Universiteit Utrecht

Markoff Theory — A Geometric Approach

by Barbara Harzevoort
under the supervision of prof. dr. Frits Beukers

Acknowledgements

This thesis has been written in order to conclude my master's program 'Mathematical Sciences' at the Utrecht University. There are some people who contributed to this thesis, in various ways.

First of all, I would like to thank my supervisor Frits Beukers. My thesis greatly benefited from his guidance over the past few months. I sincerely appreciate the time we spent figuring out how exactly things work, and all the helpful suggestions he supplied to smoothen my proofs. Furthermore I would like to thank Karma Dajani for proof-reading my work as a second examiner.

I especially would like to thank Jeroen Goudsmit. Not only was he always willing to proof-read my work, help me with LaTeX and rehearse my presentation with me. Most importantly, he was there to give me mental support whenever I needed it. Throughout my time as a student my parents never failed to stand by me, thank you for always being there.

Contents

Acknowledgements iii

Introduction vii

I. Continued Fractions 1
- I.1. Continued Fractions . . . 1
- I.2. Convergence . . . 2
- I.3. Approximation . . . 5
- I.4. Comparing Continued Fractions . . . 8
- I.5. Periodic Continued Fractions . . . 8

II. Markoff Theory 13
- II.1. The Markoff Spectrum . . . 13
  - II.1.1. Connecting the Two Definitions of $\mathbb{M}$ . . . 14
- II.2. Doubly Infinite Sequences . . . 16
- II.3. The Lagrange Spectrum . . . 24
  - II.3.1. The Spectra Compared . . . 25
- II.4. Markoff Numbers and Forms . . . 26
  - II.4.1. Markoff Forms . . . 28

III. Cutting Sequences 31
- III.1. Square Tessellation of the Euclidean Plane . . . 31
  - III.1.1. Cutting Sequences . . . 32
- III.2. Farey Tessellation of the Hyperbolic Plane . . . 36
  - III.2.1. The Upper Half-Plane . . . 36
  - III.2.2. The Farey Tessellation . . . 37
  - III.2.3. Cutting Sequences . . . 39
- III.3. A,B-Tessellation of the Hyperbolic Plane . . . 41
  - III.3.1. Cutting Sequences . . . 42
  - III.3.2. Periodic Cutting Sequences . . . 44

IV. Simple Loops on the Punctured Torus 47
- IV.1. Geodesics on the Punctured Torus . . . 47
- IV.2. Closed, Simple Geodesics on the Punctured Torus . . . 49
- IV.3. Theorem IV.1 through Examples . . . 55

V. Markoff Theory, a Geometric Approach 57
- V.1. Endpoints of Simple Geodesics . . . 57
  - V.1.1. Converting Cutting Sequences to L,R-Sequences . . . 57
  - V.1.2. Markoff Irrationalities and Closed, Simple Geodesics . . . 59
  - V.1.3. Limits of Geodesics and their Endpoints . . . 59
- V.2. Geometric Interpretation of Markoff Values . . . 61
  - V.2.1. The λ-Value of a Geodesic . . . 61

V.2.2. A Geometric Approach 62
V.3. Finding the Markoff Values 63
V.3.1. Markoff Irrationalities and Geodesics 64
V.3.2. Markoff Irrationalities and their Tails 64
V.4. A Link to the Langrange Spectrum 65
V.4.1. Horocycles and Markoff Irrationalities 66
V.5. Markoff Numbers viewed Geometrically 67
V.5.1. Fricke's Trace Identity and Markoff Numbers 67
V.5.2. Changing Generators 68

VI. Self-Intersecting Geodesics 71
VI.1. Closed Geodesics with One Self-Intersection 71
VI.2. Closed Geodesics with Two Intersections 73
VI.3. Further Generalizations 74

Bibliography 75

Index 77

Introduction

Markoff theory, as developed in Markoff (1879) and Markoff (1880), is a well-know part of number theory. Traditional Markoff theory is concerned with infima of certain quadratic forms. These infima lead to a spectrum of numbers, called the Markoff spectrum. Markoff studied this spectrum, resulting in a proof that all its elements below 3 are isolated, such numbers are called Markoff values. These values can be computed explicitly using Markoff numbers and Markoff forms. There is a tight connection between the Markoff spectrum and the Lagrange spectrum, which contains data about how well real numbers can be approximated by rational numbers in some sense. Consequently, the connection between the spectra leads to applications for Markoff theory in Diophantine approximation.

Despite its apparent number theoretical formulation, one can take a geometrical approach to Markoff theory. Such an approach has been explained first by Harvey Cohn in Cohn (1955). In this article he describes a relation between the Markoff numbers and certain types of matrices. This relation becomes apparent when realizing that the equation used to find the Markoff numbers resembles one of Fricke's trace identities. Harvey Cohn describes this view on the Markoff numbers in Cohn (1955), in Cohn (1971) he continues working on a geometric approach to Markoff theory. Caroline Series describes this approach in detail in Series (1985a).

It turns out that the Markoff spectrum can be defined as the set of suprema taken over diameters of certain classes of geodesics in the hyperbolic upper half-plane. Consequently, certain (classes of) geodesics give rise to the Markoff values. This follows quite directly from the traditional definition of the Markoff spectrum, but scarcely has geometrical content. An important result stated in Series (1985a) is that one can also describe, in purely geometric terms, precisely when a geodesic leads to a Markoff value. This allows for a translation of number theoretical results to geometrical results, and vice versa.

The thesis starts with a chapter devoted to the theory of continued fractions. Later on, we use these continued fractions to prove properties of the Markoff spectrum. They furthermore play a pivotal role in linking number theory to geometry, as described in Theorem III.2. Chapter II is concerned with Markoff theory as developed by Markoff. We start with a definition of the Markoff spectrum based on quadratic forms and show that this spectrum can also be obtained by means of associating a value to a doubly infinite sequence of positive integers. This last definition will prove to be helpful in showing properties about the Markoff spectrum. More precisely, we completely determine which doubly infinite sequences lead to a Markoff value. It then follows quite easily that the Markoff values make up a discrete set.

In Chapter III we consider tessellations of both the Euclidean plane and the hyperbolic upper half-plane. These tessellations lead to cutting sequences of the geodesics of the planes. Properties of these cutting sequences turn out to play a vital role in providing necessary and sufficient conditions for a geodesic in the upper half-plane to give rise to a Markoff value. Geodesics in the upper half-plane can be projected to the punctured torus under one of the tessellations described. Chapter IV discusses these projected geodesics, we prove that simple and closed geodesics have cutting sequences that obey certain rules. These two chapters provide us with the necessary tools to give a geometric interpretation of the Markoff spectrum, which is done in Chapter V. Finally, we briefly discuss geodesics leading to elements of the Markoff spectrum other than the Markoff values in Chapter VI as a direction of further research.

Chapter I
Continued Fractions

This chapter introduces the notion of a continued fraction expansion of a real number. The main purpose of this chapter is getting acquainted with continued fractions by proving results that will be needed in the chapters to come. When treating Markoff theory in Chapter II we heavily rely on the results obtained in this chapter. A result that is of special interest is Theorem I.2, which provides us with a lower bound for both the Markoff and Lagrange spectrum.

Several books on the theory of continued fractions are available. The results in this chapter are mainly taken from Coppel (2006) and Rockett and Szüsz (1992), Perron (1913) has been incidentally consulted.

section I.1 Continued Fractions

Consider a real number a, if a is not an integer then we can write $a = a_0 + b_1^{-1}$ where $a_0 = \lfloor a \rfloor$ and $b_1 = \frac{1}{a - a_0} > 1$ is a real number. If b_1 is not an integer, we can repeat this action: $b_1 = a_1 + b_2^{-1}$ with $a_1 = \lfloor b_1 \rfloor$ and $b_2 = \frac{1}{b_1 - a_1} > 1$. This process is called the continued fraction algorithm. It stops if at some point $b_n \in \mathbb{N}$, in this case we have written

$$a = a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\cdots + \frac{1}{a_n}}}}.$$

This is called the *continued fraction expansion* of a. For compactness we will denote this by $a = [a_0, a_1, ..., a_n]$. It is easy to see that in this case a must be a rational number. The converse is also true: if a is a rational number, then this process terminates. This because the continued fraction algorithm applied to a rational number is basically the Euclidean algorithm. Indeed, suppose we want to apply the continued fraction algorithm to $\frac{a}{b}$. We first apply the Euclidean algorithm to a and b, this gives

$$\begin{array}{rcl} a & = & q_0 b + r_1 \\ b & = & q_1 r_1 + r_2 \\ r_1 & = & q_2 r_2 + r_3 \\ & \vdots & \\ r_{n-1} & = & q_n r_n + 0 \end{array} \qquad \text{which yields} \qquad \begin{array}{rcl} \frac{a}{b} & = & q_0 + \left(\frac{b}{r_1}\right)^{-1} \\ \frac{b}{r_1} & = & q_1 + \left(\frac{r_1}{r_2}\right)^{-1} \\ \frac{r_1}{r_2} & = & q_2 + \left(\frac{r_2}{r_3}\right)^{-1} \\ & \vdots & \\ \frac{r_{n-1}}{r_n} & = & q_n. \end{array}$$

Thus $\frac{a}{b} = [q_0, q_1, ..., q_n]$ which proves that the continued fraction expansion terminates. We now also see that the continued fraction expansion applied to a doesn't terminate if and only if a is irrational. We use a similar notation for the continued fraction of an irrational number a, namely $[a_0, a_1, ...]$.

The integers a_n in the continued fraction expansion of $a \in \mathbb{R}$ are called the *partial quotients*, and the real numbers b_n the *complete quotients*. We easily see that the continued fraction expansion of b_n is $[a_n, a_{n+1}, ...]$.

Example I.1. Consider the golden ratio $\phi = \frac{1+\sqrt{5}}{2}$. As ϕ is a root of $X^2 - X - 1$ we know that $\phi^2 = \phi + 1$ and hence $\phi = 1 + \frac{1}{\phi}$. This immediately implies that the continued fraction expansion of ϕ is $[1, 1, 1, ...]$.

Example I.2. We can easily compute that $\sqrt{2} + 1$ has continued fraction expansion $[2, 2, 2, ...]$. It follows from

$$\sqrt{2} + 1 = 2 + \left(\frac{1}{\sqrt{2} - 1}\right)^{-1} = 2 + \left(\frac{\sqrt{2} + 1}{(\sqrt{2} - 1)(\sqrt{2} + 1)}\right)^{-1} = 2 + \frac{1}{\sqrt{2} + 1}.$$

We now also see that the continued fraction expansion of $\sqrt{2}$ is $[1, 2, 2, ...]$.

Example I.3. In general we have that for any $n \geq 1$ the number $\alpha_n := \frac{n+\sqrt{n^2+4}}{2}$ has continued fraction expansion $[n, n, n, ...]$, due to the computation below, keeping in mind that $\alpha_n > 1$

$$\alpha_n = n + \left(\frac{2}{-n + \sqrt{n^2 + 4}}\right)^{-1} = n + \left(\frac{2(-n - \sqrt{n^2 + 4})}{(-n + \sqrt{n^2 + 4})(-n - \sqrt{n^2 + 4})}\right)^{-1} = n + \frac{1}{\alpha_n}.$$

Example I.4. Given any positive real number a, it is easy to compute the continued fraction expansion of $\frac{1}{a}$ when given the expansion $[a_0, a_1, ...]$ of a. Indeed, if $a > 1$ we know that $0 < \frac{1}{a} < 1$, thus the first partial quotient of $\frac{1}{a}$ is 0. Yet then the first complete quotient belonging to $\frac{1}{a}$ is a and hence the continued fraction expansion of $\frac{1}{a}$ is simply $[0, a_0, a_1, ...]$. If $0 < a < 1$, then $a_0 = 0$ and hence we can write

$$a = \frac{1}{a_1 + \frac{1}{a_2 + ...}}.$$

Thus $\frac{1}{a} = a_1 + \frac{1}{a_2+...}$ which implies that $\frac{1}{a}$ has continued fraction expansion $[a_1, a_2, ...]$.

SECTION I.2 Convergence

Given a real number a and its continued fraction expansion $[a_0, a_1, ...]$, which can be both finite or infinite, we can consider only the first n terms of the expansion $r_n := [a_0, ..., a_n]$ for $n \geq 0$. As r_n has finite continued fraction expansion we know that it is a rational number, we call it the n-th *convergent* of a. This naming strongly suggests that $r_n \to a$ as $n \to \infty$, the main purpose of this section is to prove this very fact.

Define two sequences of integers $\{p_k\}_{k \geq -2}$ and $\{q_k\}_{k \geq -2}$ by

$$p_{-2} = 0, \;\; p_{-1} = 1, \;\; p_k = a_k p_{k-1} + p_{k-2} \tag{I.1}$$

$$q_{-2} = 1, \;\; q_{-1} = 0, \;\; q_k = a_k q_{k-1} + q_{k-2}. \tag{I.2}$$

We see that $\frac{p_0}{q_0} = a_0 = r_0$ and $\frac{p_1}{q_1} = \frac{a_1 a_0 + 1}{a_1} = a_0 + \frac{1}{a_1} = r_1$. The next lemma shows that this holds in general.

Lemma I.1. *If $a \in \mathbb{R}$ has continued fraction expansion $[a_0, a_1, ...]$, then for every $n \geq 0$ the n-th convergent of a is given by $\frac{p_n}{q_n}$,*

Proof. We prove the lemma by induction on n. We already know the lemma to be true for $n = 0, 1$. Suppose the lemma is true for the n-th convergent, so

$$\frac{p_n}{q_n} = \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}}.$$

The n-th convergent is an expression in $a_0, ..., a_n$, we get the $(n+1)$-th convergent by replacing every occurrence of a_n by $a_n + \frac{1}{a_{n+1}}$. Hence the $(n+1)$-th convergent is

$$\frac{\left(a_n + \frac{1}{a_{n+1}}\right) p_{n-1} + p_{n-2}}{\left(a_n + \frac{1}{a_{n+1}}\right) q_{n-1} + q_{n-2}} = \frac{(a_n p_{n-1} + p_{n-2}) + \frac{1}{a_{n-1}} p_{n-1}}{(a_n p_{n-1} + p_{n-2}) + \frac{1}{a_{n-1}} p_{n-1}} = \frac{a_{n+1} p_n + p_{n-1}}{a_{n+1} q_n + q_{n-1}} = \frac{p_{n+1}}{q_{n+1}}. \qquad \square$$

As said before, it is no coincidence that $\frac{p_n}{q_n}$ is called a convergent of a. Theorem I.1 shows that $\lim_{n\to\infty} \frac{p_n}{q_n}$ exists and that it equals a. To proof Theorem I.1 we need the following two lemma's.

Lemma I.2. *If $a \in \mathbb{R}$ has continued fraction expansion $[a_0, a_1, ...]$, complete quotients b_n and convergents $\frac{p_n}{q_n}$, then for every $n \geq 0$*

$$a = \frac{p_n b_{n+1} + p_{n-1}}{q_n b_{n+1} + q_{n-1}}.$$

Proof. We use induction to prove this lemma. The lemma holds if $n = 0$, because

$$\frac{p_0 b_1 + p_{-1}}{q_0 b_q 1 + q_{-1}} = \frac{a_0 b_1 + 1}{b_1} = a_0 + b_1^{-1} = a.$$

Suppose that the lemma holds for n, we prove that it also holds for $n+1$. By definition we have $b_{n+1} = a_{n+1} + b_{n+2}^{-1}$, hence $b_{n+1}b_{n+2} = b_{n+2}a_{n+1} + 1$. Using this we compute

$$\begin{aligned}\frac{p_{n+1}b_{n+2} + p_n}{q_{n+1}b_{n+2} + q_n} &= \frac{(a_{n+1}p_n + p_{n-1})b_{n+2} + p_n}{(a_{n+1}q_n + q_{n-1})b_{n+2} + q_n} \\ &= \frac{(b_{n+2}a_{n+1} + 1)p_n + b_{n+2}p_{n-1}}{(b_{n+2}a_{n+1} + 1)q_n + b_{n+2}q_{n-1}} \\ &= \frac{b_{n+1}b_{n+2}p_n + b_{n+2}p_{n-1}}{b_{n+1}b_{n+2}p_n + b_{n+2}p_{n-1}} \\ &= \frac{b_{n+1}p_n + p_{n-1}}{b_{n+1}q_n + q_{n-1}} = a.\end{aligned}$$ □

Lemma I.3. *If $a \in \mathbb{R}$ has continued fraction expansion $[a_0, a_1, ...]$ and convergents $\frac{p_n}{q_n}$, then for every $n \geq 0$*

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1}. \tag{I.3}$$

Proof. This lemma will also be proven by induction. If $n = 0$ we have $p_0 q_{-1} - p_{-1} q_0 = a_0 \cdot 0 - 1 \cdot 1 = -1$, hence the lemma is true in this case. Suppose the lemma is true for n, i.e. it holds that $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1}$. We then have

$$p_{n+1}q_n - p_n q_{n+1} = (a_{n+1}p_n + p_{n+1})q_n - p_n(a_{n+1}q_n + q_{n-1}) = -(p_n q_{n-1} - p_{n-1}q_n) = (-1)^{n+2}.$$ □

Lemma I.3 implies that for all $n \geq 0$

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n. \tag{I.4}$$

Indeed,

$$p_n q_{n-2} - p_{n-2} q_n = (a_n p_{n-1} + p_{n-2})q_{n-2} - p_{n-2}(a_n q_{n-1} + q_{n-2}) = a_n(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = (-1)^n a_n.$$

We are now ready to prove Theorem I.1 which shows that the convergents live up to their name.

Theorem I.1. *If $a \in \mathbb{R}$ irrational has continued fraction expansion $[a_0, a_1, ...]$, complete quotients b_n and convergents $\frac{p_n}{q_n}$, then $a > \frac{p_n}{q_n}$ if n is even, $a < \frac{p_n}{q_n}$ if n is odd and $\frac{p_n}{q_n} \to a$ as $n \to \infty$.*

Proof. As $a_n \geq 1$ for $n \geq 1$ it follows from the recurrence relation defining q_n that $1 = q_0 \leq q_1 < q_2 < ....$ This implies $q_n \geq n$ for every $n \geq 1$ and in particular $q_n > 0$ for $n \geq 0$. Consequently we can rewrite (I.3) and (I.4) to respectively

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n+1}}{q_n q_{n-1}}, \quad n \geq 1 \tag{I.5}$$

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n a_n}{q_n q_{n-2}}, \quad n \geq 2 \tag{I.6}$$

Equation (I.6) implies that the sequence $\{\frac{p_{2n}}{q_{2n}}\}_{n\geq 0}$ is increasing, while the sequence $\{\frac{p_{2n+1}}{q_{2n+1}}\}_{n\geq 0}$ is decreasing. Combining this with (I.5) gives that every element of the first sequence is less than every element of the second sequence. Indeed, suppose that there are n, m such that $\frac{p_{2n}}{q_{2n}} \geq \frac{p_{2m-1}}{q_{2m-1}}$. Equation (I.5) gives $\frac{p_{2n}}{q_{2n}} < \frac{p_{2n-1}}{q_{2n-1}}$ and the same for m. Hence we have

$$\frac{p_{2m-1}}{q_{2m-1}} \leq \frac{p_{2n}}{q_{2n}} < \frac{p_{2n-1}}{q_{2n-1}}$$

which can only happen if $m > n$ as the sequence with odd indices is decreasing. Yet then

$$\frac{p_{2n}}{q_{2n}} < \frac{p_{2m}}{q_{2m}} < \frac{p_{2m-1}}{q_{2m-1}}$$

as the sequence with even indices is increasing. This contradicts the assumption.

So $\{\frac{p_{2n}}{q_{2n}}\}_{n\geq 0}$ is an increasing sequence with an upper bound and $\{\frac{p_{2n+1}}{q_{2n+1}}\}_{n\geq 0}$ is decreasing with a lower bound. This implies that both sequences have a limit. Yet, $q_n \to \infty$ as $n \to \infty$, and hence we know from (I.5) that the difference between $\frac{p_{2n}}{q_{2n}}$ and $\frac{p_{2n-1}}{q_{2n-1}}$ becomes arbitrarily small. Thus both sequences actually have the same limit.

Lemma I.2 tells us that

$$a = \frac{p_n b_{n+1} + p_{n-1}}{q_n b_{n+1} + q_{n-1}}.$$

Hence

$$\begin{aligned} a - \frac{p_n}{q_n} &= \frac{p_n b_{n+1} + p_{n-1}}{q_n b_{n+1} + q_{n-1}} - \frac{p_n}{q_n} \\ &= \frac{q_n p_n b_{n+1} + q_n p_{n-1} - p_n q_n b_{n+1} - p_n q_{n-1}}{q_n(q_n b_{n+1} + q_{n-1})} \\ &= \frac{q_n p_{n-1} - p_n q_{n-1}}{q_n(q_n b_{n+1} + q_{n-1})} \\ &= \frac{(-1)^n}{q_n(q_n b_{n+1} + q_{n-1})}. \end{aligned} \tag{I.7}$$

This shows that $a > \frac{p_n}{q_n}$ if n is even and $a < \frac{p_n}{q_n}$ if n is odd. Now it must be the case that $\frac{p_n}{q_n} \to a$ as $n \to \infty$. □

As the convergents of a really converge to a, we now simply write $a = [a_0, a_1, ...]$. When dropping strictness, the statements of Theorem I.1 remain true for rational a. Yet, in this case there is some $N \in \mathbb{N}$ such that we really have the equality $\frac{p_N}{q_N} = a$.

We can now also work the other way around. Consider a sequence, either finite or infinite, $a_0, a_1, a_2, ...$ of integers such that $a_n \geq 1$ for $n \geq 1$. We can define integers p_n, q_n using the recurrence relations (I.1) and (I.2). By the proof of Theorem I.1 we know that the sequences $\{\frac{p_{2n}}{q_{2n}}\}_{n\geq 0}$ and $\{\frac{p_{2n+1}}{q_{2n+1}}\}_{n\geq 0}$ are respectively increasing and decreasing and have a common limit a. It now is the case that $a = [a_0, a_1, ...]$.

We end this section with a lemma that will be useful in Chapter II.

Lemma I.4. *If $\frac{p_n}{q_n}$ is a convergent of $a = [a_0, a_1, ...]$ then*

$$|q_n(q_n a - p_n)| = \frac{1}{[a_{n+1}, a_{n+2}, ...] + [0, a_n, a_{n-1}, ..., a_1]}.$$

Proof. As usual we denote the complete quotient $[a_n, a_{n+1}, ...]$ of a by b_n. Equation (I.7) tells us that

$$a - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n(q_n b_{n+1} + q_{n-1})}.$$

This gives that

$$|q_n(q_n a - p_n)| = q_n^2 \left| a - \frac{p_n}{q_n} \right| = \frac{q_n^2}{q_n(q_n b_{n+1} + q_{n-1})} = \frac{1}{b_{n+1} + \frac{q_{n-1}}{q_n}}.$$

We already know that $b_{n+1} = [a_{n+1}, a_{n+2}, ...]$, so we are done if we prove $\frac{q_{n-1}}{q_n} = [0, a_n, a_{n-1}, ..., a_1]$. From $q_n = a_n q_{n-1} + q_{n-2}$ it follows that $\frac{q_n}{q_{n-1}} = a_n + \frac{q_{n-2}}{q_{n-1}}$. We can continue this process until we find that $\frac{q_1}{q_0} = a_1$. It now immediately follows that $\frac{q_n}{q_{n-1}} = [a_n, a_{n-1}, ..., a_1]$. Example I.4 then gives $\frac{q_{n-1}}{q_n} = [0, a_n, a_{n-1}, ..., a_1]$. □

Suppose we want to approximate a real number a by rational numbers. Of course choosing the denominator very large eases the problem of finding a good approximation. One can wonder which rational number is the best approximation of a if we put an upper bound on the denominator. A rational number $\frac{p}{q}$ is called a *best approximation* of $a \in \mathbb{R}$ if $|qa - p| < |ya - x|$ for all integers x, y with $0 < y \leq q$ and $x \neq p$ if $y = q$. This means that $\frac{p}{q}$ is the rational number with denominator at most q closest to a. The next lemma shows that the convergents of a are best approximations of a.

Lemma I.5. *Suppose $a \in \mathbb{R}$ has continued fraction expansion $[a_0, a_1, ...]$ and complete quotients b_i. For $n \geq 1$ let $\frac{p_n}{q_n}$ be the n-th convergent of a and let $p, q \in \mathbb{Z}$ such that $0 < q \leq q_n$ and $p \neq p_n$ if $q = q_n$, then it holds that*

$$|qa - p| \geq |q_{n-1}a - p_{n-1}| > |q_n a - p_n| \tag{I.8}$$

$$\left|a - \frac{p}{q}\right| > \left|a - \frac{p_n}{q_n}\right|. \tag{I.9}$$

Proof. We first prove the strict inequality in (I.8), we have

$$\left|a - \frac{p_{n-1}}{q_{n-1}}\right| = \frac{1}{q_{n-1}(q_{n-1}b_n + q_{n-2})} \quad \text{and} \quad \left|a - \frac{p_n}{q_n}\right| = \frac{1}{q_n(q_n b_{n+1} + q_{n-1})}.$$

Thus $|q_{n-1}a - p_{n-1}| > |q_n a - p_n|$ holds if and only if $q_{n-1}b_n + q_{n-2} < q_n b_{n+1} + q_{n-1}$. From $b_n = [a_n, a_{n+1}, ...]$ and $b_{n+1} = [a_{n+1}, a_{n+2}, ...]$ it follows that $b_n = a_n + b_{n+1}^{-1}$, hence we have

$$\begin{aligned} q_{n-1}b_n + q_{n-2} &= q_{n-1}(a_n + b_{n+1}^{-1}) + q_{n-2} \\ &= a_n q_{n-1} + q_{n-2} + q_{n-1}b_{n+1}^{-1} \\ &= q_n + q_{n-1}b_{n+1}^{-1} \\ &< q_n b_{n+1} + q_{n-1} \end{aligned}$$

This proves the strict inequality in (I.8). Next we show that $|qa - p| \geq |q_{n-1}a - p_{n-1}|$. Consider the linear equations $xp_{n-1} + yp_n = p$ and $xq_{n-1} + yq_n = q$. We claim that this system has the integer solution $x = (-1)^{n-1}(p_n q - q_n p)$ and $y = (-1)^n(p_{n-1}q - q_{n-1}p)$. Indeed, by (I.3) it follows that

$$\begin{aligned} (-1)^{n-1}(p_n q - q_n p)p_{n-1} + (-1)^n(p_{n-1}q - q_{n-1}p)p_n &= -p((-1)^{n-1}p_{n-1}q_n + (-1)^n p_n q_{n-1}) = p; \\ (-1)^{n-1}(p_n q - q_n p)q_{n-1} + (-1)^n(p_{n-1}q - q_{n-1}p)q_n &= q((-1)^{n-1}p_n q_{n-1} + (-1)^n p_{n-1}q_n) = q. \end{aligned}$$

Suppose $x = 0$, this would imply that $p_n q = q_n p$. In the case that $q = q_n$ we would have $p = p_n$, but this contradicts the assumptions on p, q. If $q < q_n$, then $p = q\frac{p_n}{q_n}$ can never be an integer. Thus we conclude that $x \neq 0$. If $y = 0$ we have $|qa - p| = |x(q_{n-1}a - p_{n-1})| \geq |q_{n-1}a - p_{n-1}|$ which proves the desired inequality. If $y \neq 0$, then x and y must have opposite sign. For if they are both positive, then $q = xq_{n-1} + yq_n > q_n$ and if they are both negative, then $q = xq_{n-1} + yq_n < 0$. Both cannot happen as $0 < q \leq q_n$. Now we have

$$|qa - p| = |x(q_{n-1}a - p_{n-1}) + y(q_n a - p_n)| = |x(q_{n-1}a - p_{n-1})| + |y(q_n a - p_n)| \geq |q_{n-1}a - p_{n-1}|.$$

The second equality follows from the fact that $x(q_{n-1}a - p_{n-1})$ and $y(q_n a - p_n)$ have the same sign, as x, y and $q_{n-1}a - p_{n-1}, q_n a - p_n$ have opposite signs respectively. This proves (I.8). To prove (I.9) we compute

$$\left|a - \frac{p}{q}\right| = q^{-1}|qa - p| > q^{-1}|q_n a - p_n| = \frac{q_n}{q}\left|a - \frac{p_n}{q_n}\right| \geq \left|a - \frac{p_n}{q_n}\right|. \qquad \square$$

The above lemma shows that the convergents are best approximations, the next lemma gives another way in which the convergents are good rational approximations.

Lemma I.6. *Consider a real number a, if p, q are coprime integers with $q > 0$ such that*

$$|a - \frac{p}{q}| < \frac{1}{2q^2}, \tag{I.10}$$

then $\frac{p}{q}$ is a convergent of a.

Proof. We first prove the lemma in case $a \in \mathbb{Z}$. Suppose we have $|a - \frac{p}{q}| < \frac{1}{2q^2}$, this implies that $|aq - p| < \frac{1}{2q}$. Yet $q > 0$ and hence $\frac{1}{2q}$ is not an integer. The only way this inequality can be satisfied is when $|aq - p| = 0$. Thus it holds that $\frac{p}{q} = a$ and this shows that $\frac{p}{q}$ must be a convergent. Next we assume that $a \notin \mathbb{Z}$, as usual we denote the convergents of a by $\frac{p_n}{q_n}$. We proceed by contradiction, so suppose $\frac{p}{q}$ is not a convergent of a, but it still satisfies $|a - \frac{p}{q}| < \frac{1}{2q^2}$. We claim that there is a $n > 0$ with $q < q_n$. If a is irrational, then there are infinitely many convergents and $q_n \to \infty$ as $n \to \infty$. Hence there certainly is an $n > 0$ such that $q < q_n$. If a is rational, say $a = \frac{p_n}{q_n}$ with $n > 0$, then

$$\frac{1}{q_n} \leq \frac{|qp_n - pq_n|}{q_n} = |qa - p| = q\left|a - \frac{p}{q}\right| < \frac{1}{2q}.$$

Thus $2q < q_n$ and as $n > 0$ this proves the claim. As $q < q_n$ for some $n > 0$ there is a $m > 0$ such that $q_{m-1} \leq q < q_m$ and Lemma I.5 then tells us that $|q_{m-1}a - p_{m-1}| \leq |qa - p| < \frac{1}{2q}$. This gives

$$\frac{1}{qq_{m-1}} \leq \frac{|qp_{m-1} - pq_{m-1}|}{qq_{m-1}} = \left|\frac{p_{m-1}}{q_{m-1}} - \frac{p}{q}\right| \leq \left|\frac{p_{m-1}}{q_{m-1}} - a\right| + \left|a - \frac{p}{q}\right| < \frac{1}{2qq_{m-1}} + \frac{1}{2q^2}.$$

We now find that $\frac{1}{2qq_{m-1}} < \frac{1}{2q^2}$ which implies that $q < q_{m-1}$. This obviously contradicts the fact that $q_{m-1} \leq q$. Hence if $\frac{p}{q}$ is not a convergent of a, it cannot hold that $|a - \frac{p}{q}| < \frac{1}{2q^2}$. □

We can prove that at least one of any two consecutive convergents satisfies (I.10). Given two consecutive convergents of a we know that one is bigger or equal to a and the other is less or equal to a. Thus

$$\begin{aligned}\left|\frac{p_n}{q_n} - a\right| + \left|\frac{p_{n-1}}{q_{n-1}} - a\right| = \left|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right| = \left|\frac{p_nq_{n-1} - p_{n-1}q_n}{q_nq_{n-1}}\right| \\ = \frac{1}{q_nq_{n-1}} = \frac{2q_nq_{n-1}}{2q_n^2q_{n-1}^2} \\ = \frac{q_{n-1}^2 + q_n^2 - (q_{n-1} - q_n)^2}{2q_n^2q_{n-1}^2} \\ \leq \frac{q_{n-1}^2 + q_n^2}{2q_n^2q_{n-1}^2} \\ = \frac{1}{2q_{n-1}^2} + \frac{1}{2q_n^2}.\end{aligned}$$

Equality holds if and only if $q_n = q_{n-1}$ and this happens if and only if $n = n - 1$ or $n = 1$ and $q_0 = q_1 = 1$. In case $n > 1$ or $n = 1$ and $q_1 \neq 1$ we now see that at least one of $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_n}{q_n}$ satisfies (I.10).

If $n = 1$ and $q_0 = q_1 = 1$, then $a_1 = q_1 = 1$. So the continued fraction expansion of the complete quotient b_1 starts with $a_1 = 1$ and this implies $1 \leq b_1 < 2$. Furthermore we have $p_0 = a_0$ and thus $p_1 = a_1p_0 + 1 = a_0 + 1$. Combining all this gives

$$\left|a - \frac{p_1}{q_1}\right| = |a - a_0 - 1| = |a_0 + b_1^{-1} - a_0 - 1| = 1 - b_1^{-1} < \frac{1}{2} = \frac{1}{2q_1}.$$

This shows that at least one of any two consecutive convergents satisfied (I.10). In particular we see that for any irrational number a there are infinitely many rational numbers $\frac{p}{q}$ satisfying $\left|a - \frac{p}{q}\right| < \frac{1}{2q^2}$. The next theorem shows that we can replace the 2 in the denominator by $\sqrt{5}$.

Theorem I.2. *For any irrational number a there exist infinitely many rational numbers $\frac{p}{q}$ such that*

$$\left|a - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2}.$$

Proof. We are going to prove that at least one of any three consecutive convergents satisfies the inequality. This proves the theorem as an irrational number has an infinite number of convergents.

Suppose there are three consecutive convergents of a, say $\frac{p_{n-1}}{q_{n-1}}$, $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$, that do not satisfy the inequality $\left|a - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2}$. We then have

$$\frac{1}{q_{n-1}q_n} = \left|\frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n}\right| = \left|a - \frac{p_{n-1}}{q_{n-1}}\right| + \left|a - \frac{p_n}{q_n}\right| \geq \frac{1}{\sqrt{5}}\left(\frac{1}{q_n^2} + \frac{1}{q_{n-1}^2}\right).$$

Multiplying both sides with $\sqrt{5}q_{n-1}^2$ we find that

$$\left(\frac{q_{n-1}}{q_n}\right)^2 + 1 - \sqrt{5}\frac{q_{n-1}}{q_n} \leq 0.$$

We can play the same game with q_{n+1}, q_n instead of q_{n-1}, q_n and hence the same inequality holds if we replace q_{n-1} by q_{n+1}. The roots of the polynomial $X^2 - \sqrt{5}X + 1$ are $\frac{\sqrt{5}\pm 1}{2}$, these are both irrational and thus

$$\frac{\sqrt{5}-1}{2} < \frac{q_{n-1}}{q_n} < \frac{q_{n+1}}{q_n} < \frac{\sqrt{5}+1}{2}.$$

This implies that $a_{n+1} = \frac{q_{n+1}}{q_n} - \frac{q_{n-1}}{q_n} < \frac{\sqrt{5}+1}{2} - \frac{\sqrt{5}-1}{2} = 1$. We have reached a contradiction with the fact that $a_{n+1} > 0$. We conclude that at least one of any three consecutive convergents satisfies the inequality $\left|a - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2}$. Hence there are infinitely many solutions $\frac{p}{q}$. □

If we increase the constant $\sqrt{5}$, then Theorem I.2 fails to be true, i.e there are certain irrational numbers a such that $\left|a - \frac{p}{q}\right| < \frac{1}{cq^2}$ has only finitely many solutions for $c > \sqrt{5}$. Take for example $a = \frac{1+\sqrt{5}}{2}$, the golden ratio. In Example I.1 we saw that a has continued fraction expansion $[1, 1, 1, ...]$ and hence every complete quotient b_n is again a. By (I.7) we have

$$\left|a - \frac{p_n}{q_n}\right| = \frac{1}{q_n(q_n b_{n+1} + q_{n-1})} = \frac{1}{q_n^2\left(a + \frac{q_{n-1}}{q_n}\right)}.$$

In proving Lemma I.4 we saw that $\frac{q_n}{q_{n-1}} = [1, 1, ..., 1]$ where there are n ones, hence $\frac{q_n}{q_{n-1}} \to a$ as $n \to \infty$. This implies that $a + \frac{q_{n-1}}{q_n} \to a + a^{-1} = \sqrt{5}$ and thus $\left|a - \frac{p_n}{q_n}\right| \to \frac{1}{\sqrt{5}q_n^2}$ as $n \to \infty$. Hence for any $c > \sqrt{5}$ there are only finitely many convergents $\frac{p_n}{q_n}$ with $\left|a - \frac{p_n}{q_n}\right| < \frac{1}{cq_n^2}$. Yet, Lemma I.6 tells us that the convergents are the only possibly solutions of $\left|a - \frac{p}{q}\right| < \frac{1}{cq^2}$ for any $c > 2$. We can now conclude that $\left|a - \frac{p}{q}\right| < \frac{1}{cq^2}$ has only finitely many solutions if $c > \sqrt{5} > 2$.

If a is rational we can find infinitely many $\frac{p}{q}$ such that $\left|a - \frac{p}{q}\right| < \frac{1}{cq^2}$ for every $c > 0$. Indeed, if $a = \frac{x}{y}$ then $\left|a - \frac{p}{q}\right| = 0$ for $\frac{p}{q} = \frac{nx}{ny}$ with $n \in \mathbb{Z}_{>0}$. In Chapter II we study the Lagrange Spectrum which for $a \in \mathbb{R}$ consists of all possible values

$$\lambda(a) = \sup\left\{c \in \mathbb{R} : \left|a - \frac{p}{q}\right| < \frac{1}{cq^2} \text{ for infinitely many } p, q \in \mathbb{Z}\right\}.$$

We now already know that $\sqrt{5}$ belongs to the Langrange Spectrum and that it is actually the smallest element.

SECTION I.4 Comparing Continued Fractions

Up until now we have only considered results concerning the continued fraction expansion of one real number. In this section we prove two lemma's that compare the continued fraction expansions of two real numbers.

