Continued Fractions

Wieb Bosma
Cor Kraaikamp

Sandra Hommersom
Merlijn Keune
Chris Kooloos
Willem van Loon
Roy Loos
Ewelina Omiljan
Geert Popma
David Venhoek
Maaike Zwart

2012–2013

Contents

Preface

These are the notes of a course on Continued Fractions that we organized in Nijmegen in the fall semester of 2012. After a brief introduction by us, the notes contain the contents of the 18 lectures that were given by the nine student participants. Roughly speaking they correspond to topics we proposed for self-study (with literature provided by us).

The reader may notice some strange formatting in these pages; they are due to the rather hastily put together character of this document: many of the chapters were originally formatted in another LaTeX-style, and we did not invest terribly much time in re-formatting. In particular, the text may run into the margin occasionally. Also, we did not (yet) take the time to translate brief sections of the Introduction (taken from elsewhere) that were written in Dutch.

Nevertheless, this document shows very well the range of topics covered in the course. We mainly put the whole thing together for the particiapants to look back on their own, but more importantly, each other's contribution.

Thanks are due to Sandra, Merlijn, Chris, Willem, Roy, Ewelina, Geert, David and Maaike for their efforts in writing these notes, lecturing about them, and making this course into an enjoyable experience.

Wieb Bosma, Cor Kraaikamp
Nijmegen, August 2013

Chapter 1

Introduction

Wieb Bosma, Cor Kraaikamp

1.1 What is a continued fraction?

A *finite continued fraction* is a representation

$$\frac{p}{q} = a_0 + \cfrac{e_1}{a_1 + \cfrac{e_2}{a_2 + \cfrac{e_3}{\ddots \cfrac{e_n}{a_n}}}}$$

for an element p/q from the field of fractions $Q(R)$ of a commutative ring R (with unit element). Here e_i, a_i are elements from the ring R, and it is clear that $p, q \in R$ for which equality holds will always exist: we can just simplify the continued fraction ('from right to left'). Usually, certain restrictions are placed on the e_i and a_i depending on R and the type of continued fraction; we will see examples of this further on. The non-negative integer n will be called the *length* of the continued fraction.

Suppose now that $Q(R)$ is endowed with a metric, and that $\bar{Q}$ is a completion of $Q(R)$ with respect to this metric. Then we say that

$$a_0 + \cfrac{e_1}{a_1 + \cfrac{e_2}{a_2 + \cfrac{e_3}{\ddots}}}$$

is *an infinite continued fraction* if for every $n \geq 0$ the finite part

$$\frac{p_n}{q_n} = a_0 + \cfrac{e_1}{a_1 + \cfrac{e_2}{a_2 + \cfrac{e_3}{\ddots \cfrac{e_n}{a_n}}}}$$

is a finite continued fraction representation for p_n/q_n of $Q(R)$ and it holds that

$$\lim_{n\to\infty} \frac{p_n}{q_n}$$

exists as an element of $\bar{Q}$. If this limit is x, we say that the infinite continued fraction represents x. The finite truncations represent elements p_n/q_n that are called *convergents* of x.

For typographical reasons we will usually denote the above continued fraction by $[a_0, e_1/a_1, e_2/a_2, \ldots]$, or, in the (common) case that all e_i are equal to 1, by $[a_0, a_1, a_2, \ldots]$.

1.2 Finite real continued fractions

The most common type of continued fraction is that of continued fractions for real numbers: this is the case where $R = \mathbb{Z}$, so $Q(R) = \mathbb{Q}$, with the usual Euclidean metric $|\cdot|$, which yields the field of real numbers as completion. Although we do not limit ourselves to this case in the course, it will be used very often as a basic case for reference.

A *regular continued fraction expansion* for $x \in \mathbb{R}$ will be an (in)finite continued fraction of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}$$

Euclid's algorithm is very closely related to continued fractions.

Example 1.2.1. Suppose we determine the greatest common divisor of 33

and 137 using Euclid's method:

$$\begin{aligned} 137 &= 4 \cdot 33 + 5, \\ 33 &= 6 \cdot 5 + 3, \\ 5 &= 1 \cdot 3 + 2, \\ 3 &= 1 \cdot 2 + 1, \\ 2 &= 2 \cdot 1 + 0 \end{aligned}$$

which shows the g.c.d. is 1. Dividing each of the above lines of the form

$$a = q \cdot b + r$$

by b we obtain

$$\begin{aligned} \frac{137}{33} &= 4 + \frac{5}{33}, \\ \frac{33}{5} &= 6 + \frac{3}{5}, \\ \frac{5}{3} &= 1 + \frac{2}{3}, \\ \frac{3}{2} &= 1 + \frac{2}{1}, \\ \frac{2}{1} &= 2 + \frac{0}{1}. \end{aligned}$$

In these, each fraction on the right is the reciprocal of the fraction on the next line, so we get (by substitution) that

$$\frac{137}{33} = 4 + \frac{5}{33} = 4 + \frac{1}{6 + \frac{3}{5}} = 4 + \frac{1}{6 + \frac{1}{1 + \frac{2}{3}}} = 4 + \frac{1}{6 + \frac{1}{1 + \frac{1}{1 + \frac{1}{2}}}}.$$

It will be clear why the expression is called a continued fraction, and why we prefer the notation $[0; 4, 6, 1, 1, 2]$.

Definition 1.2.2. A *finite regular continued fraction* $[a_0; a_1, a_2, \ldots, a_n]$ is a repeated quotient

$$[a_0; a_1, a_2, \ldots, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots \cfrac{1}{a_n}}}},$$

with integers a_i satisfying $a_i \geq 1$ for $i \geq 1$, and $a_n \geq 2$. The integers a_i are called *partial quotients*. The continued fraction $[a_0; a_1, a_2, \ldots, a_n]$ determines a rational number p/q, called the *value* of the continued fraction. The rational numbers

$$\frac{p_0}{q_0} = [a_0;\ ], \frac{p_1}{q_1} = [a_0; a_1], \frac{p_2}{q_2} = [a_0; a_1, a_2], \ldots, \frac{p_n}{q_n} = [a_0; a_1, \ldots a_n],$$

are called the *convergents* of $p/q = [a_0; a_1, a_2, \ldots, a_n]$, and n is its *length*.

Remark 1.2.3. Note that $a_i \geq 1$ for $i \geq 1$, and $a_n \geq 2$ are natural from Euclid's algorithm. The latter restriction prohibits the alternative form of a continued fraction ending in $a_n = 1$, which is can be rewritten via

$$a_{n-1} + \frac{1}{1} = a_{n-1} + 1$$

to a regular form. We will sometimes make use of the existence of both of these expansions: one of odd and one of even length for every rational number in what follows, but only one of them is the regular continued fraction.

Theorem 1.2.4. *Every rational number p/q determines a unique finite regular continued fraction.*

PROOF Given p/q, Euclid's algorithm determines $[a_0; a_1, a_2, \ldots, a_n]$.

Note that for $t = [0; a_1, a_2, \ldots, a_n]$ holds: $0 \leq t < 1$ (with a strict inequality on the right because $a_n > 1$), and thus $a_0 \leq [a_0; a_1, a_2, \ldots, a_n] < a_0 + 1$.

Now suppose that also $p/q = [b_0; b_1, \ldots b_k]$ for another continued fraction. Then $a_0 = \lfloor p/q \rfloor = b_0$ since $\lfloor p/q \rfloor$ is the unique integer satisfying $\lfloor p/q \rfloor \leq p/q < \lfloor p/q \rfloor + 1$.

Consider

$$\frac{1}{[a_1; a_2, \ldots, a_n]} = [0; a_1, a_2, \ldots, a_n] = \frac{p}{q} - \lfloor p/q \rfloor = [0; b_1, \ldots b_k] = \frac{1}{[b_1; b_2, \ldots, b_k]},$$

then $[a_1; a_2, \ldots, a_n] = [b_1; b_2, \ldots, b_k]$ so $a_1 = b_1$ as before; etc.

Without first considering Euclid's algorithm, we now find the regular expansion of p/q in general as follows: determine the integral part a_0 and subtract it from the fraction; take the reciprocal of the result and repeat.

In other words, we iterate

$$x_{k+1} = \frac{1}{x_k - \lfloor x_k \rfloor} \tag{1.1}$$

with $x_0 = p/q$ until $x_k - \lfloor x_k \rfloor$ becomes 0. Put $a_k = \lfloor x_k \rfloor$.

The convergents are found as follows: by definition $p_0/q_0 = [a_0;\ ] = a_0/1$. Then

$$\frac{p_1}{q_1} = a_0 + \frac{1}{a_1} = \frac{a_1 a_0 + 1}{a_1},$$

and

$$\frac{p_2}{q_2} = a_0 + \frac{1}{a_1 + \dfrac{1}{a_2}} = a_0 + \frac{a_2}{a_2 a_1 + 1} = \frac{a_2 a_1 a_0 + a_2 + a_0}{a_2 a_1 + 1}.$$

Of course this follows from the previous by replacing a_1 by $a_1 + 1/a_2$. Similarly

$$\frac{p_3}{q_3} = \frac{(a_2 + \frac{1}{a_3})a_1 a_0 + a_2 + \frac{1}{a_3} + a_0}{(a_2 + \frac{1}{a_3})a_1 + 1} = \frac{a_3(a_2 a_1 a_0 + a_2 + a_0) + a_1 a_0 + 1}{a_3(a_2 a_1 + 1) + a_1}$$

and by induction we obtain the following.

Theorem 1.2.5. *For every convergent p_k/q_k of a rational number p/q:*

$$\frac{p_k}{q_k} = \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}}. \tag{1.2}$$

for $1 \le k \le n$, defining $p_{-1} = 1, q_{-1} = 0$.

Example 1.2.6. The partial quotients and convergents from our first example are summarized in the following table.

n :	-1	0	1	2	3	4	5
a_n :		0	4	6	1	1	2
p_n :	1	0	1	6	7	13	33
q_n :	0	1	4	25	29	54	137

Lemma 1.2.7. *The convergents p_k/q_k of a rational number p/q satisfy $|p_k| \ge |p_{k-1}|$ and $q_k \ge q_{k-1}$ for $k \ge 1$, and even:*

$$|p_k| > |p_{k-1}|, \quad (k \ge 3), \text{ and} \qquad q_k > q_{k-1}, \quad (k \ge 2),$$

while

$$p_{k-1}q_k - p_k q_{k-1} = (-1)^k.$$

PROOF Use equation 1.2:

$$\begin{aligned}\frac{p_{k-1}q_k - p_k q_{k-1}}{q_{k-1}q_k} &= \frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{p_{k-1}}{q_{k-1}} - \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}} = \\ &= \frac{(-1)(p_{k-2}q_{k-1} - p_{k-1}q_{k-2})}{q_{k-1}(a_k q_{k-1} + q_{k-2})},\end{aligned}$$

which by induction equals

$$\frac{(-1)^k(p_{-1}q_0 - p_0 q_{-1})}{q_{k-1}(a_k q_{k-1} + q_{k-2})} = \frac{(-1)^k}{q_{k-1}(a_k q_{k-1} + q_{k-2})}.$$

So $p_{k-1}q_k - p_k q_{k-1} = (-1)^k$ and in particular p_k and q_k will be coprime. Also, by induction $q_k = a_k q_{k-1} + q_{k-2} > q_{k-1}$ for $k \geq 2$, because $a_k \geq 1$, and $q_{k-1} > 0$ for $k \geq 1$. Moreover $|p_k| = a_k|p_{k-1}| + |p_{k-2}| > |p_{k-1}|$, for $k \geq 3$ since $|p_{k-2}| > 0$ for $k \geq 3$.

Theorem 1.2.8. *The convergents p_k/q_k of a rational number p/q satisfy for $k \geq 0$:*

$$\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_{k-1}q_k},$$

and

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \cdots < \frac{p}{q} = \frac{p_n}{q_n} < \cdots < \frac{p_3}{q_3} < \frac{p_1}{q_1} < \frac{p_{-1}}{q_{-1}};$$

also

$$\left|\frac{p}{q} - \frac{p_k}{q_k}\right| < \left|\frac{p}{q} - \frac{p_{k-1}}{q_{k-1}}\right|$$

for $0 \leq k \leq n$.

PROOF The first statement is immediate from the Lemma. That also implies that the sequence of q_i's is strictly increasing, and hence the differences between two consecutive convergents is decreasing; this proves the second part.

To prove the final statement we first note the following, for positive real numbers a, b, and integers $0 \leq k \leq n$: positieve reële getallen a, b en $0 \leq k \leq n$:

$$\begin{aligned}\frac{ap_{k-1} + p_{k-2}}{aq_{k-1} + q_{k-2}} \leq \frac{bp_{k-1} + p_{k-2}}{bq_{k-1} + q_{k-2}} \quad &\Longleftrightarrow \\ 0 \leq (a-b)(p_{k-2}q_{k-1} - p_{k-1}q_{k-2}) = (a-b)(-1)^{k-1} \quad &\Longleftrightarrow \\ k \text{ odd and } b \leq a, \quad \text{or} \quad k \text{ even and } a \leq b.&\end{aligned}$$

For odd $k < n$ it holds that

$$\frac{p_{k-1}}{q_{k-1}} < \frac{p_{k+1}}{q_{k+1}} < \frac{p}{q} < \frac{p_k}{q_k},$$

and applying the above with $b = a_{k+1} \geq 1$ and $a = 1$ we get

$$\frac{p}{q} > \frac{p_{k+1}}{q_{k+1}} = \frac{a_{k+1}p_k + p_{k-1}}{a_{k+1}q_k + q_{k-1}} \geq \frac{p_k + p_{k-1}}{q_k + q_{k-1}};$$

but then

$$\begin{aligned}
\frac{p}{q} - \frac{p_{k-1}}{q_{k-1}} &> \frac{p_{k+1}}{q_{k+1}} - \frac{p_{k-1}}{q_{k-1}} \geq \frac{p_k + p_{k-1}}{q_k + q_{k-1}} - \frac{p_{k-1}}{q_{k-1}} = \frac{-(p_{k-1}q_k - p_k q_{k-1})}{q_{k-1}(q_k + q_{k+1})} = \\
&= \frac{1}{q_{k-1}(q_k + q_{k+1})} > \frac{1}{q_k(q_k + q_{k+1})} = \frac{-(p_{k-1}q_k - p_k q_{k-1})}{q_k(q_k + q_{k+1})} = \\
&= \frac{p_k}{q_k} - \frac{p_k + p_{k-1}}{q_k + q_{k-1}} \geq \frac{p_k}{q_k} - \frac{a_{k+1}p_k + p_{k-1}}{a_{k+1}q_k + q_{k-1}} = \frac{p_k}{q_k} - \frac{p_{k+1}}{q_{k+1}} \\
&> \frac{p_k}{q_k} - \frac{p}{q}.
\end{aligned}$$

The even case is similar.

Application 1.2.9 (Gear ratios)**.** Christiaan Huygens used continued fraction convergents in his construction of a planetarium, a model of the solar system as it was known at the time. Using a single drive shaft and gears with numbers of teeth in carefully chosen ratios, all known planets should revolve with reasonable accuracy around the sun in this model. The ratios would correspond to the rations between the length of the year on each planet and that on earth. To be able to make a physical model with actual gears, the number of teeth could neither be very big nor too small.

Huygens found, for example, for the innermost planet, Mercurius, a ratio of 25335/105190. Its continued fraction is $[0; 4, 6, 1, 1, 2, 1, 1, 1, 1, 7, 1, 2]$ and initially Huygens used the fifth convergent $[0; 4, 6, 1, 1, 2] = \frac{33}{137}$. Later, he realized that, although using the ninth convergent would require too many teeth: $[0; 4, 6, 1, 1, 2, 1, 1, 1, 1] = \frac{204}{847}$, but since $204 = 12 \cdot 17$ en $847 = 7 \cdot 121$ this approximation can be used and gives a better result, when using 4 gears with $12, 17, 7$, and 121 teeth, two of these fixed to the same shaft.

Application 1.2.10 (Solving linear equations)**.** The property that consecutive regular convergents satisfy: $p_{k-1}q_k - p_k q_{k-1} = \pm 1$ can be used to solve

$$ax - by = 1, \quad a, b \in \mathbb{Z}_{\geq 1}$$

in integers x, y; note that $\gcd(a, b) = 1$ should hold for solutions to exist at all.

Expand the fraction b/a as a regular continued fraction and consider the penultimate and the ultimate convergents p_{n-1}/q_{n-1} and $p_n/q_n = b/a$. According to Lemma 1.2.7

$$p_{n-1}q_n - p_nq_{n-1} = p_{n-1}a - q_{n-1}b = (-1)^n,$$

which means that, depending on the parity of n, a solution is given by $(x_0, y_0) = (p_{n-1}, q_{n-1})$ or by $(x_0, y_0) = (-p_{n-1}, -q_{n-1})$.

If we want to consider positive solutions only, we may insist on making the length of the continued fraction even by replacing a_n by $a_n - 1, 1$, if necessary.

From the solution (x_0, y_0) we find the *general solution* by simply taking $(x, y) = (x_0 + zb, y_0 + za)$: clearly these form solutions, while on the other hand a second solution (x_1, y_1) satisfies

$$ax_1 - by_1 = 1 = ax_0 - by_0$$

so

$$a(x_1 - x_0) = b(y_1 - y_0).$$

Since a and b are coprime, a must be a divisor of $y_1 - y_0$ and b a divisor of $x_1 - x_0$; the result follows.

To solve $ax + by = \pm 1$ we apply the following: develop the continued fraction for b/a to find one solution (x_0, y_0) for $ax - by = \pm 1$. Then $(x_0, -y_0)$ will be a solution of $ax + by = \pm 1$ and the general solutin is given by $(x_0 + bz, -y_0 - az)$.

To solve equations of the form $ax \pm by = c$ with $|c| > 1$, one multiplies the solutions for $ax \pm by = 1$ by c.

Example 1.2.11. Find all solutions for the equation

$$34 \cdot x + 49 \cdot y = -13.$$

The regular continued fraction for $49/34$ is $[1; 2, 3, 1, 3]$. Modifying this to get an expansion of odd length, gives $[1; 2, 3, 1, 2, 1]$, and this provides a solution to

$$34 \cdot x + 49 \cdot y = -1$$

by looking at the penultimate convergent from the sequence

$$\frac{1}{1}, \frac{3}{2}, \frac{10}{7}, \frac{13}{9}, \frac{36}{25}, \frac{49}{34},$$

since $36 \cdot 34 - 25 \cdot 49 = -1$. The general solution will be $x = 13 \cdot 36 + 49z$, $y = 13 \cdot (-25) - 34z$.

