesic $\gamma'$ in $\mathbf{H}$ whose cutting sequence is (2.5). We can choose $\gamma'$ so that one of the pairs $X_nX_1$ in (2.5) arises from the intersections of $\gamma'$ with the sides of the fundamental domain $\mathcal{D}$ for $\Gamma'$. Further, we can choose $z_0$ and $z_1$ on $\gamma'$ so that the segment $[z_0, z_1]$ of $\gamma'$ begins in $\mathcal{D}$ and has cutting sequence $W = X_1X_2 \dots X_n$. We know from Lemma 2.1 that

$W^{-1}(z_1)$ lies in $\mathcal{D}$. The transformation $W^{-1}$ preserves cutting sequences and so $\mathbf{S}(W^{-1}(\gamma')) = \mathbf{S}(\gamma')$. Moreover, since the point $W^{-1}(z_1)$ of $W^{-1}(\gamma')$ lies in $\mathcal{D}$ we know that one of the pairs $X_nX_1$ in $\mathbf{S}(W^{-1}(\gamma'))$ results from the intersections of $W^{-1}(\gamma')$ with the sides of $\mathcal{D}$. It follows that $\mathbf{S}(W^{-1}(\gamma')) = \gamma'$ and hence $W^{-1}$ translates $\gamma'$ along itself in the direction opposite to its orientation. We conclude that $\gamma'$ is the axis of $W$. Thus the projection of $\gamma'$ to $\mathbf{T}$ covers $\gamma$ and has the same orientation as $\gamma$. By definition, $\mathbf{S}(\gamma) = \mathbf{S}(\gamma')$ and hence $\mathbf{S}(\gamma)$ is periodic with period $W$.

Conversely, suppose $\mathbf{S}$ is a periodic cutting sequence with period $W$. Clearly $W$ is cyclically reduced and hyperbolic. Let $\gamma$ be the closed geodesic on $\mathbf{T}$ defined by $W$. It follows from the first part of the proof that $\mathbf{S}(\gamma)$ is periodic with period $W$. Hence $\mathbf{S} = \mathbf{S}(\gamma)$ and the proof is complete. □

**Remark 2.1.** The proof of Theorem 2.2 shows that if $W = X_1X_2 \dots X_n$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, is a cyclically reduced hyperbolic transformation then its axis intersects the fundamental domain $\mathcal{D}$. The converse is also true. To see this, let $W$ be a hyperbolic transformation whose axis $\gamma'$ intersects $\mathcal{D}$. Choose a fundamental segment $[z_0, z_1]$ of $\gamma'$ for $W$ which begins in $\mathcal{D}$ and let its cutting sequence be $V = X_1X_2 \dots X_n$. Lemma 2.1 implies $V^{-1}(z_1)$ lies in $\mathcal{D}$. Since $W^{-1}(z_1)$ does also we can deduce that $W = V$. It is also clear that $\gamma'$ can be partitioned into the segment $[z_0, z_1]$ together with its images under $W$ and therefore $\mathbf{S}(\gamma')$ is periodic with period $X_1X_2 \dots X_n$. It follows that $W = X_1X_2 \dots X_n$ is cyclically reduced.

### *LR*-sequences

While cutting sequences provide information on the topological properties of geodesics on $\mathbf{T}$ we are also interested in their Markoff values. These are best calculated using the associated doubly infinite sequences of integers. We therefore need a means of obtaining from the cutting sequence of a given geodesic the associated sequence of integers and *vice-versa*. Series, [36], has provided the intermediate step. She begins with the Farey tessellation.

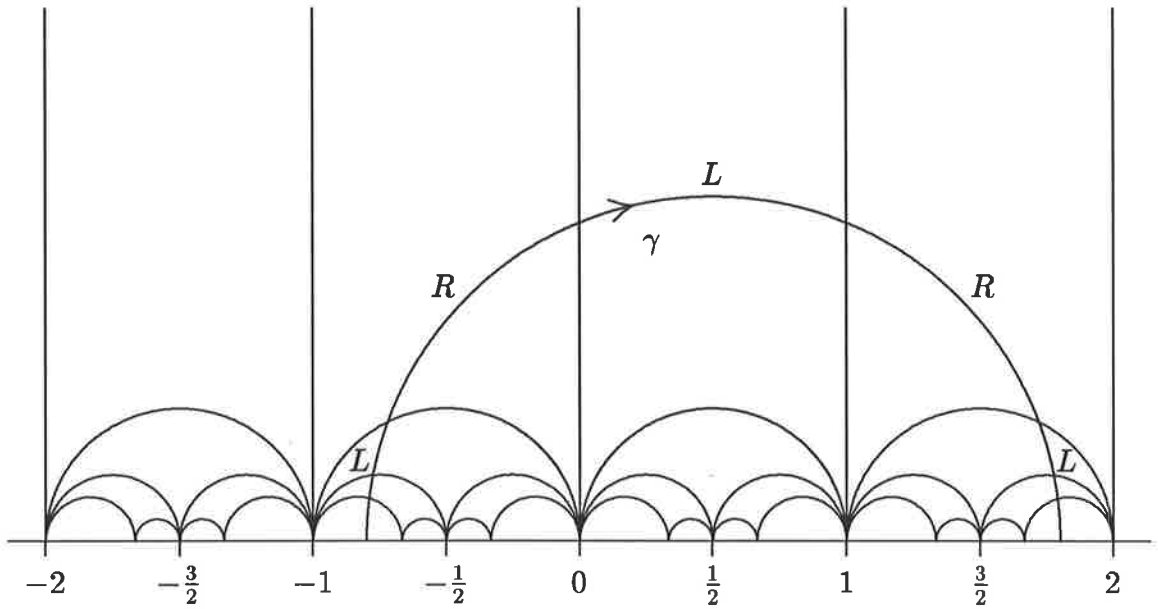


FIGURE 2.2. The Farey tessellation. The $LR$-sequence of a geodesic $\gamma$ records the manner in which it partitions the vertices of each of the triangles it passes through in the tessellation. An $L$ indicates that the isolated vertex lies to the left of $\gamma$ and an $R$ that it lies to the right.

The Farey tessellation is the tessellation of $\mathbf{H}$ obtained by taking all the images under $\Gamma$ of the ideal triangle with vertices 0,1 and $\infty$. Although this triangle is not a fundamental domain for $\Gamma$ (it is in fact a fundamental domain for a subgroup of index 3), its images under $\Gamma$ do tile $\mathbf{H}$. The resulting tessellation is invariant under the action of $\Gamma$. It derives its name from the fact that it can be described in terms of Farey sequences.

As a geodesic $\gamma$ in $\mathbf{H}$ (with irrational endpoints) traverses the Farey tessellation it cuts each of the triangles it passes through in two. Thus it divides the vertices of each triangle in its path into two sets; one set containing two vertices and the other only one. For each such triangle label the segment of $\gamma$ lying inside it by $L$ or $R$ according to whether the isolated vertex lies to the left or right, respectively, of the segment as it is traversed in the direction indicated by the orientation of $\gamma$. The resulting doubly infinite sequence of $L$'s and $R$'s is called the *$LR$-sequence* of $\gamma$. See Figure 2.2.

By adapting the methods used for cutting sequences, the following two properties can be proved. Firstly, the $LR$-sequences of two geodesics $\gamma'$ and $\gamma$ agree if and only $\gamma' = T(\gamma)$ for some $T \in \Gamma$, and secondly, a sequence of $L$'s and $R$'s is the $LR$-sequence of some geodesic if and only if neither $L^\infty$ nor $R^\infty$ occurs in it.

Series points out that it is possible to calculate the $LR$-sequence of a geodesic from its cutting sequence. The reason for this is that the tessellation of $\mathbf{H}$ by $\Gamma'$ is contained within the Farey tessellation. To be more specific, our fundamental domain $\mathcal{D}$ for $\Gamma'$ consists of the ideal triangle with vertices 0,1 and $\infty$ and its image under the transformation $T_1(z) = -\overline{z}$ and thus each tile in the tessellation $\Lambda$ consists of two triangles in the Farey tessellation. It can now be deduced that between each pair of neighbouring symbols in the cutting sequence of a geodesic there is either one or two symbols of its $LR$-sequence. It is not hard to see that only the patterns in Table 2.1 can occur. The patterns in Table 2.1 provide a recipe for calculating the $LR$-sequence of a geodesic from its cutting sequence.

| | | | |
|---|---|---|---|
| $ARB$ | $ALRA$ | $ALLB^{-1}$ | $BRRA^{-1}$ |
| $BRLB$ | $BLA$ | $A^{-1}RB^{-1}$ | $A^{-1}LRA^{-1}$ |
| $A^{-1}LLB$ | $B^{-1}RRA$ | $B^{-1}RLB^{-1}$ | $B^{-1}LA^{-1}$ |

TABLE 2.1. The patterns of symbols describing the connection between the $LR$-sequence of a geodesic and its cutting sequence.

The reason Series introduced $LR$-sequences is that it is easy to obtain from them the doubly infinite sequence of integers $\mathcal{A}$ associated with the underlying geodesic $\gamma$, see [35]. In fact, if we write the $LR$-sequence of $\gamma$ in the form

$$\ldots\ldots R^{a_{-3}}L^{a_{-2}}R^{a_{-1}}L^{a_0}R^{a_1}L^{a_2}R^{a_3}\ldots\ldots \tag{2.6}$$

then $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$. To see this, recall that the $LR$-sequence of any image of $\gamma$ under a transformation in $\Gamma$ is the same as that of $\gamma$. Clearly there is an image $\gamma'$ which contributes the first $L$ in the string $L^{a_0}$ and does so by cutting the ideal

triangle with vertices 0,1 and $\infty$. Moreover, since there is an element of $\Gamma$ which permutes the vertices of this triangle we may assume that the vertex isolated by $\gamma'$ is $\infty$. In other words, we may assume the endpoints $\eta$ and $\xi$ of $\gamma'$ satisfy $-1 < \eta < 0$ and $1 < \xi$. It is not hard to verify that in this situation

$$\eta = -[0, a_{-1}, a_{-2}, a_{-3}, \ldots] \quad \text{and} \quad \xi = [a_0, a_1, a_2, \ldots] \tag{2.7}$$

and hence $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$, as claimed.

We now have an algorithm which produces from the cutting sequence of a given geodesic the associated doubly infinite sequence of positive integers. The algorithm can be reversed. However, since there are twelve geodesics on $\mathbf{T}$ associated with a given sequence of integers, there are also twelve cutting sequences. This fact becomes apparent when one considers the reverse algorithm in detail. To this end, we let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a doubly infinite sequence of positive integers. Obviously, the geodesics with $LR$-cutting sequence (2.6) are associated with $\mathcal{A}$. However, since we are identifying $\mathcal{A}$ with the sequence obtained by shifting its indexing to the right by 1, the geodesics with $LR$-sequence

$$\ldots\ldots L^{a_{-3}} R^{a_{-2}} L^{a_{-1}} R^{a_0} L^{a_1} R^{a_2} L^{a_3} \ldots\ldots \tag{2.8}$$

are also associated with $\mathcal{A}$. We know that the connection between the cutting sequence of a geodesic and its $LR$-sequence is given by the patterns in Table 2.1. We can use the table to calculate the cutting sequence of a geodesic from its $LR$-sequence as long as we have a pattern from the table to start with. We may assume the initial pattern involves a fixed symbol, say $X_i$, in the $LR$-sequence concerned. There are exactly six possibilities for the initial pattern; two involving $X_i$ alone, two involving the pair $X_{i-1}X_i$ and two involving the pair $X_iX_{i+1}$. Thus we have accounted for the twelve cutting sequences associated with $\mathcal{A}$. By considering the details of the reverse algorithm more carefully, it is possible to determine how they are related. However, the relationship is best explained in terms of the effect of automorphisms in $\Psi$ on cutting sequences. Before we can discuss this further we need to introduce the concept of reducibility for doubly infinite sequences.

### Reducibility for doubly infinite sequences

To be able to apply automorphisms of $\Gamma'$ unambiguously to cutting sequences we must first develop the theory of reduction for doubly infinite sequences of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s. The theory is of course based on the theory of reduction for finite sequences. Recall that a word $W = X_0X_1 \dots X_n$ where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$ is *reduced* if no symbol in it is the inverse of its neighbour. Recall also that if $W$ is not reduced then it can be reduced by successively deleting all pairs of neighbouring inverses. We refer to the word which results from this process as *the reduced word equivalent to* $W$ and we note that it is the unique reduced word which is equal to $W$ in the free group $\Gamma' = F(A, B)$. (We remind the reader here that we consider $W$ to be both a sequence of symbols and an element of $\Gamma'$.) We can deal with infinite sequences and doubly infinite sequences of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s in a similar manner. We begin by viewing such sequences as the limit of a sequence of words.

Let $\mathbf{S}$ be a sequence of the form

$$\mathbf{S} = X_0X_1X_2 \dots\dots,$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. Further, let $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \mathbf{S}^{(3)}, \dots$ be a sequence of finite words and for each $j \geq 1$ write

$$\mathbf{S}^{(j)} = X_0^{(j)}X_1^{(j)} \dots X_{l(j)}^{(j)}$$

where $l(j)$ is the length of $\mathbf{S}^{(j)}$ and each $X_i^{(j)} \in \{A, B, A^{-1}, B^{-1}\}$. We say that the sequence $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \mathbf{S}^{(3)}, \dots$ *converges* to $\mathbf{S}$ and write

$$\mathbf{S} = \lim_{j \to \infty} \mathbf{S}^{(j)} \tag{2.9}$$

if for every integer $n \geq 0$ there is some $J \geq 1$ such that for all $j \geq J$ we have $l(j) \geq n$ and $X_i = X_i^{(j)}$ for $0 \leq i \leq n$. Note that the sequence $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \mathbf{S}^{(3)}, \dots$ can have at most one limit. Also, using the definition it is not hard to show that the sequence $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \mathbf{S}^{(3)}, \dots$ converges if and only if the length of the initial segment of $\mathbf{S}^{(j)}$ which agrees with $\mathbf{S}^{(j+1)}$ diverges to $\infty$ as $j$ increases. Now set

$\mathbf{S}_j = X_0X_1 \ldots X_j$ and observe that $\mathbf{S} = \lim_{j\to\infty} \mathbf{S}_j$. It is natural to use this limit to define reducibility for the sequence $\mathbf{S}$. Thus we say that $\mathbf{S}$ is *reducible* if the limit $\mathbf{S}' = \lim_{j\to\infty} \mathbf{S}'_j$ exists where $\mathbf{S}'_j$ is the reduced word which is equivalent to $\mathbf{S}_j$. Obviously, if $\mathbf{S}'$ exists then it is a reduced sequence. We refer to $\mathbf{S}'$ as *the reduced sequence equivalent to* $\mathbf{S}$.

We can extend the concept of reducibility to doubly infinite sequences. For this purpose, we now suppose $\mathbf{S}$ is a sequence of the form

$$\mathbf{S} = \ldots\ldots X_{-1}X_0X_1\ldots\ldots \tag{2.10}$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. We say that $\mathbf{S}$ is *reducible* if each of the sequences

$$X_0X_1X_2\ldots\ldots \qquad \text{and} \qquad X_{-1}X_{-2}X_{-3}\ldots\ldots$$

are reducible, to say

$$X'_0X'_1X'_2\ldots\ldots \qquad \text{and} \qquad X'_{-1}X'_{-2}X'_{-3}\ldots\ldots,$$

respectively, and if further there is some $k \geq 0$ such that $X'_{-k-1} \neq (X'_k)^{-1}$. In addition, if that is the case and $k$ is minimal then the sequence

$$\mathbf{S}' = \ldots\ldots X'_{-k-3}X'_{-k-2}X'_{-k-1}X'_kX'_{k+1}X'_{k+2}\ldots\ldots \tag{2.11}$$

is reduced. We refer to it as the reduced sequence *equivalent* to $\mathbf{S}$.

For reducibility to be a useful concept we need to verify that the reduced sequence equivalent to $\mathbf{S}$ is independent of the indexing of $\mathbf{S}$. By using induction (and replacing $\mathbf{S}$ by its reverse if necessary), it will suffice to do this only for the case where the indexing has been shifted to the right by one. Thus we assume $\mathbf{S}$ is of the form (2.10) and that it is reducible to the sequence (2.11) and we consider the sequence $\mathbf{S}^*$ obtained from $\mathbf{S}$ by shifting its indexing to the right by one. In order to decide whether or not $\mathbf{S}^*$ is reducible we need to examine the sequences

$$X_1X_2X_3\ldots\ldots \qquad \text{and} \qquad X_0X_{-1}X_{-2}\ldots\ldots$$

Given that we are assuming $X_0X_1X_2\ldots$ is reducible to $X_0'X_1'X_2'\ldots$ it is possible to deduce that $X_1X_2X_3\ldots$ is also reducible and that the result is either

$$X_1'X_2'X_3'\ldots\ldots \qquad \text{or} \qquad X_0^{-1}X_0'X_1'X_2'\ldots\ldots$$

depending on whether $X_0' = X_0$ or $X_0' \neq X_0$, respectively. The idea here is that if $X_0X_1\ldots X_j$ is reducible to say $X_0''X_1''\ldots X_i''$ then $X_1X_2\ldots X_j$ is reducible to either $X_1''X_2''\ldots X_i''$ or $X_0^{-1}X_0''X_1''\ldots X_i''$ depending on whether $X_0'' = X_0$ or $X_0'' \neq X_0$, respectively, and this property is preserved when the limit is taken. Similarly, $X_0X_{-1}X_{-2}\ldots$ is reducible and the result is either

$$X_0X_{-1}'X_{-2}'X_{-3}'\ldots\ldots \qquad \text{or} \qquad X_{-2}'X_{-3}'X_{-4}'\ldots\ldots$$

depending on whether $X_{-1}' \neq (X_0)^{-1}$ or $X_{-1}' = (X_0)^{-1}$. We shall consider the four cases separately. Firstly, we suppose $X_0' = X_0$ and $X_{-1}' \neq (X_0)^{-1}$. Since $X_0'X_1'X_2'\ldots$ is reduced and $X_0' = X_0$ we know $X_0 \neq (X_1')^{-1}$. Thus $\mathbf{S}^*$ is reducible and the result is

$$\ldots\ldots X_{-3}'X_{-2}'X_{-1}'X_0X_1'X_2'X_3'\ldots\ldots \tag{2.12}$$

Also, $k = 0$ is the smallest index such that $X_{-k-1}' \neq (X_k')^{-1}$ and $X_0' = X_0$ and so (2.11) and (2.12) are identical. Similarly, if $X_0' \neq X_0$ and $X_{-1}' = (X_0)^{-1}$ then $\mathbf{S}^*$ is reducible to

$$\ldots\ldots X_{-4}'X_{-3}'X_{-2}'X_0^{-1}X_0'X_1'X_2'\ldots\ldots$$

which is identical to (2.11). Now suppose $X_0' = X_0$ and $X_{-1}' = (X_0)^{-1}$. We are assuming $\mathbf{S}$ is reducible to (2.11) and hence $k$ is the smallest non-negative integer such that $X_{-k-1}'$ is not the inverse of $X_k'$. Clearly $k \geq 1$ and so $j = k-1 \geq 0$ is the smallest non-negative integer such that $X_{-j-2}' \neq (X_{j+1}')^{-1}$. Thus $\mathbf{S}^*$ is reducible to

$$\ldots\ldots X_{-j-4}'X_{-j-3}'X_{-j-2}'X_{j+1}'X_{j+2}'X_{j+3}'\ldots\ldots$$

Again this sequence is identical to (2.11). A similar argument works in the remaining case. This completes our verification that the reduced sequence equivalent to $\mathbf{S}$ does not depend on the indexing of $\mathbf{S}$.

To summarise, we have demonstrated that a doubly infinite sequence $\mathbf{S}$ can be reduced by splitting it in half, reducing the two halves, gluing them back together and successively cancelling any neighbouring inverses at the join. In practice however, if we already know $\mathbf{S}$ is reducible, there is a better way to proceed. It is evident from the next theorem.

**Theorem 2.3.** *Let* $\mathbf{S} = \ldots X_{-1}X_0X_1\ldots$, *where each* $X_i \in \{A, B, A^{-1}, B^{-1}\}$, *be a reducible sequence. There is an increasing sequence of indices* $\{j(i)\}_{i=-\infty}^{+\infty}$ *such that*

(a) $X_{j(i)+1}X_{j(i)+2}\ldots X_{j(i+1)-1} = \mathrm{Id}$ *holds in* $\Gamma'$ *for all* $i$ *and*

(b) *the sequence* $\mathbf{S}' = \ldots X_{j(-1)}X_{j(0)}X_{j(1)}\ldots$ *is reduced.*

*Further, if* $\{j(i)\}_{i=-\infty}^{+\infty}$ *is any increasing sequence such that* (a) *and* (b) *hold then* $\mathbf{S}'$ *is the reduced sequence equivalent to* $\mathbf{S}$.

*Proof.* We begin with the fact that the sequence $X_0, X_1, X_2, \ldots$ is reducible. Thus we set $\mathbf{S}_n = X_0X_1\ldots X_n$ for $n \geq 0$, we let $\mathbf{S}'_n$ be a reduced word equivalent to $\mathbf{S}_n$ and we write

$$\lim_{n\to\infty} \mathbf{S}'_n = X'_0X'_1X'_2\ldots\ldots$$

Recall that $\mathbf{S}'_n$ can be obtained from $\mathbf{S}_n$ by successively cancelling neighbouring inverses. Since $X'_0$ is the initial term of some $\mathbf{S}'_n$ it follows that there is an index $j(0) \geq 0$ such that

$$X_0X_1\ldots X_{j(0)-1} = \mathrm{Id} \tag{2.13}$$

in $\Gamma'$ and $X_{j(0)} = X'_0$. Now fix $m \geq 0$ and suppose we have chosen an increasing sequence of positive indices $j(0), j(1), \ldots, j(m)$ such that for each $i$ with $0 \leq i \leq m-1$ we have

$$X_{j(i)+1}X_{j(i)+2}\ldots X_{j(i+1)-1} = \mathrm{Id} \tag{2.14}$$

and $X_{j(i+1)} = X'_{i+1}$. We know $X'_0X'_1\ldots X'_{m+1}$ is the initial segment of some $\mathbf{S}'_n$ and we can assume $n > j(m)$. Our choice of $j(0), j(1), \ldots, j(m)$ implies that $\mathbf{S}'_n$ is the reduction of the word

$$X'_0X'_1\ldots X'_mX_{j(m)+1}X_{j(m)+2}\ldots X_n.$$

Hence there is an index $j(m+1)$ with $j(m) < j(m+1) \leq n$ such that

$$X_{j(m)+1}X_{j(m)+2}\cdots X_{j(m+1)-1} = \mathrm{Id}$$

and $X_{j(m+1)} = X'_{m+1}$. It follows by induction on $m$ that there is an increasing sequence of indices $\{j(i)\}_{i=0}^{+\infty}$ such that the identities (2.14) and $X_{j(i)} = X'_i$ hold for all $i \geq 0$. We can also assume $j(0) \geq 0$ and (2.13) holds.

We know the sequence $X_{-1}, X_{-2}, X_{-3}, \ldots$ is also reducible. Thus we set $\mathbf{S}_{-n} = X_{-1}X_{-2}\ldots X_{-n}$ for $n \geq 1$, we let $\mathbf{S}'_{-n}$ be reduced word equivalent to $\mathbf{S}_{-n}$ and we write

$$\lim_{n\to\infty} \mathbf{S}'_{-n} = X'_{-1}X'_{-2}X'_{-3}\cdots\cdots$$

As above, there is a decreasing sequence of negative indices $\{j(-i)\}_{i=1}^{+\infty}$ with

$$X_{j(-i)-1}X_{j(-i)-2}\cdots X_{j(-i-1)+1} = \mathrm{Id} \tag{2.15}$$

and $X_{j(-i)} = X'_{-i}$ for all $i \geq 1$. Moreover, we can assume $j(-1) \leq -1$ and

$$X_{-1}X_{-2}\ldots X_{j(-1)+1} = \mathrm{Id}. \tag{2.16}$$

Since $\mathbf{S}$ is reducible there is some $k \geq 0$ such that $X'_{-k-1} \neq (X'_k)^{-1}$. We let $k$ be the smallest such index and note that the sequence

$$\mathbf{S}' = \cdots\cdots X'_{-k-3}X'_{-k-2}X'_{-k-1}X'_kX'_{k+1}X'_{k+2}\cdots\cdots$$

is the reduced sequence equivalent to $\mathbf{S}$. We claim that the sequence

$$\ldots\ldots,\ j(-k-3),\ j(-k-2),\ j(-k-1),\ j(k),\ j(k+1),\ j(k+2),\ \ldots\ldots \tag{2.17}$$

has the required properties. Clearly this sequence is increasing. It is also easy to see that condition (b) (with an appropriate re-indexing of (2.17)) is true. Condition (a) can be deduced from (2.14) and (2.15) except that we must also verify

$$X_{j(-k-1)+1}\ldots X_{-1}X_0X_1\ldots X_{j(k)-1} = \mathrm{Id}. \tag{2.18}$$

Using (2.13) and (2.14) with $0 \leq i \leq k-1$ and (2.15) with $1 \leq i \leq k$ and (2.16) we can deduce that (2.18) is equivalent to

$$X_{j(-k)} \dots X_{j(-2)} X_{j(-1)} X_{j(0)} X_{j(1)} \dots X_{j(k-1)} = \mathrm{Id}.$$

We can rewrite the latter as

$$X'_{-k} \dots X'_{-2} X'_{-1} X'_0 X'_1 \dots X'_{k-1} = \mathrm{Id}$$

and clearly this is true.

To see that the second statement of the theorem is true, let $\{j(i)\}_{i=-\infty}^{+\infty}$ be an increasing sequence of indices and suppose that (a) and (b) hold. Now define

$$\mathbf{S}_n = X_{j(0)} X_{j(0)+1} \dots X_{j(0)+n}$$

for all $n \geq 0$ and let $\mathbf{S}'_n$ be the reduced sequence equivalent to $\mathbf{S}_n$. Conditions (a) and (b) imply that if $n = j(i) - j(0)$ for some $i \geq 0$ then $\mathbf{S}'_n$ is the word $X_{j(0)} X_{j(1)} \dots X_{j(i)}$. We are assuming $\mathbf{S}$ is reducible. Therefore $\lim_{n\to\infty} \mathbf{S}'_n$ exists and so

$$\lim_{n\to\infty} \mathbf{S}'_n = \lim_{i\to\infty} \mathbf{S}'_{j(i)-j(0)} = X_{j(0)} X_{j(1)} X_{j(2)} \cdots\cdots$$

Similarly, we define

$$\mathbf{S}_{-n} = X_{j(0)-1} X_{j(0)-2} \dots X_{j(0)-n}$$

for all $n \geq 1$ and we let $\mathbf{S}'_{-n}$ be the reduced sequence equivalent to $\mathbf{S}_{-n}$. Again, the conditions (a) and (b) imply that if $n = j(0) - j(-i)$ for some $i \geq 1$ then $\mathbf{S}'_{-n}$ is the word $X_{j(-1)} X_{j(-2)} \dots X_{j(-i)}$. It follows that

$$\lim_{n\to\infty} \mathbf{S}'_{-n} = \lim_{i\to\infty} \mathbf{S}'_{j(-i)-j(0)} = X_{j(-1)} X_{j(-2)} X_{j(-3)} \cdots\cdots$$

Obviously $X_{j(-1)} \neq (X_{j(0)})^{-1}$ and hence the sequence $\mathbf{S}'$ described in (b) is the reduced sequence equivalent to $\mathbf{S}$. $\square$

**Remark 2.2.** Let $\mathbf{S} = \ldots X_{-1}X_0X_1\ldots$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$. The existence of an increasing sequence of indices $\{j(i)\}_{i=-\infty}^{+\infty}$ for which the conditions (a) and (b) of Theorem 2.3 hold does not guarantee the reducibility of $\mathbf{S}$. For instance, suppose

$$\mathbf{S} = \ldots\ldots X_{j(-2)}W_{-2}X_{j(-1)}W_{-1}X_{j(0)}W_0X_{j(1)}W_1X_{j(2)}\ldots\ldots$$

where each word $W_i$ is of the form

$$W_i = X_{j(i)}^{-1}\ldots X_{j(1)}^{-1}X_{j(0)}^{-1}X_{j(0)}X_{j(1)}\ldots X_{j(i)}$$

and suppose also that $\mathbf{S}' = \ldots X_{j(-1)}X_{j(0)}X_{j(1)}\ldots$ is reduced. By design, conditions (a) and (b) hold. Now set $\mathbf{S}_n = X_{j(0)}X_{j(0)+1}\ldots X_{j(0)+n}$ for each $n \geq 0$ and let $\mathbf{S}'_n$ be the reduced word equivalent to $\mathbf{S}_n$. Clearly, $\mathbf{S}_n$ reduces to the trivial word whenever $n = j(i) - j(0) + i + 1$ for some $i \geq 0$. It follows that $\lim_{n\to\infty}\mathbf{S}'_n$ cannot exist and hence $\mathbf{S}$ is not reducible.

### Automorphisms of $\Gamma'$ applied to reduced sequences

Let $\mathbf{S}$ be a reduced sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s and let $G$ be an automorphism of $\Gamma'$. We shall apply $G$ to $\mathbf{S}$ by applying a substitution associated with $G$ to $\mathbf{S}$ and then reducing the resulting sequence. A *substitution* for $\mathbf{S}$ is a map

$$A \to W_A, \qquad B \to W_B$$

where $W_A$ and $W_B$ are finite words in the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$. It is associated with $G$ if $G(A, B) = (W_A, W_B)$. The result of applying this substitution to $\mathbf{S}$ is the sequence obtained from $\mathbf{S}$ by replacing each of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ in it with the words $W_A$, $W_B$, $W_A^{-1}$ and $W_B^{-1}$, respectively. In this context, $W_A^{-1}$ and $W_B^{-1}$ denote the words obtained from $W_A$ and $W_B$ by reversing them and interchanging $A$ with $A^{-1}$ and $B$ with $B^{-1}$. In the following theorem we show that if a substitution is associated with $G$ then the sequence which results when it is applied to $\mathbf{S}$ is indeed reducible and further, we show that the equivalent reduced sequence depends only on $G$ and not the particular substitution used.

**Theorem 2.4.** *Let* $\mathbf{S} = \ldots X_{-1}X_0X_1\ldots$, *where each* $X_i \in \{A, B, A^{-1}, B^{-1}\}$, *be a reduced sequence. If the sequence* $\mathbf{S}'$ *is the result of applying the substitution*

$$A \to W_A, \qquad B \to W_B \tag{2.19}$$

*to* $\mathbf{S}$ *and if the homomorphism* $G(A, B) = (W_A, W_B)$ *lies in Aut* $\Gamma'$ *then* $\mathbf{S}'$ *is reducible. Further, if that is so then the reduced sequence equivalent to* $\mathbf{S}'$ *depends only on* $G$ *and not the particular representatives* $W_A$ *and* $W_B$ *for* $G(A)$ *and* $G(B)$.

*Proof.* We write $\mathbf{S}' = \ldots X'_{-1}X'_0X'_1\ldots$ where each $X'_i \in \{A, B, A^{-1}, B^{-1}\}$. Since $\mathbf{S}'$ is the result of applying the substitution (2.19) to $\mathbf{S}$ we can also write

$$\mathbf{S}' = \ldots\ldots W_{-1}W_0W_1\ldots\ldots$$

where each $W_i$ is the result of applying (2.19) to $X_i$. Note that in $\Gamma'$ the equality $W_i = G(X_i)$ holds. We choose the indexing of $\mathbf{S}'$ so that $X'_0$ is the first term of the word $W_0$.

First we shall show that the sequence $X'_0X'_1X'_2\ldots$ is reducible. Thus we set $\mathbf{S}'_j = X'_0X'_1\ldots X'_j$ for $j \geq 0$ and we let $\mathbf{S}''_j$ be the reduced word equivalent to $\mathbf{S}'_j$. We claim that the length of $\mathbf{S}''_j$ diverges to $\infty$ as $j$ increases. Suppose not, that is, suppose there is a bound below which the lengths of infinitely many $\mathbf{S}''_j$ lie. Then infinitely many of the words $\mathbf{S}''_j$ are identical and hence infinitely many of the words $\mathbf{S}'_j$ are equal in $\Gamma'$. It follows from our choice of notation that for each word $\mathbf{S}'_j$ there is an integer $n(j)$ such that

$$\mathbf{S}'_j = W_0W_1\ldots W_{n(j)}V_j$$

where $V_j$ is an initial segment of $W_{n(j)+1}$. Therefore, in $\Gamma'$ we have

$$G(X_0X_1\ldots X_{n(j)}) = W_0W_1\ldots W_{n(j)} = \mathbf{S}'_jV_j^{-1}.$$

There are only finitely many possibilities for the word $V_j^{-1}$ and so in $\Gamma'$ infinitely many of the words $\mathbf{S}'_jV_j^{-1}$ are equal. We conclude that there is $n_1 \neq n_2$ such that

$$G(X_0X_1\ldots X_{n_1}) = G(X_0X_1\ldots X_{n_2})$$

in $\Gamma'$. However, we are assuming $\mathbf{S}'$ is reduced and therefore the words $X_0X_1\dots X_{n_1}$ and $X_0X_1\dots X_{n_2}$ are not equal in $\Gamma'$ and further, since $G$ is an automorphism their images under $G$ are also not equal in $\Gamma'$. This contradiction implies our claim is true. Thus the length of $\mathbf{S}''_j$ diverges to $\infty$ as $j$ increases. Obviously, the sequence $\mathbf{S}''_{j+1}$ agrees with $\mathbf{S}''_j$ in all but perhaps the last place of $\mathbf{S}''_j$ and hence $\lim_{j\to\infty}\mathbf{S}''_j$ exists. We write

$$\lim_{j\to\infty}\mathbf{S}''_j = X''_0X''_1X''_2\dots\dots$$

where each $X''_i \in \{A, B, A^{-1}, B^{-1}\}$. We have shown that $X'_0, X'_1, X'_2, \dots$ is reducible to the sequence $X''_0X''_1X''_2\dots$.

Next we show that the sequence $X'_{-1}X'_{-2}X'_{-3}\dots$ is reducible. Thus we set $\mathbf{S}'_{-j} = X'_{-1}X'_{-2}\dots X'_{-j}$ for $j \geq 1$ and we let $\mathbf{S}''_{-j}$ be the reduced word equivalent to $\mathbf{S}'_{-j}$. This time

$$\mathbf{S}'_{-j} = \widetilde{W}_{-1}\widetilde{W}_{-2}\dots\widetilde{W}_{n(-j)}\,\widetilde{V}_{-j}$$

where $\widetilde{W}_i$ is the reverse of $W_i$ and $\widetilde{V}_{-j}$ is an initial segment of $\widetilde{W}_{n(-j)-1}$. We write

$$\widetilde{G}(A, B) = (\widetilde{W}_A, \widetilde{W}_B).$$

Clearly, $\widetilde{G}(A, B) = R^2GR^2 \in \text{Aut}\,\Gamma'$ where $R^2(A, B) = (A^{-1}, B^{-1})$. Also, in $\Gamma'$ we have

$$\widetilde{G}(X_{-1}X_{-2}\dots X_{n(-j)}) = \widetilde{W}_{-1}\widetilde{W}_{-2}\dots\widetilde{W}_{n(-j)} = \mathbf{S}'_{-j}\,(\widetilde{V}_{-j})^{-1}.$$

Hence an argument like that above implies $\lim_{j\to\infty}\mathbf{S}''_{-j}$ exists. We write

$$\lim_{j\to\infty}\mathbf{S}''_{-j} = X''_{-1}X''_{-2}X''_{-3}\dots\dots$$

where each $X''_i \in \{A, B, A^{-1}, B^{-1}\}$. We have shown that $X'_{-1}, X'_{-2}, X'_{-3}, \dots$ is reducible to $X''_{-1}X''_{-2}X''_{-3}\dots$.

To prove $\mathbf{S}'$ is reducible, it remains to show there is $k \geq 0$ such that $X''_{-k-1} \neq (X''_k)^{-1}$. To this end, we observe that for each $i \geq 0$ the sequence $X''_0X''_1\dots X''_i$ is the initial segment of some $\mathbf{S}''_j$. Since $\mathbf{S}''_j$ can be obtained from $\mathbf{S}'_j$ by successively cancelling neighbouring inverses it follows that $X''_0X''_1\dots X''_i$ is the reduction of some initial segment of $\mathbf{S}'_j$. Hence there is some $p(i) \geq 0$ such that

$\mathbf{S}''_{p(i)} = X''_0 X''_1 \dots X''_i$ As elements of $\Gamma'$ the words $\mathbf{S}'_{p(i)}$ and $X''_0 X''_1 \dots X''_i$ are equal and hence in $\Gamma'$ we have

$$G(X_0 X_1 \dots X_{n(p(i))}) = W_0 W_1 \dots W_{n(p(i))} = \mathbf{S}'_{p(i)} V_{p(i)}^{-1} = X''_0 X''_1 \dots X''_i V_{p(i)}^{-1}.$$

A similar argument shows there is some $q(i) \geq 1$ such that

$$\widetilde{G}(X_{-1} X_{-2} \dots X_{n(-q(i))}) = X''_{-1} X''_{-2} \dots X''_{-i-1} \widetilde{V}_{q(i)}^{-1}.$$

We let $V_{q(i)}$ be the reverse of $\widetilde{V}_{q(i)}$ and we rewrite this last equality as

$$G(X_{n(-q(i))} \dots X_{-2} X_{-1}) = V_{q(i)}^{-1} X''_{-i-1} \dots X''_{-2} X''_{-1}.$$

Now suppose that $X''_{-k-1} = (X''_k)^{-1}$ for all $k \geq 0$. Then for all $i \geq 0$ we have

$$G(X_{n(-q(i))} \dots X_{-1} X_0 X_1 \dots X_{n(p(i))}) = V_{q(i)}^{-1} V_{p(i)}^{-1}. \tag{2.20}$$

The indices $p(i)$ and $q(i)$ are increasing functions of $i$. Moreover, the words

$$X_{n(-q(i))} \dots X_{-1} X_0 X_1 \dots X_{n(p(i))}$$

and hence there images under $G$ represent distinct elements of $\Gamma'$. However, there are only finitely many possibilities for the word $V_{q(i)}^{-1} V_{p(i)}^{-1}$ in (2.20) and we have a contradiction. It follows that there is some $k \geq 0$ such that $X''_{-k-1} \neq (X''_k)^{-1}$ and hence $\mathbf{S}'$ is reducible. Note that, if $k \geq 0$ is the smallest integer such that $X''_{-k-1} \neq (X''_k)^{-1}$ then

$$\mathbf{S}'' = \dots\dots X''_{-k-3} X''_{-k-2} X''_{-k-1} X''_k X''_{k+1} X''_{k+2} \dots\dots$$

is the reduced sequence equivalent to $\mathbf{S}'$.

Finally, we shall show that $\mathbf{S}''$ depends only on $G$. For this purpose, we choose the sequence $r(0), r(1), r(2), \dots$ so that $\mathbf{S}'_{r(i)} = W_0 W_1 \dots W_i$ for all $i \geq 0$. Obviously $\mathbf{S}''_{r(i)}$ is the reduced word equal to $G(X_0 X_1 \dots X_i)$ in $\Gamma'$ and so $\mathbf{S}''_{r(i)}$ depends only on $G$ and not the representatives $W_A$ and $W_B$ of $G(A)$ and $G(B)$. Since $\mathbf{S}''_{r(0)}, \mathbf{S}''_{r(1)}, \mathbf{S}''_{r(2)}, \dots$ is a subsequence of $\mathbf{S}''_0, \mathbf{S}''_1, \mathbf{S}''_2, \dots$ we know that

$$X''_0 X''_1 X''_2 \dots\dots = \lim_{i \to \infty} \mathbf{S}''_{r(i)}$$

and it follows that $X_0''X_1''X_2''\ldots$ depends only on $G$. A similar argument shows that $X_{-1}''X_{-2}''X_{-3}''\ldots$ depends only on $\widetilde{G}$ and not the representatives $\widetilde{W}_A$ and $\widetilde{W}_B$ of $\widetilde{G}(A)$ and $\widetilde{G}(B)$. The truth of the theorem follows since $\widetilde{G}$ is determined by $G$. □

Given the truth of Theorem 2.4 we can now make the following definition.

**Definition 2.1.** Let $\mathbf{S}$ be a reduced doubly infinite sequence consisting of the symbols $A$, $B$, $A^{-1}$ and $B^{-1}$ and let $G \in \text{Aut } \Gamma'$. According to Theorem 2.4 there is a (unique) reduced sequence $\mathbf{S}'$ which is equivalent to all of the sequences obtained from $\mathbf{S}$ by applying to it a substitution of the form $A \to W_A$, $B \to W_B$ where $G(A,B) = (W_A, W_B)$. We say that $\mathbf{S}'$ is *the result of applying* $G$ to $\mathbf{S}$ and we write $G\,\mathbf{S} = \mathbf{S}'$.

We complete this section by establishing some results concerning the application of automorphisms in $\Gamma'$ to reduced sequences of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s. We shall make much use of Theorem 2.3 in doing this.

**Theorem 2.5.** *Let* $\mathbf{S}$ *be a reduced sequence of the symbols* $A$, $B$, $A^{-1}$ *and* $B^{-1}$ *and let* $G$ *and* $H$ *be automorphisms of* $\Gamma'$. *Then* $HG\,\mathbf{S} = H(G\,\mathbf{S})$.

*Proof.* We need some notation. We write $G(A,B) = (W_A, W_B)$ and we let $\mathbf{S}'$ be the sequence which results when the substitution

$$A \to W_A, \qquad B \to W_B$$

is applied to $\mathbf{S}$. By definition, $G\ \mathbf{S}$ is the reduced sequence equivalent to $\mathbf{S}'$. According to Theorem 2.3, if we write

$$G\,\mathbf{S} = \ldots\ldots X'_{-1}X'_0X'_1\ldots\ldots$$

where each $X_i$ lies in $\{A, B, A^{-1}, B^{-1}\}$ then $\mathbf{S}'$ is of the form

$$\mathbf{S}' = \ldots\ldots X'_{-2}W_{-2}X'_{-1}W_{-1}X'_0W_0X'_1W_1X'_2\ldots\ldots$$

where each $W_i$ is a consecutive subsequence of $\mathbf{S}'$ which reduces to the identity. Similarly, we write $H(A,B) = (V_A, V_B)$ and we let $\mathbf{S}''$ be the result of applying the

substitution

$$A \to V_A, \qquad B \to V_B \tag{2.21}$$

to the sequence $G\,\mathbf{S}$. As above, $H(G\,\mathbf{S})$ is the reduction of $\mathbf{S}''$ and thus if we write

$$H(G\,\mathbf{S}) = \ldots\ldots X''_{-1}X''_0X''_1\ldots\ldots$$

where each $X''_i$ lies in $\{A, B, A^{-1}, B^{-1}\}$ then

$$\mathbf{S}'' = \ldots\ldots X''_{-2}V_{-2}X''_{-1}V_{-1}X''_0V_0X''_1V_1X''_2\ldots\ldots$$

where each $V_i$ is a consecutive subsequence of $\mathbf{S}''$ which reduces to the identity.

Now let $U_A$ and $U_B$ be the result of applying the substitution (2.21) to $W_A$ and $W_B$, respectively. Clearly $HG(A, B) = (U_A, U_B)$ and hence we can calculate $HG\,\mathbf{S}$ by applying the substitution

$$A \to U_A, \qquad B \to U_B$$

to $\mathbf{S}$. We denote the resulting sequence by $\mathbf{S}'''$. Note that $\mathbf{S}'''$ is also the result of applying (2.21) to $\mathbf{S}'$. Since $\mathbf{S}''$ is the image of the subsequence $G\,\mathbf{S}$ of $\mathbf{S}'$ under (2.21) it follows that $\mathbf{S}'''$ can be obtained from $\mathbf{S}''$ by inserting in it the images of the words $W_i$ under (2.21). Each $W_i$ reduces to the identity and hence so does its image under (2.21). Thus $\mathbf{S}'''$ can be obtained from $\mathbf{S}''$ by inserting in it words which are equal to the identity in $\Gamma'$. We conclude that $\mathbf{S}'''$ is of the form

$$\mathbf{S}''' = \ldots\ldots X''_{-2}U_{-2}X''_{-1}U_{-1}X''_0U_0X''_1U_1X''_2\ldots\ldots$$

where each $U_i$ is a consecutive subsequence of $\mathbf{S}'''$ which is equal to the identity in $\Gamma'$. We know that the sequence $\ldots X''_{-1}X''_0X''_1\ldots$ is reduced and therefore Theorem 2.3 implies that it is the reduced sequence equivalent to $\mathbf{S}'''$. We have shown that

$$HG\,\mathbf{S} = \ldots\ldots X''_{-1}X''_0X''_1\ldots\ldots = H(G\,\mathbf{S})$$

and the proof is complete. □

**Theorem 2.6.** *Let* $\mathbf{S}$ *be a reduced sequence of the symbols* $A$, $B$, $A^{-1}$ *and* $B^{-1}$ *and let* $G$ *be an automorphism of* $\Gamma'$.

(a) *If* $G$ *is an inner automorphism of* $\Gamma'$ *then* $G\,\mathbf{S} = \mathbf{S}$.

(b) *If* $\mathbf{S}$ *is periodic with period* $W$ *then* $G\,\mathbf{S}$ *is periodic with period* $W'$ *where* $W'$ *is any cyclically reduced word representing the conjugacy class* $[G(W)]$.

*Proof.* We prove (a) first. Thus we assume $G \in \text{Inn}\,\Gamma'$ or equivalently

$$G(A, B) = (WAW^{-1}, WBW^{-1})$$

for some word $W$ consisting solely of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s. In order to calculate $G\,\mathbf{S}$ we write $\mathbf{S} = \ldots X_{-1}X_0X_1\ldots$ and we apply the substitution

$$A \to WAW^{-1}, \qquad B \to WBW^{-1}.$$

Obviously the result is the sequence

$$\ldots\ldots WX_{-1}W^{-1}WX_0W^{-1}WX_1W^{-1}\ldots\ldots$$

An easy application of Theorem 2.3 shows that this sequence reduces to $\mathbf{S}$ and therefore $G\,\mathbf{S} = \mathbf{S}$, as claimed.

Next we prove (b). Thus we assume $\mathbf{S}$ is of the form

$$\mathbf{S} = \ldots\ldots WWW\ldots\ldots$$

for some word $W$ and we let $W'$ be a cyclically reduced word in $[G(W)]$. As usual we write $G(A, B) = (W_A, W_B)$. By definition, $G\,\mathbf{S}$ is the reduction of the sequence

$$\mathbf{S}'' = \ldots\ldots W''W''W''\ldots\ldots$$

where $W''$ is the word obtained from $W$ by applying the substitution

$$A \to W_A, \qquad B \to W_B.$$

Since $W'' \in [G(W)]$ we know that it reduces to a word of the form $VW'V^{-1}$ for some $V$. Obviously, $W''$ may be obtained from $VW'V^{-1}$ by inserting in it words which reduce to the identity. It follows that $\mathbf{S}''$ may be obtained from the sequence

$$\ldots\ldots VW'V^{-1}VW'V^{-1}VW'V^{-1}\ldots\ldots$$

in a similar manner. Theorem 2.3 implies that this last sequence reduces to

$$\mathbf{S}' = \ldots\ldots W'W'W' \ldots\ldots$$

A further application of Theorem 2.3 shows that $\mathbf{S}''$ also reduces to $\mathbf{S}'$. We conclude that $G\,\mathbf{S} = \mathbf{S}'$ and the proof is complete. □

### Automorphisms in $\Psi$ applied to cutting sequences

We can now return to the question of how the cutting sequences of geodesics which are images of one another under the isometries in $\mathbf{T}$ are related. The answer is simply that they are images of one another under the automorphisms of $\Psi$. We can prove this by showing that if $\gamma$ is a geodesic in $\mathbf{H}$ and $T$ an element of $\Gamma^*$ then

$$\mathbf{S}(T(\gamma)) = G_T\,\mathbf{S}(\gamma) \tag{2.22}$$

where $G_T$ is the automorphism defined by (1.19). The truth of our statement follows from this since the projection $\sigma$ preserves cutting sequences and maps the action of $\Gamma^*$ to that of the isometries of $\mathbf{T}$ and the image of $\Gamma^*$ under the isomorphism (1.19) is $\Psi$. Note that, Theorem 2.6 shows (2.22) remains true if we replace $G_T$ in (2.22) by any automorphism in the coset $G_T$ Inn $\Gamma'$. Corresponding to this is the fact that cutting sequences are preserved under the action of $\Gamma'$ and hence we can also replace $T$ by any transformation in the coset $T\,\Gamma'$. We have of course identified $\Gamma^*/\Gamma'$ with the isometries of $\mathbf{T}$ and we remark that the isomorphism (1.11) induced by $\sigma$ maps the coset $G_T$ Inn $\Gamma'$ to an outer automorphism of $\pi_1(\mathbf{T})$.

Observe that if (2.22) is true for both $T = T_1$ and $T = T_2$ then we have

$$\mathbf{S}(T_1T_2(\gamma)) = G_{T_1}\,\mathbf{S}(T_2(\gamma)) = G_{T_1}G_{T_2}\,\mathbf{S}(\gamma) = G_{T_1T_2}\,\mathbf{S}(\gamma)$$

and hence (2.22) is true for the transformation $T_1T_2$. Therefore, to show (2.22) is true for all $T \in \Gamma^*$ we need only verify that it is true for the generators of $\Gamma^*$. We shall do this for the particular generators $T_1(z) = -\bar{z}$ and $U_1(z) = z + 1$ and $U_2(z) = -1/z$.