Lemma I.7. *Suppose for some $n \in \mathbb{Z}_{\geq 0}$ the continued fractions expansions of x, y agree in the first n coordinates and differ in the $(n+1)$-th coordinate, say $x = [a_0, a_1, ..., a_{n-1}, b_1, b_2, ...]$ and $y = [a_0, a_1, ..., a_{n-1}, c_1, c_2, ...]$ with $b_1 \neq c_1$. Then $x > y$ if and only if $b_1 > c_1$ in case n even and $x > y$ if and only if $b_1 < c_1$ if n is odd.*

Proof. If $b_1 > c_1$, then it is easy to see that $[b_1, b_2, ...] > [c_1, c_2, ...]$. Thus we have

$$a_{n-1} + \frac{1}{[b_1, b_2, ...]} < a_{n-1} + \frac{1}{[c_1, c_2, ...]}$$

which implies that

$$a_{n-2} + \frac{1}{a_{n-1} + \frac{1}{[b_1, b_2, ...]}} > a_{n-2} + \frac{1}{a_{n-1} + \frac{1}{[c_1, c_2, ...]}}.$$

We can continue in this fashion until we finally arrive at an inequality between x and y. In every step the sign flips, starting from the inequality $[b_1, b_2, ...] > [c_1, c_2, ...]$ we have n flips. So if n is even then $b_1 > c_1$ implies that $x > y$ and if n is odd it implies that $x < y$. We can do the same procedure starting from the inequality $b_1 < c_1$ and we see that $x < y$ if n is even and $x > y$ if n is odd. □

Lemma I.8. *Suppose for some $n \in \mathbb{Z}_{\geq 0}$ the continued fractions expansions of x, y agree in the first n coordinates, say $x = [a_0, a_1, ..., a_{n-1}, b_1, b_2, ...]$ and $y = [a_0, a_1, ..., a_{n-1}, c_1, c_2, ...]$, then $|x - y| \leq 2^{-(n-3)}$.*

Proof. We know that for the convergents $\frac{p_i}{q_i}$ of $[a_0, a_1, ...a_{n-1}]$ it holds that $q_i = a_i q_{i-1} + q_{i-2}$ with $q_{-2} = 1$ and $q_{-1} = 0$. We are going to use that $q_i \geq 2^{\frac{i-1}{2}}$ for $0 \leq i \leq n-1$. This can be proven by induction. If $i = 0$ we have $q_0 = 1 \geq 2^{\frac{-1}{2}}$, so the inequality holds in this case. Suppose it holds for $i = 0, ..., j-1$, we then have

$$q_j = a_j q_{j-1} + q_{j-2} \geq q_{j-1} + q_{j-2} \geq 2q_{j-2} \geq 2 \cdot 2^{\frac{j-3}{2}} = 2^{\frac{j-1}{2}}.$$

This proves the induction step, implying that the inequality holds for all $0 \leq i \leq n-1$. It holds that $\frac{p_{n-1}}{q_{n-1}}$ is a convergent of both x and y. Thus writing d_n for the complete quotients of x we have

$$\left| x - \frac{p_{n-1}}{q_{n-1}} \right| = \frac{1}{q_{n-1}(q_{n-1}d_n + q_{n-2})} \leq \frac{1}{q_{n-1}(q_{n-1}a_n + q_{n-2})} = \frac{1}{q_{n-1}q_n} \leq \frac{1}{q_{n-1}^2} \leq \frac{1}{2^{n-2}} = 2^{-(n-2)}$$

and the same holds for y. This gives

$$|x - y| = \left| \left(x - \frac{p_{n-1}}{q_{n-1}} \right) - \left(y - \frac{p_{n-1}}{q_{n-1}} \right) \right| \leq \left| x - \frac{p_{n-1}}{q_{n-1}} \right| + \left| y - \frac{p_{n-1}}{q_{n-1}} \right| \leq 2 \cdot 2^{-(n-2)} = 2^{-(n-3)}. \quad □$$

SECTION I.5 Periodic Continued Fractions

In this section we discuss periodic continued fractions and we will see that they are closely related to the roots of a quadratic polynomial. Take for example the golden ratio ϕ. Its continued fraction expansion is $[1, 1, 1...]$ which is periodic with period 1. We also know that ϕ is the root of a quadratic polynomial, namely $X^2 - X - 1$. We call an irrational number that is the root of a quadratic polynomial with integer coefficients a *quadratic irrational*. So $a \in \mathbb{R}$ is a quadratic irrational if it is the root of some $f(X) = AX^2 + BX + C$ with $A, B, C \in \mathbb{Z}$, $A \neq 0$ and $\Delta(f) = B^2 - 4AC$ not the square of an integer. In this section we show that a is a quadratic irrational if and only if its continued fraction expansion is *periodic!continued fraction expansion!periodic*

Definition I.1 (Periodic Continued Fraction). An infinite continued fraction $a = [a_0, a_1, ...]$ is called periodic if there are $n, m \in \mathbb{Z}_{\geq 0}$ such that $a_k = a_{k+m}$ for all $k \geq n$ and it is called purely periodic if $n = 0$.

So a periodic continued fraction consists of an initial block of length n followed by a repeating block of length m. We suppose that there is no shorter repeating block and that the block of length n doesn't end with a copy of the block of length m. Then the block of length m is called the period of the continued fraction, we denote this by putting a bar over it. Thus a can be denoted by

$$[a_0, ..., a_{n-1}, \overline{a_n, ..., a_{n+m-1}}].$$

If a continued fraction is purely periodic, then the initial block doesn't exist. The next two lemma's combine to prove that $a \in \mathbb{R}$ is a quadratic irrational if and only if its continued fraction expansion is periodic.

Lemma I.9. *If $a \in \mathbb{R}$ has a periodic continued fraction expansion, then it is a quadratic irrational.*

Proof. First suppose that a is purely periodic, i.e. we can write $a = [\overline{a_0, ..., a_{m-1}}]$. Denoting the complete quotients of a by b_n we then have $a = b_{km}$ for every $k \geq 1$. Lemma I.2 then tells us that

$$a = \frac{p_{m-1}b_m + p_{m-2}}{q_{m-1}b_m + q_{m-2}} = \frac{p_{m-1}a + p_{m-2}}{q_{m-1}a + q_{m-2}}.$$

This implies that $q_{m-1}a^2 + (q_{m-2} - p_{m-1})a - p_{m-2} = 0$. We have $m > 0$ and hence $q_{m-1} \geq q_0 = 1$. This proves that a is a root of the quadratic polynomial $q_{m-1}X^2 + (q_{m-2} - p_{m-1})X - p_{m-2}$. As a has an infinite continued fraction expansion we have that a is irrational and hence it is a quadratic irrational.

Next suppose $a = [a_0, ..., a_{n-1}, \overline{a_n, ..., a_{n+m-1}}]$ with $n > 0$, we now have $b_n = b_{n+m} = b_{n+2m} = ...$ Again using Lemma I.2 we have

$$a = \frac{p_{n-1}b_n + p_{n-2}}{q_{n-1}b_n + q_{n-2}} \quad \text{and} \quad a = \frac{p_{n+m-1}b_n + p_{n+m-2}}{q_{n+m-1}b_n + q_{n+m-2}}.$$

We can rewrite these equalities to

$$b_n = \frac{p_{n-2} - q_{n-2}a}{q_{n-1}a - p_{n-1}} = \frac{p_{n+m-2} - q_{n+m-2}a}{q_{n+m-1}a - p_{n+m-1}}.$$

This leads to the equation

$$\left(\begin{array}{ll} & (q_{n-2}q_{n+m-1} - q_{n-1}q_{n+m-2})a^2 \\ + & (p_{n-1}q_{n+m-2} - p_{n-2}q_{n+m-1} + q_{n-1}p_{n+m-2} - q_{n-2}p_{n+m-1})a \\ + & p_{n-2}p_{n+m-1} - p_{n-1}p_{n+m-2} \end{array} \right) = 0.$$

If $q_{n-2}q_{n+m-1} - q_{n-1}q_{n+m-2} \neq 0$, then a is the root of a quadratic polynomial and hence it is a quadratic irrational. To prove $q_{n-2}q_{n+m-1} - q_{n-1}q_{n+m-2} \neq 0$ we first prove that $\gcd(q_k, q_{k+1}) = 1$ for all $k \geq 0$. Suppose $\gcd(q_k, q_{k+1}) = d$, then $d | q_{k+1} - a_{k+1}q_k = q_{k-1}$. We can continue in this fashion and see that $d | q_{k-2}, ..., q_0 = 1$. This can only happen if $d = 1$ and hence we see that two consecutive q_k's are coprime. Now suppose that $q_{n-2}q_{n+m-1} - q_{n-1}q_{n+m-2} = 0$, this implies that $q_{n+m-1} | q_{n-1}q_{n+m-2}$. Yet q_{n+m-1} and q_{n+m-2} are coprime and thus $q_{n+m-1} | q_{n-1}$. This can only happen if $n = m = 1$ and $q_0 = q_1$. If this is not the case, then we have found a contradiction. If this is the case we repeat the whole argument with $p_{n+2m-2}, p_{n+2m-1}, q_{n+2m-2}, q_{n+2m-1}$ instead of $p_{n+m-2}, p_{n+m-1}, q_{n+m-2}, q_{n+m-1}$, which then does lead to a contradiction. We conclude that a is a quadratic irrational. □

Lemma I.10. *If $a \in \mathbb{R}$ is a quadratic irrational, then the continued fraction expansion of a is periodic.*

Proof. Consider a quadratic irrational $a = [a_0, a_1, ...]$. It is the root of some quadratic polynomial f_0 with integer coefficients, say $A_0a^2 + B_0a + C_0 = 0$. If we denote the convergents of a by $\frac{p_n}{q_n}$ and the complete quotients by b_n, then Lemma I.2 tells us that

$$a = \frac{p_n b_{n+1} + p_{n-1}}{q_n b_{n+1} + q_{n-1}}.$$

Plugging this in in $A_0a^2 + B_0a + C_0 = 0$ we find that

$$A_0(p_n b_{n+1} + p_{n-1})^2 + B_0(p_n b_{n+1} + p_{n-1})(q_n b_{n+1} + q_{n-1}) + C_0(q_n b_{n+1} + q_{n-1})^2 = 0.$$

If we expand this we see that $A_{n+1}b_{n+1}^2 + B_{n+1}b_{n+1} + C_{n+1} = 0$ with

$$\begin{aligned} A_{n+1} &= A_0p_n^2 + B_0p_nq_n + C_0q_n^2; \\ B_{n+1} &= 2A_0p_np_{n-1} + B_0p_nq_{n-1} + B_0p_{n-1}q_n + 2C_0q_nq_{n-1}; \\ C_{n+1} &= A_0p_{n-1}^2 + B_0p_{n-1}q_{n-1} + C_0q_{n-1}^2. \end{aligned}$$

Note here that $C_{n+1} = A_n$. Also note that after a straightforward, yet tedious, computation we find that the discriminant of $f_{n+1}(X) = A_{n+1}X^2 + B_{n+1}X + C_{n+1}$ is

$$\Delta(f_{n+1}) = (B_0^2 - 4A_0C_0)(p_nq_{n-1} - p_{n-1}q_n)^2 = B_0 - 4A_0C_0 = \Delta(f_0).$$

The idea of the proof is to bound the coefficients $A_{n+1}, B_{n+1}, C_{n+1}$ from above independent of n. This leads to only finitely many possibilities for the polynomial f_{n+1}. Hence in the family $\{f_{n+1}\}_{n\geq 0}$ polynomials will repeat itself and if two polynomials are the same, then the corresponding complete quotients are the same. This proves that a is periodic. So we are done if we give a bound on $A_{n+1}, B_{n+1}, C_{n+1}$.

Note that $b_{n+1} = [a_{n+1}, a_{n+2}, ...] > a_{n+1}$, hence

$$|q_na - p_n| = q_n \left| a - \frac{p_n}{q_n} \right| = \frac{q_n}{q_n(q_nb_{n+1} + q_{n-1})} = \frac{1}{q_nb_{n+1} + q_{n-1}} < \frac{1}{a_{n+1}q_n + q_{n-1}} = \frac{1}{q_{n+1}} \leq \frac{1}{q_n}.$$

This shows that we can write $p_n = q_na + \frac{\epsilon}{q_n}$ with $|\epsilon| \leq 1$. Then

$$\begin{aligned} A_{n+1} &= A_0p_n^2 + B_0p_nq_n + C_0q_n^2 \\ &= A_0\left(q_na + \frac{\epsilon}{q_n}\right)^2 + B_0q_n\left(q_na + \frac{\epsilon}{q_n}\right) + C_0q_n^2 \\ &= (A_0a^2 + B_0a + C_0)q_n^2 + (2A_0a + B_0)\epsilon + A_0\left(\frac{\epsilon}{q_n}\right)^2 \\ &= (2A_0a + B_0)\epsilon + A_0\left(\frac{\epsilon}{q_n}\right)^2 \end{aligned}$$

Thus $|A_{n+1}| \leq |2A_0a| + |B_0| + |A_0|$ which bounds A_{n+1} independent of n. Now, $C_{n+1} = A_n$, so C_{n+1} is bounded as well. Combining these bounds with the fact that the discriminant of f_{n+1} is a constant independent of n shows that B_{n+1} is bounded, too. □

We have now established a correspondence between quadratic irrationals and periodic continued fractions. But there is more, Lemma I.11 show that the continued fraction expansion of a quadratic irrational a is purely periodic if and only if a is the largest root of a *reduced polynomial*.

Definition I.2 (Reduced Polynomial). A quadratic polynomial $f(x) = Ax^2 + Bx + C$ with real roots $r < s$ is called reduced if $-1 < r < 0$ and $s > 1$.

Lemma I.11. *Consider $a \in \mathbb{R}$ a quadratic irrational, its continued fraction expansion is purely periodic if and only if a is the largest root of a reduced polynomial. In this case, denoting the other root by r and writing $a = [\,\overline{a_0, ..., a_{m-1}}\,]$, we have $\frac{-1}{r} = [\,\overline{a_{m-1}, ..., a_0}\,]$.*

Proof. First suppose that a is purely periodic, say $a = [\,\overline{a_0, ..., a_{m-1}}\,]$. Then $a_0 = a_m \geq 1$ and hence $a > 1$. From the proof of Lemma I.9 we know that a is a root of $f(X) = q_{m-1}X^2 + (q_{m-2} - p_{m-1})X - p_{m-2}$. We compute $f(0) = -p_{m-2} < 0$ and $f(-1) = q_{m-1} - q_{m-2} + p_{m-1} - p_{m-2} > 0$, so the other root of f lies between -1 and 0. This shows that f is reduced.

Next suppose that a is the largest root of a reduced polynomial $f_0(X) = A_0X^2 + B_0X + C_0$. In the same way as in the proof of Lemma I.10 we can construct a family of polynomial $\{f_n\}_{n\geq 0}$ such that $f_n(b_n) = 0$, where b_n denotes the n-th complete quotient of a. Denote the roots of f_0 by a and r, then by construction the roots of f_n for $n \geq 1$ are

$$b_n = \frac{p_{n-2} - aq_{n-2}}{aq_{n-1} - p_{n-1}} \quad \text{and} \quad s_n := \frac{p_{n-2} - rq_{n-2}}{rq_{n-1} - p_{n-1}}.$$

As f_0 is reduced we know that $-1 < r < 0$. We can compute $s_1 = \frac{1}{r-a_0}$, so $r = a_0 + \frac{1}{s_1}$. It follows that

$$\frac{1}{s_1} = -a_0 + r < -a_0 < -1$$

and hence $-1 < s_1 < 0$. In general we have that $s_n = a_n + \frac{1}{s_{n+1}}$. This is because the polynomial f_n with roots b_n and s_n can also be seen as the polynomial g_0, if we start with b_n instead of a. Then $f_n = g_1$ has roots

$$b_{n+1} = \frac{1}{b_n - a_n} \quad \text{and} \quad s_{n+1} = \frac{1}{s_n - a_n}$$

and this implies that $s_n = a_n + \frac{1}{s_{n+1}}$. Hence if $-1 < s_n < 0$, then the same holds for s_{n+1}. Using induction this proves that $-1 < s_n < 0$ for all $n \geq 1$.

Now suppose that a is not purely periodic, say $a = [a_0, ..., a_{n-1}, \overline{a_n, ..., a_{n+m-1}}\,]$ with $n > 0$. It must hold that $a_{n-1} \neq a_{n+m-1}$, for otherwise the periodic part started sooner. As the complete quotients b_n and b_{n+m} are equal, it also is the case that $s_n = s_{n+m}$. This gives that

$$s_{n-1} - s_{n+m-1} = \left(a_{n-1} + \frac{1}{s_n}\right) - \left(a_{n+m-1} + \frac{1}{s_{n+m}}\right) = a_{n-1} - a_{n+m-1}$$

is a non-zero integer. Yet, $-1 < s_{n-1}, s_{n+m-1} < 0$ which implies that $-1 < s_{n-1} - s_{n+m-1} < 1$. We have reached a contradiction and hence a must be purely periodic.

To prove the last claim of the lemma note that $f_m = f_0$, so $s_m = r$. This gives

$$r = a_0 + \frac{1}{s_1},\ s_1 = a_1 + \frac{1}{s_2},\ ...\ s_{m-1} = a_{m-1} + \frac{1}{s_m} = a_{m-1} + \frac{1}{r}.$$

Hence, in order to compute the continued fraction expansion of $\frac{-1}{r}$ we can use the following equations.

$$\begin{array}{ccccc}
\frac{-1}{r} & = & a_{m-1} - s_{m-1} & = & a_{m-1} + \left(\frac{-1}{s_{m-1}}\right)^{-1} \\
 & & \vdots & & \\
\frac{-1}{s_2} & = & a_1 - s_1 & = & a_1 + \left(\frac{-1}{s_1}\right)^{-1} \\
\frac{-1}{s_1} & = & a_0 - r & = & a_0 + \left(\frac{-1}{r}\right)^{-1}
\end{array}$$

As $\frac{-1}{r}, \frac{-1}{s_1}, ..., \frac{-1}{s_{m-1}} > 1$ we see that $\frac{-1}{r} = [\,\overline{a_{m-1}, ..., a_0}\,]$. □

We can apply Example I.4 to $\frac{-1}{r} = [\,\overline{a_{m-1}, ..., a_0}\,]$, it shows that $-r = [0,\ \overline{a_{m-1}, ..., a_0}\,]$. Hence writing the roots $r < s$ of a reduced polynomial as $r = -[0, a_{-1}, a_{-2}, ...]$ and $s = [a_0, a_1, a_2, ...]$ we see that the doubly infinite sequence $...a_{-2}, a_{-1}, a_0, a_1, a_2, ...$ is periodic. This fact will be helpful in the next chapter.

Example I.5. The golden ratio $\phi = \frac{1+\sqrt{5}}{2} > 1$ is one of the roots of $f(X) = X^2 - X - 1$. The other root is $-1 < \frac{1-\sqrt{5}}{2} < 0$. Hence f is reduced and this agrees nicely with the fact that ϕ has the purely periodic continued fraction expansion $[\,\overline{1}\,]$. The above tells us that $\frac{1-\sqrt{5}}{2} = -[0,\ \overline{1}\,]$, this can also be computed directly. It follows from

$$-\frac{1-\sqrt{5}}{2} = \frac{\sqrt{5}-1}{2} = \left(\frac{2}{\sqrt{5}-1}\right)^{-1} = \left(\frac{2(\sqrt{5}+1)}{(\sqrt{5}-1)(\sqrt{5}+1)}\right)^{-1} = \left(\frac{\sqrt{5}+1}{2}\right)^{-1} = \phi^{-1} = [0,\ \overline{1}\,].$$

Example I.6. In Example I.3 we saw that for $n \geq 1$ it holds that $\alpha_n = \frac{n+\sqrt{n^2+4}}{2}$ has continued fraction expansion $[\,\overline{n}\,]$. Hence $\alpha_n > 1$ must be the root of a reduced polynomial for every $n \geq 1$. We can also show this without using the above theory. It is not hard to see that α_n is a zero of $f_n(X) = X^2 - nX - 1$, the other root is $\beta_n = \frac{n-\sqrt{n^2+4}}{2}$. As $\sqrt{n^2+4} > n$ it holds that $\beta_n < 0$, furthermore if $n \geq 2$ we have that $\sqrt{n^2+4} < n+1$ and hence

$$\beta_n = \frac{n - \sqrt{n^2+4}}{2} > \frac{n-n-1}{2} = \frac{-1}{2} > -1.$$

This shows that f_n is reduced for every $n \geq 2$, the example above already shows that f_1 is reduced.

Chapter II
Markoff Theory

In this chapter we introduce the Markoff spectrum. We are mainly interested in a specific part of this spectrum, the set of Markoff values. The traditional way to define the Markoff spectrum is by means of binary quadratic forms. However, in proving properties of the spectrum it is useful to have an alternative description based on doubly infinite sequences of positive integers. We start by defining the spectrum in two ways and proceed by showing how to go from one description to the other. Using the definition of the Markoff spectrum based on doubly infinite sequences we prove several properties of the Markoff values, results originally obtained by Markoff in Markoff (1879) and Markoff (1880).

We end this chapter with the definition of the Lagrange spectrum, a spectrum closely related to the Markoff spectrum, and the definition of Markoff numbers and Markoff forms. These Markoff numbers and forms turn out to provide a way to explicitly compute the Markoff values.

Most of the theory in this chapter is taken from Cusick and Flahive (1989). The theory in Subsection II.1.1, relating the two definitions of the Markoff spectrum, can be found in Ross (2007).

section II.1 The Markoff Spectrum

The Markoff spectrum can be defined in two ways, using either binary quadratic forms or doubly infinite sequences of positive integers. We first introduce the spectrum by means of the quadratic forms and then describe the Markoff spectrum using doubly infinite sequences. This last definition will prove to be very useful in proving several properties of the Markoff values.

A *binary quadratic form* $f(x,y) = ax^2 + bxy + cy^2$ with real coefficients is called indefinite if it assumes both positive and negative values. Notice that the discriminant $\Delta(f) = b^2 - 4ac$ can be expressed as the square of the difference of the two roots of f. In case f is indefinite these zeroes are real and distinct and hence $\Delta(f) > 0$. To such an indefinite binary quadratic form f we can assign a real number $\mu(f)$ by

$$\mu(f) = \sup_{(x,y)\in\mathbb{Z}^2-\{(0,0)\}} \frac{\sqrt{\Delta(f)}}{|f(x,y)|}.$$

We define the Markoff spectrum $\mathbb{M}$ as all possible values of $\mu(f)$ where f is an indefinite binary quadratic form. Markoff showed in 1879 that $\mu(f) \geq \sqrt{5}$, Markoff (1879). Recall that in Chapter I we briefly discussed the Lagrange spectrum and Theorem I.2 showed that $\sqrt{5}$ is the smallest element of this spectrum. Later in this chapter we will see that the Markoff spectrum and the Lagrange spectrum coincide if we only consider values less than 3. With this information Theorem I.2 also proves that the smallest element of $\mathbb{M}$ is $\sqrt{5}$.

Furthermore, Markoff showed that the Markoff spectrum restricted to the interval $[\sqrt{5}, 3[$ is a discrete set with limit point 3. It is precisely this part of $\mathbb{M}$ that we are interested in and we will call the values $\mu(f)$ less than 3 the *Markoff values*. To prove that the Markoff values form a discrete set with limit point 3 it is useful to have an alternative description of them in terms of continued fractions. Consider a doubly infinite sequence $A = ..., a_{-1}, a_0, a_1, ..$ of positive integers. For such a sequence A we define $A_n := [a_n, a_{n+1}, ...] + [0, a_{n-1}, a_{n-2}, ...]$ for $n \in \mathbb{Z}$. To A we can associate the number $M(A) = \sup_{n \in \mathbb{Z}} A_n$ and it is the case that the set of all possible values of $M(A)$ is precisely $\mathbb{M}$. Hence the Markoff spectrum can also be defined as

$$\mathbb{M} = \{M(A) \mid A \text{ is a doubly infinite sequence of positive integers}\}.$$

In what follows we will deduce a way to associate a doubly infinite sequence A to a binary form f such that $\mu(f) = M(A)$, but only in the case that f has integer coefficients. Later in this chapter we will see that to every Markoff value we can associate a binary form with integer coefficients. So as we are only interested in the Markoff values, the restriction to integer coefficients is not a problem.

SUBSECTION II.1.1 Connecting the Two Definitions of $\mathbb{M}$

We define an action of $\mathrm{SL}(2, \mathbb{Z})$ on the set of indefinite binary quadratic forms with integer coefficients. Consider such a form $f(x, y) = ax^2+bxy+cy^2$ and a matrix $M = \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z})$, we now define

$$Mf(x, y) := f\left(M\begin{pmatrix} x \\ y \end{pmatrix}\right) = f(m_1x + m_2y, m_3x + m_4y).$$

It is a straightforward computation to verify that $\Delta(f) = \Delta(Mf)$ and furthermore the map $\mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (m_1x + m_2y, m_3x + m_4y)$ is bijective with inverse $\mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (m_4x - m_2y, -m_3x + m_1y)$. This shows that f and Mf have the same image, which in particular implies that Mf is indefinite if f is. We now see that

$$\mu(f) = \sup_{(x,y)\in\mathbb{Z}^2-\{(0,0)\}} \frac{\sqrt{\Delta(f)}}{|f(x, y)|} = \sup_{(x,y)\in\mathbb{Z}^2-\{(0,0)\}} \frac{\sqrt{\Delta(Mf)}}{|Mf(x, y)|} = \mu(Mf).$$

Thus if two forms lie in the same orbit, then their μ-values are the same. Furthermore, notice that for $n \in \mathbb{Z} - \{0\}$ we have $f(nx, ny) = n^2f(x, y) \geq f(x, y)$. consequently, to find $\mu(f)$ we only need to consider pairs $(x, y) \in \mathbb{Z}^2 - \{(0, 0)\}$ that are relatively prime. Any pair of relatively prime integers (x, y) can be written as $M(1, 0)$ for some $M \in \mathrm{SL}(2, \mathbb{Z})$ and hence we derive

$$\mu(f) = \sup_{M\in\mathrm{SL}(2,\mathbb{Z})} \frac{\sqrt{\Delta(f)}}{|Mf(1, 0)|}.$$

Suppose we can find a specific subset of the set of all indefinite binary quadratic forms such that every form is equivalent to a form in this set. Then we need only compute the μ-values of this subset, as $\mu(f)$ is invariant under the action of $\mathrm{SL}(2, \mathbb{Z})$. We are going to prove that every form f with irrational roots[1] is equivalent to a reduced form. Recall that an indefinite binary quadratic form $f(x, y) = ax^2 + bxy + cy^2$ with roots $r < s$ is called reduced if $-1 < r < 0$ and $s > 1$. We can write $f(x, y) = a(x - ry)(x - sy)$. Ordering the roots as above $r < s$, we see that

$$s - r = \frac{-b + \sqrt{\Delta(f)}}{2a} - \frac{-b - \sqrt{\Delta(f)}}{2a} = \frac{\sqrt{\Delta(f)}}{a} = \frac{\sqrt{\Delta(f)}}{f(1, 0)}.$$

Thus denoting the roots of Mf by r_M and s_M we now have

$$\mu(f) = \sup_{M\in\mathrm{SL}(2,\mathbb{Z})} \frac{\sqrt{\Delta(f)}}{|Mf(1, 0)|} = \sup_{M\in\mathrm{SL}(2,\mathbb{Z})} |r_M - s_M|.$$

This way of writing $\mu(f)$ allows us to prove that every form f with irrational roots is equivalent to a reduced form.

Lemma II.1. *Suppose f is an indefinite binary quadratic form with irrational roots, then there is some $M \in \mathrm{SL}(2, \mathbb{Z})$ such that Mf is reduced.*

[1] The condition that $f(x, y) = ax^2 + bxy + cy^2$ has irrational roots is not a restriction, as in case f has rational roots we can easily compute $\mu(f)$. Indeed, if the roots r, s were rational, say $r = \frac{n}{m}$, then $f(n, m) = a(n - \frac{n}{m} \cdot m)(n - s \cdot m) = 0$ and thus $\mu(f) = \infty$ which is clearly not a Markoff value.

Proof. Consider an indefinite binary quadratic form $f(x,y) = ax^2 + bxy + cy^2$ with irrational roots r, s. These roots are distinct as $\Delta(f) > 0$, which means that we can order them $r < s$. We have two possibilities, either $s - r \leq 2$ or $s - r > 2$. In case that $s - r \leq 2$ we replace f by an equivalent form f' such that the roots $r' < s'$ satisfy $s' - r' > 2$. We can always find such an f'. For suppose such an f' doesn't exist, then we know that $|s_M - r_M| \leq 2$ for all $M \in \mathrm{SL}(2,\mathbb{Z})$. This gives

$$\mu(f) = \sup_{M \in \mathrm{SL}(2,\mathbb{Z})} |s_M - r_M| \leq 2$$

which contradicts the fact that $\mu(f) \geq \sqrt{5} > 2$ for all forms f. So there exists a form f' equivalent to f with roots $r' < s'$ satisfying $s' - r' > 2$. This means that without loss of generality we may assume that the roots r, s of f satisfy $s - r > 2$.

There is an integer $n \in \mathbb{Z}$ such that $n - 1 < r < n$ and thus $-1 < r + n < 0$. Define the matrix $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$, then $T^{-n} = \begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix}$ and it is easy to verify that $T^{-n}f$ has roots $r + n$ and $s + n$. This form is equivalent to f and its roots satisfy $-1 < r + n < 0$ and $1 < s + n$ where the last inequality uses $(s + n) - (r + n) = s - r > 2$. This proves the theorem. □

This lemma shows that in order to find all elements of $\mathbb{M}$ we only need to compute $\mu(f)$ for reduced f. The next step is to reduce the number of matrices we have to take the supremum over. In our last definition of μ the supremum is taken over all matrices in $\mathrm{SL}(2,\mathbb{Z})$, but of course we only need to consider matrices M such that $|s_M - r_M|$ is 'large'. This will be specified in the next theorem.

Lemma II.2. *Suppose f is a reduced indefinite binary quadratic form with irrational roots and let $R \subset \mathrm{SL}(2,\mathbb{Z})$ denote the set of all matrices M such that Mf is still reduced, then*

$$\mu(f) = \sup_{M \in R} |s_M - r_M|.$$

Proof. Consider a matrix $M \in \mathrm{SL}(2,\mathbb{Z})$ such that Mf is not reduced and denote its roots by $r_M < s_M$ as usual. We show that the contribution of $|s_M - r_M|$ to the set over which the supremum is taken in computing $\mu(f)$ is either also contributed by a reduced form or too small to make a difference.

In the same way as before there is an $n \in \mathbb{Z}$ such that $T^n Mf$ has roots $r_{T^n M} = r_M - n < s_{T^n M} = s_M - n$ with $-1 < r_{T^n M} < 0$. If $T^n Mf$ is reduced, then $T^n M \in R$ and $s_M - r_M = s_{T^n M} - r_{T^n M}$. We then have

$$\mu(f) = \sup_{N \in \mathrm{SL}(2,\mathbb{Z})} |s_N - r_N| = \sup_{N \in \mathrm{SL}(2,\mathbb{Z}) - \{M\}} |s_N - r_N|$$

because the value $s_M - r_M$ that M would contribute is still contributed by $T^n M$.[2] If $T^n Mf$ is not reduced, then we have $s_{T^n M} < 1$ and thus $s_M - r_M = s_{T^n M} - r_{T^n M} < 2$. As we already know that $\mu(f) \geq \sqrt{5} > 2$, this implies that in this case we can also remove M from the set over which the supremum is taken. consequently, in order to compute $\mu(f)$ it is enough to take the supremum over R. □

We are now finally able to rewrite $\mu(f)$ in terms of continued fractions. The continued fractions we are going to use are the continued fractions representations of the roots of f. Recall that if f is a reduced form, then its roots r, s with $-1 < r < 0$ and $s > 1$ can be written as $r = -[0, x_{-1}, x_{-2}, ...]$ and $s = [x_0, x_1, x_2, ...]$. This leads to a doubly infinite sequence $X = ..., x_{-1}, x_0, x_1, ...$ which is periodic in case that f has integer coefficients. Note here that this is the first time we use this restriction to integer coefficients.

Theorem II.1. *Suppose f is a reduced indefinite binary quadratic form with integer coefficients and irrational roots $r = -[0, x_{-1}, x_{-2}, ...]$ and $s = [x_0, x_1, x_2, ...]$, then*

$$\mu(f) = \sup_{n \in \mathbb{Z}} [x_n, x_{n+1}, ...] + [0, x_{n-1}, x_{n-2}, ...].$$

[2] Note that $n \neq 0$ as $T^n Mf$ is reduced and Mf is not.

Proof. We know that $\mu(f) = \sup_{M\in R} |s_M - r_M|$ where $R \subset \mathrm{SL}(2,\mathbb{Z})$ is the set of all matrices M such that Mf is reduced. For $M \in R$ the roots of Mf can also be written in the form $-[0, y_{-1}, y_{-2}, ...]$ and $[y_0, y_1, y_2, ...]$ such that $Y = ..., y_{-1}, y_0, y_1, ...$ is periodic. The group $\mathrm{SL}(2,\mathbb{Z})$ is generated by the two matrices T and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so every matrix in $\mathrm{SL}(2,\mathbb{Z})$ is a composition of a finite number of T and S. For a continued fraction $[z_1, z_2, ...]$ we have

$$\begin{aligned} T^n[z_1, z_2, ...] &= [z_1 + n, z_2, ...] \\ S[z_1, z_2, ...] &= -[0, z_1, z_2, ...]. \end{aligned}$$

We don't need to consider powers of S as $S^2[z_1, z_2, ...] = [z_1, z_2, ...]$. It is easy to verify that if r and s are the roots of f, then $M^{-1}r, M^{-1}s$ are the roots of Mf. As M^{-1} is a finite composition of S and T, only a finite number of entries of $-[0, x_{-1}, x_{-2}, ...]$ and $[x_0, x_1, x_2, ...]$ are changed in order to get $-[0, y_{-1}, y_{-2}, ...]$ and $[y_0, y_1, y_2, ...]$. This implies that the tails of the roots of f and Mf agree. Combining this with the fact that the doubly infinite sequence Y is periodic we conclude that X and Y are the same up to orientation and indexing.

Consequently there is some $n \in \mathbb{Z}$ such that $-[0, y_{-1}, y_{-2}, ...] = -[0, x_{n-1}, x_{n-2}, ...]$ and $[y_0, y_1, y_2, ...] = [x_n, x_{n+1}, ...]$ or, in case of orientation change, $-[0, y_{-1}, y_{-2}, ...] = -[0, x_{n+1}, x_{n+2}, ...]$ and $[y_0, y_1, y_2, ...] = [x_n, x_{n-1}, ...]$. We can forget about the orientation change. Indeed, we are only interested in the difference of the roots and

$$\begin{aligned} |[x_n, x_{n-1}, ...] - (-[0, x_{n+1}, x_{n+2}, ...])| &= |x_n + [0, x_{n-1}, x_{n-2}, ...] + [0, x_{n+1}, x_{n+2}, ...]| \\ &= |[x_n, x_{n+1}, ...] - (-[0, x_{n-1}, x_{n-2}, ...])|. \end{aligned}$$

We thus see that the difference between $-[0, x_{n-1}, x_{n-2}, ...]$ and $[x_n, x_{n+1}, ...]$ and between $-[0, x_{n+1}, x_{n+2}, ...]$ and $[x_n, x_{n-1}, ...]$ is the same. This proves that

$$\mu(f) = \sup_{n\in\mathbb{Z}} [x_n, x_{n+1}, ...] + [0, x_{n-1}, x_{n-2}, ...]. \qquad \square$$

With this we have reached our goal in determining $\mu(f)$ in terms of sums of certain continued fractions. This shows that results we can obtain on $M(A) = \sup_{n\in\mathbb{Z}}[a_n, a_{n+1}, ...] + [0, a_{n-1}, a_{n-2}, ...]$ for a doubly infinite sequences $A = ..., a_{-1}, a_0, a_1, ..$ can be translated back into results about the Markoff spectrum. In the next section we are going to determine precisely which doubly infinite sequences A have $M(A) < 3$.

SECTION II.2 Doubly Infinite Sequences

Consider a doubly infinite sequences $A = ..., a_{-1}, a_0, a_1, ..$ of positive integers and recall that for $n \in \mathbb{Z}$ we defined $A_n = [a_n, a_{n+1}, ...] + [0, a_{n-1}, a_{n-2}, ...]$. We are interested in $M(A) = \sup_{n\in\mathbb{Z}} A_n$ and more precisely, we want to know when $M(A) < 3$. It is clear that in order to have $M(A) < 3$ we need $a_n = 1, 2$ for every $n \in \mathbb{Z}$, so A is a sequence in 1's and 2's. Our goal is to determine in which pattern the 1's and 2's can appear. We easily see that the pattern $2, 1, 2$ cannot appear. If it does, then one of the A_n would be $[2, ...] + [0, 1, 2, ...] > 3$. Also the pattern $1, 2, 1$ cannot occur, as $[2, 1, ...] + [0, 1, ...] > 3$. So we already see that the 1's and 2's can never be isolated.

We start with some notation. If a subscript is attached to an entry of A, this means that the entry is repeated as many times as the subscript says. For example $...1, 2_3, 1_2, ... = ...1, 2, 2, 2, 1, 1....$ Also, as in the previous chapter, a bar over a couple of entries denotes the period of the tail. First we prove a simple lemma, which will be useful in excluding certain patterns of 1's and 2's.

Lemma II.3. *Consider $n \geq 2$ an even integer and $a, b \in \mathbb{R}_{\geq 1}$, then $[2, 1_n, a] + [0, 2, 1_{n-2}, b] = 3$ if $a = b$ and $[2, 1_n, a] + [0, 2, 1_{n-2}, b] < 3$ if and only if $b < a$.*

Proof. It is a straightforward computation that $[2, 1, 1, c] + [0, 2, c] = 3$ for any $c \in \mathbb{R}_{\geq 1}$. This proves the first statement taking $c = [1_{n-2}, a]$. The second statement then follows from Lemma I.7 which gives $[0, 2, 1_{n-2}, b] < [0, 2, 1_{n-2}, a]$ if and only if $b < a$, here we really need that n is even. $\square$

We use this lemma to prove that certain patterns of 1's and 2's cannot occur in A if $A_n < 3$ for all $n \in \mathbb{Z}$. The patterns we exclude in the lemma below are not chosen at random, they turn out to be useful in proving Theorem II.2. Note here that if a certain pattern cannot occur in A, then the pattern obtained by reversing the order can still not occur. This because reversing the order is just an orientation change under which the sums A_n are invariant. For example, in the next lemma we prove that $2, 2, 2, 1, 1, 1$ cannot occur and this implies that $1, 1, 1, 2, 2, 2$ cannot occur either.

Lemma II.4. *If for a doubly infinite sequence A we have that $A_n < 3$ for all $n \in \mathbb{Z}$, then the following patterns cannot occur in A*

(i) $2, 2, 2, 1, 1, 1$

(ii) $2, 2, 1, 1, 1, 2, 2$

(iii) $1, 1, 2, 2, 2, 1, 1$

(iv) $2, 1_i, 2, 2, 1_j, 2$ *where i is even and $j \geq i + 3$*

(v) $2_4, (1, 1, 2, 2)_i, 1_4, 2, 2, 1$ *with $i \geq 1$*

Proof. We prove that all five patterns cannot occur. All five proofs have the same structure: we suppose that the pattern does occur and find a contradiction using the Lemma II.3.

(i) Suppose the pattern $2, 2, 2, 1, 1, 1$ occurs in A. This means that, supposing the last 2 in the pattern is a_0[3], we have $A = ...a_{-4}, a_{-3}, 2, 2, 2, 1, 1, 1, a_4, a_5, ...$ We define two real numbers a, b by $a = [1, a_4, a_5, ...] > 1$ and $b = [2, a_{-3}, a_{-4}, ...] > 1$ and note that $b > a$. By lemma II.4 we now have $A_0 = [2, 1, 1, a]+[0, 2, b] > 3$ which is a contradiction.