Application 1.2.12 (Egyptian fractions)**.** Egyptian fractions are fractions with numerator 1. The Egyptians wrote every fraction (with the exception of 2/3) as a sum of these Egyptian fractions with distinct denominators. There are several algorithms to write a fraction as a sum of Egyptian fractions, and two questions arise: how large can the denominators become, and how many terms are needed?

The following method uses continued fractions to find reasonably short expansions with reasonably small denominators. More precisely: it will, for a fraction p/q, produce a sum of no more than p terms with denominators at most $q(q-1)$.

Let $0 < p/q < 1$ be given, with $\gcd(p,q) = 1$. Suppose that $p/q = [0; a_1, a_2, \ldots, a_n]$ as a regular continued fraction. With induction on the length of the continued fraction, we define an expansion in terms of Egyptian fractions as follows: if $n = 1$ then $p/q = 1/a_1$ and we are done. Next suppose that we have dealt with continued fractions of length up to $n-1$, then we coninue as follows: for odd n we have $p_{n-1}/q_{n-1} < p_n/q_n = p/q$, and

$$\frac{p}{q} - \frac{p_{n-1}}{q_{n-1}} = \frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{p_n q_{n-1} - p_{n-1} q_n}{q_{n-1} q_n} = \frac{1}{q_{n-1} q_n},$$

so we are done. For even n, we have $p_{n-2}/q_{n-2} < p/q$ and we use *intermediate approximations* or mediants:

$$\frac{p_{n-2}}{q_{n-2}} < \frac{p_{n-2} + p_{n-1}}{q_{n-2} + q_{n-1}} < \cdots < \frac{p_{n-2} + a_n p_{n-1}}{q_{n-2} + a_n q_{n-1}} = \frac{p_n}{q_n} = p/q.$$

Since

$$\frac{p_{n-2} + j p_{n-1}}{q_{n-2} + j q_{n-1}} - \frac{p_{n-2} + (j-1) p_{n-1}}{q_{n-2} + (j-1) q_{n-1}} = \frac{1}{(q_{n-2} + (j-1) q_{n-1})(q_{n-2} + j q_{n-1})}$$

we can write

$$\frac{p}{q} = \frac{p_{n-2}}{q_{n-2}} + \sum_{j=1}^{a_n} \frac{1}{(q_{n-2} + (j-1) \cdot q_{n-1})(q_{n-2} + j \cdot q_{n-1})},$$

and we are done again by induction. In all we did not use more than $1 + a_2 + \cdots + a_{n'}$ Egyptian fractions, where n' is the largest even intger less than or equal to n.

There is a modification of the algorithm that uses several mediants simultaneously, but it takes some care to avoid equal terms.

Application 1.2.13 (Sums of squares)**.** Let p be a prime number; it is well-known that the multiplicative group $\mathbb{F}_p^*$ of the finite field $\mathbb{F}_p$ is cyclic: there exists an integer g such that the powers of g modulo p produce all $p-1$ different residue classes. Of course $g^{p-1} \equiv 1 \bmod p$. The equation $x^2 - 1 = 0$ will have the two solutions 1 and $-1 \equiv g^{(p-1)/2} \bmod p$ in $\mathbb{F}_p$. The equation $x^2 + 1 = 0$ will then have either no solution in $\mathbb{F}_p$ (if $p \equiv 3 \bmod 4$), a single solution (if $p = 2$) or two different solutions $\pm g^{(p-1)/4} \bmod p$ (if $p \equiv 1 \bmod 4$).

We will use this in an attempt to write p as the sum of two squares; since this is a trivial problem if $p = 2$, we assume that p is odd. We are looing for a, b such that $p = a^2 + b^2$. If $p \equiv 3 \bmod 4$ such a, b will not exist (as otherwise $(ab^{-1})^2 \equiv -1 \bmod p$ contradicts the above).

The following finds a solution $p = a^2 + b^2$ for every prime $p \equiv 1 \bmod 4$. Suppose we have found w with $w^2 \equiv -1 \bmod p$, then we obtain a, b with continued fractions, as follows. First adapt $w \in \mathbb{Z}$ if necessary in such a way that $w^2 \equiv -1 \bmod p$ and $0 < w < p/2$. Now develop p/w as a regular continued fraction; it turns out that

$$\frac{p}{w} = [a_0; a_1, \ldots, a_m, a_m, \ldots, a_1, a_0];$$

and the solution is obtained from the convergents $p_{m-1}/q_{m-1} = [a_0; a_1, \ldots, a_{m-1}]$ and $p_m/q_m = [a_0; a_1, \ldots, a_m]$: namely, $a = p_{m-1}$ and $b = p_m$ will do.

For example, when $p = 9973$, we have $2798^2 \equiv -1 \bmod p$ and the continued fraction of $9973/2798$ is

$$[3; 1, 1, 3, 2, 1, 1, 2, 3, 1, 1, 3].$$

We find the convergents

$$\frac{3}{1}, \frac{4}{1}, \frac{7}{2}, \frac{25}{7}, \frac{57}{16}, \frac{82}{23}, \frac{139}{39}, \frac{360}{101}, \frac{1219}{342}, \frac{1579}{443}, \frac{2798}{785}, \frac{9973}{2798}.$$

The numerators of the convergents almost half-way theis expansion yield $57^2 + 82^2 = 9973$.

The reason this algorithm works can be seen when we look at the ex-

tended algorithm of Euclid. In this example the steps are

$$\begin{aligned}
1 \cdot 9973 &+ 0 \cdot 2798 &&= 9973;\\
0 \cdot 9973 &+ 1 \cdot 2798 &&= 2798;\\
1 \cdot 9973 &+ -3 \cdot 2798 &&= 1579;\\
-1 \cdot 9973 &+ 4 \cdot 2798 &&= 1219;\\
2 \cdot 9973 &+ -7 \cdot 2798 &&= 360;\\
-7 \cdot 9973 &+ 25 \cdot 2798 &&= 139;\\
16 \cdot 9973 &+ -57 \cdot 2798 &&= 82;\\
-23 \cdot 9973 &+ 82 \cdot 2798 &&= 57;\\
39 \cdot 9973 &+ -139 \cdot 2798 &&= 25;\\
-101 \cdot 9973 &+ 360 \cdot 2798 &&= 7;\\
342 \cdot 9973 &+ -1219 \cdot 2798 &&= 4;\\
-443 \cdot 9973 &+ -1579 \cdot 2798 &&= 3;\\
785 \cdot 9973 &+ -2798 \cdot 2798 &&= 1.
\end{aligned}$$

The $n+1$-st row is obtained from the rows n and $n-1$ by division with remainder in the first column. The quotients are the partial quotients. The symmetry between the second and the third columns is caused by the property $2798^2 \equiv -1 \bmod 9973$: take every row modulo $p = 9973$ and multiply by -2798. We simply find the lower half of the table from the upper half. We can stop producing rows as soon as a number less than $\sqrt{p}$ appears in the right hand column, at which point a and b appear as coefficients in the second and third columns!

1.3 Infinite real continued fractions

Broadening our outlook, we now allow arbitrary real numbers x for continued fractions: we iterate 1.1 for $x_0 = x \in \mathbb{R}$ and obtain

$$\begin{aligned}
x_0 &= a_0 + \frac{1}{x_1},\\
x_1 &= a_1 + \frac{1}{x_2},
\end{aligned}$$

etcetera. It is clear this will generate an infinite sequence $[a_0; a_1, a_2, \ldots]$ of partial fractions, unless $_0$ is rational. As before, we get (a possibly infinite) sequence of convergents

$$\frac{p_{-1}}{q_{-1}} = \frac{1}{0}, \frac{p_0}{q_0} = \frac{a_0}{1}, \frac{p_1}{q_1}, \frac{p_2}{q_2}, \cdots .$$

It is easy to see (using induction) that for $n \geq 0$

$$x = \frac{x_{n+1}p_n + p_{n-1}}{x_{n+1}q_n + q_{n-1}}; \tag{1.3}$$

since for $n = 0$ we have

$$x = \frac{x_1 a_0 + 1}{x_1} = a_0 + \frac{1}{x_1} = x_0,$$

and it holds that

$$\begin{aligned} \frac{x_{n+1}p_n + p_{n-1}}{x_{n+1}q_n + q_{n-1}} &= \frac{x_{n+1}(a_n p_{n-1} + p_{n-2}) + p_{n-1}}{x_{n+1}(a_n q_{n-1} + q_{n-2}) + q_{n-1}} = \frac{(a_n p_{n-1} + p_{n-2}) + \frac{p_{n-1}}{x_{n+1}}}{(a_n q_{n-1} + q_{n-2}) + \frac{q_{n-1}}{x_{n+1}}} \\ &= \frac{(a_n + \frac{1}{x_{n+1}})p_{n-1} + p_{n-2}}{(a_n + \frac{1}{x_{n+1}})q_{n-1} + q_{n-2}} = \frac{x_n p_{n-1} + p_{n-2}}{x_n q_{n-1} + q_{n-2}}, \end{aligned}$$

using the recursive relation for numerators and denominators of convergents given in 1.2 (the inductive proof works for the infinte expansions just as well).

As in the proof of 1.2.8 we see that the convergents provide increasingly good approximations, alternately from above and below, for x, and also again that consecutive convergents are at a distance $(q_{k-1}q_k)^{-1}$ from each other. From Theorem 1.2.8 it follows immediately that

$$x = \lim_{n\to\infty} \frac{p_n}{q_n}.$$

Theorem 1.3.1. *The convergents p_k/q_k of any irrational number x satisfy:*

$$\frac{1}{2q_k q_{k+1}} < \frac{1}{q_k(q_k + q_{k+1})} < \left| x - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}} < \frac{1}{q_k^2},$$

for $k \geq 1$.

Proof By 1.3

$$\left| x - \frac{p_k}{q_k} \right| = \left| \frac{x_{k+1}p_k + p_{k-1}}{x_{k+1}q_k + q_{k-1}} - \frac{p_k}{q_k} \right| = \left| \frac{(-1)^k}{q_k(q_k x_{k+1} + q_{k-1})} \right|.$$

Since $a_{k+1} < x_{k+1} < a_{k+1} + 1$ is $q_{k+1} < q_k x_{k+1} + q_{k-1} < q_{k+1} + q_k$:

$$\frac{1}{q_k(q_k + q_{k+1})} < \left| x - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}}.$$

The other inequalities follow from $q_k < q_{k+1}$.

Theorem 1.3.2. *Voor twee opeenvolgende convergenten $p_{k-1}/q_{k-1}, p_k/q_k$ van een irrationaal getal x geldt:*

$$\left|x - \frac{p_{k-1}}{q_{k-1}}\right| < \frac{1}{2q_{k-1}^2} \quad of \quad \left|x - \frac{p_k}{q_k}\right| < \frac{1}{2q_k^2}.$$

PROOF It would follow from

$$\left|\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}}\right| = \left|x - \frac{p_k}{q_k}\right| + \left|x - \frac{p_{k-1}}{q_{k-1}}\right|$$

and the assumption that the statement is false, that

$$\frac{1}{q_{k-1}q_k} = \left|\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}}\right| \geq \frac{1}{2q_k^2} + \frac{1}{2q_{k-1}^2}$$

which is equivalent to

$$(q_k - q_{k-1})^2 \leq 0;$$

this is a contradiction as $q_k > q_{k-1}$ for $k \geq 2$.

Theorem 1.3.3. *If a fraction p/q satisfies $0 < q \leq q_k$ for some convergent p_k/q_k of x then*

$$\frac{p}{q} \neq \frac{p_k}{q_k} \quad \Rightarrow \quad \left|x - \frac{p}{q}\right| > \left|x - \frac{p_k}{q_k}\right|.$$

PROOF Without loss of generality we assume that p and q are coprime. If $q = q_k$ then

$$\left|\frac{p}{q} - \frac{p_k}{q_k}\right| > \frac{1}{q_k}$$

but

$$\left|x - \frac{p_k}{q_k}\right| < \frac{1}{2q_k}$$

so

$$\left|x - \frac{p_k}{q_k}\right| < \left|x - \frac{p}{q}\right|.$$

Suppose that $q_{k-1} < q < q_k$; let integers e, f be defined by

$$e = (qp_{k-1} - pq_{k-1}), \quad f = (pq_k - qp_k),$$

then $f \neq 0$ and

$$\begin{aligned} ep_k + fp_{k-1} = p(p_{k-1}q_k - p_kq_{k-1}) &= \pm p, \\ eq_k + fq_{k-1} = q(p_{k-1}q_k - p_kq_{k-1}) &= \pm q, \end{aligned}$$

so (changing the sign of e, f if necessary) we may assume there are $=$-signs on the right hand side. As $eq_k + fq_{k-1} = q < q_k$, the signs of e and f are opposite, like those of $p_k - q_k x$ and $p_{k-1} - q_{k-1}x$. But then the signs $e(p_k - q_k x)$ and $f(p_{k-1} - q_{k-1}x)$ will be equal again. Moreover,

$$p - qx = e(p_k - q_k x) + f(p_{k-1} - q_{k-1}x)$$

and if $|f| = 1$ and $e \neq 0$ since $q > q_{k-1}$, and therefore

$$|p - qx| > |p_{k-1} - q_{k-1}x|\,.$$

Now Theorem 1.3.1 implies

$$|p_{k-1} - q_{k-1}x| > q_{k-1}\frac{1}{q_{k-1}(q_{k-1} + q_k)} \geq \frac{1}{q_{k+1}} > q_k\left|\frac{p_k}{q_k} - x\right| = |p_k - q_k x|\,.$$

So $|p - qx| > |p_k - q_k x|$ and the statement follows upon division by q on the left and by $q_k > q$ on the irghand the statement follows upon division by q on the left and by $q_k > q$ on the right

Theorem 1.3.4. *If p/q satisfies*

$$\left|x - \frac{p}{q}\right| < \frac{1}{2q^2}$$

then

$$\frac{p}{q} = \frac{p_k}{q_k},$$

for some convergent p_k/q_k of x.

PROOF Expand p/q in a finite continued fraction of odd length n; then $p/q = p_n/q_n$ and

$$\frac{p_n}{q_n} - x = \frac{\delta}{q_n^2}, \quad \delta < \frac{1}{2}.$$

There exists $y > 0$ such that

$$x = \frac{yp_n + p_{n-1}}{yq_n + q_{n-1}},$$

and then

$$\frac{\delta}{q_n^2} = \frac{p_n}{q_n} - x = \frac{p_n q_{n-1} - p_{n-1}q_n}{q_n(yq_n + q_{n-1}} = \frac{(-1)^{n+1}}{q_n(yq_n + q_{n-1}}),$$

so

$$\delta = \frac{q_n}{yq_n + q_{n-1}}$$

implying

$$y = \frac{1}{\delta} - \frac{q_{n-1}}{q_n} > 1.$$

According to Lemma 1.3.5, below, p_{n-1}/q_{n-1} and $p_n/q_n = p/q$ will then be consecutive convergents of x.

Lemma 1.3.5. *if*

$$x = \frac{py + r}{qy + s},$$

with $y \in \mathbb{R}$ *and* $p, q, r, s \in \mathbb{Z}$ *such that*

$$y > 1, \quad q > s > 0, \quad ps - qr = \pm 1,$$

then there exists $n \geq 0$ *with*

$$y = x_{n+1}, \quad \frac{p}{q} = \frac{p_n}{q_n}, \quad \frac{r}{s} = \frac{p_{n-1}}{q_{n-1}},$$

where $x = [a_0; a_1, \ldots]$, $x_i = [a_i; a_{i+1}, \ldots]$ *and* $p_i/q_i = [a_0; a_1, \ldots, a_i]$ *for* $i \geq 0$.

Proof Expand p/q in a continued fraction $p/q = [A_0; A_1, \ldots, A_n] = v_n/w_n$, and let $v_{n-1}/w_{n-1} = [A_0; A_1, \ldots, A_{n-1}]$. The continued fraction has been chosen in such a way that $(-1)^{n+1} = v_n w_{n-1} - v_{n-1} w_n = ps - qr = \pm 1$. Then

$$v_n w_{n-1} - v_{n-1} w_n = v_n s - v_n r,$$

hence $v_n(w_{n-1} - s) = w_n(v_{n-1} - r)$ from which (because v_n, w_n are coprime and $v_n > v_{n-1}$) follows that $s = w_{n-1}$ and $r = v_{n-1}$. But the continued fraction expansion of

$$\frac{v_n y + v_{n-1}}{w_n y + w_{n-1}} = [A_0; A_1, \ldots, A_n, y]$$

(compare 1.3) and so $[A_0; A_1, \ldots, A_n]$ is the initial part of the continued fraction of x: $[A_0; A_1, \ldots, A_n] = [a_0; a_1, \ldots, a_n]$ and $y = [A_{n+1}; A_{n+2}, \ldots] = [a_{n+1}; a_{n+2}, \ldots]$ the tail.