It is easy see (2.22) holds for the transformation $T_1$. The set of lines comprising the grid $\Lambda$ is preserved by $T_1$ and its effect on the labels is merely to interchange

$A$ with $B$ and $A^{-1}$ with $B^{-1}$. Thus, given any geodesic $\gamma$ in $\mathbf{H}$, we can obtain $\mathbf{S}(T_1(\gamma))$ from $\mathbf{S}(\gamma)$ by applying the automorphism $P(A,B) = (B,A)$. That is,

$$\mathbf{S}(T_1(\gamma)) = P\,\mathbf{S}(\gamma).$$

Since $G_{T_1} = P$ we find as claimed that (2.22) holds for $T = T_1$.

We can deal with the transformation $U_2$ in a similar manner. Again, $U_2$ preserves the set of lines comprising $\Lambda$. This time however, $U_2$ interchanges $A$ with $A^{-1}$ and $B$ with $B^{-1}$. Thus

$$\mathbf{S}(U_2(\gamma)) = R^2\,\mathbf{S}(\gamma)$$

where $R^2$ is the automorphism $R^2(A,B) = (A^{-1}, B^{-1})$. We know $G_{U_2} = R^2$ and so (2.22) holds when $T = U_2$.

In [36], Series states without proof that (2.22) is true when $T = U_1$. The proof is not trivial. We note that $G_{U_1} = SR^2$ and reformulate (2.22) with $T = U_1$ as the following theorem.

**Theorem 2.7.** *If $\gamma$ is any geodesic in $\mathbf{H}$ then $\mathbf{S}(U_1(\gamma)) = SR^2\,\mathbf{S}(\gamma)$ where $U_1(z) = z + 1$ and $SR^2(A,B) = (B^{-1}, AB)$.*

*Proof.* Let $\gamma$ be a geodesic in $\mathbf{H}$. We shall prove the theorem with the help of a new type of cutting sequence for $\gamma$. It is obtained by extending the grid $\Lambda$. We do this by first adding the line joining $O$ to $\infty$ and placing the label $E$ to its left and $E^{-1}$ to its right. We then use $\Gamma'$ to copy this labelled line to all the other tiles in the grid. We refer to the cutting sequence of $\gamma$ with respect to this new grid as its extended cutting sequence. Clearly, the extended cutting sequence of $\gamma$ can be obtained from $\mathbf{S}(\gamma)$ by inserting the label $E^{-1}$ between all pairs of the form $AB^{-1}$, $AA$, $B^{-1}B^{-1}$ and $B^{-1}A$ and the label $E$ between their inverses, namely, $BA^{-1}$, $A^{-1}A^{-1}$, $BB$ and $A^{-1}B$. Note that the resulting sequence is reduced.

The set of lines in our new grid partitions $\mathbf{H}$ into the Farey tessellation. We have noted in the section on $LR$-sequences that $U_1$ fixes this tessellation. It is not hard to verify that the effect of $U_1$ on the associated labelling is given by the substitution

$$A \to E^{-1}, \qquad B \to A, \qquad E \to B. \tag{2.23}$$

Of course, by definition this substitution also replaces the symbols $A^{-1}$, $B^{-1}$ and $E^{-1}$ by $E$, $A^{-1}$ and $B^{-1}$, respectively. To summarise, the extended cutting sequence of $U_1(\gamma)$ can be obtained from the extended cutting sequence of $\gamma$ by applying the substitution (2.23).

Now consider the effect of the substitution

$$A \to E^{-1}A, \qquad B \to BE \tag{2.24}$$

on $\mathbf{S}(\gamma)$. It inserts the label $E^{-1}$ between the pairs $AB^{-1}$, $AA$, $B^{-1}B^{-1}$ and $B^{-1}A$ and the label $E$ between their inverses. It also inserts $EE^{-1}$ between the pairs $BA$ and $A^{-1}B^{-1}$ but leaves the pairs $AB$ and $B^{-1}A^{-1}$ unaltered. This accounts for all the possible pairs in $\mathbf{S}(\gamma)$. Hence the extended cutting sequence of $\gamma$ can be obtained from its cutting sequence by applying the substitution (2.24) and then removing all occurrences of $EE^{-1}$ in the resulting sequence.

By composing the substitutions (2.24) and (2.23) we can deduce that the extended cutting sequence of $U_1(\gamma)$ is the result of applying the substitution

$$A \to B^{-1}E^{-1}, \qquad B \to AB$$

to $\mathbf{S}(\gamma)$ and then removing all occurrences of $BB^{-1}$. We already know the cutting sequence of $U_1(\gamma)$ can be obtained from its extended cutting sequence by removing all occurrences of $E$ and $E^{-1}$. It follows that $\mathbf{S}(U_1(\gamma))$ can be obtained from $\mathbf{S}(\gamma)$ by applying the substitution

$$A \to B^{-1}, \qquad B \to AB$$

and then removing all occurrences of $BB^{-1}$. The truth of the theorem is now evident. □

### Linear cutting sequences

We mentioned in Chapter 1 that the projection under Cohn's commutator map $\sigma_1$ of geodesics in $\mathbf{H}$ which correspond to the Markoff forms are precisely the geodesics in $\mathbf{P}$ which pass between the points of the lattice $\Sigma'(O)$ like those straight

lines in **P** which are parallel to the vectors joining $O$ to the points of $\Sigma'(O)$. (The meaning of this statement will be made precise in this section.) A natural extension of this result, see [21] and [36], is that the projection of the geodesics in **H** which correspond to forms with Markoff value equal to 3 are the geodesics in **P** which pass between the points of the lattice $\Sigma'(O)$ like the other straight lines in the plane **C**. The cutting sequences of such geodesics have special properties. In this section we discuss those properties. As we proceed we shall also review some of the related facts concerning the Markoff spectrum.

We begin by using Cohn's commutator map $\sigma_1$ to project the labelled grid $\Lambda$ to **P**. Recall that **P** is the plane **C** with the lattice of points $\Sigma'(O)$ removed and that $\sigma_1$ is the quotient map (1.26) which takes **H** to $\mathbf{P} = \mathbf{H}/\Gamma''$. Since the effect of $\sigma_1$ is merely to identify tiles in $\Lambda$ which are $\Gamma''$-equivalent and since the labellings of the tiles in $\Lambda$ agree under this identification the projection $\sigma_1(\Lambda)$ of $\Lambda$ is a well-defined labelled grid. It is illustrated in Figure 2.3. Note that $\sigma_1(\Lambda)$ may be obtained by using $\Sigma'$ to copy $\sigma_1(\mathcal{D})$ together with its labels to all of **P**. We define the *cutting sequence* $\mathbf{S}(\gamma)$ *of a geodesic* $\gamma$ in **P** (with respect to the grid $\sigma_1(\Lambda)$) in exactly the same way that we defined the cutting sequence of a geodesic in **H** (with respect to the grid $\Lambda$). Obviously cutting sequences of geodesics are preserved by the projection $\sigma_1$. Also, since $\sigma_1$ maps the action of $\Gamma'$ in **H** to that of $\Sigma'$ in **P**, cutting sequences of geodesics in **P** are preserved under the action of $\Sigma'$. In fact it is not hard to verify that the cutting sequences of two geodesics in **P** agree if and only if they are $\Sigma'$-equivalent.

We can also define cutting sequences for straight lines in **P**. In order that our definition and notation be consistent with that for geodesics we deal only with oriented lines. The *cutting sequence of an oriented straight line* $l$ in **P** is the sequence of labels

$$\mathbf{S}(l) = \ldots\ldots X_{-1}X_0X_1\ldots\ldots$$

which records its intersections with the grid $\sigma_1(\Lambda)$ with the usual convention that only the labels immediately after the grid lines are listed. See Figure 2.3. Again, the cutting sequences of two lines in **P** agree if and only if they are $\Sigma'$-equivalent.

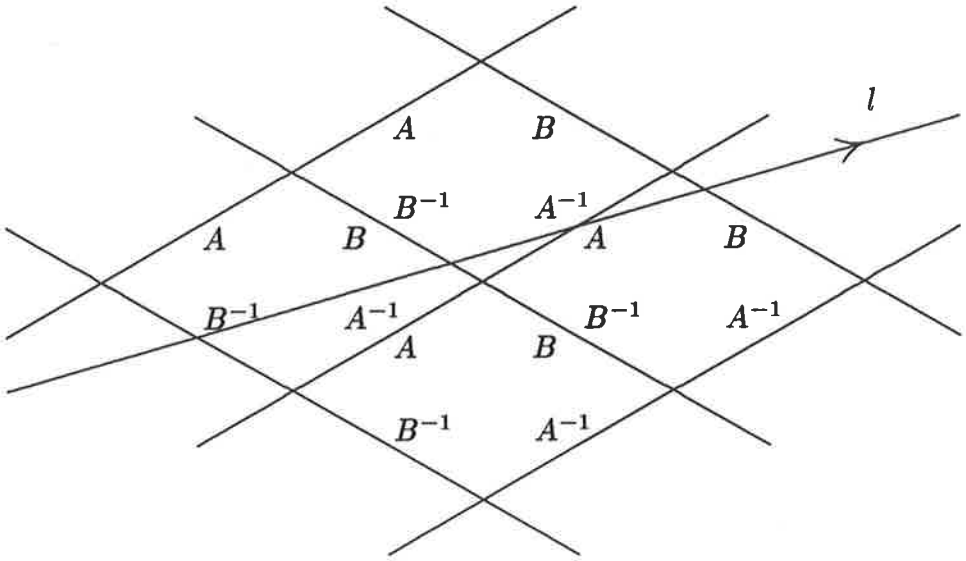


FIGURE 2.3. The projection $\sigma_1(\Lambda)$ of the labelled grid $\Lambda$ to $\mathbf{P}$. The cutting sequence of an oriented straight line $l$ in $\mathbf{P}$ is the sequence of labels $\mathbf{S}(l) = \ldots B^{-1}B^{-1}A\ldots$ which records its intersections with $\sigma_1(\Lambda)$. Again only the label following the grid line is listed in $\mathbf{S}(l)$.

Using the concept of cutting sequences for lines in $\mathbf{P}$ we can provide a rigorous formulation of the opening statement of this section. Specifically, a geodesic $\gamma$ in $\mathbf{H}$ corresponds to a form with Markoff value less than 3 if and only if $\mathbf{S}(\gamma) = \mathbf{S}(l)$ where $l$ is a line in $\mathbf{P}$ which is parallel to a vector joining $O$ to a point of $\Sigma'(O)$. We shall now outline how this result can be extended to include those geodesics which correspond to forms with Markoff value equal to 3. It is by far too large a task to provide the complete details here and our intention is only to convey the principles involved. We refer the reader to the work of Haas, [21], Series, [36] and Lunnon and Pleasants [27] for the full picture. Note that the grid which these authors use consists of all vertical and horizontal lines in the plane passing through points of the integer lattice $SL(2,\mathbf{Z})(O)$. For the purposes of studying the cutting sequences of lines, this grid and its associated labelling is equivalent to $\sigma_1(\Lambda)$ since there is an invertible linear transformation which maps it to $\sigma_1(\Lambda)$.

As a starting point, we use Haas' result that the geodesics in $\mathbf{H}$ which correspond

to forms with Markoff value less than or equal to 3 are exactly the limits of those with Markoff value less than 3. By definition, a geodesic $\gamma = [\eta, \xi]$ in $\mathbf{H}$ is the *limit* of a sequence $\{\gamma_i\}_{i=1}^{\infty}$ of geodesics if $\gamma_i = [\eta_i, \xi_i]$ and the sequences $\eta_i$ and $\xi_i$ converge to $\eta$ and $\xi$, respectively. Now suppose $\gamma$ is the limit of a sequence $\gamma_i$ of geodesics in $\mathbf{H}$ which correspond to forms with Markoff value less than 3 and let $l_i$ be a sequence of lines in $\mathbf{P}$ for which $\mathbf{S}(\gamma_i) = \mathbf{S}(\sigma_1(\gamma_i)) = \mathbf{S}(l_i)$ for all $i \geq 1$. By replacing each $l_i$ with the appropriate $\Sigma'$-translate we may assume each $l_i$ follows the same path through the lattice $\Sigma'(O)$ as the corresponding geodesic $\sigma_1(\gamma_i)$. Just as the sequence $\gamma_i$ converges to $\gamma$ in $\mathbf{H}$, the sequence $\mathbf{S}(\gamma_i)$ converges to $\mathbf{S}(\gamma)$. It follows that as $i$ increases, the paths which the geodesics $\sigma_1(\gamma_i)$ and hence the lines $l_i$ take through the lattice $\Sigma'(O)$ converge to that of $\sigma_1(\gamma)$. From this it is possible to deduce that the lines $l_i$ can be chosen so that they converge to some line $l$ in the plane $\mathbf{C}$. There are four possibilities.

**Case 1:** the line $l$ lies in $\mathbf{P}$. In this case, it is clear that $\mathbf{S}(\gamma) = \mathbf{S}(\sigma_1(\gamma)) = \mathbf{S}(l)$ where $\mathbf{S}(l)$ is as described above.

**Case 2:** the line $l$ passes through exactly one point $P$ of $\Sigma'(O)$. In this case, there are effectively two ways the lines $l_i$ can converge to $l$. They can approach $l$ from either above $P$ or below $P$. Hence $\mathbf{S}(\sigma_1(\gamma)) = \mathbf{S}(l)$ as long as we interpret $\mathbf{S}(l)$ to be the sequence obtained by distorting $l$ so that it passes above $P$ or below $P$, respectively.

**Case 3:** the line $l$ passes through infinitely many points of $\Sigma'(O)$ but is not a line in the grid $\sigma_1(\Lambda) \cup \Sigma'(O)$. In this case there are four ways the lines $l_i$ can converge to $l$. They can approach $l$ from above or below or by rotating towards it in either an anti-clockwise or a clockwise direction. In the first two instances $\mathbf{S}(\sigma_1(\gamma)) = \mathbf{S}(l')$ where $l'$ is a suitably small translation of $l$ up or down, respectively. In the second two instances $\mathbf{S}(\sigma_1(\gamma)) = \mathbf{S}(l)$ as long as we interpret $\mathbf{S}(l)$ to be the sequence obtained by distorting the right half of $l$ so that it passes above or below, respectively, the points of $\Sigma'(O)$ it contains and the left half so that it passes below or above, respectively, them.

**Case 4:** the line $l$ is a line in the grid $\sigma_1(\Lambda) \cup \Sigma'(O)$. This case is essentially

the same as the last one. The only difference is that in order to obtain $\mathbf{S}(l)$ when the lines $l_i$ converge to $l$ by rotation we need to distort $l$ everywhere rather than just at the points of $\Sigma'(O)$.

**Definition 2.2.** A doubly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is called *linear* if it is the cutting sequence $\mathbf{S}(l)$ of a line $l$ in the plane $\mathbf{C}$ where $\mathbf{S}(l)$ satisfies one of the interpretations described in the four cases above. (See also [27].)

We have just outlined why the goedesics in $\mathbf{H}$ which correspond to forms with Markoff value less than or equal to 3 have linear cutting sequences. The same argument shows that every linear sequence arises in this manner. Hence: *the Markoff value corresponding to a geodesic $\gamma$ on $\mathbf{T}$ is less or equal to 3 if and only if $\mathbf{S}(\gamma)$ is linear.* It is not hard to verify that a linear cutting sequence is periodic if and only if it is the cutting sequence of a line $l$ in $\mathbf{P}$ which is parallel to one of the vectors joining $O$ to a point in $\Sigma'(O)$. Thus we also have: *the Markoff value corresponding to a geodesic $\gamma$ on $\mathbf{T}$ is equal to 3 if and only if $\mathbf{S}(\gamma)$ is linear and aperiodic.* It is now clear how to phrase the second statement of the introduction to this section rigorously.

Of course, we know that the geodesics $\gamma$ in $\mathbf{H}$ which have Markoff value less than 3 are precisely those which cover the simple closed geodesics on $\mathbf{T}$. Thus: *a primitive closed geodesic on $\mathbf{T}$ is simple if and only if its cutting sequence is linear (and periodic).* Haas and Series point out that we also have: *if the cutting sequence of an open geodesic on $\mathbf{T}$ is aperiodic and linear then it is simple.* The reason for this is easy to see. Let $\gamma$ be a geodesic on $\mathbf{T}$ whose cutting sequence is aperiodic and linear. Clearly $\gamma$ is open. Suppose $\gamma$ is not simple and let $\tilde{\gamma}$ be a lift of $\gamma$ to $\mathbf{H}$. We know $\tilde{\gamma}$ is the limit of a sequence of geodesics $\tilde{\gamma}_i$ each of which has a periodic linear cutting sequence. We also know that each geodesic $\tilde{\gamma}_i$ covers a simple closed geodesic on $\mathbf{T}$. Our assumption that $\gamma$ is not simple implies there is some $T \in \Gamma'$ such that $T(\tilde{\gamma})$ crosses $\tilde{\gamma}$. Since the sequence $T(\tilde{\gamma}_i)$ converges to $T(\tilde{\gamma})$ we can choose $i$ so that $T(\tilde{\gamma}_i)$ crosses $\tilde{\gamma}_i$. This contradicts the fact that $\tilde{\gamma}_i$ covers a simple geodesic on $\mathbf{T}$ and it follows that $\gamma$ is simple. We shall see in Chapter 3 that the converse is not true. That is, not all simple open geodesics on $\mathbf{T}$ have

aperiodic linear cutting sequences.

The best account of linear cutting sequences is given by Lunnon and Pleasants, [27]. Series', [36], account is not intended to be rigorous and Haas, [21], does not discuss their properties. It follows from the work of Series and Lunnon and Pleasants that the cutting sequence of a line $l$ in $\mathbf{P}$ is either of the form $Y^\infty$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$ or it can be partitioned into blocks of the form $Y^n Z$ and $Y^{n+1} Z$ where $n$ is a positive integer and $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. (We remind the reader that in this context the abbreviation $Y^n$ denotes a sequence of $n$ consecutive $Y$'s and $Y^{n+1}$ is interpreted similarly.) The following definition is due to Series.

**Definition 2.3.** A doubly infinite sequence $\mathbf{S}$ of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is called *derivable* if it can be partitioned into the blocks $Y^n Z$ and $Y^{n+1} Z$, where $n \geq 1$ and $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. Moreover, if that is the case then the sequence $\mathbf{S}'$ obtained by applying the substitution

$$Y^n Z \to Z, \qquad Y^{n+1} Z \to YZ \tag{2.25}$$

to $\mathbf{S}$ is called the *derived sequence* of $\mathbf{S}$.

Of course (2.25) is applied to the sequence $\mathbf{S}$ by replacing each block of the form $Y^n Z$ with $Z$ and each block $Y^{n+1} Z$ with $YZ$. Note that $\mathbf{S}'$ is the result of applying the automorphism $G$ of $\Gamma'$ defined by

$$G(Y) = Y \qquad G(Z) = Y^{-n} Z \tag{2.26}$$

to $\mathbf{S}$. Series observes that this automorphism can be achieved by applying a linear transformation to $\mathbf{C}$ which preserves the lattice $\Sigma'(O)$. Specifically, there is a linear transformation $T$ of $\mathbf{P}$ to itself such that $\mathbf{S}(T(l)) = G\ \mathbf{S}(l)$ for all lines $l$ in $\mathbf{P}$. It follows that if $\mathbf{S}$ is the cutting sequence of a line in $\mathbf{P}$ then its derived sequence is also the cutting sequence of a line in $\mathbf{P}$. By induction we find that the cutting sequences of lines in $\mathbf{P}$ are either infinitely derivable or they are derivable to a sequence of the form $Y^\infty$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$.

The last result can be generalised to all linear sequences. While Series neglects this point, Lunnon and Pleasants do not. Using an intermediate result they show that every linear sequence is either infinitely derivable or it is derivable to a sequence of the form $Y^\infty$ or $Y^\infty ZY^\infty$ where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. One way of proving this directly is to view the linear sequences which are not the cutting sequences of lines in $\mathbf{P}$ as limits of those which are and then show that the property of being derivable is preserved under the operation of taking limits. The formal definition of such limits requires careful attention to the indexing of the sequences involved and it is not appropriate to enter into the details here. An alternative method would be to work directly with the distorted lines which generate the exceptional linear sequences. By developing the concept of the image of a distorted line under a linear transformation the arguments Series uses in the non-exceptional case could be adapted to the exceptional case.

The converse is also true. That is, any sequence $\mathbf{S}$ of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s which is either derivable infinitely often or derivable to a sequence of the form $Y^\infty$ or $Y^\infty ZY^\infty$ where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$ is linear. Series outlines a method of proving this and Lunnon and Pleasants provide the details. The idea here is to show that if a sequence $\mathbf{S}$ satisfies the hypothesis mentioned then every word $W$ contained in $\mathbf{S}$ occurs in the cutting sequence of some line. This may be done by deriving $W$ until it is of the form $X^n$ where $X \in \{A, B, A^{-1}, B^{-1}\}$, in which case it is clear that $W$ occurs in the cutting sequence of some line, and then reversing the derivation process. The result follows by choosing a sequence $W_i$ of words in $\mathbf{S}$ which converges to $\mathbf{S}$ and forming the limit of the associated lines $l_i$.

**Theorem 2.8.** *A doubly infinite sequence* $\mathbf{S}$ *of* $A$*'s,* $B$*'s,* $A^{-1}$*'s and* $B^{-1}$*'s is linear if and only if it is derivable infinitely often or it is derivable to, or is a sequence of the form* $Y^\infty$ *or* $Y^\infty ZY^\infty$ *where* $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ *and* $Z \neq Y^{\pm 1}$.

While linear sequences have many interesting properties, we only require Theorem 2.8. We complete this section with a few remarks on the geometrical aspects of Markoff's theory. Recall from Chapter 1 that the Markoff value associated with a geodesic in $\mathbf{H}$ is the supremum of the diameters of the geodesics in its $\Gamma'$-equivalence

class. Thus it is possible to phrase the qualitative aspects of Markoff's theory in completely geometrical terms. A general discussion of such matters may be found in Haas' papers, [20] and [21]. In [36], Series considers the particular statement: *a geodesic $\gamma$ in* $\mathbf{H}$ *is the limit of geodesics which cover simple closed geodesics on* $\mathbf{T}$ *if and only if the diameter of $\gamma$ and every $\Gamma'$-equivalent geodesic is less than or equal to 3.* She outlines a geometrical proof. Theorem 2.8 arises as an intermediate step. Although the statement is true, part of her work contains an error. She acknowledges the mistake in [37] and suggests a means of correcting it. The problem arises when she attempts to show that if $\mathbf{S}(\gamma)$ is not derivable infinitely often then some geodesic in the $\Gamma'$-equivalence class of $\gamma$ has diameter greater than 3. In effect, she assumes without proof that the latter property is preserved when the geodesic $\gamma$ is replaced by one whose cutting sequence is the derived sequence of $\mathbf{S}(\gamma)$. While this is true, its proof is by no means trivial.

### Half-linear cutting sequences

We know that of the primitive closed geodesics on $\mathbf{T}$, those which are simple, are characterised by the fact that their cutting sequences are linear. We can likewise characterise the simple open geodesics on $\mathbf{T}$. In order to do that we need to define a new class of cutting sequences, namely, the half-linear cutting sequences. We present the definition in this section and we study their properties. We require some preliminary material on rays first.

By a *ray* we mean any half-line in the plane $\mathbf{C}$ which does not contain its origin. We shall always orient a ray so that it faces away from its origin. We are particularly interested in those rays which lie in $\mathbf{P}$ and emanate from a point of $\Sigma'(O)$, we call them *$O$-rays.* (Recall from Chapter 1 that $\Sigma'$ is the image of $\Gamma'$ under the homomorphism $\pi$ defined by (1.25) and that $\mathbf{P}$ is the plane $\mathbf{C}$ with the lattice $\Sigma'(O)$ removed.) Just as we defined cutting sequences for lines in $\mathbf{P}$ with respect to the grid $\sigma_1(\Lambda)$, see Figure 2.3, we can define cutting sequences for $O$-rays. Thus the *cutting sequence* of an $O$-ray $r$ is the sequence of labels

$$\mathbf{S}(r) = X_1 X_2 X_3 \ldots\ldots,$$

where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, which records its intersections with the grid $\sigma_1(\Lambda)$, subject to the usual conventions. As with geodesics and lines, the cutting sequences of two $O$-rays are identical if and only if they are $\Sigma'$-equivalent. One consequence of this is that an $O$-ray which emanates from $O$ is uniquely determined by its cutting sequence. We shall often replace an $O$-ray by the $\Sigma'$-translate of it which emanates from $O$.

In the section of this chapter on reducibility for doubly infinite sequences we defined limits of the form $\mathbf{S} = \lim_{i\to\infty} \mathbf{S}_i$ where each $\mathbf{S}_i$ is a finite sequence and $\mathbf{S} = X_1X_2X_3\ldots$ is a singly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s. With minor modifications the definition is also valid in the case where each $\mathbf{S}_i$ like $\mathbf{S}$ is a singly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s. The resulting concept of limit is exactly that obtained by endowing the space of all such singly infinite sequences with its standard metric. Thus we can form limits of cutting sequences of $O$-rays.

**Definition 2.4.** A singly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is called *$O$-radial* if it is the limit of a sequence of cutting sequences of $O$-rays.

We establish some of the properties of $O$-radial sequences next. Our approach is motivated by our knowledge of linear sequences. In fact, we mention in this regard that linear sequences can be defined as the limits of the cutting sequences of lines in $\mathbf{P}$ in which case Definition 2.2 becomes in effect a theorem.

Let $\mathbf{S}$ be an $O$-radial sequence and let $\{r_i\}_{i=1}^{\infty}$ be a sequence of $O$-rays such that

$$\mathbf{S} = \lim_{i\to\infty} \mathbf{S}(r_i). \tag{2.27}$$

As mentioned above, we may assume each $r_i$ emanates from $O$. It is not hard to deduce from the convergence of their cutting sequences that the $O$-rays themselves converge in $\mathbf{C}$. Specifically, there is a ray $r$ in $\mathbf{C}$ which emanates from $O$ such that

$$r = \lim_{i\to\infty} r_i. \tag{2.28}$$

By this, we mean that the slopes of the rays $r_i$ converge to that of $r$. Note that $r$ depends only on the sequence $\mathbf{S}$ and not the particular choice of the rays $r_i$ (as long as they emanate from $O$). There are three possibilities.

**Case 1:** the ray $r$ lies in $\mathbf{P}$. In this case, $r$ is an $O$-ray and it is clear that $\mathbf{S}(r)$ is the $O$-radial sequence $\mathbf{S}$ we began with. Note that $\mathbf{S}(r)$ is aperiodic because if not, $r$ contains a $\Sigma'$-translate of itself and hence a point of $\Sigma'(O)$.

**Case 2:** the ray $r$ contains one and hence infinitely many points of $\Sigma'(O)$ but does not lie in the grid $\sigma_1(\Lambda) \cup \Sigma'(O)$. In this case there are two ways the rays $r_i$ can converge to $r$. They can approach $r$ in either an anti-clockwise or a clockwise direction. (The situation here is similar to that which arises for linear sequences in Case 3 of the previous section.) Thus $\mathbf{S} = \mathbf{S}(r)$ as long as we agree to interpret $\mathbf{S}(r)$ as the sequence obtained by distorting $r$ so that it passes to the right or to the left, respectively, of each of the points of $\Sigma'(O)$ it contains. We can be more specific about the form of $\mathbf{S}(r)$. When the points of $\Sigma'(O)$ are removed from $r$ we are left with an infinite sequence of segments, each of which lies in $\mathbf{P}$. Each segment is a $\Sigma'$-translate of the initial one and hence each segment has the same cutting sequence as the initial one. Let that cutting sequence be the word $W$. At each point $P$ of $\Sigma'(O)$ on $r$, the distortion of $r$ about $P$ crosses two grid lines at once. Thus each distortion contributes a pair of symbols $YZ$ to $\mathbf{S}$. Clearly $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. We conclude that $\mathbf{S}(r)$ is of the form $\mathbf{S}(r) = (WYZ)^{\infty}$. Note that the sequence $\mathbf{S}(r) = (WZY)^{\infty}$ is the $O$-radial sequence which results when the rays $r_i$ converge to $r$ from the opposite direction.

**Case 3:** the ray $r$ lies in $\sigma_1(\Lambda) \cup \Sigma'(O)$. As in Case 2, the rays $r_i$ can converge to $r$ from either an anti-clockwise or a clockwise direction. In either case the same sequence $bS(r)$ results. Clearly it is of the form $\mathbf{S}(r) = Y^{\infty}$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$.

To summarise, we have shown that every $O$-radial sequence can be interpreted, as described in the Cases 1 to 3 above, as the cutting sequence $\mathbf{S}(r)$ of some $O$-ray $r$ which emanates from $O$. Clearly, the converse is also true. That is, if $r$ is an $O$-ray emanating from $O$ and if $\mathbf{S}(r)$ is a cutting sequence obtained from $r$ according to the conventions in one of the Cases 1 to 3 above then $\mathbf{S}(r)$ is $O$-radial. Further properties of $O$-radial sequences may be established by adapting the techniques which Series, [36], and Lunnon and Pleasants, [27], use for linear sequences. In

particular, the application of linear transformations to $\mathbf{C}$ provides a useful tool for studying linear and hence $O$-radial sequences. However, the only property of $O$-radial sequences we shall need can be deduced directly from Theorem 2.8 with the help of the following lemma. In its proof we do use the easily verified fact that rotation through $\pi$ about $O$ fixes the lines in labelled grid $\sigma_1(\Lambda)$ and interchanges the labels $A$ and $A^{-1}$ and also the labels $B$ and $B^{-1}$.

**Lemma 2.2.** *A sequence* $\mathbf{S} = X_1X_2X_3\ldots$, *where each* $X_i \in \{A, B, A^{-1}, B^{-1}\}$, *is* $O$*-radial if and only if the sequence*

$$\ldots\ldots X_3X_2X_1\ YZ\ X_1X_2X_3\ldots\ldots \tag{2.29}$$

*is linear for some* $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ *with* $Z \neq Y^{\pm 1}$.

*Proof.* We prove the forward implication first. Let $\mathbf{S} = X_1X_2X_3\ldots$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, be an $O$-radial sequence. We know $\mathbf{S}$ arises as the cutting sequence of a ray $r$ as described in one of the cases above. If $\mathbf{S}$ arises as described in Case 3 the proof of the lemma is easy. In this case, $\mathbf{S} = Y^\infty$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$ and the result follows since each sequence of the form $Y^\infty ZY^\infty$ where $Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ is linear.

Suppose $\mathbf{S}$ arises as described in Case 1. Thus $\mathbf{S} = \mathbf{S}(r)$ where $r$ is an $O$-ray emanating from $O$. Let $l$ be the line in $\mathbf{C}$ which contains $r$ and has the same orientation as $r$. Since $l$ contains only the point $O$ of $\Sigma'(O)$ it is of the form described in Case 2 of the previous section. Thus either of the cutting sequences $\mathbf{S}(l)$ obtained by distorting $l$ so that it passes above or below $O$ are linear. Obviously $\mathbf{S} = \mathbf{S}(r)$ is the portion of $\mathbf{S}(l)$ arising as the cutting sequence of $r$. It is also easy to see that the portion of $\mathbf{S}(l)$ arising from the distortion at $O$ is a pair of symbols $YZ$, where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. Thus the lemma will be proved if we can show that the portion of $\mathbf{S}(l)$ not yet accounted for is the reverse of $\mathbf{S}$. To this end, we consider the line $l'$ obtained by rotating $l$ and its distortion through $\pi$ about $O$. Its cutting sequence can be obtained by reversing $\mathbf{S}(l)$ and interchanging $A$ and $B$ with $A^{-1}$ and $B^{-1}$, respectively. The result now follows since $l'$ is merely $l$ with its orientation reversed and the opposite distortion at $O$.

The only other possibility is that $\mathbf{S}$ arises as described in Case 2 above. In this case, $\mathbf{S} = \mathbf{S}(r)$ where $r$ is an $O$-ray emanating from $O$ which contains infinitely many points of $\Sigma'(O)$ but does not lie in $\sigma_1(\Lambda) \cup \Sigma'(O)$ and $\mathbf{S}(r)$ is the sequence obtained by distorting $r$ so that either it passes to the left of each point of $\Sigma'(O)$ on it or it passes to the right. We assume for the moment that the distortions are to the left. Again we let $l$ be the line in $\mathbf{C}$ which contains $r$ and has the same orientation as $r$. This time Case 3 of the previous section applies. Thus a linear sequence $\mathbf{S}(l)$ results when $l$ is distorted so that the half which contains $r$ passes to the left of each point of $\Sigma'(O)$ on it and the other half passes to the right of them. We allow either distortion at $O$. By design, $\mathbf{S} = \mathbf{S}(r)$ is the portion of $\mathbf{S}(l)$ arising from the ray $r$. Clearly, the portion of $\mathbf{S}(l)$ arising from the distortion at $O$ is a pair of symbols $YZ$, where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. Thus to complete the proof in this case we need only show that the remaining portion of $\mathbf{S}(l)$ is the reverse of $\mathbf{S}$. As before, this can be seen by noting that the effect of rotation through $\pi$ about $O$ on the line $l$ and its distortions is to reverse its orientation and interchange the distortions at $O$. A similar argument deals with the case where the distortions of $r$ are to the right.

To see that the converse is true, let $\mathbf{S}$ be as described and suppose there are $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ such that the sequence (2.29) is linear. By definition, the sequence (2.29) is the cutting sequence $\mathbf{S}(l)$ of a line $l$ in $\mathbf{C}$ where $\mathbf{S}(l)$ is interpreted as described in one of Cases 1 to 4 in the previous section. Since $Z \neq Y^{\pm 1}$ the pair $YZ$ in (2.29) are the result of the intersection of $l$ or some distortion of it with a pair of grid lines emanating from the same point in $\Sigma'(O)$. By replacing $l$ with the appropriate $\Sigma'$-translate we can assume that that point is $O$. Now consider the line $l'$ which results when $l$ and any distortions it may have are rotated through $\pi$ about $O$. As mentioned above, the cutting sequence of $l'$ can be obtained by reversing $\mathbf{S}(l)$ and interchanging $A$ with $A^{-1}$ and $B$ with $B^{-1}$. Thus $\mathbf{S}(l')$ is merely $\mathbf{S}(l)$ with the pair $ZY$ in place of $YZ$. Evidently the lines $l'$ and $l$, when distorted if applicable, follow the same path through the lattice $\Sigma'(O)$ except that $O$ lies between them. This can only happen if the undistorted lines $l$

and $l'$ both contain $O$ and opposing distortions are applied at $O$ to obtain their cutting sequences. Note in particular that the pair $YZ$ in $\mathbf{S}(l)$ arises from the distortion of $l$ at $O$ and the pair $ZY$ in $\mathbf{S}(l')$ arises from the distortion of $l'$ at $O$.

Now let $r$ be the ray contained in $l$ which emanates from $O$ and whose orientation agrees with that of $l$. Apply the same distortions to $r$ at the points of $\Sigma'(O)$ it contains as are applied to $l$. Thus $\mathbf{S}(r) = X_1X_2X_3\ldots = \mathbf{S}$. It remains to show that $\mathbf{S}(r)$ is $O$-radial. There are three possibilities. First suppose $r$ contains no other points of $\Sigma'(O)$. In this case, $r$ is an $O$-ray and the result is trivial. Next suppose $r$ contains infinitely many points of $\Sigma'(O)$ but does not lie in the grid $\sigma_1(\Lambda) \cup \Sigma'(O)$. According to Case 2 above, $\mathbf{S}(r)$ is $O$-radial unless $r$ is distorted to both the left and the right of points in $\Sigma'(O)$. Assume that the latter occurs. We know the distortions of $l$ and $l'$ at such points agree with those of $r$. Since $l'$ is the rotation of $l$, and it distortions, through $\pi$ about $O$ we can deduce that $l$ is distorted to both the left and the right of points in $\Sigma'(O)$ and that this happens on both sides of $O$. Clearly this contradicts the possibilities listed in Cases 1 to 4 of the previous section. Thus $\mathbf{S}(r)$ is $O$-radial. The third possibility is that $r$ lies in the grid $\sigma_1(\Lambda)$. In this case $r$ is as described in Case 3 above and an argument similar to that just completed shows $\mathbf{S}(r)$ is $O$-radial. □

**Remark 2.3.** We claim that every linear sequence which is not the cutting sequence of a line in $\mathbf{P}$ can be written in the form (2.29). This can be verified directly from the definition of linear sequences given in the previous section by considering the effect of rotation through $\pi$ about the appropriate point of $\Sigma'(O)$. We omit the details. In the same manner it may be seen that if the sequence (2.29) is linear then the sequence

$$\ldots\ldots X_3X_2X_1\ ZY\ X_1X_2X_3\ldots\ldots \tag{2.30}$$

is likewise linear. (This is also evident from the proof of Lemma 2.2.)

Our aim is to use Lemma 2.2 to obtain from Theorem 2.8 an analogous result for $O$-radial sequences. We begin with the analogous definitions.

**Definition 2.5.** A singly infinite sequence **S** of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is called *derivable* if it begins with a block of the form $Y^nZ$ and can be partitioned into the blocks $Y^nZ$ and $Y^{n+1}Z$, where $n \geq 1$ and $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. Moreover, if that is the case then we call the sequence $\mathbf{S}'$ obtained by applying the substitution

$$Y^nZ \to Z, \qquad Y^{n+1}Z \to YZ \tag{2.31}$$

to **S** the *derived sequence* of **S**.

Let $\mathbf{S} = X_1X_2X_3\ldots$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, and suppose $Y$ and $Z$ are elements of $\{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$. Observe that if **S** consists of $Y$'s and $Z$'s and is derivable then it can be partitioned into blocks either of the form $Y^nZ$ and $Y^{n+1}Z$ or of the form $Z^nY$ and $Z^{n+1}Y$ for some $n \geq 1$. In this case,

$$\ldots\ldots X_3X_2X_1\ YZ\ X_1X_2X_3\ldots\ldots \tag{2.32}$$

can be likewise partitioned and hence is also derivable. With care it can be verified that if $\mathbf{S}' = X_1'X_2'X_3'\ldots$ is the derived sequence of **S** then

$$\ldots\ldots X_3'X_2'X_1'\ YZ\ X_1'X_2'X_3'\ldots\ldots \tag{2.33}$$

is the derived sequence of (2.31). The converse is also true. That is, if (2.32) is derivable then **S** consists of $Y$'s and $Z$'s and is derivable and further, if (2.33) is the derived sequence of (2.31) then $\mathbf{S}' = X_1'X_2'X_3'\ldots$ is the derived sequence of **S**. The verification of this is straightforward (although not trivial) and we leave the details to the reader.

Now let $\mathbf{S} = X_1X_2X_3\ldots$, where each $X_i \in \{A, B, A^{-1}, B^{-1}\}$, and suppose **S** is $O$-radial. Lemma 2.2 implies (2.32) is linear for some $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ and hence we can apply Theorem 2.8. Thus either (2.32) is derivable infinitely often or it is derivable to, or is, a sequence of the form $Y^\infty$ or $Z^\infty$ or $Y^\infty ZY^\infty$ or $Z^\infty YZ^\infty$. In the first instance it follows by induction using the result just described that **S** is also derivable infinitely often. In the second instance

the sequences $Y^\infty$ and $Z^\infty$ are not possible and it follows by induction that **S** is derivable to, or is, a sequence of the form $Y^\infty$ or $Z^\infty$.

Conversely, suppose **S** is derivable infinitely often. In this case, **S** consists of $Y$'s and $Z$'s where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$. Induction using the result above shows (2.32) is derivable infinitely often. Thus Theorem 2.8 implies (2.32) is linear and Lemma 2.2 implies **S** is $O$-radial. Similarly, if **S** is derivable to, or is, a sequence of the form $Y^\infty$ or $Z^\infty$ then (2.32) is derivable to, or is, a sequence of the form $Y^\infty Z Y^\infty$ or $Z^\infty Y Z^\infty$. Again, Theorem 2.8 implies (2.32) is linear and Lemma 2.2 implies **S** is $O$-radial.

We have proved the following theorem.

**Theorem 2.9.** *A singly infinite sequence* **S** *of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is $O$-radial if and only if either it is derivable infinitely often or it is derivable to, or is, a sequence of the form $Y^\infty$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$.*

We are ready to define half-linear sequences.

**Definition 2.6.** We call a doubly infinite sequence **S** of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s *half-linear* if there are $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ such that $\mathbf{S} = (Y^{-1})^\infty Z Y^\infty$ or

$$\mathbf{S} = \ldots\ldots X_3^{-1} X_2^{-1} X_1^{-1}\ Z^{-1} Y^{-1} Z Y\ X_1 X_2 X_3 \ldots\ldots \tag{2.34}$$

where each $X_i$ is $Y$ or $Z$ and $X_1 X_2 X_3 \ldots$ is an $O$-radial sequence or

$$\mathbf{S} = \ldots\ldots X_{-3}^{-1} X_{-2}^{-1} X_{-1}^{-1}\ Z^{-1} Y^{-1} Z Y\ X_1 X_2 X_3 \ldots\ldots \tag{2.35}$$

where each $X_i$ is $Y$ or $Z$ and $X_1 X_2 X_3 \ldots$ and $X_{-1} X_{-2} X_{-3} \ldots$ are periodic $O$-radial sequences with periods of the form $WZY$ and $WYZ$, respectively, for some (possibly empty) word $W$.

We can characterise half-linear sequences in the same way that we have characterised linear sequences in Theorem 2.8 and $O$-radial sequences in Theorem 2.9.

**Definition 2.7.** We call a doubly infinite sequence **S** of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s *half-derivable* if we can choose $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ such

that **S** is of the form

$$\mathbf{S} = \ldots\ldots X_3^{-1}X_2^{-1}X_1^{-1}\, Z^{-1}Y^{-1}ZY\, X_1X_2X_3\ldots\ldots \tag{2.36}$$

where each $X_i$ is $Y$ or $Z$ and there is some $n \geq 1$ such that the sequences $X_1X_2X_3\ldots$ and $X_{-1}X_{-2}X_{-3}\ldots$ are both derivable by either the substitution

$$Y^nZ \to Z, \qquad Y^{n+1}Z \to YZ \tag{2.37}$$

or the substitution

$$Z^nY \to Y, \qquad Z^{n+1}Y \to ZY. \tag{2.38}$$

Further, if **S** is half-derivable and if $X_1'X_2'X_3'\ldots$ and $X_{-1}'X_{-2}'X_{-3}'\ldots$ are the associated derived sequences of $X_1X_2X_3\ldots$ and $X_{-1}X_{-2}X_{-3}\ldots$ then we call

$$\mathbf{S} = \ldots\ldots (X_3')^{-1}(X_2')^{-1}(X_1')^{-1}\, Z^{-1}Y^{-1}ZY\, X_1'X_2'X_3'\ldots\ldots \tag{2.39}$$

the *half-derived sequence* of **S**.

**Remark 2.4.** We observe for use in Chapter 3 that if **S** is a half-derivable sequence then its half-derived sequence **S′** may be obtained by applying an automorphism of $\Gamma'$ to **S**. In particular, if the sequence (2.36) is half-derivable to the sequence (2.39) by the the substitution (2.37) then (2.39) is the result of applying the automorphism $G$ defined by

$$G(Y) = Y \qquad \text{and} \qquad G(Z) = Y^{-n}Z \tag{2.40}$$

to (2.36). Similarly if (2.36) is half-derivable to (2.39) by (2.37) then applying the automorphism $G$ defined by

$$G(Y) = Z^{-n}Y \qquad \text{and} \qquad G(Z) = Z \tag{2.41}$$

to (2.36) yields (2.39).

**Theorem 2.10.** *A doubly infinite sequence* $\mathbf{S}$ *of* $A$*'s,* $B$*'s,* $A^{-1}$*'s and* $B^{-1}$*'s is half-linear if and only if it is half-derivable infinitely often or it is half-derivable to, or is, a sequence of the form* $(Y^{-1})^\infty\ Z^{-1}Y^{-1}Z\ Y^\infty$ *or* $(Z^{-1})^\infty\ Y^{-1}ZY\ Z^\infty$ *or* $(Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$ *or* $(Y^{-1})^\infty ZY^\infty$ *where the symbols* $Y$ *and* $Z$ *belong to the set* $\{A, B, A^{-1}, B^{-1}\}$ *and* $Z \neq Y^{\pm 1}$.

*Proof.* Suppose $\mathbf{S}$ is a half-linear sequence of the form (2.34). Since $X_1X_2X_3\ldots$ is $O$-radial and consists of $Y$'s and $Z$'s Theorem 2.9 implies that either it is derivable infinitely often or it is derivable to, or is, one of $Y^\infty$ or $Z^\infty$. It follows immediately from the definition that $\mathbf{S}$ is half-derivable infinitely often or it is half-derivable to, or is, one of

$$(2.42) \qquad (Y^{-1})^\infty\ Z^{-1}Y^{-1}ZY\ Y^\infty \qquad \text{or} \qquad (Z^{-1})^\infty\ Z^{-1}Y^{-1}ZY\ Z^\infty.$$

Conversely, suppose $\mathbf{S}$ is half-derivable infinitely often or it is half-derivable to one of (2.42). Since $\mathbf{S}$ is half-derivable it is of the form (2.36). Write $\mathbf{S}_1 = X_1X_2X_3\ldots$ and $\mathbf{S}_2 = X_{-1}X_{-2}X_{-3}\ldots$. If $\mathbf{S}$ is half-derivable infinitely often then $\mathbf{S}_1$ and $\mathbf{S}_2$ are derivable infinitely often by the same sequence of substitutions. In this case, every initial segment of $\mathbf{S}_2$ is an initial segment of $\mathbf{S}_2$ and *visa-versa* and hence $\mathbf{S}_1 = \mathbf{S}_2$. Further, Theorem 2.9 implies $\mathbf{S}_1 = \mathbf{S}_2$ is $O$-radial and thus $\mathbf{S}$ is a half-linear sequence of the form (2.34). If $\mathbf{S}$ is half-derivable to one of (2.42) then $\mathbf{S}_1$ and $\mathbf{S}_2$ are derivable by the same sequence of substitutions to one of $Y^\infty$ or $Z^\infty$. Obviously $\mathbf{S}_1 = \mathbf{S}_2$ and Theorem 2.9 implies they are $O$-radial. Again, $\mathbf{S}$ is a half-linear sequence of the form (2.34).

Now suppose $\mathbf{S}$ is a half-linear sequence of the form (2.35) and write $\mathbf{S}_1 = X_1X_2X_3\ldots$ and $\mathbf{S}_2 = X_{-1}X_{-2}X_{-3}\ldots$. We shall show that $\mathbf{S}$ is half-derivable to, or is, the sequence $(Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$. We are assuming that $\mathbf{S}_1 = (WZY)^\infty$ and $\mathbf{S}_2 = (WYZ)^\infty$ for some word $W$. If $W$ is empty then $\mathbf{S} = (Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$ and we are done. If $W$ is $Y^n$ or $Z^n$ for some $n \geq 1$ then $\mathbf{S}$ is derivable by (2.37) or (2.38), respectively, to the sequence $\mathbf{S} = (Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$ and again we are done. Thus we can assume $W$ begins with $Y^nZ$ or $Z^nY$ where $n \geq 1$. We deal with the case where $W$ begins with $Y^nZ$ first. In this case $\mathbf{S}_1$ and $\mathbf{S}_2$ both begin

with $Y^nZ$. We know $\mathbf{S}_1$ and $\mathbf{S}_2$ are derivable since they are $O$-radial. Clearly they are both derivable by the substitution (2.37). Hence both $WZ$ and $WYZ$ can be partitioned into the blocks $Y^nZ$ and $Y^{n+1}Z$. This is only possible if $W$ is of the form $W = W_1Y^n$ where $W_1$ can be partitioned into the blocks $Y^nZ$ and $Y^{n+1}Z$. Therefore the derived sequences of $\mathbf{S}_1$ and $\mathbf{S}_2$ are $\mathbf{S}_1' = (VZY)^\infty$ and $\mathbf{S}_2' = (VYZ)^\infty$, respectively, where $V$ is the result of applying (2.37) to $W_1$. In the case where $W$ begins with $Z^nY$ a similar argument shows $\mathbf{S}_1$ and $\mathbf{S}_2$ are both derivable by the substitution (2.38) and that their respective derived sequences are likewise of the form $\mathbf{S}_1' = (VZY)^\infty$ and $\mathbf{S}_2' = (VYZ)^\infty$ for some $V$. It is easily deduced from Theorem 2.9 that the derived sequence of an $O$-radial sequence is also $O$-radial. It follows that, in either case $\mathbf{S}_1'$ and $\mathbf{S}_2'$ are $O$-radial and hence $\mathbf{S}$ is half-derivable and its half-derived sequence $\mathbf{S}'$ is also a half-linear sequence of the form (2.35). We can now repeat the argument with $\mathbf{S}'$ in place of $\mathbf{S}$ and so on. Eventually, the situation where $W$ is $Y^n$ or $Z^n$ will arise. We conclude, as was claimed, that $\mathbf{S}$ is half-derivable to, or is, $(Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$.