(ii) Suppose the pattern $2, 2, 1, 1, 1, 2, 2$ occurs in A. As the patterns $2, 2, 2, 1, 1, 1$ and $2, 1, 2$ cannot occur in A the pattern $2, 2, 1, 1, 1, 2, 2$ must be preceded by $1, 1$. Hence with the second 2 as a_0 we have $A = ...a_{-4}, 1, 1, 2, 2, 1, 1, 1, 2, 2, a_6....$ We again define two numbers larger than 1, namely $a = [1, 2, 2, a_6...]$ and $b = [1, 1, a_{-4}...]$ and realize that $b > a$. It holds that $A_0 = [2, 1, 1, a] + [0, 2, b] > 3$ by the previous lemma and this contradicts the assumption that $A_n < 3$ for all $n \in \mathbb{Z}$.

(iii) If the pattern $1, 1, 2, 2, 2, 1, 1$ is a part of A, then it must be followed by $2, 2$ as the patterns $2, 2, 2, 1, 1, 1$ and $1, 2, 1$ cannot occur. With the third 2 as a_0 we can now write

$$A = ...a_{-5}, 1, 1, 2, 2, 2, 1, 1, 2, 2, a_5...$$

and we define $a = [2, 2, a_5...] > 1$ and $b = [2, 1, 1, a_{-5}...] > 1$. Notice again that $b > a$ and thus $A_0 = [2, 1, 1, a] + [0, 2, b] > 3$ which leads to a contradiction.

(iv) Suppose the pattern $2, 1_i, 2, 2, 1_j, 2$ with i even and $j \geq i + 3$ occurs in A. Putting the third 2 as a_0 we can write A as $...a_{-i-3}, 2, 1_i, 2, 2, 1_{i+2}, 1_{j-i-2}, 2....$ We define $a = [1_{j-i-2}, 2...] > 1$ and $b = [2, a_{-i-3}...] > 1$ and we again see that $b > a$. This implies that $A_0 = [2, 1_{i+2}, a] + [0, 2, 1_i, b] > 3$ which cannot happen.

(v) This last proof is a bit trickier than the previous ones. Suppose the pattern $2_4, (1, 1, 2, 2)_i, 1_4, 2, 2, 1$ with $i \geq 1$ occurs in A, because $2, 1, 2$ cannot occur the pattern must be followed by a one. Hence choosing a_0 as the first 2 after 1_4 we can write $A = ...a_{-i-6}, 2_4, (1, 1, 2, 2)_i, 1_4, 2, 2, 1, 1, a_4....$ We now define four real numbers a, b, c, d by

$$\begin{aligned} a &= [2, 2, 1, 1, a_4...] > 1 \\ b &= [(2, 2, 1, 1)_{i-1}, 2_4, a_{-i-6}...] > 1 \\ c &= [a_4, a_5...] > 1 \\ d &= [(2, 2, 1, 1)_i, 2_4, a_{-i-6}...] > 1. \end{aligned}$$

From the assumption that $A_n < 3$ for all $n \in \mathbb{Z}$ it follows that $A_0 = [2, 2, 1, 1, c] + [0, 1_4, d] = [2, 1_4, d] + [0, 2, 1, 1, c] < 3$ and thus by the previous lemma we have $c < d$. Also $A_{-5} = [2, 1_4, a] + [0, 2, 1, 1, b] < 3$

[3] Changing the choice of a_0 doesn't change $M(A)$, so we can safely choose a_0 in this way.

and consequently $b < a$. Because $a = [2,2,1,1,c]$ and $d = [2,2,1,1,b]$ we can compute that $a = \frac{12c+7}{5c+3}$ and $d = \frac{12b+7}{5c+3}$. From this we find

$$b < a = \frac{12c+7}{5c+3} < \frac{12d+7}{5d+3} = \frac{179b+105}{75b+44}.$$

This shows that $75b^2 - 135b - 105 < 0$. The biggest root of the polynomial $f(x) = 75x^2 - 135x - 105$ has continued fraction expansion $[\,\overline{2,2,1,1}\,]$ and thus $b < [\,\overline{2,2,1,1}\,]$. This contradicts the fact that $b = [(2,2,1,1)_{i-1}, 2_4...]$, so we conclude that the pattern cannot occur in A. □

We will use this lemma extensively in the next theorem. This theorem basically shows that if 1_n with $n \geq 3$ occurs in A, then 2's occur only in doubletons. And the other way around, if 2_n with $n \geq 3$ occurs in A, then a 1 can only occur in a doubleton. All this provided of course that $A_k < 3$ for all $k \in \mathbb{Z}$.

Theorem II.2. *Consider a doubly infinite sequence A which contains no infinite repetition of consecutive 1's or 2's and has $A_k < 3$ for all $k \in \mathbb{Z}$, if A contains 1_n with $n \geq 3$, then A has form* (II.1) *and if 2_n with $n \geq 3$ occurs in A then it has form* (II.2)

$$...2,2,1_{n_{-1}},2,2,1_{n_0},2,2,1_{n_1},2,2... \tag{II.1}$$

$$...1,1,2_{n_{-1}},1,1,2_{n_0},1,1,2_{n_1},1,1... \tag{II.2}$$

where $n_k \in \mathbb{Z}_{>0}$.

Proof. First suppose that somewhere in A we find 1_n with $n \geq 3$ and furthermore suppose that A is not of the form (II.1). As A does not contain an infinite repetition of 1's, this implies that 2_m with $m \geq 3$ occurs in A. The second part of the previous lemma shows that in fact $n \geq 4$ and the third part of the lemma shows the same for m. Consider a 1_n with $n \geq 4$ in A and take the 2_m with $m \geq 4$ which lies closest to 1_n. By the first part of the previous lemma 1_n and 2_m cannot immediately follow each other, so there are some doubletons 1's and 2's between them. This gives us that in A there is a pattern[4]

$$2_4, (1,1,2,2)_i, 1_n, 2, 2, 1$$

where $i \geq 1$, $n \geq 4$ and where the last 1 follows from the first part of the previous lemma. The fourth part of the lemma tells us that $n \leq 4$ and consequently $n = 4$. But now we can use the last part of the lemma which shows that a pattern $2_4, (1,1,2,2)_i, 1_4, 2, 2, 1$ cannot exist and we conclude that A must have form (II.1).

The proof that if A contains 2_n with $n \geq 3$, then A has form (II.2) is basically the same. Again we assume that A doesn't have the form (II.2) which implies that both 2_n and 1_m with $n, m \geq 4$ occur in A. If we choose the 2_n and 1_m which are closest to each other we again wind up with a pattern $2_4, (1,1,2,2)_i, 1_4, 2, 2, 1$ with $i \geq 1$ which is impossible. □

This theorem doesn't cover all possibilities for A, we have excluded the cases when (i) $A = \overline{1}$; (ii) $A = \overline{2}$; (iii) both 1 and 2 only occur in doubletons in A or (iv) A contains both the symbols 1 and 2 and an infinite repetition of consecutive 1's or 2's.

For the first three cases it holds that $M(A) < 3$. If $A = \overline{1}$, then every A_n equals $[1,1,1...] + [0,1,1...] = 2\frac{1+\sqrt{5}}{2} - 1 = \sqrt{5} < 3$. If $A = \overline{2}$ then $A_n = [2,2,2...] + [0,2,2...] = (\sqrt{2}+1) + (\sqrt{2}-1) = 2\sqrt{2} < 3$ for every $n \in \mathbb{Z}$. Thus in both cases A satisfies the condition that $A_n < 3$ for all $n \in \mathbb{Z}$. If both 1 and 2 only occur in doubletons, then $A = \overline{1,1,2,2}$ and for this sequence we can also verify that $A_n < 3$ for all $n \in \mathbb{Z}$. Note that $A = \overline{1}$ can be written in form (II.2) with $n_k = 0$ for all k and $A = \overline{2}$ is (II.1) with $n_k = 0$ for all k. Furthermore $A = \overline{1,1,2,2}$ can be written in both forms, taking $n_k = 2$ for all k.

The last case, i.e. if A contains an infinite repetition of 1 or 2, can also be written in form (II.1) or (II.2). If A contains an infinite repetition of 1 respectively 2, then it is form (II.2) respectively (II.1) with an infinite number of consecutive

[4] If 1_n lies before 2_m, then we must reverse the order of the pattern. As we know that a pattern cannot occur in A if and only if the reversed pattern cannot occur, it is enough to prove that one of them cannot occur.

n_k's equal to zero. In this case it is actually not true that $A_n < 3$ for all $n \in \mathbb{Z}$, we will be able to prove this after we have learned more properties of the n_k's.

The following lemmas tell us more about the behavior of the n_k's, in the end we have completely determined what A looks like if $A_n < 3$ for all $n \in \mathbb{Z}$. In the following, when we talk about a doubly infinite sequence A, we assume that it contains no infinite repetition of 1 or 2. From the above we know that if $A_n < 3$ for all $n \in \mathbb{Z}$, then A is of form (II.1) or (II.2) with either $n_k = 0$ for all k, $n_k = 2$ for all k or $n_k \in \mathbb{Z}_{>0}$ where at least for one k we have $n_k \geq 4$.

Lemma II.5. *Consider a doubly infinite sequence A with $A_n < 3$ for all $n \in \mathbb{Z}$. All of the integers n_k in* (II.1) *or* (II.2) *are even.*

Proof. We treat the cases that A is of from (II.1) or (II.2) separately. Both proofs have the same structure, based on contradiction. We start by assuming that both even and odd n_k occur and show that this leads to a contradiction with the fact that $A_n < 3$ for all $n \in \mathbb{Z}$. Consequently all of the n_k's are either even or odd. We then suppose that all of them are odd and again reach a contradiction.

Suppose A has form (II.1) and that some of the n_k are even and some are odd. This means that there are $n_i = 2k$ and $n_{i+1} = 2l + 1$ with integers $k, l \in \mathbb{Z}_{\geq 0}$. Write $A = ...2, 2, 1_{2k}, 2, 2, 1_{2l+1}, 2, 2...$, as $2l + 1$ is odd we have $[2, \overline{1}] < [2, 1_{2l+1}, 2, 2...]$ and as $2k$ is even we have $[0, 2, \overline{1}] < [0, 2, 1_{2k}, 2, 2...]$. Hence one of the A_n's is

$$[2, 1_{2l+1}, 2, 2...] + [0, 2, 1_{2k}, 2, 2...] > [2, \overline{1}] + [0, 2, \overline{1}] = 3$$

which is in contradiction with the assumption that $A_n < 3$ for all $n \in \mathbb{Z}$. Thus we may assume that all of the n_k are either odd or even. Suppose they are all odd and consider $n_i = 2k + 1$ and $n_{i+1} = 2l + 1$ with $k \geq l$. Then A looks like[5]

$$...2, 2, 1_{2k-2l+3}, 1_{2l-2}, 2, 2, 1_{2l}, 1, 2, 2, 1_{n_{i+2}}...$$

Define real numbers $a, b > 1$ by $a = [1, 2, 2, 1_{n_{i+1}}...]$ and $b = [1_{2k-2l+3}, 2, 2...]$, then $b > a$ and thus by Lemma II.3

$$[2, 1_{2l}, a] + [0, 2, 1_{2l-2}, b] > 3.$$

As this is one of the A_n's we have reached a contradiction. This proves that all the n_k are even.

Now suppose that A is of form (II.2) and that both odd and even n_k occur. If $n_i = 2k + 1$ is odd and $n_{i+1} = 2l$ is even, then A looks like

$$...1, 1, 2_{2k-1}, 2, 2, 1, 1, 2_{2l}, 1, 1...$$

Define $a, b \in \mathbb{R}_{>1}$ by $a = [2_{2l}, 1, 1...]$ and $b = [2_{2k-1}, 1, 1...]$, with Lemma I.7 we can show that $b > a$ holds both if $k \geq l$ and if $k < l$. Lemma II.3 then tells us that one of the A_n's is $[2, 1, 1, a] + [0, 2, b] > 3$ which cannot be. This shows that all the n_k are either even or odd.

Suppose all the n_k are odd. We can write $n_i = 2k + 1$ and $n_{i+1} = 2l + 1$ and without loss of generality we assume $k \leq l$. Indeed, if this never happens, we just look at the reversed sequence. The sequence A looks like

$$...1, 1, 2_{2k-1}, 2, 2, 1, 1, 2_{2l+1}, 1, 1...$$

We again define two real numbers $a, b > 1$ by $a = [2_{2l+1}, 1, 1...]$ and $b = [2_{2k-1}, 1, 1...]$. Because $k \leq l$ it follows that $b > a$. Thus one of the A_n has value $[2, 1, 1, a] + [0, 2, b] > 3$ which contradicts the assumptions of the theorem. This conclude the proof that all of the n_k's must be even. □

With this theorem we can show that any sequence $A \neq \overline{1}$ in form (II.2) can also be written in form (II.1). Suppose that A has form (II.2) and that $A \neq \overline{1}$, i.e $A = ...1, 1, 2_{n_{-1}}, 1, 1, 2_{n_0}, 1, 1, 2_{n_1}, 1, 1...$ with not all the n_k's equal to zero. As any n_k is even, we can write $n_k = 2m_k$ for every k. this means that 2_{n_k} can be split into m_k doubletons and if we separate these doubletons by putting 1_0 between them we have rewritten A to the form (II.1). The next lemma deals with the question how much the n_k's can differ from each other.

[5] If $k < l$, then we can consider the reversed pattern instead.

Lemma II.6. *Consider a doubly infinite sequence $A \neq \overline{1}$ with $A_n < 3$ for all $n \in \mathbb{Z}$. Writing A as*

$$...2, 2, 1_{2m_{-1}}, 2, 2, 1_{2m_0}, 2, 2, 1_{2m_1}, 2, 2...$$

we have that $m_{i+1} - m_i = -1, 0, 1$.

Proof. By Lemma II.4 we know that no pattern $2, 1_k, 2, 2, 1_l, 2$ with k even and $l \geq k+3$ can occur in A. Reversing the order of the pattern we see that $k \geq l+3$ also is a problem in case l is even. Hence as $2m_i$ and $2m_{i+1}$ are even we have $2m_{i+1} \leq 2m_i + 2$ and $2m_i \leq 2m_{i+1} + 2$. From the first inequality it follows that $m_{i+1} - m_i \leq 1$ and from the second it follows that $m_{i+1} - m_i \geq -1$. Thus we conclude that $m_{i+1} - m_i = -1, 0, 1$. □

We are now able to give sufficient and necessary conditions for a doubly infinite sequence A to have $A_n < 3$ for all $n \in \mathbb{Z}$.

Theorem II.3. *Consider a doubly infinite sequence $A \neq \overline{1}$. We have $A_n < 3$ for all $n \in \mathbb{Z}$ if and only if $A =$ $...2, 2, 1_{2m_{-1}}, 2, 2, 1_{2m_0}, 2, 2, 1_{2m_1}, 2, 2...$ with*

(i) $m_i - m_j = -1, 0, 1$ for all i, j;

(ii) If $m_{i+1} - m_i = -1, 1$ respectively, then the first $m_{i+j+1} - m_{i-j}$ for $j = 1, 2, ...$ that is non-zero equals 1, -1 respectively.

Proof. First we show that both conditions on the m_k's hold if $A_n < 3$ for all $n \in \mathbb{Z}$. To show that the first condition must hold, we assume that it doesn't holds and derive a contradiction with $A_n < 3$ for all $n \in \mathbb{Z}$. So suppose that the first condition doesn't hold, i.e. there are m_i and m_j that differ more than 1. We know that we cannot have $m_j = m_{i+1}$, so first suppose that $m_j = m_{i+2}$, i.e a pattern of the form

$$2, 2, 1_k, 2, 2, 1_{k+2}, 2, 2, 1_{k+4}, 2, 2, 1_l$$

occurs in A. The fourth part of Lemma II.4 with $i = l$ and $j = k + 4$ shows that we must have $l \geq k + 2$. Define two real numbers $a, b > 1$ by $a = [2, 2, 1_l, 2, 2...]$ and $b = [2, 2, 1_k, 2, 2...]$. Then by Lemma I.7 we have that $b > a$ and thus by Lemma II.3 one of the A_n has value

$$[2, 1_{k+4}, a] + [0, 2, 1_{k+2}, b] > 3$$

which is a contradiction. Next suppose that a pattern of the form

$$2, 2, 1_k, 2, 21_{k+2}...2, 2, 1_{k+2}, 2, 2, 1_{k+4}, 2, 2$$

occurs, where there are x repetitions of $2, 2, 1_{k+2}$ between 1_k and 1_{k+4}. Furthermore we have chosen k and x such that every $2m_i = l$ and $2m_j = l + 4$ are at least x repetitions of $2, 2, 1_{l+2}$ apart, so x is minimal. Take the 1_k in the pattern as 1_{2m_0}. As x is minimal we know that $2m_{-1}, ..., 2m_{-(x-1)}$ must be either k or $k + 2$. We now show that they are all $k + 2$. Suppose that $2m_{-l} = k$ for some $1 \leq l \leq x - 1$ and that $2m_{-j} = k + 2$ for all $1 \leq j < l$. We define two real numbers $a, b > 1$ by

$a = [2, 2, 1_{k+2}...2, 2, 1_{k+2}, 2, 2, 1_{k+4}, 2, 2...]$ where $2, 2, 1_{k+2}$ occurs $x - 1$ times before $2, 2, 1_{k+4}$
$b = [2, 2, 1_{k+2}...2, 2, 1_{k+2}, 2, 2, 1_k, 2, 2...]$ where $2, 2, 1_{k+2}$ occurs $l - 1$ times before $2, 1, 1_k$.

Because $l \leq x-1$ we know that $b > a$ and thus one of the A_n's is $[2, 1_{k+2}, a] + [0, 2, 1_k, b] > 3$. As this contradicts the assumption we conclude that $2m_{-1}, ..., 2m_{-(x-1)}$ must be $k + 2$.

We now again define two real numbers $a, b > 1$ by

$a = [2, 2, 1_{k+2}...2, 2, 1_{k+2}, 2, 2, 1_{k+4}, 2, 2...]$ where $2, 2, 1_{k+2}$ occurs $x - 1$ times before $2, 2, 1_{k+4}$
$b = [2, 2, 1_{k+2}...2, 2, 1_{k+2}, 2, 2, 1_{2m_{-x}}, 2, 2...]$ where $2, 2, 1_{k+2}$ occurs $x - 1$ times before $2, 1, 1_{2m_{-x}}$.

If $b \geq a$ then we would have a contradiction, consequently it holds that $b < a$. This can only happen if $2m_{-x} \geq k+4$. But then $2m_{-x}$ and k differ by more than 2 and there are less than x repetitions of $2, 2, 1_{k+2}$ between $1_{2m_{-x}}$ and 1_k which contradict the minimality of x. So we conclude that the first condition must hold.

Next we show that the second condition must hold if $A_n < 3$ for all $n \in \mathbb{Z}$. We again proceed by contradiction, so suppose that it doesn't hold. This means that there is an i such that $m_{i+1} - m_i = \pm 1$ and the first $m_{i+j+1} - m_{i-j}$ for $j = 1, 2, ...$ that is non-zero is also ± 1. We prove the case where both differences are 1, the other case has a similar proof. If $m_{i+1} - m_i = 1$, then there is some $k \in \mathbb{N}$ with $2m_i = k$ and $2m_{i+1} = k + 2$. Hence if $m_{i+j+1} - m_{i-j} = 1$ and $m_{i+l+1} - m_{i-l} = 0$ for $0 < l < j$ we have a pattern in A like

$$2, 2, 1_k, 2, 2, 1_x ... 2, 2, 1_x, 1_k, 2, 2, 1_{k+2}, 2, 2, 1_x ... 2, 2, 1_x, 2, 2, 1_{k+2}, 2, 2$$

where there are $j - 1$ occurrences of $2, 2, 1_x$ between the 1_k's and between the 1_{k+2}'s. Define $a, b > 1$ by

$$\begin{aligned} a &= [2, 2, 1_x ... 2, 2, 1_x, 2, 2, 1_{k+2}, 2, 2 ...] \\ b &= [2, 2, 1_x ... 2, 2, 1_x, 2, 2, 1_k, 2, 2, ...], \end{aligned}$$

where both a and b start with $j - 1$ times $2, 2, 1_x$. It now holds that $b > a$ and consequently one of the A_n is $[2, 1_{k+2}, a] + [0, 2, 1_k, b] > 3$. This is in clear contradiction with the assumptions on the A_n, implying that the second condition must hold if $A_n < 3$ for all $n \in \mathbb{Z}$.

The proof is complete when we show that $A_n < 3$ for all $n \in \mathbb{Z}$ follows from the two conditions. Consider the value of $A_n = [a_n, a_{n+1} ...] + [0, a_{n-1} ...]$. In case $a_n = 1$ we have $[a_n, a_{n+1} ...] < 2$ and $[0, a_{n-1} ...] < 1$ and hence $A_n < 3$. If $a_n = 2$ we need to do a bit more work, we distinguish two cases, $a_{n+1} = 2$ or $a_{n+1} = 1$.

Write $A = x, 1_k, 2, 2, 1_l, y$ where x, y are two sequences that start with $2, 2$. Without loss of generality we assume that $l \geq k$ and then the first condition on the m_k's gives that $k = l$ or $l = k + 2$. First suppose that $k = l$. In case $a_n = a_{n+1} = 2$ we have

$$A_n = [2, 2, 1_k, y] + [0, 1_k, x] = [2, 1_k, x] + [0, 2, 1_k, y] < 3.$$

The inequality follows from Lemma II.3 with $a = [x]$ and $b = [1, 1, y]$. Indeed, we have $b < a$ as a starts with a 2. In case $a_n = 2$ and $a_{n+1} = 1$ we have $A_n = [2, 1_k, y] + [0, 2, 1_k, x] < 3$ because $[1, 1, x] < [y]$.

Now suppose that $l = k + 2$. In case that $a_n = a_{n+1} = 2$ we have

$$A_n = [2, 2, 1_{k+2}, y] + [0, 1_k, x] = [2, 1_k, x] + [0, 2, 1_{k+2}, y] < 3.$$

The inequality follows from $[1, 1, 1, 1, y] < [x]$, as $[x]$ starts with a 2. If $a_n = 2$ and $a_{n+1} = 1$, we have $A_n = [2, 1_{k+2}, y] + [0, 2, 1_k, x]$ which is less than 3 if and only if $[x] < [y]$. This follows immediately from the second condition on the m_k's. Consequently we see that $A_n < 3$ for all $n \in \mathbb{Z}$ if the two conditions on the m_k's hold. □

With this theorem we have completely determined the form of a doubly infinite sequence A with $A_n < 3$ for all $n \in \mathbb{Z}$. Note that the theorem gives us information about the sequence of m_k's. Such a sequence can be formed for any sequence $A \neq \overline{1}$ with $A_n < 3$ for all $n \in \mathbb{Z}$. We will call this sequence of m_k's the *sequence associated to* A.

The first condition on the m_k's is very easy to interpret, it just means that there is an $m \in \mathbb{Z}_{\geq 0}$ such that m_k is either m or $m + 1$. The second condition determines in what order the m's and $m + 1$'s can occur, but it might be a bit hard to explicitly imagine the possibilities for this order. In the following we will give a little more insight in this.

Theorem II.4. *Consider a doubly infinite sequence* $A \neq \overline{1}$ *with* $A_n < 3$ *for all* $n \in \mathbb{Z}$. *There is an* $m \in \mathbb{Z}_{\geq 0}$ *such that the associated sequence* $... m_{-1}, m_0, m_1 ...$ *is either* $\overline{m}$ *or*

$$... m \pm 1, m_{p_{-1}}, m \pm 1, m_{p_0}, m \pm 1, m_{p_1}, m \pm 1 ...$$

for non-negative integers p_k *with*

(i) $p_i - p_j = -1, 0, 1$ *for all* i, j;

(ii) If $p_{i+1} - p_i = -1, 1$ *respectively, then the first* $p_{i+j+1} - p_{i-j}$ *for* $j = 1, 2, ...$ *that is non-zero equals* $1, -1$ *respectively.*

Proof. The constant sequence $\overline{m}$ clearly satisfies the two conditions of Theorem II.3. If the sequence of m_k's contains both $m, m+1$ (or $m, m-1$), then one of them must be isolated, otherwise the second condition of Theorem II.3 is not met.

Suppose that the sequence of m_k contains an infinite string of m's or $m+1$'s. If for example the sequence ends in an infinite repetition of m's, then A looks like

$$...2, 2, 1_{2m}, 2, 2, 1_{2m+2}, \overline{2, 2, 1_{2m}} ....$$

and then one of the A_n's is

$$[\overline{2, 2, 1_{2m}}] + [0, 1_{2m+2}, 2, 2, 1_{2m}...] = [0, 2, 1_{2m}, \overline{2, 2, 1_{2m}}] + [2, 1_{2m+2}, 2, 2, 1_{2m}...].$$

Defining $a = [2, 2, 1_{2m}...] > 1$ and $b = [\overline{2, 2, 1_{2m}}] > 1$ we see that $a \leq b$ and hence $A_n \geq 3$. This shows that no infinite string of m can occur and with a similar argument we can prove that an infinite string of $m \pm 1$ is impossible, too.

Hence we can write the sequence of m_k's in the form

$$...m \pm 1, m_{p_{-1}}, m \pm 1, m_{p_0}, m \pm 1, m_{p_1}, m \pm 1...$$

with $p_k \in \mathbb{Z}_{\geq 0}$ and the proof is finished if we show that the sequence of p_k's satisfies the two conditions.

We first show that $p_{i+1} - p_i = -1, 0, 1$ for all i. In the sequence of m_k's we have a pattern $m_{p_i}, m \pm 1, m_{p_{i+1}}$, suppose now that $p_i > p_{i+1} + 1$. We prove that this leads to a contradiction. If the $m \pm 1$ in the above pattern is chosen to be m_0, then $m_0 - m_{-1} = \pm 1$ and $m_j - m_{-j-1} = 0$ for all $j = 1, 2, ..., p_{i+1}$. Because $p_i > p_{i+1} + 1$ we furthermore have that $m_j - m_{-j-1} = \pm 1$ for $j = p_{i+1} + 1$. This contradicts the second condition of Theorem II.3, hence we conclude that $p_i \leq p_{i+1} + 1$. A similar argument with i and $i+1$ interchanged shows that $p_{i+1} \leq p_i + 1$. Consequently it holds that $p_{i+1} - p_i = -1, 0, 1$ for all i. We need to generalize this to Condition 1, which says that $p_i - p_j = -1, 0, 1$ for all i, j. The proof for this is similar to the proof of the first condition of Theorem II.3.

Next we show that the second condition holds. Suppose that $p_{i+1} - p_i = -1$. If we again take the $m \pm 1$ in the pattern $m_{p_i}, m \pm 1, m_{p_{i+1}}$ as m_0, then we have that $m_0 - m_{-1} = \pm 1$. For $j = 1, 2, ...$ consider now the differences $m_j - m_{-j-1}$.

As condition 2 of Theorem II.3 holds, we know that the first time this difference is non-zero it is in fact ∓ 1. This means that $m_j = m$ and $m_{-j-1} = m \pm 1$. Notice that the difference $m_j - m_{-j-1} = 0$ for $j = 1, 2...p_{i+1}$, because $p_{i+1} - p_i = -1$. So if the difference is non-zero for the first time, this means that $m_j = m$ is part of a pattern of consecutive m's which is longer than the pattern of consecutive m's following $m_{-j-1} = m \pm 1$. This implies that the first time the difference $p_{i+k+1} - p_{i-k}$ is non-zero, it is in fact 1. In case that $p_{i+1} - p_i = 1$, we can use a similar argument to prove that the first non-zero $p_{i+k+1} - p_{i-k}$ is -1. We conclude that condition 2 is satisfied. □

Recall that in Theorem II.2 four possibilities for A were excluded. We already dealt with the first three cases, in all of these cases we have that $A_n < 3$ for all $n \in \mathbb{Z}$. The fourth case was that A is not constant and it contains an infinite repetition of consecutive 1's or 2's. This is the case when an infinite number of consecutive m_k's is equal to zero. In the proof of Theorem II.4 we just saw that the sequence of m_k's cannot contain an infinite string of zeroes, unless the sequence of m_k's is constant. This shows that A cannot contain an infinite repetition of 1 or 2, unless all of the m_k are zero, i.e. $A = \overline{2}$. Thus in the fourth case we don't have $A_n < 3$ for all $n \in \mathbb{Z}$.

We will call a sequence satisfying the two conditions of Theorem II.4 *Markoff balanced*. The sequence of p_k's in Theorem II.4 is called the *derived sequence* of the sequence of m_k's. We can repeat the process, i.e. the sequence p_k has a derived sequence q_k which again satisfies the two conditions. Hence the derived sequence of a Markoff balanced sequence is again Markoff balanced. In the proof of the next theorem we see that we can only derive the sequence m_k a finite number of times before we obtain a constant sequence. With this theorem we will be able to prove Theorem II.6, which is the main result obtained by Markoff. see Markoff (1879).

Theorem II.5. *Given any $\epsilon > 0$ the set of doubly infinite sequences A with $A_n < 3 - \epsilon$ for all $n \in \mathbb{Z}$ is finite.*

In the proof we use Lemma I.8 which bounds the distance between two numbers in terms of their continued fractions expansion.

Proof. Consider an arbitrary $\epsilon > 0$ and a doubly infinite sequence $A \neq \overline{1}$ such that $A_n < 3 - \epsilon$ for all $n \in \mathbb{Z}$. If we prove that there is only a finite number of options for A, then we are done. By Theorem II.3 we know that A has the form

$$A = ...2, 2, 1_{2m_{-1}}, 2, 2, 1_{2m_0}, 2, 2, 1_{2m_1}, 2, 2...$$

such that $m_i - m_j = -1, 0, 1$ for all i, j and if $m_{i+1} - m_i = -1, 1$ respectively, then the first $m_{i+j+1} - m_{i-j}$ for $j = 1, 2, ...$ that is non-zero equals 1, -1 respectively. Furthermore, Theorem II.4 tells us that $...m_{-1}, m_0, m_1...$ is either constant or

$$...m \pm 1, m_{p_{-1}}, m \pm 1, m_{p_0}, m \pm 1, m_{p_1}, m \pm 1...$$

for non-negative integers p_k which satisfy the same two conditions. We will first show that m is bounded by a number which depends only on ϵ. Writing $A = ...a_{-1}, a_0, a_1...$ we can choose a_0 such that $a_{-1} = a_0 = 2$ and $a_1 = a_2 = 1$ and then $A_0 = [2, 1, 1, a] + [0, 2, b]$ for some $a, b \in \mathbb{R}_{\geq 1}$. Lemma II.3 tells us that $A_0 < 3$ if and only if $b < a$. In particular we have that $a \neq b$, so we cannot have $a_{i+1} = a_{-i}$ for all $i = 2, 3, ...$ Hence there is an $n \in \mathbb{N}$ such that $a_{i+1} = a_{-i}$ for all $i = 2, 3, ..., n$ and $a_{n+2} \neq a_{-n-1}$.

Define real numbers $c = [2, 1, 1, a_3, ..., a_{n+1}]$ and $d = [0, 2, a_{-2}, ..., a_{-n}]$, then by Lemma I.8 we have that $|[2, 1, 1, a] - c| \to 0$ as $n \to \infty$ and $|[0, 2, b] - d| \to 0$ as $n \to \infty$. Lemma II.3 then tells us that $c + d \to 3$ as $n \to \infty$ and hence $A_0 = [2, 1, 1, a] + [0, 2, b] \to 3$ as $n \to \infty$. Combining this with the fact that $A_0 < 3 - \epsilon$ we conclude that n is bounded from above by some number $N(\epsilon)$ depending only on ϵ, i.e. $n \leq N(\epsilon)$.

If the sequence of m_k's is constant, say $...m_{-1}, m_0, m_1... = \overline{m}$, then A looks like

$$...1_{2m}, 2, 2, 1_{2m}, 2, 2, 1_{2m}...$$

If we again choose a_0 such that $a_{-1} = a_0 = 2$ and $a_1 = a_2 = 1$, then the number n is precisely $2m - 2$. So we have $2m - 2 \leq N(\epsilon)$ which shows that m is bounded.

Now suppose the sequence of m_k's is not constant, say m and $m + 1$ occur in the sequence. We choose a_0 such that

$$a_{-2m-2} = 2, \quad a_{-2m-1} = \ldots = a_{-2} = 1, \quad a_{-1} = a_0 = 2, \quad a_1 = \ldots = a_{2m+2} = 1, \quad a_{2m+3} = 2.$$

If we let $m_{-1} = m$ correspond to $a_{-2m-1}, ..., a_{-2} = 1_{2m}$, then $m_0 = m + 1$ corresponds to $a_1, ..., a_{2m+2} = 1_{2m+2}$. We then have that $a_{i+1} = a_{-i}$ for $i \geq 2$ as long as $m_j = m_{-j-1}$. Suppose $m_j = m_{-j-1}$ for $j = 1, ..., M$ and that $m_{M+1} \neq m_{-M-2}$. As $m_j = m, m + 1$, we then have that

$$(M + 2)(2m + 2) \leq n$$

where n again denote the number such that $a_{i+1} = a_{-i}$ for $i = 2, ..., n$ and $a_{n+2} \neq a_{-n-1}$. This implies that $(M + 2)(2m + 2) \leq N(\epsilon)$ and hence both M and m are bounded. This concludes the proof that m is bounded by a number depending only on ϵ.

Next, we show that p is bounded, where $p \in \mathbb{Z}_{\geq 0}$ is such that the sequence $...p_{-1}, p_0, p_1...$ is either constant $\overline{p}$ or

$$...p \pm 1, p_{q_{-1}}, p \pm 1, p_{q_0}, p \pm 1, p_{q_1}, p \pm 1...$$

Suppose that the sequence of p_k's is constant $\overline{p}$. In case $p = 0$, there is nothing to prove. If $p > 0$ we are in the above situation that the sequence of m_k's is non-constant and we have $M = p$. Thus $(p + 2)(2m + 2) \leq N(\epsilon)$ which proves that p is bounded.

If the sequence of p_k's is non-constant, we can choose $p_{-1} = p + 1$ and $p_0 = p$. Define P such that $p_{-i-1} = p_i$ for $i = 1, ..., S$ and $p_{-P-2} \neq p_{P+1}$. Writing the sequence of m_k's as

$$...m \pm 1, m_{p_{-1}}, m \pm 1, m_{p_0}, m \pm 1...$$

and taking m_0 as the $m \pm 1$ between $m_{p_{-1}}$ and m_{p_0}, we see that $(p + 1)(P + 1) \leq M$. As M is bounded, so are p and P. Furthermore we see that $P < M$.

We can repeat this argument. If the sequence of p_k is non-constant, we can write the sequence of q_k's as $\overline{q}$ or

$$...q \pm 1, q_{r_{-1}}, q \pm 1, q_{r_0}, q \pm 1, q_{r_1}, q \pm 1...$$

and we can again show that q is bounded. Furthermore if the sequence of q_k's is non-constant, we can in the same way as above define the number Q and we have $Q < P < M$. These numbers decrease strictly and this implies that after a finite number of derivations we end up with a constant sequence. There is only a finite number of possibilities for this constant sequence, for if this sequence is $\overline{s}$, then $s \leq N(\epsilon)$. Taking such a constant sequence and working backwards we find that there are only finitely possibilities for A. □

We have shown that for a doubly infinite sequences A with $A_n < 3 - \epsilon$ for all $n \in \mathbb{Z}$ the associate sequence of m_k's will become constant after a finite number of derivations. Consider a non-constant sequence that is Markoff balanced. This sequence is purely periodic if and only if the derived sequence is purely periodic. As a constant sequence is surely purely periodic, we see that any doubly infinite sequence A with $A_n < 3 - \epsilon$ for all $n \in \mathbb{Z}$ is purely periodic. Note that this nicely agrees with Theorem II.1 which states that $\mu(f) = M(A)$ for some purely periodic sequence A if f has integer coefficients. Later in this chapter we will see that every Markoff value can be obtained as $\mu(f)$ for some binary quadratic form f with integer coefficients.

Using Theorem II.5 it is not hard to prove Theorem II.6, an important result obtained by Markoff.

Theorem II.6. *The Markoff values form a countable and discrete set with* 3 *as its only limit point.*

Proof. From Theorem II.5 it immediately follows that for every $\epsilon > 0$ the number of Markoff values less than $3 - \epsilon$ is finite. This already proves that the set of Markoff values is countable and discrete and that no number other than 3 can be a possible limit point. Hence we are left with proving that 3 is indeed a limit point.

For $k = 1, 2, \ldots$ consider the doubly infinite sequence $A(k) = \overline{2, 2, 1_{2k}}$, this sequence clearly satisfies the conditions of Theorem II.3 and hence $A(k)_n < 3$ for all $n \in \mathbb{Z}$ and for all $k \in \mathbb{N}$. As $A(k)$ is purely periodic, there are only a finite number of distinct values for $A(k)_n$. This implies that there is some $\epsilon > 0$ such that $A(k)_n < 3 - \epsilon$ for all $n \in \mathbb{Z}$ and thus $M(A(k)) < 3$. Yet, as $k \to \infty$ we have that $A(k) \to A = \overline{1}, 2, 2, \overline{1}$. We can compute $M(A) = 3$ and hence $M(A(k)) \to 3$ as $k \to \infty$. This proves that 3 is a limit point. □

Remark II.1. We have completely determined how a doubly infinite sequence A with $M(A) < 3$ looks like, it is either $\overline{1}$ or a purely periodic sequence in 1's and 2's such that its associated sequence is Markoff balanced. Theorem II.3 tells us precisely when a doubly infinite sequence A has the property that $A_n < 3$ for all $n \in \mathbb{Z}$. Indeed, such a sequence has to be a sequence in 1's and 2's such that its associated sequence is Markoff balanced. As $A_n < 3$ for all $n \in \mathbb{Z}$ it holds that $M(A) \leq 3$. The inequality $M(A) < 3$ is known to hold if and only if in addition the sequence A is purely periodic. Consequently we see that $M(A) = 3$, if A is a sequence in 1's and 2's with Markoff balanced associated sequence which is not purely periodic. It holds that A is not purely periodic if and only if its associated sequence is not purely periodic.

We claim that this associated sequence cannot have a periodic tail. Indeed, if it has a periodic tail then this tail becomes constant after a finite number of derivations. However, the sequence is not purely periodic and hence the rest of this derived sequence cannot be constant. Such a sequence can never be Markoff balanced. This is in contradiction with Theorem II.4 and consequently we conclude that the tails of a sequence A with $M(A) = 3$ cannot be periodic.