Definition 1.3.6. The regular continued fraction operator $T : [0, 1) \to [0, 1)$ is defined by

$$Tx := \frac{1}{x} - \lfloor \frac{1}{x} \rfloor, \; x \neq 0; \quad T0 := 0.$$

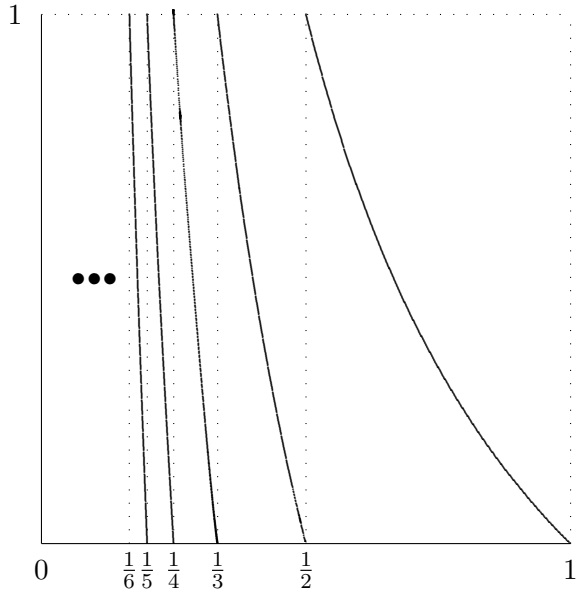


Figure 1.1: The continued fraction map T

The map T is illustrated in Figure 1.1.
Now let $x \in \mathbb{R} \setminus \mathbb{Q}$. Setting

$$a_n = a_n(x) := \lfloor \frac{1}{T_{n-1}} \rfloor, \ n \geq 1,$$

one has

$$\begin{aligned} x &= a_0 + \frac{1}{a_1 + T_1} = a_0 + \frac{1}{a_1 + \dfrac{1}{a_2 + T_2}} = \cdots \\ &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n + T_n}}} \qquad (1.4) \\ &= [\, a_0; \, a_1, \, a_2, \ldots, \, a_{n-1}, \, a_n + T_n \,], \quad n \geq 1. \end{aligned}$$

1.4 Basic properties and matrices

In this section we will derive a number of basic properties of continued fractions using 2×2 matrices. In fact, this matrix representation establishes the

connection between continued fractions and (a part of) algebraic geometry, a connection that has been beautifully explained by C. Series in [82] and [83]. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

i.e., A has integer entries a, b, c and d, and $\det(A) \in \{-1, +1\}$. The letters SL in $\mathrm{SL}_2(\mathbb{Z})$ stand for *special linear*. Now define a map $A : \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ by

$$A(x) := \frac{ax+b}{cx+d}, \ x \in \mathbb{R} \cup \{\infty\}.$$

Such a map is also known as a *Möbius transformation*. Notice that we use the same notation both for the matrix A and for its associated Möbius transformation.

Let $x \in \mathbb{R}$ be an irrational number with continued fraction expansion $x = [\, a_0; a_1, \ldots, a_n, \ldots]$. Define for $n \geq 1$ matrices A_n and M_n by

$$A_0 := \begin{bmatrix} 1 & a_0 \\ 0 & 1 \end{bmatrix}, \quad A_n := \begin{bmatrix} 0 & 1 \\ 1 & a_n \end{bmatrix}, \quad M_n := A_0 A_1 \cdots A_n, \ n \geq 1. \tag{1.5}$$

Writing

$$M_n = \begin{bmatrix} r_n & p_n \\ s_n & q_n \end{bmatrix}, \ n \geq 0,$$

it is easy to show that $(p_n, q_n) = 1$. Using $M_n = M_{n-1} A_n$ one gets

$$\begin{aligned} r_n &= p_{n-1}, \quad s_n = q_{n-1}, \\ p_{n-1}q_n - p_n q_{n-1} &= (-1)^n, \quad n \geq 1, \qquad \text{and} \\ \frac{p_n}{q_n} &= [\, a_0; a_1, \ldots, a_n\,], \ n \geq 1. \end{aligned}$$

and furthermore, that the sequences $(p_n)_{n \geq -1}$ and $(q_n)_{n \geq -1}$ satisfy the following recurrence relations

$$\begin{aligned} &p_{-1} := 1; \quad p_0 := a_0; \quad p_n = a_n p_{n-1} + p_{n-2}, \ n \geq 1, \\ &q_{-1} := 0; \quad q_0 := 1; \quad q_n = a_n q_{n-1} + q_{n-2}, \ n \geq 1. \end{aligned} \tag{1.6}$$

Finally, using (1.6) one sees that $p_n(x) = q_{n-1}(Tx)$ for all $n \geq 0$, where $T_n = T^n x$.

From the recurrence relation for the q_n's it is possible to derive

$$\frac{q_{n-1}}{q_n} = [\, 0; a_n, \ldots, a_1\,].$$

Defining

$$A_n^* := \begin{bmatrix} 0 & 1 \\ 1 & a_n + T_n \end{bmatrix}, \text{ for } n \geq 1,$$

we see that $x = (M_{n-1}A_n^*)(0)$; using

$$M_{n-1} = \begin{bmatrix} p_{n-2} & p_{n-1} \\ q_{n-2} & q_{n-1} \end{bmatrix}, \text{ for } n \geq 1,$$

and the recurrence relations for $(p_n)_{n\geq -1}$ and $(q_n)_{n\geq -1}$ one obtains

$$x = \frac{p_n + T_n p_{n-1}}{q_n + T_n q_{n-1}}, \text{ for } n \geq 1,$$

i.e., $x = M_n(T^n x)$. Finally, use the fact that $p_{n-1}q_n - p_n q_{n-1} = (-1)^n$ so

$$x - \frac{p_n}{q_n} = \frac{(-1)^n T_n}{q_n(q_n + T_n q_{n-1})}, \quad \text{for } n \geq 1. \tag{1.7}$$

Since $T_n \in [0, 1)$ we have that

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n^2}, \quad \text{for } n \geq 1. \tag{1.8}$$

The sequence $(q_n)_{n\geq 0}$ is a monotone increasing sequence of positive integers (which is the *Fibonacci sequence* $(\mathcal{F}_n)_{n\geq 1}$, given by

$$1, 1, 2, 3, 5, 8, 13, 21, \ldots$$

if all the a_i are equal to 1. Now (1.8) yields that $\lim_{n\to\infty} \frac{p_n}{q_n} = x$.

1.5 Periodic regular continued fractions

We richten ons nu op het eenvoudigste soort oneindige kettingbreuken, namelijk de repeterende. Het zal blijken dat die precies corresponderen met kwadratisch irrationale getallen, maar voordat we dat bewijzen, geven we eerst een voorbeeld.

Example 1.5.1 (wortel)**.** We bepalen de kettingbreuk voor $\sqrt{77}$. Het is belangrijk om op te merken dat we hiervoor alleen maar hoeven te weten dat $8^2 < 77 < 9^2$ en dus $8 < \sqrt{77} < 9$.

$$x_0 = x = \sqrt{77}, \text{ dus } a_0 = \lfloor x_0 \rfloor = 8.$$

Dan

$$x_1 = \frac{1}{x_0 - a_0} = \frac{1}{\sqrt{77} - 8} = \frac{\sqrt{77} + 8}{77 - 64}, \text{ dus } a_1 = \lfloor x_1 \rfloor = 1.$$

Vervolgens

$$x_2 = \frac{1}{x_1 - a_1} = \frac{1}{\frac{\sqrt{77}-5}{13}} = \frac{\sqrt{77} + 5}{\frac{77-25}{13}} = \frac{\sqrt{77} + 5}{4}, \text{ dus } a_2 = \lfloor x_2 \rfloor = 3.$$

Daarna

$$x_3 = \frac{1}{x_2 - a_2} = \frac{1}{\frac{\sqrt{77}-7}{4}} = \frac{\sqrt{77} + 7}{\frac{77-49}{4}} = \frac{\sqrt{77} + 7}{7}, \text{ dus } a_3 = \lfloor x_3 \rfloor = 2,$$

en

$$x_4 = \frac{1}{x_3 - a_3} = \frac{1}{\frac{\sqrt{77}-7}{7}} = \frac{\sqrt{77} + 7}{\frac{77-49}{7}} = \frac{\sqrt{77} + 7}{4}, \text{ dus } a_4 = \lfloor x_4 \rfloor = 3,$$

waaruit volgt

$$x_5 = \frac{1}{x_4 - a_4} = \frac{1}{\frac{\sqrt{77}-5}{4}} = \frac{\sqrt{77} + 5}{\frac{77-25}{4}} = \frac{\sqrt{77} + 5}{13}, \text{ dus } a_5 = \lfloor x_5 \rfloor = 1.$$

Tenslotte is

$$x_6 = \frac{1}{x_5 - a_5} = \frac{1}{\frac{\sqrt{77}-8}{13}} = \frac{\sqrt{77} - 8}{\frac{77-64}{13}} = \frac{\sqrt{77} + 8}{1}, \text{ dus } a_6 = \lfloor x_6 \rfloor = 16,$$

zodat

$$x_7 = \frac{1}{x_6 - a_6} = \frac{1}{\sqrt{77} - 8} = x_1,$$

en de kettingbreuk repeteert vanaf hier:

$$x + [8; \overline{1, 3, 2, 3, 1, 16}]$$

waar de overstreping oneindige herhaling van dat blok wijzergetallen aangeeft.

Definition 1.5.2. Een oneindige kettingbreuk $[a_0; a_1, a_2, \ldots]$ heet *periodiek met periodelengte* m als er een $N \geq 0$ bestaat zodanig dat voor alle $n \geq N$ geldt dat $a_n = a_{n+m}$ en er geen kleinere $m \geq 1$ met die eigenschap bestaat. De wijzergetallen $a_0, \ldots, a_{N-1}$ vormen dan de *preperiode*, de (zich steeds herhalende) $a_N, \ldots, a_{N+m-1}$ de *periode*. Een kettingbreuk heet *zuiver*

periodiek als hij periodiek is en $N = 0$ genomen kan worden, dat wil zeggen, er is geen preperiode.

Om alle identiteiten in onderstaande bewijzen ook te laten gelden wanneer $N = 0$, is het handig de (teller en noemer van de) convergent met index -2 te definiëren door $p_{-2} = 0$, en $q_{-2} = 1$. De gebruikelijke recursies (zoals $p_k = a_k p_{k-1} + p_{k-2}$) blijven dan ook geldig voor $k = 0$.

Theorem 1.5.3 (Euler)**.** *Een irrationaal getal x met een periodieke kettingbreuk is een element van $\mathbb{Q}(\sqrt{d})$ voor een $d \in \mathbb{Z}_{\geq 1}$, waar d geen kwadraat is.*

PROOF Veronderstel dat $x = [a_0; a_1, \ldots, a_{N-1}, \overline{a_N, \ldots, a_{N+m-1}}]$, en gebruik nu dat $x_N = x_{N+m}$ met relatie 1.3:

$$x = \frac{x_N p_{N-1} + p_{N-2}}{x_N q_{N-1} + q_{N-2}} = \frac{x_{N+m} p_{N+m-1} + p_{N+m-2}}{x_{N+m} q_{N+m-1} + q_{N+m-2}},$$

dan is

$$-\frac{x q_{N-2} - p_{N-2}}{x q_{N-1} - p_{N-1}} = x_N = x_{N+m-1} = -\frac{x q_{N+m-2} + p_{N+m-2}}{x q_{N+m-1} + p_{N+m-1}},$$

en daarom

$$\begin{aligned} 0 &= (q_{N-2} q_{N+m-1} - q_{N-1} q_{N+m-2}) x^2 + \\ &+ (p_{N-1} q_{N+m-2} - p_{N-2} q_{N+m-1} + p_{N+m-2} q_{N-1} - p_{N+m-1} q_{N-2}) x + \\ &+ p_{N-2} p_{N+m-1} - p_{N-1} p_{N+m-2}. \end{aligned} \tag{1.9}$$

Dit is een kwadratische vergelijking die niet ontaard is; immers de kopcoëfficiënt kan alleen nul zijn wanneer

$$q_{N-2} q_{N+m-1} = q_{N-1} q_{N+m-2},$$

maar omdat q_{N+m-2} en q_{N+m-1} onderling ondeelbaar zijn kan dat alleen indien q_{N-m+2} een deler is van q_{N-2}, hetgeen in tegenspraak is met $q_{N-m+2} > q_{N-2}$.

Definition 1.5.4. Als $x \in \mathbb{Q}(\sqrt{d})$ dan bestaan er $a, b \in \mathbb{Q}$ zodat $x = a + b\sqrt{d}$, en we noemen het element $\bar{x} = a - \sqrt{d}$ de *geconjugeerde* van x. Omdat eenvoudig is in te zien dat de geconjugeerde van de som, resp. het product van twee elementen van $\mathbb{Q}(\sqrt{d})$ de som, resp. het product van de geconjugeerden is, en de geconjugeerde van een element van $\mathbb{Q}$ het element

zelf is, volgt dat $\bar{x}$ aan dezelfde kwadratische vergelijking over $\mathbb{Q}$ voldoet als x:

$$ax^2 + bx + c = 0 \quad \Rightarrow \quad a\bar{x}^2 + b\bar{x} + c = 0.$$

voor $a, b, c \in \mathbb{Q}$.

Wanneer x kwadratisch irrationaal is, zullen we in het vervolg P, Q, d willen kiezen zodat $x = (P + \sqrt{d})/Q$, zodanig dat $P, Q, d \in \mathbb{Z}$, met $d > 0$ geen kwadraat en Q een deler van $P^2 - d$. Dat kan altijd, omdat voor zekere a, b, c geldt dat $ax^2 + bx + c = 0$, en volgens de 'wortelformule' kunnen we dan $P = -b$ nemen, $Q = 2a$ en $d = b^2 - 4ac$.

Een element x van $\mathbb{Q}(\sqrt{d})$ heet *gereduceerd* als $x > 1$ en $-1 < \bar{x} < 0$.

Theorem 1.5.5 (Lagrange)**.** *Als x irrationaal is en $x \in \mathbb{Q}(\sqrt{d})$ met $d \in \mathbb{Z}_{\geq 1}$ dan is de kettingbreuk van x periodiek.*

Proof Schrijf $x = x_0 = (P_0 + \sqrt{d})/Q_0$, met $Q_0 \,|\, P^2 - d$. Met $a_0 = \lfloor x_0 \rfloor$ krijgen we

$$x_1 = \frac{1}{x_0 - a_0} = \frac{1}{\frac{(P_0 - a_0Q_0) + \sqrt{d}}{Q_0}} = \frac{(P_0 - a_0Q_0) - \sqrt{d}}{\frac{(P_0 - a_0Q_0)^2 - d}{Q_0}} = \frac{(a_0Q_0 - P_0) + \sqrt{d}}{\frac{d - P_0^2}{Q_0} + 2a_0P_0 - a_0^2Q_0}$$

hetgeen gelijk is aan

$$\frac{P_1 + \sqrt{d}}{Q_1},$$

als we schrijven

$$P_1 = a_0Q_0 - P_0, \quad Q_1 = \frac{d - P_0^2}{Q_0} + 2a_0P_0 - a_0^2Q_0 = \frac{d - P_1^2}{Q_0};$$

dit is opnieuw van onze standaardvorm, daar duidelijk $Q_1 \,|\, d - P_1^2$. Dus is, met inductie, voor $k \geq 1$, te schrijven $x_k = (P_k + \sqrt{d})/Q_k$, waar

$$P_k = a_{k-1}Q_{k-1} - P_{k-1}, \quad Q_k = \frac{d - P_{k-1}^2}{Q_{k-1}} + 2a_{k-1}P_{k-1} - a_{k-1}^2Q_{k-1} = \frac{d - P_k^2}{Q_{k-1}},$$

met $Q_k \,|\, d - P_k^2$. We leiden een aantal ongelijkheden af, waaruit allereerst volgt dat er maar eindig veel verschillende mogelijkheden zijn voor (P_k, Q_k), maar waarvan we ook later nog gebruik zullen maken. Conjugeren we de gelijkheid

$$x = x_0 = \frac{x_k p_{k-1} + p_{k-2}}{x_k q_{k-1} + q_{k-2}},$$

dan krijgen we voor de geconjugeerde

$$\bar{x}_0 = \frac{\bar{x}_k p_{k-1} + p_{k-2}}{\bar{x}_k q_{k-1} + q_{k-2}},$$

zodat

$$\bar{x}_k = -\frac{\bar{x}_0 q_{k-2} - p_{k-2}}{\bar{x}_0 q_{k-1} - p_{k-1}} = -\frac{q_{k-2}}{q_{k-1}} \left(\frac{\bar{x}_0 - \frac{p_{k-2}}{q_{k-2}}}{\bar{x}_0 - \frac{p_{k-1}}{q_{k-1}}} \right),$$

maar omdat p_k/q_k convergeert naar $x_0 \neq \bar{x}_0$, en $q_{k-2} < q_{k-1}$ is voor k groot genoeg $-1 < \bar{x}_k < 0$ terwijl $x_k > 1$ (dus x_k is gereduceerd, voor k groot genoeg). Maar dan is

$$\frac{2\sqrt{d}}{Q_k} = x_k - \bar{x}_k > 0, \quad \text{dus} \quad Q_k > 0$$

en

$$\frac{2P_k}{Q_k} = x_k + \bar{x}_k > 0, \quad \text{dus} \quad P_k > 0.$$

Bovendien is

$$-1 < \bar{x}_k = \frac{P_k - \sqrt{d}}{Q_k} < 0$$

zodat

$$P_k < \sqrt{d} \quad \text{en} \quad \sqrt{d} - P_k < Q_k,$$

en

$$\frac{P_k + \sqrt{d}}{Q_k} = x_k > 1 \quad \text{dus} \quad Q_k < P_k + \sqrt{d},$$

en daarom

$$0 < P_k < \sqrt{d} \quad \text{en} \quad 0 < Q_k < 2\sqrt{d}. \tag{1.10}$$

Dat voltooit het bewijs. Merk nog wel op dat

$$\sqrt{d} < x_k = \frac{P_k + \sqrt{d}}{Q_k} \quad \Rightarrow \quad P_k > \sqrt{d}(Q_k - 1)$$

zodat (omdat $P_k < \sqrt{d}$) de ongelijkheid $x_k > \sqrt{d}$ precies optreedt wanneer $Q_k = 1$. In het bijzonder is

$$a_k < \sqrt{d} \quad \text{tenzij} \quad Q_k = 1, \text{ en dan } a_k < 2\sqrt{d}. \tag{1.11}$$

We mention two more theorems on periodic continued fractions without giving the proof.

Theorem 1.5.6 (Galois)**.** *A quadratic irrational number x has a purely periodic continued fraction if and only if x is reduced. Moreover, in that case $-\frac{1}{\bar{x}}$ has the reversed period:*

$$x = [\overline{a_0; a_1, \ldots, a_{m-1}}], \quad en \quad -\frac{1}{\bar{x}} = [\overline{a_{m-1}; a_{m-2}, \ldots, a_0}].$$

Theorem 1.5.7. *The regular continued fraction of $\sqrt{d}$, for $d \in \mathbb{Z}_{\geq 1}$ not a square is of the form*

$$[a_0; \overline{a_1, a_2, \ldots, a_2, a_1, 2a_0}\ ].$$

Remark 1.5.8. From what we have seen already about P_k and Q_k it follows immediately that the period length m of the regular continued fraction of $\sqrt{d}$ (and hence of any quadratic irrational in $\mathbb{Q}(\sqrt{d})$) is bounded by $2d$; a sharp upper bound is of the order $\sqrt{d}\log d$.