Apart from some trivial details which we leave to the reader, we can complete the proof by showing that if $\mathbf{S}$ is half-derivable to the sequence $(Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$, where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$, then $\mathbf{S}$ is a half-linear sequence of the form (2.35). Clearly the sequence $(Z^{-1}Y^{-1})^\infty\ (ZY)^\infty$ itself is of this form and so the result will follow by induction if we can show that $\mathbf{S}$ is a half-linear sequence of the form (2.35) whenever its half-derived sequence $\mathbf{S}'$ is. In other words it suffices to show that if $\mathbf{S}_1$ and $\mathbf{S}_2$ are derivable by the same substitution to say $\mathbf{S}_1'$ and $\mathbf{S}_2'$ and if $\mathbf{S}_1'$ and $\mathbf{S}_2'$ are periodic $O$-radial sequences with periods of the form $VZY$ and $VYZ$, respectively, for some $V$ then $\mathbf{S}_1$ and $\mathbf{S}_2$ are also periodic $O$-radial sequences and have periods of the form $WZY$ and $WYZ$, respectively, for some $W$. Suppose the hypothesis of this last statement is true. As mentioned above, it is an easy consequence of Theorem 2.9 that $\mathbf{S}_1$ and $\mathbf{S}_2$ are $O$-radial. It is also easy to see that they are periodic. To see that their periods are of the form claimed we note that the substitution involved is either (2.37) or (2.38) and we consider the inverse operations. The inverse of the substitution (2.37) is the operation which inserts

$Y^n$ in front of every $Z$. In this case the periods of $\mathbf{S}_1$ and $\mathbf{S}_2$ are $V'Y^nZY$ and $V'Y^nYZ$, respectively, where $V'$ is obtained from $V$ by inserting $Y^n$ in front of every $Z$ and we are done. In the other case, a similar argument shows the periods of $\mathbf{S}_1$ and $\mathbf{S}_2$ are $V'Z^nZY$ and $V'Z^nYZ$, respectively, where $V'$ is obtained from $V$ by inserting $Z^n$ in front of every $Y$ and the proof is complete. □

**Remark 2.5.** We conclude this section with the following easy consequence of Theorem 2.10. If a doubly infinite sequence $\mathbf{S}$ of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s is half-derivable to a sequence $\mathbf{S}'$ then $\mathbf{S}$ is half-linear if and only if $\mathbf{S}'$ is.

CHAPTER 3

# GEODESICS ON T WITH LOW SELF–INTERSECTION NUMBER

This chapter is divided into two parts. In the first part we characterise the *closed 1-intersectors* on $\mathbf{T}$. (Recall that by a closed 1-intersector we mean a closed geodesic on $\mathbf{T}$ with one self-intersection.) In the second we characterise the simple open geodesics. The closed 1-intersectors are characterised in terms of the conjugacy classes in $\Gamma'$ which define them whereas the simple open geodesics are characterised in terms of their cutting sequences. Our methods reflect this difference.

By definition, a closed 1-intersector is a closed geodesic on $\mathbf{T}$ whose parametrisations as a closed curve all have exactly one self-intersection. Thus we can begin the characterisation of the closed 1-intersectors by studying those free homotopy classes on $\mathbf{T}$ which contain loops with one self-intersection. We know closed geodesics realise the minimum number of self-intersections of all the loops in their free homotopy classes and therefore we can restrict our attention to those classes which contain a loop with one self-intersection but no simple loops. We say that a loop on $\mathbf{T}$ has a *non-trivial single self-intersection* if it has a single self-intersection and is not freely homotopic to a simple loop. To summarise then, we are interested in those free homotopy classes on $\mathbf{T}$ which contain a loop with a non-trivial single self-intersection. We have identified the free homotopy classes on $\mathbf{T}$ with the conjugacy classes of the fundamental group $\pi_1(\mathbf{T})$. In our first theorem, Theorem 3.1, we describe the conjugacy classes on $\mathbf{T}$ which contain a loop with a non-trivial single self-intersection. From the description it is easy to identify those free homotopy classes which also contain closed 1-intersectors. By using the isomorphism $\theta$ defined in (1.11) we can then convert this description into one in terms of conjugacy classes in $\Gamma'$.

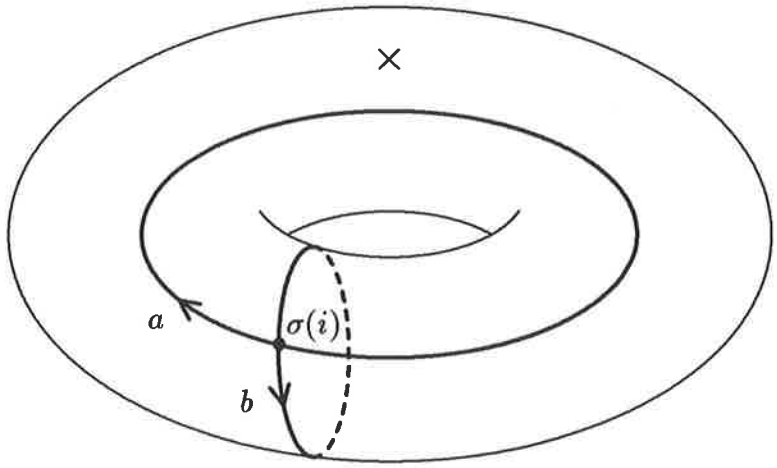


FIGURE 3.1. The punctured torus $\mathbf{T}$ together with the generators $a$ and $b$ of $\pi_1(\mathbf{T})$. The puncture is marked with a cross.

It is convenient to remind the reader here that in Chapter 1 we chose the base point of $\pi_1(\mathbf{T})$ to be the image of $i$ under the projection (1.10). We also chose the generators $a$ and $b$ of $\pi_1(\mathbf{T})$ to be the projection of the fundamental segments of the axes of $A$ and $B$, respectively, as shown in Figure 1.2 so that $a = \theta(A)$ and $b = \theta(B)$. Loops on $\mathbf{T}$ which represent $a$ and $b$ are shown in Figure 3.1.

In order to prove Theorem 3.1 we require the solution to the analogous problem for simple loops, see Birman and Series [3]. They show that the conjugacy class of a simple loop $l$ on $\mathbf{T}$ is either

(1) the identity and $l$ bounds a disc or

(2) one of $[aba^{-1}b^{-1}]$ or $[bab^{-1}a^{-1}]$ and $l$ bounds a punctured disc or

(3) $[w]$ where $w$ is a generator of $\pi_1(\mathbf{T})$ and $l$ does not separate $\mathbf{T}$.

We shall also use the fact that every automorphism of $\pi_1(\mathbf{T})$ can be induced by a homeomorphism of $\mathbf{T}$. One consequence of this last fact is that any automorphism of $\pi_1(\mathbf{T})$ at most interchanges the conjugacy classes $[aba^{-1}b^{-1}]$ and $[bab^{-1}a^{-1}]$ and likewise the conjugacy classes $[(aba^{-1}b^{-1})^2]$ and $[(bab^{-1}a^{-1})^2]$.

**Theorem 3.1.** *The conjugacy class in $\pi_1(\mathbf{T})$ of a loop on $\mathbf{T}$ with a non-trivial single self-intersection is either*

(a) $[(aba^{-1}b^{-1})^2]$ *or* $[(bab^{-1}a^{-1})^2]$ *or*

(b) $[g(a^2)]$ *or* $[g(abab^{-1})]$ *or* $[g(aaba^{-1}b^{-1})]$ *for some* $g \in Aut\ \pi_1(\mathbf{T})$.

*Conversely, each of these conjugacy classes contains such a loop.*

*Proof.* Let $l$ be a loop on $\mathbf{T}$ with a non-trivial single self-intersection. By using a free homotopy which preserves the point set covered by $l$ and the intersection property of $l$, we can assume without loss of generality that the origin of $l$ is its point of self-intersection. There are two simple loops $l_1, l_2$ determined by $l = l_1 l_2$ and two possibilities for their relative orientations. Either $l_1$ and $l_2$ are exactly as shown in Figure 3.2 or they are as shown in Figure 3.2 but with the orientation of $l_2$ reversed. In the former case, we say that the intersection is *transverse*. In the latter case, $l$ is homotopic to a simple loop and the self-intersection is trivial. By hypothesis then, the intersection is transverse.

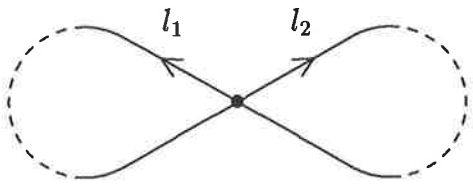


FIGURE 3.2. The simple loops $l_1$ and $l_2$ determined by $l = l_1 l_2$.

For each pair of points on $\mathbf{T}$ there is an isotopy of $\mathbf{T}$ which maps one to the other. By using such an isotopy if necessary we can assume that the point of self-intersection of $l$ is the base point $\sigma(i)$ of $\pi_1(\mathbf{T})$. There is no loss of generality in doing this since isotopies are homeomorphisms and therefore preserve free homotopy classes. Now let $w, w_1, w_2$ be the homotopy classes of $l, l_1, l_2$ respectively. The corresponding free homotopy classes are $[w]$, $[w_1]$ and $[w_2]$. Note that $l = l_1 l_2$ and therefore $w = w_1 w_2$. Since $l_1$ and $l_2$ are simple loops we know the possibilities for $[w_1]$ and $[w_2]$. Not all of them need to be considered in detail.

**Case 1:** $[w_1] = [\mathrm{Id}]$.

Obviously $w_1 = \mathrm{Id}$ and so $w = w_2$. Thus $l$ is homotopic to $l_2$ and the self-intersection is trivial. Case 1 does not arise.

**Case 2:** $[w_1] = [aba^{-1}b^{-1}] = [w_2]$.

Here both $l_1$ and $l_2$ bound discs on $\mathbf{T}$ containing the puncture. If $l_2$ does not lie entirely inside the disc bounded by $l_1$ then it must lie entirely outside, bounding a disc containing both $l_1$ and the puncture. In either event, $l_1$ is homotopic to $l_2$ or $l_2^{-1}$. Only the first possibility occurs since the intersection in $l$ is transverse. Thus $w_1 = w_2$ and $[w] = [w_1^2] = [(aba^{-1}b^{-1})^2]$.

**Case 3:** $[w_1] = [aba^{-1}b^{-1}]$ and $[w_2] = [bab^{-1}a^{-1}]$.

By the argument of Case 2 we have again $w_1 = w_2$. This is impossible however since $[w_1] \neq [w_2]$. Clearly Case 3 does not arise.

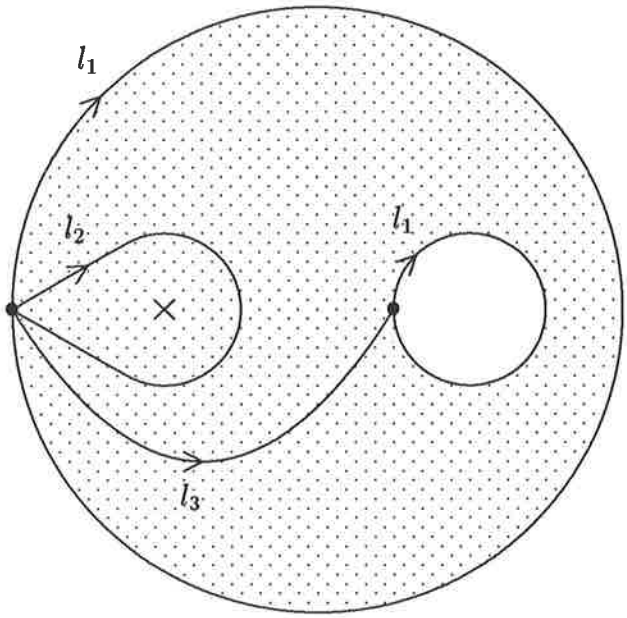


FIGURE 3.3. The torus $\mathbf{T}$ disected along the loop $l_1$ in Case 4.

**Case 4:** $[w_1] = [\text{generator}]$ and $[w_2] = [aba^{-1}b^{-1}]$.

The loop $l_1$ does not separate $\mathbf{T}$. Cut $\mathbf{T}$ along $l_1$ to obtain a disc bounded by $l_1$ containing the puncture and a hole also bounded by $l_1$. On this surface $l_2$ bounds a disc containing the puncture, see Figure 3.3. Note that the loops have the relative orientations indicated since $\mathbf{T}$ is orientable and the self-intersection is transverse. The path $l_3$ shown in Figure 3.3 projects to a simple non-separating loop on $\mathbf{T}$. Let $w_3$ denote its homotopy class. By cutting along $l_3$ it is also not hard to deduce that $w_1$ and $w_3$ are in fact a generating pair for $\pi_1(\mathbf{T})$. That is, there

is an automorphism $g$ with $g(a) = w_1$ and $g(b) = w_3$. We remark here that the homeomorphism of $\mathbf{T}$ which induces $g$ is given by the natural identification of the 'rectangle' obtained by disecting $\mathbf{T}$ along $a$ and $b$ with that obtained by disecting $\mathbf{T}$ along $w_1$ and $w_3$. It is evident from Figure 3.3 that $l_2$ is homotopic to $l_1 l_3 l_1^{-1} l_3^{-1}$ and so $w_2 = w_1 w_3 w_1^{-1} w_3^{-1}$. Thus $[w] = [w_1 w_2] = [w_1 w_1 w_3 w_1^{-1} w_3^{-1}] = [g(aaba^{-1}b^{-1})]$.

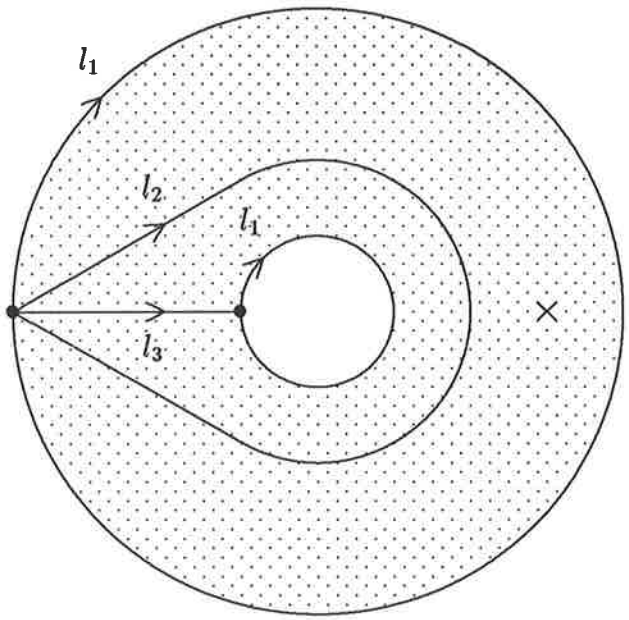


FIGURE 3.4. The torus $\mathbf{T}$ disected along the loop $l_1$ in Case 5.

**Case 5:** $[w_1]$ = [generator] and $[w_2]$ = [generator].

Again cut $\mathbf{T}$ along $l_1$ to obtain a disc bounded by $l_1$ which contains both the puncture and a hole also bounded by $l_1$. Clearly $l_2$ is a simple loop on this surface and since it does not separate $\mathbf{T}$ it must bound a disc containing the hole. There are two possibilities for the location of the puncture. The first is shown in Figure 3.4. In this case, the puncture lies 'outside' $l_2$. As before, the path $l_3$ shown projects to a loop with homotopy class $w_3$ and there is an automorphism $g$ with $g(a) = w_1$ and $g(b) = w_3$. This time $l_2$ is homotopic to $l_3 l_1 l_3^{-1}$ and so $w_2 = w_3 w_1 w_3^{-1}$ and $[w] = [w_1 w_2] = [w_1 w_3 w_1 w_3^{-1}] = [g(abab^{-1})]$. The other possibility is that the puncture lies 'inside' $l_2$. In this case, $l_1$ is homotopic to $l_2$ and so $w_1 = w_2$ and $[w] = [w_1^2] = [g(a)^2] = [g(a^2)]$ for some automorphism $g$.

All remaining cases can be expressed in terms of these first five as follows. Since $l = l_1 l_2$ is freely homotopic to $l_2 l_1$ we can interchange the roles of $l_1$ and $l_2$ without changing the conjugacy class of $w$. This leaves only the case where $[w_1] = [bab^{-1}a^{-1}] = [w_2]$ and the case where $[w_1] =$ [generator] and $[w_2] = [bab^{-1}a^{-1}]$ to consider. For these we apply the homeomorphism of $\mathbf{T}$ which induces the automorphism $h$ defined by $h(a) = b$ and $h(b) = a$ to obtain Case 2 and Case 4, respectively. This gives us the conjugacy class of the image of $l$ under the homeomorphism. Applying the inverse automorphism yields the conjugacy class of the original loop $l$, namely, $[(bab^{-1}a^{-1})^2]$ in the first case and $[g'(aaba^{-1}b^{-1})]$ where $g' = hg$ in the second.

To complete the proof it remains to demonstrate loops with non-trivial single self-intersections for each of the listed conjugacy classes. We actually do this only for the classes $[a^2]$, $[abab^{-1}]$, $[aaba^{-1}b^{-1}]$ and $[(aba^{-1}b^{-1})^2]$ and note that the rest follow since every automorphism of $\pi_1(\mathbf{T})$ can be induced by a homeomorphism of $\mathbf{T}$. The property of generators of $\pi_1(\mathbf{T})$ given in Theorem 5.1 of [3] shows that none of these classes contain a simple loop. It suffices then to merely find a single self-intersection loop for each of the classes mentioned. Figures 3.5 and 3.6 show such loops for the classes $[abab^{-1}]$ and $[aaba^{-1}b^{-1}]$. By considering simple loops for $[a]$ and $[aba^{-1}b^{-1}]$ it is not hard to demonstrate single self-intersection loops for $[a^2]$ and $[(aba^{-1}b^{-1})^2]$. The proof is complete. □

We can now consider the closed 1-intersectors. It is clear from the discussion in Chapter 1 on self-intersection numbers that closed 1-intersectors are primitive. Now recall that primitive closed geodesics realise the minimum number of self-intersections of all the loops in their free homotopy classes. It follows in particular that the intersection of a closed 1-intersector is non-trivial. Therefore we can apply Theorem 3.1. By precluding the non-primitive classes, we conclude that the conjugacy classes in $\pi_1(\mathbf{T})$ of the closed 1-intersectors on $\mathbf{T}$ are all of the form $[g(abab^{-1})]$ or $[g(aaba^{-1}b^{-1})]$ where $g \in \text{Aut } \pi_1(\mathbf{T})$. We can express this statement in terms of $\Gamma'$ rather than $\pi_1(\mathbf{T})$ by using the isomorphism $\theta$ defined in (1.11). Recall that $\theta$ maps each hyperbolic conjugacy class in $\Gamma'$ to the free homotopy class of the closed

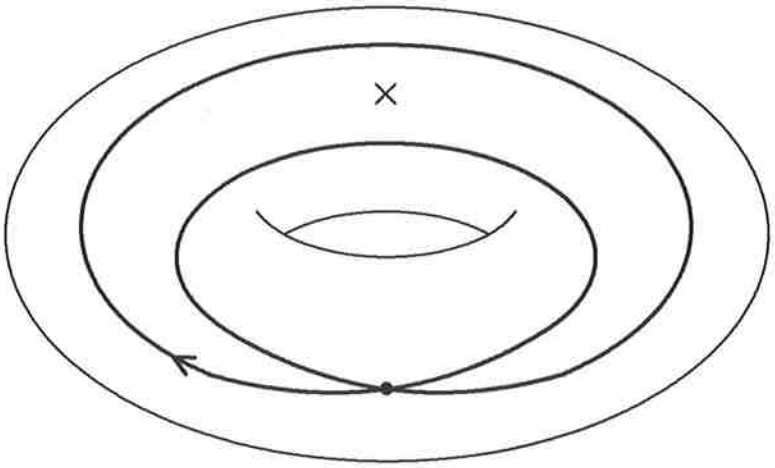

FIGURE 3.5. The geodesic on $\mathbf{T}$ in the free homotopy class $[abab^{-1}]$.

geodesic on $\mathbf{T}$ which it defines. It follows that each conjugacy class in $\Gamma'$ which defines a closed 1-intersector is of the form $[G(ABAB^{-1})]$ or $[G(AABA^{-1}B^{-1})]$ for some $G \in \text{Aut } \Gamma'$. This in fact characterises the closed 1-intersectors. To see why, observe first that all such classes are hyperbolic. (The only non-hyperbolic classes of $\Gamma'$ are of the form $[(ABA^{-1}B^{-1})^n]$ where $n$ is a non-zero integer.) Thus the corresponding free homotopy classes $[g(abab^{-1})]$ and $[g(aaba^{-1}b^{-1})]$ contain closed geodesics. These geodesics are primitive since $ABAB^{-1}$ and $AABA^{-1}B^{-1}$ are. According to Theorem 3.1 the classes $[g(abab^{-1})]$ and $[g(aaba^{-1}b^{-1})]$ also contain loops with a single self-intersection but no simple loops. Since geodesics realise the minimum number of self-intersections of all loops in their free homotopy classes, the classes $[g(abab^{-1})]$ and $[g(aaba^{-1}b^{-1})]$ must contain closed 1-intersectors. We have proved the following theorem.

**Theorem 3.2.** *A closed geodesic on* $\mathbf{T}$ *is a closed 1-intersector if and only if it is defined by a conjugacy class in* $\Gamma'$ *of the form* $[G(ABAB^{-1})]$ *or* $[G(AABA^{-1}B^{-1})]$ *where* $G \in \text{Aut } \Gamma'$.

Because every automorphism of $\pi_1(\mathbf{T})$ can be induced by a homeomorphism of $\mathbf{T}$ we know that the closed 1-intersectors which are defined by the conjugacy classes of the form $[G(AABA^{-1}B^{-1})]$ where $G \in \text{Aut } \Gamma'$ are topologically all the same as the one shown in Figure 3.6. In particular, they all contain a subloop which

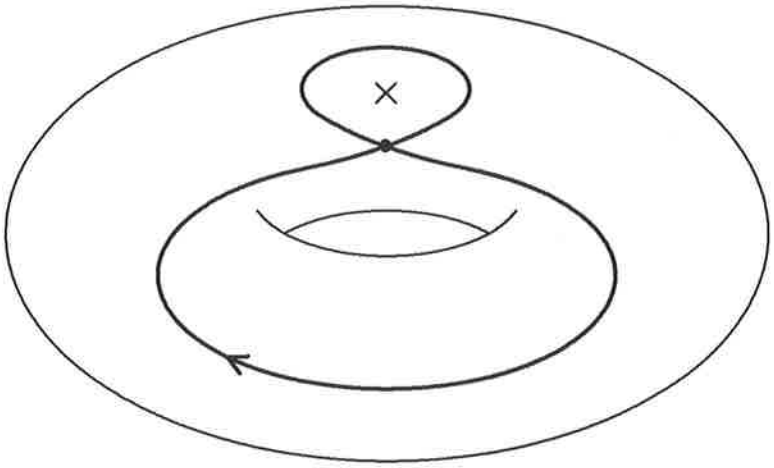

FIGURE 3.6. The geodesic on $\mathbf{T}$ in the free homotopy class $[aaba^{-1}b^{-1}]$.

bounds a disc containing the puncture. As stated in the introduction, we call such geodesics *improper closed 1-intersectors*. The only other closed 1-intersectors are defined by conjugacy classes of the form $[G(ABAB^{-1})]$. We refer to them as *proper closed 1-intersectors*. They are all topologically the same as the one shown in Figure 3.5

**Remark 3.1.** We can now explain why the improper closed 1-intersectors have Markoff values greater than 6. Suppose $\gamma$ is a improper closed 1-intersector defined by the conjugacy class $[G(AABA^{-1}B^{-1})]$ where $G \in \text{Aut } \Gamma'$. We know $\gamma$ is a loop in the free homotopy class $[g(aaba^{-1}b^{-1})]$ where $g = \theta(G)$ and $\theta$ is the isomorphism defined by (1.11). It follows from the proof of Theorem 3.1 that by choosing the initial point of $\gamma$ correctly we can write $\gamma = \gamma_1\gamma_2$ where $\gamma_2$ is a subloop which bounds a disc containing the puncture. Moreover, assuming we have done so, the free homotopy classes of $\gamma_1$ and $\gamma_2$ are $[g(a)]$ and $[g(aba^{-1}b^{-1})]$, respectively. Since $[g(aba^{-1}b^{-1})]$ is either $[aba^{-1}b^{-1}]$ or $[bab^{-1}a^{-1}]$ the lift $\tilde{\gamma}_2$ of $\gamma_2$ to $\mathbf{H}$ is a curve from some $z$ to $W(z)$ where $W \in [ABA^{-1}B^{-1}]$ or $W \in [BAB^{-1}A^{-1}]$. Thus we set

$$T(z) = A^{-1}B^{-1}AB(z) = z + 6.$$

Since $W \in [T]$ or $W \in [T^{-1}]$ there is some $V \in \Gamma'$ such that $VWV^{-1}$ is $T$ or $T^{-1}$. The geodesic segment $V(\tilde{\gamma}_2)$ is a lift of $\gamma_2$ from $V(z)$ to $VW(z)$. Our choice of $V$

implies $VW(z)$ is $TV(z)$ or $T^{-1}V(z)$. Hence $V(\tilde{\gamma}_2)$ passes through both the point $V(z)$ and its translation by $+6$ or $-6$. It follows that $\gamma$ has a lift to $\mathbf{H}$ which, as a semi-circle, has a diameter greater than 6. The Markoff value of $\gamma$ is the supremum of such diameters and therefore must also be greater than 6.

In the proof of Theorem 3.1 we used the fact that every automorphism $g$ of $\pi_1(\mathbf{T})$ can be induced by a homeomorphism of $\mathbf{T}$. This allowed us to replace a conjugacy class in $\pi_1(\mathbf{T})$ by its image under $g$ without altering our hypothesis concerning the topological properties of loops in the associated free homotopy classes. We thereby simplified the task of characterising those conjugacy classes containing loops with a non-trivial single self-intersection. We shall use a similar technique to characterise the simple open geodesics on $\mathbf{T}$ but with cutting sequences in place of conjugacy classes. We provide the following theorem as motivation for that technique.

**Theorem 3.3.** *Let $\gamma_1$ and $\gamma_2$ be primitive closed geodesics on $\mathbf{T}$ and suppose $\mathbf{S}(\gamma_2) = G\,\mathbf{S}(\gamma_1)$ for some $G \in Aut\ \Gamma'$. Then $\gamma_1$ and $\gamma_2$ have the same number of self-intersections.*

*Proof.* Let $W_1$ be a cyclically reduced representative of the conjugacy class defining $\gamma_1$ and choose $H \in \text{Inn } \Gamma'$ so that $W_2 = HG(W_1)$ is a cyclically reduced word. It is clear from Theorem 2.6 that $\mathbf{S}(\gamma_2) = HG\,\mathbf{S}(\gamma_1)$. We know from Theorem 2.2 that $\mathbf{S}(\gamma_1)$ is periodic with period $W_1$. Therefore Theorem 2.6 implies $\mathbf{S}(\gamma_2)$ is periodic with period $W_2$. Now observe that since $\gamma_1$ is primitive and defined by $[W_1]$ we know $W_1$ is primitive. Obviously automorphisms preserve the property of being primitive and hence $W_2$ is also primitive. It follows that $W_2$ is primitive and a period of $\mathbf{S}(\gamma_2)$. Using Theorem 2.2 and the fact that $\gamma_2$ is primitive it can be seen that any cyclically reduced word in the conjugacy class defining $\gamma_2$ is also primitive and a period of $\mathbf{S}(\gamma_2)$. Clearly primitive periods of $\mathbf{S}(\gamma_2)$ are cyclic permutations of one another and hence $W_2$ lies in the conjugacy class defining $\gamma_2$. We have shown that $[W_2]$ defines $\gamma_2$.

Now recall from Chapter 1 that $\gamma_1$ and $\gamma_2$ lie in the free homotopy classes $w_1 = \theta(W_1)$ and $w_2 = \theta(W_2)$, respectively, where $\theta$ is the isomorphism (1.11). We know that $W_2 = HG(W_1)$ and therefore $w_2 = g'(w_1)$ for some $g' \in \pi_1(\mathbf{T})$. Using the fact

that every automorphism of $\pi_1(\mathbf{T})$ is induced by a homeomorphism of $\mathbf{T}$ we can now deduce that the free homotopy classes $[w_1]$ and $[w_2]$ are homeomorphic images of one another. Since primitive closed geodesics realise the minimum number of self-intersections of all loops in their free homotopy classes, it follows that $\gamma_1$ and $\gamma_2$ have the same number of self-intersections. □

Theorem 3.3 is also true for open geodesics on $\mathbf{T}$. To be more precise, we claim that if $\gamma_1$ and $\gamma_2$ are open geodesics on $\mathbf{T}$ and if $\mathbf{S}(\gamma_2) = G\ \mathbf{S}(\gamma_1)$ for some $G \in \text{Aut } \Gamma'$ then $\gamma_1$ and $\gamma_2$ have the same number of self-intersections. As indicated above, we shall use this result to help characterise the simple open geodesics on $\mathbf{T}$. First we must prove that it is true. We cannot use the homeomorphisms of $\mathbf{T}$ to do this since the image of an open geodesic under a homeomorphism is not in general a geodesic and we have no theory dealing with homotopy and cutting sequences for arbitrary open curves on $\mathbf{T}$. While we believe it is possible to develop such a theory there is an easier way to obtain the desired result. We shall need the following background material.

By applying the automorphism $R(A, B) = (B^{-1}, A)$ to boundary expansions we can define a map

$$f: \quad \mathbf{R}\backslash\mathbf{Q} \quad \longrightarrow \quad \mathbf{R}\backslash\mathbf{Q} \tag{3.1}$$

from the set of irrationals $\mathbf{R}\backslash\mathbf{Q}$ to itself. Specifically, given an irrational $\xi$ with boundary expansion $\mathbf{S}(\xi)$ we define $f(\xi)$ to be the irrational with boundary expansion $R\ \mathbf{S}(\xi)$. By $R\ \mathbf{S}(\xi)$ we mean of course the image of $\mathbf{S}(\xi)$ under the substitution

$$A \to B^{-1}, \qquad B \to A. \tag{3.2}$$

In other words, $f$ is defined by

$$\mathbf{S}(f(\xi)) = R\ \mathbf{S}(\xi) \tag{3.3.}$$

Obviously $f$ is a well-defined bijection. We claim that it cyclically permutes the natural ordering of the irrationals. Specifically, we claim that if $\xi_1, \xi_2, \ldots, \xi_n$ are any irrationals with

$$\xi_1 < \xi_2 < \cdots < \xi_n$$

then $f(\xi_1), f(\xi_2), \ldots, f(\xi_n)$ satisfy some cyclic permutation of the ordering

$$f(\xi_1) < f(\xi_2) < \cdots < f(\xi_n).$$

The reason for this is that the substitution (3.2) cyclically permutes the ordering (2.4). Therefore (3.2) cyclically permutes the associated lexicographic ordering of boundary expansions and the result now follows from Theorem 2.1.

The map $f$ induces a bijection from the set of geodesics in $\mathbf{H}$ (with irrational endpoints) to itself. Thus we define

$$T_R: \quad \gamma = [\eta, \xi] \quad \longmapsto \quad T_R(\gamma) = [f(\eta), f(\xi)]. \tag{3.4}$$

The significance of this map lies in the fact that

$$\mathbf{S}(T_R(\gamma)) = R\,\mathbf{S}(\gamma) \tag{3.5}$$

To see that this is true, suppose $\eta$ and $\xi$ have boundary expansions

$$\mathbf{S}(\eta) = X_{-1}X_{-2}X_{-3}\ldots\ldots \qquad \text{and} \qquad \mathbf{S}(\xi) = X_0X_1X_2\ldots\ldots,$$

respectively, and recall that

$$\mathbf{S}(\gamma) = \ldots\ldots X_{k-3}^{-1}X_{k-2}^{-1}X_{k-1}^{-1}X_kX_{k+1}X_{k+2}\ldots\ldots$$

where $k \geq 0$ is the smallest integer such that $X_{k-1} \neq X_k$. A similar relationship holds between the boundary expansions of $f(\eta)$ and $f(\xi)$ and the cutting sequence of $T_R(\gamma)$. Since $\mathbf{S}(f(\eta))$ and $\mathbf{S}(f(\xi))$ are the images of $\mathbf{S}(\eta)$ and $\mathbf{S}(\xi)$, respectively, under the substitution (3.2) it follows that $\mathbf{S}(T_R(\gamma))$ is likewise the image of $\mathbf{S}(\gamma)$ under (3.2). The truth of (3.5) is now evident.

We can deduce from (3.5) that $T_R$ maps $\Gamma'$-equivalent geodesics to $\Gamma'$-equivalent geodesics. To see this, let $\gamma$ and $\gamma'$ be $\Gamma'$-equivalent. We have

$$\mathbf{S}(T_R(\gamma)) = R\,\mathbf{S}(\gamma) = R\,\mathbf{S}(\gamma') = \mathbf{S}(T_R(\gamma'))$$

and so $T_R(\gamma)$ and $T_R(\gamma')$ are $\Gamma'$-equivalent. Note also that since $R^4 = \mathrm{Id}$ and hence $f^4 = \mathrm{Id}$ and $T_R^4 = \mathrm{Id}$, it is clear that the converse is true. Another important

property of $T_R$ is that it maps intersecting geodesics to intersecting geodesics. In other words, for any geodesics $\gamma$ and $\gamma'$ in $\mathbf{H}$ we have

$$\gamma \cap \gamma' \neq \emptyset \qquad \Longleftrightarrow \qquad T_R(\gamma) \cap T_R(\gamma') \neq \emptyset \tag{3.6}$$

To see this, observe that if $\gamma \neq \gamma'$ then $\gamma \cap \gamma' \neq \emptyset$ if and only if the endpoints of $\gamma$ and $\gamma'$ separate one another as points of the real axis. Similarly, if $T_R(\gamma) \neq T_R(\gamma')$ then $T_R(\gamma) \cap T_R(\gamma') \neq \emptyset$ if and only if the endpoints of $T_R(\gamma)$ and $T_R(\gamma')$ separate one another. Since the endpoints of $T_R(\gamma)$ and $T_R(\gamma')$ are the images of those of $\gamma$ and $\gamma'$ under the map $f$ and since $f$ preserves the cyclic ordering of the irrationals we know that the endpoints of $\gamma$ and $\gamma'$ separate one another if and only if the endpoints of $T_R(\gamma)$ and $T_R(\gamma')$ do also. The truth of (3.6) is now apparent.

We can now prove the following theorem.

**Theorem 3.4.** *Let $\gamma_1$ and $\gamma_2$ be open geodesics on $\mathbf{T}$ and suppose that $\mathbf{S}(\gamma_2) = G\,\mathbf{S}(\gamma_1)$ for some $G \in$ Aut $\Gamma'$. Then $\gamma_1$ and $\gamma_2$ have the same number of self-intersections.*

*Proof.* We shall deal with the case where $G \in \Psi$ first. In this case, we know from Chapter 1 that $G = G_T$ for some $T \in \Gamma^*$. Let $\tilde{\gamma}_1$ be a lift of $\gamma_1$ to $\mathbf{H}$ and note that (2.22) implies $\mathbf{S}(T(\tilde{\gamma}_1)) = G_T\,\mathbf{S}(\tilde{\gamma}_1)$. Thus

$$\mathbf{S}(T(\tilde{\gamma}_1)) = G\,\mathbf{S}(\tilde{\gamma}_1) = G\,\mathbf{S}(\gamma_1) = \mathbf{S}(\gamma_2)$$

and so $T(\tilde{\gamma}_1)$ is a lift of $\gamma_2$ to $\mathbf{H}$. It follows that $\gamma_2$ is the image of $\gamma_1$ under the isometry of $\mathbf{T}$ induced by the action of $T$ on $\mathbf{H}$ and the truth of the theorem is evident in this case.

It is easy to verify that if the theorem is true for both $G = G_1$ and $G = G_2$ then it is also true for $G = G_1G_2$. We know $R$ and $\Psi$ generate Aut $\Gamma'$ and therefore we can complete the proof by showing the theorem is true for $G = R$. Thus we now assume $G = R$. We can also assume neither $\gamma_1$ nor $\gamma_2$ covers a closed geodesic. To see this, observe that if $\gamma_1$ covers a closed geodesic then its cutting sequence is periodic. In this case, Theorem 2.6 implies $\mathbf{S}(\gamma_2)$ is periodic and Theorem 2.2 implies $\gamma_2$ covers a closed geodesic. Hence $\gamma$ and $\gamma'$ both have a continuum of

self-intersections and the theorem is true. A similar argument using the fact that $\mathbf{S}(\gamma_1) = G^{-1}\,\mathbf{S}(\gamma_2)$ shows the theorem is true if $\gamma_2$ covers a closed geodesic.

Again, let $\tilde{\gamma}_1$ be a lift of $\gamma_1$ to $\mathbf{H}$. We are assuming $\mathbf{S}(\gamma_2) = R\,\mathbf{S}(\gamma_1)$ and since (3.5) is true we have

$$\mathbf{S}(T_R(\tilde{\gamma}_1)) = R\,\mathbf{S}(\tilde{\gamma}_1) = R\,\mathbf{S}(\gamma_1) = \mathbf{S}(\gamma_2).$$

Hence the geodesic $T_R(\tilde{\gamma}_1)$ is a lift of $\gamma_2$ to $\mathbf{H}$. Now recall from Chapter 1 that the self-intersection number of $\gamma_1$ is half the number of geodesics in $\mathbf{H}$ which are $\Gamma'$-equivalent to and intersect $\tilde{\gamma}_1$. Similarly, the self-intersection number of $\gamma_2$ is half the number of geodesics which are $\Gamma'$-equivalent to and intersect $T_R(\tilde{\gamma}_1)$. We have seen above that the map $T_R$ preserves $\Gamma'$-equivalence of geodesics and their intersection properties. Clearly then the number of geodesics which are $\Gamma'$-equivalent to $\tilde{\gamma}_1$ and intersect $\tilde{\gamma}_1$ is the same as the number of geodesics which are $\Gamma'$-equivalent to $T_R(\tilde{\gamma}_1)$ and intersect $T_R(\tilde{\gamma}_1)$. It follows that $\gamma_1$ and $\gamma_2$ have the same number of self-intersections and the proof is complete. □

We are now ready to characterise the simple open geodesics on $\mathbf{T}$. We present our results in the form of two theorems. In the first theorem, Theorem 3.5, we show that their cutting sequences are linear and aperiodic or half-linear. In the second theorem, Theorem 3.6, we show that the converse is true. The proof Theorem 3.5 is algebraic in nature. It relies on the properties of linear and half-linear cutting sequences discussed in Chapter 2. In contrast to this the proof of Theorem 3.6 is geometrical. It uses the lines and rays associated with linear and half-linear cutting sequences. We remind the reader that Haas, [21], has already provided a topological characterisation of the simple open geodesics on $\mathbf{T}$.

**Theorem 3.5.** *Let $\gamma$ be an open geodesic on $\mathbf{T}$. If $\gamma$ is simple then $\mathbf{S}(\gamma)$ is linear and aperiodic or half-linear.*

*Proof.* Let $\gamma$ be a simple open geodesic on $\mathbf{T}$. It follows from our conventions regarding self-intersections numbers that $\gamma$ cannot cover a closed geodesic. Thus $\mathbf{S}(\gamma)$ aperiodic. As a consequence of this $\mathbf{S}(\gamma)$ is not of the form $Y^\infty$ where $Y \in \{A, B, A^{-1}, B^{-1}\}$. There are three other possibilities, namely, $\mathbf{S}(\gamma)$ is composed of

two, three or four symbols in $\{A, B, A^{-1}, B^{-1}\}$. We shall consider each possibility separately. Before we do so, we shall establish some general restrictions on the form of $\mathbf{S}(\gamma)$. For this purpose it is convenient to refer to a pair of symbols $Y, Z$ with $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$ as *elementary*. The elementary pairs are exactly the images of the pair $A, B$ under the automorphisms in the subgroup

$$\{\mathrm{Id},\ R,\ R^2,\ R^3,\ P,\ PR,\ PR^2,\ PR^3\} \tag{3.7}$$

of Aut $\Gamma'$. Theorem 3.4 implies we can replace $\gamma$ by any geodesic whose cutting sequence is the image of $\mathbf{S}(\gamma)$ under an automorphism in the set (3.7) without changing our hypothesis that $\gamma$ is simple. It is also clear from Theorems 2.8 and 2.10 or otherwise that the automorphisms in (3.7) permute the set of linear and half-linear sequences. Hence the conclusion of the theorem is likewise un-effected by the application of automorphisms in (3.7) to $\mathbf{S}(\gamma)$.

We begin with the restriction:-

(R1) there is no elementary pair $Y, Z$ such that
both $Y^2$ and $Z^2$ occur in $\mathbf{S}(\gamma)$.

To see this suppose otherwise. By applying the appropriate automorphism in the subgroup (3.7) we may assume $Y = A$ and $Z = B$. Now let $\tilde{\gamma}_1$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses the fundamental domain $\mathcal{D}$ shown in Figure 1.2 and contributes $A^2$ to $\mathbf{S}(\gamma)$ as it does so and let $\tilde{\gamma}_2$ be the lift which likewise crosses $\mathcal{D}$ and contributes $B^2$ to $\mathbf{S}(\gamma)$. By referring to Figure 2.1, for instance, it is not hard to see that $\tilde{\gamma}_1 = [\eta_1, \xi_1]$ where $\eta_1 < -1$ and $0 < \xi_1 < 1$ and $\tilde{\gamma}_2 = [\eta_2, \xi_2]$ where $1 < \eta_2$ and $-1 < \xi_2 < 0$. Hence $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ intersect. Such an intersection projects to a self-intersection of $\gamma$ and we have a contradiction of our assumption that $\gamma$ is simple.

Similarly, we have the restriction:-

(R2) not both $ZY^nZ$ and $Y^{n+2}$ occur in $\mathbf{S}(\gamma)$ where
$n \geq 1$ and the pair $Y, Z$ is elementary.

Suppose otherwise. As in (R1), we may assume $Y = A$ and $Z = B$. This time we let $\tilde{\gamma}_1$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes $BA$ to the word $BA^nB$

in $\mathbf{S}(\gamma)$ as it does so and we let $\tilde{\gamma}_2$ be the lift which crosses $\mathcal{D}$ and contributes the initial $A^2$ to the word $A^{n+2}$. Clearly, $\tilde{\gamma}_1 = [\eta_1, \xi_1]$ where $\eta_1 > 1$ and $0 < \xi_1 < 1$ and $\tilde{\gamma}_2 = [\eta_2, \xi_2]$ where $\eta_2 < -1$ and $0 < \xi_2 < 1$. The boundary expansions of $\xi_1$ and $\xi_2$ begin with $A^nB$ and $A^{n+1}$, respectively. Since $A^nB\ldots < A^{n+1}\ldots$ in the lexicographic ordering of boundary expansions Theorem 2.1 implies that $\xi_1 < \xi_2$. Again, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ intersect contradicting our assumption that $\gamma$ is simple.

Now suppose $\mathbf{S}(\gamma)$ is composed of exactly two symbols in $\{A, B, A^{-1}, B^{-1}\}$, say $Y$ and $Z$. Since $\mathbf{S}(\gamma)$ is reduced, $Z \neq Y^{\pm 1}$ and as usual, we may assume $Y = A$ and $Z = B$. Restriction (R1) implies not both $A^2$ and $B^2$ occur in $\mathbf{S}(\gamma)$. Thus one of the symbols $A$ or $B$ is isolated in $\mathbf{S}(\gamma)$. By applying the automorphism $P$, if necessary, we may assume $B$ is the isolated symbol. We may also assume that a word of the form $BA^nB$ with $n \geq 1$ occurs, else $\mathbf{S}(\gamma) = A^\infty BA^\infty$ and hence is linear. Let the integer $n$ be minimal. Restriction (R2) implies there are no occurrences of $A^{n+2}$. It follows that $\mathbf{S}(\gamma)$ can be partitioned into the blocks $A^nB$ and $A^{n+1}B$ and therefore is derivable. Recall from the section of Chapter 2 on linear sequences and (2.25) and (2.26) in particular that the derived sequence of $\mathbf{S}(\gamma)$ is $G\,\mathbf{S}(\gamma)$ where $G$ is defined by $G(A) = A$ and $G(B) = A^{-n}B$. Let $\gamma'$ be the open geodesic on $\mathbf{T}$ whose cutting sequence $\mathbf{S}(\gamma')$ is the derived sequence of $\mathbf{S}(\gamma)$. Theorem 3.4 implies $\gamma'$ is simple. Since $G(A^nB) = B$ and $G(A^{n+1}B) = AB$ we know $\mathbf{S}(\gamma')$ is composed of $B$'s and $AB$'s. Equivalently, $\mathbf{S}(\gamma')$ is composed of $A$'s and $B$'s and the occurrences of $A$ are isolated. We can now repeat the argument with $\gamma'$ in place of $\gamma$ and the roles of $A$ and $B$ interchanged. It follows that either $\mathbf{S}(\gamma') = B^\infty AB^\infty$ or $\mathbf{S}(\gamma')$ is derivable. Further, if $\mathbf{S}(\gamma')$ is derivable and $\gamma''$ the open geodesic on $\mathbf{T}$ whose cutting sequence $\mathbf{S}(\gamma'')$ is the derived sequence of $\mathbf{S}(\gamma')$ then $\mathbf{S}(\gamma'')$ is composed of $A$'s and $B$'s and the occurrences of $B$ are isolated. In the latter case, we repeat the argument with $\gamma''$ in place of $\gamma$ and so on. By continuing in this manner we conclude that either $\mathbf{S}(\gamma)$ is derivable infinitely often or it is derivable to a sequence of the form $A^\infty BA^\infty$ or $B^\infty AB^\infty$. Theorem 2.8 implies $\mathbf{S}(\gamma)$ is linear.

Before we consider the cases where $\mathbf{S}(\gamma)$ is composed of three and four symbols

we establish three further restrictions on its form. We claim that:-

(R3) there is no elementary pair $Y, Z$ such that
$YZ^nY^{-1}$ occurs in $\mathbf{S}(\gamma)$ with $n \geq 2$.

As usual, we suppose otherwise and assume $Y = A$ and $Z = B$. Let $\tilde{\gamma}_1$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes $AB$ to the word $AB^n$ in $\mathbf{S}(\gamma)$ and let $\tilde{\gamma}_2$ be the lift which crosses $\mathcal{D}$ and contributes the initial $B^2$ to the word $B^nA^{-1}$. Then $\tilde{\gamma}_1 = [\eta_1, \xi_1]$ where $\eta_1 < -1$ and $-1 < \xi_1 < 0$ and $\tilde{\gamma}_2 = [\eta_2, \xi_2]$ where $\eta_2 > 1$ and $-1 < \xi_2 < 0$. Since the boundary expansions of $\xi_1$ and $\xi_2$ begin with $B^n$ and $B^{n-1}A^{-1}$, respectively, and $B^n \ldots > B^{n-1}A^{-1} \ldots$ we know $\xi_1 > \xi_2$. Thus $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ intersect and we have a contradiction.

We also have:-

(R4) not both $ZY^nZ$ and $Z(Y^{-1})^mZ$ occur in $\mathbf{S}(\gamma)$ where
$n \geq 1$ and $m \geq 1$ and the pair $Y, Z$ is elementary.

Suppose otherwise and assume $Y = A$ and $Z = B$. Let $\tilde{\gamma}_1$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes $BA$ to the word $BA^nB$ and let $\tilde{\gamma}_2$ be the lift which crosses $\mathcal{D}$ and contributes $A^{-1}B$ to the word $B(A^{-1})^mB$. Then $\tilde{\gamma}_1 = [\eta_1, \xi_1]$ where $\eta_1 > 1$ and $0 < \xi_1 < 1$ and $\tilde{\gamma}_2 = [\eta_2, \xi_2]$ where $0 < \eta_2 < 1$ and $-1 < \xi_2 < 0$. The boundary expansions of $\xi_1$ and $\eta_2$ begin with $A^nB$ and $A^mB^{-1}$, respectively, and since $A^nB \ldots < A^mB^{-1} \ldots$ we know $\xi_1 < \eta_2$. Thus $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ intersect and we have a contradiction.

Our final restriction is that:-

(R5) not both $ZY^nZ$ and $(Y^{-1})^{n+2}$ occur in $\mathbf{S}(\gamma)$ where
$n \geq 1$ and the pair $Y, Z$ is elementary.

Suppose otherwise and assume $Y = A$ and $Z = B$. Let $\tilde{\gamma}_1$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes $BA$ to the word $BA^nB$ and let $\tilde{\gamma}_2$ be the lift which crosses $\mathcal{D}$ and contributes the final $(A^{-1})^2$ to the word $(A^{-1})^{n+2}$. Then $\tilde{\gamma}_1 = [\eta_1, \xi_1]$ where $\eta_1 > 1$ and $0 < \xi_1 < 1$ and $\tilde{\gamma}_2 = [\eta_2, \xi_2]$ where $0 < \eta_2 < 1$ and $\xi_2 < -1$. The boundary expansions of $\xi_1$ and $\eta_2$ begin with $A^nB$ and $A^{n+1}$, respectively, and since $A^nB \ldots < A^{n+1} \ldots$ we know $\xi_1 < \eta_2$. Thus $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ intersect and we have a contradiction.