SECTION II.3 The Lagrange Spectrum

Up until now we have dealt with the Markoff spectrum $\mathbb{M}$ which is obtained by associating a value in $\mathbb{R} \cup \{\infty\}$ to a indefinite binary quadratic form. We now discuss the Lagrange spectrum which, unlike $\mathbb{M}$, isn't obtained from binary forms but from real numbers. Consider a real number a, with arbitrary $p, q \in \mathbb{Z}$ we define $\lambda(a)$ by

$$\lambda(a) = \limsup_{q \to \infty} \frac{1}{|q(qa - p)|}.$$

We now define the Lagrange spectrum by $\mathbb{L} = \{\lambda(a) | a \in \mathbb{R}\}$. In case $\lambda(a) < \infty$ we see that for every $0 < c < \lambda(a)$ the inequality $\frac{1}{|q(qa-p)|} \geq c$ has infinitely many solutions $(p, q) \in \mathbb{Z}^2$. Rewriting this we see that $|a - \frac{p}{q}| \leq \frac{1}{cq^2}$ has infinitely many solutions. Hence we can also define

$$\lambda(a) = \sup \left\{ c \in \mathbb{R} : \left| a - \frac{p}{q} \right| < \frac{1}{cq^2} \text{ for infinitely many } p, q \in \mathbb{Z} \right\}.$$

Recall Theorem I.2, we already saw in Chapter I that $\sqrt{5}$ is the smallest element of the Lagrange spectrum and that the λ-value of any rational number is ∞. A real number a with the property that $\lambda(a) < 3$ is thus always irrational, we call such a number a *Markoff irrationality*.

SUBSECTION II.3.1 The Spectra Compared

The Lagrange spectrum and the Markoff spectrum are not the same. It can be proven that $\mathbb{L} \subsetneq \mathbb{M}$, see for example Cusick and Flahive (1989). However, the spectra coincide on the interval $[\sqrt{5}, 3[$. To prove this we first show that $\mathbb{L}$ can also be defined in terms of doubly infinite sequences. Recall that for a doubly infinite sequences $A = ..., a_{-1}, a_0, a_1, ..$ of positive integers we have $A_n = [a_n, a_{n+1}, ...] + [0, a_{n-1}, a_{n-2}, ...]$ for $n \in \mathbb{Z}$. Define $L(A) = \limsup_{n \in \mathbb{Z}} A_n$, then $\mathbb{L}$ has the following alternative description.

Theorem II.7. *We have* $\mathbb{L} = \{L(A) \mid A$ *is a doubly infinite sequence of positive integers*$\}$.

Proof. For $a = [b_0, b_1, ...] \in \mathbb{R}$ consider the inequality $|q(qa - p)| < \frac{1}{2}$ in $p, q \in \mathbb{Z}$. Lemma I.6 tells us that $\frac{p}{q}$ is a convergent of a if p, q satisfy this inequality. Hence we have $\frac{p}{q} = \frac{p_i}{q_i} = [b_0, b_1, ..., b_i]$ for some $i \in \mathbb{Z}$.

In computing $\lambda(a)$ we compute $\limsup_{q \to \infty}(|q(qa - p)|)^{-1}$. This means that in order to get $\lambda(a)$ we only need to consider $(p, q) \in \mathbb{Z}^2$ such that $|q(qa - p)|$ is small, by the above we can thus assume that $(p, q) = (p_i, q_i)$ such that $\frac{p_i}{q_i}$ is a convergent of a. Recall Lemma I.4 which states that

$$|q_i(q_i a - p_i)| = \frac{1}{[b_{i+1}, b_{i+2}, ...] + [0, b_i, b_{i-1}, ..., b_1]}.$$

We first prove that we can find an $a \in \mathbb{R}$ to every doubly infinite sequence A such that $\lambda(a) = L(A)$. For a doubly infinite sequence $A = ..., a_{-1}, a_0, a_1, ..$ we have either $L(A) = \limsup_{n \to \infty} A_n$ or $L(A) = \limsup_{n \to -\infty} A_n$. In the first case put $a = [a_0, a_1, ...]$ and in the second case put $a = [a_0, a_{-1}, ...]$. Using Lemma I.4 we see in the first case that

$$\begin{aligned} \lambda(a) &= \limsup_{i \to \infty} \frac{1}{|q_i(q_i a - p_i)|} \\ &= \limsup_{i \to \infty} [a_{i+1}, a_{i+2}, ...] + [0, a_i, a_{i-1}, ..., a_1] \\ &= L(A) \end{aligned}$$

and in the second case we see

$$\begin{aligned} \lambda(a) &= \limsup_{i \to \infty} \frac{1}{|q_i(q_i a - p_i)|} \\ &= \limsup_{i \to \infty} [a_{-(i+1)}, a_{-(i+2)}, ...] + [0, a_-i, a_{-(i-1)}, ..., a_1] \\ &= \limsup_{i \to -\infty} [a_{i+1}, a_{i+2}, ...] + [0, a_i, a_{i-1}, ..., a_1] \\ &= L(A). \end{aligned}$$

This proves that all of the values $L(A)$ are contained in $\mathbb{L}$. If we prove that there is a doubly infinite sequence A for every $a \in \mathbb{R}$ such that $L(A) = \lambda(a)$ then we are done. So consider $a = [b_0, b_1, b_2, ...]$ and define a doubly infinite sequence by $A = ...b_2, b_1, b_0, b_1, b_2, ...$ Again using Lemma I.4 we see $L(A) = \limsup_{i \to \infty} A_i = \lambda(a)$. This concludes the proof. □

With this result it is not hard to show that $\mathbb{L}$ and $\mathbb{M}$ agree if we only consider elements less than 3. We know that every Markoff value is given by $M(A) = \sup_{n\in\mathbb{Z}} A_n$ where A is a purely periodic doubly infinite sequence. Furthermore we know that every element of $\mathbb{L}$ is given as $L(A) = \limsup_{n\in\mathbb{Z}} A_n$ for some doubly infinite sequence A. Now, if A is purely periodic, then taking a supremum or a limit superior is the same, i.e. $L(A) = M(A)$. This shows that $\mathbb{L}$ and $\mathbb{M}$ coincide in the interval $[\sqrt{5}, 3[$. Consequently we can find a real number a and an indefinite binary quadratic form f such that $\mu(f) = m = \lambda(a)$ for each Markoff value m.

This gives us a way to associate a class of indefinite binary quadratic forms to a class of real numbers. Consider a class of indefinite binary quadratic forms with the same μ-value which is less than 3. We know that two indefinite binary quadratic forms have the same μ-value if they are equivalent under the $\mathrm{SL}(2,\mathbb{Z})$-action defined in Section II.1. By Lemma II.1 we know that this class contains a reduced form g and Theorem II.1 tells us that $\mu(g) = M(A)$ with $A = ..., a_{-1}, a_0, a_1, ...$ such that $-[0, a_{-1}, a_{-2}, ...]$ and $[a_0, a_1, ...]$ are the roots of g. By the proof of Theorem II.7 we now have that $M(A) = L(A) = \lambda(a)$ for $a = [a_0, a_1, ...]$. The proof of Theorem II.7 gives that $\lambda(a) = \lambda(b)$ for $a, b \in \mathbb{R}$ if and only if the tails of the continued fraction expansions of a and b agree. So we see that to the class of forms equivalent to g we can associate the class of real numbers whose tails agree with the tail of a.

We know that $M(A) < 3$ if and only if A is a purely periodic sequence in 1's and 2's such that its associated sequence is Markoff balanced. Hence $\lambda(a) < 3$ if and only if the tail of a agrees with the tail of such a sequence A. With this we have completely determined when $a \in \mathbb{R}$ is a Markoff irrationality.

Remark II.2. We can also determine when a real number a has a λ-value equal to 3. Remark II.1 tells us that a sequence A has the property $M(A) = 3$, if it is a sequence in 1's and 2's with Markoff balanced associated sequence, which is not purely periodic. For such a sequence we have that $M(A) = \sup_{n\in\mathbb{Z}} A_n = 3$ and hence

$$L(A) = \limsup_{n\in\mathbb{Z}} A_n \le \sup_{n\in\mathbb{Z}} A_n = 3.$$

It cannot hold that $L(A) < 3$. Indeed, the proof of Theorem II.7 tells us that one of the tails of a doubly infinite sequence B and one of those of a sequence C agree if and only if $L(B) = L(C)$. Which tail agrees with which tail depends on whether we use $\limsup_{n\to\infty}$ or $\limsup_{n\to-\infty}$ to compute the L-values. This implies that a sequence B with $L(B) < 3$ must have a periodic tail. So in case $L(A) < 3$ it holds that a tail of A is periodic. This contradicts Remark II.1 which states that the tails of A cannot be periodic. Consequently we have that $L(A)$ must be equal to 3. Using the proof of Theorem II.7 we can now conclude that a real number a has the property that $\lambda(a) = 3$, if the tail of a agrees with the tail of a non-periodic sequence A in 1's and 2's with Markoff balanced associated sequence.

SECTION II.4 Markoff Numbers and Forms

We conclude this chapter with the introduction of Markoff numbers and Markoff forms. Consider the equation

$$x^2 + y^2 + z^2 = 3xyz. \tag{II.3}$$

We are interested in solutions $(x, y, z) \in \mathbb{Z}^3_{>0}$ and for such a solution the integers x, y, z are called *Markoff numbers*. Obviously, if (x, y, z) is a solution to this equation then so is $\sigma(x, y, z)$ for every $\sigma \in S_3$, the permutation group of 3 elements. Furthermore, if (x, y, z) is a solution then so are $(3yz - x, y, z)$, $(x, 3xz - y, z)$ and $(x, y, 3xy - z)$. These are called the *neighbors* of (x, y, z). It is not hard to prove that these three tuples are also solutions of (II.3). For example, if (x, y, z) is a solution, then x is a zero of $f(X) = X^2 - 3yzX + y^2 + z^2$. Writing a for the other zero of f we find $f(X) = (X - x)(X - a) = X^2 - (x + a)X + ax = X^2 - 3yzX + y^2 + z^2$ and hence $a = 3yz - x$.

We immediately see that $(1, 1, 1)$ and $(1, 1, 2)$ are solutions. We call these solutions singular as their coefficients are not distinct.

Lemma II.7. *Up to permutation the only singular solutions to* (II.3) *are* $(1, 1, 1)$ *and* $(1, 1, 2)$.

Proof. Consider a singular solution (x, y, z), say with $x = y$. Then (II.3) comes down to $z^2 = (2 - 3z)x^2$ and thus $x|z$. If we put $z = kx$, then equation (II.3) becomes $(2 + k^2)x^2 = 3kx^3$ and hence $2 + k^2 = 3kx$. It now holds that $x = \frac{k^2+2}{3k} = \frac{k}{3} + \frac{2}{3k} \in \mathbb{Z}$. If $k \equiv 0 \mod 3$, then $\frac{k}{3} \in \mathbb{Z}$ and $\frac{2}{3k} \notin \mathbb{Z}$, which contradicts the fact that $x \in \mathbb{Z}$.

If $k \equiv 1 \mod 3$, then we can write $k = 3l + 1$ and we find $x = l + \frac{1}{3} + \frac{2}{9l+3} = m + \frac{3l+3}{9l+3}$. This can only be an integer if $l = 0$, so $k = 1$. We then see that $x = y = z$ and (II.3) becomes $3x^2 = 3x^3$. The only positive solution is $x = 1$. This leads to the singular solution $(1, 1, 1)$.

If $k \equiv 2 \mod 3$, we write $k = 3l + 2$ and find $x = l + \frac{6l+6}{9l+6}$. For this to be an integer we need $l = 0$ and thus $k = 2$. Now (II.3) comes down to $6x^2 = 6x^3$ which has only $x = 1$ as positive solution. We find the singular solution $(1, 1, 2)$. □

Starting with the solution $(1, 1, 1)$ we can generate all solutions of (II.3) using just two operations.

Theorem II.8. *All solutions to* (II.3) *can be found from* $(1, 1, 1)$ *using the operations* $r : (x, y, z) \to (z, x, y)$ *and* $s : (x, y, z) \to (x, 3xy - z, y)$.

Proof. The proof consists of three steps. We first prove that up to permutation every solution can be found from $(1, 1, 1)$ by successively taking neighbors. The next step is to show that up to permutation the three neighbors of a solution (x, y, z) can be obtained by using the operations r and s on (x, y, z). We have then proved the theorem up to permutation. Thus the last step is to show that every permutation of a solution can be obtained with the operations r and s.

Consider a non-singular solution (x, y, z), arranged such that $x > y > z$. The polynomial $f(X) = X^2 - 3yzX + y^2 + z^2$ has x and $3yz - x$ as zeros and using $y > z$ and $z \geq 1$ we see

$$f(y) = 2y^2 + z^2 - 3y^2z = (2 - 3z)y^2 + z^2 \leq -y^2 + z^2 < 0.$$

Hence y lies strictly between x and $3yz - x$, together with $x > y$ this gives $y > 3yz - x$. This means that the neighbor $(3yz - x, y, z)$ has y as maximal element and consequently the maximal element of this neighbor is smaller than the maximal element of (x, y, z).

From the fact that $x > y, z$ it immediately follows that $3xz - y > x$ and $3xy - z > z$. Hence the maximal elements of the other two neighbors $(x, 3xz - y, z)$ and $(x, y, 3xy - z)$ are bigger than the maximal element of (x, y, z). So every non-singular solution of (II.3) has precisely one neighbor with smaller maximal element. Given a non-singular solution we can successively take neighbors with smaller maximal element and as all solutions are in $\mathbb{Z}^3_{>0}$ this process must terminate. The only way the process can terminate is if we ultimately end up with a singular solution. As the singular solutions $(1, 1, 1)$ and $(1, 1, 2)$ are neighbors this concludes the first step of the proof: up to permutation every solution can be found from $(1, 1, 1)$ by successively taking neighbors.

Next we show that up to permutation the three neighbors of a solution (x, y, z) can be obtained by using the operations r and s on (x, y, z). Of course $s(x, y, z) = (x, 3xy - z, y)$ gives already one of the neighbors. The other two are given by $s \circ r(x, y, z) = s(z, x, y) = (z, 3xz - y, x)$ and $s \circ r \circ r(x, y, z) = s \circ r(z, x, y) = s(y, z, x) = (y, 3yz - x, z)$.

We conclude by proving that every permutation of a solution can be obtained with the operations r and s. Given a solution (x, y, z) we get two permutations by $r(x, y, z) = (z, x, y)$ and $r \circ r(x, y, z) = (y, z, x)$. Note that this gives already all permutations of $(1, 1, 2)$. For a non-singular solution (x, y, z) we have three other permutations, namely (x, z, y), (y, z, x) and (z, y, x). The solution (x, y, z) has one neighbor, say $(x, y, 3xy - z)$, with smaller maximal element. Suppose that all of the permutations of this neighbor can be obtained with r and s. Then in particular we have the permutation $(x, y, 3xy - z)$ and we find

$$\begin{aligned} s(x, y, 3xy - z) &= (x, 3xy - 3xy + z, y) = (x, z, y); \\ r \circ s(x, y, 3xy - z) &= (y, x, z); \\ r \circ r \circ s(x, y, 3xy - z) &= (z, y, x). \end{aligned}$$

So if all permutations of $(x, y, 3xy - z)$ can be obtained using r and s, then the same is true for (x, y, z). Working backwards by successively taking neighbors with smaller maximal element we end up in $(1, 1, 1)$. As surely all permutations of $(1, 1, 1)$ can be obtained from r and s we see by induction that the same is true for all solutions of (II.3). □

SUBSECTION II.4.1 Markoff Forms

Using the Markoff numbers we can construct the *Markoff forms*, indefinite binary quadratic forms with integer coefficients such that their μ-values are Markoff values. With these forms we can finally fulfill the promise that to every Markoff value we can associate a binary form with integer coefficients. We only construct the Markoff forms and state several properties of them, proofs can be found in Cusick and Flahive (1989).

Consider (k, l, m) a solution of (II.3), arranged such that m is the maximal element. To this solution we will associate a Markoff form f_m. Let r be the least positive residue of $\pm\frac{k}{l} \mod m$. For r to be well-defined we need to check two things, namely that r is invariant under interchanging k and l and that the inverses of k and l exist modulo m. First we prove that k, l are invertible modulo m, i.e. we need to show that $\gcd(k, m) = \gcd(l, m) = 1$. Suppose $\gcd(k, m) = d$, then it follows from (II.3) that also $d|l$. But if k, l, m are all divisible by d, then so are all the elements of the neighbors of (k, l, m). We know that by repeatedly taking the neighbor with smaller maximal element we end up in $(1, 1, 1)$. As the elements of this solution clearly have greatest common divisor 1, we conclude that $d = 1$. The proof that $\gcd(l, m) = 1$ is similar.

Next we show that the least positive residue of $\pm\frac{k}{l} \mod m$ is the same as the least positive residue of $\pm\frac{l}{k} \mod m$. The least positive residue of $\pm\frac{k}{l} \mod m$ is either $\frac{k}{l} \mod m$ or $\frac{-k}{l} \mod m$. From (II.3) it follows that $k^2 + l^2 \equiv 0 \mod m$ and thus $\frac{k^2}{l^2} \equiv -1 \mod m$. This gives that

$$-\frac{l}{k} \equiv \frac{k^2}{l^2} \cdot \frac{l}{k} \equiv \frac{k}{l} \mod m.$$

Of course we then also have $\frac{l}{k} \equiv -\frac{k}{l} \mod m$. This proves that r is invariant under interchanging k and l.

We define s by $s := \frac{r^2+1}{m}$, note that s is an integer as $r^2 + 1 \equiv \frac{k^2}{l^2} + 1 \equiv -1 + 1 \equiv 0 \mod m$. The *Markoff form* f_m is now given by

$$f_m(x, y) = mx^2 + (3m - 2r)xy + (s - 3r)y^2.$$

Markoff proved that $\mu(f_m) = \frac{\sqrt{9m^2-4}}{m}$. Note that $\frac{\sqrt{9m^2-4}}{m} < \frac{\sqrt{9m^2}}{m} = 3$ which means that $\mu(f_m)$ is a Markoff value. It even holds that all Markoff values are of the form $\frac{\sqrt{9m^2-4}}{m}$ where m is the maximal element of a solution of (II.3). Consequently we can associate a binary form with integer coefficients to every Markoff value. It is easy to compute that $\Delta(f_m) = 9m^2 - 4$, thus the definition of μ tells us that $\min_{(x,y)\in\mathbb{Z}^2-\{(0,0)\}} |f_m(x, y)| = m$.

Now it is not hard to compute some Markoff values and the Markoff forms and doubly infinite sequences corresponding to them. The smallest Markoff value is $\sqrt{5}$, it is obtained from the solution $(1, 1, 1)$ of (II.3). We can easily compute that in this case $r = 1$ and $s = 2$ and so $f_1(x, y) = x^2 + xy - y^2$. Its roots are $\frac{-1\pm\sqrt{5}}{2}$, hence f_1 is not reduced, but $T^{-1}f_1(x, y) = x^2 - xy - y^2$ is. The doubly infinite sequence belonging to $\sqrt{5}$ is thus the sequence obtained from the continued fraction expansions of the roots of this form. We already saw in Chapter I that the polynomial $f(X) = X^2 - X - 1$ has the golden ratio $\frac{1+\sqrt{5}}{2} = [\,\overline{1}\,]$ as one of its roots and hence the doubly infinite sequence belonging to $\sqrt{5}$ is just $\overline{\,1\,}$. Note that this can easily be verified, as for $A = \overline{\,1\,}$ we have $A_n = [\,\overline{1}\,] + [0, \overline{1}\,] = \sqrt{5}$ for every $n \in \mathbb{Z}$.

The second Markoff value is $\sqrt{8}$, obtained from the solution $(1, 1, 2)$. The Markoff form corresponding to $\sqrt{8}$ is $f_2(x, y) = 2x^2 + 4xy - 2y^2$. Its roots are $-1 \pm \sqrt{2}$, which means that f_2 itself is not reduced. As the difference between the roots is more than 2 and the smallest root lies between -3 and -2, we see that $T^{-2}f_2$ is reduced. We have $T^{-2}f_2 = 2x^2 - 4xy - 2y^2$ with roots $1 \pm \sqrt{2}$. In Example I.2 we computed that $1 + \sqrt{2} = [\,\overline{2}\,]$ and consequently the doubly infinite sequence belonging to $\sqrt{5}$ is $A = \overline{\,2\,}$. We can verify this by computing $A_n = [\,\overline{2}\,] + [0, \overline{2}\,] = 2\sqrt{2} = \sqrt{8}$ for every $n \in \mathbb{Z}$.

In a similar way we can compute the values in Table II.1.

Solution of (II.3)	Markoff value	Markoff form	Doubly infinite sequence
$(1,1,1)$	$\sqrt{5}$	x^2+xy-y^2	$\overline{1}$
$(1,1,2)$	$\sqrt{8}$	$2x^2+4xy-2y^2$	$\overline{2}$
$(1,2,5)$	$\frac{\sqrt{221}}{5}$	$5x^2+11xy-5y^2$	$\overline{2,2,1,1}$
$(1,5,13)$	$\frac{\sqrt{1517}}{13}$	$13x^2+29xy-13y^2$	$\overline{2,2,1,1,1,1}$
$(2,5,29)$	$\frac{\sqrt{7565}}{29}$	$29x^2+63xy-31y^2$	$\overline{2,2,2,2,1,1}$

Table II.1.: Markoff values with associated forms and sequences

Chapter III
Cutting Sequences

We have learned about Markoff theory, a theory that belongs to the field of number theory. The purpose of this thesis is looking at the Markoff values in a geometric way. The geometric approach to the Markoff values we will describe in chapters to come relies on properties of geodesics and tessellations of the hyperbolic upper half-plane.

In this chapter we consider geodesics of both the Euclidean plane and the hyperbolic upper half-plane. When tessellating these planes the geodesics cut the sides of the tessellation, leading to cutting sequences. We prove several properties of these cutting sequences. The theory we thus obtain will be used to give a geometric interpretation of Markoff theory in Chapter V.

The results of this chapter are mainly taken from Series (1985a) and Series (1985b).

section III.1 Square Tessellation of the Euclidean Plane

We will begin with the Euclidean plane $\mathbb{R}^2 = \{(x, y)|x, y \in \mathbb{R}\}$. We can subdivide the plane into squares by drawing vertical lines $\{x = n|n \in \mathbb{Z}\}$ and horizontal lines $\{y = n|n \in \mathbb{Z}\}$. This gives a *tessellation* of the Euclidean plane in the following sense.

Definition III.1 (Tessellation). A tessellation of a plane is a covering of this plane by congruent figures such that there are no overlaps and no gaps.

We will call the tessellation mentioned above the *square tessellation* of $\mathbb{R}^2$, as the congruent figures covering $\mathbb{R}^2$ are squares. This tessellation can also be obtained in the following way. Consider an action of $\mathbb{Z}^2$ on $\mathbb{R}^2$, given $(n, m) \in \mathbb{Z}^2$ and $(x, y) \in \mathbb{R}^2$ we put

$$(n, m)(x, y) = (x + n, y + m).$$

The orbit of $(x, y) \in \mathbb{R}^2$ is the set $\{(n, m)(x, y)|(n, m) \in \mathbb{Z}^2\}$. Given an action of a group G on a set X we can divide X into equivalence classes by saying that two elements of X are equivalent if and only if they lie in the same orbit. In some sense we are dividing X by G in this way. It is often important to have a well-behaved set of representatives of X/G, the notion of a *fundamental domain* captures this.

Definition III.2 (Fundamental Domain). A fundamental domain F for an action of a group G on a set X is a connected subset of X such that every orbit under G has one or two points in F. If an orbit has two points in F, then both of these points lie on the boundary ∂F.

It is not hard to see that $[0, 1] \times [0, 1]$ is a fundamental domain for the action of $\mathbb{Z}^2$ on $\mathbb{R}^2$. So the square tessellation can also be obtained by dividing the plane in fundamental domains of this action. Note that the above action of $\mathbb{Z}^2$

on $\mathbb{R}^2$ gives rise to an action of a group isomorphic to $\mathbb{Z}^2$ on $\mathbb{R}^2$. In the following we replace $\mathbb{Z}^2$ by an isomorphic group which gives us a natural way to label the sides of the tessellation. Consider the maps

$$a : \mathbb{R}^2 \to \mathbb{R}^2,\ \ a(x, y) = (x + 1, y) \quad \text{and} \quad b : \mathbb{R}^2 \to \mathbb{R}^2,\ \ b(x, y) = (x, y + 1).$$

The free group generated by a and b consists of all maps of the form $\mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (x + n, y + m)$ for some $n, m \in \mathbb{Z}$. It is easy to see that this group is isomorphic to $\mathbb{Z}^2$, thus we can also see the square tessellation as a division of $\mathbb{R}^2$ in fundamental domains of an action of $\langle a, b\rangle$ on $\mathbb{R}^2$.

Let us now focus on the behavior of the action on the border of this fundamental domain $[0, 1] \times [0, 1]$. We see that the left side of this square is mapped to the right side by a and that the right side is mapped to the left side with a^{-1}. In the same way the lower side is mapped to the upper side with b and the upper side is mapped to the lower side with b^{-1}. Therefore it is a natural choice to label the sides of the squares as in Figure III.1.

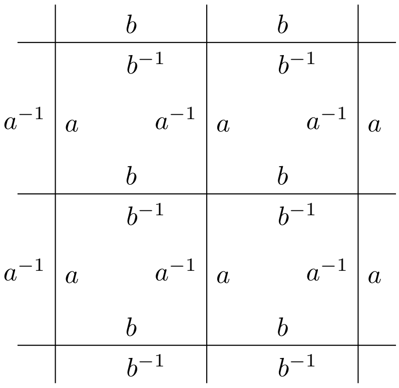


Figure III.1.: The square tessellation with a, b-labeling.

SUBSECTION III.1.1 Cutting Sequences

Consider the square tessellation, it divides the Euclidean plane in labeled squares. If we follow a curve through the plane, we meet sides of these squares. As we have labeled these sides, we obtain a sequence in a, b, a^{-1}, b^{-1}. Note that every side has two labels, as it is the side of two squares. Hence orientation is important, it does not matter which label we choose as long as we choose consistently. Once we have chosen a direction for the curve, we can consistently pick the labels we encounter entering a square or the labels encountered leaving a square. We now agree to choose the labels leaving the squares. The sequence obtained in this way is called the *cutting sequence* of the curve. In the following we will examine cutting sequences of straight lines, it is possible to determine precisely which sequences in a, b, a^{-1}, b^{-1} can occur as such a cutting sequence.

Consider a straight line with positive slope, it can be seen as going from the lower left-quadrant to the upper right-quadrant or the other way around. In the first case we will only encounter a's and b's and in the second case only a^{-1}'s and b^{-1}'s. We now restrict our attention to lines going from left to right. Every result we obtain on these lines can be translated to a result on lines going from right to left by simply replacing a, b by a^{-1}, b^{-1}.

It is possible that a line cuts a vertex at some point, in this case it is not clear what to record. If this happens we can avoid the vertices by moving the line a little bit to the left or right. If we keep the original line and the translated line close together, the lines cut the same sides of the square tessellation, except at the places where the original line meets a vertex. The translated line is moved a little bit away from the vertex and we record ab if we translate to the right and ba if we translate to the left. Hence it seems natural to record either ab or ba if a line meets a vertex. It does not matter which one we choose, yet we have to be consistent.

Lines with negative slope go from either the upper left-quadrant to the lower right-quadrant, in which case we encounter only a's and b^{-1}'s, or from the lower right-quadrant to the upper left-quadrant and then we only encounter a^{-1}s and b's. We now again restrict our attention to the first case, i.e. lines going from left to right. If we encounter a vertex, we denote this by ab^{-1} or $b^{-1}a$.

The only lines we haven't considered yet are horizontal and vertical lines. In the first case we get a constant a-sequence or a constant a^{-1}-sequence, depending on direction. Similarly, in the second case we get a constant b-sequence or a constant b^{-1}-sequence.

We will now restrict ourselves to lines with positive slope, and we will determine precisely which sequences of a and b can occur as cutting sequences. This is not a restriction, every result we obtain on lines with positive slope can be translated into a result on lines with negative slopes by replacing b by b^{-1}. Hence in the end we have determined which sequences in two symbols can occur as cutting sequences of straight lines. Depending on orientation and slope these two symbols can be $a, b, \quad a^{-1}, b, \quad a, b^{-1}$ or a^{-1}, b^{-1}. Multiple consecutive occurrences of the same symbol will from now on be denoted by a superscript, for example b^3 means bbb.

Example III.1. For $n \in \mathbb{Z}_{>0}$ consider the line $l_n : y = nx$ of slope n. We want to determine its cutting sequence. As l_n goes through the origin $(0, 0)$ we record a ab there. After that we cut $n - 1$ times a b-side before we go through the vertex $(1, n)$. This pattern repeats itself: the only vertices we encounter are (k, kn) for every $k \in \mathbb{Z}$ and in between any consecutive two of these vertices we record $n - 1$ times a b. We see that we get the sequence $...ab\, b^{n-1}\, ab... = ...ab^n ab^n...$, thus the cutting sequence is periodic with period ab^n. In the same way we can show that for $n \in \mathbb{Z}_{>0}$ the line $y = \frac{1}{n}x$ has periodic cutting sequence with period $a^n b$.

Consider a line l with positive slope $\lambda > 0$. If $\lambda > 1$ we can never have two consecutive a's in the cutting sequence of l and if $\lambda < 1$ we can never have two consecutive b's. So in the first case a is isolated and in the second case b is. This also works the other way around, if we have two consecutive a's in the cutting sequence then $\lambda < 1$ and if bb occurs in the cutting sequence, then $\lambda > 1$. Note that this implies that the cutting sequence of a line with slope $\lambda = 1$ must have periodic cutting sequence with period ab. We already showed this directly in Example III.1.

Lemma III.1. *Consider the cutting sequence of a line l with slope λ. If $\lambda > 1$, then between any two occurrences of a there are $\lfloor \lambda \rfloor$ or $\lfloor \lambda \rfloor + 1$ occurrences of b.*

Proof. Suppose there are $n \leq \lfloor \lambda \rfloor - 1$ occurrences of b between two a's. This means that when going one square to the right the line l goes less than $n + 1$ up. So the slope of l is less than $n + 1 = \lfloor \lambda \rfloor$ which cannot happen.

If there are $n \geq \lfloor \lambda \rfloor + 1$ occurrences of b between two a's, then going one square to the right we go more than n up. This implies that $\lambda > n \geq \lfloor \lambda \rfloor + 1$. This contradicts the fact that $\lambda < \lfloor \lambda \rfloor + 1$. Hence we see that there can only be $\lfloor \lambda \rfloor$ or $\lfloor \lambda \rfloor + 1$ occurrences of b between any two occurrences of a. □

In the same way we can prove that if $\lambda < 1$, then between any two occurrences of b there are $\lfloor \lambda^{-1} \rfloor$ or $\lfloor \lambda^{-1} \rfloor + 1$ occurrences of a. Consider $n \in \mathbb{Z}_{>0}$, we will call any sequence in two symbols *almost constant with value* n, if one of the symbols is isolated and if between every two occurrences of this symbol there are n or $n + 1$ occurrences of the other symbol. So, the cutting sequence of a line with slope $\lambda > 0$ is almost constant with value $\lfloor \lambda \rfloor$ if $\lambda > 1$ and almost constant with value $\lfloor \lambda^{-1} \rfloor$ if $\lambda < 1$.

Consider an almost constant sequence with value n in the symbols a, b and let a be isolated. After every occurrence of a there are n or $n + 1$ occurrences of b, hence we can also see the sequence as a sequence in $a' = ab^n$ and b. This sequence of a' and b is called the *derived sequence* of the original sequence. One can wonder whether this sequence is again almost constant, this turns out to be the case for the derived sequence of a cutting sequence.

Lemma III.2. *Consider a line l with slope $\lambda > 1$ and cutting sequence s_l. If $\lambda \in \mathbb{Z}$, then the derived sequence of s_l is constant and if $\lambda = [\lambda_0, \lambda_1, ...]$ with $\lambda_1 \geq 1$, then the derived sequence of s_l is almost constant with value λ_1.*

Proof. We get the cutting sequence s_l by looking at l relative to the square tessellation obtained by the maps a and b. For the derived sequence we now consider $a' = ab^{\lambda_0}$ instead of a. So we look at l relative to another tessellation of $\mathbb{R}^2$. This tessellation is obtained from $a' : (x, y) \mapsto (x + 1, y + \lambda_0)$ and $b : (x, y) \mapsto (x, y + 1)$, it consists of parallelograms. We can obtain the grid of this tessellation by a linear map on the square grid. This map has to take $(1, 0)$ to $(1, \lambda_0)$ and $(0, 1)$ to $(0, 1)$. Hence it is given precisely by the matrix

$$M = \begin{pmatrix} 1 & 0 \\ \lambda_0 & 1 \end{pmatrix}.$$

The derived sequence of s_l is thus itself a cutting sequence, namely the cutting sequence of l relative to this new tessellation. This is the same as the cutting sequence of $M^{-1}l$ relative to our original tessellation. For some $\alpha \in \mathbb{R}$ we know that $l = \{(x, \lambda x + \alpha) | x \in \mathbb{R}\}$, hence $M^{-1}l$ is the collection of all points

$$M^{-1} \begin{pmatrix} x \\ \lambda x + \alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\lambda_0 & 1 \end{pmatrix} \begin{pmatrix} x \\ \lambda x + \alpha \end{pmatrix} = \begin{pmatrix} x \\ (\lambda - \lambda_0)x + \alpha \end{pmatrix}.$$

So $M^{-1}l$ has slope $\lambda - \lambda_0$. If $\lambda = \lambda_0$, then $M^{-1}l$ has constant cutting sequence. Thus in this case the derived sequence of s_l is constant. If $\lambda > \lambda_0$, then the derived sequence is the cutting sequence of a line with slope $0 < \lambda - \lambda_0 < 1$. We now see that the derived sequence is almost constant with value $\lfloor (\lambda - \lambda_0)^{-1} \rfloor$. As $\lambda - \lambda_0 = [0, \lambda_1, \lambda_2, ...]$ it follows from Example I.4 that $(\lambda - \lambda_0)^{-1} = [\lambda_1, \lambda_2, ...]$. We conclude that the derived sequence has value λ_1. □

We can now repeat the argument of the previous lemma. If $\lambda_2 = 0$, then deriving s_l twice leads to a constant sequence. If $\lambda_2 > 0$, then it leads to an almost constant sequence of value λ_2. We will call an almost constant sequence *characteristic* if it either becomes constant after a finite number of derivations or can be derived arbitrarily many times. Repeating the argument of Lemma III.2 shows that cutting sequences are characteristic. Furthermore we see that if a line has slope $[\lambda_0, \lambda_1, \lambda_2 ...]$, then the values of the successive derived sequences are $\lambda_1, \lambda_2, ...$ Hence the cutting sequence of a line becomes constant after a finite number of derivations if and only if the slope is rational. With Theorem III.1 we have determined precisely which sequences can occur as cutting sequences.

Theorem III.1. *A sequence in two symbols is a cutting sequence if and only if it is characteristic. A cutting sequence coming from a line with slope* $[\lambda_0, \lambda_1, \lambda_2, ...]$ *corresponds precisely to a characteristic sequence of value* λ_0 *with the successive derived sequences having* $\lambda_1, \lambda_2, ...$ *as values.*

Proof. We already saw that a cutting sequence coming from a line with slope $[\lambda_0, \lambda_1, \lambda_2, ...]$ is a characteristic sequence of value λ_0 and that the successive derived sequences have $\lambda_1, \lambda_2, ...$ as their values. To prove the other implication, consider a characteristic sequence of value λ_0 with successive derived sequences having $\lambda_1, \lambda_2, ...$ as values. We will prove that there is precisely one such sequence. Say the sequence is in the symbols a and b and let a be the isolated one. The value of the sequence is λ_0 and hence there are λ_0 or $\lambda_0 + 1$ occurrences of b between every two occurrences of a. This implies that the sequence is made up of ab^{λ_0} and b such that the occurrences of b are isolated. As the derived sequence has value λ_1 there must be λ_1 or $\lambda_1 + 1$ occurrences of ab^{λ_0} between any two occurrences of b. This tells us that the sequence is made up of $(ab^{\lambda_0})^{\lambda_1}b$ and $(ab^{\lambda_0})^{\lambda_1+1}b$. We can continue this process: as the next derived sequence has value λ_2 we know that $(ab^{\lambda_0})^{\lambda_1+1}b$ is isolated and that between every two occurrences of it there are λ_2 or $\lambda_2 + 1$ occurrences of $(ab^{\lambda_0})^{\lambda_1}b$. This process completely determines in which order the ab^{λ_0} and b can occur in the sequence. Consequently there is precisely one such sequence. Yet, any line with slope $[\lambda_0, \lambda_1, ...]$ has this sequence as its cutting sequence and hence every characteristic sequence is a cutting sequence. □

Corollary III.1. Given $\lambda \in \mathbb{R}_{>0}$ every two lines with slope λ have the same cutting sequence. Furthermore, the cutting sequence of any line is symmetric.

Proof. We just saw that a cutting sequence coming from a line with slope $\lambda = [\lambda_0, \lambda_1, \lambda_2, ...]$ is a characteristic sequence of value λ_0 with the successive derived sequences having values $\lambda_1, \lambda_2, ...$ Yet, the above proof shows us that there is only one such characteristic sequence. Hence two lines with the same slope give equal cutting sequences.

To prove that a cutting sequence is symmetric, consider a line l and its cutting sequence s_l. Directing l from left to right gives us s_l. Directing l from right to left gives a cutting sequence in a^{-1} and b^{-1}. If we replace a^{-1} with a and b^{-1} with b we obtain the reversal of s_l. We want to show that this is actually s_l itself.

We now rotate the plane 180°, the reversed cutting sequence is precisely the cutting sequence of the image of l under this rotation. But l and its image have the same slope and hence the same cutting sequence. We conclude that the reversal of s_l is again s_l. This proves that s_l is symmetric. □

In particular we see that a line and any image of this line under the action of $\mathbb{Z}^2$ have the same cutting sequence. This means that instead of looking at lines in the plane, we can also consider projections of these lines onto the fundamental domain. The opposite sides of the fundamental domain are identified via the maps a and b, in this way we obtain a torus T. Hence we now consider projections of lines on the torus, we can obtain a cutting sequence of such a line by looking at the cutting sequence of any of its lift to the plane. As all of the lifts have the same cutting sequence this gives a well-defined way to consider the cutting sequence of a projected line.

In Chapter V we need to link characteristic sequences to Markoff balanced sequences. Consider a characteristic sequence in a, b with value n and a isolated. The b's occur in n-tuples or in $(n+1)$-tuples and hence we can associate a sequence in n and $n+1$ to the characteristic sequence. The next lemma shows that this sequence is Markoff balanced.

Lemma III.3. *The sequence in n and $n+1$ associated to a characteristic sequence of value n is Markoff balanced, i.e. if we write the sequence as $\ldots, n_{-1}, n_0, n_1, \ldots$ then*

(i) $n_i - n_j = -1, 0, 1$ *for every $i, j \in \mathbb{Z}$;*

(ii) If $n_{i+1} - n_i = -1, 1$ respectively, then the first $n_{i+j+1} - n_{i-j}$ for $j = 1, 2, \ldots$ that is non-zero equals 1, -1 respectively.