It is also not hard to prove that

$$Q_k = 1 \quad \Longleftrightarrow \quad m|k.$$

Application 1.5.9 (Pell)**.** De Pell-vergelijking is de vergelijking $x^2 - dy^2 = 1$, met $d \in \mathbb{Z}_{\geq 2}$ geen kwadraat; er worden niet-negatieve gehele oplossingen voor x, y gezocht. We betrekken ook de vergelijking $x^2 - dy^2 = -1$ direct mee in de beschouwing. Omdat

$$x^2 - dy^2 = (x - y\sqrt{d})(x + y\sqrt{d}) = (x - \sqrt{d})^2 + 2y\sqrt{d},$$

zien we dat voor oplossingen (x, y) van de vergelijkingen geldt

$$0 < \left|\frac{x}{y} - \sqrt{d}\right| < \frac{1}{2y^2\sqrt{d}} < \frac{1}{2y^2}.$$

Volgens Stelling 1.3.4 geldt dan dat x/y een convergent van $\sqrt{d}$ moet zijn. Dus alle oplossingen van de vergelijkingen $x^2 - dy^2 = \pm 1$ zijn te vinden onder de convergenten van $\sqrt{d}$.

Om alle oplossingen te bepalen gebruiken we weer de relatie

$$x = \frac{x_{n+1}p_n + p_{n-1}}{x_{n+1}q_n + q_{n-1}},$$

met $x = \sqrt{d}$ en $x_{n+1} = (P_{n+1} + \sqrt{d})/Q_{n+1}$. Hieruit volgt

$$(q_{n-1}Q_{n+1} + q_nP_{n+1} - p_n)\sqrt{d} = p_nP_{n+1} + p_{n-1}Q_{n+1} - q_nd,$$

en dat kan alleen als links en rechts nul staat, dus

$$p_n = q_{n-1}Q_{n+1} + q_nP_{n+1}, \quad \text{en} \quad q_nd = p_{n-1}Q_{n+1} + p_nP_{n+1}.$$

Maar dan is

$$\begin{aligned} p_n^2 - q_n^2 d &= p_n(q_{n-1}Q_{n+1} + q_nP_{n+1}) - q_n(p_{n-1}Q_{n+1} + p_nP_{n+1}) = \\ &= (p_nq_{n-1} - p_{n-1}q_n)Q_{n+1} = (-1)^{n+1}Q_{n+1}. \end{aligned}$$

Volgens de voorgaande opmerking kan dat laatste alleen ± 1 zijn indien $n+1$ een veelvoud is van de periodelengte m van de kettingbreuk voor $\sqrt{d}$. Is die periodelengte m *even*, dan zijn voor $k = 0, 1, 2, 3, \ldots$ de tellers en noemers (p_{km-1}, q_{km-1}) van de convergenten van $\sqrt{d}$ dus precies alle oplossingen van de vergelijking $x^2 - dy^2 = 1$ en zijn er geen oplossingen voor de vergelijking met -1; is de periodelengte oneven, dan zijn beide vergelijkingen oplosbaar en vormen (p_{km-1}, q_{km-1}) voor $k = 1, 3, 5, \ldots$ alle oplossingen voor $x^2 - dy^2 = -1$ en (p_{km-1}, q_{km-1}) voor $k = 0, 2, 4, \ldots$ alle oplossingen voor $x^2 - dy^2 = 1$.

Application 1.5.10 (factorisatie)**.** Een aantal van de beste methoden om een gegeven getal N in factoren te ontbinden is gebaseerd op het idee dat wanneer je twee gehele getallen x en y hebt met $0 < x < y < N$ zodat modulo N geldt $x^2 \equiv y^2$, dan zal $\gcd(N, x - y)$ een factor voor N opleveren omdat dan N een deler is van $x^2 - y^2 = (x - y)(x + y)$. Die factor kan triviaal zijn, wanneer $x \equiv -y \bmod N$, (een geval dat wel op moet treden wanneer N priem is), maar als N minstens twee verschillende priemdelers heeft kan het zijn dat sommige priemfactoren in $x + y$ zitten en andere in $x - y$ zodat we een niet-triviale factor detecteren.

Het grote probleem is natuurlijk om x en y te *vinden*. Fermat probeerde $x^2 - y^2 = N$ op te lossen door systematisch te zoeken, beginnend bij $x = \lfloor\sqrt{N}\rfloor$, en x telkens met 1 ophogend, naar een kwadraat van de vorm $x^2 - N$. Dit werkt aardig wanneer N het product is van twee priemgetallen die heel dicht bij elkaar liggen, maar hopeloos als de twee ver uiteen lopen, bijvoorbeeld $p \approx \sqrt[3]{N}$.

Een succesvolle (en tot in de jaren 1980 veel gebruikte) aanpassing maakt gebruik van kettingbreuken, als volgt. We gebruiken de in het vorige voorbeeld gevonden identiteit $p_n^2 - Nq_n^2 = (-1)^{n+1}Q_{n+1}$, voor de convergenten p_n/q_n van $\sqrt{N}$, met $x_n = (P_n + \sqrt{N})/Q_n$, voor $n \geq 0$, en de ongelijkheid $Q_{n+1} < 2\sqrt{N}$. Het nut is gelegen in de congruentie $p_n^2 \equiv (-1)^{n+1}Q_{n+1} \bmod N$. Om ook rechts een kwadraat te krijgen vereist wat veel geluk, maar we kunnen wel congruenties proberen te *combineren* tot

een kwadraat, en daarvoor willen we Q_{n+1} klein hebben om deze te kunnen ontbinden in factoren. Het idee is als volgt: leg een lijst van kleine priemgetallen aan (te beginnen met $-1, 2, 3, 5, \cdots$, zie echter onder), voer een stap in de kettingbreukontwikkeling van N uit, en zie (via deling met rest door de priemen) of Q_{n+1} te ontbinden is met behulp van uitsluitend de priemen uit de lijst. Bepaal dan de exponenten k_i in

$$Q_{n+1} = (-1)^{k_0} p_1^{k_1} \cdots p_r^{k_r},$$

en herhaal dit proces. Het doel is om zo een matrix met als rijen de gevonden exponenten modulo 2 op te bouwen en in deze matrix een afhankelijheid tussen de rijen te vinden: zo'n afhankelijkheid modulo 2 betekent namelijk dat er een product van overeenkomstige Q's bestaat waarvan de exponenten in de factorisatie bij alle p_i en bij -1 even zijn. Met andere woorden, dit product is een kwadraat! Omtrent de *factor basis* (de lijst van priemgetallen tot een te kiezen grens B) is het nuttig op te merken dat natuurlijk eerst gekeken wordt of N door één van die kleine p_i deelbaar is, maar ook dat slechts ongeveer de helft van de priemgetallen van nut zijn. Immers, een priemgetal p dat een Q_{n+1} deelt, deelt $p_n^2 - Nq_n^2$, dus $N \equiv (p_n/q_n)^2 \bmod p$. Met andere woorden, N moet een kwadraatrest modulo p zijn, een eigenschap die eenvoudig te verifiëren is. Het kan zijn dat de periodelengte van de kettingbreuk van $\sqrt{N}$ klein is, en in dat geval treden maar weinig verschillende waarden Q_{n+1} op. Een manier om dat te verhelpen is door naar de kettingbreuk van $\sqrt{kN}$ voor kleine veelvouden van N te kijken. Niet alleen kan de periode zo groter worden, maar k kan ook nog eens zo geselecteerd worden dat kN kwadraatrest is voor veel van de priemgetallen tot B.

Chapter 2

Planetaria

Maaike Zwart

An application of continued fractions in engineering

Astronomy is possibly the oldest science around. Ever since the dawn of mankind people have wondered about the sun, moon and stars, about what they are, what they are made of. Various tales and myths are built upon this curiosity. It has encouraged many a man to make models of it reflecting their view. Some are built on myths, like the turtle carrying the earth, others purely on observations, and some on both, like Kepler's model of the Platonic Universe.

Figure 2.1: The world carried by elephants and a turtle (Hindu Myth) and the Mysterium Cosmographicum by Kepler.

With the invention of the telescope in the 1600's, the accuracy of the scien-

tific models rapidly increased. More planets were discovered, some of them carrying their own moons with them. In 1682, Huygens built a table top size planetarium for the Académie Royale des Sciences in Paris. His design uses state of the art engineering as well as some interesting mathematics. A century later, Eise Eisinga turned his house into a museum by converting the ceiling of his living room into a planetarium. Both designs are breathtaking. I will give a brief description of them, after which I will discuss the main mathematical issue in designing such a model of our solar system; approximating the periods of the planets around the sun.

2.1 Huygens's Planetarium

Christiaan Huygens, born in 1629, was one of the most brilliant scientists of his day. He contributed greatly to physics, mathematics and astronomy. In 1682 he designed a planetarium intended for the Académie Royale des Sciences in Paris, of which he had been head until the previous year. The planetarium was crafted by Johannes van Ceulen, a craftsman from The Hague. It contains the planets Mercury, Venus, Earth, Mars, Jupiter and Saturn. Apart from showing the motions of the planets, the design indicates the date and time, which constellations are in the sky, and many things more, far too many to describe here. There are various descriptions of this planetarium, some are translations of Huygens's own description, such as [33]. To fully appreciate the beauty of this planetarium, I think its best to take one of those descriptions to the Boerhave museum in Leiden, where the planetarium is on display, and watch the real thing while reading the description. As I obviously cannot do this for you in these notes, I'll settle with an ancient wisdom: "a picture says more than a thousand words", see figures 2.2 and 2.3.

2.2 Eisinga's Planetarium

A century later, Eise Eisinga invested seven years of his spare time to construct a magnificent planetarium in the ceiling of his home. While Huygens was a professional scientist (as far as such a profession existed in that time), Eisinga was just an interested layman. His profession was woolcomber, like his father. To relax his mind, he spend his evening hours studying physics and mathematics.

On the 8^{th} of May, 1774, the Moon and the planets Jupiter, Mars, Mercury and Venus all aligned in the constellation Aries. This extraordinary

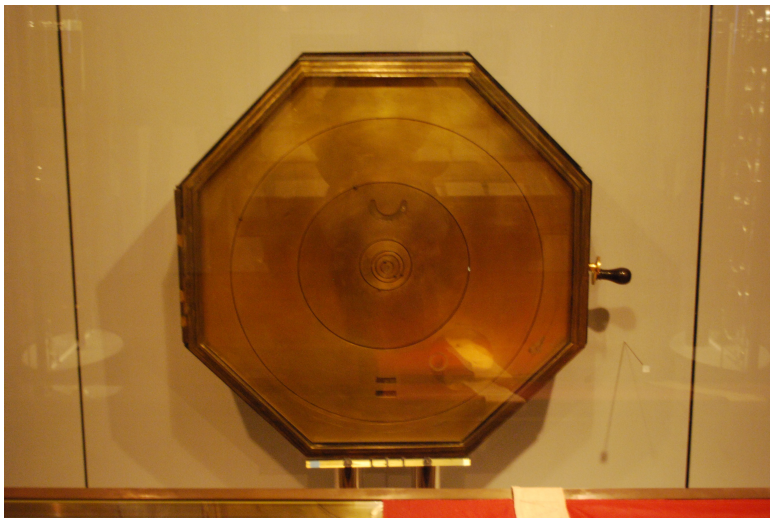


Figure 2.2: Huygens's planetarium, front view.

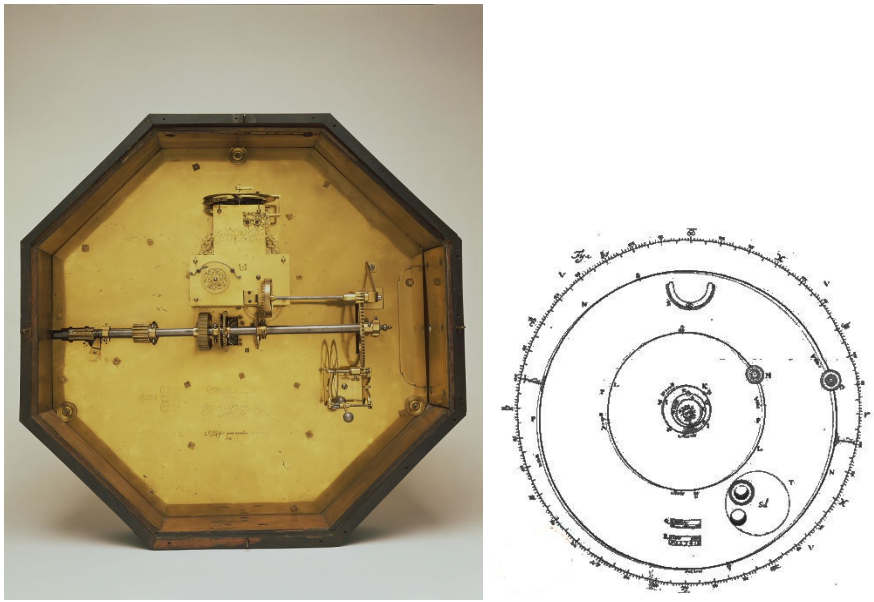


Figure 2.3: Huygens's planetarium, inside view and schematic drawing.

phenomenon gave rise to end-of-the world myths much like the expiring Maya calendar on 21-12-12. Eise Eisinga regretted the ignorance of his fellow villagers. It was then that he decided to built his own planetarium. His planetarium should accurately show the positions of the heavenly bodies at any given time. He did a magnificent job. Although his planetarium is a little less accurate than the one Huygens made [87], this one is truly breathtaking. The planets move along the ceiling, following circular trails. In the attic above this ceiling, Eisinga built the whole machinery. Everything is handmade out of wood and steel. Apart from the planets (again, Mercury, Venus, Earth, Mars, Jupiter and Saturn) there are indescribably many displays and other clockworks giving information about constellations, date, lunar phases, time of sunrise and sunset, etc...

Again, the only way to fully appreciate this work is to visit it, which I haven't been able to yet. But these pictures should give you a fair impression.

2.3 Mathematical Issues

Apart from technical difficulties, Huygens and Eisinga faced a mathematical challange. The planetaria are driven by a single power source, for example a spring (remember, there was no electricity back then), driving one wheel. Interlocking gears are then used to give each planet its own period around the sun. The idea is simple: if the main wheel makes a full revolution in x minutes, and for example Mercury should revolve around the sun in y minutes, then we should mount a gearwheel with x teeth on the axis of the main wheel, and a gearwheel with y teeth on the axis connected to Mercury. However, x and y are generally not whole numbers. This issue could be easily solved by taking the nearest integers of x and y, but still technical issues get in the way. A gearwheel with a single tooth is not functional, one with a million teeth is not manufacturable. Say that a gearwheel must have a number of teeth between 7 and 200[1]. The question is to find two of these numbers $7 \leq n, m \leq 200$, such that $\frac{n}{m}$ approximates $\frac{x}{y}$ as closely as possible. Then the two gears, one with n teeth and one with m teeth, will approximate Mercury's period as good as possible. As Huygens formulates it:

> De geheele kwestie komt dus hierop neer: wanneer twee groote

[1] [34] refers to a book from 1947 by Merrit, wherein 127 is given as upper limit. I thought 200 would be more reasonable

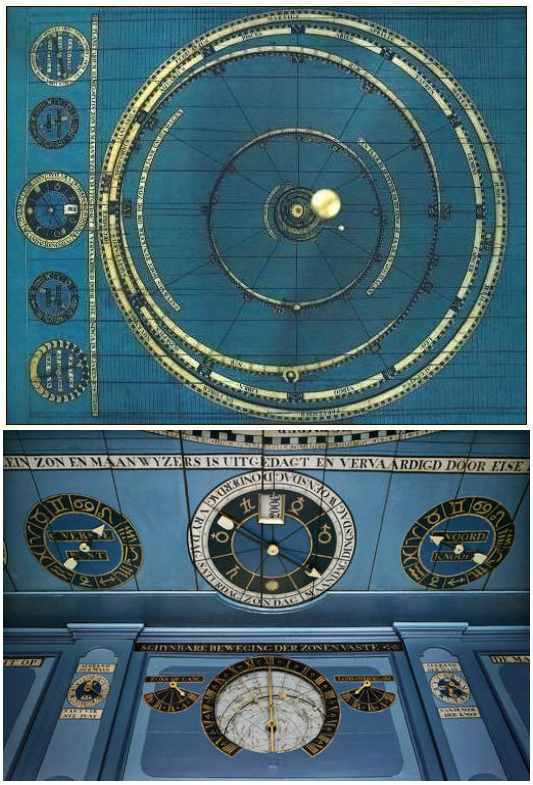


Figure 2.4: Eisinga's planetarium, planet discs and constellations clocks.

> getallen gegeven zijn die in een bepaalde verhouding tot elkander staan, andere kleinere te vinden voor de radertanden die niet ongeschikt zijn door hun grootte en die dezelfde verhouding met een zoo groote nauwkeurigheid opleveren, dat geen andere kleinere getallen een betere benadering geven. [33]

That is:[2]:
Given a $r \in \mathbb{Q}$**, find natural numbers** n **and** m **such that** $7 \leq n, m \leq 200$ **and** $\frac{n}{m} \approx r$**, and such that there are no other** $7 < n', m' < 200$ **with** $|r - \frac{n'}{m'}| < |r - \frac{n}{m}|$.

[2]We may assume x and y are both rationals, so that the resulting fraction (r) is again a rational number, because both x and y are measured quantities, and measurements always yield a rational quantities.

I will discuss two methods for finding the required n and m. Then, I will discuss a method using more than two gears to make an even better approximation.

2.3.1 Take some Educated Guesses

This may seem too simple a solution, far from efficient and not usable at all. But a little common sense at a bit of trial and error can get you a long way. Nowadays, the use of a computer has made this method the most favourable of all. During my lecture, David actually wrote a program that tried every possible n and m and returned the best approximation. It returned the same value Huygens found using the second method, involving continued fractions.

2.3.2 Continued Fractions

As we know from Sandra's notes about decimals and continued fractions, the latter are very suitable for approximating numbers. In the construction of continued fractions, the partial quotients $\frac{p_n}{q_n}$ converge to r, while p_n and q_n are both increasing with each step. So all we have to do is make the continued fraction expansion of the number r we want to approximate, keeping track of the partial quotients, and stop as soon as either $p_n > 200$ or $q_n > 200$.