Now we suppose $\mathbf{S}(\gamma)$ is composed of exactly three symbols in $\{A, B, A^{-1}, B^{-1}\}$, say $Y$, $Y^{-1}$ and $Z$. Clearly one of $Y^{-1}Z^nY$ or $YZ^nY^{-1}$ with $n \geq 1$ occurs in $\mathbf{S}(\gamma)$. In either case, (R3) implies $n = 1$. By applying the appropriate element of (3.7) we can assume that $Y = A$ and $Z = B$ and $A^{-1}BA$ occurs. The symbol $B$ cannot occur to both the left and right of $A^{-1}BA$ in $\mathbf{S}(\gamma)$ else words of the form $B(A^{-1})^mB$ and $BA^nB$ with $n \geq 1$ and $m \geq 1$ occur contradicting (R4). Thus $\mathbf{S}(\gamma)$ contains one of $(A^{-1})^\infty BA$ or $A^{-1}BA^\infty$. In either case, there can be no other $B$'s in $\mathbf{S}(\gamma)$ else (R5) is contradicted. It follows that $\mathbf{S}(\gamma) = (A^{-1})^\infty BA^\infty$ and hence is half-linear.

Finally, we suppose $\mathbf{S}(\gamma)$ is composed of all four symbols in $\{A, B, A^{-1}, B^{-1}\}$. Using (R3) it is not hard to see that a word of the form $Y^{-1}ZY$ where the pair $Y, Z$ is elementary occurs. As usual, we assume $Y = A$ and $Z = B$ and hence $A^{-1}BA$ occurs. Since (R4) implies no word of the form $B(A^{-1})^mBA^nB$ with $n \geq 1$ and $m \geq 1$ occurs and since at least one $B^{-1}$ occurs we conclude that one of the words $B^{-1}(A^{-1})^nBA$ or $A^{-1}BA^nB^{-1}$ where $n \geq 1$ occurs. Restriction (R3) implies $n = 1$. By applying the appropriate element of (3.7) we can assume $B^{-1}A^{-1}BA$ occurs. We write

$$\mathbf{S}(\gamma) = \ldots\ldots X_{-3}^{-1}X_{-2}^{-1}X_{-1}^{-1}\, B^{-1}A^{-1}BA\, X_1X_2X_3 \ldots\ldots$$

Suppose for the moment that $X_1 = B^{-1}$. Then $\gamma$ has a lift $\tilde{\gamma} = [\eta, \xi]$ to $\mathbf{H}$ with $\eta < -5$ and $\xi > 1$. Clearly $\tilde{\gamma}$ has a diameter greater than 6 and so intersects $U_1^6(\tilde{\gamma})$ where $U_1^6(z) = z + 6$. Since $U_1^6 \in \Gamma'$ this contradicts our assumption that $\gamma$ is simple. A similar argument shows $X_{-1}^{-1} \neq A$. Thus $X_1$ is $A$ or $B$ and $X_{-1}^{-1}$ is $A^{-1}$ or $B^{-1}$. It follows that $\gamma$ has a lift $\tilde{\gamma} = [\eta, \xi]$ to $\mathbf{H}$ with $-7 < \eta < -5$ and $-1 < \xi < 1$. We claim that this implies neither $AB^{-1}$ nor $BA^{-1}$ occurs in $\mathbf{S}(\gamma)$. If $AB^{-1}$ occurs then $\gamma$ has a lift $\tilde{\gamma}' = [\eta', \xi']$ to $\mathbf{H}$ with $\eta' < -1$ and $\xi' > 1$. In this case, $\eta'$ must be less than $-7$ else $\tilde{\gamma}$ and $\tilde{\gamma}'$ intersect. But then the diameter of $\tilde{\gamma}'$ is greater than 6 which have just seen is impossible. Similarly, if $BA^{-1}$ occurs then $\gamma$ has a lift $\tilde{\gamma}' = [\eta', \xi']$ with $\eta' > -5$ and $\xi' < -7$. In this case, $\eta' > 1$ else $\tilde{\gamma}$ and $\tilde{\gamma}'$ intersect but again this implies $\tilde{\gamma}'$ has a diameter greater than 6 which is impossible. It follows that our claim is true. Note that our claim implies

that $X_1, X_2, X_3 \ldots$ is composed entirely of $A$'s and $B$'s and $\ldots X_{-3}^{-1} X_{-2}^{-1} X_{-1}^{-1}$ is composed entirely of $A^{-1}$'s and $B^{-1}$'s.

We know from (R1) that either the symbols $A$ and $A^{-1}$ are isolated in $\mathbf{S}(\gamma)$ or the symbols $B$ and $B^{-1}$ are. We shall assume first that $B$ and $B^{-1}$ are isolated. If $X_1, X_2, X_3 \ldots$ is $A^\infty$ then (R5) implies $\ldots X_{-3}^{-1} X_{-2}^{-1} X_{-1}^{-1}$ is $(A^{-1})^\infty$ in which case, $\mathbf{S}(\gamma)$ is half-linear. Thus we also assume $X_1, X_2, X_3 \ldots$ begins with $A^n B$ for some $n \geq 0$. A similar argument allows us to assume $\ldots X_{-3}^{-1} X_{-2}^{-1} X_{-1}^{-1}$ ends with $B^{-1}(A^{-1})^m$. Note that $m \geq 1$ since $(B^{-1})^2$ cannot occur. Since $B^{-1}(A^{-1})^m B^{-1}$ and $BA^{n+1}B$ occur in $\mathbf{S}(\gamma)$, restriction (R5) implies $m = n$ or $m = n+1$ or $m = n+2$. We consider two cases.

Suppose that $m = n+1$ or $m = n+2$. Let $\gamma'$ be the open geodesic on $\mathbf{T}$ whose cutting sequence is the image of $\mathbf{S}(\gamma)$ under the automorphism $G(A, B) = (A, A^{-(n+1)}B)$. Recall that, $\mathbf{S}(\gamma')$ is obtained from $\mathbf{S}(\gamma)$ by applying the substitution $A \to A$, $B \to (A^{-1})^{n+1}B$ and reducing the resulting sequence. In this process only the symbols $A$ and $A^{-1}$ are added to or removed from $\mathbf{S}(\gamma)$. It follows that the occurrence of the string

$$B^{-1}(A^{-1})^m B^{-1} A^{-1} B A^{n+1} B$$

in $\mathbf{S}(\gamma)$ implies the occurrence of

$$B^{-1}(A^{-1})^k\, B^{-1} A^{-1} B B, \tag{3.8}$$

where $k = m - (n+1)$, in $\mathbf{S}(\gamma')$. We know $\mathbf{S}(\gamma')$ is simple and we have shown that $B^{-1}A^{-1}B$ occurs. According to the arguments above if all four symbols occur in $\mathbf{S}(\gamma')$ then $A$ must occur immediately before or after $B^{-1}A^{-1}B$. It is evident from (3.8) that that is impossible. Hence only three symbols occur in $\mathbf{S}(\gamma')$. The arguments above imply $\mathbf{S}(\gamma') = (B^{-1})^\infty A^{-1} B^\infty$ and therefore

$$\mathbf{S}(\gamma) = (B^{-1}A^{-1}(A^{-1})^n)^\infty\, B^{-1}A^{-1}BA\,(A^n BA)^\infty$$

where $n \geq 0$. Clearly $\mathbf{S}(\gamma)$ is half-linear in this case.

The other possibility is that $n = m \geq 1$. In this case, the words $B^{-1}(A^{-1})^n B^{-1}$ and $BA^{n+1}B$ occur in $\mathbf{S}(\gamma)$. It follows from (R2) and (R5) that $X_1, X_2, X_3 \ldots$ can

be partitioned into the blocks $A^nB$ and $A^{n+1}B$. Similarly, $X_{-1}, X_{-2}, X_{-3}\ldots$ can be likewise partitioned. Since $X_1, X_2, X_3\ldots$ and $X_{-1}, X_{-2}, X_{-3}\ldots$ both begin with $A^nB$ they are both derivable by the substitution $A^nB \to B$, $A^{n+1}B \to AB$. Let their derived sequences be $X'_1, X'_2, X'_3\ldots$ and $X'_{-1}, X'_{-2}, X'_{-3}\ldots$, respectively. Also, let $\gamma'$ be the open geodesic on $\mathbf{T}$ whose cutting sequence is the image of $\mathbf{S}(\gamma)$ under the automorphism $G(A,B) = (A, A^{-n}B)$. With care, it can be verified that

$$\mathbf{S}(\gamma') = \ldots\ldots (X'_{-3})^{-1}(X'_{-2})^{-1}(X'_{-1})^{-1}\ B^{-1}A^{-1}BA\ X'_1X'_2X'_3\ldots\ldots$$

Note that the sequence $X'_1X'_2X'_3\ldots$ is composed of $A$'s and $B$'s and that the sequence $\ldots(X'_{-3})^{-1}(X'_{-2})^{-1}(X'_{-1})^{-1}$ is composed of $A^{-1}$'s and $B^{-1}$'s. Further, the symbols $A$ and $A^{-1}$ are isolated in $\mathbf{S}(\gamma')$. Note also that $\gamma'$ is a simple open geodesic and $\mathbf{S}(\gamma')$ is the half-derived sequence of $\mathbf{S}(\gamma)$.

To summarise, we have shown that if $B$ and $B^{-1}$ are isolated in $\mathbf{S}(\gamma)$ then either $\mathbf{S}(\gamma)$ is half-linear or it is half-derivable to the cutting sequence of a simple open geodesic on $\mathbf{T}$ which is of the same form as $\mathbf{S}(\gamma)$ except that $A$ and $A^{-1}$ are the isolated symbols. A similar argument shows that in the case where $A$ and $A^{-1}$ are isolated in $\mathbf{S}(\gamma)$ either $\mathbf{S}(\gamma)$ is half-linear or it is half-derivable to the cutting sequence of a simple open geodesic on $\mathbf{T}$ which is of the same form as $\mathbf{S}(\gamma)$ except that $B$ and $B^{-1}$ are the isolated symbols. (The truth of the latter statement may also be deduced from the former by reversing the orientation of $\gamma$ and applying $P$ to $\mathbf{S}(\gamma)$.) It follows by induction that either $\mathbf{S}(\gamma)$ is half-derivable infinitely often or it is half-derivable to, or is, a half-linear sequence. In the first case, Theorem 2.10 implies $\mathbf{S}(\gamma)$ is half-linear and in the second case, Remark 2.5 does. The proof is complete. □

**Theorem 3.6.** *Let $\gamma$ be an open geodesic on $\mathbf{T}$. If $\mathbf{S}(\gamma)$ is linear and aperiodic or half-linear then $\gamma$ is simple.*

*Proof.* We have already noted in Chapter 2 that Haas, [21], has proved the theorem is true if $\mathbf{S}(\gamma)$ is aperiodic and linear. Thus we shall assume $\mathbf{S}(\gamma)$ is half-linear. We point out however that our method also works in the case where $\mathbf{S}(\gamma)$ is linear. Half-linear sequences are aperiodic and hence $\gamma$ does not cover a closed

geodesic on **T**. It follows from our discussion in Chapter 1 on the self-intersection numbers of geodesics that we can prove $\gamma$ is simple by showing there is a lift $\tilde{\gamma}$ of $\gamma$ to **H** which is not intersected by any $\Gamma'$-equivalent geodesic. We shall prove the latter is true by assuming otherwise and obtaining a contradiction. Specifically, we assume that for every lift $\tilde{\gamma}$ of $\gamma$ to **H** there is some non-trivial element $T$ of $\Gamma'$ such that $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$.

It is evident from Theorem 3.4 that we can replace $\gamma$ by any geodesic whose cutting sequence is an image of $\mathbf{S}(\gamma)$ under an automorphism of $\Gamma'$ without changing the conclusion of the theorem. We also know from Remark 2.4 that the half-derived sequence of $\mathbf{S}(\gamma)$ can be obtained by applying an automorphism of $\Gamma'$ to $\mathbf{S}(\gamma)$. Hence we can replace $\mathbf{S}(\gamma)$ by any of the sequences it is half-derivable to. It now follows from Theorem 2.10 that we may assume either $\mathbf{S}(\gamma)$ is half-derivable infinitely often or it is one of

$$(3.9) \qquad (Y^{-1})^{\infty}\, Z^{-1}Y^{-1}Z\, Y^{\infty} \qquad \text{or} \qquad (Z^{-1})^{\infty}\, Y^{-1}ZY\, Z^{\infty}$$

or one of

$$(3.10) \qquad (Z^{-1}Y^{-1})^{\infty}\, (ZY)^{\infty} \qquad \text{or} \qquad (Y^{-1})^{\infty} ZY^{\infty},$$

where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$.

We shall deal with the case where $\mathbf{S}(\gamma)$ is half-derivable infinitely often first. In this case, with reference to Definitions 2.6 and 2.7 and by noting that periodic singly infinite sequences can only be derived a finite number of times it is clear that $\mathbf{S}(\gamma)$ is of the form

$$\mathbf{S} = \ldots\ldots X_3^{-1}X_2^{-1}X_1^{-1}\; Z^{-1}Y^{-1}ZY\; X_1X_2X_3\ldots\ldots$$

where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ and each $X_i$ is $Y$ or $Z$ and $X_1X_2X_3\ldots$ is an aperiodic $O$-radial sequence. Every aperiodic $O$-radial sequence arises as described in Case 1 of the section on half-linear sequences in Chapter 2. Hence we may assume there is some $O$-ray $r$ emanating from $O$ such that

$$\mathbf{S}(r) = X_1X_2X_3\ldots.$$

By applying the appropriate element of the set

$$\{\mathrm{Id},\ R,\ R^2,\ R^3,\ P,\ PR,\ PR^2,\ PR^3\} \tag{3.11}$$

to $\mathbf{S}(\gamma)$ we may also assume $Y = A^{-1}$ and $Z = B^{-1}$.

Now let $\tilde{\gamma}$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes the pair $AB^{-1}$ to the word $BAB^{-1}A^{-1}$ in $\mathbf{S}(\tilde{\gamma}) = \mathbf{S}(\gamma)$ as it does so. We shall examine the projection of $\tilde{\gamma}$ to $\mathbf{P}$ by Cohn's commutator map $\sigma_1$. With reference to the description of $\sigma_1$ given in Chapter 1, it can be seen that $\sigma_1(\tilde{\gamma})$ circles around $O$ in an anti-clockwise direction contributing the word $BAB^{-1}A^{-1}$ to $\mathbf{S}(\sigma_1(\tilde{\gamma})) = \mathbf{S}(\gamma)$ as it does so. The remaining right and left portions of $\mathbf{S}(\sigma_1(\tilde{\gamma}))$ are $\mathbf{S}(r)$ and its 'inverse', respectively. Therefore $\sigma_1(\tilde{\gamma})$ follows the reverse of $r$ through the lattice $\Sigma'(O)$ until it reaches $O$, at which point it circles around $O$ in an anti-clockwise direction and then follows $r$ back out through the lattice. Consequently, there is a lift $\tilde{r}$ of $r$ to $\mathbf{H}$ which emanates from $\infty$ and follows the portion of $\tilde{\gamma}$ with cutting sequence $\mathbf{S}(r)$ through the grid $\Lambda$. Similarly, there is a lift $\tilde{r}'$ of $r$ which emanates from $\infty$ and follows the portion of the reverse of $\tilde{\gamma}$ with cutting sequence $\mathbf{S}(r)$. In other words, if we write $\tilde{\gamma} = [\eta, \xi]$ then there are lifts $\tilde{r}$ and $\tilde{r}'$ of $r$ which join $\infty$ to $\xi$ and $\eta$, respectively. We know from the properties of $\sigma_1$ that $\tilde{r}'$ is $\Gamma''$-equivalent to $\tilde{r}$. It is not hard to see that $\tilde{r}'$ is the image of $\tilde{r}'$ under the translation $BAB^{-1}A^{-1}(z) = z - 6$ and therefore $\tilde{\gamma}$ has diameter 6.

We are assuming there is $T \in \Gamma'$ such that $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$. Hence, on the real axis, the endpoints $T(\eta)$ and $T(\xi)$ of $T(\tilde{\gamma})$ separate and are separated by the endpoints $\eta$ and $\xi$ of $\tilde{\gamma}$. Now observe that $T(\tilde{r}')$ joins $T(\infty)$ to $T(\eta)$ and $T(\tilde{r})$ joins $T(\infty)$ to $T(\xi)$. We know $T(\infty) \neq \infty$ since $\tilde{\gamma}$ has diameter 6 and the stabiliser of $\infty$ in $\Gamma'$ is the cyclic subgroup generated by translation by $ABA^{-1}B^{-1}(z) = z + 6$. Thus one of the points $T(\eta)$ or $T(\xi)$ and the point $T(\infty)$ separate and are separated by the points $\eta$ and $\xi$. It follows that one of $T(\tilde{r})$ or $T(\tilde{r}')$ crosses one of $\tilde{r}$ or $\tilde{r}'$. This intersection property is preserved by the projection $\sigma_1$. However, the pair $\tilde{r}$ and $\tilde{r}'$ project to $r$ and the pair $T(\tilde{r})$ and $T(\tilde{r}')$ project to $t(r)$ where $t \in \Sigma'$ is the image of $T \in \Gamma'$ under the isomorphism $\pi$ defined in (1.25). We know $\Sigma'$ is

generated by the translations (1.22) and so $t$ is a translation. Since $t(r)$ and $r$ are parallel they cannot intersect and we have the required contradiction.

Next we assume $\mathbf{S}(\gamma)$ is one of the sequences displayed in (3.9) where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. By reversing the orientation of $\gamma$ (which does not change its self-intersection number and replaces $\mathbf{S}(\gamma)$ by its 'inverse') and then applying the automorphism $G$ defined by $G(Y) = Z$ and $G(Z) = Y$ to $\mathbf{S}(\gamma)$ we may assume that $\mathbf{S}(\gamma)$ is the sequence $(Y^{-1})^\infty\, Z^{-1}Y^{-1}ZY\, Y^\infty$. As usual, by applying the appropriate element of (3.11) we may also assume $Y = A^{-1}$ and $Z = B^{-1}$. In other words, we may assume

$$\mathbf{S}(\gamma) = A^\infty\, BAB^{-1}\, (A^{-1})^\infty. \tag{3.12}$$

As above, we let $\tilde{\gamma} = [\eta, \xi]$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes the pair $AB^{-1}$ to the word $BAB^{-1}A^{-1}$ in $\mathbf{S}(\tilde{\gamma}) = \mathbf{S}(\gamma)$ as it does so and we examine the projection of $\tilde{\gamma}$ to $\mathbf{P}$ by $\sigma_1$. Again, $\sigma_1(\tilde{\gamma})$ circles around $O$ in an anti-clockwise direction contributing $BAB^{-1}A^{-1}$ to $\mathbf{S}(\sigma_1(\tilde{\gamma})) = \mathbf{S}(\gamma)$. This time the remaining left and right portions of $\mathbf{S}(\sigma_1(\tilde{\gamma}))$ are $A^\infty$ and $(A^{-1})^\infty$, respectively. Let $r$ be the ray in $\sigma_1(\Lambda) \cup \Sigma'(O)$ which emanates from $O$ and faces north west. (Equivalently, $r$ is the ray in $\mathbf{C}$ emanating from $O$ which passes through the point $a^{-1}(O)$ where $a$ is defined by (1.22).) The portion of $\sigma_1(\tilde{\gamma})$ with cutting sequence $(A^{-1})^\infty$ traverses, in the same direction as $r$, the chain of tiles in $\mathbf{P}$ which lies above and borders $r$ and the portion of $\sigma_1(\tilde{\gamma})$ with cutting sequence $A^\infty$ traverses this same chain of tiles in the opposite direction. We can lift this chain of tiles to similar chains of tiles in $\mathbf{H}$. Obviously there is such a chain which contains the portion of $\tilde{\gamma}$ with cutting sequence $(A^{-1})^\infty$ and another chain which contains the portion of $\tilde{\gamma}$ with cutting sequence $A^\infty$. By considering the edges of the first chain of tiles it can be deduced that $r$ lifts to a chain of geodesics $\{\tilde{r}_i\}_{i=1}^\infty$ in the grid $\Lambda$ with the property that each $\tilde{r}_i = [P_i, P_{i+1}]$ where $P_1 = \infty$ and the sequence $P_i$ converges to the endpoint $\xi$ of $\tilde{\gamma}$. Similarly, consideration of the second chain of tiles shows $r$ also lifts to a chain of geodesics $\{\tilde{r}'_i\}_{i=1}^\infty$ in $\Lambda$ such that $\tilde{r}'_i = [P'_i, P'_{i+1}]$ where $P'_1 = \infty$ and the sequence $P'_i$ converges to the endpoint $\eta$ of $\tilde{\gamma}$. We shall denote these chains by $\tilde{r}$ and $\tilde{r}'$, respectively. As above, it is not hard to see that the second chain of tiles is

the image of the first under the translation $BAB^{-1}A^{-1}(z) = z - 6$. Thus $\tilde{r}'$ is the translate of $\tilde{r}$ to the left by 6 and $\tilde{\gamma}$ has diameter 6. We remark that the sequences $P_i$ and $P_i'$ converge to $\xi$ and $\eta$, respectively, from above.

Now suppose there is $T \in \Gamma'$ such that $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$. By arguing as above we can assume $T(\infty) \neq \infty$ and one of the points $T(\eta)$ or $T(\xi)$ together with the point $T(\infty)$ separate and are separated by the points $\eta$ and $\xi$. Since the endpoints of the chains $T(\tilde{r})$ and $T(\tilde{r}')$ start at $T(\infty)$ and converge to $T(\xi)$ and $T(\eta)$, respectively, it follows that one of the chains $T(\tilde{r})$ or $T(\tilde{r}')$ crosses one of $\tilde{r}$ or $\tilde{r}'$. Of course, we must allow the possibility that the chains cross at a point of $\mathbf{R}$ rather than at a point in $\mathbf{H}$. This possibility causes no difficulty because no matter how the crossing occurs, it is clear that one of the chains $T(\tilde{r})$ or $T(\tilde{r}')$ enters or leaves one of the chains of tiles which neighbour $\tilde{r}$ and $\tilde{r}'$. By projecting to $\mathbf{P}$ we can conclude as before that $t(r)$ crosses $r$ where $t = \pi(T)$. Again, this is impossible since $t(r)$ is parallel to $r$ and we have the required contradiction.

The final possibility is that $\mathbf{S}(\gamma)$ is one of (3.10) where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$. Since the image of $(Z^{-1}Y^{-1})^{\infty}$ $(ZY)^{\infty}$ under the automorphism $G$ define by $G(Y) = Z^{-1}Y$ and $G(Z) = Z$ is $(Y^{-1})^{\infty}ZY^{\infty}$ we may assume $\mathbf{S}(\gamma)$ is the latter sequence. As usual, we also assume $Y = A^{-1}$ and $Z = B^{-1}$ so that

$$\mathbf{S}(\gamma) = A^{\infty}\ B^{-1}\ (A^{-1})^{\infty}. \tag{3.13}$$

We choose $\tilde{\gamma} = [\eta, \xi]$ be the lift of $\gamma$ to $\mathbf{H}$ which crosses $\mathcal{D}$ and contributes the pair $AB^{-1}$ to the word $BAB^{-1}A^{-1}$ in $\mathbf{S}(\tilde{\gamma}) = \mathbf{S}(\gamma)$ as it does so. In this case, the geodesic $\sigma_1(\tilde{\gamma})$ only contributes $AB^{-1}A^{-1}$ to its cutting sequence as it circles around $O$. The remaining left and right portions of $\mathbf{S}(\sigma_1(\tilde{\gamma})) = \mathbf{S}(\gamma)$ are $A^{\infty}$ and $(A^{-1})^{\infty}$, respectively. As in the second case, we take $r$ to be the ray in $\sigma_1(\Lambda) \cup \Sigma'(O)$ which emanates from $O$ and faces north west. Again, the portion of $\sigma_1(\tilde{\gamma})$ with cutting sequence $(A^{-1})^{\infty}$ traverses, in the same direction as $r$, the chain of tiles which lies above and borders $r$. This time however, the portion of $\sigma_1(\tilde{\gamma})$ with cutting sequence $A^{\infty}$ traverses, in the opposite direction to $r$, the chain of tiles which lies below and borders $r$. Apart from this difference the argument proceeds in much the same way as it did before. Thus the chain of tiles lying above $r$ lifts

to a chain of tiles in $\mathbf{H}$ which contains the portion of $\tilde{\gamma}$ with cutting sequence $(A^{-1})^{\infty}$ and the chain of tiles below $r$ lifts to a chain which contains the portion of $\tilde{\gamma}$ with cutting sequence $A^{\infty}$. Note that the two chains of tiles in $\mathbf{H}$, being lifts of different chains in $\mathbf{P}$, are certainly not $\Gamma''$-equivalent. As before, by considering the appropriate edges of these chains of tiles it can be seen that $r$ lifts to a chain $\tilde{r}$ of geodesics in $\Lambda$ which starts at $\infty$ and converges to $\xi$ and also to a chain $\tilde{r}'$ which starts at $\infty$ and converges to $\eta$. A straightforward calculation shows $\tilde{\gamma}$ has diameter less than 6. We remark that while the points $P_i$ converge to $\xi$ from above as before, the points $P_i'$ converge to $\eta$ from below.

As usual, we suppose there is $T \in \Gamma'$ such that $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$. Since $\gamma$ has diameter less than 6 we can assume $T(\infty) \neq \infty$. The argument of the second case now applies and we conclude that one of the chains $T(\tilde{r})$ or $T(\tilde{r}')$ enters or leaves one of the chains of tiles which neighbour $\tilde{r}$ and $\tilde{r}'$. By considering the projection of this situation to $\mathbf{P}$ we produce the usual contradiction. $\square$

The second half of the proof of Theorem 3.6 consists of showing that the open geodesics $\gamma$ on $\mathbf{T}$ with cutting sequences (3.12) and (3.13) are simple. The proof we have presented involves the use of Cohn's commutator map $\sigma_1$. We believe that the details of the proof provide useful insight into the properties of $\sigma_1$ and especially in relation to its application to problems arising in connection with the Markoff spectrum. However, there is a simpler alternative. It is not hard to demonstrate smooth curves on $\mathbf{T}$ with cutting sequences equal to (3.12) and (3.13) which are simple, see for instance [36]. The existence of such curves implies that the associated geodesics are simple. The reason is outlined in the next remark.

**Remark 3.2.** In this remark, we discuss open curves on $\mathbf{T}$ other than geodesics. Of course, we have not defined cutting sequences for such curves and our intention here is only to convey an idea. Let $\gamma$ be an open geodesic on $\mathbf{T}$ which does not cover a closed geodesic and suppose $\delta$ is a 'well-behaved' open curve on $\mathbf{T}$ whose cutting sequence is equal to $\mathbf{S}(\gamma)$. For each lift $\tilde{\gamma} = [\eta, \xi]$ of $\gamma$ to $\mathbf{H}$ there is a lift $\tilde{\delta}$ of $\delta$ to $\mathbf{H}$ which also has endpoints $\eta$ and $\xi$. We know that a geodesic $T(\tilde{\gamma})$, where $T \in \Gamma'$, intersects $\tilde{\gamma}$ if and only if the endpoints of $T(\tilde{\gamma})$ and $\tilde{\gamma}$ separate one another

on the real axis. It follows that if $T(\tilde{\gamma})$ intersects $\tilde{\gamma}$ then $T(\tilde{\delta})$ intersects $\tilde{\delta}$. Since the self-intersection numbers of $\gamma$ and $\delta$ can be expressed in terms of the number of curves which are $\Gamma'$-equivalent to and intersect $\tilde{\gamma}$ and $\tilde{\delta}$ we conclude that the number of self-intersections of $\delta$ must be greater than or equal to that of $\gamma$.

**Remark 3.3.** We complete this chapter by demonstrating that the Markoff values of the simple open geodesics $\mathbf{T}$ with half-linear cutting sequences lie in Hall's ray. Let $\gamma$ be such a geodesic and let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be the associated doubly infinite sequence of positive integers. Our aim is to show that $M(\mathcal{A})$ lies in Hall's ray. We shall do this by using the algorithm described in Chapter 2 to establish certain properties of $\mathcal{A}$ from the properties of $\mathbf{S}(\gamma)$. To this end recall that we can replace $\mathbf{S}(\gamma)$ by any of its images under $\Psi$ without changing $\mathcal{A}$. Since $\mathbf{S}(\gamma)$ half-linear we know it contains a block of the form $Z^{-1}Y^{-1}ZY$ or $Y^{-1}Y^{-1}ZYY$ where $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. By applying the appropriate element of $\Psi$ we may assume either $B^{-1}A^{-1}BA$ or $B^{-1}ABA^{-1}$ occurs. It follows, with reference to Table 2.1, that the combined cutting and $LR$-sequence of $\gamma$ contains the blocks

$$B^{-1}LA^{-1}LLBLA \qquad \text{or} \qquad B^{-1}RRARBRRA^{-1},$$

respectively. Hence the $LR$-sequence of $\gamma$ contains a block of the form $L^4$ or $R^5$. In the latter case, we know some $a_i$ is 5 or more and so $M(\mathcal{A}) \geq \lambda_i(\mathcal{A}) > 5$ and we are done. Thus we assume $L^4$ occurs. Clearly we can also assume that it is preceded and followed by $R$. The block $L^4R^3$ cannot occur else we have

$$B^{-1}LA^{-1}LLBLARBRRA^{-1}$$

implying that $B^{-1}A^{-1}BABA^{-1}$ occurs in $\mathbf{S}(\gamma)$ which contradicts the fact that $\mathbf{S}(\gamma)$ is half-linear. Similarly, the block $R^3L^4$ cannot occur else

$$BRRA^{-1}RB^{-1}LA^{-1}LLBLA$$

implying that $BA^{-1}B^{-1}A^{-1}BA$ occurs in $\mathbf{S}(\gamma)$. We conclude that there is some index $i$ such that $a_i = 4$ and $a_{i-1} \leq 2$ and $a_{i+1} \leq 2$. In this case,

$$M(\mathcal{A}) \geq \lambda_i(\mathcal{A}) > [4, 2, 1] + [0, 2, 1] = 4.666...$$

and $M(\mathcal{A})$ lies in Hall's ray.

CHAPTER 4

# A RIGHT TRANSVERSAL FOR $\Psi$ IN Aut $\Gamma'$

We begin this chapter by motivating the need for a right transversal for $\Psi$ in Aut $\Gamma'$. We saw in the last chapter that the proper closed 1-intersectors on $\mathbf{T}$ are defined by the primitive conjugacy classes of the form $[G(V)]$ where

$$V = ABAB^{-1} \qquad \text{and} \qquad G \in \text{Aut } \Gamma'.$$

In the next chapter we shall calculate the representatives $f_W$ where $W = G(V)$ of the corresponding classes of forms and their Markoff values. However, redundancies will be encountered in this process. In particular, we know from the section of Chapter 1 dealing with the automorphisms of $\Gamma'$ that if $G \in \Psi$ then $f_{G(V)}$ is equivalent to $f_V$. Since we shall only want one representative for each equivalence forms we shall only need to consider the conjugacy classes of the form $[G(V)]$ where $G$ lies in a right transversal for $\Psi$ in Aut $\Gamma'$. We remark that $\Psi$ is not normal in Aut $\Gamma'$.

In order to describe the transversals we shall be using we introduce the semigroup $\Omega$ of Aut $\Gamma'$ generated by

$$SR(A,B) = (AB,\ B) \qquad \text{and} \qquad S^2R^3\,(A,B) = (A,\ AB).$$

Specifically, we define

$$\Omega = \{G_1G_2 \ldots G_n \ :\ n \geq 0,\ G_i = SR \text{ or } S^2R^3 \text{ for } i = 1,2,\ldots,n\}. \tag{4.1}$$

This set is most easily comprehended when arranged as a tree. We use two arrangements, the first is shown in Figure 4.1 and the second in Figure 4.2. These arrangements arise from the two methods of recursively building the typical element $G_1G_2 \ldots G_n$ of $\Omega$ from the identity. In the first tree, we view $G_1G_2 \ldots G_n$ as

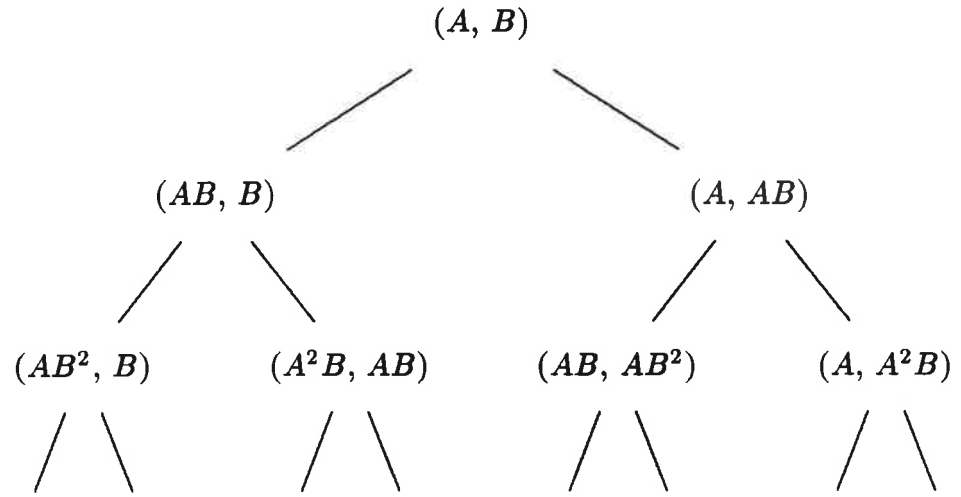


FIGURE 4.1. The set $\Omega$ of automorphisms of $\Gamma'$. The tree continues by substituting $AB$ for $A$ to branch left and $AB$ for $B$ to branch right.

being the result of composing $G_1$ with $G = G_2G_3 \dots G_n$. When $G_1 = SR$ the tree branches left and when $G_1 = S^2R^3$ it branches right. The branching operations themselves are a consequence of the fact the automorphism $SR$ may be applied to an element $W$ of $\Gamma'$ by writing $W$ as a word in the generators $A$ and $B$ of $\Gamma'$ and then substituting $AB$ for $A$ everywhere in $W$, and similarly, $S^2R^3$ may be applied to $W$ by substituting $AB$ for $B$ everywhere. In the second tree, we view the element $G_1G_2 \dots G_n$ as being the result of composing $G = G_1G_2 \dots G_{n-1}$ with $G_n$. Again, when $G_n = SR$ the tree branches left and when $G_n = S^2R^3$ it branches right. This time the branching operations are a consequence of

$$GSR(A, B) = (G(A)G(B),\ G(B))$$

and

$$GS^2R^3(A, B) = (G(A),\ G(A)G(B)).$$

In our main result of this chapter, Theorem 4.1, we shall show that the set

$$\mathcal{T} = \{RG\ :\ G \in \Omega\} \cup \{RGR\ :\ G \in \Omega\}$$

of automorphisms is a right transversal for $\Psi$ in Aut $\Gamma'$. The proof of Theorem 4.1 relies on the following three lemmas. While the second lemma is not essential to the proof we require it in Chapter 5.

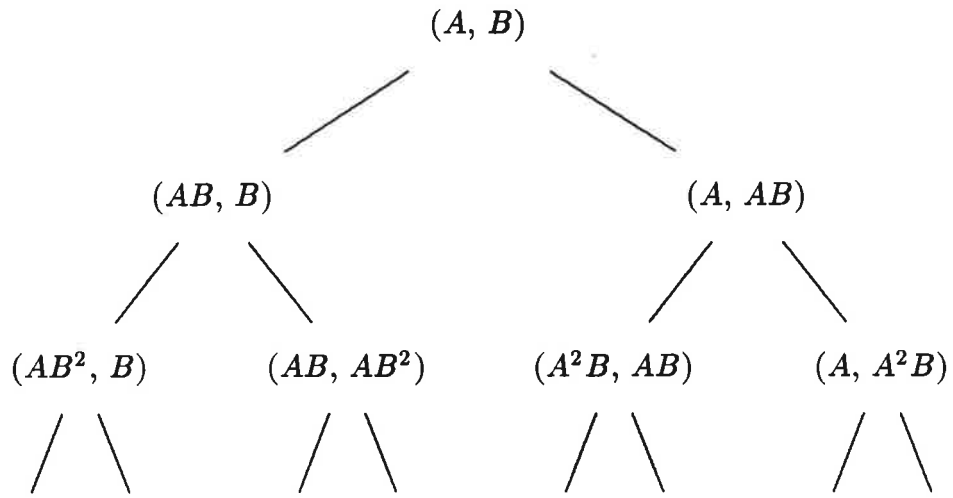


FIGURE 4.2. Another arrangement of the set $\Omega$ as a tree. The tree continues from the node $(W_1, W_2)$ by forming $(W_1W_2, W_2)$ to branch left and $(W_1, W_1W_2)$ to branch right.

**Lemma 4.1.** *The coset equality $\Psi\, G\, R^2 = \Psi\, G$ holds for all $G \in$ Aut $\Gamma'$.*

*Proof.* We know from Chapter 1 that $P$, $R$ and $S$ generate Aut $\Gamma'$ and that $P^{-1} = P$, $R^{-1} = R^3$ and $S^{-1} = S^2$. Therefore, any $G \in$ Aut $\Gamma'$ can be written as

$$G = G_1 G_2 \ldots G_n$$

where each $G_i$ is one of $P$, $R$ or $S$. We prove the lemma for all such words by using induction on the length $n$.

For $n = 0$ or 1 the truth of the lemma is trivial since $P$, $R^2$ and $S$ belong to $\Psi$. For $n = 2$ the lemma is again trivial except when $G = RP$ and $G = RS$. For the first exception, using the relation $PR^2 = R^2P$, we have

$$\Psi\, G\, R^2 = \Psi\, RPR^2 = \Psi\, RR^2P = \Psi\, G$$

and for the second, using the relation $RSR^2 = R^2S^2R^2SR^3S$, we have

$$\Psi\, G\, R^2 = \Psi\, RSR^2 = \Psi\, R^2S^2R^2SR^3S = \Psi\, RS = \Psi\, G.$$

In either case, the lemma is true.

Now suppose $n \geq 3$ and the lemma is true for all words of length less than $n$. We must establish the lemma for $G = G_1 G_2 \dots G_n$. If $G_n = R$ the truth of the lemma follows directly from our inductive hypothesis and if $G_n = P$ it follows after using the relation $PR^2 = R^2P$. Thus we assume $G_n = S$ and consider the possibilities for $G_{n-1}$. It will be convenient to write

$$H_i = G_1 G_2 \dots G_i, \qquad 1 \leq i \leq n.$$

First we suppose $G_{n-1} = S$. If also $G_{n-2} = S$ then since $S^3 = 1$ we know $G = H_{n-3}$ and our inductive hypothesis shows the lemma is true. If $G_{n-2} = R$ then using the relation $RS^2R^2 = SR^2S^2RS^2$ we have

$$\Psi\, G\, R^2 = \Psi\, H_{n-3}\, RS^2R^2 = \Psi\, H_{n-3}\, SR^2S^2RS^2$$

and by induction

$$\Psi\, G\, R^2 = \Psi\, H_{n-3}\, S^3RS^2 = \Psi\, H_{n-3}\, RS^2 = \Psi\, G$$

and so the lemma is true. The only remaining possibility is $G_{n-2} = P$ in which case using the relations of Aut $\Gamma'$ and our inductive hypothesis we have

$$\begin{aligned} \Psi\, G\, R^2 &= \Psi\, H_{n-3}\, PS^2R^2 = \Psi\, H_{n-3}\, R^2SP \\ &= \Psi\, H_{n-3}\, SPR^2 = \Psi\, H_{n-3}\, R^2PS^2 \\ &= \Psi\, H_{n-3}\, PS^2 = \Psi\, G \end{aligned}$$

and again the lemma is true.

Next we suppose $G_{n-1} = R$. Here the lemma follows from

$$\begin{aligned} \Psi\, G\, R^2 &= \Psi\, H_{n-2}\, RSR^2 = \Psi\, H_{n-2}\, R^2S^2R^2SR^3S \\ &= \Psi\, H_{n-2}\, S^3R^3S = \Psi\, H_{n-2}\, R^3S \\ &= \Psi\, H_{n-2}\, RS = \Psi\, G. \end{aligned}$$

The only remaining possibility is that $G_{n-1} = P$ in which case

$$\begin{aligned} \Psi\, G\, R^2 &= \Psi\, H_{n-2}\, PSR^2 = \Psi\, H_{n-2}\, R^2S^2P \\ &= \Psi\, H_{n-2}\, S^2R^2P = \Psi\, H_{n-2}\, R^2PS \\ &= \Psi\, H_{n-2}\, PS = \Psi\, G. \end{aligned}$$

The proof of the lemma is complete. $\square$

**Lemma 4.2.** *The identities*

$$\Psi\, G\, SRP = \Psi\, G\, P\, S^2R^3 \qquad \text{and} \qquad \Psi\, G\, S^2R^3\, P = \Psi\, G\, P\, SR$$

*hold for every automorphism $G$ in Aut $\Gamma'$.*

*Proof.* Let $G \in$ Aut $\Gamma'$. Using the relations of Aut $\Gamma'$ and Lemma 4.1 we have

$$\Psi\, G\, SRP = \Psi\, G\, SPR^3 = \Psi\, G\, R^2PS^2R = \Psi\, G\, P\, S^2R^3.$$

Similarly,

$$\Psi\, G\, S^2R^3\, P = \Psi\, G\, S^2PR = \Psi\, G\, PR^2SR^3 = \Psi\, G\, P\, SR$$

and we are done. □

**Lemma 4.3.** *The set of automorphisms of $\Gamma'$ defined by*

$$\mathcal{T} = \{RG \,:\, G \in \Omega\} \cup \{RGR \,:\, G \in \Omega\}$$

*contains a right transversal for $\Psi$ in Aut $\Gamma'$.*

*Proof.* Let $G$ be an arbitrary element of Aut $\Gamma'$. We must show there is $G' \in \mathcal{T}$ such that $\Psi\, G = \Psi\, G'$. As stated in the proof of Lemma 4.1, we can write

$$G = G_1G_2 \ldots G_k$$

where each $G_i$ is one of $P$, $R$ or $S$. By repeatedly using the relations $RP = PR^3$ and $SP = PR^2S^2R^2$ of Aut $\Gamma'$, any occurrence of $P$ in $G_1G_2 \ldots G_k$ may be shifted to the left hand end without changing the coset $\Psi\, G$. Since $P \in \Psi$ it follows that we may assume $P$ does not occur at all. Likewise, according to Lemma 4.1, we can successively remove all occurrences of $R^2$ without changing the coset $\Psi\, G$. Any remaining occurrences of $S^3$ can be deleted since $S^3 =$ Id. Thus we can write

$$\Psi\, G = \Psi\, G_1'G_2' \ldots G_l'$$

where each $G_i'$ is either $R$ or $S$ and there are no occurrences of $S^3$ or $R^2$. In addition, since $S \in \Psi$, we may assume that either $l = 0$ or $G_1' = R$. It is not hard to deduce now that either $\Psi\, G = \Psi$ or $\Psi\, G$ is of the form

$$\Psi\, G = \Psi\, RS^{j(1)}RS^{j(2)} \ldots RS^{j(n)} \tag{4.2}$$

where $n \geq 1$ and each $j(i)$ is 1 or 2 except $j(n)$ which may also be 0. Using Lemma 4.1, we can insert in an $R^2$ immediately to the right of each $S^2$ in (4.2). If $j(n) \neq 0$ we also add an additional $R^2$ after the term $S^{j(n)}$. It follows that if we let $G_i'' = SR$ when $j(i) = 1$ and $G_i'' = S^2R^3$ when $j(i) = 2$ then we have

$$\Psi\, G = \Psi\, R G_1'' G_2'' \ldots G_{n-1}''$$

if $j(n) = 0$ and

$$\Psi\, G = \Psi\, R G_1'' G_2'' \ldots G_n'' R$$

otherwise. The case $\Psi\, G = \Psi$ is covered also since $\Psi = \Psi\, R \mathrm{Id}\, R$. The proof of the lemma is complete. □

**Theorem 4.1.** *The set of automorphisms of* $\Gamma'$ *defined by*

$$\mathcal{T} = \{RG \,:\, G \in \Omega\} \cup \{RGR \,:\, G \in \Omega\} \tag{4.3}$$

*is a right transversal for* $\Psi$ *in Aut* $\Gamma'$.

*Proof.* We know already from Lemma 4.3 that the automorphisms described include representatives of every coset. It remains to show that they represent distinct cosets. We do this by first showing that if two cosets are identical then there is some $G \in \Omega$ such that $\Psi\, RG = \Psi$. We then complete the proof by showing that this is impossible.

There are three ways two cosets can be identical. The first possibility we shall consider is that there is $n \geq 0$ and $m \geq 0$ such that

$$\Psi\, R G_1 G_2 \ldots G_n = \Psi\, R G_1' G_2' \ldots G_m' R$$

where each $G_i$ and $G_i'$ is either $SR$ or $S^2R^3$. If $m = 0$ then

$$\Psi\, R G_1 G_2 \ldots G_n = \Psi\, RR = \Psi$$

and we are done. Thus we assume $m \neq 0$. If $G_m' = SR$ then

$$\Psi\, R G_1 G_2 \ldots G_n = \Psi\, R G_1' G_2' \ldots G_{m-1}' SRR$$

which implies, after composition with $S^2R^3 = S^2R^2R$ and an application of Lemma 4.1, that

$$\Psi\, RG_1G_2 \ldots G_n\, S^2R^3 = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, R.$$

While, if $G_m' = S^2R^3$ then

$$\Psi\, RG_1G_2 \ldots G_n = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, S^2R^3R$$

which implies, after composition with $SR$, that

$$\Psi\, RG_1G_2 \ldots G_n\, SR = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, R.$$

In either case we have

$$\Psi\, RG_1G_2 \ldots G_nG_{n+1} = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, R$$

where $G_{n+1} = SR$ or $G_{n+1} = S^2R^3$ and we can repeat the argument. Eventually $m = 0$ will arise. It follows, as claimed, that $\Psi\, RG = \Psi$ for some $G \in \Omega$.

The next possibility we consider is that there is $n \geq 0$ and $m \geq 0$ such that

$$\Psi\, RG_1G_2 \ldots G_n = \Psi\, RG_1'G_2' \ldots G_m'$$

where as usual each $G_i$ and $G_i'$ is $SR$ or $S^2R^3$. We cancel $G_n$ and $G_m'$ if they are equal. By repeating this cancellation as often as is necessary we may assume that either $G_n \neq G_m'$ and $n \geq 1$ and $m \geq 1$ or exactly one of $n$ or $m$ is 0. In the first case we can assume without loss of generality that $G_n = SR$ and $G_m' = S^2R^3$, so that,

$$\Psi\, RG_1G_2 \ldots G_{n-1}\, SR = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, S^2R^3.$$

Using Lemma 4.1 to remove $R^2$, then cancelling $SR$ and using Lemma 4.1 again we have

$$\Psi\, RG_1G_2 \ldots G_{n-1} = \Psi\, RG_1'G_2' \ldots G_{m-1}'\, SRR$$

and we can apply the argument above to obtain the desired result. In the second case we can assume without loss of generality that $m = 0$ and $n \neq 0$. If $G_n = SR$ then

$$\Psi\, RG_1G_2 \ldots G_{n-1}\, SR = \Psi\, R$$

and by cancelling $R$ and composing with $S^2$ we have $\Psi\ RG_1G_2\dots G_{n-1} = \Psi$ while if $G_n = S^2R^3$ then

$$\Psi\ RG_1G_2\dots G_{n-1}\,S^2R^3 = \Psi\ R$$

and by cancelling $R$, applying Lemma 4.1 and composing with $S$ we again find that $\Psi\ RG_1G_2\dots G_{n-1} = \Psi$. In all cases we have the desired result.

The final possibility is that there is $n \geq 0$ and $m \geq 0$ such that

$$\Psi\ RG_1G_2\dots G_n\,R = \Psi\ RG_1'G_2'\dots G_m'\,R$$

where each $G_i$ and $G_i'$ is $SR$ or $S^2R^3$. Cancelling $R$ we have

$$\Psi\ RG_1G_2\dots G_n = \Psi\ RG_1'G_2'\dots G_m'$$

and hence can apply the argument just completed.

To complete the proof we need to show there is no automorphism $G \in \Omega$ such that $\Psi\ RG = \Psi$. Since $S \in \Psi$ it is equivalent to show that there is no $G \in \Omega$ such that $\Psi\ SRG = \Psi$. Clearly it suffices to prove this only for the relevant images under the natural homomorphism

$$\theta: \quad \text{Aut } \Gamma' \longrightarrow GL(2,\mathbf{Z}). \tag{4.4}$$

This homomorphism is as described by Cohn in [9] except that, since we are interpreting the composition $GH$ of automorphisms as $H$ followed by $G$ rather than $G$ followed by $H$, we must use the transpose of the matrices Cohn describes. The image of $\Psi$ is

$$\theta(\Psi) = \left\{ \begin{array}{l} \pm\begin{pmatrix}1&0\\0&1\end{pmatrix}, \pm\begin{pmatrix}0&-1\\1&-1\end{pmatrix}, \pm\begin{pmatrix}-1&1\\-1&0\end{pmatrix} \\ \pm\begin{pmatrix}0&1\\1&0\end{pmatrix}, \pm\begin{pmatrix}-1&0\\-1&1\end{pmatrix}, \pm\begin{pmatrix}1&-1\\0&-1\end{pmatrix} \end{array} \right\} \cong \Psi/\text{Inn } \Gamma'. \tag{4.5}$$

The image of the automorphisms of the form $SRG$ where $G \in \Omega$, can be calculated by starting with the image $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$ of $RS$ and post-multiplying successively by the image $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$ of $RS$ and the image $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ of $R^3S^2$. Post-multiplying a matrix by $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$ replaces its first column by the sum of its columns and post-multiplying by

$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ replaces its second column by the sum of its columns. It is clear that all matrices arising in this manner except $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ itself have an entry which is greater than or equal to 2 and hence do not lie in $\theta(\Psi)$. Hence there is no $G \in \Omega$ such that $\theta(SRG) \in \theta(\Psi)$. It follows that there is no $G \in \Omega$ such that $SRG \in \Psi$. The theorem is proved. □

**Remark 4.1.** In the proof of Theorem 4.1, we have shown that each automorphism of the form $RG_1G_2 \dots G_n$, where each $G_i$ is $SR$ or $S^2R^3$, represents a distinct right coset of $\Psi$. Consequently the automorphisms themselves are distinct elements of Aut $\Gamma'$. It follows that for each element $G$ of $\Omega$ there is only one expression of the form $G = G_1G_2 \dots G_n$ where each $G_i$ is $SR$ or $S^2R^3$. In other words, there is no duplication of automorphisms amongst the nodes of either of the trees shown in Figures 4.1 and 4.2.