Proof. The first condition certainly holds as we have $n_i = n$ or $n_i = n+1$ for every $i \in \mathbb{Z}$. To prove the second condition, suppose $n_{i+1} - n_i = 1$ and that the first $n_{i+j+1} - n_{i-j}$ for $j = 1, 2, \ldots$ that is non-zero is 1, too. We are going to show that this leads to a contradiction. We have $n_i = n$ and $n_{i+1} = n+1$ and because $n+1$ is isolated we thus have a pattern

$$n+1, n, \ldots, n_i = n, n_{i+1} = n+1, n, \ldots, n, n+1$$

We have assumed that the first non-zero difference $n_{i+j+1} - n_{i-j}$ is 1. This cannot happen in the above pattern, for then the last $n+1$ in the pattern would be grouped with the first n in the pattern. This can only happen if the pattern looks like $n+1, n_{m+2}, n+1, n_m, n+1$ for some $m > 0$. This contradicts the definition of a characteristic sequence: between every two $(n+1)$'s there are m or $m+1$ occurrences of n for some $m \geq 0$. As the first non-zero $n_{i+j+1} - n_{i-j}$ is 1, we now know that there is some $p > 0$ such that the $n, n+1$-sequence looks like

$$\ldots, n_{m+1}, n+1, (n_m, n+1)_p, n_{m+1}, n_{i+1} = n+1, (n_m, n+1)_{p+1}, n_m, n+1, \ldots$$

Hence there is a pattern $n_{m+1}, n+1, (n_m, n+1)_p, n_{m+1}, n_{i+1} = n+1, (n_m, n+1)_{p+2}$. Again this contradict the definition of a characteristic sequence, between every two occurrences of $m+1$ there are p or $p+1$ occurrences of m for some $p \geq 0$. The case that both $n_{i+1} - n_i$ and the first non-zero difference $n_{i+j+1} - n_{i-j}$ are -1 can be treated in a similar way. □

It can easily be seen that an almost constant sequence of value n is characteristic if and only if its associated sequence in n and $n+1$ is characteristic. Indeed, suppose a sequence s is characteristic. We can derive it and the derived sequence will basically have the same structure as the associated sequence in $n, n+1$. If the derived sequence is an almost constant of value k, then the associated sequence is almost constant of value $k-1$. More precisely, if the derived sequence is a sequence A and B with B isolated, then we get the associated sequence by replacing A^k by A^{k-1}, A^{k+1} by A^k and after that we replace every occurrence of A by n and every occurrence of B by $n+1$. It follows immediately that the derived sequence of s, and hence also s itself, is characteristic if and only if the associated sequence of s is characteristic.

Lemma III.3 now states that a characteristic sequence in n and $n+1$ for some $n \geq 0$ is Markoff balanced if it is characteristic. The converse statement can easily be proven, if we restrict ourselves to purely periodic sequences.

Lemma III.4. *A purely periodic, Markoff balanced sequence is characteristic.*

Proof. Consider such a purely periodic, Markoff balanced sequence s. There is an $n \geq 0$ such that s is a sequence in n and $n \pm 1$ with $n \pm 1$ isolated. As s is purely periodic, it will become constant after a finite number of derivations. Of course, a constant sequence is characteristic. Consequently we can prove that s is characteristic by proving that the *anti-derivatives* of a characteristic sequence are again characteristic.

Consider a Markoff balanced sequence $\bar{n} = ..., n_{-1}, n_0, n_1, ...$, we will define its anti-derivatives to be sequences of the form $..., m \pm 1, m_{n_{-1}}, m \pm 1, m_{n_0}, m \pm 1, m_{n_1}, m \pm 1, ...$ for some $m \in \mathbb{Z}_{\geq 0}$. Note that deriving any sequence of this form will lead to the sequence $\bar{n}$, hence the name anti-derivative. If we can prove that the anti-derivatives of a characteristic sequence are again characteristic, then we can work backwards from the characteristic constant sequence to the sequence s.

Suppose that the sequence $n = ..., n_{-1}, n_0, n_1, ...$ in $n, n+1$ is characteristic and consider one of its anti-derivatives $\bar{m} = ..., m \pm 1, m_{n_{-1}}, m \pm 1, m_{n_0}, m \pm 1, m_{n_1}, m \pm 1, ...$, note that $\bar{m}$ is almost constant of value n and hence we can associate a sequence in $n, n+1$ to it in the same way we did above. This associated sequence is precisely $\bar{n}$ which we know to be characteristic. Yet, we have just seen that a sequence is characteristic if and only if its associated sequence is characteristic. Consequently we now see that the sequence $\bar{m}$ is characteristic. □

SECTION III.2 Farey Tessellation of the Hyperbolic Plane

In this section we will consider another tessellation, this time of the hyperbolic upper half-plane

$$\mathbb{H} = \{z \in \mathbb{C} | \Im(z) > 0\}.$$

First we will shortly discuss some properties of $\mathbb{H}$. After that we introduce the Farey tessellation, this tessellation consists of triangles rather than squares.

SUBSECTION III.2.1 The Upper Half-Plane

The upper half-plane $\mathbb{H}$ can be made into a metric space with the *Poincaré metric* $ds^2 = \frac{dx^2+dy^2}{y^2}$. The metric space $\mathbb{H}$ then becomes a model for *hyperbolic geometry*, a special type of *non-Euclidean geometry*. With *geodesics* of a metric space we mean curves such that shortest paths between two points in the space are segments of these curves. For example, given two points in $\mathbb{R}^2$ the shortest path between them is a segment of a straight line. Hence in the Euclidean plane it holds that straight lines are precisely the geodesics. For $\mathbb{H}$ it is the case that geodesics are half-lines perpendicular to the real axis or semi-circles with their center on the real line. Note that for the maps a, b considered in the previous section it holds that they send straight lines to straight lines. We will be interested in similar maps $\mathbb{H} \to \mathbb{H}$, i.e maps that have the property that the image of a geodesic under such a map is again a geodesic. We can use the group $\mathrm{SL}(2, \mathbb{Z})$ to find such maps.

$\mathrm{SL}(2, \mathbb{Z})$ acts on the hyperbolic upper half-plane $\mathbb{H}$ is the following way. Given a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z})$ and a point $z \in \mathbb{H}$ we define

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

For this to be an action we need that this is again a point in $\mathbb{H}$, to prove this we really need the fact that $ad - bc = 1$:

$$\Im\left(\frac{az+b}{cz+d}\right) = \frac{\Im((az+b)(c\bar{z}+d))}{|cz+d|^2} = \frac{\Im(acz\bar{z}+bd+adz+bc\bar{z})}{|cz+d|^2} = \frac{\Im((ad-bc)z)}{|cz+d|^2} = \frac{\Im(z)}{|cz+d|^2} > 0.$$

Hence to a matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z})$ we can associate a map $M : \mathbb{H} \to \mathbb{H}$, $z \mapsto \frac{az+b}{cz+d}$. The natural way to extend this with ∞ is $M : \infty \mapsto \frac{a}{c}$ if $c \neq 0$ and $M : \infty \mapsto \infty$ if $c = 0$. These maps are examples of *Möbius transformations*, it can be shown that they send geodesics to geodesics.

We will now introduce a tessellation of $\mathbb{H}$ in triangles. Note that in $\mathbb{R}^2$ a triangle has three sides which are segments of straight lines, in the same way triangles in $\mathbb{H}$ have sides that are segments of geodesics.

It is well-known that the standard fundamental domain of the action of SL$(2,\mathbb{Z})$ on $\mathbb{H}$ consists of all $z \in \mathbb{C}$ with $|\Re(z)| \leq \frac{1}{2}$ and $|z| \geq 1$. We are going to use this domain to construct one of the triangles of the tessellation, this triangle will actually consists of three fundamental domains of the action of SL$(2,\mathbb{Z})$ on $\mathbb{H}$.

To construct the desired triangle, note that the imaginary axis divides the fundamental domain in two halves. Call the left half L and the right half R. In Chapter I we came across the matrix $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in$ SL$(2,\mathbb{Z})$ and it holds that $F = R \cup T(L)$ is an alternative fundamental domain of the action of SL$(2,\mathbb{Z})$ on $\mathbb{H}$, it is a quadrilateral with vertices $i, i+1, \rho = \frac{1}{2} + \sqrt{\frac{3}{4}}i$ and ∞. Consider the matrix $S = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \in$ SL$(2,\mathbb{Z})$, it corresponds to the map $S : z \mapsto \frac{-1}{z-1}$. We can compute that

$$S(i) = \frac{1}{2} + \frac{1}{2}i, \;\; S(i+1) = i, \;\; S(\rho) = \rho, \;\; \text{and} \;\; S(\infty) = 0.$$

Hence $S(F)$ is a quadrilateral with vertices $\frac{1}{2} + \frac{1}{2}i, i, \rho, 0$. We know that S sends geodesics to geodesics, so as the sides of F are parts of geodesics $S(F)$ must look like in Figure III.2.

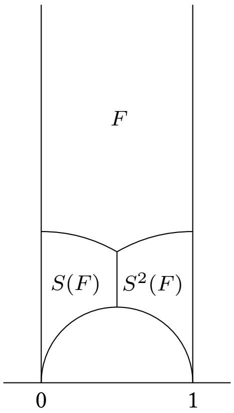


Figure III.2.: The fundamental domain F

In a similar way we can compute that

$$S^2(i) = i+1, \;\; S^2(i+1) = \frac{1}{2} + \frac{1}{2}i, \;\; S^2(\rho) = \rho, \;\; \text{and} \;\; S^2(\infty) = 1.$$

So $S^2(F)$ is precisely the quadrilateral in Figure III.2. We see that $D := F \cup S(F) \cup S^2(F)$ is a triangle with vertices 0, 1 and ∞. We call such a triangle *ideal*, its vertices all lie either at ∞ or on the real line. This triangle D can be used to construct a tessellation of $\mathbb{H}$.

Consider all images of D under matrices $M \in$ SL$(2,\mathbb{Z})$. First note that these images cover the whole plane $\mathbb{H}$, because the images under SL$(2,\mathbb{Z})$ of the standard fundamental domain already cover the whole plane. Furthermore, note that two matrices $M, N \in$ SL$(2,\mathbb{Z})$ map D to the same image if and only if either $M = NS$ or $M = NS^2$. This shows that two images that are not equal can only overlap on the boundaries. Hence the images of D under SL$(2,\mathbb{Z})$ make up a tessellation of $\mathbb{H}$ in ideal triangles. We will call this tessellation of $\mathbb{H}$ the *Farey tessellation*, this name will be explained below.

Note that D is made up of three fundamental domains of the action of SL$(2,\mathbb{Z})$ on $\mathbb{H}$, as such D is not a fundamental domain of this action. It is however a fundamental domain of the same action on $\mathbb{H}$ restricted to

$$\Gamma_0(2) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{SL}(2,\mathbb{Z}) | c \equiv 0 \mod 2 \right\}.$$

This follows from the fact that $\mathrm{SL}(2,\mathbb{Z}) = \Gamma_0(2) \cup S\,\Gamma_0(2) \cup S^2\,\Gamma_0(2)$, which can easily be deduced from $\mathrm{SL}(2,\mathbb{Z})$ being generated by S and $T \in \Gamma_0(2)$.

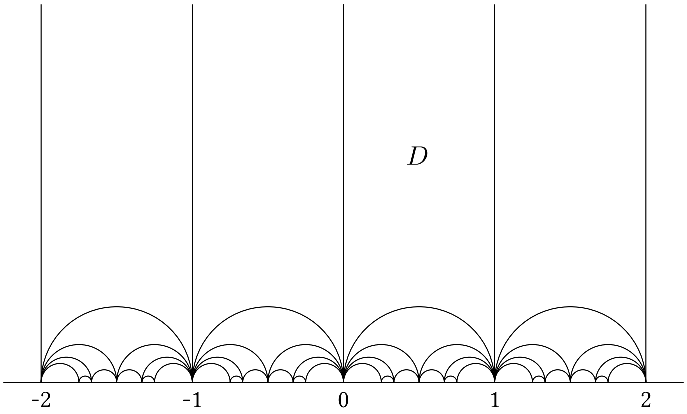


Figure III.3.: The Farey tesselation

We claim that a geodesic is a side of one of the triangles in the Farey tessellation if and only if this geodesic is the image of the imaginary axis under some $M \in \mathrm{SL}(2,\mathbb{Z})$. Indeed, S sends the imaginary axis to the semi-circle joining 0 and 1 and S^2 sends the imaginary axis to the vertical line through 1. Hence all the sides of D can be obtained as images of the imaginary axis and all the other sides in the Farey tessellation are just $\mathrm{SL}(2,\mathbb{Z})$-images of these three sides.

Lemma III.5. *The set of all vertices in the Farey tessellation is* $\mathbb{Q} \cup \{\infty\}$ *and two fractions* $\frac{p}{q}$ *and* $\frac{r}{s}$ *in their lowest terms are joined by a side of the Farey tessellation if and only if* $ps - qr = \pm 1$.

Proof. Note that a real number is a vertex of one of the triangles in the Farey tessellation if and only if it is the image of 0 or ∞ under some $M \in \mathrm{SL}(2,\mathbb{Z})$. These images all lie either at ∞ or are rational. To show that all rational numbers are vertices, consider $\frac{p}{q} \in \mathbb{Q}$ in lowest terms. We then have $\gcd(p,q) = 1$ and hence there are $r, s \in \mathbb{Z}$ such that $ps - qr = 1$. This implies that $\frac{p}{q}$ is the image of ∞ under $\begin{pmatrix} p & r \\ q & s \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$. So every rational number is a vertex and hence the set of all vertices is precisely $\mathbb{Q} \cup \{\infty\}$.

Next we want to prove that two fractions $\frac{p}{q}$ and $\frac{r}{s}$ in their lowest terms are joined by a side of the Farey tessellation if and only if $ps - qr = \pm 1$. If $\frac{p}{q}$ and $\frac{r}{s}$ are joined by a side in the Farey tessellation, then they are the image of 0 and ∞ under some $M \in \mathrm{SL}(2,\mathbb{Z})$. Suppose $M(0) = \frac{p}{q}$ and $M(\infty) = \frac{r}{s}$, this gives four possibilities for M:

$$\begin{aligned} M &: z \mapsto \tfrac{rz+p}{sz+q}; \\ M &: z \mapsto \tfrac{-rz-p}{-sz-q}; \\ M &: z \mapsto \tfrac{-rz+p}{-sz+q}; \\ M &: z \mapsto \tfrac{rz-p}{sz-q}. \end{aligned}$$

Computing the determinant leads in the first two cases to $sp - qr = -1$ and in the last two cases to $sp - qr = 1$. This shows that $ps - qr = \pm 1$.

Now suppose that $ps - qr = \pm 1$. If $ps - qr = 1$, then $\begin{pmatrix} p & r \\ q & s \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$. It sends 0 to $\frac{r}{s}$ and ∞ to $\frac{p}{q}$. Hence the geodesic joining $\frac{p}{q}$ and $\frac{r}{s}$ is the image of the imaginary axis under this matrix. This proves that it is one of the sides in the Farey tessellation. If $ps - qr = -1$ we can repeat this argument using the matrix $\begin{pmatrix} r & p \\ s & q \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$. □

Using the above lemma we can explain why the tessellation is called the Farey tessellation, it is because sides in the tessellation can be obtained by joining adjacent points in *Farey series*. For every $n \in \mathbb{Z}_{\geq 1}$ we define the n-th order

Farey series F_n as the set of all rational numbers $\frac{p}{q}$ in lowest terms with $|p|, |q| \leq n$, arranged in increasing order. For example F_1 is $-1, 0, 1$, the Farey series are named after the British geologist John Farey, Sr. It can be proven, see Hardy/Wright, that two fractions $\frac{p}{q}, \frac{r}{s}$ are adjacent in some Farey series if and only if $rq - ps = \pm 1$. The above lemma tells us that this happens if and only if $\frac{p}{q}$ and $\frac{r}{s}$ are joined by a side in the Farey tessellation. Hence the Farey tessellation can also be obtained by drawing vertical lines through all integers and joining all rational that are adjacent in some Farey series. The following lemma will be important in the next section.

Lemma III.6. *For $a, b \in \mathbb{R}$ consider $\mathbb{H}_{a,b} = \{z \in \mathbb{H} | a < \Re(z) < b\}$. For every $a, b \in \mathbb{R}$ and $\alpha > 0$ there is only a finite number of semi-circles of radius bigger than α in $\mathbb{H}_{a,b}$ belonging to the Farey tessellation.*

Proof. We know that $\frac{p}{q}$ and $\frac{r}{s}$ in their lowest terms are connected in the Farey tessellation if and only if $ps - rq = \pm 1$. If they are connected, then they are the endpoints of a semi-circle of radius $\frac{1}{2}|\frac{p}{q} - \frac{r}{s}| = \frac{1}{2|qs|}$. Now, if this radius is bigger that α, then this implies $2|qs| < \alpha^{-1}$. There are only finitely many solutions $(q, s) \in \mathbb{Z}^2$ to this inequality. For such a solution (q, s) there are only finitely many p, r such that $a < \frac{p}{q}, \frac{r}{s} < b$. So there are only finitely many $\frac{p}{q}$ and $\frac{r}{s}$ that are the endpoints of a semi-circle with radius bigger than α in $\mathbb{H}_{a,b}$ belonging to the Farey tessellation. □

SUBSECTION III.2.3 Cutting Sequences

In the Euclidean plane we considered cutting sequences of straight lines, the geodesics of $\mathbb{R}^2$. So in $\mathbb{H}$ it is natural to consider cutting sequences of the geodesics there, i.e. vertical lines or semi-circles centered on $\mathbb{R}$. Unlike in the square tessellation we will not label the sides of the triangles of the Farey tessellation. Instead we will 'label' the vertices. Again we will work with directed geodesics. Consider such a geodesic, if it goes through a triangle of the Farey tessellation, then it cuts two of the three sides. These two sides meet in a vertex of the triangle. As the geodesic is directed, this vertex lies either to the left or to the right of the geodesic. We will label the vertex L or R respectively. Following the geodesic through the hyperbolic plane we can get a cutting sequence in R and L, we will call this sequence the L, R-sequence. Note that all the vertices lie at ∞ or in $\mathbb{Q}$, so if a geodesic cuts a triangle in a vertex, then this vertex is either the starting-point or the endpoint of the geodesic. We can record either an L or an R here, it does not matter as long as we're consistent. Also note that the L, R-sequence of a geodesic is doubly infinite if and only if the geodesic neither starts nor ends in a vertex.

Under the action of $\mathrm{SL}(2, \mathbb{Z})$ on $\mathbb{H}$ we know that geodesics are send to geodesics. As the map $M : \mathbb{H} \to \mathbb{H}$ belonging to some $M \in \mathrm{SL}(2, \mathbb{Z})$ is orientation-preserving, it follows that two geodesics are in the same orbit if and only if they have the same L, R-sequence. Hence in this case we can also consider geodesics projected on the triangle D, instead of geodesics in the hyperbolic plane.

The L, R-sequences of a geodesic is related to the continued fraction expansion of its endpoint. This is specified in Theorem III.2.

Theorem III.2. *For $a \in \mathbb{R}_{>0}$ consider (part of) a geodesic joining any point $b \neq (0, 0)$ on the imaginary axis with a. Directing this geodesic segment γ_a from b to a we can read off the L, R-sequence $L^{a_0} R^{a_1} L^{a_2}...$, with possibly $a_0 = 0$, and this sequence has the property that $[a_0, a_1, a_2, ...] = a$.*

Proof. Figure III.4 illustrates the proof given below for $a = 3\frac{3}{7} = [3, 2, 3]$. It might be clarifying to have a look at this figure while reading the proof.

First suppose that a is an integer. We see that we go through $a - 1$ vertical lines before we meet the vertex a. Going through a vertical line gives an L and meeting the vertex a gives either an L or an R. In the first case we find that $a_0 = a$ and $a_n = 0$ for all $n > 0$, so $[a_0, a_1, ...] = [a] = a$. In the second case we find that $a_0 = a - 1$, $a_1 = 1$ and $a_n = 0$ for all $n > 1$. This gives $[a_0, a_1, ...] = [a - 1, 1] = a$. So the theorem holds for $a \in \mathbb{Z}$. Next suppose that a is not an integer. If $a < 1$, then the first side we cut is the side joining 0 to 1. The vertex we cut off lies at the right of the geodesic, and hence the L, R-sequence starts with an R. This gives that $a_0 = 0 = \lfloor a \rfloor$ which also is the zeroth partial quotient of the continued fraction expansion of a.

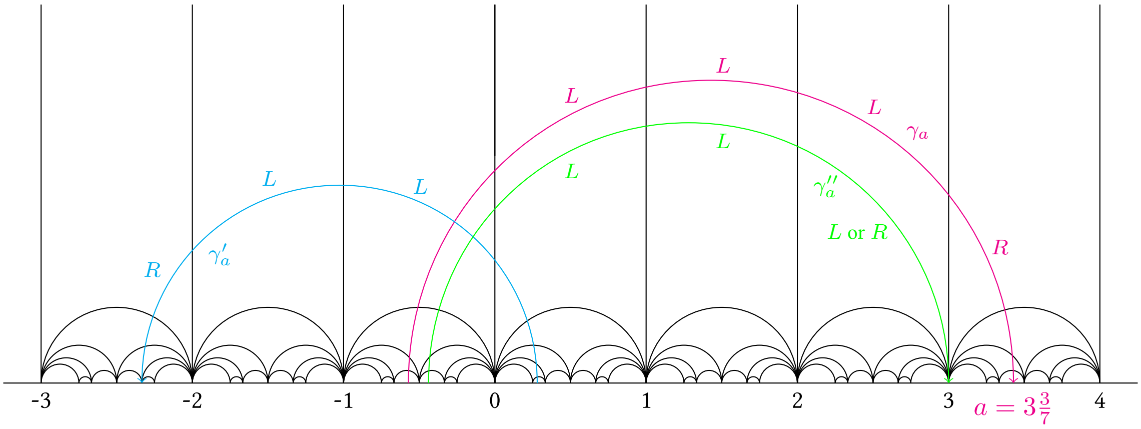


Figure III.4.: Illustration of the proof of Theorem III.2 for $a = 3\frac{3}{7} = [3, 2, 3]$

If $a > 1$, then we go through $\lfloor a \rfloor$ vertical lines before we go through a semi-circle. As long as we go through vertical lines we record L and the first time we go through a semi-circle we record an R. Hence we see that $a_0 = \lfloor a \rfloor$, the zeroth partial quotient of a.

To prove that a_1 coincides with the first partial quotient of a we consider the image γ_a' of γ_a under the map $M_1 : z \mapsto \frac{-1}{z - a_0}$. Note that the matrix M_1 corresponding to this map belongs to $\mathrm{SL}(2, \mathbb{Z})$ and hence the L, R-sequences of γ_a and γ_a' agree. Consider the point c where γ_a meets the vertical line through a_0. Then in particular the L, R-sequences of the geodesic segment from c to a, which equals $R^{a_1} L^{a_2} \ldots$, is the same as the L, R-sequence of the segment from $M_1(c)$ to $M_1(a)$.

As c lies on the vertical line through a_0 we have that $M_1(c)$ lies on the imaginary axis. Furthermore it holds that $M_1(a) = \frac{-1}{a - a_0} < -1$. If $M_1(a)$ is an integer, then the geodesic segment from $M_1(c)$ to $M_1(a)$ first cuts $\frac{1}{a - a_0} - 1$ vertical lines and then it means the vertex $\frac{1}{a - a_0}$. Going through vertical lines we record an R and meeting the vertex we record either an L or an R. In the first case we have that $a_1 = \frac{1}{a - a_0} - 1$, $a_2 = 1$ and $a_n = 0$ for all $n > 2$. This leads to $[a_0, a_1, a_2, \ldots] = \left[a_0, \frac{1}{a - a_0} - 1, 1\right] = a$. In the second case we find that $a_1 = \frac{1}{a - a_0}$ and $a_n = 0$ for all $n > 1$, so $[a_0, a_1, \ldots] = \left[a_0, \frac{1}{a - a_0}\right] = a$.

If $M_1(a) \notin \mathbb{Z}$ this means that the geodesic segment from $M_1(c)$ to $M_1(a)$ first cuts $\left\lfloor \frac{1}{a - a_0} \right\rfloor$ vertical lines after which it cuts a semi-circle. Going through vertical lines we record an R and going through the first semicircle we record an L. So, as the geodesic segment from $M_1(c)$ to $M_1(a)$ has L, R-sequence $R^{a_1}, L^{a_2} \ldots$, we know that $a_1 = \left\lfloor \frac{1}{a - a_0} \right\rfloor$. This is precisely the first partial quotient of a.

We can repeat this argument and apply the map $M_2 : z \mapsto \frac{-1}{z + a_1}$ to γ_a' resulting in a geodesic γ_a''. Using this geodesic we can prove that a_2 is the second partial quotient of a. We can continue in this fashion, a_3 is the third partial quotient of a and so on. If a is a fraction, it has a finite continued fraction expansion. This is in accordance with the fact that the L, R-sequence terminates if the endpoint a of γ_a is rational. Consequently the sequence $a_0, a_1, \ldots$ is finite, say of length $n + 1$, and repeating the above argument we find $[a_0, a_1, a_2, \ldots, a_n] = a$.

If a is irrational, it has an infinite continued fraction expansion and the L, R-sequence of γ_a is infinite. Hence the above argument never terminates. However, in this case we can apply the above procedure to the convergents $\frac{p_n}{q_n}$ of a. We know that a_n is the n-th partial quotient of a for every n, this implies that for every convergent $\frac{p_n}{q_n}$ we have $\frac{p_n}{q_n} = [a_0, \ldots, a_n]$. Theorem I.1 tells us that $\frac{p_n}{q_n} \to a$ as $n \to \infty$ and hence in this case we also have $[a_0, a_1, a_2, \ldots] = a$. □

A,B-Tessellation of the Hyperbolic Plane

In this section we consider yet another tessellation of $\mathbb{H}$. This tessellation will have several similarities to the square tessellation of $\mathbb{R}^2$. Note that we obtained this tessellation by dividing the plane in fundamental domains for the action of $\langle a, b\rangle$ on $\mathbb{R}^2$. Furthermore the opposite sides of the fundamental domain are identified by a and b. We will tessellate the hyperbolic plane in a similar manner, we will divide it in squares that are the fundamental domain for the action of a certain free group $\langle A, B\rangle$ on $\mathbb{H}$ such that the opposite sides of the fundamental domain are identified by the maps A and B.

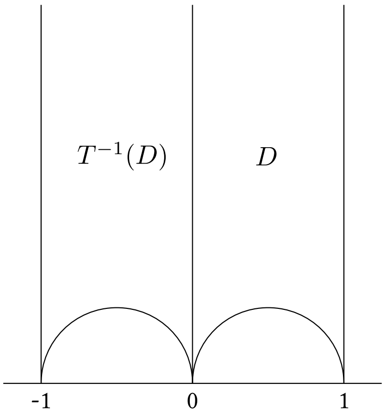


Figure III.5.: The square S is made up of D and $T^{-1}(D)$

Recall the Farey tessellation, two of its triangles are D, with vertices $0, 1, \infty$, and $T^{-1}(D)$, with vertices $-1, 0, \infty$. The union $D \cup T^{-1}(D)$ then makes up a square S with vertices $-1, 0, 1, \infty$, see Figure III.5. We want this square S to be the fundamental domain of the action of some free group $\langle A, B\rangle$ on $\mathbb{H}$. Consider the maps

$$A : z \mapsto \frac{z+1}{z+2} \qquad \text{and} \qquad B : z \mapsto \frac{z-1}{-z+2}.$$

We see that $A(-1) = 0$ and $A(\infty) = 1$, hence A sends the side of S from -1 to ∞ to the side from 0 to 1. Furthermore $B(1) = 0$ and $B(\infty) = -1$, thus B sends the side of S from 1 to ∞ to the side joining 0 to -1. This shows that A and B identify the opposite sides of S. We now claim that S is the fundamental domain for $G = \langle A, B\rangle$ acting on $\mathbb{H}$.

Consider all the images of S under the free group generated by A and B. As A sends the side from -1 to ∞ to the side joining 0 and 1, we see that S and $A(S)$ have the side from 0 to 1 in common. Furthermore, A is orientation preserving, hence the other vertices of $A(S)$ lie between 0 and 1. We can do the same for B, A^{-1} and B^{-1} and we see that S is surrounded by four squares each of which have one side in common with S. Furthermore, the opposite sides of all four these squares are identified by A and B. Hence we can do the same for $A(S)$, $A^{-1}(S)$, $B(S)$ and $B^{-1}(S)$ and we see that they are each surrounded by four squares which all have the property that the opposite sides are identified by A and B. Continuing in this way we find that the images of S under G are either disjoint or share a side and that they fill up the whole hyperbolic plane. Thus S is a fundamental domain for $\langle A, B\rangle$ acting on $\mathbb{H}$. From now on we will used A' to denote A^{-1} and B' to denote B^{-1}.

Consequently the images of S under G tessellate the hyperbolic plane, we will call this tessellation the A, B-*tessellation*. Note that the A, B-tessellation can also be obtained from the Farey tessellation by removing all the G-images of the geodesic joining 0 and ∞. This makes the A, B-tessellation a subtessellation of the Farey tessellation. This might lead one to think that G is a subgroup of $\Gamma_0(2)$, as the Farey tessellation consists of fundamental domains of the action of $\Gamma_0(2)$ on $\mathbb{H}$. This is however not true, both A and B do not belong to $\Gamma_0(2)$.

We want to consider cutting sequences of geodesics relative to the A, B-tessellation. There is a natural way to label the sides of the squares. As A sends the side joining -1 and ∞ to the side from 0 to 1 it is a natural choice to label the side from -1 to ∞ by A and the side joining 0 and 1 by A'. Similarly we label the side from 1 to ∞ by B and the side joining -1 and 0 by B'. Note here that every side has two labels, as it is the side of two different squares.

As opposite sides are identified we see that every side has the labels A and A' or the labels B and B', see Figure III.6.

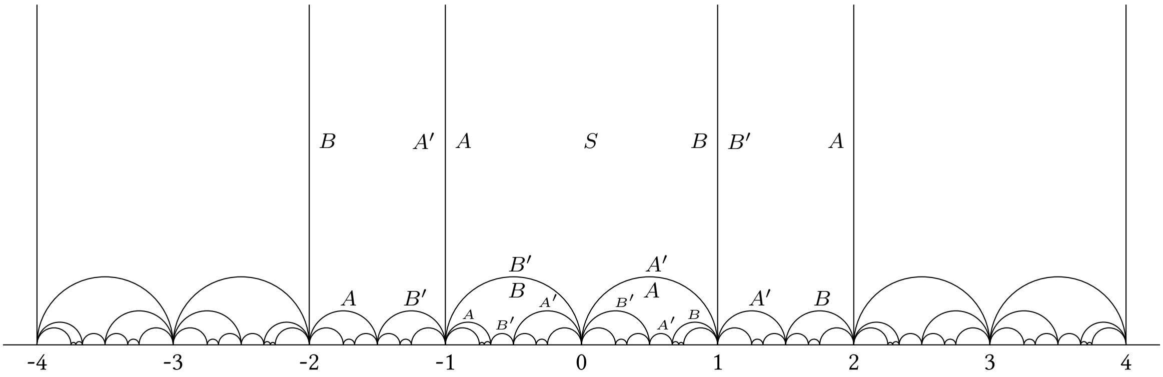


Figure III.6.: The A, B-tesselation

SUBSECTION III.3.1 Cutting Sequences

Just as in the Euclidean case we can now consider cutting sequences of directed geodesics. If a geodesic cuts a side of the tessellation, there are two squares this side belongs to. We now agree to record the label belonging to the side of the square we are leaving. We can wonder what sequences in A, B, A', B' occur as the cutting sequences of the geodesics.

First we are going to prove that every doubly infinite *reduced* sequence in A, B, A', B' that doesn't begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ is the cutting sequence of some geodesic in $\mathbb{H}$. With a reduced sequence we mean a sequence in which a symbol is never immediately followed by its inverse, e.g. AA' will never occur. Consider such a sequence s, starting from our central square S we can make a polygonal path with cutting sequence s. We simply pick a point in S and a starting point in the sequence s. If the first symbol after this starting point is A, then we pick a point in the square adjacent to S, such that connecting the point from S with this point we cut the edge labeled A, and so on. In this way we can construct a polygonal path p_s which cutting sequence is equal to s, see Figure III.7.

Lemma III.7. *Given a doubly infinite reduced sequence s in A, B, A', B' that doesn't begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ we have that the associated polygonal path p_s converges to two points on the real line.*

Proof. Note that s is a reduced sequence, this implies that going from one square to an adjacent square, we do not go back to the square we came from in the next step. So, once we have entered a half-disk, we never leave this half-disk

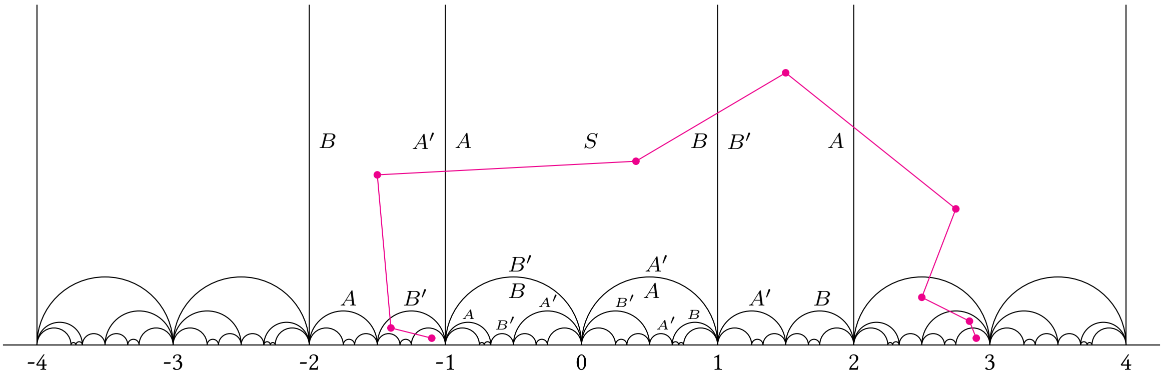


Figure III.7.: Example of a polygonal path

again. The only way to avoid entering a half-disk is an infinite repetition of either $ABA'B'$ or $BAB'A'$. Since we have excluded this, we know that p_s enters a half-disk at some point. In the next step we enter a half-disk with smaller radius, and continuing this we get an infinite sequence of half-disks. If we can show that the associated sequence of radii tends to zero, we are done. Realize that the A, B-tessellation is a subtessellation of the Farey tessellation. We know that the sequence of radii is a strictly decreasing sequence bounded from below by zero. This means that it must have a limit $\alpha \geq 0$. Suppose $\alpha > 0$ and take a and b as the endpoints of the first half-disk we enter. By Lemma III.6 we know that in $\mathbb{H}_{a,b}$ there are only finitely many half-disks of the tessellation with radius bigger than α. But our infinite sequence of half-disks supplies us with an infinite number of those. We have reached a contradiction and thus we conclude that $\alpha = 0$. Since s is doubly infinite, we can apply this procedure twice to get two limit points $l_1, l_2 \in \mathbb{R}$. □

We are going to prove two properties of l_1 and l_2, namely that they do not belong to the vertex-set of the A, B-tessellation and that $l_1 \neq l_2$.

Lemma III.8. *Consider a doubly infinite reduced sequence s in A, B, A', B' that doesn't begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ and the associated polygonal path p_s. The two limit points l_1, l_2 of p_s do not belong to the vertex-set of the A, B-tessellation.*

Proof. Suppose l_1 does belong to the vertex-set of the A, B-tessellation. After a finite number of steps we have reached a square with l_1 as a vertex. Because s is doubly infinite, we have to leave this square in the next step. But l_1 is a limit point, so l_1 remains a vertex of all the squares we reach after this point. The only way to do this is to end in an infinite repetition of $ABA'B'$ or $BAB'A'$. But this was excluded, so we can conclude that l_1 is not a limit point. The proof for l_2 is exactly the same. □

Lemma III.9. *Consider a doubly infinite reduced sequence s in A, B, A', B' that doesn't begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ and the associated polygonal path p_s. For the two limit points l_1, l_2 of p_s we have $l_1 \neq l_2$.*

Proof. Once we are in a half-disk, we know that the limit point must lie strictly between the endpoints of the semicircle bounding this half-disk. The strictness comes from Lemma III.8. So once the two ends of p_s end up in different half-disks, the limit points must be different. That the two ends of p_s cannot enter the same half-disk simply follows from the fact that s is reduced. □

We are now ready to prove that s is the cutting sequence of some geodesic in $\mathbb{H}$. More precisely, we are going to prove that the geodesic γ_s connecting l_1 and l_2 has cutting sequence s.

Corollary III.2. Consider a doubly infinite reduced sequence s in A, B, A', B' that doesn't begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ and the associated polygonal path p_s with endpoints l_1 and l_2. The geodesic γ_s joining l_1 and l_2 has cutting sequence s.

Proof. Suppose that γ_s doesn't have cutting sequence s. Then at some point the cutting sequence of γ_s and s differ, forcing γ_s and p_s to end up in different squares. But once they are in different squares, p_s can never have the same limit points as γ_s, which is in contradiction with the construction of γ_s. □

We note here that a reduced doubly infinite sequence s that does begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ cannot be the cutting sequence of some geodesic in $\mathbb{H}$. We will be able to prove this in a while. So now we have found all doubly infinite sequences that can occur as the cutting sequence of a geodesic. What about terminating sequences? For such a sequence s we can make a polygonal path p_s in the same way as above. The difference is that at some point this path p_s stops, ending up in a square. Arranging the four vertices v_1, v_2, v_3, v_4 of this square in increasing order $v_1 < v_2 < v_3 < v_4$, we may choose v_2 or v_3 as 'limit point'. If s is finite, then at both sides we end up in a square and we have to choose both the limit points in this way. If s is infinite, then we get the other limit point in the usual way. Here again we have to exclude the infinite tail of $ABA'B'$ or $BAB'A'$. So also with terminating sequences s we can associate two limit points and again, the geodesic connecting these limit points has precisely s as cutting sequence. The proof is basically the same as in the doubly infinite case.

Recall that in the square tessellation we could also consider the geodesics projected on the fundamental domain. This is possible because two geodesics in the same orbit have the same cutting sequence. In the following we will see that the same holds in the A, B-tessellation. To this end we need to show that the construction of the cutting sequence of a geodesic is invariant under G.

Theorem III.3. *Two geodesics in $\mathbb{H}$ have the same cutting sequence if and only if they are in the same G-equivalence class.*

Proof. Take two geodesics γ_1 and γ_2 which are in the same G-equivalence class, i.e. there is an $M \in G$ with $\gamma_1 = M\gamma_2$. As M preserves the tessellation and the labeling we know that γ_1 and γ_2 must have the same cutting sequence.