Huygens invented this method for the design of his planetarium. The axis driving the Earth was his main axis. In getting the period of Mercury right, he took the following steps: The ratio between the periods of Mercury and Earth he found to be $21038 : 5067$ [33]. The continued fraction of $\frac{5067}{21038}$ is $[0; 4, 6, 1, 1, 2, 1, 1, 1, 1, 7, 1, 2]$. The table with the subsequent convergents $\frac{p_n}{q_n}$ is given below. This table is the result of using the recursive relations for p_n and q_n:

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

Table 2.1: The convergents of the continued fraction expansion of $\frac{5067}{21038}$

n	-1	0	1	2	3	4	5	6	7	8	9	10	11	12
a_n		0	4	6	1	1	2	1	1	1	1	7	1	2
p_n	1	0	1	6	7	13	33	46	79	125	204	1553	1757	5067
q_n	0	1	4	25	29	54	137	191	328	519	847	6448	7295	21038

As we see from the table, the most useful convergent would be $\frac{46}{191}$. Would be, because Huygens used an additional clever trick; the concept of gear trains. But before I discuss those, I want to turn your attention to the consequences of measuring errors. The periods of the various planets around the sun are all measured values. Measurements always come with errors. The ratio of 21038 : 5067 comes from the measured ratio of values of 525950 minutes for the Earth's period and 126675 minutes for Mercury's. If instead, Mercury's period was measured two minutes slower, 126677 minutes, then the continued fraction of $\frac{126677}{525950}$ is: $[0; 46, 1, 1, 2, 2, 266, 1, 5]$:

Table 2.2: The convergents of the continued fraction expansion of $\frac{126677}{525950}$

n	-1	0	1	2	3	4	5	6	7	8	9
a_n		0	4	6	1	1	2	2	266	1	5
p_n	1	0	1	6	7	13	33	79	21047	21126	126677
q_n	0	1	4	25	29	54	137	328	87385	87713	525950

This small measuring error leads to the choice of $\frac{33}{137}$ instead of $\frac{46}{191}$. This illustrates that due to measuring errors, the error of the model is always larger than the error of the approximations using continued fractions suggests.

2.3.3 Gear Trains

The approximation Huygens picked for the ratio of the periods of Earth and Mercury is 204 : 847 [33]. These numbers are far too large to craft as teeth of a gearwheel. However, Huygens realised that $204 = 12 \cdot 17$, and $847 = 121 \cdot 7$. Using 4 gears instead of 2, he constructed a 'gear train', see figure 2.5 below.

Suppose gear A has 121 teeth, B has 12 teeth, C has 7 teeth and D has 17 teeth. Then, as A makes one full revolution, B makes $\frac{121}{12}$ revolutions. As C is mounted on the same axis as B, it too makes $\frac{121}{12}$ revolutions. The last one then, D, makes $\frac{7}{17}$ as many revolutions as C. So as A makes one full revolution, D makes $\frac{121}{12} \cdot \frac{7}{12} = \frac{121 \cdot 7}{12 \cdot 17} = \frac{847}{204}$ revolutions. This is exactly the ratio Huygens wanted to use for the periods of Mercury and Earth.

So gear trains can get you a more precise approximation at the cost of using more material. However, this trick only works if this better approximation consists of numbers that factorise conveniently.

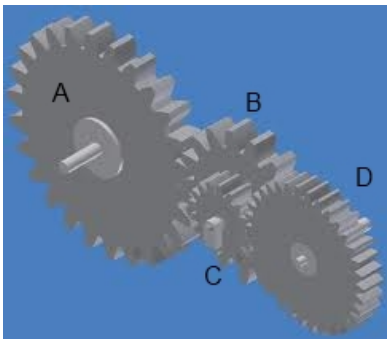


Figure 2.5: A four stage gear train

2.3.4 Some last Comments

Huygens clearly indicated how he used continued fractions to approximate the periods of the planets around the sun. Altough Eisinga built his planetarium a century later, it appears to be less accurate [87], which is odd. I do not yet know which method Eisinga used to get his approximations. He might not have known about Huygens' work. The planetarium itself is much more impressive than Huygens', as it has many more details.

In both planetaria, as much as possible is made to scale. However, the sizes of the planets are greatly exaggerated to make them visible. If you would like to experience the true emptiness of our solar system, I would recommend to pay a visit to het Melkwegpad, near Dwingeloo/Westerbork. Here, all planets are to scale. Beware, it's quite a walk reaching Pluto.

Chapter 3

The Stern-Brocot algorithm

Maaike Zwart

In the 1860's, Moritz Stern and Achille Brocot independently developed an interesting algorithm. Number theorist Stern used it to construct all rational numbers, Clockmaker Brocot as a method to approximate real numbers. The algorithm is closely related to Euclids algorithm, and therefore also to continued fractions. In this chapter I will give Stern's definition and discuss Brocot's application in his clockwork, linking this chapter to my previous topic, Planetaria, and showing the similarity with Euclid's algorithm. I will finish with an unexpected link between Stern-Brocot and continued fractions.[1]

3.1 Constructing the Rationals

The rationals are produced in a 'hanging tree'. Starting with two boundaries, every rational number in between them appears in a tree that is constructed by taking *mediants*.

Definition 3.1.1. Mediants
Let $\frac{p}{q}$ and $\frac{p'}{q'}$ be two rational numbers (expressed as fractions in their lowest terms). Then their mediant is given by:

$$\frac{p}{q} \oplus \frac{p'}{q'} = \frac{p+p'}{q+q'}$$

The Stern-Brocot tree is then the result of the following iterative process:

[1] The definition of both Stern and Brocot I learned from [34], a nice article I definitely recommend reading for leisure time.

Definition 3.1.2. Stern-Brocot tree

- **Stage 0:** Choose rational boundaries and write them as fractions in their lowest terms.
- **Stage 1:** Add the mediant of the boundaries.
- **Stage n+1:** Add the mediants of all consecutive fractions in the tree, including the boundaries.

Choosing the somewhat peculiar expressions $\frac{0}{1}$ and $\frac{1}{0}$ as boundaries, the latter expressing ∞, yield all the positive rational numbers:

$$\frac{0}{1} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \frac{1}{0}$$

$$\frac{0}{1} \qquad\qquad\qquad\qquad\qquad\qquad\qquad \frac{1}{1} \qquad\qquad\qquad\qquad\qquad\qquad\qquad \frac{1}{0}$$

$$\frac{0}{1} \qquad\qquad\qquad \frac{1}{2} \qquad\qquad\qquad \frac{1}{1} \qquad\qquad\qquad \frac{2}{1} \qquad\qquad\qquad \frac{1}{0}$$

$$\frac{0}{1} \qquad \frac{1}{3} \qquad \frac{1}{2} \qquad \frac{2}{3} \qquad \frac{1}{1} \qquad \frac{3}{2} \qquad \frac{2}{1} \qquad \frac{3}{1} \qquad \frac{1}{0}$$

$$\frac{0}{1} \quad \frac{1}{4} \quad \frac{1}{3} \quad \frac{2}{5} \quad \frac{1}{2} \quad \frac{3}{5} \quad \frac{2}{3} \quad \frac{3}{4} \quad \frac{1}{1} \quad \frac{4}{3} \quad \frac{3}{2} \quad \frac{5}{3} \quad \frac{2}{1} \quad \frac{5}{2} \quad \frac{3}{1} \quad \frac{4}{1} \quad \frac{1}{0}$$

$$\vdots$$

The 'hanging tree' structure becomes more apparent when only the newborns of each stage (i.e. the fractions added to the tree in that stage) are depicted:

$$\frac{0}{1} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \frac{1}{0}$$

$$\frac{1}{1}$$

$$\frac{1}{2} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \frac{2}{1}$$

$$\frac{1}{3} \qquad\qquad\qquad \frac{2}{3} \qquad\qquad\qquad \frac{3}{2} \qquad\qquad\qquad \frac{3}{1}$$

$$\frac{1}{4} \qquad \frac{2}{5} \qquad \frac{3}{5} \qquad \frac{3}{4} \qquad \frac{4}{3} \qquad \frac{5}{3} \qquad \frac{5}{2} \qquad \frac{4}{1}$$

$$\vdots$$

I claimed every positive rational number appears in this Stern-Brocot tree (with boundaries $\frac{0}{1}$ and $\frac{1}{0}$, from now on referred to simply as the Stern-Brocot tree). The proof of this fact uses some uses some properties of the

Stern-Brocot tree. As these are interesting properties on their own, I will first discuss these, and then justify my claim.

The first property is one about consecutive fractions, meaning that in a certain stage in the tree, there is no fraction in between them, e.g. $\frac{0}{1}$ and $\frac{1}{1}$ are consecutive in stage 1.

Proposition 3.1.3. *Any any stage of the Stern-Brocot tree, two consecutive fractions $\frac{p}{q} < \frac{p'}{q'}$ have the property that:*

$$qp' - pq' = 1$$

Proof. By induction to the stages of the Stern-Brocot tree.
Base case: At the first stage we only have the fractions $\frac{0}{1}$ and $\frac{1}{0}$. As $1 \cdot 1 - 0 \cdot 0 = 1$, this case is ok.
Induction step: Suppose at stage n of the Stern Brocot tree, all consecutive fractions have this property. At stage $n+1$, all mediants of the consecutive fractions are added to the tree. Let $\frac{p}{q}$ and $\frac{p'}{q'}$ be two consecutive fractions of stage $n+1$ of the Stern-Brocot tree. Then one these two fractions already existed in stage n (suppose $\frac{p}{q}$), and the other is a newborn. The newborn must then be the mediant of $\frac{p}{q}$ and $\frac{p''}{q''}$, that is: $\frac{p'}{q'} = \frac{p+p''}{q+q''}$. Note that $\frac{p}{q}$ and $\frac{p''}{q''}$ are consecutive fractions in stage n. Suppose wlog that $\frac{p}{q} < \frac{p'}{q'} < \frac{p''}{q''}$. Then:

$$\begin{aligned} qp' - pq' &= q(p + p'') - p(q + q'') \\ &= qp + qp'' - pq - pq'' \\ &= qp'' - pq'' \\ &= 1 \end{aligned}$$

Where in the last step, the induction hypothesis is applied. □

The next proposition concerns the newborns of stages.

Proposition 3.1.4. *At stage n, the sum of the numerator and denominator of the newborns is at least $n+1$.*

Proof. By induction.
Base case: At the first stage (n=0), we have the fractions $\frac{0}{1}$ and $\frac{1}{0}$, for both fractions, the sum of their numerator and denominator equals 1, which is $n+1$.
Induction step: Suppose the sum of the numerator and denominator of the newborns of stage n is at least $n+1$. Let $\frac{a}{b}$ be an arbitrary newborn of

stage $n+1$. Then it is the mediant of a newborn of stage n ($\frac{p}{q},\ p+q \geq n+1$) and an older fraction ($\frac{p'}{q'},\ p'+q' \geq 1$): $\frac{a}{b} = \frac{p+p'}{q+q'}$. The sum of the numerator and denominator of $\frac{a}{b}$ is hence at least $(n+1)+1$. □

These two propositions enable us to prove the claim that every positive rational number appears in the Stern-Brocot tree. Even better, it shows up only once. The proof is a slightly edited version of a proof by A. Bogomolny and W.McWorter [10].

Proposition 3.1.5. *Every positive rational number appears exactly once in the Stern-Brocot tree with starting fractions* $\frac{0}{1}$ *and* $\frac{1}{0}$.

Proof. Each mediant lies strictly in between its two parents. As all rationals in the tree are constructed by taking mediants, it is impossible for a rational to appear more than once in this tree. To prove each rational appears at least once takes some more work. Let $\frac{p}{q}$ be any fraction expressed in its lowest terms and suppose it does not appear in the tree. Then, at any stage of the tree, there are two consecutive fractions $\frac{n}{m}$ and $\frac{n'}{m'}$ such that $\frac{n}{m} < \frac{p}{q} < \frac{n'}{m'}$. In particular, these fractions can be found in stage $p+q$ of the tree. We then have:

$$\frac{n}{m} < \frac{p}{q} < \frac{n'}{m'}$$

Dividing by $\frac{p}{q}$ yields:

$$\begin{aligned} &\frac{nq}{mp} < 1 < \frac{n'q}{m'p} \\ &mp - nq > 0 \qquad\qquad \Rightarrow \quad mp - nq \geq 1 \\ &n'q - m'p > 0 \qquad\qquad \Rightarrow \quad n'q - m'p \geq 1 \end{aligned}$$

Which can be rewritten as:

$$(n'+m')(mp-nq) \geq n'+m' \quad \text{and} \quad (n+m)(n'q-m'p) \geq n+m$$
$$n'mp + m'mp - n'nq - m'nq \geq n'+m' \quad \text{and} \quad nn'q + mn'q - nm'p - mm'p \geq n+m$$

Adding these two expressions and reducing the result to a readable form:

$$p(n'm - nm') + q(mn' - m'n) \geq n' + m' + n + m$$

As $\frac{n}{m}$ and $\frac{n'}{m'}$ are consecutive fractions, we may use $mn' - nm' = 1$ (proposition 3.1.3), yielding:

$$p + q \geq n' + m' + n + m$$

As $\frac{n}{m}$ and $\frac{n'}{m'}$ are consecutive fractions in stage $p+q$ of the Stern-Brocot tree, one of them is a newborn. By proposition 3.1.4, the sum of its numerator and denominator is at least $p+q+1$, leading to the result that:

$$p+q > p+q+1$$

Which is clearly a contradiction. Hence the assumption "$\frac{p}{q}$ does not appear in the first in the first $p+q$ stages of the Stern-Brocot tree" is false. Hence $\frac{p}{q}$ does appear in the Stern-Brocot tree. More specifically, it does so within the first $p+q$ stages. □

Now that we an interesting tree to work with, I will show how Brocot used this tree for making approximations needed in his clockwork.

3.2 Application: Approximating Fractions

As discussed in the chapter about planetaria, gears used to drive clockwork are limited in number of teeth. I have already discussed some methods for approximating fractions that are inappropriate for gears. Here is how Brocot tackled the problem. I will first give the algorithm for an arbitrary fraction $\frac{p}{q}$, then I will give an example, clarifying the different steps taken. Then I will show this procedure is equivalent to following a specific path down the Stern-Brocot tree.

Definition 3.2.1. Brocot's algorithm for approximating a fraction

- Start with $\lfloor \frac{p}{q} \rfloor$ and $\lceil \frac{p}{q} \rceil$ and calculate the errors of these two approximations as follows: $e(\lceil \frac{p}{q} \rceil) = \lceil \frac{p}{q} \rceil \cdot q - p$, and similarly $e(\lfloor \frac{p}{q} \rfloor) = \lfloor \frac{p}{q} \rfloor \cdot q - p$. Make a table with three columns as follows:

numerator (n)	denominator (d)	error (e)
$\lceil \frac{p}{q} \rceil$	1	$\lceil \frac{p}{q} \rceil \cdot q - p$
$\lfloor \frac{p}{q} \rfloor$	1	$\lfloor \frac{p}{q} \rfloor \cdot q - p$

  The error in the table divided by the denominator gives the total error of the approximation in terms of the denominator of the original fraction. That is, ($\frac{n}{d} - \frac{p}{q} = \frac{e/d}{q}$).

- add the two rows column-wise to create a third row in between them:

numerator (n)	denominator (d)	error (e)
$\lceil \frac{p}{q} \rceil$	1	$\lceil \frac{p}{q} \rceil \cdot q - p$
$\lceil \frac{p}{q} \rceil + \lfloor \frac{p}{q} \rfloor$	2	$\lceil \frac{p}{q} \rceil \cdot q - p + \lfloor \frac{p}{q} \rfloor \cdot q - p$
$\lfloor \frac{p}{q} \rfloor$	1	$\lfloor \frac{p}{q} \rfloor \cdot q - p$

- If the new error e is positive, add this row to the row below it (which has a negative error term) to create a row in between them. If the new error term is negative, add this row to the row above it (which has a positive error term) to create a row in between them.

- Repeat the last step until you are satisfied with the approximation or until the error term is zero, then you have recovered the original fraction.

Example 3.2.2. Hayes uses the example of $\frac{191}{23}$ [34]. As $\lfloor \frac{191}{23} \rfloor = 8$ and $\lceil \frac{191}{23} \rceil = 9$, the first table looks like:

numerator (n)	denominator (d)	error (e)
9	1	16
8	1	-7

Adding these rows results in:

numerator (n)	denominator (d)	error (e)
9	1	16
17	2	9
8	1	-7

Adding the middle row to the lower row:

numerator (n)	denominator (d)	error (e)
9	1	16
17	2	9
25	3	2
8	1	-7

etc... each time adding the rows in such a way that the error decreases. The finished table looks like this:

numerator (n)	denominator (d)	error (e)
9	1	16
17	2	9
25	3	2
108	13	1
191	23	0
83	10	-1
58	7	-3
33	4	-5
8	1	-7

In this example, the best choice for approximating $\frac{191}{23}$ with the restriction that both the numerator and the denominator should be at least 7 and at most 100 is $\frac{83}{10}$, with an error of $\frac{-1/10}{23} = -1/230$.

Proposition 3.2.3. *Brocot's algorithm is sound, that is, the error in the table gained by adding the errors of the parent rows, is the true error for the fraction in that row compared to the fraction that is approximated. In mathematical terms:* $\frac{e/d}{q} = \frac{n}{d} - \frac{p}{q}$

Proof. Note: if $e = nq - dp$, then $\frac{e/d}{q} = \frac{nq-dp}{dq} = \frac{n}{d} - \frac{p}{q}$, so it is enough to prove that $e = nq - dp$. This is done by induction to the number of iterations of the algorithm. The base case is the start of the algorithm. As e is here defined to be $nq - dp$ (as d is 1 here), this case is ok. In the subsequent rows, the error is computed as the sum of the errors of the parents. That is, $e(\frac{n'+n''}{d'+d''}) = e(\frac{n'}{d'}) + e(\frac{n''}{d''})$. As Induction hypothesis we may

assume $e(\frac{n'}{d'}) = n'q - d'p$ and $e(\frac{n''}{d''}) = n''q - d''p$, so that:

$$\begin{aligned} e(\frac{n'+n''}{d'+d''}) &= e(\frac{n'}{d'}) + e(\frac{n''}{d''}) \\ &= n'q - d'p + n''q - d''p \\ &= (n'+n'')q - (d'+d'')p \end{aligned}$$

So indeed, $e = nq - dp$, and Brocot's algorithm gives the right error terms. □

3.2.1 Stern - Brocot

Brocot's algorithm is very mechanic, while Stern's work is very theoretic. Still, these two gentlemen actually hit upon the same idea. Brocot picks a very specific path down Stern's tree. When approximating a fraction $\frac{p}{q}$, he picks $\lfloor \frac{p}{q} \rfloor$ and $\lceil \frac{p}{q} \rceil$ as boundary values for the tree. He then constructs the tree by taking mediants. However, he is not interested in the entire tree, just the part that approximates the fraction $\frac{p}{q}$. So he in step n, he takes the left branch if the approximation is larger than the original fraction, the right branch if the approximation is smaller.