**Remark 4.2.** The automorphisms in $\mathcal{T}$ are not the most convenient to calculate with. For such purposes either one of the right transversals

$$\mathcal{T}_1 = \{SRG \ : \ G \in \Omega\} \cup \{SRGR^3P \ : \ G \in \Omega\} \tag{4.6}$$

or

$$\mathcal{T}_2 = \{S^2R^3G \ : \ G \in \Omega\} \cup \{S^2R^3GR^3P \ : \ G \in \Omega\} \tag{4.7}$$

is better. That these are transversals can be seen by applying the relation $R^3P = PR$ and Lemma 4.2 to them and then noting that $S, P, R^2 \in \Psi$. We can use the arrangement of the set $\Omega$ shown in Figure 4.2 to explain the structure of these sets. First note that

$$GR^3P(A, B) = (G(A)^{-1}, G(B)).$$

Now observe that when the base node $(A, B)$ of the tree in Figure 4.2 is deleted two trees are left behind. The left hand tree comprises the set $\{SRG \ : \ G \in \Omega\}$ and the right hand tree comprises $\{S^2R^3G \ : \ G \in \Omega\}$. Clearly $\mathcal{T}_1$ consists of all pairs of the form $(W_1, W_2)$ or $(W_1^{-1}, W_2)$ where $(W_1, W_2,)$ lies in the left hand tree. Similarly, $\mathcal{T}_2$ consists of all pairs of the form $(W_1, W_2)$ or $(W_1^{-1}, W_2)$ where

$(W_1, W_2,)$ lies in the right hand tree. We remark that this process is not so simple for the tree in Figure 4.1.

We conclude this chapter with a description of a left transversal for $\Psi$ in Aut $\Gamma'$. We iterate that $\Psi$ is not normal in Aut $\Gamma'$ and the set $\mathcal{T}$ defined in Theorem 4.1 is not a left transversal. However, with the help of an anti-isomorphism of Aut $\Gamma'$ we can obtain such a transversal from $\mathcal{T}$. It is described in Theorem 4.2. Before we state and prove Theorem 4.2 we introduce the anti-isomorphism involved.

**Remark 4.3.** Recall that Aut $\Gamma'$ is generated by $P$, $R$ and $S$ and hence the typical element $G$ can be written in the form $G = G_1 G_2 \dots G_n$ where each $G_i$ is one of $P$, $R$ and $S$ or their inverses. The anti-isomorphism we are referring to is the map from Aut $\Gamma'$ to itself defined by

$$\theta: \quad G = G_1 G_2 \dots G_n \quad \longmapsto \quad \theta(G) = G_n \dots G_2 G_1. \tag{4.8}$$

To see that $\theta$ is well-defined it suffices to check that when the relations defining Aut $\Gamma'$ are reversed, the resulting words are also relations of Aut $\Gamma'$. The only relations changed by reversing are $(RP)^2 = \mathrm{Id}$ and $(R^2SP)^2 = \mathrm{Id}$. They become $(PR)^2 = \mathrm{Id}$ and $(PSR^2)^2 = \mathrm{Id}$, respectively. Conjugating the first by $P$ yields $(RP)^2 = \mathrm{Id}$ again, whilst conjugating the second by $R^2P$ yields

$$R^2SR^2PSR^2PR^2 = \mathrm{Id}$$

which is a consequence of $(R^2SP)^2 = \mathrm{Id}$ since $R^2P = PR^2$ and $R^4 = \mathrm{Id}$. Thus $\theta$ is a well-defined map from Aut $\Gamma'$ to itself. It is onto since it maps the generators $P$, $R$ and $S$ to themselves and it is a bijection since $\theta^2$ is the identity. Obviously

$$\theta(GH) = \theta(H)\theta(G)$$

for all $G, H \in$ Aut $\Gamma'$ and hence $\theta$ is an anti-isomorphism. Note that in effect we have shown that the presentation (1.17) of Aut $\Gamma'$ is independent of whether the composition of automorphisms $GH$ is interpreted to mean $H$ followed by $G$ as is our convention, or $G$ followed by $H$ as Cohn does in [9].

**Theorem 4.2.** *The set of automorphisms of $\Gamma'$ defined by*

$$\mathcal{T}' = \Omega \cup \{RG \,:\, G \in \Omega\} \tag{4.9}$$

*constitutes a left transversal for $\Psi$ in Aut $\Gamma'$.*

*Proof.* As indicated in the preamble, we shall prove this theorem by applying the anti-isomorphism $\theta$ defined in Remark 4.3 to the right transversal $\mathcal{T}$ defined by (4.3). The reason we can do this is that $\theta$ maps the generators $P$, $R^2$ and $S$ of $\Psi$ to themselves and hence preserves $\Psi$, that is, $\theta(\Psi) = \Psi$. Because $\theta$ is an anti-isomorphism, it follows immediately that $\theta(\mathcal{T})$ is a left transversal for $\Psi$ in Aut $\Gamma'$. We shall complete the proof by demonstrating a bijection between $\theta(\mathcal{T})$ and $\mathcal{T}'$ which preserves left cosets of $\Psi$. Before we begin note that by applying $\theta$ to the identity in Lemma 4.1 we know

$$R^2\, G\, \Psi = G\, \Psi \tag{4.10}$$

for all automorphisms $G$ in Aut $\Gamma'$.

We consider the elements of $\theta(\mathcal{T})$ of the form $\theta(R\,G)$, where $G \in \Omega$, first. According to the definition of $\Omega$ we can write $G = G_1 G_2 \ldots G_n$ where each $G_i$ is $SR$ or $S^2R^3$. The left coset associated with $\theta(R\,G)$ is

$$\theta(R\,G_1G_2\ldots G_n)\,\Psi = \theta(G_n)\ldots\theta(G_2)\theta(G_1)\,R\,\Psi.$$

We know each $\theta(G_i)$ is $RS$ or $R^3S^2$ depending on whether $G_i$ is $SR$ or $S^2R^3$, respectively. By repeatedly using the identity (4.10) it follows that

$$\theta(R\,G_1G_2\ldots G_n)\,\Psi = G'_n\ldots G'_2G'_1\,R\,\Psi$$

where $G'_i$ is $RS$ or $RS^2R^2$ depending on whether $G_i$ is $SR$ or $S^2R^3$, respectively. It is now clear that

$$\theta(R\,G_1G_2\ldots G_n)\,\Psi = R\,G_n\ldots G_2G_1\,\Psi$$

and we have a bijection between the sets $\{\theta(R\,G) \,:\, G \in \Omega\}$ and $\{RG \,:\, G \in \Omega\}$ which preserves left cosets of $\Psi$.

Next we consider the elements of $\theta(\mathcal{T})$ of the form $\theta(R\,G\,R)$ where $G \in \Omega$. Again, we let $G = G_1 G_2 \ldots G_n$ where each $G_i$ is $SR$ or $S^2R^3$. The left coset associated with $\theta(R\,G\,R)$ is

$$\theta(R\,G_1 G_2 \ldots G_n\,R)\,\Psi = R\,\theta(G_n)\ldots\theta(G_2)\theta(G_1)\,R\,\Psi.$$

By arguing as above, we know that

$$\theta(R\,G_1 G_2 \ldots G_n\,R)\,\Psi = RRG_n \ldots G_2 G_1\,\Psi$$

and hence the identity (4.10) shows

$$\theta(R\,G_1 G_2 \ldots G_n\,R)\,\Psi = G_n \ldots G_2 G_1\,\Psi.$$

Thus we also have a bijection between the sets $\{\theta(R\,G\,R) : G \in \Omega\}$ and $\Omega$ which preserves the left cosets of $\Psi$. By combining this bijection with the first one we obtain the desired bijection between $\theta(\mathcal{T})$ and $\mathcal{T}'$ and the proof is complete. $\square$

CHAPTER 5

# MARKOFF VALUES FOR THE PROPER CLOSED 1–INTERSECTORS

In this chapter we provide the means of calculating the Markoff values of the proper closed 1-intersectors. We achieve this by describing both the associated doubly infinite sequences of positive integers and representatives of the corresponding classes of forms. We do not include the improper closed 1-intersectors because, as was demonstrated in Remark 3.1, their Markoff values are greater than 6 and hence lie in Hall's ray. Although the results for the simple closed geodesics are already known, we outline at the end of the chapter how they may be dealt with in a similar manner. We do this in order to shed more light on Cohn's work.

We have seen in Chapter 3 that the proper closed 1-intersectors on $\mathbf{T}$ are defined by the conjugacy classes in $\Gamma'$ of the form $[G(ABAB^{-1})]$ where $G \in \text{Aut } \Gamma'$. In order to simplify the calculations in this chapter we replace $ABAB^{-1}$ by $A^2B^2$. We are able to do this since

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

Thus the proper closed 1-intersectors are defined by the conjugacy classes

$$[G(A^2B^2)] \qquad \text{where} \qquad G \in \text{Aut } \Gamma'. \tag{5.1}$$

The corresponding equivalence classes of forms are represented by the forms $f_W$ where $W = G(A^2B^2)$. (Recall here that the definition of $f_W$ is given in (1.15) and does not depend on the choice of matrix for $W$.) As mentioned in the introduction to Chapter 4, there is some redundancy amongst these representatives and it suffices to take only the forms $f_W$ where $W = G(A^2B^2)$ and $G$ lies in a right transversal for $\Psi$. However, there is further redundancy in such a set of forms. More restrictions

on the automorphisms $G$ involved are needed. Our next theorem describes a set of automorphisms for which there is no duplication of representatives. In its proof we make use of the right transversal $\mathcal{T}$ described in Theorem 4.1.

**Theorem 5.1.** *The forms $f_W$ where $W = G(A^2B^2)$ for some $G \in \mathcal{K}$ and*

$$\mathcal{K} = \{SR^2S^2, SR^3\} \cup \{SR^3SRG \ : \ G \in \Omega\} \tag{5.2}$$

*represent all classes of forms which map to proper closed 1-intersectors on* $\mathbf{T}$.

*Proof.* It is clear from the discussion above that the theorem will be proven if we can show that for every $G \in \mathcal{T}$ there exist $H, H' \in \Psi$ and $G' \in \mathcal{K}$ such that $HG(A^2B^2) = H'G'(A^2B^2)$. We consider the two possibilities for $G \in \mathcal{T}$ separately.

First suppose $G = R\,G_1G_2 \ldots G_n$ where $n \geq 0$ and each $G_i$ is $SR$ or $S^2R^3$. If $n = 0$ then $H = SR^2$, $H' = \mathrm{Id}$ and $G' = SR^3$ satisfy the required conditions. Similarly, if $n \geq 1$ and $G_1 = SR$ then $H = SR^2$, $H' = \mathrm{Id}$ and $G' = SR^3SR\,G_2G_3 \ldots G_n$ will do. Now assume $n \geq 1$ and $G_1 = S^2R^3$. Let $H$ be conjugation by $G(A^{-2}) = (G(A))^{-2}$ so that

$$HG(A^2B^2) = G(B^2A^2) = G\,P(A^2B^2) = R\,G_1G_2 \ldots G_n\,P(A^2B^2)$$

and note that $H \in \Psi$ since $\mathrm{Inn}\,\Gamma' \subset \Psi$. Using Lemma 4.2 and the relation $RP = PR^3$ we can deduce that

$$\Psi\ R\,G_1G_2 \ldots G_n\,P = \Psi\ PR^3\,G_1'G_2' \ldots G_n' = \Psi\ R\,G_1'G_2' \ldots G_n'$$

where $G_i' = S^2R^3$ if $G_i = SR$ and $G_i' = SR$ if $G_i = S^2R^3$. Hence we can choose $H_1 \in \Psi$ so that

$$HG(A^2B^2) = H_1\,R\,G_1'G_2' \ldots G_n'(A^2B^2).$$

Since $G_1' = SR$ the first part of our argument shows there is $H_2, H_3 \in \Psi$ and $G' \in \mathcal{K}$ such that

$$H_2\,R\,G_1'G_2' \ldots G_n'(A^2B^2) = H_3G'(A^2B^2).$$

It follows that $HG(A^2B^2) = H'G'(A^2B^2)$ where $H' = H_1H_2^{-1}H_3 \in \Psi$, as required.

Now suppose $G = RG_1G_2 \dots G_n R$ where $n \geq 0$ and each $G_i$ is $SR$ or $S^2R^3$. If $n = 0$ then $H = SR^2S^2R^2$, $H' = \mathrm{Id}$ and $G' = SR^2S^2$ will do. Thus we may assume $n \geq 1$. We consider two possibilities. First we suppose that $G_n = SR$. Let $H$ be conjugation by $G(B)$. Then $H \in \Psi$ and

$$HG(A^2B^2) = G(BA^2B) = RG_1G_2 \dots G_{n-1} SR^2(BA^2B)$$

and so

$$HG(A^2B^2) = RG_1G_2 \dots G_{n-1}(AB^{-1}AB).$$

Since $S^2R^3(AB^{-1}AB) = AB^{-1}AB$ we know that either

$$HG(A^2B^2) = R(AB^{-1}AB) = B^{-1}A^{-1}B^{-1}A = S^2(A^2B^2)$$

or there is $j$ with $1 \leq j \leq n-1$ such that $G_j = SR$. In the first instance, $HG(A^2B^2) = H'G'(A^2B^2)$ where $H' = R^2S^2$ and $G' = SR^2S^2$ as required. Thus we assume the latter is true. We choose $j$ to be maximal and observe that

$$HG(A^2B^2) = RG_1G_2 \dots G_j(AB^{-1}AB) = RG_1G_2 \dots G_{j-1}(A^2B^2).$$

In this case, the results already established imply that there are $H_1, H_2 \in \Psi$ and $G' \in \mathcal{K}$ such that

$$H_1 RG_1G_2 \dots G_{j-1}(A^2B^2) = H_2G'(A^2B^2)$$

and hence $HG(A^2B^2) = H'G'(A^2B^2)$ where $H' = H_1^{-1}H_2$.

The other possibility is that $G_n = S^2R^3$. In this case, we let $H \in \Psi$ be conjugation by $G(A^{-2})$ so that

$$HG(A^2B^2) = G(B^2A^2) = GP(A^2B^2) = RG_1G_2 \dots G_n RP(A^2B^2).$$

As above,

$$\Psi\, RG_1G_2 \dots G_n RP = \Psi\, RG_1G_2 \dots G_n PR^3 = \Psi\, RG_1'G_2' \dots G_n' R$$

where $G_i' = S^2R^3$ if $G_i = SR$ and $G_i' = SR$ if $G_i = S^2R^3$. Hence there is some $H_1 \in \Psi$ such that

$$HG(A^2B^2) = H_1\, RG_1'G_2' \ldots G_n'\, R(A^2B^2).$$

Since $G_n' = SR$ we can apply the argument just completed to find the required $H'$ and $G'$. The proof is complete. □

Theorem 5.1 allows us to list one representative from each of the classes of forms which map to proper closed 1-intersectors on **T**. (It will become apparent, when we describe the associated sequences of integers that there is exactly one representative for each class, see Remark 5.1.) However, the set $\mathcal{K}$ involved is not unique in this respect. By forming compositions of the type $HG$ where $H \in \Psi$ and $G \in \mathcal{K}$ we can produce other such sets. We have chosen the particular set $\mathcal{K}$ in the theorem for two reasons. Firstly, for each $G \in \mathcal{K}$, the transformations $G(A)$ and $G(B)$ can be expressed as finite continued fractions and secondly, there is a simple formula for $G(A)$ and $G(B)$ in terms of the solutions to Markoff's equation. The first property allows us to describe the doubly infinite sequences of positive integers arising from the proper closed 1-intersectors, and the second leads to a formula for the representatives of the associated classes of forms. Before we discuss such things, we shall examine the set $\mathcal{K}$ more closely.

Again, the set $\mathcal{K}$ is best appreciated when arranged as a tree. We simplify this process by first introducing the transformations $C$ and $D$ defined by

$$C(z) = B^{-1}A^{-1}B^{-1}(z) = \frac{5z+2}{2z+1} \tag{5.3}$$

and

$$D(z) = B^{-1}(z) = \frac{2z+1}{z+1}. \tag{5.4}$$

Note that $SR^3SR(A,B) = (C,D)$ and hence $C$ and $D$ generate $\Gamma'$.

As with the set $\Omega$ there are two useful arrangements of $\mathcal{K}$ as a tree. The first arrangement is shown in Figure 5.1. It is obtained by placing $SR^2S^2(A,B) = (CD^{-2},D)$ and $SR^3(A,B) = (CD^{-1},D)$ at the top. Next comes

$$SR^3SR(A,B) = (C,D).$$

The remaining elements of $\mathcal{K}$ are all of the form $SR^3SRG_1G_2\dots G_n$ where $n \geq 1$. We view each such element as being the result of composing

$$(SR^3SR)\,G_1\,(SR^3SR)^{-1}$$

with

$$G = (SR^3SR)\,G_2G_3\dots G_n.$$

When $G_1 = SR$ the tree branches left and when $G_1 = S^2R^3$ it branches right. As in Figure 4.1, the branching operations are a consequence of

$$(SR^3SR)\,SR(SR^3SR)^{-1}(C,D) = (CD,D) \tag{5.5}$$

and

$$(SR^3SR)\,S^2R^3\,(SR^3SR)^{-1}(C,D) = (C,CD) \tag{5.6}$$

(Here we have extended our notation for automorphisms so that they can act on any pair of elements of $\Gamma'$.) Since the tree begins branching from the node $SR^3SR(A,B) = (C,D)$ induction shows that at every subsequent node $G(A,B) = (G(A),G(B))$ both $G(A)$ and $G(B)$ can be written as words consisting solely of $C$'s and $D$'s. It follows that the tree branches to the left by substituting $CD$ for $C$ everywhere in both $G(A)$ and $G(B)$ and to the right by substituting $CD$ for $D$. We refer to this tree as $\mathcal{K}$ *arranged by substitution.*

The second tree, shown in Figure 5.2, is obtained by again placing the automorphisms $SR^2S^2$, $SR^3$ and $SR^3SR$ at the top. This time however we view the typical element $SR^3SRG_1G_2\dots G_n$ of $\mathcal{K}$ with $n \geq 1$ as being the result of composing

$$G = SR^3SRG_1G_2\dots G_{n-1}$$

with $G_n$. When $G_n = SR$ the tree branches to the left and when $G_n = S^2R^3$ the tree branches to the right. For this tree, the branching operations are a consequence of the formulae

$$G\,SR(A,B) = (G(A)G(B),\ G(B)) \tag{5.7}$$

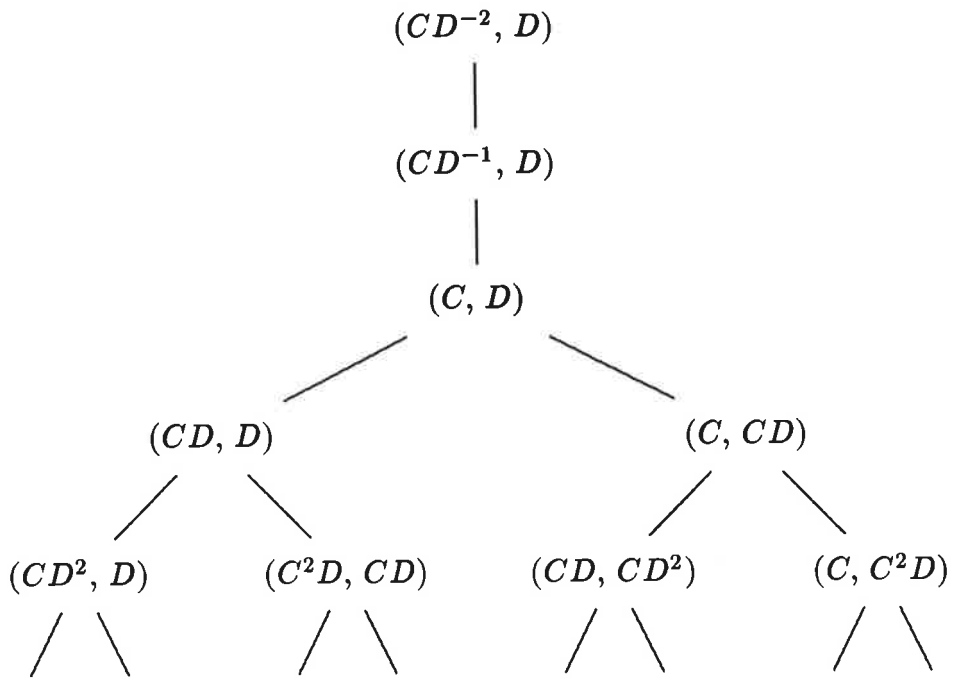


FIGURE 5.1. The set $\mathcal{S}$ arranged by substitution. Here $(C, D) = (B^{-1}A^{-1}B^{-1}, B^{-1})$. The tree continues by substituting $CD$ for $C$ to branch left and $CD$ for $D$ to branch right.

and

$$G\,S^2R^3(A,B) = (G(A),\ G(A)G(B)). \tag{5.8}$$

We refer to this tree as $\mathcal{K}$ *arranged by juxtaposition.*

The primary reason for our particular choice of $\mathcal{K}$ is that, the transformations $C$ and $D$ can be expressed as finite continued fractions, that is,

$$C(z) = 2 + \cfrac{1}{2 + \cfrac{1}{z}} = [2, 2, z] \tag{5.9}$$

and

$$D(z) = 1 + \cfrac{1}{1 + \cfrac{1}{z}} = [1, 1, z]. \tag{5.10}$$

Here $[a_0, a_1, \ldots, a_n, z]$ denotes the finite continued fraction with partial quotients $a_0, a_1, \ldots, a_n$ and $z$. The composition of such transformations is easy. If

$$W(z) = [a_0, a_1, \ldots, a_n, z] \qquad \text{and} \qquad V(z) = [b_0, b_1, \ldots, b_m, z]$$

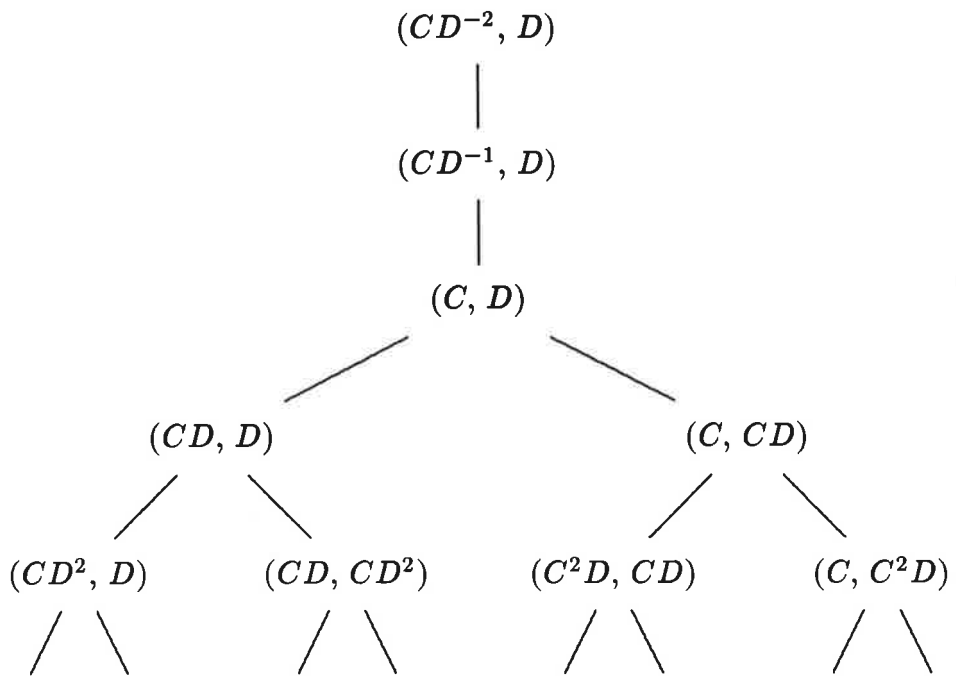


FIGURE 5.2. The set $\mathcal{S}$ arranged by juxtaposition. This tree continues from the node $(W_1, W_2)$ by forming $(W_1W_2, W_2)$ to branch left and $(W_1, W_1W_2)$ to branch right.

then

$$WV(z) = [a_0, \ldots, a_n, [b_0, \ldots, b_m, z]] = [a_0, \ldots, a_n, b_0, \ldots, b_m, z].$$

Now let $G(A, B) = (G(A), G(B))$ be an automorphism in $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$. We have seen from the arrangements of $\mathcal{K}$ by substitution and by juxtaposition that both $G(A)$ and $G(B)$ can be written as a words consisting solely of $C$'s and $D$'s. It follows that the transformation

$$W = G(A^2B^2) = G(A)G(A)G(B)G(B)$$

can likewise be written as a word consisting solely of $C$'s and $D$'s. By expressing $C$ and $D$ as continued fractions and using the rule above for the composition of such transformations, we find that $W(z)$ is of the form

$$W(z) = [a_0, a_1, \ldots, a_m, z]. \tag{5.11}$$

Of course, given that Theorem 5.1 is true, we are interested in the associated form $f_W$ and its Markoff value. With this in mind, we observe that it is possible to obtain

the sequence of positive integers corresponding to $f_W$ directly from the expression (5.11). Specifically, we claim that under the correspondence (1.8) established in Chapter 1 we have

$$\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_0, a_1, \ldots, a_m}\} \quad \longmapsto \quad \mu f_W(x, y) \tag{5.12}$$

where $\mu$ is some positive real number. As is standard, $\{\overline{a_0, a_1, \ldots, a_m}\}$ denotes the periodic doubly infinite sequence with period $a_0, a_1, \ldots, a_m$. To see that the statement is true, it suffices to show that $f_W$ has first root $\eta = -[0, a_{-1}, a_{-2}, a_{-3}, \ldots]$ and second root $\xi = [a_0, a_1, a_2, \ldots]$. By the elementary properties of continued fractions, $\xi$ is a fixed point of $W$. Similarly, $\eta$ is a fixed point of its inverse

$$W^{-1}(z) = -[0, a_m, \ldots, a_1, a_0, -1/z]$$

and hence also of $W$. It follows that $\eta$ and $\xi$ are the roots of $f_W$. To see that they are correctly ordered, note that $W$ has a matrix with all positive entries. Hence the leading coefficient $\mu$ of $f_W$ is positive and the result follows since $\mu(\xi - \eta) > 0$.

We have now established a direct relationship between each element $G$ of $\mathcal{K}$ and the sequence of positive integers which corresponds to the form $f_W$ where $W = G(A^2B^2)$. This relationship allows to describe an algorithm which lists the periods of all such sequences of positive integers. It is based on the arrangement of $\mathcal{K}$ by substitution. In order to describe it we let $G(A, B) = (G(A), G(B))$ be an element of $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$. We have seen above that both $G(A)$ and $G(B)$ and hence the transformation $W = G(A^2B^2)$ can be written as a words consisting solely of $C$'s and $D$'s. Consequently $W$ is of the form (5.11) and

$$a_0, a_1, \ldots, a_m$$

is the period of the corresponding sequence of positive integers. The arrangement of $\mathcal{K}$ by substitution continues from $G$ by substituting $CD$ for $C$ everywhere in both $G(A)$ and $G(B)$ to branch left and $CD$ for $D$ to branch right. We denote the resulting automorphisms by $G'$ and $G''$, respectively. Clearly the transformation $W' = G'(A^2B^2)$ can be obtained from $W$ by substituting $CD$ for $C$ everywhere in

$W$. It follows that $W'(z) = [a'_0, a'_1, \ldots, a'_k, z]$ where $a'_0, a'_1, \ldots, a'_k$ is obtained from $a_0, a_1, \ldots, a_m$ by substituting $2,2,1,1$ for $2,2$. We define the substitution $\phi$ by

$$\phi(1,1) = 1,1 \qquad \text{and} \qquad \phi(2,2) = 2,2,1,1 \tag{5.13}$$

and we write

$$a'_0, a'_1, \ldots, a'_k = \phi(a_0, a_1, \ldots, a_m).$$

Note that (5.12) implies $a'_0, a'_1, \ldots, a'_k$ is the period of the integer sequence associated with the form $f_{W'}$. We deal with the transformation $W'' = G''(A^2B^2)$ in a similar manner. It can be obtained from $W$ by substituting $CD$ for $D$ everywhere. Hence $W''(z) = [a''_0, a''_1, \ldots, a''_l, z]$ where $a''_0, a''_1, \ldots, a''_l$ is the result of substituting $2,2,1,1$ for $1,1$ in $a_0, a_1, \ldots, a_m$. We define the substitution $\psi$ by

$$\psi(1,1) = 2,2,1,1 \qquad \text{and} \qquad \psi(2,2) = 2,2 \tag{5.14}$$

and write

$$a''_0, a''_1, \ldots, a''_l = \psi(a_0, a_1, \ldots, a_m).$$

Clearly $a''_0, a''_1, \ldots, a''_l$ is a period of the integer sequence associated with the form $f_{W''}$. We now repeat the process with $G'$ and $G''$ in place of $G$ and so on. Since the arrangement of $\mathcal{K}$ by substitution begins branching from the automorphism $SR^3SR$ and since

$$SR^3SR(A^2B^2)(z) = [2,2,2,2,1,1,1,1,z],$$

our algorithm begins with the period $2,2,2,2,1,1,1,1$.

To summarise, we have shown that by starting with $2,2,2,2,1,1,1,1$ and repeatedly applying the substitutions $\phi$ and $\psi$ we obtain all the periods of the integer sequences which correspond to the forms $f_W$ where $W = G(A^2B^2)$ and $G$ is an element of $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$. It follows from Theorem 5.1 that our algorithm accounts for all the sequences of positive integers which map to proper closed 1-intersectors on $\mathbf{T}$ except those which correspond to the forms $f_W$ where $W = G(A^2B^2)$ and $G$ is $SR^2S^2$ or $SR^3$. For these automorphisms we have

$$SR^2S^2(A^2B^2)(z) = CD^{-2}C(z) = [3,3,z]$$

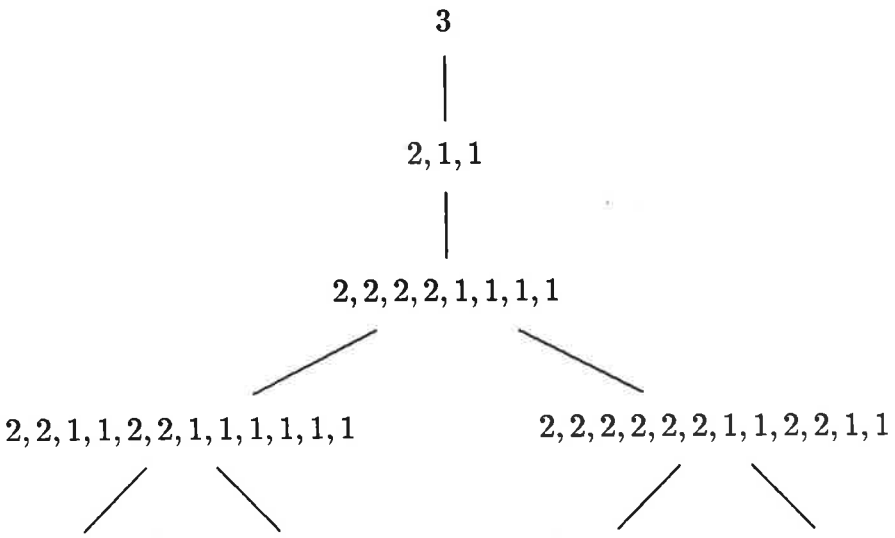


FIGURE 5.3. Periods of the sequences of positive integers which map to proper closed 1-intesectors on $\mathbf{T}$. The tree continues by substituting $2, 2, 1, 1$ for $2, 2$ to branch left and $2, 2, 1, 1$ for $1, 1$ to branch right.

and

$$SR^3(A^2B^2)(z) = CD^{-1}CD(z) = [2, 1, 1, 2, 1, 1, z].$$

Periods of the associated sequences of integers are 3 and $2, 1, 1$, respectively. Since we obtained our algorithm from the the arrangement of $\mathcal{K}$ by substitution it is natural to present it in the form of a tree. The tree is shown in Figure 5.3.

**Theorem 5.2.** *A doubly infinite sequence of positive integers maps to a proper closed 1-intersector on* $\mathbf{T}$ *if and only if it is periodic and has a period appearing somewhere in the tree shown in Figure 5.3.*

Just as the tree of periods in Figure 5.3 was derived from the arrangement of $\mathcal{K}$ by substitution, we can derive a re-arrangement of it from the arrangement of $\mathcal{K}$ by juxtaposition. However, there is no simple formulae for the branching operations in this second tree. The problem is that the way a new period in it is formed from an existing period depends on the way the existing was formed from the period preceding it. While it is possible to overcome this problem by inserting markers in the periods, this is not necessary for our purposes and we omit the details.

In the next theorem, Theorem 5.3, we shall describe some of the properties

of the periods displayed in Figure 5.3 and we shall show how to calculate the Markoff values of the associated periodic doubly infinite sequences. We can reduce the amount of work involved by taking advantage of a symmetry of tree. The symmetry we shall use is best seen by introducing another operation on sequences of $1,1$'s and $2,2$'s. If $a_0, a_1, \ldots, a_m$ is a sequence composed solely of $1,1$'s and $2,2$'s then we define $\tau$ by

$$\tau(a_0, a_1, \ldots, a_m) = a_1', a_2', \ldots, a_m' \tag{5.15}$$

where $a_1', a_2', \ldots, a_m'$ is obtained from $a_0, a_1, \ldots, a_m$ by reversing it and interchanging each $2,2$ with $1,1$ and *vica-versa*. It is straightforward to show that

$$\tau^2 = \mathrm{Id} \qquad \text{and} \qquad \tau \circ \phi \circ \tau = \psi. \tag{5.16}$$

These equations together with the remark that the first non-singular sequence in the tree in Figure 5.3, namely $2,2,2,2,1,1,1,1$, is invariant under $\tau$ show that the non-singular part of the tree is invariant under $\tau$. In other words, if the sequence $a_0, a_1, \ldots, a_m$ lies in the tree then so does $\tau(a_0, a_1, \ldots, a_m)$. The symmetry we are referring to is the fact that the location of $\tau(a_0, a_1, \ldots, a_m)$ in the tree is a reflection of that of $a_0, a_1, \ldots, a_m$ in the vertical line which passes through the base node. Note in particular that $\tau(a_0, a_1, \ldots, a_m)$ and $a_0, a_1, \ldots, a_m$ are at the same level in the tree.

Before we begin the proof of Theorem 5.3 it is also convenient to extend the definition of the operations $\phi$, $\psi$ and $\tau$ so that they act on doubly infinite sequences of $1,1$'s and $2,2$'s. (Recall that $\phi$ and $\psi$ are defined by (5.13) and (5.14).) For the operations $\phi$ and $\psi$ it is clear how this is done. One applies $\phi$ by replacing each occurrence of $2,2$ with $2,2,1,1$ and one applies $\psi$ by replacing each $1,1$ with $2,2,1,1$. Further, it is not hard to verify that if $a_0, a_1, \ldots, a_m$ is a sequence of $1,1$'s and $2,2$'s then

$$\phi\big(\{\overline{a_0, a_1, \ldots, a_m}\}\big) = \{\overline{\phi(a_0, a_1, \ldots, a_m)}\} \tag{5.17}$$

and

$$\psi\big(\{\overline{a_0, a_1, \ldots, a_m}\}\big) = \{\overline{\psi(a_0, a_1, \ldots, a_m)}\}. \tag{5.18}$$

The operation $\tau$ is also easily extended. Given a doubly infinite sequence $\mathcal{A}$ of $1,1$'s and $2,2$'s, we define $\tau(\mathcal{A})$ to be the sequence obtained by reversing $\mathcal{A}$ and replacing each occurrence of $1,1$ with $2,2$ and *vice-versa*. As above, we have

$$\tau\big(\{\overline{a_0, a_1, \ldots, a_m}\}\big) = \{\overline{\tau(a_0, a_1, \ldots, a_m)}\}. \tag{5.19}$$

Note also that the relationships (5.16) still hold for the extended operations.

**Remark 5.1.** We can now show that there is no duplication amongst the equivalence classes represented by the forms described in Theorem 5.1. This is possible because, as can be deduced from the construction of the tree in Figure 5.3, it is equivalent to show that there is no duplication amongst the periodic sequences of integers which have a period appearing in the tree. To prove the latter is true we shall use the fact that every such sequence, other than $\{\overline{3}\}$ and $\{\overline{2,1,1}\}$, can be obtained from $\{\overline{2,2,2,2,1,1,1,1}\}$ by applying the substitutions $\phi$ and $\psi$. First note that if $\mathcal{A}$ is a sequence consisting solely of $1,1$'s and $2,2$'s then $\phi(\mathcal{A})$ can be partitioned into the sequences $1,1$ and $2,2,1,1$ and hence any occurrence of $2,2$ in $\phi(\mathcal{A})$ is isolated. Similarly, if $\mathcal{A}'$ is another sequence of $1,1$'s and $2,2$'s then any occurrence of $1,1$ in $\psi(\mathcal{A}')$ is isolated. Therefore,

$$\phi(\mathcal{A}) = \psi(\mathcal{A}') \qquad \Longleftrightarrow \qquad \mathcal{A} = \{\overline{2,2}\} \quad \text{and} \quad \mathcal{A}' = \{\overline{1,1}\}.$$

It is also apparent that

$$\phi(\mathcal{A}) = \phi(\mathcal{A}') \qquad \Longleftrightarrow \qquad \mathcal{A} = \mathcal{A}' \qquad \Longleftrightarrow \qquad \psi(\mathcal{A}) = \psi(\mathcal{A}').$$

Using these identities it is not hard to deduced that there is at most one way for any given sequence to be obtained from the $\{\overline{2,2,2,2,1,1,1,1}\}$ by applying the substitutions $\phi$ and $\psi$. It follows, as claimed, that the sequences which have a period appearing in the tree in Figure 5.3 are all distinct.

**Theorem 5.3.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_0, a_1, \ldots, a_m}\}$ be a periodic sequence of positive integers, other than $\{\overline{3}\}$ or $\{\overline{2,1,1}\}$, whose period $a_0, a_1, \ldots, a_m$ occurs somewhere in the tree shown in Figure 5.3. Then, $m+1 = 4n$ for some $n \geq 2$ and $a_0, a_1, \ldots, a_m$ is of the form*

$$2, 2, a_2, a_3, \ldots, a_{2n-3}, 2, 2, 1, 1, a_{2n-3}, \ldots, a_3, a_2, 1, 1 \tag{5.20}$$

*where $a_2, a_3, \ldots, a_{2n-3}$ is symmetric. Further $M(\mathcal{A}) = \lambda_0(\mathcal{A})$.*

*Proof.* We prove the first statement of the theorem by induction on the level of the sequence $a_0, a_1, \ldots, a_m$ in the tree. By the level of $a_0, a_1, \ldots, a_m$ we mean the number of nodes between it and the sequence $2,2,2,2,1,1,1,1$. The statement is true with $n = 2$ if $a_0, a_1, \ldots, a_m$ is $2,2,2,2,1,1,1,1$ itself. Hence, we assume $a_0, a_1, \ldots, a_m$ is at some arbitrary position in the tree and that it is of the form (5.20) where $m+1 = 4n \geq 8$ and $a_2, a_3, \ldots, a_{2n-3}$ is symmetric and we prove that the two sequences immediately below, namely

$$\phi(a_0, a_1, \ldots, a_m) \qquad \text{and} \qquad \psi(a_0, a_1, \ldots, a_m),$$

are of the same form. Since $\tau$ preserves the symmetry of (5.20) and $\psi = \tau \circ \phi \circ \tau$ it suffices to do this only for the sequence $\phi(a_0, a_1, \ldots, a_m)$. We write

$$a'_0, a'_1, \ldots, a'_k = \phi(a_0, a_1, \ldots, a_m).$$

The symmetry of $a_2, a_3, \ldots, a_{2n-3}$ implies that $a'_0, a'_1, \ldots, a'_k$ is of the form

$$2,2,1,1,a'_4,a'_5,\ldots,a'_{2r-3},2,2,1,1,1,1,a'_4,a'_5,\ldots,a'_{2r-3},1,1$$

where $4r = k+1$ and

$$a'_4, a'_5, \ldots, a'_{2r-3} = \phi(a_2, a_3, \ldots, a_{2n-3}).$$

To see that $a'_0, a'_1, \ldots, a'_k$ has the required property we need only verify that

$$1,1,a'_4,a'_5,\ldots,a'_{2r-3} = a'_{2r-3},\ldots,a'_5,a'_4,1,1. \tag{5.21}$$

For this purpose, we let $\widetilde{\phi}$ denote the substitution

$$\widetilde{\phi}(1,1) = 1,1 \qquad \text{and} \qquad \widetilde{\phi}(2,2) = 1,1,2,2. \tag{5.22}$$

Since $\widetilde{\phi}(1,1)$ and $\widetilde{\phi}(2,2)$ are the reverse of $\phi(1,1)$ and $\phi(2,2)$, respectively, it is clear that $\widetilde{\phi}(a_{2n-3}, \ldots, a_3, a_2)$ is the reverse of $\phi(a_2, a_3, \ldots, a_{2n-3})$. Thus,

$$\widetilde{\phi}(a_{2n-3}, \ldots, a_3, a_2) = a'_{2r-3}, \ldots, a'_5, a'_4.$$

Using the symmetry of $a_2, a_3, \ldots, a_{2n-3}$ we can now re-write (5.21) as

$$1, 1, \phi(a_2, a_3, \ldots, a_{2n-3}) = \widetilde{\phi}(a_2, a_3, \ldots, a_{2n-3}), 1, 1.$$

This equality is easily verified by noting that $\phi$ and $\widetilde{\phi}$ both increase by one the number of $1, 1$'s between each pair of $2, 2$'s in $a_2, a_3, \ldots, a_{2n-3}$ and that $\phi$ adds an extra $1, 1$ on the right where as $\widetilde{\phi}$ adds an extra $1, 1$ on the left.

The other claim of the theorem is that $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. We shall prove a somewhat stronger result. We shall prove that for all integers $i$

$$\xi_0(\mathcal{A}) \geq \xi_{2i}(\mathcal{A}) \qquad \text{and} \qquad \eta_0(\mathcal{A}) \leq \eta_{2i}(\mathcal{A}). \tag{5.23}$$

Recall that the functions $\xi_{2i}$ and $\eta_{2i}$ are defined by (1.6) in Chapter 1. To see that these inequalities imply $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ note that they certainly imply

$$\lambda_0(\mathcal{A}) = \xi_0(\mathcal{A}) - \eta_0(\mathcal{A}) \geq \xi_{2i} - \eta_{2i} = \lambda_{2i}(\mathcal{A})$$

for all $i$. Now observe that the the period (5.20) of $\mathcal{A}$ has a centre of symmetry between the terms $a_{n-1}$ and $a_n$. Thus $a_{n+j} = a_{n-1-j}$ for all integers $j$. Further,

$$\xi_{n+j}(\mathcal{A}) = a_{n-1-j} - \eta_{n-1-j}(\mathcal{A})$$

and

$$\eta_{n+j}(\mathcal{A}) = a_{n-1-j} - \xi_{n-1-j}(\mathcal{A})$$

and so $\lambda_{n+j}(\mathcal{A}) = \lambda_{n-1-j}(\mathcal{A})$ for all integers $j$. Putting $j = 2i + 1 - n$ we have

$$\lambda_{2i+1}(\mathcal{A}) = \lambda_{2n-2i-2}(\mathcal{A}).$$

Since $2n - 2i - 2$ is even we know $\lambda_0(\mathcal{A}) \geq \lambda_{2n-2i-2}(\mathcal{A})$ and therefore $\lambda_0(\mathcal{A}) \geq \lambda_{2i+1}(\mathcal{A})$ for all integers $i$. It follows that $M(\mathcal{A}) = \lambda_0(\mathcal{A})$.

We prove the inequalities (5.23) by induction on the level of $a_0, a_1, \ldots, a_m$ in the tree. All we require to make the relevant comparisons of the continued fractions is the fact that

$$[b_1, b_1, b_2, b_2, \ldots, b_j, b_j, 2, 2, \ldots] > [b_1, b_1, b_2, b_2, \ldots, b_j, b_j, 1, 1 \ldots]$$

and

$$[0, b_1, b_1, b_2, b_2, \ldots, b_j, b_j, 2, 2, \ldots] < [0, b_1, b_1, b_2, b_2, \ldots, b_j, b_j, 1, 1 \ldots]$$

for any $j \geq 0$ and positive integers $b_1, b_2, \ldots, b_j$. With this in mind it is easy to verify that (5.23) is true for all $i$ when $a_0, a_1, \ldots, a_m$ is the sequence $2, 2, 2, 2, 1, 1, 1, 1$. Having established the basis for induction we now assume the period $a_0, a_1, \ldots, a_m$ is at an arbitrary position in the tree and that (5.23) holds. We shall prove that for all $i$ we have

$$\xi_0(\mathcal{A}') \geq \xi_{2i}(\mathcal{A}') \qquad \text{and} \qquad \eta_0(\mathcal{A}') \leq \eta_{2i}(\mathcal{A}') \tag{5.24}$$

where $\mathcal{A}' = \{a_i'\}_{i=-\infty}^{+\infty} = \{\overline{a_0', a_1', \ldots, a_k'}\}$ and $a_0', a_1', \ldots, a_k' = \phi(a_0, a_1, \ldots, a_m)$ and

$$\xi_0(\mathcal{A}'') \geq \xi_{2i}(\mathcal{A}'') \qquad \text{and} \qquad \eta_0(\mathcal{A}'') \leq \eta_{2i}(\mathcal{A}'') \tag{5.25}$$

where $\mathcal{A}'' = \{a_i''\}_{i=-\infty}^{+\infty} = \{\overline{a_0'', a_1'', \ldots, a_l''}\}$ and $a_0'', a_1'', \ldots, a_l'' = \psi(a_0, a_1, \ldots, a_m)$.

We deal the inequality $\xi_0(\mathcal{A}') \geq \xi_{2i}(\mathcal{A}')$ first. To this end, we observe from (5.17) that $\mathcal{A}' = \phi(\mathcal{A})$. Note in particular that

$$a_0', a_1', a_3', \ldots\ldots = \phi(a_0, a_1, a_3, \ldots\ldots).$$

Now suppose that $a_{2i}', a_{2i+1}' = 2, 2$. Since $\mathcal{A}' = \phi(\mathcal{A})$ we know that the $2, 2$'s in $\mathcal{A}'$ are isolated. Thus $a_{2i+2}', a_{2i+3}' = 1, 1$ and the sequence $a_{2i}', a_{2i+1}', a_{2i+2}', a_{2i+3}'$ is the image under $\phi$ of some pair $a_{2j}, a_{2j+1} = 2, 2$ in $\mathcal{A}$. It follows that

$$a_{2i}', a_{2i+1}', a_{2i}', \ldots\ldots = \phi(a_{2j}, a_{2j+1}, a_{2j+3}, \ldots\ldots).$$

Our inductive hypothesis is that

$$\xi_0(\mathcal{A}) = [a_0, a_1, a_2, \ldots\ldots] \geq [a_{2j}, a_{2j+1}, a_{2j+2}, \ldots\ldots] = \xi_{2j}(\mathcal{A}). \tag{5.26}$$

Since the image of $2, 2$ under $\phi$ begins with $2, 2$ and the image of $1, 1$ is $1, 1$ it is clear that the inequality (5.26) is preserved by the substitution $\phi$. Thus

$$\xi_0(\mathcal{A}') = [a_0', a_1', a_2', \ldots\ldots] \geq [a_{2i}', a_{2i+1}', a_{2i+2}', \ldots\ldots] = \xi_{2i}(\mathcal{A}')$$

and we are done. The only other possibility is that $a'_{2i}, a'_{2i+1} = 1,1$. In this case the inequality is trivial since $a'_0, a'_1 = 2,2$.