The other way around: suppose that two geodesics γ_1 and γ_2 have the same cutting sequence. Pick a square that is cut by γ_1 and take this square as the starting square to construct a polygonal path with the same cutting sequence as γ_1. There is an $M \in G$ such that $M\gamma_2$ also cuts this square, and in exactly the same point of the cutting sequence as γ_1. Yet as the cutting sequences are the same, we can now use the same polygonal path for both geodesics. This implies that both γ_1 and $M\gamma_2$ are equal to the geodesic connecting the two limit points of the polygonal path. So we conclude that γ_1 and γ_2 are in the same G-equivalence class if their cutting sequences agree. □

So instead of looking at geodesics in the upper half-plane $\mathbb{H}$, we can also consider projections onto the fundamental domain S. As the four vertices of S all lie at ∞ or in $\mathbb{Q}$, we see that identifying opposite sides of S we find a torus minus one point. We will call this a *punctured torus*, denoted by T^*. The cutting sequence of a projected geodesic is defined to be the cutting sequence of any of its lifts to $\mathbb{H}$.

Using Theorem III.3 we can prove that (doubly infinite or terminating) sequences that begin or end in an infinite repetition of $ABA'B'$ or $BAB'A'$ cannot be the cutting sequence of some geodesic in $\mathbb{H}$. Consider such a sequence s, for example suppose it ends in an infinite repetition of $ABA'B'$. We make a polygonal path with s as its cutting sequence by starting from the square S and as starting point in s we choose the point where the infinite repetition of $ABA'B'$ begins. So starting from S we cut the edge labeled A and see that we go one square to the left, here we cut the edge labeled B and again this means going one square to the left. In this way we see that we keep on going one square to the left each time and never end up in a half-disk. This implies that the polygonal path doesn't have two limit points which means that we cannot associate a geodesic to it in the way we did above. Now suppose that s is the cutting sequence of a geodesic γ. Then there is a $M \in G$ such that $M\gamma$ cuts S at exactly the point in the cutting sequence where the infinite repetition starts. Then $M\gamma$ must follow the polygonal path constructed above which implies that $M\gamma$ cannot be a geodesic. Yet an element of G sends geodesics to geodesics, which means that γ cannot exist.

subsection III.3.2 Periodic Cutting Sequences

We have now completely determined which sequences can occur as cutting sequences of geodesics. In particular we see that every periodic sequence is a cutting sequence, except if the period is $ABA'B'$ or $BAB'A'$. It is quite easy to find a geodesic with a given periodic cutting sequence. Consider a periodic sequence s with period C, a finite word in A, B, A' and B'. When constructing the polygonal path p_s we have to choose a starting point in s. This choice gave us a geodesic with cutting sequence s. Changing the starting point leads to a different geodesic with the same cutting sequence. Hence these two geodesics are in the same G-equivalent class.

Example III.2. Suppose the pattern AB occurs in s. We could choose to take B as a starting point in S, i.e. starting from a point in S we cut the side labeled B. Note that this is the same as taking A as a starting point in $A(S)$. Hence the geodesic we obtain from starting with B in S is the same as the A-image of the geodesic we get when starting with A in S. This explains how changing the starting point leads to G-equivalent geodesics.

As s is periodic, then there are only finitely many different starting points to choose from, precisely as many as the length $l(C)$ of the period C. In particular we see from the above example that a geodesic γ with cutting sequence s has the property that $C\gamma = \gamma$. This implies that the endpoints of γ must be fixed points of C. Note that there are different ways to describe the same period, for example the periods AB and BA amount to the same sequence. In general, there are just as many ways to write C as the length $l(C)$. We can thus find $l(C)$ different geodesics with cutting sequence s by looking at the fixed points of the different ways to write C.

Example III.3. Consider the cutting sequence s with period AB or BA. The period AB corresponds to the map $AB : z \mapsto \frac{1}{3-z}$ and the period BA gives us the map $BA : z \mapsto \frac{-1}{z+3}$. We can compute the fixed points of both of these maps. The map AB has fixed points $\frac{3\pm\sqrt{5}}{2}$ and thus the geodesic γ_{AB} connecting these two points has cutting sequence s. Similarly BA has fixed points $\frac{-3\pm\sqrt{5}}{2}$ which shows that the geodesic γ_{BA} joining these point has cutting sequence s. Note that $B\gamma_{AB} = \gamma_{BA}$ and $A\gamma_{BA} = \gamma_{AB}$. Figure III.8 depicts γ_{AB} and γ_{BA}. If we project either γ_{AB} or γ_{BA} on S we get exactly the union of the two parts of γ_{AB} and γ_{BA} that lie in S. We see that this projection is closed and has non self-intersections.

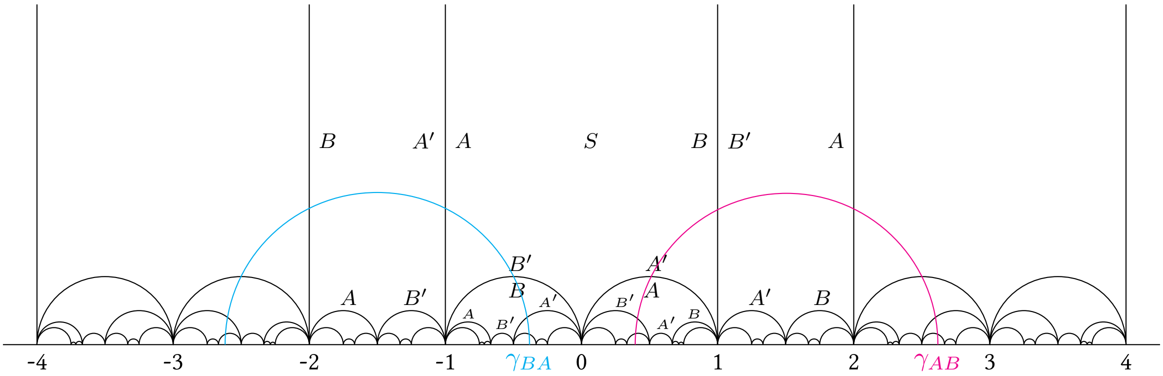


Figure III.8.: The geodesics γ_{AB} and γ_{BA} with cutting sequence $\overline{AB}$

Example III.4. Consider a cutting sequence s with period $ABA'B$, we can look at this period in four different ways. We can compute that $ABA'B$ corresponds to a map with fixed points $\frac{5\pm\sqrt{13}}{6}$ which implies that the geodesic $\gamma_{ABA'B}$ joining these points has cutting sequence s. In the same way we can compute that $\gamma_{BA'BA}$ connects $\frac{-5\pm\sqrt{13}}{2}$, that $\gamma_{A'BAB}$ is a geodesic between $\frac{1\pm\sqrt{13}}{2}$ and finally that $\gamma_{BABA'}$ joins $\frac{1\pm\sqrt{13}}{6}$. Figure III.9 depicts these four geodesics in the upper half-plane. Again we can consider the projection of any of these geodesics to S, it is precisely the union of the four parts of the geodesics lying in S. Note that this projected geodesic cuts itself on the boundary of S. Hence, if we identify the opposite sides of S to obtain the punctured torus T^*, we find a closed projected geodesic with one self-intersection.

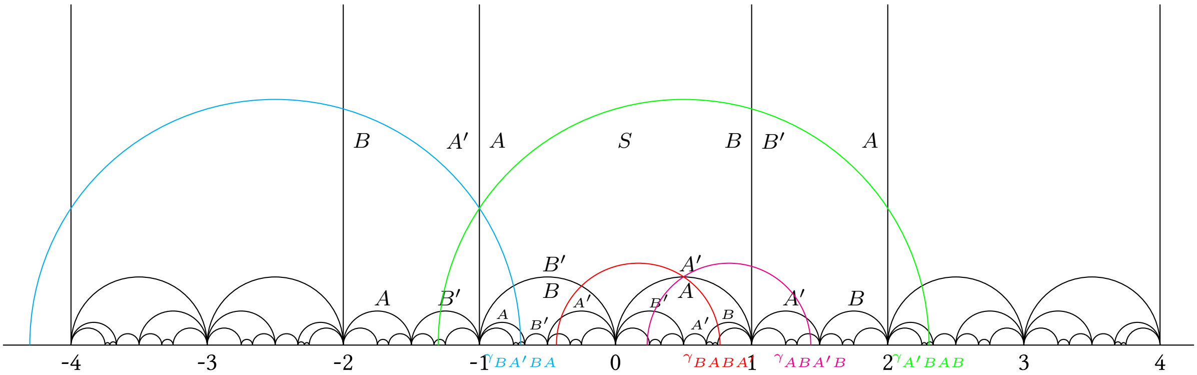


Figure III.9.: Four geodesics with cutting sequence $\overline{ABA'B}$

In the next chapter we will examine projected geodesics on T^* in more detail. It turns out to be no coincidence that the projection of a geodesic with cutting sequence $\overline{AB}$ has no self-intersections while the projection of a geodesic with cutting sequence $\overline{ABA'B}$ intersects itself.

Chapter IV
Simple Loops on the Punctured Torus

In the previous chapter we have considered geodesics in $\mathbb{H}$ and their cutting sequences coming from the A, B-tessellation. In this chapter we focus on the punctured torus T^*, instead of on the hyperbolic upper half-plane $\mathbb{H}$. We know that the punctured torus is obtained by identifying opposite sides of the square S of the A, B-tessellation. By a *geodesic on* T^* we mean the projection of a geodesic in $\mathbb{H}$ on the punctured torus. The other way around, given a geodesic on T^* we can consider its lifts to the upper half-plane. There are different lifts, yet we know that they all have the same cutting sequence. This allows us to associate a cutting sequence to a geodesic on T^*. Note that we needed to direct geodesics in $\mathbb{H}$ in order to get a notion of a cutting sequence. In the same way we assume geodesics on T^* to be oriented, the orientation comes from the orientation of the lifts. In the following we consider closed and *simple* geodesics on the punctured torus, where by simple we mean 'without self-intersections'.

This chapter revolves around proving that the cutting sequences of these geodesics are periodic and characteristic. Note that we have defined the notion of a characteristic sequence only for sequences of two symbols. Yet, the cutting sequence of a geodesic can contain four symbols in which case it is unclear what a characteristic sequence is. We will see that the cutting sequence of a closed, simple geodesic can contain at most two symbols, consequently we can use the same definition of a characteristic sequence as before.

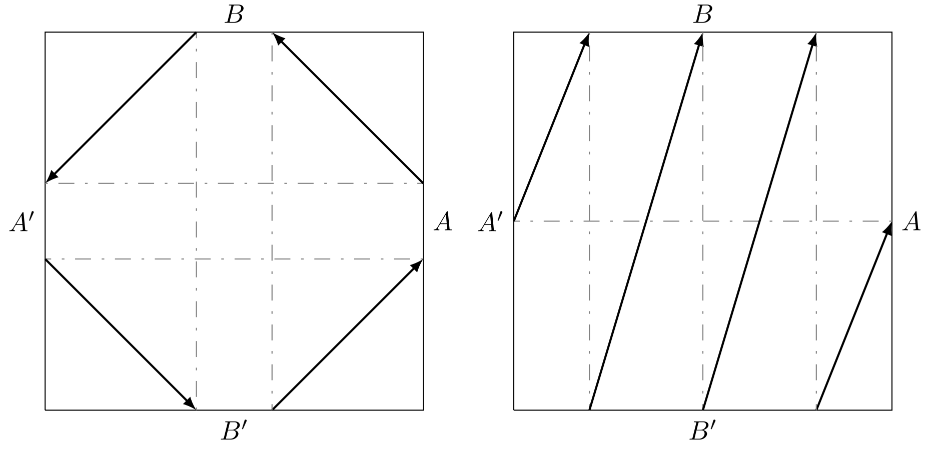
section IV.1 Geodesics on the Punctured Torus

We can find the cutting sequence of a geodesic γ on T^* by considering any of its lifts to $\mathbb{H}$. However, there is another way to find this cutting sequence, which turns out to be more convenient for us. We can consider the geodesic γ projected on the square S rather than on T^*. This gives the same result as the projection on T^*, only in this case the opposite sides of S are not yet identified. These sides are labeled A, B, A', B' just as before, and we can find the cutting sequence of γ by looking at the order in which it cuts these sides.

We will use this way of finding the cutting sequence of a certain geodesic several times in the proofs below. Most of the results in this chapter are proven by using a picture of the square S and drawing a geodesic with a certain cutting sequence in this square. We do this because these pictures give a good idea of what is going on. In most of the proofs we would like to show that a certain pattern of A, B, A' and B' cannot occur in the cutting sequence of a closed, simple geodesic. To this end we assume that the pattern does occur and find a contradiction with the closedness and simpleness of the geodesic. We find this contradiction by drawing the segment of the geodesic with this pattern as a cutting sequence in the square S and exploring all the options for this geodesic segment to continue. The picture then illustrates that we can never close the geodesic without allowing self-intersections.

These proofs depending on pictures might not be as rigorous as one would like a proof to be. However, the pictures give a lot more insight in the situation as would any proof without pictures do. For clarity we will depict S by a usual square and the geodesic by straight lines. We can safely do this as it doesn't change the essence of the situation. As we agreed on recording the label of the side of the square we are leaving, we label the outer sides of S. See Figure IV.1 for two examples of closed geodesics on S.

Remark IV.1. The lemma's and theorems in this chapter involve geodesics. As said above we depict these geodesics as straight lines in a square. Following the proofs of the results in this chapter one sees that it is not necessary to restrict these results to geodesics. In fact, all the results in this chapter obtained on geodesics holds for continuous curves in general. More concretely, instead of using straight lines we could have used curved lines as well. The essence lies in these lines forming a simple and closed path, whether they be straight or curved. We restrict ourselves to geodesics because those are exactly the paths we need to apply the results to in the chapter to come.



(a) A geodesic with period cutting sequence $\overline{BAB'A'}$ (b) A geodesic with cutting sequence $\overline{ABBB}$

Figure IV.1.: Examples of projected geodesics on S.

SECTION IV.2 Closed, Simple Geodesics on the Punctured Torus

We now embark on a journey towards proving that a geodesic on T^* is closed and simple if and only if its cutting sequence is periodic and characteristic. This result is split up into several lemma's.

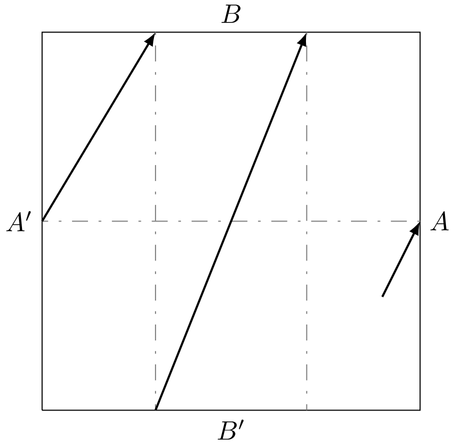
Lemma IV.1. *A geodesic on T^* is closed if and only if its cutting sequence is periodic.*

Proof. Consider a geodesic γ on T^* and view it as a geodesic on S, as discussed above. We assumed γ to be oriented, so we can walk along γ in the direction given by this orientation. The geodesic γ is closed precisely when the entire geodesic is determined by a connected and bounded segment. In more concrete terms, this means that walking along γ, starting at any point s, always leads back to s, and from then on the path repeats itself. Hence after a finite number of times intersecting the sides, we repeat the pattern in A, B, A' and B' we encountered thus far. From this description the equivalence between γ being closed and its cutting sequence being periodic is clear. □

Lemma IV.2. *A closed and simple geodesic on T^* has a cutting sequence containing at most two symbols, unless its cutting sequence is $\overline{ABA'B'}$ or $\overline{BAB'A'}$.*

Proof. We will prove that A and A' cannot both occur in the cutting sequence of a closed and simple geodesic, unless this cutting sequence is $\overline{ABA'B'}$ or $\overline{BAB'A'}$. In the same way it then holds that B and B' cannot both occur in such a cutting sequence. This proves that the cutting sequence of a closed, simple geodesic can contain at most two symbols, if its period is not $ABA'B'$ or $BAB'A'$.

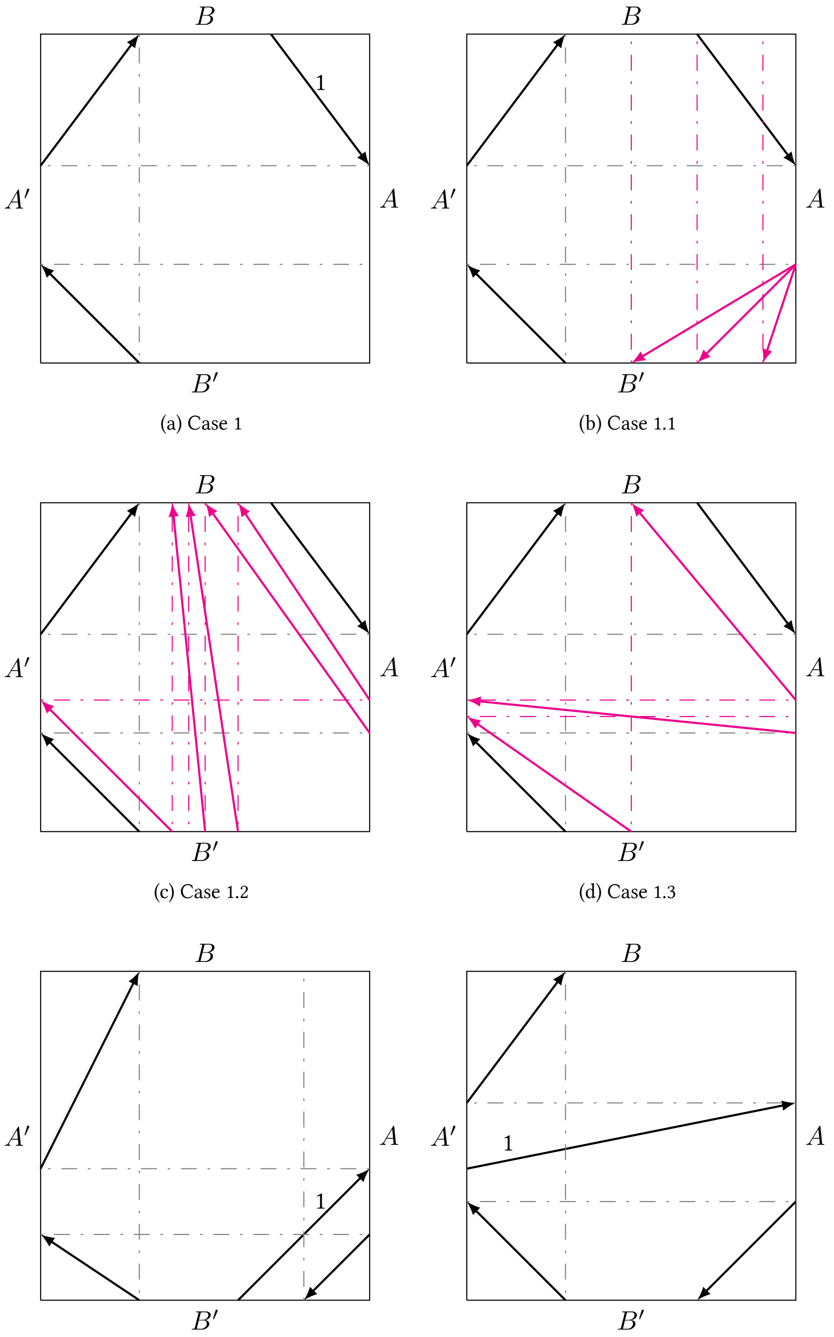
We proceed by contradiction. Suppose both A and A' occur in the cutting sequence of a closed, simple geodesic on T^*. We can consider occurrences of A and A' that are as close to each other as possible. This means that there are either only B's or only B''s between them. We restrict ourselves to the case where there are only B's between A and A', the other case can be treated similarly. It is easy to prove that the cutting sequence of a closed, simple geodesic cannot contain a pattern AB^nA' for $n > 1$. In case $n = 1$ we need to do a bit more work.



Suppose the cutting sequence of a closed, simple geodesic contains the pattern AB^nA' for some $n > 1$. We can draw the geodesic segment which has cutting sequence AB^nA' in the square S. In the figure we have depicted a geodesic segment with cutting sequence ABB. Depending on n we need to draw geodesic segments going from the B'-side to the B-side a couple more times. Yet, at some point we arrive at the side labeled B' and have to go to the side labeled A'. It is clear from the picture that we need to allow self-intersections to do so. Hence a simple geodesic never has AB^nA' with $n > 1$ as part of its cutting sequence. In the same way we can prove that $A'B^nA$ with $n > 1$ cannot occur in the cutting sequence of a simple geodesic, simply flip the picture to see this.

We have proven that AB^nA' cannot occur in the cutting sequence of a closed, simple geodesic if $n > 1$. Now suppose there is one B between A and A', i.e. ABA' occurs in the cutting sequence. This also contradicts the closedness and simpleness of the geodesic, unless the cutting sequence is $\overline{ABA'B'}$. To prove this we make several case distinctions. The sequence ABA' can be preceded by B, B' or A. Realize that the sequence can never be preceded by A', because we only consider reduced sequences. These three cases will be treated separately. The proofs all have the same structure, we will assume that a certain pattern exists and, unless this pattern is $\overline{ABA'B'}$, find that at some point we cannot make the geodesic closed without using self-intersections.

We call the situations where ABA' is preceded by a B', B and A Case 1, 2 and 3 respectively. The next page contains the accompanying figures, they provide a picture of the situation described in these cases.



(a) Case 1

(b) Case 1.1

(c) Case 1.2

(d) Case 1.3

(e) Case 2

(f) Case 3

Case 1: Consider Case 1, so suppose ABA' is preceded by a B'. This means that the arrow labeled '1' goes from the B-side to the A-side. We have three options for the next cut, we can cut B', B and A. This leads to Case 1.1, 1.2 and 1.3 respectively.

Case 1.1: Suppose we record B' after $B'ABA'$, there are three different ways to do this. We can cut the B'-side precisely below the point where the arrow labeled '1' intersects B, which closes the geodesic and yields the cutting sequence $\overline{ABA'B'}$. The other options are cutting B' to the left or right of this point. If we cut to the right, we are forced to walk in cycles $ABA'B'$ in order to keep the geodesic simple, and we can never get a closed geodesic in this way.

In case we cut to the left we can reverse the situation by checking the options left for the arrow that comes before arrow 1. This arrow must go from the A-side to the B'-side in order to keep the geodesic simple. The arrow before that must go from the B'-side to the A'-side for the same reason. Continuing this we see that we must walk in cycles $ABA'B'$ to keep the geodesic simple. This can never give rise to a closed geodesic.

Case 1.2: If $B'ABA'$ is followed by B, we end up at the B'-side and can choose A, A' or B. If we pick A we end up in circles $BAB'A'$ in order to keep the geodesic simple, which cannot lead to a closed geodesic. If we choose A', we are forced to choose B after that and again end up at the B'-side in a similar situation as before. Hence we can assume to choose B, after which we have a choice between A' and B. We can cut B a finite number of times more, but at some point we have to cut A'. Indeed, if we choose B an infinite number of times we can never get a closed geodesic.

In order to keep the geodesic simple we now have to cut B. After that we have a choice between A and B. Picking A leads to cycles $BAB'A'$, so we choose B. At this point we have a choice between A' and B, and neither of these choices changes the situation essentially. For if we pick A', we are forced to record BB to keep the geodesic simple, and if we pick B we have to choose $A'BB$ for the same reason, both end up in the same situation as before. This process can never yield a closed geodesic, because this would require cutting B' at some point.

Case 1.3: Now suppose that $B'ABA'$ is followed by A'. We could cut A' some more times, but if we want the geodesic to be closed, we need to cut B at some point. This leads us to the B'-side, where we can choose between A and A'. If we choose A we end up in cycles $BAB'A'$ after that and we can never obtain a closed geodesic. Thus we pick A' to be cut next. We now have two options, we either cut A' or B, neither change the situation essentially. Indeed, if we cut A' we are forced to cut BA', and if we cut B we have to record A', both leading to a similar situation. Consequently we from now on only record A' and B, which does not give rise to a closed geodesic.

This deals with Case 1, the case in which ABA' is preceded by a B'. Note that the resulting sequence $B'ABA'$ is a commutator. Consequently the above reasoning shows that this particular commutator cannot occur in the cutting sequence of a closed, simple geodesic and in the same way one can prove that any commutator in the symbols A, B, A' and B' cannot occur in such a cutting sequence. This fact can be used in proving the next two cases.

Case 2: Suppose ABA' is preceded by a B. This means that the arrow labeled '1' goes from the B'-side to the A-side. In order to have no self-intersection the pattern $BABA'$ must be followed by a B', i.e. we obtain $BABA'B'$. This results in a situation as depicted in (a). Realize that $ABA'B'$ is a commutator, hence using the above argumentation we deduce that this sequence can never be part of the cutting sequence of a closed and simple geodesic.

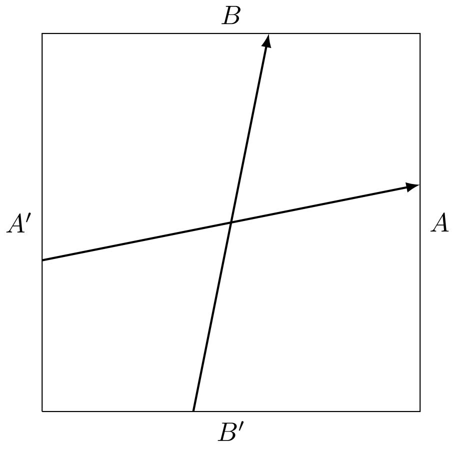
Case 3: Suppose that ABA' is preceded by A, this leads to the arrow labeled '1' going from the A'-side to the A-side. The pattern $AABA'$ we obtain in this way can be followed by A' or B'. If we choose A' our only option is to choose A' over and over again, which does not lead to a closed geodesic. If we choose B' we obtain a sequence in which a commutator occurs. Such a sequence can not be part of the cutting sequence of a closed, simple geodesic.

We see that ABA' can never occur in the cutting sequence of a closed, simple geodesic, unless the cutting sequence is $\overline{ABA'B'}$. In a similar way we can prove that $A'BA$ can never occur, unless the cutting sequence is $\overline{BAB'A'}$. □

A geodesic with cutting sequence $\overline{ABA'B'}$ or $\overline{BAB'A'}$ circles around the puncture. With these two exceptions the cutting sequence of a closed, simple geodesic on T^* has at most two symbols. In case the cutting sequence is non-constant, it is easy to see that one of the symbols must be isolated.

Lemma IV.3. *Consider a simple geodesic whose cutting sequence consists of two symbols. Now one of these symbols must be isolated.*

Proof. Suppose the cutting sequence of a closed simple geodesic on T^* contains the symbols A and B. The proof for a cutting sequence in $A, B',\ \ A', B$ or A', B' is entirely similar.



If A and B are both not isolated, then at some point AA and BB occur in the cutting sequence. This results in a situation as depicted here, which is to say that two line segments similar to the ones depicted here must be part of the geodesic. Such a geodesic clearly contains a self-intersection. □

Eventually we wish to prove that the cutting sequence of a closed and simple geodesic on T^* is periodic and characteristic. In Lemma IV.1 we already saw that the cutting sequence of such a geodesic is periodic, hence it remains to be shown that it is characteristic. First we prove that such a cutting sequence is either constant or almost constant. Recall that a sequence in two symbols is called almost constant of value n if one of the symbols is isolated and if between every two occurrences of this symbol there are n or $n+1$ occurrences of the other symbol.

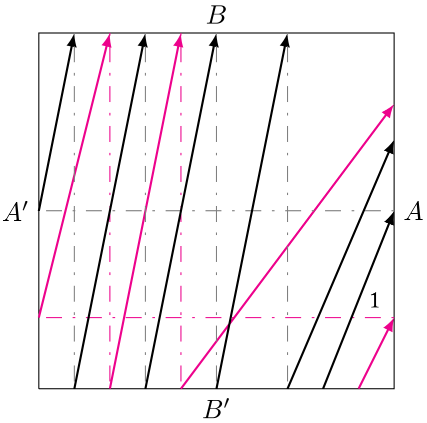
Lemma IV.4. *Consider a closed and simple geodesic γ on T^* whose cutting sequence contains two symbols. This cutting sequence is almost constant of value n for some $n \geq 1$.*

Proof. Without loss of generality we assume that the cutting sequence contains the symbols A and B, the proof for a cutting sequence in $A, B',\ \ A', B$ or A', B' being similar. In Lemma IV.3 we proved that one of the symbols must be isolated, suppose that A is this isolated symbol. We proceed by contradiction and assume the cutting sequence not to be almost constant. This means that at some point in the cutting sequence AB^kA and AB^lA occur with $|k-l| > 1$. Without loss of generality we may assume that $k > l$.

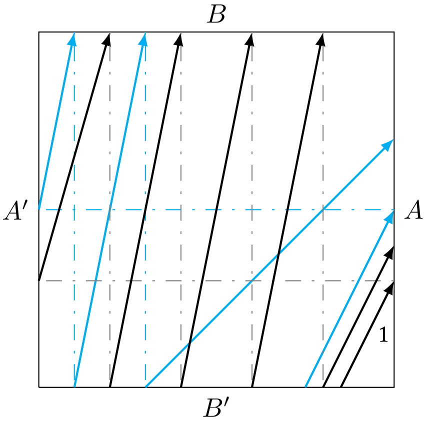
We take segments γ_k and γ_l of the geodesic γ with cutting sequences AB^kA and AB^lA respectively. First observe that as A is isolated in the cutting sequence of γ, we know that both γ_k and γ_l must originate from the B'-side. Indeed, we might as well have taken the segments associated to BAB^kA and BAB^lA. The l intersections of γ_l with the B-side of S have to be interleaved with those of γ_k in such a way that γ_l and γ_k do not intersect. For such an intersection would lead to a self-intersection in γ, a clear contradiction with the assumption that γ is simple.

There are k intersections of γ_k with the B-side, resulting in $k-1$ intervals between the intersections. It is clear that if $l < k-1$, then after γ_l recorded AB^l, it has to intersect γ_k to record A. Now as $l < k$ and $|k-l| > 1$ we know that $l < k-1$, deriving the desired contradiction. □

The above proof is best illustrated by an example, we consider the case where $k = 4$ and $l = 2$. Note that nothing essential changes when using different numbers.



The geodesic segment γ_4 with cutting sequence AB^4A is depicted in black. We try to draw the geodesic segment γ_2 of AB^2A in there without intersections. Note that A is isolated, hence the pattern AB^2A is preceded by B. This leads to an arrow from the B'-side to the A-side, we can choose two different starting points for this arrow. If we choose a starting point such that the tip of the arrow is below the tip of arrow labeled '1', we follow the magenta path. We see that there is no other way but to self-intersect. The only way to have no self-intersections is putting $k = 3$ or $k = 4$ instead of $k = 2$.



If we choose the arrow from the B'-side to the A-side such that the tip of the arrow is above the tip of the arrow labeled '1', we end up with the cyan path. In this case it is even worse, we have two self-intersections. The only way to avoid this, is to set the value of k equal to either 5 or 4.

The above implies that if AB^4A occurs in the cutting sequence of a closed and simple geodesic, then either B occurs only in third and fourth powers or it occurs in only fourth and fifth powers. So a cutting sequence in which AB^4A occurs is either almost constant of value 4 or almost constant of value 5. The argument for AB^nA with arbitrary $n \geq 1$ is entirely the same.

We are now ready to prove that the cutting sequence of a closed, simple geodesic is characteristic.

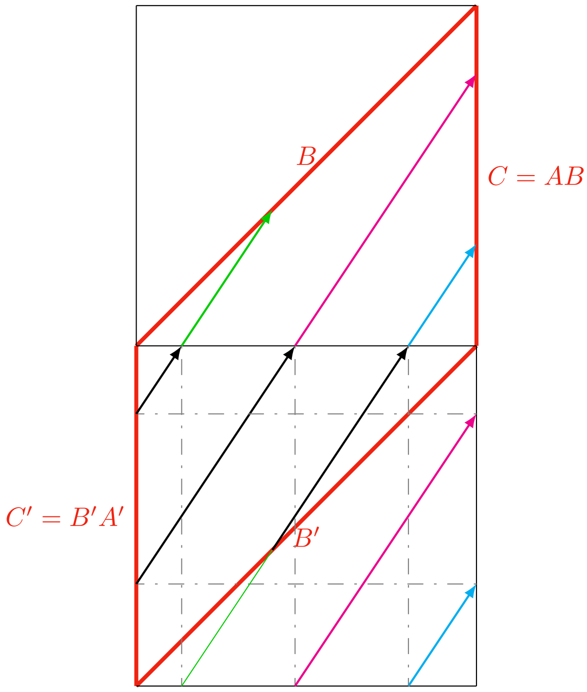
Lemma IV.5. *A closed, simple geodesic on T^* which doesn't have cutting sequence $\overline{ABA'B'}$ or $\overline{BAB'A'}$ has a periodic and characteristic cutting sequence.*

Proof. Consider such a geodesic γ, by Lemma IV.1 we know that its cutting sequence is periodic. Thus we are left with proving that the cutting sequence is characteristic. Lemma IV.2 tells us that the cutting sequence of γ contains at most two symbols. In case it contains just one symbol, we have a constant cutting sequence which is characteristic. So suppose that the cutting sequence contains two symbols. Without loss of generality we assume that these symbols are A and B and by Lemma IV.3 we can assume that A is isolated. Lemma IV.4 shows us that there is some $n \geq 1$ such that the cutting sequence is made up of AB^n's and isolated B's. We need to show that the derived sequence of the cutting sequence is either constant or again almost constant for some $m \geq 1$.

As we only encounter AB^n's and B's we could, instead of using the square S, also use a quadrilateral of which opposite sides are identified by AB^n and B. We can make such a quadrilateral out of S. Consider the left side of S, this is the side usually labeled A'. Under AB^n this side is mapped to the right side of the n-th square 'above' S. Connecting these two sides we obtain a quadrilateral. By construction, its left and right side are identified by AB^n. It can easily be seen that the other sides are identified by B. Consequently we label the sides of this quadrilateral $B, B', C = AB^n$ and $C' = B'^nA'$. This quadrilateral is again a fundamental domain for the action of G on $\mathbb{H}$.

The closed, simple geodesic γ on S can be viewed on this quadrilateral instead of on S and we obtain a closed, simple geodesic γ' whose cutting sequence is precisely the derived sequence of the cutting sequence of γ. If this derived sequence is not constant, then Lemma IV.4 implies that it is again almost constant for some value $m \geq 1$. Repeating this argument we have shown that the cutting sequence of γ is characteristic. □

Again, this proof is better understood when considering a concrete example. We inspect the geodesic with cutting sequence $\overline{ABABB}$.



We draw a geodesic γ with cutting sequence $\overline{ABABB}$ in the usual manner. The quadrilateral whose opposite sides are identified by AB and B is depicted in red. There are three parts of γ that do not lie in this quadrilateral. We can transfer them to the red area, obtaining a geodesic within this quadrilateral.

This geodesic is closed and simple, as γ has the same properties. We can read of that the new geodesic has cutting sequence $\overline{BBC}$ and this is exactly the derived sequence of $\overline{ABABB}$.

The converse of Lemma IV.5 is also true, if a geodesic has a periodic, characteristic cutting sequence, then it is closed and simple.

Lemma IV.6. *A geodesic on T^* with a periodic, characteristic cutting sequence is closed and simple.*

Proof. Consider such a geodesic γ with cutting sequence s_γ. As this cutting sequence is periodic, Lemma IV.1 tells us that γ must be closed. Thus we only need to prove that γ is simple.

We know that s_γ is periodic and characteristic and hence there is a straight line l in $\mathbb{R}^2$ with rational slope and s_γ as its cutting sequence. We can choose l such that it never goes through a vertex of the square tessellation. It is possible to consider l projected on the torus T. In a similar way as we did with geodesics in $\mathbb{H}$ we can also consider this projection of l on the square $[0,1] \times [0,1]$ instead of on T. Opposite sides of this square are identified under the maps a and b and this gives a labeling a, a^{-1}, b and b^{-1} which is very similar to the labeling A, A', B, B' of S.

Consider l projected on the square, by construction it does not go through any of the vertices. Changing the metric on the square we can also consider the square to be S and then the line l becomes a geodesic in $\mathbb{H}$ with cutting sequence s_γ. As the projection of l to the square has no self-intersection we see that this geodesic doesn't have self-intersections either. The geodesic thus constructed has the same cutting sequence as γ. As there is only one geodesic on T^* with a given cutting sequence we see the constructed simple geodesic must actually equal γ, proving the desired. □

Putting everything together we have proven the main result of this chapter, namely that a geodesic on T^* is closed and simple if and only if its cutting sequence is periodic and characteristic. This fact turns out to be very important in providing a link between Markoff irrationalities and closed, simple geodesics on the punctured torus, which will be given in the next chapter.

Theorem IV.1. *A geodesic on T^* is closed and simple if and only if it has a periodic and characteristic cutting sequence.*

SECTION IV.3 Theorem IV.1 through Examples

Recall Example III.3 which showed that a geodesic on T^* with cutting sequence $\overline{AB}$ is closed and simple. Furthermore, we saw in ExampleIII.4 that a geodesic with $\overline{ABA'B}$ as its cutting sequence has a self-intersection. This is completely in accordance with Theorem IV.1. Indeed, the cutting sequence $\overline{AB}$ is periodic and characteristic, while the sequence $\overline{ABA'B}$ fails to be characteristic. Example IV.1, IV.2 and IV.3 provide us with three more instances of Theorem IV.1.

Example IV.1. Consider the cutting sequence s with period $ABABB$. As this sequence is periodic and characteristic, the geodesic on T^* with s as a cutting sequence must be closed and simple. We can compute the fixed points of the five different ways to describe this period and this leads to five geodesics in $\mathbb{H}$ with s as their cutting sequence. Figure IV.2 depicts $\gamma_1 = \gamma_{ABABB}$, $\gamma_2 = \gamma_{BABBA}$, $\gamma_3 = \gamma_{ABBAB}$, $\gamma_4 = \gamma_{BBABA}$ and $\gamma_5 = \gamma_{BABAB}$. We see that the projection of any of these geodesics to T^* is closed and simple.