But Stern is not the only one whose work is closely related to Brocot's approximation algorithm. Where the first two columns of Brocot's table pointed to Stern's tree, the operations on the error column remind of Euclid.

3.2.2 Brocot - Euclid

Brocot's procedure is actually a a primitive variant of Euclid's algorithm. When only considering the error terms, starting with error $+e$ and $-f$, the procedure adds $-f$ to $+e$ 'once too many times', so that the outcome is negative, while adding $-f$ once less is still positive. In other words, it writes e as $e = b|f| - r$. After some more iterations:

$$\begin{aligned} e &= b_0|f| + r_0 \\ &= (b_0+1)|f| - r'_0 \\ r'_0 &= b_1 r_0 + r_1 \\ &= (b_1+1)r_0 - r'_1 \\ r'_1 &= b_2 r_1 + r_3 \\ &= (b_2+1)r_1 - r' \\ &\vdots \end{aligned}$$

Which is clearly a variant of Euclid's algorithm. Only in Brocot's algorithm, subtracting $r_n\ b_{n+1}+1$ times from r'_n is done once at a time, which could result in cumbersome work if one error term is relatively small and the other relatively large.

The connection between Stern-Brocot and continued fractions is not too hard to find now. However, there is another, less obvious connection between the two.

3.3 An Amusing Property of the Stern-Brocot Sequence

This section contains my interpretarion of a lemma (lemma 2.1) from [44]. It is the same proof as given there, but I have expanded it a bit to clarify some details.

Preciously, the boundaries were either $\frac{0}{1}$ and $\frac{1}{0}$ or two consecutive natural numbers. In this part, I will take a closer look at the tree resulting from the boundaries $\frac{0}{1}$ and $\frac{1}{1}$, which includes every positive rational smaller than 1. The different stages of the tree are represented by sets in the Stern-Brocot *sequence*. These sets $\mathcal{I}_n$ contain all the rational numbers constructed by the algorithm up to and including stage n. i.e. the first few elements of the Stern-Brocot sequence are:

$$\mathcal{I}_0 = \left\{\frac{0}{1}, \frac{1}{1}\right\}$$
$$\mathcal{I}_1 = \left\{\frac{0}{1}, \frac{1}{2}, \frac{1}{1}\right\}$$
$$\mathcal{I}_2 = \left\{\frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}\right\}$$
$$\mathcal{I}_3 = \left\{\frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}\right\}$$

To give a precise definition of this sequence is somewhat less obvious, but we will need in proofs later on. Kesseböhmer and Stratmann [44] formulate the definition as stated below, I haven't found a better way to express it.

Definition 3.3.1. Stern-Brocot sequence

The Stern-Brocot sequence is a sequence of sets $\mathcal{I}_n$,

$$\mathcal{I}_n = \left\{ \frac{s_{n,k}}{t_{n,k}} \mid k = 1, \ldots 2^n + 1 \right\}$$

where $s_{n,k}$ and $t_{n,k}$ are inductively defined as:

$$s_{0,1} = 0 \qquad t_{0,1} = 1 \tag{3.1}$$
$$s_{0,2} = 1 \qquad t_{0,2} = 1 \tag{3.2}$$
$$\tag{3.3}$$
$$s_{n+1,k} = \begin{cases} s_{n,l} & \text{if } k = 2l - 1 \\ s_{n,l} + s_{n,l+1} & \text{if } k = 2l \end{cases} \tag{3.4}$$
$$t_{n+1,k} = \begin{cases} t_{n,l} & \text{if } k = 2l - 1 \\ t_{n,l} + t_{n,l+1} & \text{if } k = 2l \end{cases} \tag{3.5}$$

Taking a close look at the sets in the Stern-Brocot sequence, we see this definition makes sense. For odd k, the elements of the previous set are included, and for even k, the algorithm of adding numerators and denominators is applied.

By taking differences between consecutive sets of the sequence, we gain sets containing the newborns of each stage:

$$\mathcal{I}_1 \backslash \mathcal{I}_0 = \left\{ \frac{1}{2} \right\}$$
$$\mathcal{I}_2 \backslash \mathcal{I}_1 = \left\{ \frac{1}{3}, \frac{2}{3} \right\}$$
$$\mathcal{I}_3 \backslash \mathcal{I}_2 = \left\{ \frac{1}{4}, \frac{2}{5}, \frac{3}{5}, \frac{3}{4} \right\}$$

These difference sets have an amusing property. If we express all the elements of $\mathcal{I}_n \backslash \mathcal{I}_{n-1}$ as continued fractions, then the sum of the quotients

a_i is $n+1$. Take for example $\mathcal{I}_3\backslash\mathcal{I}_2$:

$$\begin{aligned}\mathcal{I}_3\backslash\mathcal{I}_2 &= \left\{\frac{1}{4}, \frac{2}{5}, \frac{3}{5}, \frac{3}{4}\right\} \\ &= \left\{\frac{1}{4}, \frac{1}{2+\frac{1}{2}}, \frac{1}{1+\frac{1}{1+\frac{1}{2}}}, \frac{1}{1+\frac{1}{3}}\right\} \\ &= \{[0;4], [0;2,2], [0;1,1,2], [0;1,3]\}\end{aligned}$$

$$\Rightarrow \quad \sum_i a_i = 4 = 3+1$$

To state this more precisely, consider the following sets:

$$\mathcal{A}_k^n = \left\{ [0; a_1, \ldots, a_k] \;\middle|\; \sum_{i=1}^{k} a_i = n, a_k \neq 1 \right\}$$

The set $\mathcal{A}_k^n$ contains all continued fractions with k quotients adding up to n. The union $\bigcup_{k=1}^{n-1} \mathcal{A}_k^n$ contains all continued fractions of which the quotients have sum n. Our theorem expressed in the example above, then becomes:

Theorem 3.3.2.

$$\mathcal{I}_n\backslash\mathcal{I}_{n-1} = \bigcup_{k=1}^{n} \mathcal{A}_k^{n+1} \tag{3.6}$$

As the sets $\mathcal{I}_n$ are recursively defined, a proof by induction is quite obvious. However, to carry through the induction, we need a relation between the elements of $\mathcal{I}_n\backslash\mathcal{I}_{n-1}$ and $\mathcal{I}_{n-1}\backslash\mathcal{I}_{n-2}$. The following Lemma provides us with such a relation.

Lemma 3.3.3.

$$\mathcal{I}_n\backslash\mathcal{I}_{n-1} = \left\{ \frac{1}{x+1}, \frac{x}{x+1} \mid x \in \mathcal{I}_{n-1} \right\}$$

To my regret, I did not succeed in proving this Lemma[2]. So for the time being I will just assume it is true and use it in the proof of Theorem 3.3.2:

[2] [11] claims to prove it, but I do not understand the reasoning presented there.

Proof. By induction.
Base: n=1

$$\begin{aligned}\mathcal{I}_1\backslash\mathcal{I}_0 &= \left\{\frac{1}{2}\right\}\\ &= \{\,[0;2]\,\}\\ &= \mathcal{A}_1^2\\ &= \bigcup_{k=1}^{1}\mathcal{A}_k^2\end{aligned}$$

To prove the induction step, I will first show that $\mathcal{I}_n\backslash\mathcal{I}_{n-1} \subseteq \bigcup_{k=1}^{n}\mathcal{A}_k^{n+1}$. Then, I will note the two sets have an equal cardinality. As both sets are finite, and one is completely contained in the other, this proves they are the same.

In proving the inclusion we need Lemma 3.3.3. I want to show that:

$$\mathcal{I}_n\backslash\mathcal{I}_{n-1} \subseteq \bigcup_{k=1}^{n}\mathcal{A}_k^{n+1}$$

By Lemma 3.3.3 we have:

$$\mathcal{I}_n\backslash\mathcal{I}_{n-1} = \left\{\frac{1}{x+1}, \frac{x}{x+1} \mid x \in \mathcal{I}_{n-1}\right\}$$

So I need to show that, for each $x \in \mathcal{I}_{n-1}\backslash\mathcal{I}_{n-2}$, both $\frac{1}{x+1}$ and $\frac{x}{x+1}$ are elements of $\bigcup_{k=1}^{n}\mathcal{A}_k^{n+1}$.

So fix any $x \in \mathcal{I}_{n-1}\backslash\mathcal{I}_{n-2}$ arbitrarily. By the induction hypothesis, we know that $\mathcal{I}_{n-1}\backslash\mathcal{I}_{n-2} = \bigcup_{k=1}^{n-1}\mathcal{A}_k^{n}$. So we know the sum of the quotients of x equals n.

Suppose $x = [0; a_1, \ldots, a_k]$ (so that $\sum_{i=1}^{k} a_i = n$). Consider $\frac{1}{x+1}$:

$$\begin{aligned}\frac{1}{x+1} &= \frac{1}{[0;a_1,\ldots,a_k]+1}\\ &= \cfrac{1}{1+0+\cfrac{1}{a_1+\cfrac{1}{\ddots+\cfrac{1}{a_k}}}}\\ &= [0;1,a_1,\ldots,a_k]\end{aligned}$$

Hence the sum of the quotients of $\frac{1}{x+1}$ equals $n+1$. That is, $\frac{1}{x+1} \in \mathcal{A}_{k+1}^{n+1}$

A similar argument holds for $\frac{x}{x+1}$:

$$\begin{aligned}
\frac{x}{x+1} &= \cfrac{1}{1+\cfrac{1}{x}} \\
&= \cfrac{1}{1+\cfrac{1}{[0;a_1,\ldots,a_k]}} \\
&= \cfrac{1}{1+\cfrac{1}{\cfrac{1}{a_1+\cfrac{1}{\ddots+\cfrac{1}{a_k}}}}} \\
&= \cfrac{1}{1+a_1+\cfrac{1}{\ddots+\cfrac{1}{a_k}}} \\
&= [0;a_1+1,a_2\ldots,a_k]
\end{aligned}$$

Hence the sum of the quotients of $\frac{x}{x+1}$ equals $n+1$. That is, $\frac{x}{x+1} \in \mathcal{A}_k^{n+1}$. Hence both $\frac{1}{x+1}$ and $\frac{x}{x+1}$ are elements of $\bigcup_{k=1}^{n} \mathcal{A}_k^{n+1}$. We may conclude that indeed

$$\mathcal{I}_n \backslash \mathcal{I}_{n-1} \subseteq \bigcup_{k=1}^{n} \mathcal{A}_k^{n+1}$$

For equality of these sets, we will compute their cardinality:
$\#\mathcal{I}_n = 2^n + 1$, which is easily proven by induction. So:

$$\begin{aligned}
\#(\mathcal{I}_n \backslash \mathcal{I}_{n-1}) &= \#\mathcal{I}_n \ - \ \#\mathcal{I}_{n-1} \\
&= 2^n + 1 - (2^{n-1}+1) \\
&= 2^{n-1}
\end{aligned}$$

Computing the cardinality of $\bigcup_{k=1}^{n} \mathcal{A}_k^{n+1}$ will take some more work. First note that:

$$\#\mathcal{A}_k^{n+1} = \binom{n-1}{k-1}$$

which I will explain using balls and sticky walls. To get the number of elements in $\mathcal{A}_k^{n+1}$, I need to count the number of possible sequences $(a_1, \ldots, a_k)$ such that $[0; a_1, \ldots, a_k]$ is a continued fraction (that is, each $a_i \geq 1$ and $a_k > 1$), and $\sum_{i=1}^{k} a_i = n + 1$.

Suppose I have $n + 1$ balls in a row:

$$\bullet \ \bullet \ \cdots \ \bullet$$

I want to divide them into k groups, the number of balls in group i representing the value of a_i. This is done by picking $k - 1$ balls and stick a wall to their left:

$$\bullet \ \bullet \ \big| \circ \ \bullet \ \bullet \ \bullet \ \big| \circ \ \bullet \ \bullet \ \cdots \ \big| \circ$$

The $k - 1$ walls divide the balls into k groups. The stickiness ensures each group contains at least one ball (you cannot stick two walls to the same ball). The number of ways these walls can be placed is equal to the number of ways I can pick $k - 1$ balls out of $n + 1$ balls, that is: $\binom{n+1}{k-1}$. However, there are some more constraints. Using this construction, I could pick the very fist ball in the row to stick a wall to its left, resulting in $a_1 = 0$. To prevent this, I take 1 ball apart before placing the walls, and add it to the first group once the walls are in place, ensuring $a_1 \geq 1$. In the same way, I reserve one ball beforehand for the last group, a_k. As the walls are placed on the left side of the balls, the procedure gives me $a_k \geq 1$, adding my reserved ball yields the desired $a_k > 1$. So I am left with $n + 1 - 2 = n - 1$ balls to stick the $k - 1$ walls to, yielding $\binom{n-1}{k-1}$ ways to do so. Hence

$$\#\mathcal{A}_k^{n+1} = \binom{n-1}{k-1}$$

From this fact, it is only a small step to $\#\bigcup_{k=1}^{n} \mathcal{A}_k^{n+1}$. Newton's Bino-

mial Theorem is used here:

$$\begin{aligned}
\#\bigcup_{k=1}^{n}\mathcal{A}_k^{n+1} &= \sum_{k=1}^{n}\#\mathcal{A}_k^{n+1} \\
&= \sum_{k=1}^{n}\binom{n-1}{k-1} \\
&= \sum_{k=1}^{n}\binom{n-1}{k-1}1^{k-1}\cdot 1^{n-1-(k-1)} \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k}1^{k}\cdot 1^{n-1-k} \\
&= (1+1)^{n-1} \\
&= 2^{n-1}
\end{aligned}$$

So we have:

$$\#(\mathcal{I}_n\backslash\mathcal{I}_{n-1}) = 2^{n-1} = \#\bigcup_{k=1}^{n}\mathcal{A}_k^{n+1}$$

As each element of $\mathcal{I}_n\backslash\mathcal{I}_{n-1}$ is also an element of $\bigcup_{k=1}^{n}\mathcal{A}_k^{n+1}$, and both sets have the same number of elements, they must be equal. □

This proves that the elements of $\mathcal{I}_n\backslash\mathcal{I}_{n-1}$ as continued fractions, then the sum of the quotients a_i is $n+1$.

3.3.1 Some last Comments

For a brilliant explanation of the Stern-Brocot tree and its properties, please go to `http://www.ams.org/samplings/feature-column/fcarc-stern-brocot` [3]. Here, David Austin explains the whole subject way better than I ever could, incuding nice geometrical interpretations.

Chapter 4

Sums of squares

Chris Kooloos

Question 1. Let $n \in \mathbb{N}$. Can we find a and b in $\mathbb{N}$ such that $a^2 + b^2 = n$?

First of all we have the following lemma, which tells us that we only have to consider the special case where n is a prime.

Lemma 4.0.4. *The product of two natural numbers that each are the sum of two squares is again a sum of two squares.*

Proof:
Let a, b, x and y be natural numbers. We have:

$$(a^2 + b^2) \cdot (x^2 + y^2) = a^2x^2 + a^2y^2 + b^2x^2 + b^2y^2 =$$

$$a^2x^2 - 2abxy + b^2y^2 + a^2y^2 + 2abxy + b^2x^2 = (ax - by)^2 + (ay + bx)^2.$$

Suppose we have a and b such that $a^2 + b^2 = p$ for some prime p. All squares equal 0 or 1 modulo 4 so $a^2 + b^2$ equals $0, 1$ or 2 modulo 4. This implies that we only have to consider primes p such that $p = 2$ or $p \equiv_4 1$. Of course, we have $2 = 1^2 + 1^2$.

4.0.2 Theorem 1

et p be a prime such that $p \equiv_4 1$. Then we can find $a, b \in \mathbb{N}$ such that $a^2 + b^2 = p$. **Proof:**
Find w such that $w^2 \equiv_p -1$ and $0 < w < \frac{p}{2}$. This is always possible:
The multiplicative group $\mathbb{F}_p^*$ is cyclic. Therefore, there exists a generator g for this group. $g^{p-1} \equiv_p 1$ and $g^{\frac{p-1}{2}} \equiv_p -1$.
Now, both $g^{\frac{p-1}{4}}$ and $g^{\frac{3p-3}{4}}$ are roots of $x^2 + 1$ over $\mathbb{F}_p[x]$ and one of them

has a representative w such that $w < \frac{p}{2}$.
Consider the regular continued fraction expansion of the rational number $\frac{w}{p}$.
The sequence $q_0, q_1, \cdots$ of nominators of the convergents is strictly increasing and $q_0 = 1$. Therefore we can find an m such that $q_m < \sqrt{p} < q_m + 1$. We know that:

$$| \frac{w}{p} - \frac{p_m}{q_m} | < \frac{1}{q_{m+1} \cdot q_m}.$$

If we set $a := w \cdot q_m - p \cdot p_m$, then:

$$|\frac{a}{p \cdot q_m}| = |\frac{w}{p} - \frac{p_m}{q_m} | < \frac{1}{q_{m+1} \cdot q_m}.$$

So we have:

$$|a| < \frac{p}{\sqrt{p}} = \sqrt{p}.$$

Together with $q_m < \sqrt{p}$, we have:

$$a^2 + q_m^2 < p + p = 2p.$$

Furthermore, $a = w \cdot q_m - p \cdot p_m \equiv_p w \cdot q_m$. So, because we have $w^2 \equiv_p -1$:

$$a^2 + q_m^2 \equiv_p 0.$$

Therefore, if we set $b = q_m$, we have:

$$a^2 + b^2 = p.$$

The proof gives us an algorithm to find a and b in which we calculate the continued fraction expansion of a rational number until the nominators of the partial quotiënts get bigger than $\sqrt{p}$. The problem is to find the number w or to find the generator for $\mathbb{F}_p^*$ so that we can calculate w.

We now consider a shorter algorithm using the Euclidean algorithm. Unfortunately, this algorithm also starts with finding a w as above. However, in this algorithm we don't have to calculate partial quotiënts.