For the inequality $\eta_0(\mathcal{A}') \leq \eta_{2i}(\mathcal{A}')$ we make use of the substitution $\widetilde{\phi}$ defined by (5.22). We have

$$\ldots\ldots a'_{-3}, a'_{-2}, a'_{-1} = \phi(\ldots\ldots a_{-3}, a_{-2}, a_{-1})$$

and therefore

$$a'_{-1}, a'_{-2}, a'_{-3}, \ldots\ldots = \widetilde{\phi}(a_{-1}, a_{-2}, a_{-3}, \ldots\ldots).$$

Now suppose that $a'_{2i-2}, a'_{2i-1} = 1,1$. Since $\mathcal{A}' = \phi(\mathcal{A})$ we know that either $a'_{2i-2}, a'_{2i-1}$ is the image under $\phi$ of a pair $a_{2j-2}, a_{2j-1} = 1,1$ or $a'_{2i-4}, a'_{2i-3} = 2,2$ and the sequence $a'_{2i-4}, a'_{2i-3}, a'_{2i-2}, a'_{2i-1}$ is the image under $\phi$ of a pair $a_{2j-2}, a_{2j-1} = 2,2$. In either case, we have

$$\ldots\ldots a'_{2i-3}, a'_{2i-2}, a'_{2i-1} = \phi(\ldots\ldots a_{2j-3}, a_{2j-2}, a_{2j-1})$$

and hence

$$a'_{2i-1}, a'_{2i-2}, a'_{2i-3}, \ldots\ldots = \widetilde{\phi}(a_{2j-1}, a_{2j-2}, a_{2j-3}, \ldots\ldots)$$

for some integer $j$. Our inductive hypothesis is that

(5.27)

$$\eta_0(\mathcal{A}) = -[0, a_{-1}, a_{-2}, a_{-3}, \ldots\ldots] \leq -[0, a_{2j-1}, a_{2j-2}, a_{2j-3}, \ldots\ldots] = \xi_{2j}(\mathcal{A}).$$

Since the images of both the sequences $1,1,1,1$ and $1,1,2,2$ under $\widetilde{\phi}$ begin with $1,1,1,1$ and the image of $2,2$ is $1,1,2,2$ it is clear that the inequality (5.27) is preserved by the substitution $\widetilde{\phi}$. Thus

$$\eta_0(\mathcal{A}') = -[0, a'_{-1}, a'_{-2}, a'_{-3}, \ldots\ldots] \leq -[0, a'_{2i-1}, a'_{2i-2}, a'_{2i-3}, \ldots\ldots] = \eta_{2i}(\mathcal{A}')$$

and we have the desired result. The only other possibility is that $a'_{2i-2}, a'_{2i-1} = 2,2$ in which case the inequality is trivial since $a'_{-2}, a'_{-1} = 1,1$.

We complete the proof by using the symmetry induced by $\tau$ and the truth of (5.24) to show that the inequalities (5.25) hold. Thus we let $\tilde{\mathcal{A}} = \{\tilde{a}_i\}_{i=-\infty}^{+\infty} = \{\overline{\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_l}\}$ where $\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_l = \tau(a_0'', a_1'', \ldots, a_l'')$ and we note from (5.19) that $\tilde{\mathcal{A}} = \tau(\mathcal{A}'')$. It is not hard to deduce from the definition of $\tau$ that

$$\xi_0(\mathcal{A}'') \geq \xi_{2i}(\mathcal{A}'') \qquad \Longleftrightarrow \qquad \eta_0(\tilde{\mathcal{A}}) \leq \eta_{-2i}(\tilde{\mathcal{A}}).$$

To see that the right hand side is true we observe that

$$\tilde{\mathcal{A}} = \tau\mathcal{A}'' = \tau \circ \psi(\mathcal{A}) = \phi \circ \tau(\mathcal{A}) = \phi(\tau\mathcal{A}).$$

Since $\tau\mathcal{A}$ is at the same level in the tree as $\mathcal{A}$ our inductive hypothesis applies and the argument above shows $\eta_0(\phi(\tau\mathcal{A})) \leq \eta_{-2i}(\phi(\tau\mathcal{A}))$. We conclude that $\xi_0(\mathcal{A}'') \geq \xi_{2i}(\mathcal{A}'')$. The inequality $\eta_0(\mathcal{A}'') \leq \eta_{2i}(\mathcal{A}'')$ can be dealt with in a similar manner. □

While Theorem 5.3 allows us to calculate the Markoff values arising from the proper closed 1-intersectors, there is a better way. Just as the Markoff forms and their Markoff values can be expressed in terms the solutions to Markoff's equation, so too can the forms and values arising from the proper closed 1-intersectors. The connection here is that the automorphisms of $\mathcal{K}$ themselves can be so expressed. We describe how to do this in the next theorem. Lekkerkerker, [26], has already established such a result for a different but closely related set of automorphisms. We shall elaborate on Lekkerkerker's work at the end of this chapter.

**Theorem 5.4.** *There is a bijection between the elements $G$ of $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$ and the solutions $(m, m_1, m_2)$ in positive integers to the Markoff equation*

$$m^2 + m_1^2 + m_2^2 = 3mm_1m_2, \qquad m \geq \max\{m_1, m_2\} \tag{5.28}$$

*other than $(1,1,1)$ and $(2,1,1)$, for which*

$$G(A,B) = \left( \begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix}, \begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix} \right) \tag{5.29}$$

*where the integer $k$ and the ordering of $m_1$ and $m_2$ are uniquely determined by*

$$m_1 k \equiv m_2 \pmod{m}, \qquad m/2 < k < m \tag{5.30}$$

*and $k_1$ and $k_2$ are the integers which satisfy*

$$m_1 k - m k_1 = m_2 \qquad \text{and} \qquad m k_2 - m_2 k = m_1 \tag{5.31}$$

*and $p_1$ and $p_2$ are chosen so that the matrices have determinant 1.*

*Proof.* Throughout this proof, when we say $(m, m_1, m_2)$ is a *solution*, we shall mean that $m$, $m_1$ and $m_2$ are positive integers satisfying (5.28). The set of all such solutions is described by Cassels in §3 of Chapter II of his book, [5]. We have summarised his work in the section of Chapter 1 on Markoff forms.

It is appropriate to begin the proof by outlining why all the integers mentioned in the theorem are well-defined. Recall that the solutions $(1,1,1)$ and $(2,1,1)$ are called singular and that for each non-singular solution $(m, m_1, m_2)$ there is exactly one integer $k$ with $0 < k < m/2$ and one ordering of $m_1$ and $m_2$ such that $m_1 k \equiv m_2 \pmod{m}$. By replacing $k$ by $m - k$ and interchanging $m_1$ and $m_2$ it can be seen that there is exactly one integer $k$ and one ordering of $m_1$ and $m_2$ such that (5.30) holds. It is also readily seen from Cassels' work that there are integers $k_1$ and $k_2$ which satisfy (5.31) and that $p_1$ and $p_2$ can be chosen as described. We refer, as Cassels does, to the collection of integers

$$(m, k;\ m_1, k_1;\ m_2, k_2) \tag{5.32}$$

as an *ordered Markoff set.* We stress that our definition includes the restriction $m/2 < k < m$ rather than $0 < k < m/2$, and therefore, our ordering of $m_1$ and $m_2$ is the reverse of Cassels'. We shall also require the identity

$$m_1 k_2 - m_2 k_1 = 3 m_1 m_2 - m. \tag{5.33}$$

It may be found in Cassels' exposition and is an consequence of (5.31) and (5.28).

We shall actually prove that there is a bijection between the non-singular ordered Markoff sets (5.32) and the elements $G$ of $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$ which has

the property (5.29). In order to define the bijection we first need to arrange the order Markoff sets into a tree. Cassels' has already done this for the solutions of (5.28). His tree is shown in Figure 1.1. Obviously we can obtain a similar arrangement of the ordered Markoff sets by replacing each solution with the corresponding set. More importantly however, we can extend Cassels' branching operations (1.4) and (1.5) so that they produce the associated tree of ordered Markoff sets. Before we describe the extended branching operations we must modify Cassels' tree.

As already indicated, we need to interchange $m_1$ and $m_2$. It is also necessary to interchange the branching operations. The top of Cassels' tree is unaltered by these changes, it still contains the singular solutions $(1,1,1)$ and $(2,1,1)$. However, immediately below is the non-singular solution $(5,2,1)$. The modified tree then continues from each such non-singular solution $(m, m_1, m_2)$ by forming the solution $(m_1', m, m_2)$ where $m_1' = 3mm_2 - m_1$ to branch left and the solution $(m_2', m_1, m)$ where $m_2' = 3mm_1 - m_2$ to branch right. The associated tree of ordered Markoff sets is obtained by replacing each non-singular solution with the corresponding set. It is shown in Figure 5.4. It continues from the set $(m, k;\ m_1, k_1;\ m_2, k_2)$ by forming the set

$$(5.34) \qquad (m_1', k_1';\ m, k;\ m_2, k_2), \qquad m_1' = 3mm_2 - m_1, \qquad k_1' = 3km_2 - k_1$$

to branch left and the set

$$(5.35) \qquad (m_2', k_2';\ m_1, k_1;\ m, k), \qquad m_2' = 3mm_1 - m_2, \qquad k_2' = 3km_1 - k_2$$

to branch right.

Of course we need to verify that the sets (5.34) and (5.35) are indeed ordered Markoff sets. For the first set, using the definitions of $k_1$, $k_1'$ and $m_1'$ we have

$$mk_1' - m_1'k = m_1k - mk_1 = m_2$$

and

$$m_1'k_2 - m_2k_1' = 3m_2(mk_2 - m_2k) + m_2k_1 - m_1k_2 = m.$$

Thus

$$mk_1' \equiv m_2 \pmod{m_1'}.$$

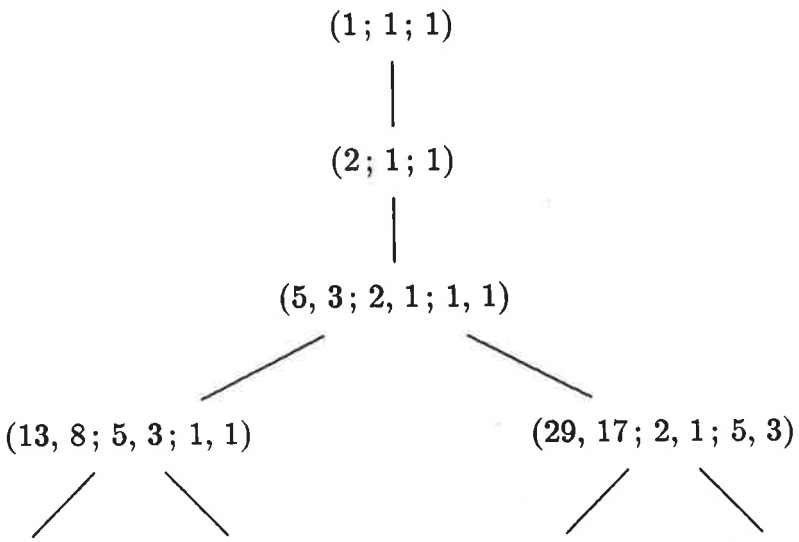


FIGURE 5.4. The tree of ordered Markoff sets described in the proof of Theorem 5.4. The tree continues from $(m, k;\ m_1, k_1;\ m_2, k_2)$ by forming (5.34) to branch left and (5.35) to branch right.

We also have

$$k_1' = 3km_2 - (m_1k - m_2)/m = m_1'k/m + m_2/m$$

and since $m/2 < k < m$ and $m_2/m < 1$ it follows that $m_1'/2 < k_1' \leq m_1'$. Note that, $k_1' \neq m_1'$ because $m_2$ and $m_1'$ are coprime. The proof for the other set is similar.

We can now use the congruence of shape between the tree of ordered Markoff sets shown in Figure 5.4 and the arrangement of $\mathcal{K}$ by juxtaposition shown in Figure 5.2 to define the bijection we are interested in. The bijection simply identifies elements which occupy the same relative positions in the trees. To see that the bijection is well-defined, we note that the automorphisms in the tree in Figure 5.2 are all distinct since Remark 4.1 shows that the same is true of the automorphisms the tree in Figure 4.2. It is evident from Cassels' work that the solutions to (5.28) in the tree in Figure 1.1 are likewise all distinct and therefore there is no duplication of ordered Markoff sets in the tree in Figure 5.4.

We can complete the proof by showing that if under the bijection described a non-singular ordered Markoff set $(m, k;\ m_1, k_1;\ m_2, k_2)$ corresponds to an automor-

phism $G$ and if $p_1$ and $p_2$ are chosen so that the matrices in (5.29) have determinant 1 then (5.29) is true. The proof is by induction on the level of the set in the tree. When

$$(m, k;\, m_1, k_1;\, m_2, k_2) = (5,\, 3;\, 2,\, 1;\, 1,\, 1)$$

the corresponding automorphism is

$$SR^3SR(A, B) = \left(\begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

and (5.29) is true and we have a basis for induction. Now let $(m, k;\, m_1, k_1;\, m_2, k_2)$ be an arbitrary non-singular ordered Markoff set and $G$ the corresponding automorphism of $\mathcal{K}$ and suppose that the matrices in (5.29) have determinant 1 and (5.29) is true. The tree of ordered Markoff sets continues from $(m, k;\, m_1, k_1;\, m_2, k_2)$ by forming the set (5.34) to branch left and (5.35) to branch right. Likewise, the arrangement of $\mathcal{K}$ by juxtaposition continues from $G$ by forming $GSR$ to branch left and $GS^2R^3$ to branch right. Therefore, we must prove that

$$GSR(A, B) = \left(\begin{pmatrix} 3m - k & p \\ m & k \end{pmatrix}, \begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix}\right)$$

and

$$GS^2R^3(A, B) = \left(\begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix}, \begin{pmatrix} 3m - k & p \\ m & k \end{pmatrix}\right)$$

where $p$ is chosen so that the associated matrix has determinant is 1. Our inductive hypothesis is that (5.29) holds and since

$$GSR(A, B) = (G(A)G(B), G(B))$$

and

$$GS^2R^3(A, B) = (G(A), G(A)G(B))$$

it suffices to show that

$$\begin{pmatrix} 3m - k & p \\ m & k \end{pmatrix} = \begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix} \begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix}. \tag{5.36}$$

To this end note that (5.33) implies $3m_2 - k_2 = (m - m_2k_1)/m_1$ and define

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix} \begin{pmatrix} (m - m_2k_1)/m_1 & p_2 \\ m_2 & k_2 \end{pmatrix}.$$

Clearly

$$c = m - m_2k_1 + m_2k_1 = m$$

and since $m_2p_2 = (m - m_2k_1)k_2/m_1 - 1$ we have

$$m_2d = m_1m_2p_2 + m_2k_1k_2 = (m - m_2k_1)k_2 - m_1 + m_2k_1k_2 = m_2k$$

and since $m_1p_1 = (3m_1 - k_1)k_1 - 1$ we also have

$$\begin{aligned} m_1a &= (3m_1 - k_1)(m - m_2k_1) + ((3m_1 - k_1)k_1 - 1)m_2 \\ &= 3mm_1 - k_1m - m_2 \\ &= m_1(3m - k). \end{aligned}$$

Thus $c = m$, $d = k$ and $a = 3m - k$. That $b = p$ follows from $ad - bc = 1$ and the proof is complete. □

**Remark 5.2.** If we dispense with the restriction $m/2 < k < m$ then it is possible to define ordered Markoff sets for the two non-singular solutions to (5.28). The set corresponding to the solution $(1,1,1)$ is $(1,0;1,1;1,1)$ and the set for $(2,1,1)$ is $(2,1;1,0;1,1)$. These sets can be obtained by applying the reverse of the branching operation (5.34) to the set $(5,3;2,1;1,1)$. (The sets produced by reverse of (5.35) are different.) When the bijection described in Theorem 5.4 is extended in the obvious manner to include these new sets, the identity (5.29) remains true.

**Theorem 5.5.** *Let $G \in \mathcal{K}$ and $W = G(A^2B^2)$. If $(m, m_1, m_2)$ is the triple of positive integers corresponding to $G$ under the bijection described in Theorem 5.4 and if $k$, $k_1$ and $k_2$ are as defined there then*

$$f_W(x,y) = 3m^2x^2 + (2r - 9m^2)xy - sy^2 \tag{5.37}$$

*where $r = 9m_1m_2k - 3m_1k_1 - 3m_2k_2$ and $s$ satisfies $3m^2s = 9m^2(r+1) - r^2$. Further,*

$$M(f_W) = \sqrt{9 + \frac{4}{m^2}}. \tag{5.38}$$

*Proof.* We know from (5.11) that we can write

$$W(z) = \frac{az+b}{cz+d} = [a_0, a_1, \ldots, a_n, z].$$

Now set $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_0, a_1, \ldots, a_n}\}$ and recall that we proved (5.12) is consistent with the correspondence (1.8) established in Chapter 1 by showing that $f_W$ has first and second roots

$$\eta = -[0, a_{-1}, a_{-2}, a_{-3}, \ldots] \qquad \text{and} \qquad \xi = [a_0, a_1, a_2, \ldots],$$

respectively. The significance of (1.8) is that $M(f_W) = M(\mathcal{A})$. We know from Theorem 5.3 that $M(A) = \lambda_0(A)$ and hence $M(f_W) = \lambda_0(A) = \xi - \eta$. Since we can assume all of the integers $a$, $b$, $c$ and $d$ are positive it is not hard to verify that

$$\eta = \frac{(a-d) - \sqrt{(d-a)^2 + 4bc}}{2c} \qquad \text{and} \qquad \xi = \frac{(a-d) + \sqrt{(d-a)^2 + 4bc}}{2c}.$$

It follows that

$$M(f_W) = \frac{\sqrt{(d-a)^2 + 4bc}}{c} = \sqrt{\frac{(a+d)^2 - 4}{c^2}}.$$

The last equality being due to $ad - bc = 1$. Thus we only need show

$$W = \begin{pmatrix} 9m^2 - r + 1 & s \\ 3m^2 & r+1 \end{pmatrix}$$

in order to prove the theorem is true.

We know $W = G(A^2B^2) = (G(A))^2(G(B))^2$. Hence Theorem 5.4 implies

$$W = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix}^2 \begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix}^2$$

where $p_1$ and $p_2$ are chosen so that the determinants of the matrices are 1. We shall use (5.28) and (5.31) to show that $a = 9m^2 - r + 1$, $c = 3m^2$ and $d = r + 1$. The definition of $p_1$ implies $m_1 p_1 = (3m_1 - k_1)k_1 - 1$ and hence

$$\begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix}^2 = \begin{pmatrix} 3m_1(3m_1 - k_1) - 1 & 3k_1(3m_1 - k_1) - 3 \\ 3m_1^2 & 3m_1 k_1 - 1 \end{pmatrix}.$$

Similarly, using the rearrangement of (5.33) to $3m_2 - k_2 = (m - m_2k_1)/m_1$ we have

$$\begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix}^2 = \begin{pmatrix} 3m_2(m - m_2k_1)/m_1 - 1 & 3k_2(m - m_2k_1)/m_1 - 3 \\ 3m_2^2 & 3m_2k_2 - 1 \end{pmatrix}.$$

It follows that

$$\begin{aligned} c &= 3m_1 3m_2(m - m_2k_1) - 3m_1^2 + 9m_1m_2^2k_1 - 3m_2^2 \\ &= 9mm_1m_2 - 3m_1^2 - 3m_2^2 \\ &= 3m^2 \end{aligned}$$

and

$$\begin{aligned} d &= 3m_1 3k_2(m - m_2k_1) - 9m_1^2 + (3m_1k_1 - 1)(3m_2k_2 - 1) \\ &= 9m_1(mk_2 - m_1) - 3m_1k_1 - 3m_2k_2 + 1 \\ &= r + 1 \end{aligned}$$

and

$$\begin{aligned} a =& (3m_1(3m_1 - k_1) - 1)\left(\frac{3m_2(m - m_2k_1)}{m_1} - 1\right) \\ &+ 3m_2^2 3k_1(3m_1 - k_1) - 9m_2^2 \\ =& 9m_2(3m_1 - k_1)(m - m_2k_1) - 3m_1(3m_1 - k_1) - \frac{3m_2(m - m_2k_1)}{m_1} + 1 \\ &+ 3m_2^2 3k_1(3m_1 - k_1) - 9m_2^2 \\ =& 9mm_2(3m_1 - k_1) - 3m_1(3m_1 - k_1) - \frac{3m_2(m - m_2k_1)}{m_1} + 1 - 9m_2^2 \\ =& 27mm_1m_2 - 9m_2(mk_1 + m_2) - 3m_1(3m_1 - k_1) - 3m_2(3m_2 - k_2) + 1 \\ =& 27mm_1m_2 - 9m_1m_2k - 9m_1^2 + 3m_1k_1 - 9m_2^2 + 3m_2k_2 + 1 \\ =& 9m^2 - r + 1. \end{aligned}$$

That $b = s$ follows from $ad - bc = 1$. The proof is complete. □

We conclude this chapter with a brief review of the situation for simple closed geodesics and related to this, Lekkerkerker's version of Theorem 5.4.

By considering the conjugacy classes in $\pi_1(\mathbf{T})$ which contain simple loops, as described in the preamble to Theorem 3.1, it can be shown that the conjugacy

classes in $\Gamma'$ which define the simple closed geodesics are of the form $[G(A)]$ where $G \in \text{Aut } \Gamma'$. Since $AB = SR(A)$ the class $[G(A)]$ can be replaced by $[G(AB)]$. From this and the arguments of Theorem 5.1 with appropriate modifications it can be deduced that the forms $f_W$ where

$$W = G(AB) \qquad \text{and} \qquad G \in \mathcal{K} \tag{5.39}$$

represent all classes of forms which map to the simple closed geodesics on $\mathbf{T}$. In other words the forms $f_W$ represent all classes containing Markoff forms. The discussion following Theorem 5.1 is still relevant. As before, each transformation $W = G(AB) = G(A)G(B)$ where $G$ is an element of $\mathcal{K}$ other than $SR^2S^2$ and $SR^3$ can be expressed as a continued fraction whose sequence of partial quotients is comprised solely of the blocks $2,2$ and $1,1$. Further, since

$$SR^3SR(AB)(z) = [2,2,1,1,z]$$

that sequence can be obtained by successively applying the substitutions $\phi$ and $\psi$ to the initial sequence $2,2,1,1$. Thus the periods of the sequences which map to simple closed geodesics can be arranged as a tree like that in Figure 5.3. For the singular entries in the tree we have $SR^3(AB)(z) = C(z) = [2,2,z]$ and, replacing $SR^2S^2$ by $S^2$ which is allowable since $SR^2 \in \Psi$, we also have $S^2(AB)(z) = D(z) = [1,1,z]$. Details of the construction and properties of the general period

$$a_0, a_1, \ldots, a_m$$

be found in [8], [14] and [17], for example. The properties of the associated doubly infinite sequences $A = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_0, a_1, \ldots, a_m}\}$ are similar to those of the sequences arising from the proper closed 1-intersectors, as described in Theorem 5.3. To be precise, if $m$ is minimal and if $A$ is indexed so that $M(A) = \lambda_0(A)$ then we can choose the period $a_0, \ldots, a_m$ so that it is of the form

$$2, 2, a_2, a_3, \ldots, a_{m-2}, 1, 1 \tag{5.40}$$

where $a_2, a_3, \ldots, a_{m-2}$ is symmetric. Finally, an application of Theorem 5.4 yields the analogue of Theorem 5.5. In particular, if $(m,k;\, m_1,k_1;\, m_2,k_2)$ is the non-singular ordered Markoff set corresponding to $G$ and if $W = G(AB) = G(A)G(B)$

then

$$W = \begin{pmatrix} 3m_1 - k_1 & p_1 \\ m_1 & k_1 \end{pmatrix} \begin{pmatrix} 3m_2 - k_2 & p_2 \\ m_2 & k_2 \end{pmatrix} = \begin{pmatrix} 3m - k & p \\ m & k \end{pmatrix}$$

and so

$$f_W(x,y) = mx^2 - (3m - 2k)xy - py^2 \tag{5.41}$$

and

$$M(f_W) = \sqrt{9 - \frac{4}{m^2}}. \tag{5.42}$$

While $f_W$ is not the usual representative of the class of Markoff forms associated with the triple $(m, m_1, m_2)$ it is closely related. We describe the connection after discussing Lekkerkerker's work.

Lekkerkerker, [26], has produced a concise formulation of the usual representatives for the Markoff forms from Cohn's work with triples of matrices, [6], and Cassels' solution to the diophantine equation (5.28). An indication of Lekkerkerker's results follows. If in the definition of an ordered Markoff set given in the proof of Theorem 5.4 we replace the restriction $m/2 < k < m$ by $0 < k < m/2$ and interchange the order of $m_1$ and $m_2$ then the ordered Markoff set (5.32) becomes

$$(m, \tilde{k};\ m_2, \tilde{k}_2;\ m_1, \tilde{k}_1)$$

where

$$\tilde{k} = m - k, \qquad \tilde{k}_1 = m_1 - k_1, \qquad \tilde{k}_2 = m_2 - k_2.$$

These new sets are the sets originally defined by Cassels, [5]. It is not hard to verify that

$$\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3m - k & p \\ m & k \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \tilde{k} & \tilde{p} \\ m & 3m - \tilde{k} \end{pmatrix}$$

where as usual $p$ and $\tilde{p}$ are chosen so that the the determinants of the matrices are 1. The same is true for the other pairs with the appropriate choice of $p_1$, $\tilde{p}_1$ and $p_2$, $\tilde{p}_2$ and it follows that

$$\begin{pmatrix} \tilde{k}_1 & \tilde{p}_1 \\ m_1 & 3m_1 - \tilde{k}_1 \end{pmatrix} \begin{pmatrix} \tilde{k}_2 & \tilde{p}_2 \\ m_2 & 3m_2 - \tilde{k}_2 \end{pmatrix} = \begin{pmatrix} \tilde{k} & \tilde{p} \\ m & 3m - \tilde{k} \end{pmatrix}.$$

The corresponding set of automorphisms

$$G(A,B) = \left( \begin{pmatrix} \tilde{k}_1 & \tilde{p}_1 \\ m_1 & 3m_1 - \tilde{k}_1 \end{pmatrix}, \begin{pmatrix} \tilde{k}_2 & \tilde{p}_2 \\ m_2 & 3m_2 - \tilde{k}_2 \end{pmatrix} \right)$$

is the focus of Lekkerkerker's work. To see the direct connection between Lekkerkerker's set and ours, note that $\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ is the matrix of the transformation $U_1^{-2}$ where $U_1(z) = z + 1$. We know from the section of Chapter 1 dealing with $\Psi$ that conjugation by $U^{-2}$ can be achieved by applying the automorphism $(SR^2)^{-2} = (R^2S^2)^2$. Thus the non-singular elements of Lekkerkerker's set are of the form $(R^2S^2)^2\, G$ where $G$ is a non-singular element of $\mathcal{K}$. We define

$$\widetilde{\mathcal{K}} = \{(R^2S^2)^2\, S^2\} \cup \{(R^2S^2)^2\, G \ : \ G \in \mathcal{K},\ G \neq SR^2S^2\}.$$

Since $(R^2S^2)^2 \in \Psi$ it is clear that the forms $f_W$ where

$$W = G(AB) \qquad \text{and} \qquad G \in \widetilde{\mathcal{K}} \tag{5.43}$$

represent the same classes of forms as do the forms $f_W$ where (5.39) is true. Hence they represent all classes containing Markoff forms. Now observe that if $(m, \tilde{k};\ m_2, \tilde{k}_2;\ m_1, \tilde{k}_1)$ is the ordered Markoff set corresponding to $G \in \widetilde{\mathcal{K}}$ and if $W = G(AB)$ then

$$f_W(x,y) = mx^2 + (3m - 2\tilde{k})xy - \tilde{p}y^2. \tag{5.44}$$

These forms are exactly the Markoff forms described by Cassels.

**Remark 5.3.** In terms of the size of coefficients the forms (5.41) are just as efficient at representing their equivalence class as the forms (5.44). Specifically,

$$m + (3m - 2k) + p = m + (3m - 2\tilde{k}) + \tilde{p}.$$

Given that the transformations $W$ which define the forms (5.41) can be written in the form $W(z) = [a_0, a_1, \ldots, a_m, z]$ where $a_0, a_1, \ldots, a_m$ is the period of the associated sequence of integers we would argue that the forms (5.41) are preferable representatives.

**Remark 5.4.** We can now explain our comment in Chapter 1 that $\Psi$ is the largest subgroup of Aut $\Gamma'$ which preserves the Markoff values of the conjugacy classes in $\Gamma'$. By the Markoff value of a conjugacy class $[W]$ we mean of course the Markoff value of the form $f_W$. Suppose $G$ does not belong to $\Psi$. We shall show that $G$ does not preserve the Markoff value of the class $[AB]$. We know $G = HG'$ for some $H \in \Psi$ and $G' \in \widetilde{\mathcal{K}}$. Clearly $G' \neq (R^2S^2)^2S^2$ else $G' \in \Psi$. Since $H \in \Psi$ the Markoff value of $[G(AB)]$ is the same as that of $[G'(AB)]$. The discussion above shows that the Markoff value $[G'(AB)]$ is of the form $\sqrt{(9-4/m^2)}$ where $m$ is a Markoff number other than 1. The Markoff value of $[AB]$ is of course $\sqrt{5}$. We conclude as required that the Markoff values of $[G(AB)]$ and $[AB]$ differ.

CHAPTER 6

# ISOLATION RESULTS

Our primary aim in this chapter is to provide evidence for our conjecture that the Markoff values of the proper closed 1-intersectors are isolated in the Markoff spectrum. However, having established the means of doing this, we shall also be able to prove the existence of two new families of values which are isolated in the spectrum. Our calculations are based on the description of the spectrum in terms of doubly infinite sequences of positive integers. While we cannot prove in general that the Markoff value of a proper closed 1-intersector is isolated we can, in effect, describe a large class of integer sequences whose Markoff values are bounded away from the given one. By estimating the possible Markoff values of the remaining integer sequences we verify that the first few proper closed 1-intersectors do have isolated Markoff values.

The first restriction we can place on the integer sequences whose Markoff values are close to those of the proper closed 1-intersectors is easy to obtain. We have seen that the integer sequences arising from the proper closed 1-intersectors are composed solely of 1's and 2's. It follows that their Markoff values all lie in the portion of the spectrum below $\sqrt{12}$. It is also well-known that the Markoff value of an integer sequence is greater than or equal to $\sqrt{13}$ if and only if at least one of its terms is 3 or larger. Thus we can restrict our attention to sequences consisting solely of 1's and 2's. This restriction will also apply to the two families of isolated values mentioned above for exactly the same reasons. There is another elementary restriction we can make on the sequences we consider, if desired. It arises from the fact that the Markoff spectrum (excluding $\infty$) is covered by the Markoff values of the sequences of positive integers $\mathcal{A}$ for which $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. In order that our results be as general as possible, we will not always apply these conditions.

We shall of course be dealing with continued fractions. The following background material and notation will be useful. Details may be found in [4], [14], [15], [16] and [19]. As in Cusick and Flahive's book [14], we denote the numerator of the finite continued fraction $[0, a_1, a_2, \ldots, a_n]$ by $K(a_1, a_2, \ldots, a_n)$ so that

$$[a_0, a_1, a_2, \ldots, a_n] = \frac{K(a_0, a_1, \ldots, a_n)}{K(a_1, a_2, \ldots, a_n)}.$$

It can be calculated recursively from

$$K(a_1, a_2, \ldots, a_n) = K(a_1, a_2, \ldots, a_{n-2}) + a_n K(a_1, a_2, \ldots, a_{n-1}),$$

where $n \geq 3$, the recursion being initiated with $K(a_1) = a_1$ and $K(a_1, a_2) = a_1 a_2 + 1$. Note that

$$K(a_1, \ldots, a_n, 2) = K(a_1, \ldots, a_n, 1, 1)$$

and

$$K(a_1, a_2, \ldots, a_n) = K(a_n, \ldots, a_2, a_1).$$

We shall also make use of functions of the form

$$f(x) = \frac{ax + b}{cx + d} = [a_0, a_1, \ldots, a_n, x]$$

where $n \geq 0$ and $a_0, a_1, \ldots, a_n$ is a sequence of positive integers (except that possibly $a_0 = 0$). It is well-known that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots\cdots \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}.$$

Further, $a/c = [a_0, a_1, \ldots, a_n]$ and $b/d = [a_0, a_1, \ldots, a_{n-1}]$ and in particular,

$$(6.1) \qquad c = K(a_1, a_2, \ldots, a_n) \qquad \text{and} \qquad d = K(a_1, a_2, \ldots, a_{n-1})$$

A simple calculation using the fact that $ad - bc = (-1)^{n+1}$ shows

$$(6.2) \qquad f(x) - f(y) = \frac{(-1)^{n+1}(x - y)}{(cx + d)(cy + d)}.$$

It is easy to deduce from this that $f(x)$ is an increasing function of $x$ if $n$ is odd and a decreasing function of $x$ if $n$ is even. We shall use this fact often without further comment. We also observe that $f'(x) = (-1)^{n+1}/(cx+d)^2$. Thus $|f'(x)| < 1$ for $x > 1$ and so the function

$$g(x) = x + [a_0, a_1, \ldots, a_n, x] \tag{6.3}$$

is strictly increasing for $x \geq 1$.

We now state and prove three lemmas. Although the first is not new in its essence, our restriction to sequences of 1's and 2's allows us to obtain the specific constant $\delta$ mentioned.

**Lemma 6.1.** *Let $\mathcal{A} = \{a_i\}_{i=0}^{+\infty}$ and $\mathcal{B} = \{b_i\}_{i=0}^{+\infty}$ be sequences of 1's and 2's and set $\delta = 8\sqrt{3}/3 - 4$. If there is an integer $n \geq 0$ such that $b_i = a_i$ for $0 \leq i \leq n$ then*

$$|[a_0, a_1, a_2, \ldots] - [b_0, b_1, b_2, \ldots]| \leq \frac{\delta}{(K(a_1, a_2, \ldots, a_n, 1))^2}$$

*with equality possible only when $n = 1$ and $a_1 = 1$. Further, if $b_{n+1} \neq a_{n+1}$ then*

$$|[a_0, a_1, a_2, \ldots] - [b_0, b_1, b_2, \ldots]| \geq \frac{\delta}{(K(a_1, a_2, \ldots, a_n, 2))^2}$$

*with equality possible only when $n = 0$.*

*Proof.* Let $\mathcal{A}$ and $\mathcal{B}$ be as described and suppose $b_i = a_i$ for $0 \leq i \leq n$. Set

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots\cdots \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}$$

so that

$$f(x) = \frac{ax+b}{cx+d} = [a_0, a_1, \ldots, a_n, x].$$

It is not hard to deduce from (6.1) and the recursion formula for $K$ that

$$K(a_1, a_2, \ldots, a_n, 1) = c + d \qquad \text{and} \qquad K(a_1, a_2, \ldots, a_n, 2) = 2c + d.$$

We also set

$$\alpha = [a_{n+1}, a_{n+2}, a_{n+3}, \ldots] \qquad \text{and} \qquad \beta = [b_{n+1}, b_{n+2}, b_{n+3}, \ldots],$$

so that

$$|[a_0, a_1, a_2, \ldots] - [b_0, b_1, b_2, \ldots]| = |f(\alpha) - f(\beta)|.$$

The first inequality of the lemma can be re-written as $|f(\alpha) - f(\beta)| \leq \delta/(c+d)^2$. By interchanging $\mathcal{A}$ and $\mathcal{B}$ if necessary, we may assume $\beta \leq \alpha$ and thus

$$(1+\sqrt{3})/2 = [\overline{1,2}] \leq \beta \leq \alpha \leq [\overline{2,1}] = 1+\sqrt{3}.$$

Since $f'(x)$ is non-zero on $x > 0$, it follows that

$$|f(\alpha) - f(\beta)| \leq |f(1+\sqrt{3}) - f((1+\sqrt{3})/2)|$$

By (6.2) we have

$$|f(1+\sqrt{3}) - f((1+\sqrt{3})/2)| = \frac{(1+\sqrt{3})/2}{((1+\sqrt{3})c+d)((1+\sqrt{3})c/2+d)}.$$

and therefore to prove $|f(\alpha) - f(\beta)| \leq \delta/(c+d)^2$ it suffices to show

$$\frac{(1+\sqrt{3})/2}{((1+\sqrt{3})c+d)((1+\sqrt{3})c/2+d)} \leq \frac{\delta}{(c+d)^2}$$

or equivalently (when $d \neq 0$)

$$(6.4)\quad (1+\sqrt{3})(c/d+1)^2/2 \leq (8\sqrt{3}/3-4)((1+\sqrt{3})c/d+1)((1+\sqrt{3})(c/d)/2+1).$$

Clearly $d = 0$ if and only if $n = 0$ and if that is the case then $c = 1$ and

$$|f(1+\sqrt{3}) - f((1+\sqrt{3})/2)| = 1/(1+\sqrt{3}) < 8\sqrt{3}/3 - 4 = \delta/(c+d)^2.$$

Assuming $d \neq 0$, it can be deduced from (6.1) that $c/d = [a_n, a_{n-1} \ldots, a_1]$ and hence $1 \leq c/d \leq 3$. Using this it is not hard to verify that (6.4) is true and that there is equality only when $c/d = 1$. The first statement of the lemma follows since $c/d = 1$ implies $n = 1$ and $a_1 = 1$.

Now suppose that $b_{n+1} \neq a_{n+1}$ and re-write the second inequality of the lemma as $|f(\alpha) - f(\beta)| \geq \delta/(2c+d)^2$. Again we assume $\beta \leq \alpha$ and thus $b_{n+1} = 1$ and $a_{n+1} = 2$ and

$$\beta \leq [1, \overline{1,2}] = \sqrt{3} \leq (3+\sqrt{3})/2 = [2, \overline{2,1}] \leq \alpha$$

and so

$$|f(\alpha) - f(\beta)| \geq |f((3+\sqrt{3})/2) - f(\sqrt{3})| = \frac{(3-\sqrt{3})/2}{((3+\sqrt{3})c/2+d)(\sqrt{3}c+d)}.$$

It remains to show that

$$\frac{(3-\sqrt{3})/2}{((3+\sqrt{3})c/2+d)(\sqrt{3}c+d)} \geq \frac{\delta}{(2c+d)^2}. \tag{6.5}$$

If $d = 0$ then $n = 0$ and $c = 1$ and so the two terms are in fact equal. When $d \neq 0$ we re-arrange (6.5) to

$$(3-\sqrt{3})(2c/d+1)^2/2 \geq (8\sqrt{3}/3-4)((3+\sqrt{3})(c/d)/2+1)(\sqrt{3}c/d+1)$$

and observe that again this is true (without equality) since $1 \leq c/d \leq 3$. □

**Lemma 6.2.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ and $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ be sequences of 1's and 2's and suppose there are integers $m, n \geq 0$ such that $b_i = a_i$ for $-m \leq i \leq n$. If further, $(-1)^n(a_{n+1} - b_{n+1}) > 0$ and $(-1)^m(a_{-m-1} - b_{-m-1}) > 0$ then*

$$\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) > \min\left\{\frac{1}{(K(a_1,a_2,\ldots,a_n,2))^2}, \frac{1}{(K(a_{-1},a_{-2},\ldots,a_{-m},2))^2}\right\}.$$

*Proof.* From the definition of $\lambda_0(\mathcal{B})$ and $\lambda_0(\mathcal{A})$, see Chapter 1, and our assumption that $b_i = a_i$ for $-m \leq i \leq n$ we know that $\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) = d_1 + d_2$ where

$$d_1 = [a_0, a_1, \ldots, a_n, b_{n+1}, b_{n+2}, \ldots] - [a_0, a_1, \ldots, a_n, a_{n+1}, a_{n+2}, \ldots]$$

and

$$d_2 = [a_0, a_{-1}, \ldots, a_{-m}, b_{-m-1}, b_{-m-2}, \ldots] - [a_0, a_{-1}, \ldots, a_{-m}, a_{-m-1}, a_{-m-2}, \ldots].$$

Our assumption that $(-1)^n(a_{n+1} - b_{n+1}) > 0$ implies $d_1 > 0$. Similarly, the inequality $(-1)^m(a_{-m-1} - b_{-m-1}) > 0$ implies $d_2 > 0$. Therefore $d_1 = |d_1|$ and $d_2 = |d_2|$ and an application of Lemma 6.1 yields

$$d_1 \geq \frac{\delta}{(K(a_1,a_2,\ldots,a_n,2))^2} \quad \text{and} \quad d_2 \geq \frac{\delta}{(K(a_{-1},a_{-2},\ldots,a_{-m},2))^2}.$$

The truth of the lemma follows from this and the fact that $\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) = d_1 + d_2$ since $\delta > 1/2$. □

**Lemma 6.3.** *Let* $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ *and* $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ *be sequences of 1's and 2's and suppose there are integers* $m, n \geq 0$ *such that* $b_i = a_i$ *for* $-m \leq i \leq n$ *and suppose also that* $(-1)^n(a_{n+1} - b_{n+1}) > 0$. *Then* $\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) > 0$ *if*

$$K(a_1, a_2, \ldots, a_n, 2) \leq K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1) \neq 2$$

*and*

$$\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) > \frac{1}{(K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1))^4}$$

*if*

$$K(a_1, a_2, \ldots, a_n, 2) < K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1).$$

*Proof.* From the definitions and our assumption that $b_i = a_i$ for $-m \leq i \leq n$ it is clear that $\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) = d_1 - d_2$ where

$$d_1 = [a_0, a_1, \ldots, a_n, b_{n+1}, b_{n+2}, \ldots] - [a_0, a_1, \ldots, a_n, a_{n+1}, a_{n+2}, \ldots]$$

and

$$d_2 = [a_0, a_{-1}, \ldots, a_{-m}, a_{-m-1}, a_{-m-2}, \ldots] - [a_0, a_{-1}, \ldots, a_{-m}, b_{-m-1}, b_{-m-2}, \ldots].$$

Since $(-1)^n(a_{n+1} - b_{n+1}) > 0$ we know that $d_1 > 0$ and hence $d_1 = |d_1|$. Clearly $d_2 \leq |d_2|$. It follows from Lemma 6.1 that

$$d_1 \geq \frac{\delta}{(K(a_1, a_2, \ldots, a_n, 2))^2} \qquad \text{and} \qquad d_2 \leq \frac{\delta}{(K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1))^2}$$

and equality is possible in the first term only when $n = 0$ and in the second term only when $m = 1$ and $a_{-1} = 1$. Since we are not interested in the case where

$$K(a_1, a_2, \ldots, a_n, 2) = K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1) = 2,$$

we can deduce that

$$d_1 - d_2 > \delta \frac{(K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1))^2 - (K(a_1, a_2, \ldots, a_n, 2))^2}{(K(a_1, a_2, \ldots, a_n, 2)K(a_{-1}, a_{-2}, \ldots, a_{-m}, 1))^2}.$$

The truth of the lemma is now clear. □

**Remark 6.1.** Although the statement of Lemma 6.3 is asymmetrical in the indices $m$ and $n$, it can be seen by considering the reverse sequences $\overline{\mathcal{A}} = \{a_{-i}\}_{i=-\infty}^{+\infty}$ and $\overline{\mathcal{B}} = \{b_{-i}\}_{i=-\infty}^{+\infty}$, that the same result holds with the roles of $m$ and $n$ interchanged.

In Theorem 5.3 we gave a description of the doubly infinite sequences of positive integers other than $\{\overline{3}\}$ and $\{\overline{2,2,1}\}$ which arise from the proper closed 1-intersectors. We saw that each such sequence $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ has a period of length $4n$ where $n \geq 2$ and that we can choose the indexing of $\mathcal{A}$ so that the period $a_0, a_1, \ldots, a_{4n-1}$ is of the form

$$2, 2, a_2, a_3, \ldots, a_{2n-3}, 2, 2, 1, 1, a_{2n-3}, \ldots, a_3, a_2, 1, 1 \tag{6.6}$$

where $a_2, a_3, \ldots, a_{2n-3}$ is symmetric. We also saw that if this is the case then $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. It is convenient to use a different indexing here. Thus we write

$$\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}}\}$$

where $a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}$ is the period described by (6.6). Note that with this change $a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}$ is of the form

$$2, 2, a_{2n-2}, \ldots, a_4, a_3, 2, 2, 1, 1, a_3, a_4, \ldots, a_{2n-2}, 1, 1$$

where $a_3, a_4, \ldots, a_{2n-3}, a_{2n-2}$ is symmetric. It is evident from the symmetry of $\mathcal{A}$ that we still have $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. The following lemma deals with all sequences of this type. Its proof relies on Lemmas 6.2 and 6.3.

**Lemma 6.4.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}}\}$ where $n \geq 2$ be a periodic sequence of 1's and 2's such that $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and suppose that the period $a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}$ of $\mathcal{A}$ is of the form*

$$2, 2, a_{2n-2}, a_{2n-3}, \ldots, a_4, a_3, 2, 2, 1, 1, a_3, a_4, \ldots, a_{2n-3}, a_{2n-2}, 1, 1$$

*where $a_3, a_4, \ldots, a_{2n-3}, a_{2n-2}$ is symmetric. If $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is any sequence of positive integers other than $\mathcal{A}$ which satisfies $b_i = a_i$ for $-(2n-2) \leq i \leq 2n$ then*

$$M(\mathcal{B}) - M(\mathcal{A}) \geq \delta$$

*where* $\delta = \min\{(K(a_1, a_2, \ldots, a_{4n-1}, 1))^{-4}, (K(a_{-1}, a_{-2}, \ldots, a_{-(4n-1)}, 1))^{-4}\}$.

*Proof.* We shall prove the theorem by showing that if $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is a sequence of positive integers which satisfies $b_i = a_i$ for $-(2n-2) \leq i \leq 2n$ and if $M(\mathcal{B}) - M(\mathcal{A}) < \delta$ then $\mathcal{B} = \mathcal{A}$. We shall often make use of the facts that $a_{4n+i} = a_i$ and $a_{2n+1-i} = a_i$ for all integers $i$.

First we assume only that $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is a sequence for which $M(\mathcal{B}) - M(\mathcal{A}) < \delta$ and we shall describe certain restrictions on the form of $\mathcal{B}$. As indicated in the introduction, if some $b_i \geq 3$ then $M(\mathcal{B}) \geq \sqrt{13}$. We also know $M(\mathcal{A}) \leq \sqrt{12}$ and since $\delta < \sqrt{13} - \sqrt{12}$ we can assume $\mathcal{B}$ is a sequence of 1's and 2's. Next we have the restriction:-

(R1) there is no index $j$ such that $b_{j+i} = a_i$ for $-(4n-1) \leq i \leq k$ where $0 \leq k \leq 4n-1$ and $(-1)^k(a_{k+1} - b_{j+k+1}) > 0$.

To see that this is true suppose such a $j$ exists. If $k = 4n-1$ then $(-1)^k(a_{k+1} - b_{j+k+1}) > 0$ implies $-(2 - b_{j+4n}) > 0$ which is impossible. Thus $k \leq 4n-2$ and

$$\begin{aligned}
K(a_1, a_2, \ldots, a_k, 2) \leq & K(a_1, a_2, \ldots, a_{4n-2}, 2) \\
= & K(1, 1, a_3, \ldots, a_{2n-2}, 1, 1, 2, 2, a_{2n-2}, \ldots, a_3, 2) \\
= & K(2, a_3, \ldots, a_{2n-2}, 2, 2, 1, 1, a_{2n-2}, \ldots, a_3, 1, 1) \\
< & K(2, a_3, \ldots, a_{2n-2}, 2, 2, 1, 1, a_{2n-2}, \ldots, a_3, 1, 1, 1) \\
= & K(a_{-1}, a_{-2}, \ldots, a_{-(4n-1)}, 1).
\end{aligned}$$

It follows from Lemma 6.3 that $\lambda_j(\mathcal{B}) - \lambda_0(\mathcal{A}) > (K(a_{-1}, a_{-2}, \ldots, a_{-(4n-1)}, 1))^{-4}$. But $M(\mathcal{B}) \geq \lambda_j(\mathcal{B})$ and hence $M(\mathcal{B}) - M(\mathcal{A}) \geq \lambda_j(\mathcal{B}) - \lambda_0(\mathcal{A}) > \delta$, contradicting our initial assumption about $\mathcal{B}$.

We also have:-

(R2) there is no index $j$ such that $b_{j+i} = a_i$ for $-k \leq i \leq 4n-1$ where $-(4n-1) \leq -k \leq 0$ and $(-1)^k(a_{-(k+1)} - b_{j-(k+1)}) > 0$.

Again, suppose otherwise. If $k = 4n-1$ then $(-1)^k(a_{-(k+1)} - b_{j-(k+1)}) > 0$ implies $-(2 - b_{j-4n}) > 0$ which is impossible. Similarly, if $k = 4n-2$ then $1 - b_{j-(4n-1)} > 0$

which is also impossible. Thus $k \leq 4n-3$ and

$$\begin{aligned}
K(a_{-1},a_{-2},\ldots,a_{-k},2) \leq & K(a_{-1},a_{-2},\ldots,a_{-4n+3},2) \\
= & K(2,a_3,\ldots,a_{2n-2},2,2,1,1,a_{2n-2},\ldots,a_3,2) \\
= & K(1,1,a_3,\ldots,a_{2n-2},1,1,2,2,a_{2n-2},\ldots,a_3,2) \\
< & K(1,1,a_3,\ldots,a_{2n-2},1,1,2,2,a_{2n-2},\ldots,a_3,2,1) \\
= & K(a_1,a_2,\ldots,a_{4n-1},1).
\end{aligned}$$

As before, an application of Lemma 6.3 (with $\mathcal{A}$ and $\mathcal{B}$ replaced by their reverses) shows $\lambda_j(\mathcal{B})-\lambda_0(\mathcal{A}) > (K(a_1,a_2,\ldots,a_{4n-1},1))^{-4}$. Hence $M(\mathcal{B})-M(\mathcal{A}) \geq \lambda_j(\mathcal{B})-\lambda_0(\mathcal{A}) > \delta$ and we have a contradiction.