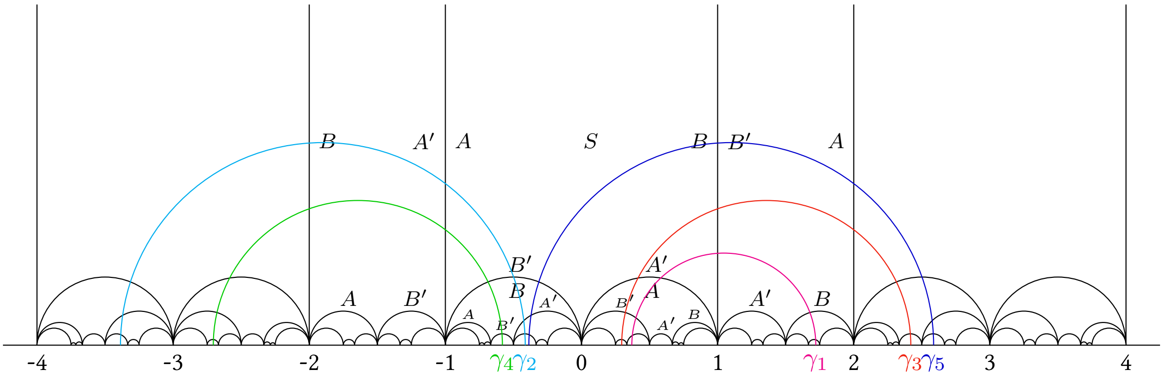


Figure IV.2.: Five geodesics with cutting sequence $\overline{ABABB}$

Example IV.2. We consider a cutting sequence s with period $ABAB^3$. This sequence is periodic albeit not characteristic. We deduce that the geodesic on T^* with s as cutting sequence is closed, but not simple. Computing the fixed points of the six different ways to write the period, we find six geodesics $\gamma_1 = \gamma_{ABAB^3}$, $\gamma_2 = \gamma_{BAB^3A}$, $\gamma_3 = \gamma_{AB^3AB}$, $\gamma_4 = \gamma_{B^3ABA}$, $\gamma_5 = \gamma_{B^2ABAB}$ and $\gamma_6 = \gamma_{BABAB^2}$ on $\mathbb{H}$. These geodesics are depicted in Figure IV.3. Note that the projection of any of these geodesics to S cuts itself on the boundary of this square. This means that a geodesic on T^* with cutting sequence s is not simple, it is a closed geodesic with one self-intersection.

Example IV.3. We have seen two examples, namely Example III.4 and IV.2, of self-intersecting geodesics for which the intersection happens on the boundary of S. Of course the intersection can also happen in the interior of S, this is the case for the geodesic on T^* with cutting sequence $\overline{AABB}$. In Figure IV.3 we obtain $\gamma_1 = \gamma_{AABB}$, $\gamma_2 = \gamma_{ABBA}$, $\gamma_3 = \gamma_{BBAA}$ and $\gamma_4 = \gamma_{BBAA}$, four geodesics in $\mathbb{H}$ with cutting sequence $\overline{AABB}$. The projection of any of these geodesics to S cuts itself in the interior of S.

Looking at all five examples we might note that the diameters of all the geodesics in Examples III.3 and III.4 are less than 3. Yet, both in Example III.4, IV.2 and IV.3 we find a geodesic with a diameter which is bigger than 3. The value 3 might remind us of the Markoff spectrum, as in Chapter II we had a special interest in the μ-values below 3. In the next chapter we discuss the Markoff spectrum in a geometrical way and we will see that the above observation on the diameters of the geodesics is something that holds in greater generality.

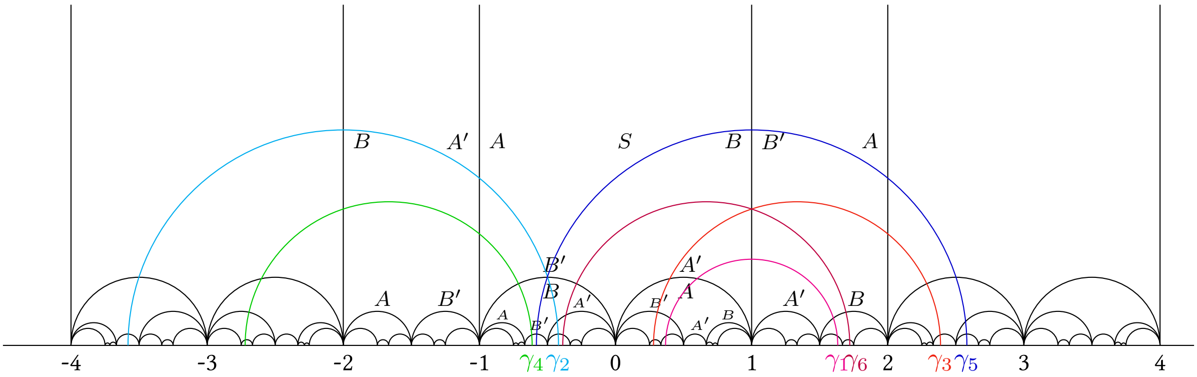


Figure IV.3.: Six geodesics with cutting sequence $\overline{ABAB^3}$

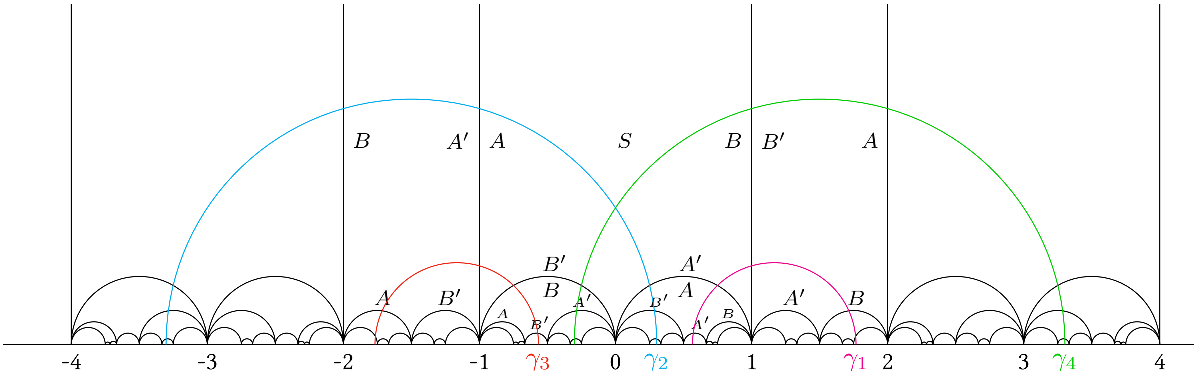


Figure IV.4.: Four geodesics with cutting sequence $\overline{AABB}$

Chapter V

Markoff Theory, a Geometric Approach

Traditional Markoff theory, as discussed in Chapter II, is mainly concerned with doubly infinite sequences of integers. In this chapter we cover a connection between Markoff theory and geometry, important notions of Chapter II have geometric counterparts. Chapters III and IV provide us with the machinery needed to give such a geometric interpretation of Markoff Theory.

We start by describing a way to associate a Markoff irrationality to a closed and simple geodesic on T^*. After that we discuss the notion of a λ-value of a geodesic. This allows us to define the set of Markoff values as the set of λ-values below 3. It turns out that these Markoff values can be computed explicitly. We end this chapter by giving a geometric interpretation of the Lagrange spectrum and the Markoff numbers. This chapter makes use of the theory in Series (1985a), Series (1985b) and Cohn (1955).

section V.1 Endpoints of Simple Geodesics

In this section we will associate a class of Markoff irrationalities to a closed and simple geodesic on T^*. Recall that a Markoff irrationality is a real number a such that $\lambda(a) < 3$ and the latter holds precisely if the tail of a agrees with the tail of a doubly infinite, purely periodic sequence A that satisfies the two conditions of Theorem II.3. There is a correspondence between these Markoff irrationalities and lifts of closed, simple geodesics. More precisely, Markoff irrationalities are the endpoints of these lifts. Recall Theorem III.2:

For $a \in \mathbb{R}_{>0}$ consider (part of) a geodesic joining any point $b \neq (0, 0)$ on the imaginary axis with a. Directing this geodesic segment γ_a from b to a we can read of the L, R-sequence $L^{a_0}R^{a_1}L^{a_2}...$, with possibly $n_0 = 0$, and this sequence has the property that $[a_0, a_1, a_2, ...] = a$.

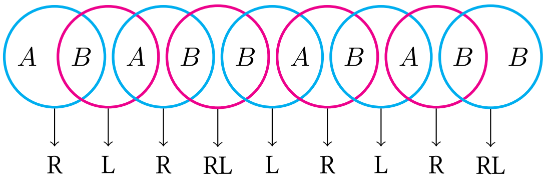
This theorem describes a relation between the endpoint of a geodesic and its L, R-sequence. It is this relation we are going to use in proving the above correspondence. Note that at this point we only have some information about the A, B-cutting sequence of a closed, simple geodesics, namely that this cutting sequence is periodic and characteristic. If we want to use the above theorem we need some information on its L, R-sequence as well.

subsection V.1.1 Converting Cutting Sequences to L,R-Sequences

The A, B-cutting sequence of geodesic is obtained using the A, B-tessellation. Another subdivision of the upper half-plane $\mathbb{H}$ is the Farey tessellation, hence each geodesic can also be assigned an L, R-sequence. The A, B-tessellation is in fact a subtessellation of this tessellation and consequently we can convert cutting sequences in A, B, A', B' to L, R-sequences. For example, if we cut a side labeled A and next a side labeled B we pass one triangle of the Farey tessellation and the vertex we cut of lies to the right. Consequently we replace every occurrence of AB by R. The full conversion table is

$$\begin{array}{cccccc} AB, A'B' & BA, B'A' & AA, A'A' & BB, B'B' & AB', A'B & BA', B'A \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \text{R} & \text{L} & \text{LR} & \text{RL} & \text{LL} & \text{RR} \end{array}$$

Example V.1. Consider a geodesic with cutting sequence $\overline{ABABB}$, we can convert this sequence into an L, R-sequence by using the conversion table above.



We see that a geodesic with cutting sequence $\overline{ABABB}$ has L, R-sequence $\overline{RLRRLL}$.

Our main interest lies in geodesics on T^* that are closed and simple. Such geodesics have a periodic, characteristic cutting sequence, so let's examine the conversion of such sequences.

Consider a characteristic[1] sequence s in A and B such that A is isolated and B occurs in n-tuples or $(n+1)$-tuples for some $n > 0$. Given the pattern AB^nA, the occurrence of B^n contains $n-1$ tuples BB and each of these tuples is converted to RL. Furthermore, AB is converted to R and BA to L. This implies that AB^nA is converted to $R(RL)^{n-1}L$, which equals $RR(LR)^{n-2}LL$ in case $n > 1$. In the same way $AB^{n+1}A$ is converted to $R(RL)^nL = RR(LR)^{n-1}LL$.

For $n > 1$ we thus see that the sequence s converts to a sequence in $RR(LR)^{n-2}LL$ and $RR(LR)^{n-1}LL$. In case $n = 1$ it holds that s gets converted to a sequence in RL and $RRLL$.

If the roles of A and B were reversed, i.e. if B were isolated and A occurred in n-tuples or $(n+1)$-tuples, then we would have to convert BA^nB and $BA^{n+1}B$. In the same way as above we see that BA^nB is converted to $L(LR)^{n-1}R$ and that $BA^{n+1}B$ converts to $L(LR)^nR$. This implies that s becomes a sequence in $LL(RL)^{n-2}RR$ and $LL(RL)^{n-1}RR$ if $n > 1$ and a sequence in LR and $LLRR$ if $n = 1$.

We see that the conversion of a characteristic sequence is build up of two patterns of L and R, the order in which these patterns occur depends on the characteristic sequence itself. If a characteristic sequence is almost constant of value n, we can associate a sequence in n and $n+1$, say ..., n_{-1}, n_0, n_1, ..., to the characteristic sequence. This sequence determines the order in which the two patterns of L and R occur. More explicitly, the characteristic sequence ...$AB^{n_{-1}}AB^{n_0}AB^{n_1}A$... which is almost constant of value $n > 1$ gets converted to

$$...RR(LR)^{n_{-1}-2}LLRR(LR)^{n_0-2}LLRR(LR)^{n_1-2}LL...$$

Reversing the roles of A and B leads to the L, R-sequence

$$...LL(RL)^{n_{-1}-2}RRLL(RL)^{n_0-2}RRLL(RL)^{n_1-2}RR...$$

We could also consider characteristic sequences in A, B', A', B or A', B', they turn out to convert to a similar L, R-sequence. Hence we have determined what the L, R-sequence of a closed and simple geodesic on T^* can look like.

[1] For our purposes we only need to convert periodic, characteristic sequence. Yet, the process of converting a general characteristic sequence, periodic or not, is basically the same as the conversion of a periodic, characteristic sequence. Hence we discuss the conversion of a general characteristic sequence.

SUBSECTION V.1.2 Markoff Irrationalities and Closed, Simple Geodesics

The conversion of A, B-sequences in L, R-sequences allows us to prove that a lift of a closed, simple geodesic has a Markoff irrationality as its endpoint.

Consider γ, an arbitrary lift of a closed, simple geodesic $\overline{\gamma}$ on T^*, and denote its endpoint by e. Using Theorem III.2 we know that a geodesic segment joining any point $a \neq (0,0)$ on the imaginary axis with e has an L, R-sequence L, R-sequence $L^{e_0} R^{e_1} L^{e_2} ...$, with possibly $e_0 = 0$, such that $[e_0, e_1, e_2, ...] = e$. As this geodesic segment and γ have the same endpoint, their L, R-sequences will agree at some point. This implies that we can read of the tail of the continued fraction expansion of e from the tail of the L, R-sequence of γ.

For example, suppose γ has a periodic, characteristic cutting sequence in A and B which is almost constant of value $n > 1$ and with associated sequence in n and $n+1$ denoted by $..., n_{-1}, n_0, n_1, ...$ In this case the tail of e looks like

$$2, 2, (1, 1)_{n_i - 2}, 2, 2, (1, 1)_{n_{i+1} - 2}, 2, 2, ...$$

for some $i \geq 0$ and we get a similar result for any characteristic cutting sequence in two other symbols. As γ is the lift of a closed geodesic, we know that γ has a periodic cutting sequence. This implies that its associated sequence is also periodic and hence the tail of e becomes periodic at some point. Furthermore, Lemma III.3 shows precisely that the associated sequence of a characteristic sequence is Markoff balanced. This implies that the tail of e agrees with the tail of a doubly infinite, purely periodic sequence A in 1's and 2's with Markoff balanced associated sequence. Realize now that e is a Markoff irrationality if and only if precisely these conditions on its tail are met, implying that e is a Markoff irrationality. This proves the fact that every lift of a closed, simple geodesic has a Markoff irrationality as its endpoint.

The above allows one to associate a Markoff irrationality to a lift of a closed, simple geodesic. Every closed, simple geodesic on T^* has different lifts, and to all these lifts we can associate a Markoff irrationality. One might wonder how these irrationalities relate to each other.

It holds that two different lifts of a geodesic on T^* have the same cutting sequence. Hence, putting every lift of a certain closed, simple geodesic through the procedure above, we see that the tails of the continued fractions expansions of the endpoints of all lifts of the same closed, simple geodesic agree. As two geodesics in $\mathbb{H}$ have the same cutting sequence if and only if they project to the same geodesic on T^*, we also see that the tails of two geodesics that come from two different geodesics on T^* can never agree.

The above enables us to associate a class of Markoff irrationalities to a closed, simple geodesic on T^*. Indeed, two Markoff irrationalities are in the same class if and only if the have the same λ-value. We know this to hold if and only if the irrationalities have the same tail. Hence to a closed and simple geodesic we can associate the class of Markoff irrationalities whose tails agree with the tails of the endpoints of the lifts of the geodesic.

SUBSECTION V.1.3 Limits of Geodesics and their Endpoints

We have established that lifts of closed, simple geodesics have Markoff irrationalities as endpoints. Markoff irrationalities correspond to Markoff values. Recall that the Markoff values have 3 as limit point, this result carries over to the geometric setting in the sense that *limits of closed, simple geodesics* that are not closed itself have endpoints with λ-value 3.

Consider geodesics on T^*, we would like to have a notion of a limit of a sequence of such geodesics. To do this it is helpful to first consider geodesics in $\mathbb{H}$, instead of on T^*. A geodesic in $\mathbb{H}$ has a starting point and an endpoint on $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Consequently, a sequence of geodesics $\{\gamma_n\}_{n\geq 0}$ leads to two sequences of values in $\overline{\mathbb{R}}$, a sequence of starting points $\{s_n\}_{n\geq 0}$ and a sequence of endpoints $\{e_n\}_{n\geq 0}$. We say that the sequence of geodesics $\{\gamma_n\}$ in $\mathbb{H}$ converges if both of these sequences converge, say $s_n \to s$ and $e_n \to e$ as $n \to \infty$. The limit of $\{\gamma_n\}$ is then defined to be the geodesic from s to e.

If we want to find the limit of a sequence of geodesics on T^* we need to consider their lifts to $\mathbb{H}$. Note that a geodesic on T^* can be lifted to $\mathbb{H}$ in different ways. Thus a sequence $\{\overline{\gamma_n}\}_{n\geq 0}$ of geodesics on T^* gives us different sequences of geodesics in $\mathbb{H}$. These sequences may not all have a limit in the way described above, but if there exists a sequence of lifts $\{\gamma_n\}$ that does have a limit γ, then we say that $\{\overline{\gamma_n}\}_{n\geq 0}$ converges. In this case we can project γ to T^* to obtain the limit of the sequence of geodesics on T^*.

We restrict our attention again to simple, closed geodesics. Recall the main result of Chapter IV, Theorem IV.1.

A geodesic on T^* is closed and simple if and only if it has a periodic and characteristic cutting sequence.

Consider a geodesic $\overline{\gamma}$ on T^* which is the limit of a sequence of simple, closed geodesics $\{\overline{\gamma_n}\}_{n\geq 0}$. Theorem IV.1 tells us that the cutting sequence of $\overline{\gamma_n}$, and hence also of its lifts, is periodic and characteristic for every $n \geq 0$. The property of periodicity might be lost under taking limits, as the next lemma shows.

Lemma V.1. *A geodesic $\overline{\gamma}$ on T^* which is the limit of closed and simple geodesics $\{\overline{\gamma_n}\}_{n\geq 0}$ is simple and its cutting sequence is characteristic. However, $\overline{\gamma}$ need not be closed and its cutting sequence is not necessarily periodic.*

Proof. First we show that the cutting sequence of $\overline{\gamma}$ need not be periodic, Lemma IV.1 then tells us that it need not be closed. As $\overline{\gamma}$ is the limit of $\{\overline{\gamma_n}\}$ there is a sequence of lifts $\{\gamma_n\}$ to $\mathbb{H}$ that converges to, say, γ. We know that the cutting sequence of γ_n is periodic and characteristic for every $n \geq 0$. Periodicity implies that the starting point s_n and endpoint e_n of γ_n are fixed points of the period of the cutting sequence. We deduce that s_n and e_n must be roots of a quadratic polynomial with integer coefficients. As the sequence $\{\gamma_n\}$ is assumed to have a limit γ, we know that both sequences $\{s_n\}$ and $\{e_n\}$ converge to a real number, s and e respectively. However, the numbers s and e need not be roots of a quadratic polynomial with integer coefficients. In case they are not the cutting sequence of γ is not periodic. This implies that the cutting sequence of $\overline{\gamma}$ is not necessarily periodic.

Next we prove that the cutting sequence of γ, and thus also that of $\overline{\gamma}$, is characteristic. We proceed by contradiction, so suppose that the cutting sequence of γ is not characteristic. We are able to see this by checking out a sufficiently long finite block of the cutting sequence. Indeed, if the sequence is not characteristic, we have several possibilities

- The sequence consists of more than two symbols;
- The sequence consists of at most two symbols which are both not isolated;
- The sequence consists of at most two symbols of which one is isolated, yet the sequence is not almost constant;
- The sequence is almost constant, yet after a finite number of derivations we arrive at a sequence which fails to be almost constant.

Each of these possibilities displays itself in a sufficiently long finite block B of the cutting sequence. As γ is the limit of $\{\gamma_n\}$, it holds that the γ_n get arbitrarily close to γ. This implies that the cutting sequences of γ and γ_n are much alike for big n. By choosing n sufficiently big these cutting sequences can be made to agree on an arbitrarily long block. Hence if we choose n sufficiently big the finite block B is also a part of the cutting sequence of γ_n, implying that this cutting sequence cannot be characteristic. Note however that the cutting sequence of γ_n is characteristic for every $n \geq 0$ by assumption, a clear contradiction. Consequently the cutting sequence of γ, and $\overline{\gamma}$, must be characteristic.

To prove that $\overline{\gamma}$ is simple, recall Lemma IV.6. It states that a geodesic on T^* with a periodic and characteristic cutting sequence is simple and closed. Note that in proving that the geodesic was simple we did not rely on the fact that the cutting sequence was periodic. Hence it holds that a geodesic on T^* is simple if its cutting sequence is characteristic. As we have just seen that the cutting sequence of $\overline{\gamma}$ is characteristic, it must hold that $\overline{\gamma}$ is simple. □

Remark V.1. We just saw that a geodesic on T^* is simple if its cutting sequence is characteristic, the converse is partially true. Recall Lemma IV.3, IV.4 and IV.5, the proofs of these lemma's did not need the fact that we were considering a closed geodesic. We only used the fact that the cutting sequence had at most two symbols. Consequently it is the case that a simple geodesic on T^* has a characteristic cutting sequence, if the cutting sequence consists of two symbols.

We obtain the result that a geodesic on T^* with a cutting sequence in two symbols is simple if and only if its cutting sequence is characteristic.

Markoff Limits and Limits of Closed, Simple Geodesics

We have now proven that a geodesic $\overline{\gamma}$ on T^* is the limit of closed and simple geodesics if and only if its cutting sequence is characteristic. Consider such a limit of closed, simple geodesics and one of its lifts γ to $\mathbb{H}$. If the cutting sequence of γ is periodic, then $\overline{\gamma}$ is itself a closed, simple geodesic. Such a geodesic is known to have a Markoff irrationality as endpoint.

Note that Markoff irrationalities have a tail which agrees with the tail of a purely periodic sequence in 1's and 2's with Markoff balanced associated sequence. Remark II.2 shows that a real number a with $\lambda(a) = 3$ has a similar tail, it agrees with the tail of a sequence in 1's and 2's with Markoff balanced associated sequence which, however, is not periodic. We will call $a \in \mathbb{R}$ a *Markoff limit* if $\lambda(a) = 3$. Consider again a limit of closed, simple geodesics $\overline{\gamma}$ on T^* and one of its lifts γ to $\mathbb{H}$. We claim that the endpoint e of γ is a Markoff limit if and only if the cutting sequence of γ is not periodic.

We know the cutting sequence of γ to be characteristic. Such a characteristic sequence can be converted into an L, R-sequence, the same way as we did in Section V.1.1. We can then read of the tail of the continued fraction expansion of e from this L, R-sequence. If the cutting sequence of γ is not periodic, then we see that the tail of e agrees with the tail of a doubly infinite sequence in 1's and 2's with Markoff balanced associated sequence that is not periodic. By Remark II.2 this implies that e is a Markoff limit. Consequently we see that a lift of a limit of closed, simple geodesics, which is not closed itself, has a Markoff limit as its endpoint.

Remark V.2. Realize that we constructed a limit of closed, simple geodesics by considering limits of starting points and endpoints of lifts of closed, simple geodesics. The endpoints of these lifts were actually Markoff irrationalities. Consequently we see that if a sequence of Markoff irrationalities converges, then it converges to a Markoff limit.

section V.2 Geometric Interpretation of Markoff Values

We have established that one can associate a class of Markoff irrationalities to every closed, simple geodesic on T^*. In Chapter II we associated a λ-value to every Markoff irrationality and we proved some properties of the set of these λ-values, the Markoff values. It seems natural to also associate a λ-value to a closed, simple geodesic, namely the λ-value of all the elements of the class of Markoff irrationalities associated to it. It turns out that this λ-value also has a geometric interpretation and one can wonder whether we can use this geometric interpretation to prove the same results on the Markoff values we proved in Chapter II in a number theoretical setting.

subsection V.2.1 The λ-Value of a Geodesic

In Chapter II we discussed both the Markoff spectrum and the Lagrange spectrum, they turned out to agree on the interval $[\sqrt{5}, 3[$. The Markoff spectrum was obtained by associating a μ-value to an indefinite binary quadratic form and for the Lagrange spectrum we looked at $\lambda(a)$ for a real number a. As the two spectra are the same when it comes to Markoff values, we can associate a class of forms to a class of real numbers. It turned out that μ-value of the class of a form f with roots $r < s$ equals the λ-value of the class of s. In Section II.1 we discussed several ways to obtain $\lambda(s) = \mu(f)$, in particular we saw that $\mu(f)$ can be obtained as the supremum of the difference of the roots of all Mf. Thus, writing r_M and s_M for the roots of Mf we have that

$$\lambda(s) = \mu(f) = \sup_{M \in \mathrm{SL}(2,\mathbb{Z})} |r_M - s_M|.$$

Yet, it holds that the roots of Mf are given by $M^{-1}r$ and $M^{-1}s$ and hence we see

$$\lambda(s) = \mu(f) = \sup_{M\in \mathrm{SL}(2,\mathbb{Z})} |Mr - Ms|.$$

This difference of roots might remind us of the (Euclidean) diameter of a geodesic. This observation enables us to give a geometric interpretation of $\lambda(s)$. Indeed, consider the geodesic $\gamma_{r,s}$ joining r and s, the roots of f. The Euclidean diameter of $\gamma_{r,s}$ is $\mathrm{diam}(\gamma_{rs}) = |r - s|$ and as $M\gamma_{rs}$ has endpoints Mr and Ms we see that $\mathrm{diam}(M\gamma_{r,s}) = |Mr - Ms|$ for every $M \in \mathrm{SL}(2,\mathbb{Z})$. Consequently we have

$$\lambda(s) = \mu(f) = \sup_{M\in \mathrm{SL}(2,\mathbb{Z})} \mathrm{diam}(M\gamma_{rs}).$$

This leads the way to define the λ-value of a geodesic γ, namely

$$\lambda(\gamma) := \sup_{M\in \mathrm{SL}(2,\mathbb{Z})} \mathrm{diam}(M\gamma).$$

It follows from this definition that all the lifts of a geodesic have the same λ-value, which actually allows one to associate a λ-value to a geodesic on T^*.

Consider a lift γ of a closed, simple geodesic, we would like its λ-value to agree with the λ-value of its endpoint. For in that case the above definition of the λ-value of a geodesic is a generalization of the definition we gave before, namely to relate the λ-value of all the elements of a class of Markoff irrationalities to the closed, simple geodesic it is associated to.

The lift γ has a periodic cutting sequence, and consequently the starting point d and endpoint e of γ are fixed points of a certain quadratic polynomial. Yet, then d and e can also be obtained as the roots of another quadratic polynomial. This quadratic polynomial leads to an indefinite binary quadratic form g with roots d and e and for this form it holds that

$$\lambda(\gamma) = \mu(g) = \lambda(e).$$

Consequently we see that the λ-values of a lift of a closed, simple geodesic and its endpoint agree. The λ-value of a geodesic γ which is a lift of a limit of closed, simple geodesics is the limit of the λ-values of these geodesics. If γ is not closed itself, then this limit of λ-values must be 3. Such a lift of a limit of closed, simple geodesics has a Markoff limit as its endpoint. Hence we see that the λ-value of such a geodesic equals the λ-value of its endpoint, too.

SUBSECTION V.2.2 A GEOMETRIC APPROACH

From the definition of the λ-value of a geodesic it immediately follows that all the lifts of a closed, simple geodesic have a diameter strictly smaller than 3. Actually, for every closed, simple geodesic there is an $\epsilon > 0$ such that all of its lifts have diameter less than $3 - \epsilon$. Furthermore all the lifts of a limit of closed, simple geodesics are in diameter less than or equal to 3. We have thus proved the following theorem.

Theorem V.1. *If a geodesic γ in $\mathbb{H}$ is the lift of a closed, simple geodesic or of a limit of closed, simple geodesics, then* $\mathrm{diam}(M\gamma) \leq 3$ *for every $M \in G$.*

The proof of this theorem as given above might feel a bit unsatisfactory in the sense that we implicitly still used a lot of information from Chapter II, information that we obtained in a number theoretical setting. A more geometric approach would be to prove the theorem just using the properties of geodesics and tessellations. It is possible to give such a geometric proof of Theorem V.1.

Geometric proof of Theorem V.1. Consider the translation $t : z \mapsto z + 3$ of the hyperbolic upper half-plane. It can be seen to leave the A, B-tessellation intact. Yet, the labeling changes: A and A' are interchanged and so are B and B'.

A geodesic on T^* is simple if and only if all its lifts to $\mathbb{H}$ are disjoint. As the lifts of a certain geodesic form a G-equivalence class, we see that two G-equivalent geodesics are disjoint, if the geodesic on T^* they project to is simple. Now, consider a geodesic $\overline{\gamma}$ on T^* which is closed and simple or a limit of closed, simple geodesics and one of its lifts γ.

It holds that γ has a characteristic cutting sequence s_γ in (at most) two symbols, say X and Y. The cutting sequence $s_{t(\gamma)}$ of $t(\gamma)$ is obtained from s_γ by replacing every occurrence of X by X' and every occurrence of Y by Y'. There are two ways to direct $t(\gamma)$. The cutting sequence $s_{t(\gamma)}$ in X', Y' comes from the orientation inherited by the orientation of γ. If we direct $t(\gamma)$ the other way around, we find its cutting sequence by reading $s_{t(\gamma)}$ backwards and replace X' by X and Y' by Y. Consequently the cutting sequence of $t(\gamma)$ with the alternative orientation is the reversal of s_γ.

Recall Corollary III.1 which implies that any characteristic sequence is symmetric. Hence $t(\gamma)$ with the alternative orientation, called γ' from now on, also has cutting sequence s_γ. Note that this implies that γ and γ' are G-equivalent. As γ is the lift of a simple geodesic, it now holds that γ and γ' must be disjoint. One can easily see that diam$(\gamma) \leq 3$ if and only if $\gamma \cap \gamma' = \emptyset$. As γ was chosen to be an arbitrary lift of $\overline{\gamma}$ we can conclude that diam$(\gamma) \leq 3$ for any lift γ of $\overline{\gamma}$. □

In case we consider only closed, simple geodesics, we can make the theorem a bit stronger, we can prove that all the lifts of such a geodesic have a diameter strictly less than 3. Indeed, suppose γ is the lift of a closed, simple geodesic $\overline{\gamma}$, we would like to prove that diam$(\gamma) < 3$. We proceed by contradiction, so assume diam$(\gamma) = 3$. This implies that γ and γ', as defined in the above proof, are disjoint, they meet however at the real line. Hence these two geodesics become arbitrarily close to each other.

Projecting γ and γ' to T^* this means that we can find points on $\overline{\gamma}$, arbitrarily close to each other, that have the property that the neighborhood containing a geodesic segment connecting these points cannot be chosen arbitrarily small. This can not happen, because the closedness of $\overline{\gamma}$ implies that there is an $\epsilon > 0$ such that the distance between any two points with this property is at least ϵ. Consequently it holds that all the lifts of $\overline{\gamma}$ have a diameter less than 3, actually the above implies that all its lifts have a diameter less than $3 - \epsilon$ for some $\epsilon > 0$. This proves Theorem V.2.

Theorem V.2. *If a geodesic γ in $\mathbb{H}$ is the lift of a closed, simple geodesic, then there is an $\epsilon > 0$ such that* diam$(M\gamma) < 3 - \epsilon$ *for every $M \in G$.*

The converse of Theorem V.1 is also true, Theorem V.3. A (geometric) proof of this is given in Series (1985a). This proof proceeds by contradiction, it proves that there exists a geodesic γ with cutting sequence s and diam$(\gamma) > 3$ for every non-characteristic sequence s. Note that Theorem V.1 and V.3 combined imply that closed, simple geodesics are precisely the geodesics whose λ-values are Markoff values and that the limits of closed, simple geodesics, that are not closed itself, are exactly the geodesics with λ-value 3.

Theorem V.3. *If a geodesic γ in $\mathbb{H}$ is such that* diam$(M\gamma) \leq 3$ *for every $M \in G$, then γ is the lift of either a closed, simple geodesic or of a limit of closed, simple geodesics.*

SECTION V.3 Finding the Markoff Values

In the previous section we have seen that the λ-values of closed, simple geodesics are Markoff values. This follows quite directly from the definition of the λ-value, but it can also be proven using only properties of geodesics and tessellations. We are already familiar with some properties of these Markoff values, in the end of Chapter II we were able to explicitly compute them.

We would like to gain similar results about the Markoff values, now defined as all possible values $\lambda(\gamma) < 3$, in a purely geometric way. To be able to do this we first need to show that we lose no information. More precisely, we need to know whether all Markoff values, in the old definition, are still obtained using the new definition. Concretely this means that every class of Markoff irrationalities can be associated to a closed, simple geodesic.

SUBSECTION V.3.1 Markoff Irrationalities and Geodesics

In order to associate every Markoff irrationality to a closed, simple geodesic, it is enough to show that every class of Markoff irrationalities has an element that occurs as the endpoint of a lift of a closed, simple geodesic. To prove this it helps to consider the classes of Markoff irrationalities in a bit more detail.

Every two elements of the same class have the same tail. Consequently we can associate to every class of Markoff irrationalities an infinite periodic sequence, the tail of every number in this class. This sequence contains 1's and 2's such that the 2's only come in 2-tuples and the 1's in $2n$- or $2(n+1)$-tuples for some $n \geq 0$. This sequence of 1's and 2's is completely determined by the sequence in which the n's and $(n+1)$'s occur. Consequently we can associate an infinite periodic sequence of n's and $(n+1)$'s to a class of Markoff irrationalities.

The periodic tail of a class is known to agree with the tail of a purely periodic sequence A whose associated sequence is purely periodic and Markoff balanced. The infinite periodic sequence of n's and $(n+1)$'s we just found is then the tail of this associated sequence. Recall Lemma III.4 which shows that a periodic Markoff balanced sequence is characteristic. Hence it now holds that the infinite periodic sequence of n's and $(n+1)$'s coming from the tail of any class of Markoff irrationalities is the tail of a periodic, characteristic sequence.

Consider now a class of Markoff irrationalities and its associated infinite periodic sequence of n's and $(n+1)$'s. We just saw that this sequence is the tail of a periodic, characteristic sequence in n's and $(n+1)$'s. We can make this into a periodic, characteristic sequence in $(n+2)$'s and $(n+3)$'s by replacing every occurrence of n by $n+2$ and every occurrence of $n+1$ by $n+3$. Let this sequence in $(n+2)$'s and $(n+3)$'s be the associated sequence of a periodic sequence in A and B, this sequence must then be characteristic. This implies that we can find a closed and simple geodesic with this sequence as a cutting sequence. Putting any lift γ through the procedure described in Section V.1.2 we find that the endpoint of γ is a Markoff irrationality which has the same tail as the tail of the class of Markoff irrationalities we began with. Hence, we have found an element of this class that occurs as the endpoint of a lift of a closed and simple geodesic. This proves that every class of Markoff irrationalities can be associated to a closed, simple geodesic.

Consequently we can safely consider the Markoff values from our new, geometric, viewpoint, without losing any information. These Markoff values can be computed explicitly, in Haas (1986) they are expressed in the hyperbolic length of geodesic on T^*.

SUBSECTION V.3.2 Markoff Irrationalities and their Tails

Every class of Markoff irrationalities has a representative that occurs as the endpoint of a closed, simple geodesic. However, it is not true that every Markoff irrationality occurs as an endpoint of a lift of a closed, simple geodesic. To prove this recall Theorem III.3, it states that two geodesics in $\mathbb{H}$ have the same cutting sequence if and only if they are in the same G-equivalence class. This implies that two geodesics are lifts of the same geodesic on T^* if and only if they are G-equivalent. Now, if a geodesic γ has endpoint e, then $M\gamma$ has endpoint Me for $M \in G$. Consequently we see that it can only be the case that every element of a class of Markoff irrationalities occurs as the endpoint of a lift of a closed, simple geodesic, if all of these elements are equivalent under G. Lemma V.2 proves that this is not true.

Definition V.1. Two continued fractions $a = \pm[a_0, a_1, ...]$ and $b = \pm[b_0, b_1, ...]$ are said to have the *same tail modulo* 2 if there exist m, n such that $a_{m+k} = b_{n+k}$ for all $k \geq 0$ and $m + n$ is even if $ab > 0$ and $n + m$ is odd if $ab < 0$.

Lemma V.2. *Two real numbers $a, b \in \mathbb{R}$ are* $\mathrm{SL}(2, \mathbb{Z})$*-equivalent if and only if they have the same tails modulo* 2*.*

Note that a class of Markoff irrationalities consists of all numbers with a certain tail, yet not all of these numbers have the same tail modulo 2. Consequently Lemma V.2 shows that the elements of a certain class of Markoff irrationalities are not all $\mathrm{SL}(2, \mathbb{Z})$-equivalent, and hence in particular not all G-equivalent.

Proof. Consider a, b which have the same tails modulo 2. Suppose they are both positive, say $a = [a_0, a_1, ...]$ and $b = [b_0, b_1, ...]$. The maps $a_0 : z \to \frac{-1}{z-a_0}$ and $b_0 : z \to \frac{-1}{z-b_0}$ both come from matrices $a_0, b_0 \in \mathrm{SL}(2,\mathbb{Z})$, and we can compute $a_0(a) = -[a_1, a_2, ...] := a'$ and $b_0(b) = -[b_1, b_2, ...] := b'$. As a_0, b_0 are matrices in $\mathrm{SL}(2,\mathbb{Z})$ it now holds that a and a' are $\mathrm{SL}(2,\mathbb{Z})$-equivalent and b and b', too. In the same way we can now define maps $a_1 : z \to \frac{-1}{z+a_0}$ and $b_1 : z \to \frac{-1}{z-b_0}$ coming from matrices $a_1, b_1 \in \mathrm{SL}(2,\mathbb{Z})$, and it holds that $a_1(a') = [a_2, a_3, ...] := a''$ and $b_1(b') = [b_2, b_3 ...] := b''$. This tells us that a, a'' are $\mathrm{SL}(2,\mathbb{Z})$-equivalent and b, b'', too.

We can continue in this fashion, the plus- and minus signs flip at every step. As a and b have the same tail modulo 2, we see that at some point $a^{(m)}$ and $b^{(n)}$ are equal for some n, m which are either both even or both odd. This parity of m and n is needed for the signs of $a^{(m)}$ and $b^{(n)}$ to agree. So a and b are then $\mathrm{SL}(2,\mathbb{Z})$-equivalent to the same number, which implies that a and b are itself $\mathrm{SL}(2,\mathbb{Z})$-equivalent. Note that the proof for a and b which are both negative is completely the same. If one of a and b is positive and the other is negative, then the same also holds for a', b' and so on. Note that in this case having the same tail modulo 2 means that at some point $a^{(n)}$ and $b^{(m)}$ are equal for m, n of which one is even and one is odd. Again this parity of m, n is needed for the signs to agree.

To prove the converse statement, consider $a, b \in \mathbb{R}$ which are $\mathrm{SL}(2,\mathbb{Z})$-equivalent. Thus there is some $M \in \mathrm{SL}(2,\mathbb{Z})$ such that $Ma = b$. First suppose that both a and b are positive, this means that we can write $a = [a_0, a_1, ...]$ and $b = [b_0, b_1, ...]$. Now, pick $c \in \mathbb{R}_{<0}$ and consider the geodesics γ_a connecting c to a and γ_b joining c and b. Theorem III.2 tells us that the L, R-sequence of γ_a ends in $L^{a_0}R^{a_1}L^{a_2}...$, and that $L^{b_0}R^{b_1}L^{b_2}...$ is the tail of the L, R-sequence of γ_b.