Carry out the Euclidean algorithm on $\frac{p}{w}$ (instead of $\frac{w}{p}$), producing the sequence $r_1, r_2, \cdots$ of remainders. When we first encounter a remainder r_k such that $r_k < \sqrt{p}$, we calculate the next remainder r_{k+1} and stop after doing so.

Claim 1.

$$\begin{aligned} p &= r_k^2 + r_{k+1}^2 && \text{if } r_1 > 1 \\ p &= w^2 + 1 && \text{if } r_1 = 1. \end{aligned}$$

Before proving this claim, we prove a lemma about the shape of the continued fraction expansion of $\frac{p}{w}$.

Lemma 4.0.5. *Let $\frac{p}{w}$ be an irreducible fraction where $p > w > 1$ and $w^2 \equiv_p -1$. Then $\frac{p}{w}$ has a palindromic (symmetric) continued fraction expansion with an even number of partial quotiëns.*

Proof:
We have $w^2 + 1 = v \cdot p$ for some $v \in \mathbb{N}$.
Consider the continued fraction expansion of $\frac{p}{w}$:

$$\frac{p}{w} = [a_0; a_1, a_2, \cdots a_n] = \frac{p_n}{q_n}$$

where we can assume that n is odd. The greatest common divisor of p and w is 1 by assumption and the greatest common divisor of p_n and q_n is always 1 so we have $p = p_n$ and $w = q_n$. We know that:

$$p_{n-1} \cdot q_n - p_n \cdot q_{n-1} = (-1)^n = -1$$

So

$$1 = p_n \cdot q_{n-1} - p_{n-1} \cdot q_n$$

which we submit in $w^2 + 1 = v \cdot p$ to see:

$$q_n^2 + p_n \cdot q_{n-1} - p_{n-1} \cdot q_n = v \cdot p_n$$

$$q_n \cdot (q_n - p_{n-1}) = p_n \cdot (v - q_{n-1})$$

This implies $p_n | (q_n - p_{n-1})$.
We have

1. $p_n > q_n > 0$
2. $p_n = a_n \cdot p_{n-1} + p_{n-2} > p_{n-1} > 0$

Together these imply $p_n > |q_n - p_{n-1}|$ so we can conclude $q_n - p_{n-1} = 0$, and therefore, $q_n = p_{n-1}$.
Now we have:

$$\frac{p_n}{p_{n-1}} = \frac{p_n}{q_n} = [a_0; a_1, \cdots a_n]$$

But we also have:

$$\frac{p_n}{p_{n-1}} = [a_n; a_{n-1}, \cdots, a_1, a_0]$$

Because:

$$p_n = a_n \cdot p_{n-1} + p_{n-2}$$

$$\frac{p_n}{p_{n-1}} = a_n + \frac{p_{n-2}}{p_{n-1}}$$
$$\frac{p_{n-1}}{p_{n-2}} = a_{n-1} + \frac{p_{n-3}}{p_{n-2}}$$

and so on..
So now we have

$$[a_0; \cdots, a_n] = [a_n; \cdots, a_0]$$

And n is odd, say $n = 2k+1$. We have:

$$\frac{p}{w} = [a_0; a_1, \cdots, a_k, a_k, \cdots, a_0] = \frac{p_{2k+1}}{p_{2k}}$$

Which proves the lemma.

Observe:
$\frac{p}{w}$ has continued fraction expansion $[a_0; \cdots, a_0]$ with convergents $\frac{p_0}{q_0}, \frac{p_1}{q_1}, \cdots$.
$\frac{w}{p}$ is equal to $0 + \frac{1}{\frac{p}{w}}$ so $\frac{w}{p}$ has continued fraction expansion $[0; a_0, a_1, \cdots, a_0]$.
So if $\frac{p'_0}{q'_0}, \frac{p'_1}{q'_1}, \cdots$ is the sequence of convergents of the continued fraction expansion of $\frac{w}{p}$, which we used in the first algorithm in the theorem above, we have, for all m:

$$\frac{p'_{m+1}}{q'_{m+1}} = \frac{q_m}{p_m}.$$

Lemma 4.0.6. $p_k^2 + p_{k-1}^2 = p$

Proof:
This proof is given in Perron (blz. 29).

Proof of Claim 1:
Suppose $r_1 > 1$. We want to prove $p = r_k^2 + r_{k+1}^2$.
We have $p_{2k+1} = p$ and $p_{2k} = w$.
The recursion formula for p_n gives us:

$$p = a_0 \cdot w + p_{2k-1}$$

$$w = a_1 \cdot p_{2k-1} + p_{2k-2}$$

and so on..
These equations are identical to those in the Euclidean algorithm for $\frac{p}{w}$.
Hence,

$$p_{2k-1} = r_1,\ p_{2k-2} = r_2, \cdots,\ p_{k+1} = r_{k-1},\ p_k = r_k,\ p_{k-1} = r_{k+1}, \cdots$$

Together with Lemma 1.3 this gives us:

$$p = {r_k}^2 + {r_{k+1}}^2.$$

From this we conclude $r_k < \sqrt{p}$. To prove that k is indeed the smallest k_0 such that $r_{k_0} < \sqrt{p}$, first assume $k = 1$. Then surely 1 is the smallest k_0 such that $r_{k_0} < \sqrt{p}$.
Now assume $k > 1$. We have: $r_{k-1} = p_{k+1}$.
From the observation above, we have $p_{k+1} = {q'}_{k+2}$, where ${q'}_{k+2}$ is the nominator of the $k+2$'nd convergent of the continued fraction expansion of $\frac{w}{p}$.
There is a natural number z such that:

$${q'}_{k+2}^{\,2} = ({q'}_{k+1} \cdot z + {q'}_k)^2 = {q'}_{k+1}^{\,2} + 2z \cdot {q'}_{k+1} \cdot {q'}_k + {q'}_k^{\,2} >$$

$${q'}_{k+1}^{\,2} + {q'}_k^{\,2} = {p_k}^2 + {p_{k+1}}^2 = p.$$

So $r_{k-1} = p_{k+1} = {q'}_{k+2} > \sqrt{p}$ which proves that r_k is the first remainder smaller than $\sqrt{p}$.
Now assume $r_1 = 1$.

Then $p = w \cdot a_0 + 1$.
Also $\frac{p}{w} = [a_0; a_0]$, so $w = a_0$.
We conclude $p = w^2 + 1$.

Chapter 5

Pell equation

Merlijn Keune

A Pell equation is an equation of the form $x^2 - dy^2 = \pm 1$, with $d \in \mathbb{Z}_{\geq 2}$ not a perfect square. The equation is named after the English mathematician John Pell (1610-1685), who had nothing to do with this equation. In fact William Brouncker (1620-1684) came up with a solution method, but Euler mistakenly credited Pell for this. Brouncker's method is in substance identical to a method known to Indian mathematicians at least six centuries earlier. The equation also occurred in Greek mathematics, but there is no evidence that they were able to solve it.

In this chapter we will use the continued fraction expansion of $\sqrt{d}$ and quadratic irrationals in general to solve the Pell equation. The first proposition makes clear why this is a sensible thing to do.

Proposition 5.0.7. *Let $(p, q) \in \mathbb{Z}^2$ be a solution to $x^2 - dy^2 = \pm 1$. Then $p = p_k$ and $q = q_k$ for some convergent $\frac{p_k}{q_k}$ of $\sqrt{d}$.*

Proof $\frac{p^2}{q^2} = d \pm \frac{1}{q^2}$, so $p \geq q$. Therefore

$$|p - \sqrt{d}q| = \frac{1}{p + \sqrt{d}q} \leq \frac{1}{(1 + \sqrt{d})q} < \frac{1}{2q}.$$

So

$$|\frac{p}{q} - \sqrt{d}| < \frac{1}{2q^2}.$$

Furthermore $\gcd(p, q) = 1$, so the proposition follows from what we already know.

Before we can have a good look at the continued fraction expansion of quadratic irrationals we need some definitions and a few immediate results.

Definitions 5.0.8. *Let α be a quadratic irrational. We can write $\alpha = a + b\sqrt{d}$ with $a, b \in \mathbb{Q}$ and d some square-free natural number. Then $a - b\sqrt{d}$ is its conjugate, denoted by $\overline{\alpha}$. There are unique $a \in \mathbb{N}^*, b, c \in \mathbb{Z}$, $\gcd(a, b, c) = 1$ such that α is a root of ax^2+bx+c. Its discriminant is $\operatorname{disc}(\alpha) = b^2-4ac$. α is called reduced if $\alpha > 1$ and $-1 < \overline{\alpha} < 0$. For the continued fraction expansion we shall use the map*

$$\varphi : \mathbb{R} \setminus \mathbb{Q} \to \mathbb{R} \setminus \mathbb{Q},\ \alpha \mapsto \frac{1}{\alpha - \lfloor \alpha \rfloor}.$$

So we have

$$\alpha = [\lfloor \alpha \rfloor; \varphi(\alpha)] = [\lfloor \alpha \rfloor; \lfloor \varphi(\alpha) \rfloor, \varphi^2(\alpha)] = \dots.$$

Proposition 5.0.9. *Let α be a quadratic irrational. Then $\varphi(\alpha)$ is a quadratic irrational and $\operatorname{disc}(\alpha) = \operatorname{disc}(\varphi(\alpha))$. If α is reduced, then $\varphi(\alpha)$ is reduced.*

Proof $\alpha \in \mathbb{Q}(\sqrt{d})$ for some $d \in \mathbb{N}^*$, so obviously $\varphi(\alpha) \in \mathbb{Q}(\sqrt{d})$. Let α be a root of $ax^2 + bx + c$, so $\operatorname{disc}(\alpha) = b^2 - 4ac$. We show that the discriminant is invariant under the maps $\alpha \mapsto \alpha - 1$ and $\alpha \mapsto \frac{1}{\alpha}$. $\alpha - 1$ is a root of $ax^2 + (2a + b)x + a + b + c$:

$$a(\alpha-1)^2+(2a+b)(\alpha-1)+a+b+c = a\alpha^2-2a\alpha+a+2a\alpha-2a+b\alpha-b+a+b+c = 0.$$

Also $\gcd(a, 2a+b, a+b+c) = 0$ and a simple calculation shows that $\operatorname{disc}(\alpha - 1) = a^2 - 4ac$. $\frac{1}{\alpha}$ is a root of $cx^2 + bx + a$ (divide everything by α^2) which obviously gives the same discriminant.

Because α is reduced, we have the following inequalities:

$$\lfloor \alpha \rfloor < \lfloor \alpha \rfloor - \overline{\alpha} < \lfloor \alpha \rfloor + 1,$$

$$\frac{1}{\lfloor \alpha \rfloor + 1} < \frac{1}{\lfloor \alpha \rfloor - \overline{\alpha}} < \frac{1}{\lfloor \alpha \rfloor},$$

$$-1 \le -\frac{1}{\lfloor \alpha \rfloor} < \overline{\varphi(\alpha)} < -\frac{1}{\lfloor \alpha \rfloor + 1} < 0.$$

So $\varphi(\alpha)$ is also reduced. (Note that $\varphi(\alpha) > 1$.)

Lemma 5.0.10. *There are only finitely many reduced quadratic irrationals β with a given discriminant d.*

PROOF Let β be a root of $ax^2 + bx + c$. Then

$$\beta\overline{\beta} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{b^2 - b^2 + 4ac}{4a^2} = \frac{c}{a}.$$

If β is reduced, then $\beta\overline{\beta} < 0$, so $\frac{c}{a} < 0$ and thus $c < 0$. But

$$b^2 - 4ac = b^2 + 4a(-c) = d$$

with b^2 and $4a(-c)$ positive, so a, b, c are all bounded.

We are now ready to prove some classical results about reduced quadratic irrationals.

Theorem 5.0.11. *A reduced quadratic irrational α has a strictly periodic continued fraction expansion. That is: $\alpha = [\overline{a_0; a_1, \ldots, a_{n-1}}]$*

PROOF The set of quadratic irrationals with discriminant $d = \mathrm{disc}(\alpha)$ is invariant under φ. We show that the restriction of φ to this set is injective.

Suppose $\varphi(\beta_1) = \varphi(\beta_2)$, so

$$\frac{1}{\beta_1 - \lfloor\beta_1\rfloor} = \frac{1}{\beta_2 - \lfloor\beta_2\rfloor},\ \beta_1 - \lfloor\beta_1\rfloor = \beta_2 - \lfloor\beta_2\rfloor.$$

So also the conjugates are the same:

$$\overline{\beta_1} - \lfloor\beta_1\rfloor = \overline{\beta_2} - \lfloor\beta_2\rfloor.$$

But since $-1 < overline\beta_1, \overline{\beta_2} < 0$ this implies $\lfloor\beta_1\rfloor = \lfloor\beta_2\rfloor$, and thus $\beta_1 = \beta_2$.

Because the set is finite there must occur repetition in the sequence

$$\alpha, \varphi(\alpha), \varphi^2(\alpha), \ldots$$

Injectivity ensures this sequence is strictly periodic.

Proposition 5.0.12. *If α is reduced with $\alpha = [\overline{a_0; a_1, \ldots, a_{n-1}}]$, then*

$$-\frac{1}{\overline{\alpha}} = [\overline{a_{n-1}; a_{n-2}, \ldots, a_0}].$$

PROOF Let us write $\alpha_i = \varphi^i(\alpha)$. Then

$$\alpha = \alpha_0 = \lfloor\alpha\rfloor + \alpha_0 - \lfloor\alpha\rfloor = a_0 + \frac{1}{\alpha_1}.$$

Continuing like this gives

$$\alpha_0 = a_0 + \frac{1}{\alpha_1}, \alpha_1 = a_1 + \frac{1}{\alpha_2}, \ldots, \alpha_{n-1} = a_{n-1} + \frac{1}{\alpha_n} = a_{n-1} + \frac{1}{\alpha_0},$$

so for the conjugates we have

$$\overline{\alpha_0} = a_0 + \frac{1}{\overline{\alpha_1}}, \overline{\alpha_1} = a_1 + \frac{1}{\overline{\alpha_2}}, \ldots, \overline{\alpha_{n-1}} = a_{n-1} + \frac{1}{\overline{\alpha_n}} = a_{n-1} + \frac{1}{\overline{\alpha_0}}.$$

Rewriting this gives

$$-\frac{1}{\overline{\alpha_1}} = a_0 + (-\overline{\alpha_0}), -\frac{1}{\overline{\alpha_2}} = a_1 + (-\overline{\alpha_1}), \ldots, -\frac{1}{\overline{\alpha_n}} = a_{n-1} + (-\overline{\alpha_{n-1}}) = -\frac{1}{\overline{\alpha_0}}.$$

So a_{n-1} is the floor of $-\frac{1}{\overline{\alpha}}$ and

$$\varphi(-\frac{1}{\overline{\alpha}}) = -\frac{1}{\overline{\alpha_{n-1}}}$$

and so on.

We now turn our attention to a special quadratic irrational, being $\sqrt{d}$. We need its continued fraction expansion to solve the Pell equation $x^2 - dy^2 = \pm 1$.

Lemma 5.0.13. *$\varphi(\sqrt{d})$ is reduced.*

PROOF We have

$$\varphi(\sqrt{d}) = \frac{1}{\sqrt{d} - \lfloor \sqrt{d} \rfloor} > 1$$

and

$$\overline{\varphi(\sqrt{d})} = \frac{1}{-\sqrt{d} - \lfloor \sqrt{d} \rfloor} = -\frac{1}{\sqrt{d} + \lfloor \sqrt{d} \rfloor}.$$

Theorem 5.0.14. *The continued fraction expansion of $\sqrt{d}$ is of the form*

$$\sqrt{d} = [a_0; \overline{a_1, a_2, \ldots, a_2, a_1, 2a_0}].$$

PROOF $\varphi(\sqrt{d}) = [\overline{a_1; a_2, \ldots, a_n}]$, so $\sqrt{d} = [a_0; \overline{a_1, a_2, \ldots, a_n}]$. Then

$$\sqrt{d} + a_0 = [2a_0; \overline{a_1, \ldots, a_{n-1}, a_n}].$$

But applying the previous proposition to $\varphi(\sqrt{d})$ also gives

$$\sqrt{d} + a_0 = [\overline{a_n; a_{n-1}, \ldots, a_1}].$$

So $a_n = 2a_0$ and $a_i = a_{n-i}$ for $1 \leq i \leq n-1$.

This brings us to the main theorem on this topic. We already know that all solutions to the Pell equation can be found in the continued fraction expansion of $\sqrt{d}$, but we can also specify where exactly they can be found. We use some equalities we encountered earlier on.

Theorem 5.0.15. *Let $\sqrt{d} = [a_0; \overline{a_1, \ldots, a_m}]$, where m is the smallest period. Then (p_n, q_n) is a solution of $x^2 - dy^2 = \pm 1$ if and only if $m|n+1$. Moreover, it is a solution to $x^2 - dy^2 = (-1)^{n+1}$.*

PROOF $\sqrt{d} = [a_0; a_1, \ldots, a_n, \alpha_{n+1}]$, so

$$\sqrt{d} = \frac{p_n\alpha_{n+1} + p_{n-1}}{q_n\alpha_{n+1} + q_{n-1}}.$$

Substituting $\sqrt{d}$ gives

$$p_n - q_n\sqrt{d} = \frac{p_nq_n\alpha_{n+1} + p_nq_{n-1} - p_nq_n\alpha_{n+1} - q_np_{n-1}}{q_n\alpha_{n+1} + q_{n-1}} = \frac{(-1)^{n+1}}{q_n\alpha_{n+1} + q_{n-1}}.$$

Since $p_n^2 - dq_n^2 = (p_n - \sqrt{d}q_n)(p_n + \sqrt{d}q_n)$ and $p_n + \sqrt{d}q_n$ is positive, we now know that if (p_n, q_n) is a solution, it is a solution to $x^2 - dy^2 = (-1)^{n+1}$.