The restrictions (R1) and (R2) also apply to the sequence $\overline{\mathcal{B}} = \{b_{-i}\}_{i=-\infty}^{+\infty}$ obtained by reversing $\mathcal{B}$. To see this, observe that $\lambda_i(\overline{\mathcal{B}}) = \lambda_{-i}(\mathcal{B})$ for all integers $i$ and hence $M(\overline{\mathcal{B}}) = \lambda_0(\overline{\mathcal{B}}) = M(\mathcal{B})$ and $|M(\overline{\mathcal{B}}) - M(\mathcal{A})| < \delta$. The restriction (R1) applied to $\overline{\mathcal{B}}$ implies that there is no index j such that $b_{-(j+i)} = a_i$ for all $-(4n-1) \leq i \leq k$ where $0 \leq k \leq 4n-1$ and $(-1)^k(a_{k+1} - b_{-(j+k+1)}) > 0$. By replacing $i$ by $-i$ and $j$ by $-j$ and using the fact that $a_{-i} = a_{2n+1+i}$ we can rewrite this as:-

(R3) there is no index $j$ such that $b_{j+i} = a_{2n+1+i}$ for $-k \leq i \leq 4n-1$ where $-(4n-1) \leq -k \leq 0$ and $(-1)^k(a_{2n+1-(k+1)} - b_{j-(k+1)}) > 0$.

Similarly, (R2) implies:-

(R4) there is no index $j$ such that $b_{j+i} = a_{2n+1+i}$ for $-(4n-1) \leq i \leq k$ where $0 \leq k \leq 4n-1$ and $(-1)^k(a_{2n+1+k+1} - b_{j+k+1}) > 0$.

Now we assume $\mathcal{B}$ also satisfies $b_i = a_i$ for all $-(2n-2) \leq i \leq 2n$. In order to prove the theorem we must show $\mathcal{B} = \mathcal{A}$. We assume $\mathcal{B} \neq \mathcal{A}$ and we shall obtain a contradiction. Clearly, either there is $r \geq 2n$ such that $b_{r+1} \neq a_{r+1}$ or there is $s \geq 2n-2$ such that $b_{-(s+1)} \neq a_{-(s+1)}$ or both. By choosing $r$ and $s$ to be minimal (if they exist) it follows that at least one of the following conditions holds:-

(C1) $b_i = a_i$ for $-(2n-2) \leq i \leq r$ and $b_{r+1} \neq a_{r+1}$ for some $r \geq 2n$.

(C2) $b_i = a_i$ for $-s \leq i \leq 2n$ and $b_{-(s+1)} \neq a_{-(s+1)}$ for some $s \geq 2n-2$.

We claim that if (C1) is true then

(C3) $(-1)^r(a_{r+1} - b_{r+1}) > 0$ and

(C4) $r \leq 4n - 1$ and

(C5) $b_i \neq a_i$ for some $i$ with $-(4n-1) \leq i \leq -(2n-1)$.

To see that (C3) is true suppose not, that is, suppose $(-1)^r(a_{r+1} - b_{r+1}) < 0$. Note that $r \neq 2n$ else $(-1)^r(a_{r+1} - b_{r+1}) < 0$ implies $2 - b_{2n+1} > 0$ which is impossible. Thus we can choose an integer $d \geq 0$ such that $4nd+2n+1 \leq r < 4n(d+1)+2n+1$. Set $j = 4nd+2n+1$ and $k = r-j$. Then $b_{j+i} = a_{j+i} = a_{2n+1+i}$ for $-(4n-1) \leq i \leq k$ and $0 \leq k \leq 4n-1$ and $(-1)^k(a_{2n+1+k+1} - b_{j+k+1}) = -(-1)^r(a_{r+1} - b_{r+1}) > 0$. This contradicts (R4) and hence we can assume (C3) holds. Now suppose (C4) is false, that is, suppose $r \geq 4n$. Let $d \geq 1$ be the integer satisfying $4nd \leq r < 4n(d+1)$ and set $j = 4nd$ and $k = r-j$. Then $b_{j+i} = a_{j+i} = a_i$ for $-(4n-1) \leq i \leq k$ and $0 \leq k \leq 4n-1$ and $(-1)^k(a_{k+1} - b_{j+k+1}) = (-1)^r(a_{r+1} - b_{r+1}) > 0$, contradicting (R1). Hence we (C4) holds. Finally, if $b_i = a_i$ for $-(4n-1) \leq i \leq -(2n-1)$ then since (C1), (C3) and (C4) are true we have a contradiction of (R1) with $j = 0$ and $k = r$. Thus (C5) is true also.

Similarly, we can show that if (C2) is true then

(C6) $(-1)^s(a_{-(s+1)} - b_{-(s+1)}) > 0$ and

(C7) $s \leq 4n - 1$ and

(C8) $b_i \neq a_i$ for some $i$ with $2n + 1 \leq i \leq 4n - 1$.

Again we prove this by supposing otherwise and obtaining a contradiction. First suppose $(-1)^s(a_{-(s+1)} - b_{-(s+1)}) < 0$. Note that $s \neq 2n - 2$ else $(-1)^r(a_{r+1} - b_{r+1}) < 0$ implies $2 - b_{-(2n-1)} < 0$ which is impossible. Thus we can choose an integer $d \geq 0$ such that $4nd+2n-1 \leq s < 4n(d+1)+2n-1$. Set $j = -(4nd+2n-1)$ and $k = s+j$. Then $b_{j+i} = a_{j+i} = a_{2n+1+i}$ for $-k \leq i \leq 4n-1$ and $0 \leq k \leq 4n-1$ and $(-1)^k(a_{2n+1-(k+1)} - b_{j-(k+1)}) = -(-1)^s(a_{-(s+1)} - b_{-(s+1)}) > 0$, contradicting (R3). Hence (C6) is true. Now suppose $s \geq 4n$. Let $d \geq 1$ be the integer satisfying $4nd \leq s < 4n(d+1)$ and set $j = -4nd$ and $k = s + j$. Then $b_{j+i} = a_{j+i} = a_i$ for $-k \leq i \leq 4n - 1$ and $0 \leq k \leq 4n - 1$ and $(-1)^k(a_{-(k+1)} - b_{j-(k+1)}) = (-1)^s(a_{-(s+1)} - b_{-(s+1)}) > 0$, contradicting (R2). Hence (C7) holds. Since (C2),

(C6) and (C7) are all true it follows that if $b_i = a_i$ for $2n+1 \leq i \leq 4n-1$ then we have a contradiction of (R2) with $j = 0$ and $k = s$. Thus (C8) holds also.

We know that one of (C1) or (C2) is true. If (C1) is true then (C5) is true and hence so is (C2) and if (C2) is true then so is (C8) and hence (C1). Evidently the only possibility is that all of (C1) through to (C8) are true. Hence there is $r$ with $2n \leq r \leq 4n-1$ and $s$ with $2n-2 \leq s \leq 4n-1$ such that $b_i = a_i$ for $-s \leq i \leq r$ and $(-1)^r(a_{r+1} - b_{r+1}) > 0$ and $(-1)^s(a_{-(s+1)} - b_{-(s+1)}) > 0$. Applying Lemma 6.2 we conclude that

$$\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) > \min\left\{(K(a_1, a_2, \ldots, a_r, 2))^{-2},\ (K(a_{-1}, a_{-2}, \ldots, a_{-s}, 2))^{-2}\right\}.$$

This together with $r \leq 4n-1$ and $s \leq 4n-1$ implies $\lambda_0(\mathcal{B}) - \lambda_0(\mathcal{A}) \geq \delta$ and hence $M(\mathcal{B}) - M(\mathcal{A}) \geq \delta$. This contradiction completes the proof. □

Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be as described in Lemma 6.4. It follows from Lemma 6.4 that there is a constant $\delta > 0$ such that if $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is any sequence with $b_i = a_i$ for $-n \leq i \leq n$ then either $M(\mathcal{B}) = M(\mathcal{A})$ or the distance from $M(\mathcal{B})$ to $M(\mathcal{A})$ is at least $\delta$. We can express this more succinctly by introducing the natural topology on the space of sequences of positive integers. The topology is induced by defining the distance between two sequences $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ and $\mathcal{B}' = \{b_i'\}_{i=-\infty}^{+\infty}$ to be

$$d(\mathcal{B}, \mathcal{B}') = \begin{cases} 0 & \text{if } \mathcal{B} = \mathcal{B}' \\ 1/(k+1) & \text{if } \mathcal{B} \neq \mathcal{B}' \end{cases}$$

where $k \geq 0$ is the largest integer such that $b_i = b_i'$ for $-k < i < k$. With this definition in mind, our earlier statement is equivalent to saying that $M(\mathcal{A})$ is an isolated point of the image under the function $M$ of some neighbourhood $\mathcal{N}$ of $\mathcal{A}$. Davis and Kinney, [15], have considered such sequences before. The following definition is due to them.

**Definition 6.1.** Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a sequence of positive integers which satisfies $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. The function $M$ is said to have a *locally isolated value* at $\mathcal{A}$ if there is a constant $\delta > 0$ and an integer $n \geq 0$ such that if $\mathcal{B}$ is any sequence of positive integers with $d(\mathcal{B}, \mathcal{A}) < 1/(n+2)$ then either $M(\mathcal{B}) = M(\mathcal{A})$ or $|M(\mathcal{B}) - M(\mathcal{A})| > \delta$.

It follows from Davis and Kinney's work that $M$ takes locally isolated values at the integer sequences $\mathcal{A}$ of the form described in Lemma 6.4. It is convenient to restate the relevant result here.

**Remark 6.2.** In Theorem 5 of [15], Davis and Kinney show that if $\mathcal{A}$ is a periodic sequence of 1's and 2's with $M(\mathcal{A}) = \lambda_0(\mathcal{A})$, and if there is some odd integer $k$ such that $a_{k-i} = a_i$ for all integers $i$ then $M$ takes a locally isolated value at $\mathcal{A}$. (It is evident from their discussion of Theorem 5 that its statement contains an error, namely, the condition that $k$ be odd is not present. Note also that their restriction that $\mathcal{A}$ have a period of even length is redundant.)

Although, like Lemma 6.4, Davis and Kinney's work shows that the value of $M$ at the sequences $\mathcal{A}$ of the form described in Lemma 6.4 is locally isolated, the estimate it provides of the size of the neighbourhood $\mathcal{N}$ of $\mathcal{A}$ in which this is true is not as good. To be more precise the class of sequences whose Markoff values are shown to be bounded away from $M(\mathcal{A})$ by Lemma 6.4 is substantially larger than corresponding class in Davis and Kinney's work. In order to be able to discover isolated points of the Markoff spectrum we are of course interested in determining exactly how large this class can be made and especially in the case where $\mathcal{A}$ arises from a proper closed 1-intersector. Note also that the constant $\delta$ specified in Lemma 6.4 is larger than the corresponding constant described by Davis and Kinney.

Davis and Kinney point out that if $M$ has a locally isolated value at $\mathcal{A}$ but $M(\mathcal{A})$ is not isolated in the Markoff spectrum then there is a sequence $\mathcal{B}$ other than $\mathcal{A}$ and its reverse such that $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. We provide a brief explanation of why this is so.

**Remark 6.3.** If $M(\mathcal{A}) < \infty$ is not isolated in the Markoff spectrum then there is a sequence of integer sequences, $\mathcal{B}^{(1)}, \mathcal{B}^{(2)}, \mathcal{B}^{(3)}, \ldots$, such that $M(\mathcal{B}^{(j)}) \neq M(\mathcal{A})$ and

$$M(\mathcal{A}) = \lim_{j \to \infty} M(\mathcal{B}^{(j)}).$$

As noted at the beginning of the chapter, we can assume $M(\mathcal{B}^{(j)}) = \lambda_0(\mathcal{B}^{(j)})$.

Since the metric space consisting of all sequences of positive integers is compact there is a subsequence $\mathcal{B}^{(j(k))}$ which converges to some integer sequence $\mathcal{B}$. It is not hard to verify that $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and

$$M(\mathcal{B}) = \lim_{k\to\infty} M(\mathcal{B}^{(j(k))}) = M(\mathcal{A}).$$

(The details may be found in Lemma 6 and Theorem 8 of Chapter 1 of [14] for instance.) If also, $M$ takes a locally isolated value at $\mathcal{A}$ then by definition there is some neighbourhood $\mathcal{N}$ of $\mathcal{A}$ such that $M(\mathcal{A})$ is an isolated point of $M(\mathcal{N})$. Because $M(\mathcal{B}^{(j)})$ converges to $M(\mathcal{A})$ it follows that there is some integer $J \geq 1$ such that $\mathcal{B}^{(j)} \notin \mathcal{N}$ for all $j \geq J$. Therefore $\mathcal{B} \notin \mathcal{N}$ and in particular $\mathcal{B} \neq \mathcal{A}$. Obviously $M$ also takes a locally isolated value at the reverse of $\mathcal{A}$ and the same argument shows $\mathcal{B}$ is not the reverse of $\mathcal{A}$.

Our main result for the sequences $\mathcal{A}$ of the form described in Lemma 6.4 is presented in Theorem 6.1. In it we establish constraints on the sequences $\mathcal{B}$ for which $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. Our motivation is that if it can be shown that there are no such sequences (other than $\mathcal{A}$ and its reverse) then according to Remark 6.3 the value $M(\mathcal{A})$ is isolated in the Markoff spectrum. While there are some sequences $\mathcal{A}$ such that $M(\mathcal{A})$ is an isolated point of the spectrum and $M(\mathcal{A}) = M(\mathcal{B}) = \lambda_0(\mathcal{B})$ for some sequence $\mathcal{B}$ other than $\mathcal{A}$ and its reverse, we are presuming, at least in the case where $\mathcal{A}$ arises from a proper closed 1-intersector, that this is not so. In connection with this presumption we mention the well-known conjecture about the uniqueness of Markoff numbers. That conjecture is equivalent to our presumption in the case where $\mathcal{A}$ arises from a simple closed geodesic. Even if our presumption is not correct Theorem 6.1 is still useful. We demonstrate this in the remark following it where we deduce from it information about the class of sequences whose Markoff values are bounded away from $M(\mathcal{A})$.

Before we state and prove Theorem 6.1 we record, for use in its proof and elsewhere, some simple consequences of the results in Bumby's paper, [4].

**Remark 6.4.** Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ and $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ be sequences of 1's and 2's with $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and $M(\mathcal{B}) = \lambda_0(\mathcal{B})$. We claim that if $a_{-1}, a_0, a_1 = 2, 2, 1$ then

$M(\mathcal{B})$ is bounded away from $M(\mathcal{A})$ unless $b_{-1}, b_0, b_1 = 2,2,1$ or $b_{-1}, b_0, b_1 = 1,2,2$. To see this, observe that according to Bumby's work, [4], if $\mathcal{A}$ is as described then

$$\sqrt{221}/5 = M(\overline{2,2,1,1}) \leq M(\mathcal{A}) \leq M(\overline{2,2,2,1}) = 40\sqrt{30}/7. \tag{6.7}$$

If $b_0 = 1$ then $M(\mathcal{B}) = M(\overline{1}) = \sqrt{5} < \sqrt{221}/5$, if $b_{-1}, b_0, b_1 = 2,2,2$ then $M(\mathcal{B}) = M(\overline{2}) = \sqrt{8} < \sqrt{221}/5$ and if $b_{-1}, b_0, b_1 = 1,2,1$ then $M(\mathcal{B}) \geq M(\overline{2,1,1}) = \sqrt{10} > 40\sqrt{30}/7$. In each of these cases $M(\mathcal{B})$ is bounded away from $M(\mathcal{A})$. The only other possibility is that $b_{-1}, b_0, b_1$ is $2,2,1$ or $1,2,2$ and hence the claim is true.

Similarly, we claim that if $a_{-1}, a_0, a_1, a_2 = 2,2,1,1$ then $M(\mathcal{B})$ is bounded away from $M(\mathcal{A})$ unless $b_{-1}, b_0, b_1, b_2 = 2,2,1,1$ or $b_{-2}, b_{-1}, b_0, b_1 = 1,1,2,2$. This time, as well as (6.7), we have

$$M(\mathcal{A}) \leq M(\overline{2,2,2,1,1,1}) = \sqrt{3360}/19.$$

We have already seen that $b_{-1}, b_0, b_1 = 2,2,1$ or $b_{-1}, b_0, b_1 = 1,2,2$. It follows from Bumby's work that if $b_{-1}, b_0, b_1, b_2 = 2,2,1,2$ or $b_{-2}, b_{-1}, b_0, b_1 = 2,1,2,2$ then

$$M(\mathcal{B}) \geq M(\overline{2,2,2,2,1,2,2,1}) = \sqrt{233285}/155 > \sqrt{3360}/19$$

and $M(\mathcal{B})$ is bounded away from $M(\mathcal{A})$. The truth of the claim is now evident.

**Theorem 6.1.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}}\}$ where $n \geq 2$ be a periodic sequence of 1's and 2's such that $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and suppose that the period $a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n}$ of $\mathcal{A}$ is of the form*

$$2,2,a_{2n-2},a_{2n-3},\ldots,a_4,a_3,2,2,1,1,a_3,a_4,\ldots,a_{2n-3},a_{2n-2},1,1 \tag{6.8}$$

*where $a_3, a_4, \ldots, a_{2n-3}, a_{2n-2}$ is symmetric. If $M(\mathcal{A})$ is not an isolated value of the Markoff spectrum then there is a sequence $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ of 1's and 2's other than $\mathcal{A}$ such that $b_{-1}, b_0, b_1, b_2 = 2,2,1,1$ and $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. Further, for each such sequence there is $r$ with $2 \leq r \leq 2n-3$ such that $b_i = a_i$ for $-(r-1) \leq i \leq r$ and $b_{-r} = b_{r+1} \neq a_{r+1} = a_{-r}$.*

*Proof.* Let $\mathcal{A}$ be as described and suppose $M(\mathcal{A})$ is not isolated. It is clear from Lemma 6.4 or Remark 6.2 that $M$ takes a locally isolated value at $\mathcal{A}$. Thus

Remark 6.3 implies there is a sequence of integers $\mathcal{B}$, other than $\mathcal{A}$ or its reverse, satisfying $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. Since $M(\mathcal{A}) < \sqrt{12}$ we know $\mathcal{B}$ consists solely of 1's and 2's. Further, it is evident from Remark 6.4 that either $b_{-1}, b_0, b_1, b_2 = 2, 2, 1, 1$ or $b_{-2}, b_{-1}, b_0, b_1 = 1, 1, 2, 2$. By reversing $\mathcal{B}$ if necessary we can assume the former is true. It remains to show that there is an integer $r$ as described.

Lemma 6.4 applies to $\mathcal{A}$ and since $M(\mathcal{B}) = M(\mathcal{A})$ we can conclude that $b_i \neq a_i$ for some $i$ with $-(2n-2) \leq i \leq 2n$. There are two possibilities:-

(C1) $b_i = a_i$ for $i = 3, 4, \ldots, r$ and $b_{r+1} \neq a_{r+1}$ where $2 \leq r \leq 2n-1$.

(C2) $b_i = a_i$ for $i = -2, -3, \ldots, -s$ and $b_{-(s+1)} \neq a_{-(s+1)}$ where $1 \leq s \leq 2n-3$.

We can complete the proof by showing that both (C1) and (C2) are true and that $r = s + 1 \leq 2n - 3$ since the condition $b_{-r} = b_{r+1} \neq a_{r+1} = a_{-r}$ is an easy consequence of this and the symmetry of the period of $\mathcal{A}$.

We claim first that if (C2) holds then so does (C1) and $r < s + 2$. Suppose not. That is, suppose (C2) holds and that either (C1) does not or (C1) holds with $r \geq s + 2$. Then $b_i = a_i$ for $-s \leq i \leq s + 2$. Since $b_{-(s+1)} \neq a_{-(s+1)}$ and

$$\begin{aligned} K(a_{-1}, a_{-2}, \ldots, a_{-s}, 2) &= K(2, a_3, a_4, \ldots, a_{s+1}, 2) \\ &\leq K(1, 1, a_3, a_4, \ldots, a_{s+1}, a_{s+2}, 1) \\ &= K(a_1, a_2, \ldots, a_{s+2}, 1) \neq 2 \end{aligned}$$

we can apply Lemma 6.3 (with $\mathcal{A}$ and $\mathcal{B}$ replaced by their reverses). We conclude that $\lambda_0(\mathcal{B}) \neq \lambda_0(\mathcal{A})$ and hence $M(\mathcal{B}) \neq M(\mathcal{A})$ and we have a contradiction.

Next, we claim that if (C1) holds then so does (C2) and $s < r$. Again, suppose not. That is, suppose (C1) holds and that either (C2) does not or (C2) holds with $s \geq r$. Assuming for the moment that $r \leq 2n - 2$ we have $b_i = a_i$ for $-r \leq i \leq r$. Since $b_{r+1} \neq a_{r+1}$ and

$$\begin{aligned} K(a_1, a_2, \ldots, a_r, 2) &= K(1, 1, a_3, a_4, \ldots, a_r, 2) \\ &= K(2, a_{-2}, a_{-3}, \ldots, a_{-(r-1)}, 2) \\ &\leq K(a_{-1}, a_{-2}, \ldots, a_{-(r-1)}, a_{-r}, 1) \neq 2 \end{aligned}$$

Lemma 6.3 again implies $\lambda_0(\mathcal{B}) \neq \lambda_0(\mathcal{A})$. Hence $M(\mathcal{B}) \neq M(\mathcal{A})$ and we have a contradiction. It remains to consider the case where $r = 2n-1$. In this case, $b_i = a_i$ for $-(2n-2) \leq i \leq 2n-1$. If also $b_{-(2n-1)} = a_{-(2n-1)}$ then since

$$\begin{aligned} K(a_1, a_2, \ldots, a_{2n-1}, 2) &= K(1, 1, a_3, a_4, \ldots, a_{2n-2}, 1, 2) \\ &< K(2, a_3, a_4, \ldots, a_{2n-2}, 2, 2, 1) \\ &= K(a_{-1}, a_{-2}, \ldots, a_{-(2n-1)}, 1) \end{aligned}$$

and $b_{r+1} \neq a_{r+1}$, an application of Lemma 6.3 leads as usual to a contradiction. If $b_{-(2n-1)} \neq a_{-(2n-1)}$ then since $a_{-(2n-1)} = 2$ we have $(-1)^{2n-2}(a_{-(2n-1)} - b_{-(2n-1)}) > 0$. We already know $b_{r+1} \neq a_{r+1}$ and since $a_{2n} = 1$ we also have $(-1)^{2n-1}(a_{2n} - b_{2n}) > 0$. In this case, Lemma 6.2 implies $\lambda_0(\mathcal{B}) \neq \lambda_0(\mathcal{A})$ and again we have a contradiction. There are no more possibilities and the claim is proven.

Since one of (C1) or (C2) is true it follows from the two claims above that both (C1) and (C2) are true. Further, $s < r < s+2$ and so $r = s+1$. We know $r \leq 2n-2$ since $s \leq 2n-3$ and therefore it only remains to show that $r \leq 2n-3$. Suppose not, that is, suppose $r = 2n-2$. Then $b_i = a_i$ for $-(2n-3) \leq i \leq 2n-2$ and $b_{2n-1} \neq a_{2n-1} = 1$ and $b_{-(2n-2)} \neq a_{-(2n-2)} = 2$. Further, $(-1)^{2n-2}(a_{2n-1} - b_{2n-1}) < 0$ and $(-1)^{2n-3}(a_{-(2n-2)} - b_{-(2n-2)}) < 0$ and so Lemma 6.2 (with $\mathcal{A}$ and $\mathcal{B}$ interchanged) implies $\lambda_0(\mathcal{B}) \neq \lambda_0(\mathcal{A})$ and hence $M(\mathcal{B}) \neq M(\mathcal{A})$. This contradiction completes the proof. □

**Remark 6.5.** Let $\mathcal{A}$ be as described in Theorem 6.1. We claim that if $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is a sequence of 1's and 2's other than $\mathcal{A}$ with $b_{-1}, b_0, b_1, b_2 = 2, 2, 1, 1$ and $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and if there is no integer $r$ such that $2 \leq r \leq 2n-3$ and $b_i = a_i$ for $-(r-1) \leq i \leq r$ and $b_{-r} = b_{r+1} \neq a_{r+1} = a_{-r}$ then $M(\mathcal{B})$ is bounded away from $M(\mathcal{A})$. To see that this true, suppose otherwise. In this case, there is a sequence of sequences, $\mathcal{B}^{(1)}, \mathcal{B}^{(2)}, \mathcal{B}^{(3)}, \ldots$, such that, firstly, $M(\mathcal{A}) = \lim_{j\to\infty} M(\mathcal{B}^{(j)})$ and, secondly, for each $j \geq 1$ the sequence $\mathcal{B}^{(j)} = \{b^{(j)}{}_i\}_{i=-\infty}^{+\infty}$ is a sequence of 1's and 2's other than $\mathcal{A}$ which satisfies $b_{-1}^{(j)}, b_0^{(j)}, b_1^{(j)}, b_2^{(j)} = 2, 2, 1, 1$ and $M(\mathcal{B}^{(j)}) = \lambda_0(\mathcal{B}^{(j)})$ and for which there is no integer $r$ with $2 \leq r \leq 2n-3$ such that $b_i^{(j)} = a_i$ for

$-(r-1) \leq i \leq r$ and $b_{-r}^{(j)} = b_{r+1}^{(j)} \neq a_{r+1} = a_{-r}$. As was done in Remark 6.3, we let $\mathcal{B}$ be the limit of some subsequence $\mathcal{B}^{(j(k))}$ so that $M(\mathcal{B}) = M(\mathcal{A})$. It is not hard to verify that $\mathcal{B}$ satisfies the same conditions as each $\mathcal{B}^{(j(k))}$ and hence we have a contradiction of Theorem 6.1.

We can now describe how we have tested our conjecture that the Markoff values of the proper closed 1-intersectors are isolated points of the spectrum. For this purpose we let $\mathcal{A}$ be a sequence of integers arising from some proper closed 1-intersector. We know that $\mathcal{A}$ satisfies the hypothesis of Theorem 6.1. Therefore, in order to prove $M(\mathcal{A})$ is an isolated point of the spectrum, it suffices to show that $M(\mathcal{B}) \neq M(\mathcal{A})$ for every sequence $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ of 1's and 2's other than $\mathcal{A}$ with $b_{-1}, b_0, b_1, b_2 = 2, 2, 1, 1$ and $M(\mathcal{B}) = \lambda_0(\mathcal{B})$. Thus we assume $M(\mathcal{B})$ is such a sequence and we attempt to show $M(\mathcal{B}) \neq M(\mathcal{A})$. We only need to consider the case where there is some $r$ with $2 \leq r \leq 2n-3$ such that $b_i = a_i$ for $-(r-1) \leq i \leq r$ and $b_{-r} = b_{r+1} \neq a_{r+1} = a_{-r}$ because it follows immediately from Theorem 6.1 that $M(\mathcal{B}) \neq M(\mathcal{A})$ if no such $r$ exists. For each such $r$ we consider the possibilities for the segments

$$b_{-m}, \ldots, b_{-(r+2)}, b_{-(r+1)} \qquad \text{and} \qquad b_{r+2}, b_{r+3}, \ldots, b_n \tag{6.9}$$

of $\mathcal{B}$ as $n$ and $m$ increase to $\infty$. For each choice of the segments (6.9) we calculate bounds on the range of $M(\mathcal{B})$. If for some choice of $m$ and $n$ we find that $M(\mathcal{A})$ lies outside all the relevant bounds then it is not possible that $M(\mathcal{B}) = M(\mathcal{A})$ and we can consider the next $r$. (By considering Lemma 6.3 it can be seen that during this process is it desirable to choose $m$ and $n$ so that $K(b_{-1}, b_{-2}, \ldots, b_{-m})$ is about the same size as $K(b_1, b_2, \ldots, b_n)$.) Clearly the algorithm will terminate only if it is impossible to choose $\mathcal{B}$ so that $M(\mathcal{B}) = M(\mathcal{A})$ and therefore if it does $M(\mathcal{A})$ is an isolated point of the spectrum. Our conjecture in a slightly stronger form is that the algorithm will always terminate.

We have implemented such an algorithm on a computer and thereby tested our conjecture for the sequences in the first five non-singular levels of the tree in Figure 5.3. There are 63 such sequences. Explicit details for the first three are presented after the following remark.

**Remark 6.6.** For use in the examples below and latter, we let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a sequence of 1's and 2's with $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and we note the following. Firstly, if $a_{-1}, a_0, a_1 = 2, 2, 1$ then there is no index $i$ such that $a_{i-1}, a_i, a_{i+1} = 1, 2, 1$. To see this, suppose such an index $i$ exists. Then

$$\lambda_i(\mathcal{A}) \geq [2, 1, \overline{1, 2}] + [0, 1, \overline{1, 2}] = 2 + 2\sqrt{3}/3.$$

However, we have seen in Remark 6.4 that $M(\mathcal{A}) \leq M(\overline{2, 2, 2, 1, 1, 1}) = 4\sqrt{30}/7$. Since $4\sqrt{30}/7 < 2 + 2\sqrt{3}/3$ we have $M(\mathcal{A}) < \lambda_i(\mathcal{A})$ which is impossible.

Likewise, if $a_{-1}, a_0, a_1, a_2 = 2, 2, 1, 1$ then there is no $i$ such that $a_i, a_{i+1}, a_{i+2} = 2, 1, 2$. Again, suppose otherwise. The argument above shows $1, 2, 1$ cannot occur in $\mathcal{A}$. Hence $a_{i-1} = 2$ and

$$\lambda_i(\mathcal{A}) \geq [2, 1, 2, \overline{2, 1}] + [0, 2, \overline{1, 2}] = (45 + 13\sqrt{3})/22.$$

We know from Remark 6.4 that $M(\mathcal{A}) \leq M(\overline{2, 2, 2, 1, 1, 1}) = \sqrt{3360}/19$ and again we have a contradiction.

**Example 6.1.** The value $M(\mathcal{A})$ where $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-3}, \ldots, a_4}\}$ and

$$a_{-3}, a_{-2}, \ldots, a_4 = 2, 2, 2, 2, 1, 1, 1, 1$$

is an isolated value of the Markoff spectrum. This is an easy consequence of Theorem 6.1 since there is no integer $r$ with $2 \leq r \leq 4 - 3$.

**Example 6.2.** The value $M(\mathcal{A})$ where $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-5}, \ldots, a_6}\}$ and

$$a_{-5}, \ldots, a_6 = 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1$$

is an isolated value of the Markoff spectrum. We know $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and therefore $M(\mathcal{A}) = \sqrt{1525/169}$. By Theorem 6.1, it suffices to show that if $\mathcal{B}$ is a sequence of 1's and 2's with $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and if there is an integer $r$ with $2 \leq r \leq 3$ such that $b_i = a_i$ for $-(r-1) \leq i \leq r$ and $b_{-r} = b_{r+1} \neq a_{-r} = a_{r+1}$ then $M(\mathcal{B}) \neq M(\mathcal{A})$. We assume $\mathcal{B}$ is such a sequence. Remark 6.6 shows $1, 2, 1$

and $2,1,2$ cannot occur in $\mathcal{B}$. Hence the only possibility is that $r = 2$ and the initial segment $b_{-2},\ldots,b_0,\ldots,b_3$ of $\mathcal{B}$ is $2,2,2,1,1,2$. If $b_{-3} = 2$ then

$$M(\mathcal{B}) < [2,1,1,2,3] + [0,2,2,2,3] = 2093/697 < M(\mathcal{A}),$$

while if $b_{-3} = 1$ we know $b_{-4} = 1$ and $b_4 = 2$ else $1,2,1$ or $2,1,2$ occurs and so

$$M(\mathcal{B}) > [2,1,1,2,2,3] + [0,2,2,1,1,3] = 5296/1763 > M(\mathcal{A}).$$

Since $M(\mathcal{B}) \neq M(\mathcal{A})$ in all cases, $M(\mathcal{A})$ is isolated.

**Example 6.3.** The value $M(\mathcal{A})$ where $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-5},\ldots,a_6}\}$ and

$$a_{-5},\ldots,a_6 = 2,2,2,2,2,2,1,1,2,2,1,1$$

is an isolated value of the Markoff spectrum. In this case, $M(\mathcal{A}) = \sqrt{7573/841}$. Again, we suppose $\mathcal{B}$ is a sequence of 1's and 2's with $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and that there is an integer $r$ with $2 \leq r \leq 3$ such that $b_i = a_i$ for $-(r-1) \leq i \leq r$ and $b_{-r} = b_{r+1} \neq a_{-r} = a_{r+1}$ and we shall show that $M(\mathcal{B}) \neq M(\mathcal{A})$. By reasoning as in Example 6.2, we can assume the initial segment $b_{-2},\ldots,b_0,\ldots,b_3$ of $\mathcal{B}$ is $1,2,2,1,1,1$. Further, $b_{-3} = 1$ since Remark 6.6 shows $2,1,2$ cannot occur. If $b_4 = 2$ then

$$M(\mathcal{B}) > [2,1,1,1,2,3] + [0,2,1,1,1] = 649/216 > M(\mathcal{A}).$$

Note that we have shown $2,1,1,1,2,2,1,1$ cannot occur in $\mathcal{B}$. Now we suppose $b_4 = 1$. If $b_5 = 2$ then, either $b_{-4} = 1$ and

$$M(\mathcal{B}) < [2,1,1,1,1,2,3] + [0,2,1,1,1,1] = 1715/572 < M(\mathcal{A}),$$

or $b_{-4} = 2$ and $b_{-5} = 2$ (else $1,2,1$ occurs) and

$$M(\mathcal{B}) < [2,1,1,1,1,2,3] + [0,2,1,1,2,2,3] = 6997/2332 < M(\mathcal{A}).$$

Thus we can assume $b_5 = 1$. We can also assume $b_{-4} = 1$ else $b_{-4} = 2$ and

$$M(\mathcal{B}) > [2,1,1,1,1,1,1] + [0,2,1,1,2,3] = 1717/572 > M(\mathcal{A}).$$

Further, since $2,1,1,1,2,2,1,1$ cannot occur $b_{-5}=1$. If $b_6=2$ then $b_{-6}=1$ and

$$M(\mathcal{B}) > [2,1,1,1,1,1,2,3]+[0,2,1,1,1,1,1,3] = 16195/5396 > M(\mathcal{A}),$$

or $b_{-6}=2$ and

$$M(\mathcal{B}) > [2,1,1,1,1,1,2,3]+[0,2,1,1,1,1,2,3] = 24514/8165 > M(\mathcal{A}).$$

We assume $b_6=1$. If $b_7=2$ then

$$M(\mathcal{B}) < [2,1,1,1,1,1,1,2,3]+[0,2,1,1,1,1,3] = 16217/5405 < M(\mathcal{A}).$$

Now assume $b_7=1$. If $b_{-6}=1$ then

$$M(\mathcal{B}) < [2,1,1,1,1,1,1,1,3]+[0,2,1,1,1,1,1,1] = 3877/1292 < M(\mathcal{A}).$$

Finally, assume $b_{-6}=2$ and note that $b_{-7}=2$ (else $1,2,1$ occurs). If $b_8=2$ then

$$M(\mathcal{B}) > [2,1,1,1,1,1,1,1,2,3]+[0,2,1,1,1,1,2,2,1] = 64189/21390 > M(\mathcal{A}),$$

and if $b_8=1$ then

$$M(\mathcal{B}) < [2,1,1,1,1,1,1,1,1,1]+[0,2,1,1,1,1,2,2,3] = 28199/9405 < M(\mathcal{A}).$$

We have shown $M(\mathcal{B}) \neq M(\mathcal{A})$ in all cases and hence $M(\mathcal{A})$ is isolated.

In the remainder of this chapter we describe two new families of isolated points of the Markoff spectrum and the integer sequences which give rise to them. Our motivation came from Gbur's work, [19]. There she showed that for all $n \geq 1$ the Markoff values $M(\mathcal{A})$ where $\mathcal{A}$ is periodic with period

$$2,\overbrace{1,\ldots,1}^{2n}$$

are isolated points of the Markoff spectrum. We note that this is also a consequence of Theorem 6 in Davis and Kinney's paper [15]. It is natural to examine in a similar manner the Markoff values of the sequences which have a period of the form

$$1,\overbrace{2,\ldots,2}^{2n}.$$

These sequences constitute our first family. We prove that their Markoff values are isolated in Theorem 6.2. Our second family of sequences is closely related. We described them in Theorem 6.3. Of the first family, only the case where $n = 1$ has been discussed before, see [4]. The second family is completely new. Our methods are more direct than Gbur's. One reason for this is that we are only interested in establishing that the values are isolated rather than determining the endpoints of the neighbouring gaps. Like Davis and Kinney we make use of the fact that the value of $M$ at $\mathcal{A}$ is locally isolated.

We begin by stating the following simplified version of Proposition 2 in Bumby's paper, [4]. (A similar result may also be found in Gbur's paper.)

**Lemma 6.5.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a sequence of positive integers and suppose $n \geq 1$ and $a_1, a_2, \ldots, a_{n-1}$ is symmetric. Then $\lambda_0(\mathcal{A}) \geq \lambda_n(\mathcal{A})$ if and only if*

$$[a_0, a_{-1}, a_{-2}, \ldots] \geq [a_n, a_{n+1}, a_{n+2}, \ldots].$$

Using Lemma 6.5 and Davis and Kinney's results on local isolation we can prove the following general result.

**Lemma 6.6.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-(m+1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_n}\}$ be a sequence of 1's and 2's satisfying $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and suppose both the sequences*

$$a_{-m}, \ldots, a_{-2}, a_{-1} \qquad \text{and} \qquad a_1, a_2, \ldots, a_n$$

*are symmetric and $n, m \geq 0$ are both even and $a_{-(m+1)} = a_0$. Then there is a constant $\delta > 0$ such that if $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is any sequence of 1's and 2's other than $\mathcal{A}$ satisfying $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and $b_i = a_i$ for $-m \leq i \leq n$ then $M(\mathcal{B}) - M(\mathcal{A}) \geq \delta$.*

*Proof.* Let $\mathcal{A}$ be as described and suppose $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is a sequence of 1's and 2's other than $\mathcal{A}$ satisfying $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and $b_i = a_i$ for $-m \leq i \leq n$. We shall begin by showing $M(\mathcal{B}) > M(\mathcal{A})$. For this purpose we set

$$\alpha_1 = [a_{n+1}, a_{n+2}, a_{n+3}, \ldots] \qquad \alpha_2 = [a_{-(m+1)}, a_{-(m+2)}, a_{-(m+3)}, \ldots]$$

and

$$\beta_1 = [b_{n+1}, b_{n+2}, b_{n+3}, \ldots] \qquad \beta_2 = [b_{-(m+1)}, b_{-(m+2)}, b_{-(m+3)}, \ldots]$$

so that

$$M(\mathcal{A}) = [a_0, a_1, \ldots, a_n, \alpha_1] + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \alpha_2]$$

and

$$M(\mathcal{B}) = [a_0, a_1, \ldots, a_n, \beta_1] + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \beta_2].$$

Note that the periodicity and symmetry of $\mathcal{A}$ implies

$$\alpha_1 = [a_0, a_{-1}, \ldots, a_{-m}, \alpha_2] \qquad \text{and} \qquad \alpha_2 = [a_0, a_1, \ldots, a_n, \alpha_1].$$

Since $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ we know $\lambda_0(\mathcal{B}) \geq \lambda_{n+1}(\mathcal{B})$. We are assuming $a_1, a_2, \ldots, a_n$ is symmetric and therefore an application of Lemma 6.5 to $\mathcal{B}$ implies

$$\beta_1 \leq [a_0, a_{-1}, \ldots, a_{-m}, \beta_2]. \tag{6.10}$$

Similarly, by applying Lemma 6.5 to the reverse of $\mathcal{B}$ we have

$$\beta_2 \leq [a_0, a_1, \ldots, a_n, \beta_1]. \tag{6.11}$$

Now suppose $\beta_1 < \alpha_1$ and set $\beta_3 = [a_0, a_1, \ldots, a_n, \beta_1]$. Constraint (6.11) can be re-written as $\beta_2 \leq \beta_3$. We are assuming $m$ is even and therefore

$$\begin{aligned} M(\mathcal{B}) &\geq [a_0, a_1, \ldots, a_n, \beta_1] + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \beta_3] \\ &= \beta_3 + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \beta_3]. \end{aligned}$$

The right hand side of this inequality is a function of the form (6.3) and hence increases with $\beta_3$. Since $n$ is even we know

$$\beta_3 = [a_0, a_1, \ldots, a_n, \beta_1] > [a_0, a_1, \ldots, a_n, \alpha_1] = \alpha_2$$

and so

$$M(\mathcal{B}) > \alpha_2 + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \alpha_2] = M(\mathcal{A}).$$

Similarly, suppose $\beta_2 < \alpha_2$ and set $\beta_3 = [a_0, a_{-1}, \ldots, a_{-m}, \beta_2]$. In this case, (6.10) implies $\beta_1 \leq \beta_3$ and so

$$\begin{aligned} M(\mathcal{B}) &\geq [a_0, a_1, \ldots, a_n, \beta_3] + [0, a_{-1}, a_{-2}, \ldots, a_{-m}, \beta_2] \\ &= \beta_3 + [0, a_1, a_2, \ldots, a_n, \beta_3]. \end{aligned}$$

We also have

$$\beta_3 = [a_0, a_{-1}, \ldots, a_{-m}, \beta_2] > [a_0, a_{-1}, \ldots, a_{-m}, \alpha_2] = \alpha_1$$

and therefore

$$M(\mathcal{B}) > \alpha_1 + [0, a_1, a_2, \ldots, a_n, \alpha_1] = M(\mathcal{A}).$$

The only other possibility is that $\beta_1 \geq \alpha_1$ and $\beta_2 \geq \alpha_2$. In this case,

$$\alpha_2 = [a_0, a_1, \ldots, a_n, \alpha_1] \geq [a_0, a_1, \ldots, a_n, \beta_1] \geq \beta_2 \geq \alpha_2$$

implying that $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2$. However, this contradicts our assumption that $\mathcal{B} \neq \mathcal{A}$. It follows that $M(\mathcal{B}) > M(\mathcal{A})$ as was claimed.

Now suppose the lemma is not true. Then for every $\delta > 0$ we can choose the sequence $\mathcal{B}$ so that it also satisfies $M(\mathcal{A}) + \delta > M(\mathcal{B}) > M(\mathcal{A})$. Hence there is a sequence of sequences $\mathcal{B}^{(j)} = \{b_i^{(j)}\}_{i=-\infty}^{+\infty}$, where $j = 1, 2, 3 \ldots$, each of which consists only of 1's and 2's and none of which is $\mathcal{A}$ such that, firstly, for each $j \geq 1$ we have $M(\mathcal{B}^{(j)}) = \lambda_0(\mathcal{B}^{(j)})$ and $b_i^{(j)} = a_i$ for all $-m \leq i \leq n$, and secondly,

$$M(\mathcal{A}) = \lim_{j \to \infty} M(\mathcal{B}^{(j)}).$$

As noted in Remark 6.3, there is a subsequence $\mathcal{B}^{(j(k))}$ which converges to a sequence $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ and further,

$$M(\mathcal{B}) = \lambda_0(\mathcal{B}) = \lim_{k \to \infty} M(\mathcal{B}^{(j(k))}) = M(\mathcal{A}).$$

It is not hard to verify that, like each of the sequences $\mathcal{B}^{(j(k))}$, the sequence $\mathcal{B}$ consists only of 1's and 2's and $b_i = a_i$ for all $-m \leq i \leq n$. Since $M(\mathcal{B}) = M(\mathcal{A})$ the sequence $\mathcal{B}$ cannot satisfy the hypothesis of our initial claim. It follows that $\mathcal{B} = \mathcal{A}$ and hence the sequence $\mathcal{B}^{(j(k))}$ converges to $\mathcal{A}$. We also know $M(\mathcal{B}^{(j(k))})$ converges to $M(\mathcal{A})$ and therefore there is no neighbourhood $\mathcal{N}$ of $\mathcal{A}$ such that $M(\mathcal{A})$ is an isolated point of the set $M(\mathcal{N})$. But $\mathcal{A}$ is a periodic sequence of 1's and 2's with $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and $m + 1$ is odd and $a_{-(m+1)-i} = a_i$ for all integers $i$ and therefore, according to Remark 6.2, $M$ takes a locally isolated value at $\mathcal{A}$. This contradiction completes the proof. □

**Remark 6.7.** In Lemma 6.6 we have assumed the sequences $\mathcal{A}$ and $\mathcal{B}$ consist of 1's and 2's rather than arbitrary positive integers. The reason we have done so is that such an assumption is made in the result of Davis and Kinney cited in the proof. We believe that that result, namely Theorem 5 of [15], does not require such an assumption and therefore it is not necessary in Lemma 6.6 either.

**Remark 6.8.** The sequences $\mathcal{A}$ discussed in Lemma 6.4 satisfy the hypothesis of Lemma 6.6. Thus a weaker version of Lemma 6.4 is deducible from Lemma 6.6. However, the proof of Lemma 6.6 provides no estimate of the size of the constant $\delta$ involved. As mentioned before, the size of $\delta$ is of interest because, in the case where $M(\mathcal{A})$ is isolated, it provides information about the size of the gap in the spectrum above $M(\mathcal{A})$.

The two families of sequences generating the Markoff values we shall be considering satisfy the hypothesis of Lemma 6.6. We shall use Lemma 6.6 to help prove their Markoff values are isolated. We need one final technical lemma before we can proceed.

**Lemma 6.7.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ and $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ be sequences of 1's and 2's such that $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ and $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and suppose*

$$\mathcal{A} = \ldots\ldots, 1, \overbrace{2, \ldots, 2}^{2n}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n}, \ldots\ldots$$

*and*

$$\mathcal{B} = \ldots\ldots, 2, \overbrace{2, \ldots, 2}^{2n}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n}, \ldots\ldots$$

*where $n \geq 1$ and the asterisk distinguishes the term with index zero. Then*

$$M(\mathcal{A}) - M(\mathcal{B}) > 0.0009\,(K(\overbrace{2, \ldots, 2}^{2n-1}))^{-2}.$$

*Proof.* Set

$$f(x) = [0, \overbrace{2, \ldots, 2}^{2n}, x] \qquad \text{and} \qquad g(x) = [0, 1, \overbrace{2, \ldots, 2}^{2n}, x]$$

so that $M(\mathcal{A}) = 2 + f(\alpha_1) + g(\alpha_2)$ and $M(\mathcal{B}) = 2 + f(\beta_1) + g(\beta_2)$ where

$$\alpha_1 = [a_{-(2n+1)}, a_{-(2n+2)}, a_{-(2n+3)}, \dots] \qquad \alpha_2 = [a_{2n+2}, a_{2n+3}, a_{2n+4}, \dots]$$

and

$$\beta_1 = [b_{-(2n+1)}, b_{-(2n+2)}, b_{-(2n+3)}, \dots] \qquad \beta_2 = [b_{2n+2}, b_{2n+3}, b_{2n+4}, \dots].$$

Hence

$$M(\mathcal{A}) - M(\mathcal{B}) = f(\alpha_1) - f(\beta_1) + g(\alpha_2) - g(\beta_2).$$

According to Remark 6.6 the sequence $1, 2, 1$ cannot occur in $\mathcal{A}$. Since $a_{-2n+1} = 1$ we have $\alpha_1 < [1, 1, 2, 2] = 12/7$. Similarly, $\alpha_2 > [1, 2, 2] = 7/5$ and $\beta_1 > [2, 2, 1] = 7/3$. We also have $\beta_2 < [2, 2, 2, 1] = 17/7$ because if not $\beta_2 = [2, 1, \dots]$. To see that the latter is impossible note that the sequence consisting of the single term $a_1$ is symmetric and hence Lemma 6.5 implies

$$[2, \overbrace{2, \dots, 2}^{2n}, 2, \dots] \geq [\overbrace{2, \dots, 2}^{2n}, \beta_2]$$

or equivalently $\beta_2 \leq [2, 2, \dots]$. The functions $f$ and $g$ are decreasing and increasing, respectively, and therefore

$$M(\mathcal{A}) - M(\mathcal{B}) > f(12/7) + g(7/5) - f(7/3) - g(17/7).$$

We shall prove the lemma by obtaining a lower bound for the right hand side.

As usual, we know from the definitions of $f$ and $g$ as continued fractions that

$$f(x) = \frac{cx + d}{ax + b} \qquad \text{and} \qquad g(x) = \frac{ax + b}{(a + c)x + (b + d)}$$

where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \overbrace{\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \cdots\cdots \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}}^{2n}.$$

Note that $a/c = \overbrace{[2, \dots, 2]}^{2n}$ and $d/b = [0, \overbrace{2, \dots, 2}^{2n-1}]$ and $b = c = K(\overbrace{2, \dots, 2}^{2n-1})$. Using $ad - bc = (-1)^{2n} = 1$ it is not hard to verify that

$$f(12/7) - f(7/3) = \frac{-12/7 + 7/3}{(12a/7 + b)(7a/3 + b)}$$

and

$$g(7/5) - g(17/7) = \frac{7/5 - 17/7}{(7(a+c)/5 + (b+d))(17(a+c)/7 + (b+d))}.$$

We also know $a/b = d/b + 2$ and $c/b = 1$ and hence can deduce that

$$f(12/7) - f(7/3) = \frac{13/21}{(12/7(d/b) + 31/7)(7/3(d/b) + 17/3)}\, b^2$$

and

$$g(7/5) - g(17/7) = \frac{36/35}{(24/7(d/b) + 58/7)(12/5(d/b) + 26/5)}\, b^2.$$

It follows that

$$f(12/7) - f(7/3) + g(7/5) - g(17/7) = h(d/b)\, K(\overbrace{2, \ldots, 2}^{2n-1})$$

where

$$h(x) = \frac{13/21}{(12x/7 + 31/7)(7x/3 + 17/3)} - \frac{36/35}{(24x/7 + 58/7)(12x/5 + 26/5)}.$$

To complete the proof it suffices to show $h(d/b) > 0.0009$.