As equivalent geodesics have the same L, R-sequence we know that the tails of γ_a and $M\gamma_a$ agree. Yet, it holds that $Ma = b$ and hence $M\gamma_a$ and γ_b both end in b. This implies that the tails of $M\gamma_a$ and γ_b will agree at some point. Consequently the sequences $L^{a_0}R^{a_1}L^{a_2}...$ and $L^{b_0}R^{b_1}L^{b_2}...$ will become equal eventually. This can only be the case if there are m and n with $m + n$ even such that $a_{m+k} = b_{n+k}$ for all $k \geq 0$, i.e. a and b have the same tails modulo 2.

In case a and b are both negative the proof is exactly the same, we need to compare tails $R^{a_0}L^{a_1}R^{a_2}$ and $R^{b_0}L^{b_1}R^{b_2}$ in this case. If a and b have different sign, say a is positive and b is negative, then we have to compare $R^{a_0}L^{a_1}R^{a_2}$ to $L^{a_0}R^{a_1}L^{a_2}$. If these tails agree at some point then there are m, n with $m + n$ odd such that $a_{m+k} = b_{n+k}$ for all $k \geq 0$. Consequently we also see in this case that a and b have the same tails modulo 2. □

Remark V.3. The main purpose of this chapter is to consider the Markoff theory, part of the field of number theory, in a geometrical setting. It is also possible to consider continued fractions in a more geometrical way. For example we see that the proof of Lemma V.2 above depends on the L, R-sequences of certain geodesics. Irwin (1989) approaches continued fractions in a geometric way and proves some of the results discussed in Chapter I in a geometric setting. Furthermore, Series (1985b) gives a proof of Lemma I.11, Lemma I.10 and Lemma I.9 using geodesics.

Linking Number Theory to Geometry

In the past three sections we have considered a geometric way to look at the Markoff values and we saw that it is possible to compute these values explicitly via this geometric approach. This gives us two ways to look at the Markoff values. We can consider them number theoretically, namely through the Markoff or Lagrange spectrum, or geometrically, using geodesics in $\mathbb{H}$. This situation allows one to convert number theoretical results to geometrical results, and vise versa.

The remainder of this chapter is devoted to a nice geometrical interpretation of the λ-value of a real number and to a geometrical way to look at the Markoff numbers.

section V.4 A Link to the Langrange Spectrum

In this chapter we have seen that all the Markoff values can be found by considering all possible λ-values $\lambda(\gamma)$ below 3. The definition of this value $\lambda(\gamma)$ can be derived from the μ-value of a certain binary quadratic form. This leads us back to the Markoff spectrum, hence in a way the definition of the λ-value of a geodesic comes from the definition of the μ-values making up the Markoff spectrum $\mathbb{M}$.

Besides $\mathbb{M}$ we have also considered the Lagrange spectrum $\mathbb{L}$, and we actually saw that $\mathbb{M}$ and $\mathbb{L}$ agree on the interval $[\sqrt{5}, 3[$. The λ-value of a certain real number a is defined as

$$\lambda(a) = \sup \left\{ c \in \mathbb{R} : |a - \frac{p}{q}| < \frac{1}{cq^2} \text{ for infinitely many } p, q \in \mathbb{Z} \right\}.$$

In this definition we consider differences between a real number and rational numbers. In a way we have to check whether a certain rational number lies in an open interval, or ball, around a. This interpretation with open balls brings us back to geometry, which leads to a nice geometric way of checking whether some $a \in \mathbb{R}$ is a Markoff irrationality.

SUBSECTION V.4.1 Horocycles and Markoff Irrationalities

For $p, q \in \mathbb{Z}$ consider the open circle in $\mathbb{H}$ tangent at $\frac{p}{q}$, of (Euclidean) radius $\frac{1}{3q^2}$. We call this the *horocycle* belonging to p, q. Letting coprime p and q run through $\mathbb{Z}$ we obtain a set of horocycles, say C. Now, for $a \in \mathbb{R}$ consider the vertical line $l(a)$ joining a and ∞. If this line cuts the horocycle belonging to p, q it holds that $|a - \frac{p}{q}| < \frac{1}{3q^2}$. Consequently it is the case that $\lambda(a)$ is a Markoff value if and only if $l(a)$ cuts only a finite number of the horocycles in C and is furthermore arbitrarily close to only a finite number of elements of C.

We also know that $\lambda(a) < 3$ if and only if the tail of the continued fraction expansion of a agrees with the tail of a periodic, characteristic sequence. This proves the following lemma.

Lemma V.3. *For $a \in \mathbb{R}$ the vertical line $l(a)$ joining a and ∞ cuts only a finite number of the horocycles in C and is furthermore arbitrarily close to only a finite number of elements of C if and only if the tail of a agrees with the tail of a periodic, characteristic sequence.*

The proof of this lemma depends on facts from Chapter II, but Theorem V.2 can be used to give more insight in this equivalence in a geometric way. This theorem states that a geodesic γ in $\mathbb{H}$ is the lift of a closed, simple geodesic, or the limit of closed, simple geodesics if and only if $\operatorname{diam}(M\gamma) \leq 3$ for all $M \in G$. This means that such a geodesic γ and all of its images under G stay out of the region $H = \{z \in \mathbb{H} | \Im(z) > \frac{3}{2}\}$. It then holds that the image $\overline{\gamma}$ of γ on T^* avoids the image $\overline{H}$ of H in T^*. Yet, then the lifts of $\overline{\gamma}$, i.e. γ and its G-images, do not only avoid H but all the lifts of $\overline{H}$ to $\mathbb{H}$. These lifts of $\overline{H}$ turn out to be exactly the horocycles in C.

The lifts of $\overline{H}$ to $\mathbb{H}$ are precisely $M(H)$ for $M \in G = \langle A, B \rangle$. Writing $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$ one can compute that the image of H under M is an open circle which is tangent at $\frac{a}{c}$ of Euclidean radius $\frac{1}{3c^2}$. Thus the image $M(H)$ is a horocycle belonging to a, c. Indeed, it is a straightforward computation that the difference between $\frac{a}{c} + \frac{i}{3c^2}$, the center of the horocycle belonging to $\frac{a}{c}$, and any $x + yi$ is given by

$$\sqrt{\frac{1}{9c^4} + \frac{3 - 2y}{3c^2((d + cx)^2 + c^2y^2)}}.$$

Here we do not really need the fact that $M \in G$, but we do need that the determinant of M equals 1. Hence the difference between $\frac{a}{c} + \frac{i}{3c^2}$ and a number with imaginary-part equal to $\frac{3}{2}$ is precisely $\frac{1}{3c^2}$, while the difference between $\frac{a}{c} + \frac{i}{3c^2}$ and $z \in \mathbb{H}$ with $\Im(z) > \frac{3}{2}$ is strictly less than $\frac{1}{3c^2}$. Along with $M(\infty) = \frac{a}{c}$ this results in the fact that $M(H)$ is precisely the open circle tangent at $\frac{a}{c}$ of radius $\frac{1}{3c^2}$. We thus find that the lifts of $\overline{H}$ to $\mathbb{H}$ are precisely the horocycles belonging to a, c such that there is an $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$.

We claim that for every rational number $\frac{a}{c}$ there is a horocycle $M(H)$ tangent to it. Indeed, Lemma III.5 tells us that the vertex-set of the Farey-tessellation is $\mathbb{Q} \cup \{\infty\}$ and from the construction of the A, B-tessellation as a subtessellation of the Farey tessellation it immediately follows that the A, B-tessellation has the same set of vertices. This set of vertices can be obtained by considering $M(\infty)$ for $M \in G$ and this is exactly the rational number the horocycle $M(H)$ is tangent to. From this it is clear that every element of $\mathbb{Q}$ has a horocycle $M(H)$ tangent to it. Furthermore, for $M \in G \subset \mathrm{SL}(2, \mathbb{Z})$ it holds that the horocycle $M(H)$ belongs to a, c that are coprime. We can thus conclude that the set of horocycles $M(H)$ equals the set C. These observations allows us to gain more insight in one of the implications of Lemma V.3:

For $a \in \mathbb{R}$ the vertical line $l(a)$ joining a and ∞ cuts only a finite number of the elements in C and is arbitrarily close to only a finite number of horocycles in C if the tail of a agrees with the tail of a periodic, characteristic sequence.

For such a there might not be a closed, simple geodesic with a as its endpoint. If it does hold however that a is the endpoint of a closed, simple geodesic γ_a, then we can give an alternative proof of the above implication.

Indeed, as γ_a is closed and simple we can use Theorem V.2 to deduce that there is an $\epsilon > 0$ with $\mathrm{diam}(M\gamma_a) < 3-\epsilon$ for every $M \in G$. From the above this implies that γ_a avoids the horocycles in C, and more, it even stays at bounded distance from them. Consider now the vertical line $l(a)$ joining a and ∞. It meets γ_a in a and as such at some point $l(a)$ becomes a bounded distance from the horocycles in C, too. This can only be the case if the line $l(a)$ cuts only a finite number of these horocycles and is arbitrarily close to only a finite number of them.

We can apply the same argument to $a \in \mathbb{R}$ that do not have a tail agreeing with the tail of a periodic, characteristic sequence. For example, consider a real number a which has a tail that agrees with the tail of a characteristic sequence which is not periodic. Such a characteristic sequence corresponds to a limit of closed, simple geodesics. In case a is the endpoint of such a limit γ we can use Theorem V.1 to deduce that $\mathrm{diam}(M\gamma) \leq 3$ for every $M \in G$. We can view γ as a limit of closed, simple geodesics $\{\gamma_n\}_{n\geq 0}$, each of these geodesics stays a bounded distance from the horocycles in C, yet this distance tends to zero as $n \to \infty$. This implies that the limit γ approaches the horocycles in C arbitrarily close. As the vertical line $l(a)$ and γ have the same endpoint we can conclude that $l(a)$ cuts only a finite number of the horocycles in C, but is arbitrarily close to infinitely many of them. This nicely agrees with the fact that $\lambda(a) = 3$.

We can also gain information about the behavior of $l(a)$ for $a \in \mathbb{R}$ which has a non-characteristic tail. Indeed, any geodesic with a as its endpoint cannot be closed and simple or a limit of closed, simple geodesics. Theorem V.2 then tells us that such a geodesic γ has at least one G-equivalent geodesic that penetrates the region H. This implies that γ cuts infinitely many of the horocycles in C and consequently $l(a)$ will do the same thing. Again this agrees nicely with known facts about a, namely that $\lambda(a) > 3$. This shows that we can find information about the λ-values of real numbers by considering geometric objects as horocycles and geodesics.

SECTION V.5 Markoff Numbers viewed Geometrically

This chapter is about a geometric interpretation of the Markoff values. As part of Chapter II about Markoff theory we also discussed the Markoff numbers. It turns out that these Markoff numbers have geometric counterparts, too. There is a correspondence between triples of Markoff numbers, i.e. triples of positive numbers that form a solution to (II.3)

$$x^2 + y^2 + z^2 = 3xyz,$$

and triples of traces of three special matrices. We know that we can find all the solutions of (II.3) from the solution $(1, 1, 1)$ by applying the operations $r : (x, y, z) \to (z, x, y)$ and $s : (x, y, z) \to (x, 3xy - z, y)$, see Theorem II.8. This special solution $(1, 1, 1)$ can be associated to a triple of matrices and these matrices turn out to be precisely A, B and $B^{-1}A^{-1}$.

SUBSECTION V.5.1 Fricke's Trace Identity and Markoff Numbers

The matrices A and B are known to identify opposite sides of the domains of the A, B-tessellation, in particular they identify the sides of the square with vertices -1, 0, 1 and ∞. It can easily be seen that the images of ∞ under AB and BA are equal, and hence ∞ is a fixed point of the commutator $B^{-1}A^{-1}BA$. Consider now in more generality two matrices $M, N \in \mathrm{SL}(2, \mathbb{Z})$ that identify opposite sides of a square $S_{M,N}$ in $\mathbb{H}$, one of whose vertices is ∞ and the other vertices are rational.

In the same way as with A and B it holds that $K := M^{-1}N^{-1}MN$ has ∞ as fixed point. Combining this with the fact that $K \in \mathrm{SL}(2,\mathbb{Z})$ we find that K can be written as $\begin{pmatrix} \pm 1 & b \\ 0 & \pm 1 \end{pmatrix}$ for some $b \in \mathbb{Z}$. In fact, the ±-signs must always be minus-signs, as can be shown by straightforward computation. From $MN = NMK$ it follows in particular that the lower left-hand entries of these matrices must be equal. Writing $M = (m_{ij})_{i,j=1,2}$ and $N = (n_{ij})_{i,j=1,2}$ this amounts to the equality $m_{21}n_{11} + m_{22}n_{21} = \pm(m_{11}n_{21} + m_{21}n_{22})$. Furthermore, one can compute that

$$N(\infty) - NM(\infty) = \frac{n_{11}}{n_{21}} - \frac{m_{11}n_{11} + m_{21}n_{12}}{m_{11}n_{21} + m_{21}n_{22}} = \frac{m_{21}}{n_{21}(m_{11}n_{21} + m_{21}n_{22})};$$

$$MN(\infty) - M(\infty) = \frac{m_{11}n_{11} + m_{12}n_{21}}{m_{21}n_{11} + m_{22}n_{21}} - \frac{m_{11}}{m_{21}} = \frac{-n_{21}}{m_{21}(m_{21}n_{11} + m_{22}n_{21})}.$$

This leads to

$$-(m_{21}n_{11} + m_{22}n_{21})(m_{11}n_{21} + m_{21}n_{22}) = (N(\infty) - NM(\infty))(MN(\infty) - M(\infty)) > 0.$$

The positivity comes from the fact that $MN(\infty) = NM(\infty)$ lies between $N(\infty)$ and $M(\infty)$, which can be seen by viewing them as three vertices of $S_{M,N}$. We conclude that $m_{21}n_{11} + m_{22}n_{21}$ and $(m_{11}n_{21} + m_{21}n_{22})$ must have opposite signs, and hence the plus-sign cannot occur in the equality $m_{21}n_{11} + m_{22}n_{21} = \pm(m_{11}n_{21} + m_{21}n_{22})$.

Consider now $G_{M,N}$, the free group generated by M and N. Of course $G_{M,N}$ is also generated by the triple M, N and $M^{-1}N^{-1}$. We will show how to connect this triple to a triple of Markoff numbers, this can be done using one of Fricke's trace identities for matrices $X, Y \in \mathrm{SL}(2,\mathbb{Z})$

$$\mathrm{tr}(X)^2 + \mathrm{tr}(Y)^2 + \mathrm{tr}(X^{-1}Y^{-1})^2 = \mathrm{tr}(X)\mathrm{tr}(Y)\mathrm{tr}(X^{-1}Y^{-1}) + \mathrm{tr}(X^{-1}Y^{-1}XY) + 2.$$

This identity can be simplified a bit when applied to the matrices M and N. The above implies that the trace of $M^{-1}N^{-1}MN$ equals -2 and hence the trace identity of Fricke becomes

$$\mathrm{tr}(X)^2 + \mathrm{tr}(Y)^2 + \mathrm{tr}(X^{-1}Y^{-1})^2 = \mathrm{tr}(X)\mathrm{tr}(Y)\mathrm{tr}(X^{-1}Y^{-1}) \tag{V.1}$$

for matrices X, Y similar to M and N. This identity might remind us a bit of (II.3), and actually it is now an easy task to transform (V.1) into (II.3). To do this, consider (V.1) modulo 3. It can easily be seen that (V.1) cannot hold modulo 3 if $\mathrm{tr}(X)$, $\mathrm{tr}(Y)$ and $\mathrm{tr}(X^{-1}Y^{-1})$ are not all divisible by 3. This allows us to write $\mathrm{tr}(X) = 3x$, $\mathrm{tr}(Y) = 3y$ and $\mathrm{tr}(X^{-1}Y^{-1}) = 3z$ and this transforms (V.1) into

$$x^2 + y^2 + z^2 = 3xyz$$

which we recognize as (II.3). We conclude that the traces of M, N and $M^{-1}N^{-1}$ divided by 3 make up a triple of Markoff numbers.

SUBSECTION V.5.2 Changing Generators

We now know that the traces of X, Y and $X^{-1}Y^{-1}$ divided by 3 make up a Markoff triple for matrices X and Y similar to M and N discussed above. The Markoff triples can be found by considering images of $(1,1,1)$ under two maps, r and s. An endomorphism of $G_{M,N}$ induces a map on the traces of the matrices in this group. It turns out that the endomorphisms inducing r and s are in fact automorphisms, i.e. a map that sends generators to generators.

To prove this we are interested in the automorphisms of $G_{M,N}$. It can be shown that the automorphism group of $G_{M,N}$ is generated by the *Nielsen transformations*, see Nielsen (1924). In case of $G_{M,N}$ there are three Nielsen transformations f, g and h, given by

$$f(M) = N \text{ and } f(N) = M,\ g(M) = M^{-1} \text{ and } g(N) = N,\ h(M) = MN \text{ and } h(N) = N.$$

These transformations on M and N induce operations on the traces of these matrices. As these traces are related to Markoff numbers we thus obtain a operations on these numbers. One can wonder how the induced operations of f, g and h operate on a triple of Markoff numbers. As f sends M to N and vice versa, it holds that the operation induced by f is such that

$$\text{tr}(M) \mapsto \text{tr}(N),\ \text{tr}(N) \mapsto \text{tr}(M) \text{ and } \text{tr}(M^{-1}N^{-1}) \mapsto \text{tr}(N^{-1}M^{-1}) = \text{tr}(M^{-1}N^{-1}).$$

We hence have that the triple (x, y, z) of Markoff numbers given by tr$(M) = 3x$, tr$(N) = 3y$ and tr$(M^{-1}N^{-1}) = 3z$ is sent to (y, x, z).

In the same way we can compute how the operations induced by g and h affect Markoff numbers. To this end we use another trace identity of Fricke:

$$\text{tr}(YXY) + \text{tr}(X) = \text{tr}(Y)\text{tr}(XY). \tag{V.2}$$

The transformation g is such that tr$(M) \mapsto$ tr$(M^{-1}) =$ tr(M) and tr$(N) \mapsto$ tr(N). The effect on tr$(M^{-1}N^{-1})$ needs a little more computation, we use (V.2). With $X = N^{-1}M$ and $Y = M^{-1}$ this identity becomes tr$(MN^{-1}) =$ tr(M^{-1})tr$(N^{-1}) -$ tr$(M^{-1}N^{-1})$. Thus we find

$$\text{tr}(M^{-1}N^{-1}) \mapsto \text{tr}(MN^{-1}) = \text{tr}(M^{-1})\text{tr}(N^{-1}) - \text{tr}(M^{-1}N^{-1}).$$

This implies that the triple (x, y, z) of Markoff numbers given by tr$(M) = 3x$, tr$(N) = 3y$ and tr$(M^{-1}N^{-1}) = 3z$ is sent to $(x, y, 3xy - z)$. Finally, h induces an operation with

$$\text{tr}(M) \mapsto \text{tr}(MN) = \text{tr}(M^{-1}N^{-1}) \text{ and } \text{tr}(N) \mapsto \text{tr}(N).$$

To compute the effect h has on tr$(M^{-1}N^{-1})$ we consider (V.2) with $X = M^{-1}$ and $Y = N^{-1}$:

$$\text{tr}(M^{-1}N^{-1}) \mapsto \text{tr}(N^{-1}M^{-1}N^{-1}) = \text{tr}(N^{-1})\text{tr}(M^{-1}N^{-1}) - \text{tr}(M^{-1}) = \text{tr}(N)\text{tr}(M^{-1}N^{-1}) - \text{tr}(M).$$

We find that the triple (x, y, z) of Markoff numbers gets sent to $(z, y, 3yz-x)$. These three maps on triples of Markoff numbers clearly have similarities with the operations $r : (x, y, z) \to (z, x, y)$ and $s : (x, y, z) \to (x, 3xy - z, y)$. In fact, one can express r and s as compositions of the maps induced by f, g and h which we will now call f, g and h, too. One can compute that $r = f \circ g \circ f \circ h \circ f$ and $s = f \circ r \circ g$. Indeed,

$$\begin{aligned} f \circ g \circ f \circ h \circ f(x, y, z) &= f \circ g \circ f \circ h((y, x, z) = f \circ g \circ f(z, x, 3xz - y) \\ &= f \circ g(x, z, 3xz - y) = f(x, z, y) = (z, x, y) = r(x, y, z); \\ f \circ r \circ g(x, y, z) &= f \circ r(x, y, 3xy - z) = f(3xy - z, x, y) = (x, 3xy - z, y) = s(x, y, z). \end{aligned}$$

This means that we can also find all the solutions of (II.3) from $(1, 1, 1)$ using the operations f, g and h.

Starting from the matrices M and N and their corresponding triple of Markoff numbers (x, y, z) we know from Theorem II.8 that (x, y, z) can be found by applying a certain composition of r and s to $(1, 1, 1)$. Applying the inverse of the map that induces this composition to the matrices M, N and $M^{-1}N^{-1}$ we find a triple of matrices whose corresponding Markoff triple is exactly $(1, 1, 1)$. Consequently all three of these matrices must have trace 3. Using this information on the traces we can explicitly compute these matrices, they turn out to be

$$B = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix},\ A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \text{ and } B^{-1}A^{-1} = \begin{pmatrix} 3 & -1 \\ 1 & 0 \end{pmatrix},$$

precisely the matrices A and B we used to construct the A, B-tessellation. As these are also generators for the group $G_{M,N}$ we see that this group is just G.

The above argumentation shows that we can now associate the simplest solution $(1, 1, 1)$ of (II.3) to the A, B-tessellation. Different generators of the group G lead the way to tessellations similar to the A, B-tessellation. To such a tessellation we can also associate a triple of Markoff numbers, which then brings us back to Markoff theory again.

Chapter VI
Self-Intersecting Geodesics

In the last chapters we have restricted ourselves to closed, simple geodesics and their limits. We saw that the λ-values of such geodesics are precisely the Markoff values in case of closed and simple geodesics, their limits yield a λ-value 3. Consequently, the fact that a geodesic has λ-value less or equal to 3 can be captured in some restrictions on its behavior. One can wonder what happens if we alter those restrictions. For example, can we describe the set of λ-values that closed geodesics with one self-intersection can attain? The answer to this question is given in Crisp and Moran (1993) and Crisp et al. (1998) deals with closed geodesics that intersect themselves twice.

Theorem V.3 already tells us that the λ-values of such geodesics are bigger than 3. As these geodesics are closed we know that the starting point and endpoint of any of its lifts to $\mathbb{H}$ are roots of a certain binary form. It then follows immediately from the definition of the λ-value of a geodesic that the λ-value of closed geodesics is in the Markoff spectrum. Consequently, results about the Markoff spectrum, this time not restricted to the interval $[\sqrt{5}, 3[$, may help in obtaining results on these closed, self-intersecting geodesics.

The behavior of the Markoff spectrum above 3 is not as well-known as that below 3, however, some results have been proven. For example, it is known that all real numbers greater than or equal to $\mu := 4+[0,3,2,1,1,\ \overline{3,1,3,1,2,1}\,]+[0,4,3,2,2,\ \overline{3,1,3,1,2,1}\,] \approx 4.5278$ belong to the Markoff spectrum. The interval $[\mu, \infty[$ is known as *Hall's ray*. The behavior of the Markoff spectrum between 3 and μ is still rather mysterious. We do know that certain numbers, such as $\sqrt{13}$, belong to the spectrum and that they are in fact isolated.

We briefly discuss the results obtained in Crisp and Moran (1993) and Crisp et al. (1998). It turns out that we can completely determine the cutting sequences closed geodesics with one or two self-intersections can have. This information on the cutting sequences brings us back to the Markoff and Lagrange spectrum. Indeed, consider such a geodesic $\overline{\gamma}$ and any of its lifts γ to $\mathbb{H}$. Just as in V.1.1 we can convert its cutting sequences to an L, R-sequence. With this L, R-sequence we can determine the tail of the continued fraction expansion of the endpoint of γ. This allows us to associate a class of real numbers to $\overline{\gamma}$, namely precisely all those numbers whose continued fractions expansions have this tail. All these numbers have the same λ-value, leading to a single element of the Lagrange spectrum $\mathbb{L}$. This is also an element of the Markoff spectrum as $\mathbb{L} \subset \mathbb{M}$. The value found is exactly the λ-value of the geodesic $\overline{\gamma}$.

section VI.1 Closed Geodesics with One Self-Intersection

We state results of Crisp and Moran (1993) without proof, the proofs of these results are mainly based on group theory. The closed single self-intersection geodesics, *cssi geodesics* for short, can be grouped into two sets. If a cssi geodesic encircles the puncture of T^*, then we call this geodesic *improper* and in case a cssi geodesic does not encircle the puncture we call it *proper*. One can quite easily see that the λ-value of an improper cssi geodesic lies in Hall's ray. Indeed, if a geodesic $\overline{\gamma}$ encircles the puncture, then its cutting sequence must contain a commutator, say

$AB'A'B$. Consequently one of its lift looks like the geodesic in Figure VI.1. The diameter of this geodesic is larger than 5 and consequently $\lambda(\overline{\gamma})$ is bigger than 5, which lies in Hall's ray.

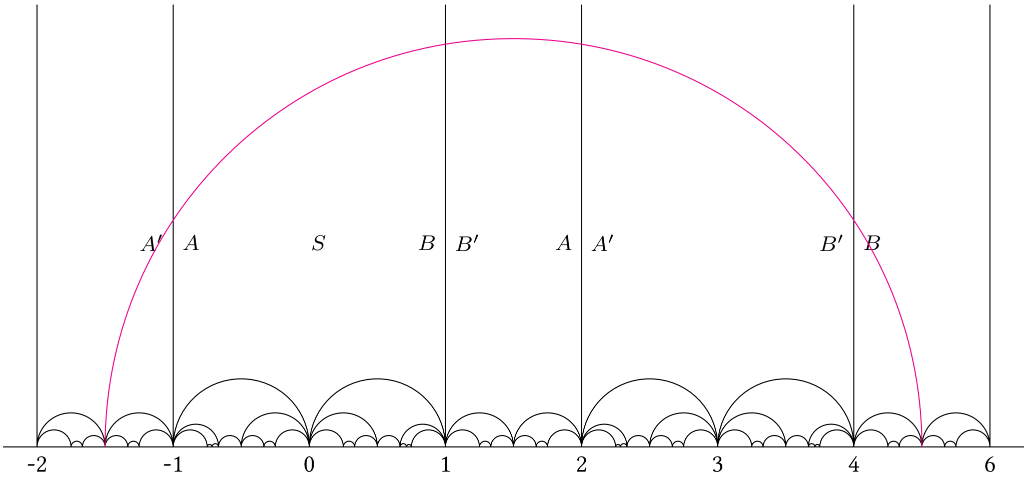


Figure VI.1.: Example of a geodesic whose cutting sequence contains a commutator

The proper CSSI geodesic are more noteworthy. In can be shown that a geodesic is a proper CSSI geodesic precisely if its cutting sequence has period X^2Y^2 such that (X, Y) is a generating pair of the group $G = \langle A, B\rangle$. Consider a geodesic $\overline{\gamma}$ with such a cutting sequence and any of its lifts γ. The cutting sequence is known to be convertible to an L, R-sequence and from this L, R-sequence we can read of the tail of the continued fractions expansion of the endpoint of γ. Different generating pairs (X, Y) may lead to the same tail. When this redundancy is removed we end up with two maps

$$\phi : (A, B) \to (AB, B) \text{ and } \psi : (A, B) \to (A, AB)$$

such that all essentially different generating pairs can be obtained from A and B by repeatedly applying ϕ and ψ. In Figure VI.2 we see a tree containing all essentially different generating pairs and a tree with the tails obtained from these pairs. The comma's between the 1's and 2's in the right tree are omitted for aesthetic reasons.

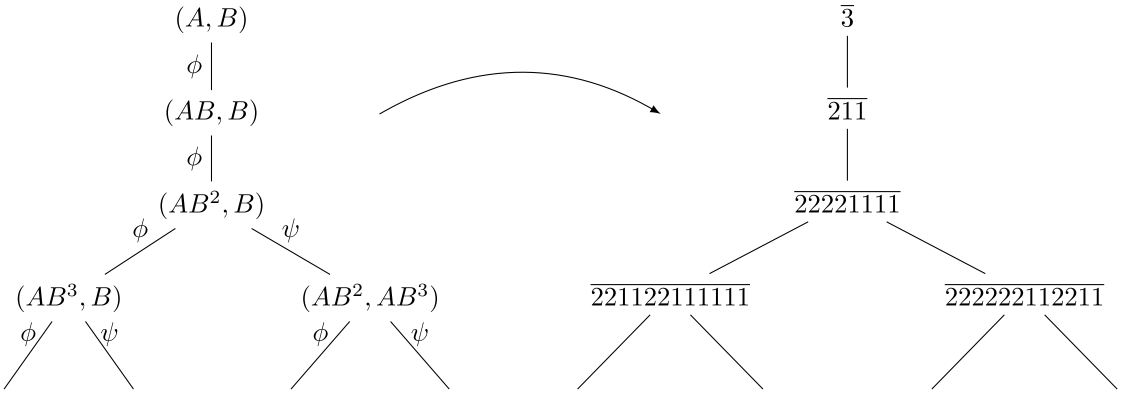
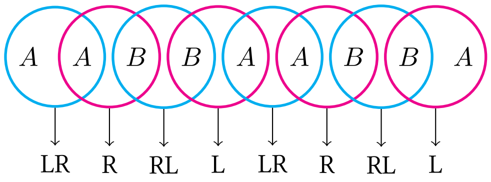


Figure VI.2.: Tree of essentially different generating pairs and tree of tails obtained from them

For example, the generating pair (A, B) leads to a cutting sequence $\overline{A^2B^2}$. Any geodesic in $\mathbb{H}$ with this sequence as a cutting sequence has an endpoint whose continued fractions expansion ends in $\overline{3}$. This follows from Theorem III.2 and the following diagram.



We can use Example I.3 to compute that the λ-value of a number with tail $\overline{3}$ equals $[\,\overline{3}\,] + [0,\,\overline{3}\,] = \sqrt{13}$. Consequently, the geodesic on T^* with $\overline{A^2B^2}$ as cutting sequence has λ-value $\sqrt{13}$. This nicely agrees with Figure IV.4 depicting four geodesics with cutting sequence $\overline{A^2B^2}$ of which the biggest has diameter $\sqrt{13}$. In a same way we can convert all the cutting sequences obtained from the pairs in the left tree of Figure VI.2, leading to the tree of tails. This tree can be seen to continue by substituting 2211 for 22 to branch to the left and 2211 for 11 to branch to the right. This implies that, walking down through the tree, the tails encountered will contain increasingly long blocks of 1's and 2's that can also be found in tails of Markoff irrationalities. From this it is clear that the λ-values that can be found from the tree of tails will converge to 3. Consequently the λ-values that can be found from proper CSSI geodesics are all less or equal to $\sqrt{13}$ and converge to 3 from above. All of these values are elements of the Markoff spectrum, it is conjectured in [crisp] that they in fact all lie isolated in the spectrum. For some of the values, such as $\sqrt{13}$, this is known to be true.

SECTION VI.2 Closed Geodesics with Two Intersections

The closed doubly self-intersection geodesics, *CDSI geodesics* for short, can be treated in the same way as the CSSI geodesics. Again we can group them into two sets. A CDSI geodesic that encircles the puncture of T^* is called improper and a CDSI geodesic which does not encircle the puncture is again called proper. We saw that improper CSSI have λ-values that in Hall's ray and the same also holds for improper CDSI geodesics. It turns out that proper CDSI geodesics have a cutting sequence with period X^3Y^2, X^2YXY^{-1} or $X^2YX^{-2}Y^{-1}$ such that (X, Y) is a generating pair for the group G. The geodesics with cutting sequences $\overline{X^3Y^2}$ or $\overline{X^2YXY^{-1}}$ have λ-values lying between 3 and 6, with 3 as a limit point. Geodesics with cutting sequence $\overline{X^2YX^{-2}Y^{-1}}$ also lead to λ-values between 3 and 6, yet their unique limit point is 6.

Example VI.1. Consider the geodesic $\overline{\gamma}$ on T^* with cutting sequence $\overline{A^3B^2}$, from the above we know that $\overline{\gamma}$ is a proper CDSI geodesic whose λ-value lies strictly between 3 and 6. There are five different ways to write the period of the cutting sequence and this leads to the five geodesics in $\mathbb{H}$ depicted in Figure VI.3. One can compute that the geodesics with the biggest diameter have diameter $\frac{5}{7}\sqrt{29}$. This is exactly the λ-value of the geodesic with cutting sequence $\overline{A^3B^2}$ and lies indeed between 3 and 6.

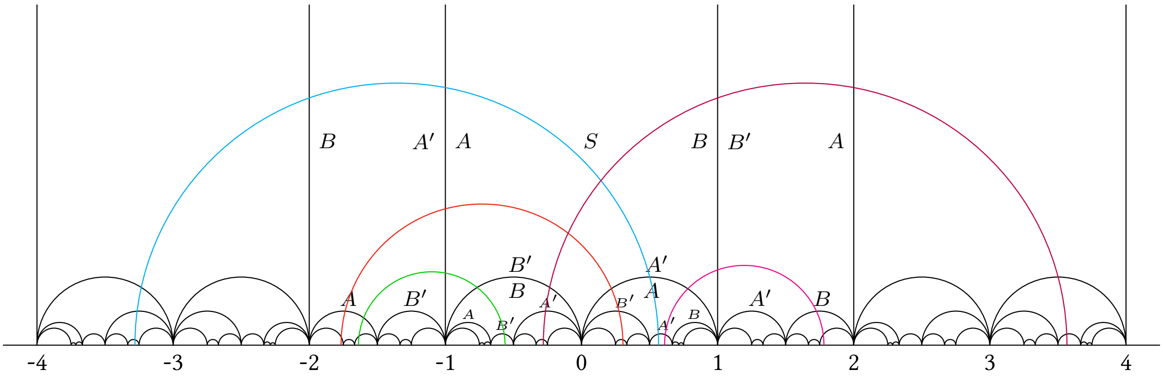


Figure VI.3.: Four geodesics with cutting sequence $\overline{A^3B^2}$.

Further Generalizations

We now have a complete picture of the cutting sequences that closed geodesics with zero, one or two self-intersections have. It allows us to compute the corresponding λ-values, which give information about both the geodesics and elements of the Markoff spectrum. The methods in Crisp et al. (1998) can be generalized to obtain information about geodesics with higher intersection numbers. An important result of Crisp et al. (1998) is that any geodesic in $\mathbb{H}$ whose cutting sequence only consists of A's and B's has a diameter less than 6. Note that these geodesics are necessarily proper, as their cutting sequence cannot contain a commutator. One can wonder whether there is a bound on the diameters of all proper geodesics, even those whose cutting sequences also contain A' and B'. Such results obtained on the geodesics on T^* can be used to gain more insight in the structure of the Markoff spectrum. The other way around, new results on the Markoff spectrum may uncover new properties of geodesics on the punctured torus.

Bibliography

Cohn, Harvey (1955). "Approach to Markoff's Minimal Forms through Modular Functions". In: *Annals of Mathematics* 61.1, pp. 1–12. (Cit. on pp. vii, 57).

— (1971). "Representation of Markoff's Binary Quadratic Forms by Geodesics on a Perforated Torus". In: *Acta Arithmetica* 18, pp. 125–136. (Cit. on p. vii).

Coppel, William A. (2006). *Number Theory, An Introduction to Mathematics: Part A*. Springer. isbn: 0-387-29851-7. (Cit. on p. 1).

Crisp, David et al. (1998). "Closed Curves and Geodesics with Two Self-Intersections on the Punctured Torus". In: *Monatshefte für Mathematik* 125.3, pp. 189–209. (Cit. on pp. 71, 74).

Crisp, David J. and William Moran (1993). "Number Theory with an Emphasis on the Markoff Spectrum". In: CRC Press. Chap. Single Self-Intersection Geodesics and the Markoff Spectrum, pp. 83–93. isbn: 978-0824789022. (Cit. on p. 71).

Cusick, Thomas W. and Mary E. Flahive (1989). *The Markoff and Lagrange Spectra*. American Mathematical Society. isbn: 0-8218-1531-8. (Cit. on pp. 13, 25, 28).

Haas, Andrew (1986). "Diophantine Approximation on Hyperbolic Riemann Surfaces". In: *Acta Mathematica* 156.1, pp. 33–82. (Cit. on p. 64).

Irwin, Michael C. (1989). "Geometry of Continued Fractions". In: *The American Mathematical Monthly* 96.8, pp. 696–703. (Cit. on p. 65).

Markoff, Andrey (1879). "Sur les Formes Quadratiques Binaires Indéfinies". In: *Mathematische Annalen* 15.3, pp. 381–406. (Cit. on pp. vii, 13, 22).

— (1880). "Sur les Formes Quadratiques Binaires Indéfinies". In: *Mathematische Annalen* 17.3, pp. 379–399. (Cit. on pp. vii, 13).

Nielsen, Jakob (1924). "Die Isomorphismengruppe der freien Gruppen". In: *Mathematische Annalen* 91.3-4, pp. 169–209. (Cit. on p. 68).

Perron, Oskar (1913). *Die Lehre von den Kettenbrüchen*. B.G. Teubner. (Cit. on p. 1).

Rockett, Andrew M. and Peter Szüsz (1992). *Continued Fractions*. World Scientific. isbn: 981-02-1047-7. (Cit. on p. 1).

Ross, Adam G. (2007). *Closed Geodesics on the Punctured Torus*. Senior Independent Study. (Cit. on p. 13).

Series, Caroline (1985a). "The Geometry of Markoff Numbers". In: *The Mathematical Intelligencer* 7.3, pp. 20–29. (Cit. on pp. vii, 31, 57, 63).

— (1985b). "The Modular Surface and Continued Fractions". In: *Journal of the London Mathematical Society* 2.31, pp. 69–80. (Cit. on pp. 31, 57, 65).

Index

anti-derivatives, 35

best approximation, 5
binary quadratic form, 13

continued fraction expansion, 1
convergent, 2

Farey series, 38
fundamental domain, 31

geodesic, 36
  -on T^*, 47
  CDSI-, 73
  CSSI-, 71
  improper-, 71
  limit of closed and simple-, 59
  proper-, 71
  simple-, 47
geometry
  hyperbolic-, 36
  non-Euclidian-, 36

Hall's ray, 71
horocycle, 66

ideal triangle, 37

Markoff
  -balanced, 22
  -form, 28
  -irrationality, 25
  -limit, 61
  -number, 26
  -values, 13
  neighbors of - numbers, 26

Poincaré metric, 36
punctured torus, 44

quadratic irrational, 8
quotients
  complete-, 1
  partial-, 1

reduced polynomial, 10

same tail modulo 2, 64
sequence
  almost constant-, 33
  associated-, 21
  characteristic-, 34
  cutting-, 32
  derived-, 22, 33
  reduced-, 42

tesselation
  A, B-, 41
  Farey-, 37
  square-, 31
tessellation, 31
transformations
  Möbius, 36
  Nielsen-, 68