Rearranging the expression for $h$ we have

$$h(x) = \frac{48x^2/49 + 668x/245 + 632/735}{(12x/7 + 31/7)(7x/3 + 17/3)(24x/7 + 58/7)(12x/5 + 26/5)}. \tag{6.12}$$

It is evident from the continued fraction expansion of $d/b$ that $d/b > [0, 2, 2] = 2/5$ and $d/b < [0, 2] = 1/2$. Hence a crude lower bound for $h(d/b)$ may be found by substituting $x = 2/5$ in the numerator of (6.12) and $x = 1/2$ in the denominator. In this manner we find $h(d/b) > 0.0009$ as claimed. □

**Theorem 6.2.** *If* $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-2n}, \ldots, a_{-1}, a_0}\}$ *where* $n \geq 1$ *and*

$$a_{-2n}, \ldots, a_{-1}, a_0 = 1, \overbrace{2, \ldots, 2}^{2n}$$

*then* $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ *and* $M(\mathcal{A})$ *is an isolated point of the Markoff spectrum.*

*Proof.* Let $\mathcal{A}$ be as described. Using Lemma 6.5 and the periodicity and symmetry of $\mathcal{A}$ (or otherwise), it is not hard to verify that $\lambda_0(\mathcal{A}) \geq \lambda_i(\mathcal{A})$ for $-2n \leq i \leq 0$.

Hence $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. Because $\mathcal{A}$ is a periodic sequence of 1's and 2's and $a_{-(2n-1)-i} = a_i$ for all integers $i$, Remark 6.2 implies that the function $M$ has a locally isolated value at $\mathcal{A}$. It follows from Remark 6.3 that if $M(\mathcal{A})$ is not isolated in the spectrum then there is a sequence $\mathcal{B}$ other than $\mathcal{A}$ and its reverse such that $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. We shall prove the theorem by showing no such sequence $\mathcal{B}$ exists.

Suppose $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is a sequence other than $\mathcal{A}$ and its reverse which satisfies $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$. Clearly, we can assume $\mathcal{B}$ consists solely of 1's and 2's. Further, since $a_{-1}, a_0, a_1 = 2, 2, 1$, Remark 6.4 implies that either $b_{-1}, b_0, b_1 = 2, 2, 1$ or $b_{-1}, b_0, b_1 = 1, 2, 2$. By replacing $\mathcal{B}$ by its reverse if necessary we can assume the former is true. We shall divide the possibilities for $\mathcal{B}$ into several cases. Before we begin, observe that

$$\mathcal{A} = \ldots\ldots, 1, \overbrace{2, \ldots, 2}^{2n-1}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n}, 1, \ldots\ldots$$

where the asterisk marks the term $a_0$.

Now, suppose $\mathcal{B}$ is of the form

$$\mathcal{B} = \ldots\ldots, \overbrace{2, \ldots, 2}^{j+1}, 2^*, 1, \overbrace{2, \ldots, 2}^{j}, 1, \ldots\ldots$$

where $0 \leq j \leq 2n-2$ and the asterisk marks the term $b_0$ in $\mathcal{B}$. The sequence consisting of the single term $b_1$ is symmetric and so Lemma 6.5 applied to $\mathcal{B}$ shows

$$[2, \overbrace{2, \ldots, 2}^{j+1}, b_{-(j+2)}, b_{-(j+3)}, b_{-(j+4)}, \ldots] \geq [\overbrace{2, \ldots, 2}^{j}, 1, b_{j+3}, b_{j+4}, b_{j+5}, \ldots].$$

Consequently $j$ is even. Also, $b_i = a_i$ for $-(j+1) \leq i \leq j+1$ and since $(-1)^{j+1}(b_{j+2} - a_{j+2}) > 0$ and

$$K(1, \overbrace{2, \ldots, 2}^{j}, 2) = K(\overbrace{2, \ldots, 2}^{j+1}, 1)$$

Lemma 6.3 (applied with $\mathcal{A}$ and $\mathcal{B}$ interchanged) implies $M(\mathcal{A}) > M(\mathcal{B})$. This contradicts our assumption that $M(\mathcal{B}) = M(\mathcal{A})$.

Next, suppose

$$\mathcal{B} = \ldots\ldots, 1, \overbrace{2, \ldots, 2}^{j}, 2^*, 1, \overbrace{2, \ldots, 2}^{j}, \ldots\ldots$$

where $1 \leq j \leq 2n-2$. In this case, $j$ is odd else Lemma 6.7 implies $M(\mathcal{A}) > M(\mathcal{B})$. As above, Lemma 6.5 implies

$$[2, \overbrace{2, \ldots, 2}^{j}, 1, b_{-(j+2)}, b_{-(j+3)}, b_{-(j+4)}, \ldots] \geq [\overbrace{2, \ldots, 2}^{j}, b_{j+2}, b_{j+3}, b_{j+4}, \ldots]$$

and hence $b_{j+2} = 2$. It follows that $b_i = a_i$ for $-j \leq i \leq j+2$. Since $(-1)^j(b_{-(j+1)} - a_{-(j+1)}) > 0$ and

$$K(\overbrace{2, \ldots, 2}^{j}, 2) < K(1, \overbrace{2, \ldots, 2}^{j}, 2, 1)$$

Lemma 6.3 (applied with $\mathcal{A}$ and $\mathcal{B}$ interchanged and replaced by their reverses) implies $M(\mathcal{A}) > M(\mathcal{B})$. Again we have a contradiction.

The only other possibility is that $\mathcal{B}$ is of the form

$$\mathcal{B} = \ldots\ldots, \overbrace{2, \ldots, 2}^{2n-1}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n-1}, \ldots\ldots$$

As usual, we can apply Lemma 6.5. Thus

(6.13)
$$[2, \overbrace{2, \ldots, 2}^{2n-1}, b_{-2n}, b_{-(2n+1)}, b_{-(2n+2)}, \ldots] \geq [\overbrace{2, \ldots, 2}^{2n-1}, b_{2n+1}, b_{2n+2}, b_{2n+3}, \ldots]$$

and so $b_{2n+1} = 2$. It follows that $b_{-2n} = 1$ because if not $(-1)^{2n-1}(a_{-2n} - b_{-2n}) > 0$ in which case, since $a_i = b_i$ for $-(2n-1) \leq i \leq 2n+1$ and

$$K(\overbrace{2, \ldots, 2}^{2n-1}, 2) < K(1, \overbrace{2, \ldots, 2}^{2n-1}, 2, 1),$$

Lemma 6.3 (applied with $\mathcal{A}$ and $\mathcal{B}$ replaced by their reverses) implies $M(\mathcal{B}) > M(\mathcal{A})$. Further, we can now deduce from the inequality (6.13) that $b_{2n+2} = 1$.

We know that $\mathcal{A} = \{\overline{a_{-(2n-1)}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n+2}}\}$ and $M(\mathcal{A}) = \lambda_0(\mathcal{A})$,

$$a_{-(2n-2)}, \ldots, a_{-2}, a_{-1} = \overbrace{2, \ldots, 2}^{2n-2}, \qquad a_1, a_2, \ldots, a_{2n+2} = 1, \overbrace{2, \ldots, 2}^{2n}, 1$$

and $a_{-(2n-1)} = a_0$. Therefore $\mathcal{A}$ satisfies the hypothesis of Lemma 6.6. We are assuming $\mathcal{B}$ is a sequence of 1's and 2' other than $\mathcal{A}$ with $M(\mathcal{B}) = \lambda_0(\mathcal{B})$ and we have shown that $b_i = a_i$ for $-(2n-2) \leq i \leq 2n+2$. It follows from Lemma 6.6 that $M(\mathcal{B}) \neq M(\mathcal{A})$. This contradicts our assumption that $M(\mathcal{B}) = M(\mathcal{A})$ and the proof is complete. □

**Theorem 6.3.** *If* $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty} = \{\overline{a_{-(4n+3)}, \ldots, a_{-1}, a_0}\}$ *where* $n \geq 1$ *and*

$$a_{-(4n+3)}, \ldots, a_{-1}, a_0 = 1, \overbrace{2, \ldots, 2}^{2n}, 1, \overbrace{2, \ldots, 2}^{2n+2}$$

*then* $M(\mathcal{A}) = \lambda_0(\mathcal{A})$ *and* $M(\mathcal{A})$ *is an isolated point of the Markoff spectrum.*

*Proof.* The proof of this theorem is very similar to that of Theorem 6.2 and we shall refer to that proof for certain details. As was done there, we shall use an asterisk to mark terms which have index zero.

Let $\mathcal{A}$ be as described. Since the sequence consisting of the single term $a_1$ is symmetric and

$$[\overbrace{2, \ldots, 2}^{2n+2}, \ldots] \geq [\overbrace{2, \ldots, 2}^{2n}, 1, \ldots]$$

Lemma 6.5 implies $\lambda_0(\mathcal{A}) > \lambda_2(\mathcal{A})$. By noting the periodicity and symmetry of $\mathcal{A}$ and using Lemma 6.5 in a similar manner it is not hard to verify that one of $\lambda_0(\mathcal{A}) > \lambda_i(\mathcal{A})$ or $\lambda_2(\mathcal{A}) > \lambda_i(\mathcal{A})$ is true for all $-(4n+3) \leq i \leq 0$ and hence $M(\mathcal{A}) = \lambda_0(\mathcal{A})$. Also, $\mathcal{A}$ is a periodic sequence of 1's and 2's and $a_{-(2n+1)-i} = a_i$ for all integers $i$ and so again Remark 6.2 implies the function $M$ has a locally isolated value at $\mathcal{A}$. As before, to prove the theorem it suffices to show there is no sequence $\mathcal{B}$ other than $\mathcal{A}$ and its reverse such that $M(\mathcal{B}) = \lambda_0(\mathcal{B}) = M(\mathcal{A})$.

Suppose $\mathcal{B} = \{b_i\}_{i=-\infty}^{+\infty}$ is such a sequence. Since $\mathcal{A}$ is of the form

$$\mathcal{A} = \ldots\ldots, 1, \overbrace{2, \ldots, 2}^{2n-1}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n}, 1, \ldots\ldots,$$

the arguments in this first part of the proof of Theorem 6.2 apply *verbatim.* Thus $\mathcal{B}$ consists solely of 1's and 2's and, by replacing $\mathcal{B}$ by its reverse if necessary, we can assume $b_{-1}, b_0, b_1 = 2, 2, 1$. Similarly, we know that $\mathcal{B}$ is neither of the form

$$\mathcal{B} = \ldots\ldots, \overbrace{2, \ldots, 2}^{j+1}, 2^*, 1, \overbrace{2, \ldots, 2}^{j}, 1, \ldots\ldots$$

where $0 \leq j \leq 2n-1$ nor of the form

$$\mathcal{B} = \ldots\ldots, 1, \overbrace{2, \ldots, 2}^{j}, 2^*, 1, \overbrace{2, \ldots, 2}^{j}, \ldots\ldots$$

where $1 \leq j \leq 2n$. (Note the range of the index $j$ in each case.) It follows that $\mathcal{B}$ is of the form

$$\mathcal{B} = \ldots\ldots, \overbrace{2, \ldots, 2}^{2n+1}, 2^*, 1, \overbrace{2, \ldots, 2}^{2n}, \ldots\ldots$$

Further, $b_{2n+2} = 1$ because if not $(-1)^{2n+1}(a_{2n+2} - b_{2n+2}) > 0$ in which case, since $a_i = b_i$ for $-(2n+1) \leq i \leq 2n+1$ and

$$K(1, \overbrace{2, \ldots, 2}^{2n}, 2) = K(\overbrace{2, \ldots, 2}^{2n+1}, 1),$$

Lemma 6.3 implies $M(\mathcal{B}) > M(\mathcal{A})$.

Clearly $\mathcal{A} = \{\overline{a_{-(2n+1)}, a_{-2n}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{2n+2}}\}$ and $M(\mathcal{A}) = \lambda_0(\mathcal{A})$,

$$a_{-2n}, \ldots, a_{-2}, a_{-1} = \overbrace{2, \ldots, 2}^{2n}, \qquad a_1, a_2, \ldots, a_{2n+2} = 1, \overbrace{2, \ldots, 2}^{2n}, 1$$

and $a_{-(2n+1)} = a_0$. Again, $\mathcal{A}$ satisfies the hypothesis of Lemma 6.6 and hence $M(\mathcal{B}) \neq M(\mathcal{A})$ contradicting our initial assumption about $\mathcal{B}$. The proof is complete. □

APPENDIX

# SOME SIMPLIFICATIONS TO MARKOFF'S THEORY

Markoff's original work,[29] and [30], has been refined by several authors. See for instance the books by Dickson, [17], and Cusick and Flahive, [14]. Here, we present another improvement to some of the theory. Specifically, we shall give a shortened proof of a theorem which is equivalent to Theorem 64 in [17] and Theorem 3 in Chapter 1 of [14]. Having done this, it will also be appropriate to make some comments on the connection between the cutting sequences of lines in the plane and the sequences of positive integers associated with the Markoff forms.

We shall use the notation introduced in the section of Chapter 2 on doubly infinite sequences of positive integers. Recall in particular that if $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ is a sequence of positive integers then we write $\overline{\mathcal{A}} = \{a_{-i}\}_{i=-\infty}^{+\infty}$ and note that $\lambda_i(\overline{\mathcal{A}}) = \lambda_{-i}(\mathcal{A})$ for all $i$ so that $M(\overline{\mathcal{A}}) = M(\mathcal{A})$. We shall also need the following two lemmas. The first is well-known. The second is easy to verify and may be found in any of the references mentioned.

**Lemma A.1.** *Let $\alpha = [a_0, a_1, a_2, \ldots]$ and $\beta = [b_0, b_1, b_2, \ldots]$. Then $\alpha > \beta$ if and only if there is $n \geq 0$ such that $a_i = b_i$ for $0 \leq i \leq n-1$ and $(-1)^n(a_n - b_n) > 0$.*

**Lemma A.2.** *Let $\alpha$ and $\beta$ be any real numbers. Then*

$$\alpha \leq \beta \iff [2, 2, \alpha] + [0, 1, 1, \beta] \leq 3$$

*with equality on the left if and only if there is equality on the right.*

We can state and prove the theorem referred to above.

**Theorem A.1.** *Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a sequence of positive integers. Then $M(\mathcal{A}) \leq 3$ if and only if $\mathcal{A}$ is $(1,1)^\infty$ or $(1,1)^\infty, 2, 2, (1,1)^\infty$ or $\mathcal{A}$ is of the form*

$$\text{(A.1)} \qquad \ldots\ldots, 2, 2, (1,1)^{r(-1)}, 2, 2, (1,1)^{r(0)}, 2, 2, (1,1)^{r(-1)}, 2, 2, \ldots\ldots$$

*where* $\{r(i)\}_{i=-\infty}^{+\infty}$ *is a sequence of non-negative integers for which*

(a) $|r(i+1)-r(i)| \leq 1$ *for all* $i$,

(b) *if* $r(i+1)-r(i)$ *is* $+1$ *or* $-1$, *respectively, for some* $i$ *then either all the differences* $r(i+1+j)-r(i-j)$ *where* $j = 1,2,3,\ldots$ *are zero or the first which is non-zero is negative or positive, respectively.*

*Proof.* Suppose $M(\mathcal{A}) \leq 3$. Since $M(\mathcal{A}) \geq \lambda_i(\mathcal{A}) > a_i$ for all $i$, we know that each $a_i$ is 1 or 2. Further, there is no $i$ such that $a_{i-1}, a_i, a_{i+1} = 1,2,1$ else

$$M(\mathcal{A}) \geq \lambda_i(\mathcal{A}) = \xi_i(\mathcal{A}) - \eta_i(\mathcal{A}) > [2,1,1] + [0,1,1] = 3,$$

and similarly, there is no $i$ such that $a_i, a_{i+1}, a_{i+2} = 2,1,2$ else

$$M(\mathcal{A}) \geq \lambda_i(\mathcal{A}) = \xi_i(\mathcal{A}) - \eta_i(\mathcal{A}) > [2,1,2] + [0,2,1] = 3.$$

It follows that we can write $\mathcal{A}$ in the form

$$\ldots\ldots, \overbrace{2,\ldots,2}^{s(-1)}, \overbrace{1,\ldots,1}^{t(-1)}, \overbrace{2,\ldots,2}^{s(0)}, \overbrace{1,\ldots,1}^{t(1)}, \overbrace{2,\ldots,2}^{s(1)}, \overbrace{1,\ldots,1}^{t(1)}, \ldots\ldots$$

where each $s(j)$ and each $t(j)$ is an integer greater than or equal to 2. Note that it is possible that the sequence $\ldots, s(-1), t(-1), s(0), t(0), s(1), t(1), \ldots$ terminates in one or both directions with $\infty$.

We claim that each $s(j)$ is even (or $\infty$). To see this suppose not and let $s \geq 3$ be the smallest odd integer which occurs among the $s(j)$. Clearly, there is some $i$ such that

$$a_{i-2}, a_{i-1}, a_i, \ldots = 1,1,2,2,\overbrace{2,\ldots,2}^{s-2},1,1,\ldots.$$

By Lemma A.2, we know $\lambda_i(\mathcal{A}) \leq 3$ if and only if

$$[\overbrace{2,\ldots,2}^{s-2},1,1,\ldots] \leq [a_{i-3}, a_{i-4}, a_{i-5}, \ldots].$$

Since $s-2$ is odd, it can be deduced from Lemma A.1 that there is some odd $s' \leq s-2$ such that

$$a_{i-3}, a_{i-4}, a_{i-5}, \ldots = \overbrace{2,\ldots,2}^{s'},1,1,\ldots.$$

However, this contradicts the minimality of $s$ and the claim is true.

Similarly, we claim that each $t(j)$ is even (or $\infty$). Again suppose not and let $t \geq 3$ be the smallest odd integer which occurs among the $t(j)$ and let $i$ be the index for which

$$a_{i-1}, a_i, a_{i+1}, \ldots = 2, 2, 1, 1, \overbrace{1, \ldots, 1}^{t-2}, 2, 2, \ldots.$$

An application of Lemma A.2 to $\overline{\mathcal{A}}$ (the reverse of $\mathcal{A}$) shows that $\lambda_i(\mathcal{A}) \leq 3$ if and only if

$$[a_{i+2}, a_{i+3}, a_{i+4}, \ldots] \leq [\overbrace{1, \ldots, 1}^{t-2}, 2, 2, \ldots].$$

Again, since $t-2$ is odd, Lemma A.1 implies that there is some odd $t' \leq t-2$ such that

$$a_{i+2}, a_{i+3}, a_{i+4}, \ldots = \overbrace{1, \ldots, 1}^{t'}, 2, 2, \ldots$$

and the minimality of $t$ is contradicted.

If $(1,1)^\infty$ occurs in $\mathcal{A}$ and $\mathcal{A} \neq (1,1)^\infty$ then, by reversing $\mathcal{A}$ if necessary, we can assume that for some $i$

$$a_{i-1}, a_i, a_{i+1}, \ldots = 2, 2, 1, 1, 1, 1, \ldots.$$

Applying Lemma A.2 to $\overline{\mathcal{A}}$ shows that $\lambda_i(\mathcal{A}) \leq 3$ if and only if

$$[a_{i-3}, a_{i-4}, a_{i-5}, \ldots] \leq [1, 1, 1, 1, \ldots].$$

Since each $t(j)$ is even, this last relation is an equality and $\mathcal{A} = (1,1)^\infty, 2, 2, (1,1)^\infty$.

It follows that if $\mathcal{A}$ is not $(1,1)^\infty$ or $(1,1)^\infty, 2, 2, (1,1)^\infty$ then $(1,1)^\infty$ does not occur in $\mathcal{A}$ in which case, we can express $\mathcal{A}$ in the form (A.1) where $\{r(i)\}_{i=-\infty}^{+\infty}$ is a doubly infinite sequence of non-negative integers. If (a) is not true then, by reversing $\mathcal{A}$ if necessary, we can assume there is some $j$ such that $r(j+1) \leq r(j)-2$ and hence that there is some $i$ such that

$$a_i, a_{i+1}, a_{i+2}, \ldots = 2, 2, (1,1)^{r(j+1)}, 2, 2, \ldots$$

and

$$a_{i-1}, a_{i-2}, a_{i-3}, \ldots = 1, 1, (1,1)^{r(j+1)}, 1, 1, \ldots.$$

However, we are assuming $\lambda_i(\mathcal{A}) \leq 3$ and Lemma A.2 implies

$$[(1,1)^{r(j+1)}, 2, 2, \ldots] \leq [(1,1)^{r(j+1)}, 1, 1, \ldots]$$

which contradicts Lemma A.1.

Similarly, if (b) is not true then, by reversing $\mathcal{A}$ if necessary, we can assume there is some $j$ and some $k \geq 1$ such that $r(j+1) = r(j) - 1$ and $r(j+1+k) < r(j-k)$ and $r(j+1+l) = r(j-l)$ for $l = 1, 2, \ldots, k-1$ and hence that there is some $i$ such that $a_i, a_{i+1}, a_{i+2}, \ldots$ is the sequence

$$2, 2, (1,1)^{r(j+1)}, 2, 2, (1,1)^{r(j+2)}, \ldots, 2, 2, (1,1)^{r(j+1+k)}, 2, 2, \ldots\ldots$$

and $a_{i-1}, a_{i-2}, a_{i-3}, \ldots$ is the sequence

$$1, 1, (1,1)^{r(j+1)}, 2, 2, (1,1)^{r(j+2)}, \ldots, 2, 2, (1,1)^{r(j+1+k)}, 1, 1, \ldots\ldots.$$

Again, since $\lambda_i(\mathcal{A}) \leq 3$, Lemma A.2 implies

$$\begin{aligned}[(1,1)^{r(j+1)}, 2, 2, (1,1)^{r(j+2)}, \ldots, 2, 2, (1,1)^{r(j+1+k)}, 2, 2, \ldots\ldots] \\ \leq [(1,1)^{r(j+1)}, 2, 2, (1,1)^{r(j+2)}, \ldots, 2, 2, (1,1)^{r(j+1+k)}, 1, 1, \ldots\ldots].\end{aligned}$$

which contradicts Lemma A.1. The proof of the forward implication is complete.

To prove the reverse implication we assume $\mathcal{A}$ is $(1,1)^\infty$ or $(1,1)^\infty, 2, 2, (1,1)^\infty$ or of the form (A.1) where $\{r(i)\}_{i=-\infty}^{+\infty}$ is a sequence of non-negative integers satisfying (a) and (b) and we prove that $M(\mathcal{A}) \leq 3$ by showing $\lambda_i(\mathcal{A}) \leq 3$ for all $i$.

If $a_i = 1$ then $\lambda_i(\mathcal{A}) < [1,1] + [0,1] = 3$ and if $a_{i-1}, a_i, a_{i+1} = 2, 2, 2$ then $\lambda_i(\mathcal{A}) < [2,2] + [0,2] = 3$. Since there are no isolated 2's in $\mathcal{A}$ the only other possibilities are $a_{i-1}, a_i, a_{i+1} = 1, 2, 2$ or $a_{i-1}, a_i, a_{i+1} = 2, 2, 1$. By reversing $\mathcal{A}$, if necessary, we can assume that the former is true. In this case, either $\mathcal{A} = (1,1)^\infty, 2, 2, (1,1)^\infty$ and Lemma A.2 shows that $\lambda_i(\mathcal{A}) = 3$ or $\mathcal{A}$ is of the form (A.1) and we can choose an integer $j$ so that $r(j) \geq 1$ and

$$\begin{aligned}\lambda_i(\mathcal{A}) = &[2, 2, (1,1)^{r(j+1)}, 2, 2, (1,1)^{r(j+2)}, 2, 2, (1,1)^{r(j+3)}, 2, 2, \ldots\ldots] \\ &+ [1, 1, (1,1)^{r(j)-1}, 2, 2, (1,1)^{r(j-1)}, 2, 2, (1,1)^{r(j-2)}, 2, 2, \ldots\ldots].\end{aligned}$$

In the latter situation it suffices, by Lemma A.2, to show that

$$[(1,1)^{r(j+1)},2,2,(1,1)^{r(j+2)},2,2,(1,1)^{r(j+3)},2,2,\ldots\ldots]$$
$$\leq [(1,1)^{r(j)-1},2,2,(1,1)^{r(j-1)},2,2,(1,1)^{r(j-2)},2,2,\ldots\ldots].$$

That this is true is an easy consequence of Lemma A.1 and the fact that $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the properties (a) and (b). We leave the details to the reader. □

**Remark A.1.** Although our statement of Theorem A.1 is asymmetrical with respect to the terms 1, 1 and 2, 2 this is for convenience only. It is not hard to show that if $\mathcal{A}$ is of the form (A.1) where $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the conditions described. and if $\mathcal{A}$ is not $(2,2)^\infty$ or $(2,2)^\infty,1,1,(2,2)^\infty$ then $\mathcal{A}$ can also be written in the form

$$\ldots\ldots,1,1,(2,2)^{r'(-1)},1,1,(2,2)^{r'(0)},1,1,(2,2)^{r'(-1)},1,1,\ldots\ldots$$

where $\{r'(i)\}_{i=-\infty}^{+\infty}$ is a sequence of non-negative integers which satisfies the same conditions as $\{r(i)\}_{i=-\infty}^{+\infty}$. Moreover, either $\{r(i)\}_{i=-\infty}^{+\infty}$ consists of positive integers and $\{r'(i)\}_{i=-\infty}^{+\infty}$ is

$$\ldots\ldots 1,\overbrace{0,\ldots,0}^{r(i-1)-1},1,\overbrace{0,\ldots,0}^{r(i)-1},1,\overbrace{0,\ldots,0}^{r(i+1)-1},1,\ldots\ldots$$

or this is true with the roles of $\{r(i)\}_{i=-\infty}^{+\infty}$ and $\{r'(i)\}_{i=-\infty}^{+\infty}$ reversed.

**Remark A.2.** Theorem A.1 with conditions (a) and (b) replaced by

(a)′ $|r(j)-r(i)|\leq 1$ for all $i$ and $j$,

(b)′ if $r(i+1)-r(i)$ is $+1$ or $-1$, respectively, for some $i$ then either all the differences $r(i+1+j)-r(i-j)$ where $j=1,2,3,\ldots$ are zero or the first which is non-zero is $-1$ or $+1$, respectively.

is essentially Theorem 3 in Chapter 1 of Cusick and Flahive's book, [14]. The only difference being that we also allow sequences $\mathcal{A}$ with $\lambda_i(\mathcal{A})=3$ for some $i$. We claim that the two pairs of conditions are equivalent. Clearly, (a)′ and (b)′ imply (a) and (b). Likewise, (a)′ and (b) imply and (b)′. Therefore to prove the claim we

need only demonstrate that (a) and (b) imply (a)′. Suppose not, that is, suppose (a) and (b) are true and (a)′ is not. Then there exist integers $i < j$ such that $|r(j) - r(i)| \geq 2$. Choose $i$ and $j$ so that in addition $j - i$ is minimal. Note that (a) implies $j - i \geq 2$. By reversing $\mathcal{A}$, if necessary, we can assume $r(j) \geq r(i) + 2$. Since $j - i$ is minimal we know

$$r(k) = r(i) + 1, \qquad k = i + 1, i + 2, \ldots, j - 1$$

and thus (a) implies $r(j) = r(i) + 2$. Since $r(j) = r(i) + 2$, the minimality of $j - i$ also implies

$$r(j + l) \geq r(i) + 1, \qquad l = 1, 2, \ldots, j - i - 1$$

and hence

$$r(j + l) \geq r(j - 1 - l), \qquad l = 1, 2, \ldots, j - i - 2$$

and $r(j + l) > r(j - 1 - l)$ when $l = j - i - 1$. It follows that the first of the differences

$$r(j + l) - r(j - 1 - l), \qquad l = 1, 2, \ldots, j - i - 1$$

which is not zero is positive. However, $r(j) - r(j - 1)$ is $+1$ and we have a contradiction of condition (b). We conclude that (a) and (b) imply (a)′ and the claim is true.

We shall complete this appendix by converting the characterisation of the integer sequences $\mathcal{A}$ with $M(\mathcal{A}) \leq 3$ given in Theorem A.1 into a similar characterisation of the cutting sequences of the corresponding geodesics on $\mathbf{T}$. We shall use the algorithm involving $LR$-sequences as described in Chapter 2 to do this. We remind the reader that we use the abbreviation

$$W^n = \overbrace{W \ldots W}^{n}$$

where $n \geq 0$ is an integer and $W$ is a word in the symbols $\{A, B, A^{-1}, B^{-1}\}$.

**Theorem A.2.** *A doubly infinite sequence* $\mathbf{S}$ *of* $A$*'s,* $B$*'s,* $A^{-1}$*'s and* $B^{-1}$*'s is the cutting sequence of a geodesic on* $\mathbf{T}$ *which corresponds to a sequence of positive*

*integers $\mathcal{A}$ with $M(\mathcal{A}) \leq 3$ if and only if there are $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$ such that $\mathbf{S} = Y^\infty$ or $\mathbf{S} = Y^\infty Z Y^\infty$ or $\mathbf{S}$ is of the form*

$$\text{(A.2)} \qquad \ldots\ldots ZY^{s(-1)}ZY^{s(0)}ZY^{s(1)}Z\ldots\ldots$$

*where $\{s(i)\}_{i=-\infty}^{+\infty}$ is a sequence of positive integers for which*

(a) $|s(i+1) - s(i)| \leq 1$ *for all* $i$,

(b) *if $s(i+1) - s(i)$ is $+1$ or $-1$, respectively, for some $i$ then either all the differences $s(i+1+j) - s(i-j)$ where $j = 1, 2, 3, \ldots$ are zero or the first which is non-zero is negative or positive, respectively.*

*Proof.* We prove the forward implication first. Let $\mathcal{A} = \{a_i\}_{i=-\infty}^{+\infty}$ be a sequence of positive integers with $M(\mathcal{A}) \leq 3$. Theorem A.1 describes the possibilities for $\mathcal{A}$. We suppose first that $\mathcal{A}$ is of the form (A.1) where $\{r(i)\}_{i=-\infty}^{+\infty}$ is as described. According to the section of Chapter 2 on $LR$-sequences we can produce the cutting sequence of a geodesic on $\mathbf{T}$ which corresponds to $\mathcal{A}$ by first forming the $LR$-sequence

$$\text{(A.3)} \qquad \ldots\ldots L^2R^2(LR)^{r(-1)}L^2R^2(LR)^{r(0)}L^2R^2(LR)^{r(1)}L^2R^2\ldots\ldots$$

We then form a pattern from Table 2.1 somewhere in (A.3) and extend it to the whole sequence using only patterns in Table 2.1. There are six possibilities. One of them is

$$\ldots B^{-1}R^2A(LRA)^{r(-1)}L^2B^{-1}R^2A(LRA)^{r(0)}L^2B^{-1}R^2A(LRA)^{r(1)}L^2B^{-1}\ldots$$

which yields the cutting sequence

$$\text{(A.4)} \qquad \ldots\ldots B^{-1}A^{r(-1)+1}B^{-1}A^{r(0)+1}B^{-1}A^{r(1)+1}B^{-1}\ldots\ldots$$

This is of the form (A.2) with $Y = A$ and $Z = B^{-1}$ and $s(i) = r(i) + 1$ for all $i$. It is not hard to deduce from the fact that $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the conditions (a) and (b) of Theorem A.1 that $\{s(i)\}_{i=-\infty}^{+\infty}$ satisfies the corresponding conditions of this theorem. Hence (A.4) is of the form described in this theorem.

We know that altogether there are 12 cutting sequences which correspond to $\mathcal{A}$. The others may be obtained from (A.4) by applying to it automorphisms which represent the cosets of $\Psi/\text{Inn}\,\Gamma'$. We take as a transversal the set

$$\{\text{Id}, S, S^2, R^2, R^2S, R^2S^2, P, PS, PS^2, PR^2, PR^2S, PR^2S^2\}.$$

Thus all cutting sequences which correspond to $\mathcal{A}$ may be obtained by applying the automorphisms Id, $P$, $R^2$ and $PR^2$ to (A.4) and its images under $S$ and $S^2$. The image of (A.4) under $S(A,B) = (B, B^{-1}A^{-1})$ is

$$\text{(A.5)} \qquad \ldots\ldots AB^{r(-1)+2}AB^{r(0)+2}AB^{r(1)+2}A\ldots\ldots$$

This is of the form (A.2) with $Y = B$ and $Z = A$ and $s(i) = r(i) + 2$ for all $i$. Again, it is easy to deduce that $\{s(i)\}_{i=-\infty}^{+\infty}$ satisfies the conditions (a) and (b) of this theorem from the fact that $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the corresponding conditions of Theorem A.1. Hence (A.5) is of the form described in this theorem.

A second application of $S$ to (A.4) yields the sequence

$$\text{(A.6)} \qquad \ldots(A^{-1}B^{-1})^{r(-1)+1}A^{-1}(A^{-1}B^{-1})^{r(0)+1}A^{-1}(A^{-1}B^{-1})^{r(1)+1}A^{-1}\ldots.$$

We can re-write (A.6) in the form

$$\text{(A.7)} \qquad \ldots\ldots B^{-1}(A^{-1})^{s(-1)}B^{-1}(A^{-1})^{s(0)}B^{-1}(A^{-1})^{s(1)}B^{-1}\ldots\ldots$$

where $\{s(i)\}_{i=-\infty}^{+\infty}$ is the sequence

$$\text{(A.8)} \qquad \ldots\ldots 2, \overbrace{1,\ldots,1}^{r(i-1)}, 2, \overbrace{1,\ldots,1}^{r(i)}, 2, \overbrace{1,\ldots,1}^{r(i+1)}, 2, \ldots\ldots.$$

Clearly condition (a) holds in this case. Suppose condition (b) does not. By reversing $\{s(i)\}_{i=-\infty}^{+\infty}$ and $\{r(i)\}_{i=-\infty}^{+\infty}$, if necessary, we can assume there are integers $i$ and $j \geq 1$ such that $s(i) = 1$ and $s(i+1) = 2$ and

$$s(i-l) = s(i+1+l), \qquad l = 1, 2, \ldots, j-1$$

and $s(i-j) = 1$ and $s(i+1+j) = 2$. Re-index $\{r(i)\}_{i=-\infty}^{+\infty}$ so that $r(i+1)$ corresponds to the block of 1's immediately following $s(i+1) = 2$ and let $k \geq 0$

be the number of indices $l$ with $i+1 < l < i+1+j$ and $s(l) = 2$. If $k = 0$ then $r(i) \geq 2 = r(i+1) + 2$ contradicting (a) in Theorem A.1 and if $k \geq 1$ then $r(i) = r(i+1) + 1$ and

$$r(i-l) = r(i+1+l), \qquad l = 1, 2, \ldots, k-1$$

and $r(i-k) \geq r(i+1+k) + 1$ contradicting (b) in Theorem A.1. Hence condition (b) also holds for $\{s(i)\}_{i=-\infty}^{+\infty}$ and (A.7) is of the required form.

Since $P(A,B) = (B,A)$ and $R^2(A,B) = (A^{-1}, B^{-1})$, it is trivial that the images of the sequences (A.4), (A.5) and (A.7) under $P$, $R^2$ and $PR^2$ are of the required form. Therefore to complete the proof of the forward implication we need only consider the sequences $(1,1)^\infty$ and $(1,1)^\infty, 2, 2, (1,1)^\infty$. Corresponding to these are the $LR$-sequences $(LR)^\infty$ and $(LR)^\infty L^2 R^2 (LR)^\infty$, respectively, and hence the cutting sequences $A^\infty$ and $A^\infty B^{-1} A^\infty$. The images of the latter under $S$ are $B^\infty$ and $B^\infty A B^\infty$ and a second application of $S$ yields $(B^{-1}A^{-1})^\infty$ and $(B^{-1}A^{-1})^\infty A^{-1} (B^{-1}A^{-1})^\infty$. Obviously all these sequences and their images under $P$, $R^2$ and $PR^2$ are of the required form.

To prove the reverse implication we let $\mathbf{S}$ be a doubly infinite sequence of $A$'s, $B$'s, $A^{-1}$'s and $B^{-1}$'s and we let $\mathcal{A}$ be the corresponding integer sequence. We suppose first that $\mathbf{S}$ is of the form (A.2) where $\{s(i)\}_{i=-\infty}^{+\infty}$ is as described and $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ and $Z \neq Y^{\pm 1}$. We shall use Theorem A.1 to show that $M(\mathcal{A}) \leq 3$. We can replace $\mathbf{S}$ by any of its images under $P$, $R^2$ and $PR^2$ without altering the sequence $\mathcal{A}$. Hence we can assume $Z = B^{-1}$. If also $Y = A$ then $\mathbf{S}$ is of the form (A.4) where $r(i) = s(i) - 1$ for all $i$. In this case, the corresponding $LR$-sequence is (A.3) and so $\mathcal{A}$ is of the form (A.1). It is not hard to verify that $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the same conditions as $\{s(i)\}_{i=-\infty}^{+\infty}$ (except that some of the $r(i)$'s may be zero). Therefore Theorem A.1 implies $M(\mathcal{A}) \leq 3$ and we are done. If $Y = A^{-1}$ we consider two subcases. The first subcase is that $s(i) \geq 2$ for all $i$. In this subcase, we replace $\mathbf{S}$ by its image under $PR^2$ so that it is of the form (A.5) where $r(i) = s(i) - 2$ for all $i$. The corresponding $LR$-sequence is (A.3) and $\mathcal{A}$ is of the form (A.1). Again, $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the same conditions as $\{s(i)\}_{i=-\infty}^{+\infty}$ and so Theorem A.1 implies $M(\mathcal{A}) \leq 3$ and we are done.

It remains to consider the situation where $Y = A^{-1}$ and $Z = B^{-1}$ and some $s(i)$ is 1. In Remark A.2 we reformulated the conditions (a) and (b) of Theorem A.1 to the conditions (a)$'$ and (b)$'$. Clearly this reformulation is equally applicable to the conditions which $\{s(i)\}_{i=-\infty}^{+\infty}$ satisfies. It follows from the condition corresponding to (a)$'$ that $\{s(i)\}_{i=-\infty}^{+\infty}$ is a sequence of 1's and 2's. If $\{s(i)\}_{i=-\infty}^{+\infty}$ is

$$\ldots, 1, 1, 1, \ldots \qquad \text{or} \qquad \ldots, 1, 1, 1, 2, 1, 1, 1, \ldots$$

then $\mathbf{S}$ is $(B^{-1}A^{-1})^{\infty}$ or $(B^{-1}A^{-1})^{\infty}A^{-1}(B^{-1}A^{-1})^{\infty}$, respectively, and so $\mathcal{A}$ is $(1,1)^{\infty}$ or $(1,1)^{\infty}, 2, 2, (1,1)^{\infty}$ and $M(\mathcal{A}) \leq 3$. The only other possibility is that more than one 2 occurs in $\{s(i)\}_{i=-\infty}^{+\infty}$. In this case, it can be deduced from condition (b) that every block of 1's in $\{s(i)\}_{i=-\infty}^{+\infty}$ is of finite length. Hence $\{s(i)\}_{i=-\infty}^{+\infty}$ is of the form (A.8) where $\{r(i)\}_{i=-\infty}^{+\infty}$ is a sequence of non-negative integers. We are assuming $\mathbf{S}$ is of the form (A.7) and therefore we can re-write $\mathbf{S}$ in the form (A.6). The corresponding $LR$-sequence is (A.3) and as usual $\mathcal{A}$ is of the form (A.1). To see that $M(\mathcal{A}) \leq 3$ we need only show that the sequence $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies the conditions (a) and (b) of Theorem A.1.

Suppose condition (a) of Theorem A.1 does not hold. By reversing $\{r(i)\}_{i=-\infty}^{+\infty}$ and $\{s(i)\}_{i=-\infty}^{+\infty}$, if necessary, we may assume there is some $i$ such that $r(i+1) \geq r(i)+2$. Re-index $\{s(i)\}_{i=-\infty}^{+\infty}$ so that $s(i) = 2$ and $r(i+1)$ corresponds to the block of 1's immediately following $s(i)$. Set $k = r(i) + 1$. Then $s(i) = 2$ and $s(i+1) = 1$ and

$$s(i-l) = s(i+1+l) = 1, \qquad l = 1, 2, \ldots, k-1$$

and $s(i-k) = 2$ and $s(i+1+k) = 1$. This contradicts the property (b) of this theorem. Now suppose condition (b) of Theorem A.1 does not hold. By reversing the sequences, if necessary, we may assume there are integers $i$ and $j \geq 1$ such that $r(i+1) = r(i) + 1$ and

$$r(i-l) = r(i+1+l), \qquad l = 1, 2, \ldots, j-1$$

and $r(i+1+j) \geq r(i-j) + 2$. Again, re-index $\{s(i)\}_{i=-\infty}^{+\infty}$ so that $s(i) = 2$ and $r(i+1)$ corresponds to the block of 1's immediately following $s(i) = 2$. This time

set

$$k = r(i) + 1 + r(i-1) + 1 + \cdots + r(i-j) + 1.$$

Then $s(i) = 2$ and $s(i+1) = 1$ and

$$s(i-l) = s(i+1+l), \qquad l = 1, 2, \ldots, k-1$$

and $s(i-k) = 2$ and $s(i+1+k) = 1$. Again this contradicts the property (b) of this theorem. We conclude, as required, that $\{r(i)\}_{i=-\infty}^{+\infty}$ satisfies both the conditions of Theorem A.1.

To complete the proof it remains to consider the situation where $\mathbf{S} = Y^\infty$ or $\mathbf{S} = Y^\infty Z Y^\infty$ for some $Y, Z \in \{A, B, A^{-1}, B^{-1}\}$ with $Z \neq Y^{\pm 1}$. In this case, by applying one of Id, $P$, $R^2$ or $PR^2$ we may assume that $\mathbf{S}$ is one of $A^\infty$ or $A^\infty B A^\infty$ or $A^\infty B^{-1} A^\infty$. The corresponding integer sequence $\mathcal{A}$ is $(1,1)^\infty$ or $(1,1)^\infty, 2, 2, (1,1)^\infty$ or $(1,1)^\infty, 2, 2, (1,1)^\infty$, respectively. In all cases Theorem A.1 implies $M(\mathcal{A}) \leq 3$ and the proof is complete. □

**Remark A.3.** We saw in the section of Chapter 2 on linear sequences that the integer sequences $\mathcal{A}$ with $M(\mathcal{A}) \leq 3$ correspond exactly to the geodesics on $\mathbf{T}$ with linear cutting sequences. It follows that Theorem A.2 provides a characterisation of linear sequences. There are many other characterisations. An example is Theorem 2.8. Further examples may be found in [27].

## BIBLIOGRAPHY


1. A.F. Beardon, J. Lehner and M. Sheingorn, *Closed geodesics on a Riemann surface with application to the Markoff spectrum*, Trans. Amer. Math. Soc. **295** (1986), 635–647.

2. A.F. Beardon, *The geometry of discrete groups*, Graduate Texts in Mathematics 91, Springer-Verlag, New York Heidelberg and Berlin, 1983.

3. J.S. Birman and C. Series, *An algorithm for simple curves on surfaces*, J. London Math. Soc. (2) **29** (1984), 331–342.

4. R.T. Bumby, *Structure of the Markoff spectrum below* $\sqrt{12}$, Acta Arith. **29** (1976), 299–307.

5. J.W.S. Cassels, *An introduction to diophantine approximation*, Cambridge University Press, Cambridge, 1965.

6. H. Cohn, *Approach to Markoff's minimal forms through modular functions*, Ann. of Math. (2) **61** (1955), 1–12.

7. H. Cohn, *Conformal mapping on Riemann surfaces*, McGraw-Hill Book Company, New York, 1967.

8. H. Cohn, *Representation of Markoff's binary quadratic forms by geodesics on a perforated torus*, Acta Arith. **18** (1971), 125–136.

9. H. Cohn, *Markoff forms and primitive words*, Math. Ann. **196** (1972), 8–22.

10. H. Cohn, *Some direct limits of primitive homotopy words and of Markoff geodesics*, Discontinuous groups and Riemann surfaces (L. Greenberg, ed.), Ann. of Math. Studies No. 79, Princeton Univ. Press, Princeton N.J., 1974, pp. 81–98.

11. H. Cohn, *Mathematical microcosm of geodesics, free groups and Markoff forms*, Classical and quantum models and arithmetic problems (D.V. Chudnovsky et al., eds.), Lecture Notes in Pure and Applied Mathematics 92, Dekker, New York, 1984, pp. 69–97.

12. D.J. Crisp and W. Moran, *Single self-intersection geodesics and the Markoff spectrum*, Number theory with an emphasis on the Markoff spectrum (A.D. Pollington and W. Moran, eds.), Lecture Notes in Pure and Applied Mathematics 147, Dekker, New York, 1993, pp. 83–93.

13. T.W. Cusick, *The largest gaps in the lower Markoff spectrum*, Duke Math. J. **41** (1974), 453–463.

14. T.W. Cusick and M.E. Flahive, *The Markoff and Lagrange spectra*, Mathematical Surveys and Monographs 30, American Mathematical Society, Providence R.I., 1989.

15. N. Davis and J.R. Kinney, *Quadratic irrationals in the lower Lagrange spectrum*, Canad. J. Math. **25** (1973), 578–584.

16. L.E. Dickson, *Introduction to the theory of numbers*, The University of Chicago Press, Chicago, 1929.

17. L.E. Dickson, *Studies in the theory of numbers*, The University of Chicago Press, Chicago, 1930.

18. L.R. Ford, *Automorphic functions*, McGraw-Hill Book Company, New York, 1929.

19. M.E. Gbur (now Flahive), *On the lower Markoff spectrum*, Monatsh. Math. **81** (1976), 95–107.

20. A. Haas, *The geometry of Markoff forms*, Number Theory New York 1984-1985 (D.V. Chudnovsky et al., eds.), Lecture Notes in Mathematics 1240, Springer-Verlag, Berlin and New York, 1987, pp. 135–144.

21. A. Haas, *Diophantine approximation on hyperbolic Riemann surfaces*, Acta Math. **156** (1986), 33–82.

22. A. Haas, *Diophantine approximation on hyperbolic orbifolds*, Duke Math. J. **56** (1988), 531–547.

23. M. Hall, Jr., *The Markoff spectrum*, Acta Arith. **18** (1971), 387–399.

24. J. Lehner, *Discontinuous groups and automorphic functions*, Mathematical Surveys and Monographs 8, American Mathematical Society, Providence R.I., 1964.

25. J. Lehner and M. Sheingorn, *Simple closed geodesics on $H^+/\Gamma(3)$ arise from the Markov spectrum*, Bull. Amer. Math. Soc. (N.S.) **11** (1984), 359–362.

26. C.G. Lekkerkerker, *Geometry of numbers*, Wolters-Noordhoff Publishing, Groningen and North-Holland Publishing Company, Amsterdam, 1969.

27. W.F. Lunnon and P.A.B. Pleasants, *Characterization of two-distance sequences*, J. Austral. Math. Soc. (Series A) **53** (1992), 198–218.

28. W. Magnus, *Noneuclidean tesselations and their groups*, Academic Press, New York and London, 1974.

29. A.A. Markoff, *Sur les formes quadratiques binaires indéfines*, Math. Ann. **15** (1879), 381–406.

30. A.A. Markoff, *Sur les formes quadratiques binaires indéfines II*, Math. Ann. **17** (1880), 379–400.

31. R.A. Rankin, *Modular forms and functions*, Cambridge University Press, Cambridge and New York and Melbourne, 1977.

32. A.L. Schmidt, *Minimum of quadratic forms with respect to Fuchsian groups. I*, J. Reine Angew. Math. **286/287** (1976), 341–368.

33. A.L. Schmidt, *Minimum of quadratic forms with respect to Fuchsian groups. II*, J. Reine Angew. Math. **292** (1977), 109–114.

34. C. Series, *The infinite word problem and limit sets in Fuchsian groups*, Ergod. Th. & Dynam. Sys. **1** (1981), 337–360.

35. C. Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), 69–80.

36. C. Series, *The geometry of Markoff numbers*, Mathematical Intelligencer **7** (1985), 20–29.

37. C. Series, *The Markoff spectrum in the Hecke group* $G_5$, Proc. London Math. Soc. (3) **57** (1988), 151–181.

38. C. Series, *Symbolic dynamics and diophantine equations*, Number theory and dynamical systems (M.M. Dodson et al., eds.), London Mathematical Society Lecture Note Series 134, Cambridge University Press, Cambridge and New York, 1989, pp. 49–67.

39. C. Series, *Geometrical methods of symbolic coding*, Ergodic theory, symbolic dynamics and hyperbolic spaces (T. Bedford, M. Keane and C. Series, eds.), Oxford University Press, Oxford, 1991, pp. 125–151.

40. M. Sheingorn, *Characterization of simple closed geodesics on Fricke surfaces*, Duke Math. J. **52** (1985), 535–545.

41. T.H. Southard, *Weierstrass elliptic and related functions*, Handbook of mathematical functions (M. Abramowitz and I.A. Stegun, eds.), Dover Publications, New York, 1972 reprint of 1965 first edition, pp. 627–683.