Also, by eliminating the denominator, we get

$$\sqrt{d}(q_n\alpha_{n+1} + q_{n-1}) = p_n\alpha_{n+1} + p_{n-1},$$

so

$$(p_n - q_n\sqrt{d})\alpha_{n+1} = q_{n-1}\sqrt{d} - p_{n-1}.$$

Multiplying by $p_n + q_n\sqrt{d}$ gives

$$\begin{aligned}(p_n^2 - dq_n^2)\alpha_{n+1} &= (q_{n-1}\sqrt{d} - p_{n-1})(p_n + q_n\sqrt{d}) \\ &= (p_nq_{n-1} - p_{n-1}q_n)\sqrt{d} + k \\ &= (-1)^{n+1}\sqrt{d} + k\end{aligned}$$

for some $k \in \mathbb{Z}$.
$\Rightarrow$: If (p_n, q_n) is a solution, then $(-1)^{n+1}\alpha_{n+1} = (-1)^{n+1}\sqrt{d} + k$, so $\alpha_{n+1} = \sqrt{d} + k$. Then

$$\alpha_{n+2} = \frac{1}{\alpha_{n+1} - \lfloor\alpha_{n+1}\rfloor} = \frac{1}{\sqrt{d} - \lfloor\sqrt{d}\rfloor} = \alpha_1.$$

So $m|n+1$.

$\Leftarrow$: Suppose $m|n+1$. Then $\sqrt{d} = [a_0; \overline{a_1, \ldots, a_{n+1}}]$ and $a_{n+1} = 2a_0$. So

$$\begin{aligned}\sqrt{d} &= [a_0; a_1, \ldots, a_{n+1}, \overline{a_1, \ldots, a_{n+1}}] \\ &= [a_0; a_1, \ldots, a_n, a_0 + \sqrt{d}] \\ &= \tfrac{p_n(a_0+\sqrt{d})+p_{n-1}}{q_n(a_0+\sqrt{d})+q_{n-1}}.\end{aligned}$$

Multiplying by the denominator and rearranging gives

$$(q_n a_0 + q_{n-1} - p_n)\sqrt{d} = p_n a_0 + p_{n-1} - dq_n.$$

Since $\sqrt{d}$ is irrational, both $q_n a_0 + q_{n-1} - p_n$ and $p_n a_0 + p_{n-1} - dq_n$ must be equal to 0. Multiplying by p_n and q_n gives

$$p_n q_n a_0 + p_n q_{n-1} - p_n^2 = 0,\ p_n q_n a_0 + q_n p_{n-1} - dq_n^2 = 0.$$

By subtracting these equalities we get

$$p_n^2 - dq_n^2 + p_{n-1}q_n - p_n q_{n-1} = 0,$$

so

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

Example 5.0.16. Let us solve the equation $x^2 - 14y^2 = \pm 1$.

Calculating the continued fraction expansion of $\sqrt{14}$ gives us:

$$\sqrt{14} = [3; \overline{1, 2, 1, 6}].$$

Notice that this indeed is of the form we proved it must have. Since the length of the period is 4, we already know that the equation $x^2 - 14y^2 = -1$ has no solutions. As seen earlier, we can determine p_n and q_n as follows:

n :	-2	-1	0	1	2	3	4	5	6	7
a_n :			3	1	2	1	6	1	2	1
p_n :	0	1	3	4	11	15	101	116	333	449
q_n :	1	0	1	1	3	4	27	31	89	120

The smallest solution to the equation can be found in the table under $n = 3$: $15^2 - 14 \cdot 4^2 = 1$. $n = 7$ gives us the next solution: $449^2 - 14 \cdot 120^2 = 1$. Continuing like this, we see that the solutions grow rather fast; the sixth solution that will be found is $362074049^2 - 14 \cdot 96768360^2 = 1$.

Remark 5.0.17. Obviously this method works fine for finding small solutions, but if you are interested in lets say the 37th smallest solution, the calculation gets quite nasty. However, if (x, y) is a solution of $x^2 - dy^2 = \pm 1$, then $x - \sqrt{d}y$ is a unit of the ring $\mathbb{Z}[\sqrt{d}]$. In algebraic number theory, *Dirichlet's unit theorem* states that the group of units of $\mathbb{Z}[\sqrt{d}]$ is of the form $\langle -1, \alpha \rangle$, with α of infinite order. For our use this means that once you found the first solution (also called the *fundamental* solution) (p_{m-1}, q_{m-1}), you get all other solutions by calculating the unit $(p_{m-1} - \sqrt{d}q_{m-1})^k$ for $k \in \mathbb{N}^*$.

Chapter 6

Markov numbers

Chris Kooloos

Consider the following equation of integers:

$$m^2 + m_1^2 + m_2^2 = 3mm_1m_2$$

We will call this the *Markov equation.*
The solutions of this equation are triples of integers, and we consider only triples of positive integers. The numbers that occur in such a solution are called *Markov numbers.* The first two solutions are $(1, 1, 1)$ and $(1, 1, 2)$. These two are called *singular solutions.*

Claim 2. The singular solutions are the only solutions where m, m_1, m_2 are not distinct.

Proof:
Let (m, m_1, m_2) be a solution and assume $m_1 = m_2$.
Then we have

$$m^2 + 2m_1^2 = 3m{m_1}^2 \Rightarrow m^2 = (3m - 2)m_1^2 \Rightarrow m_1|m.$$

There is a $k \in \mathbb{N}$ such that $m = km_1$.
We have:

$$k^2m_1^2 + m_1^2 + m_1^2 = 3km_1m_1m_1 \Rightarrow (k^2 + 2) = 3km_1.$$

Thus, $k = 1$ or $k = 2$, which leads us to conclude that we started with one of the singular solutions.

From a given solution (m, m_1, m_2) we can obtain three other solutions, which we will call it's neighbours, namely:

$$(m', m_1, m_2),\ (m, m_1', m_2),\ (m, m_1, m_2')$$

defined by:

$$m' = 3m_1m_2 - m,\ m_1' = 3mm_2 - m_1,\ m_2' = 3mm_1 - m_2.$$

Indeed, if $m^2 + m_1^2 + m_2^2 = 3mm_1m_2$, then:

$$m'^2 + m_1^2 + m_2^2 = 9m_1^2m_2^2 - 6mm_1m_2 + m^2 + m_1^2 + m_2^2$$

$$= 9m_1^2m_2^2 - 3mm_1m_2 = 3m_1m_2(3m_1m_2 - m) = 3m_1m_2m'$$

and similar for m_1' and m_2'.

Theorem 6.0.18. *1. All of the positive integer solutions (m, m_1, m_2) of the Markov equation are obtained by succesively taking neighbours, starting with $(1, 1, 1)$.*

2. The three integers in any solution are pairwise relatively prime.

Proof:
(i): Consider the quadratic polynomial f given by:

$$f(x) = x^2 - 3m_1m_2x + m_1^2 + m_2^2.$$

If $m \in \mathbb{Z}_{>0}$ is a root of f (i.e. (m, m_1, m_2) a solution of the Markov equation), then so is $3m_1m_2 - m$, and these are the only two roots. If (m, m_1, m_2) is a nonsingular solution, with say $m_1 > m_2$, then:

$$f(m_1) = 2m_1^2 - 3m_2m_1^2 + m_2^2 < 3m_2^2 - 3m_2m_1^2 < 0.$$

Also, $f(m_1) = (m_1 - m)(m_1 - m')$. So we have $(m_1 - m)(m_1 - m') < 0$.
Now we drop the assumption $m_1 > m_2$.
Suppose that m is the largest of m, m_1, m_2. Then $\max(m_1, m_2) - m < 0$ so $\max(m_1, m_2) > m'$, i.e. $m' < \max(m_1, m_2) < m$. We also have:

$$m_1' = 3mm_2 - m_1 > 3m - m > m$$

$$m_2' = 3mm_1 - m_2 > 3m - m > m.$$

We can conclude that for any solution of the Markov equation, there is one neighbour with a smaller maximal element and there are two neighbours

with a larger maximal element. Given any nonsingular solution, we can walk backwards from this solution through succesive neighbours with smaller maximal elements, a process which must terminate in a singular solution.
We observe that $(1,1,1)$ has only one neighbour, which in it's turn has only two neighbours; $(1,1,1)$ and $(1,2,5)$. Thus, the solutions can be arranged in a tree (fig. 1).
(ii): Suppose that (m, m_1, m_2) is a solution where two of the integers have a common divisor d. Looking at the Markov equation, we see that d must also be a divisor of the third integer. Therefore, d will also be a divisor of the neighbors. Walking back in the tree we conclude that d must equal 1.

Quadratic Forms

Quadratic forms are homogeneous polynomials of degree two.
We will consider real indefinite quadratic forms:

$$f(x,y) = ax^2 + bxy + cy^2$$

where the discriminant $d(f) = b^2 - 4ac$ is positive. Such an f has two roots.
We define, for two quadratic forms f and g; $f \sim g$ if and only if we can obtain f out of g by applying a linear transformation $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ to $\begin{pmatrix} x \\ y \end{pmatrix}$ with integer coëfficients and determinant $\alpha\delta - \beta\gamma = \pm 1$.
If $f \sim g$ then $f(x,y) = g(\alpha x + \beta y, \gamma x + \delta y)$ and $d(f) = f(g)$.
One may easlily check that $\sim$ is an equivalence relation. We define:

$$\mu(f) = \inf_{(x,y)\in\mathbb{Z}\setminus\{(0,0)\}} |f(x,y)|$$

If $f \sim g$ then $\mu(f) = \mu(g)$. For all $\lambda \neq 0$, we have $\mu(\lambda f) = \lambda\mu(f)$ and $d(\lambda f) = \lambda^2 d(f)$. Therefore, the value of $\frac{\mu(f)}{\sqrt{d(f)}}$ is the same for all equivalent forms and non-zero multiples of them.
The *Markov spectrum* is the set of all possible values for $\frac{\mu(f)}{\sqrt{d(f)}}$.
An *automorph* of a quadratic form is a linear transformation with determinant $+1$ which leaves the form unaltered.
A quadratic form $f(x,y) = ax^2 + bxy + cy^2$ is *primitive* if it's coëfficients have no common divisor, it is *reduced* if it's roots ξ and η satisfy $1 > \xi > 0$ and $\eta < -1$.

Theorem 6.0.19. *If a quadratic form $f(x,y) = ax^2 + bxy + cy^2$ has integer coëfficients and is primitive, then the following are equivalent:*

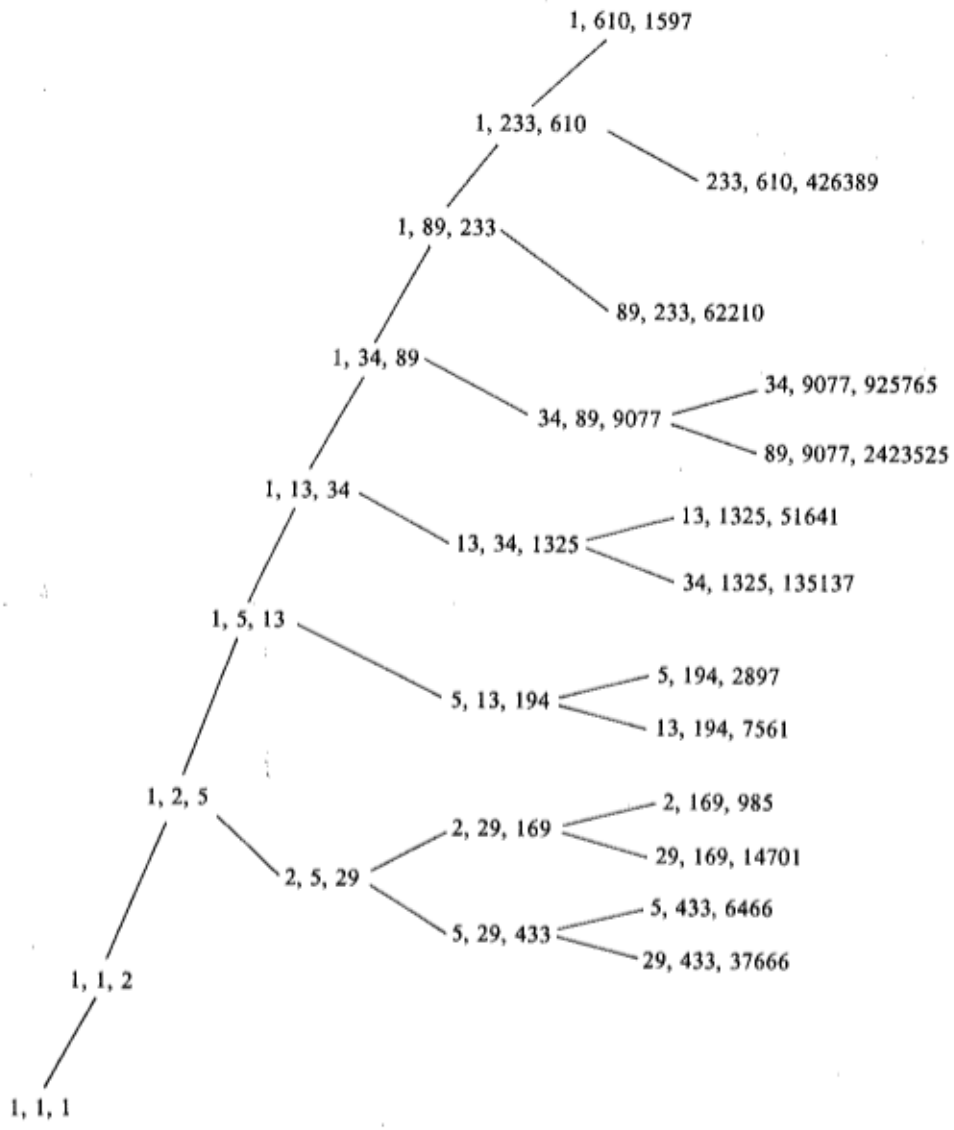


Figure 6.1: A part of the Markov tree

1. $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ *is an automorph of* f.

2. $\alpha = \frac{r-bs}{2}, \quad \beta = -cs, \quad \gamma = as \quad \delta = \frac{r+bs}{2}$
where r, s *is an integer solution of* $r^2 - d(f)s^2 = 4$.

Proof:
Both directions follow immediately from the given definitions.

Theorem 6.0.20. *Let* $f(x,y) = ax^2 + bxy + cy^2$ *be a quadratic form with integer coëfficients which is primitive and reduced. Let the continued fraction expansion of root* ξ*, which is purely periodic by Galois, be given by* $\xi = [0; \overline{a_1, a_2, \cdots, a_n}]$,
where n *is taken to be even and* $a_1, a_2, \cdots, a_n$ *is the shortest period. Let* $\frac{p_1}{q_1}, \frac{p_2}{q_2}, \cdots$ *be the sequence of convergents of the continued fraction expansion of* ξ.

If we define $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} q_{n-1} & q_n \\ p_{n-1} & p_n \end{pmatrix}$,
we have determinant: $p_n q_{n-1} - q_n p_{n-1} = (-1)^n = 1$.
Furthermore this is an automorph of f*, which satisfies:*
$(*)$ $\alpha = \frac{r-bs}{2}, \quad \beta = -cs, \quad \gamma = as \quad \delta = \frac{r+bs}{2}$
where r, s *is the* <u>least</u> *positive integer solution of* $r^2 - d(f)s^2 = 4$.

(sketch of) **Proof:**
$\xi = \left[0; a_1, \cdots, a_n, \xi^{-1}\right]$, so $\xi^{-1} = \frac{p_n \xi^{-1} + p_{n-1}}{q_n \xi^{-1} + q_{n-1}} = \frac{p_n + \xi p_{n-1}}{q_n + \xi q_{n-1}}$.
This can be rewritten as $\gamma\xi^2 + (\delta - \alpha)\xi - \beta = 0$, so ξ is also a root of $g(x,y) = \gamma x^2 + (\delta - \alpha)xy - \beta y^2$ and therefore $g(x,y)$ has integer coëfficients which are multiples of the coëfficients of f, say $\gamma = as,\ \delta - \alpha = bs,\ \beta = -cs$. Because f is primitive, s must be an integer. If we define $r = \alpha + \delta$, one may show that the equations in $(*)$ hold, and conclude that we have an automorph of f. Since this integer solution r, s is associated with the shortest period of ξ, the theory of Pell equations tells us that it is indeed the least positive integer solution.

Markov Forms

Let (m, m_1, m_2) be a nonsingular solution with m the largest of the integers m, m_1, m_2. From Theorem 1 we know that m and m_1 are relative prime. By Euclid's algorithm we can find integers w, u such that $wm + um_1 = m_2$ and therefore $um_1 \equiv_m m_2$ with $u < m$. From the Markov equation we see

that $m_1^2 + m_2^2 \equiv_m 0$ from which we can show that $um_2 \equiv_m -m_1$. Observe that:

$$m_1^2(u^2+1) = m_1^2u^2 + m_1^2 \equiv_m m_2^2 + m_1^2 \equiv_m 0$$

So we can find $v < m$ such that $mv = u^2 + 1$.
We consider the quadratic form:

$$f(x,y) = mx^2 + (3m-2u)xy + (v-3u)y^2.$$

If u' is another integer such that $0 < u' < m$ and $u' \equiv_m 0$,
then $(u+u')m_2 \equiv_m m_1 + -m_1 \equiv_m 0$, hence u' must be equal to $m-u$. Considering the quadratic form f' which we obtain by using this u' instead of u, we see that

$$f'(x,y) = mx^2 + (3m-2u')xy + (v'-3u')y^2 = f(x+2y,-y)$$

so $f \sim f'$. Thus, we can define the quadratic form associated to a solution (m, m_1, m_2) as above. Also, this form only depends (up to equivalence) upon the largest element m and we can assume $0 \leq 2u \leq m$.

The *Markov Form* associated to a solution (m, m_1, m_2), is defined by:

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

Where u and v satisfy $m_1u \equiv_m m_2$, $0 \leq 2u \leq m$ and $mv = u^2 + 1$.

Theorem 6.0.21. *Let (m, m_1, m_2) be a solution and $f_m(x,y)$ it's Markov form.*

1. *The discriminant $d(f_m)$ is $9m^2 - 4$.*
2. *The form $f_m(x,y)$ is properly equivalent to $-f_m(x,y)$.*
3. *$\mu(f_m) = f_m(1,0) = m$.*
4. *If α and β denote the roots of $f_m(x,1)$ with $\alpha > \beta$, then $1 > \alpha > 0$, $\beta < -1$ and α has a continued fraction expansion of the form $[0; \overline{a_1, \cdots, a_k}]$.*
5. *If m is odd, then f_m is primitive.*
   If m is even, then $\frac{1}{2}f_m$ has integer coëfficients and is primitive.

Proof of *(iv)*:
$f_m(x,y) = mx^2 + (3m-2u)x + v - 3u$, so the roots are:

$$\alpha = \frac{-3m+2u+\sqrt{9m^2-4}}{2m} \quad \text{and} \quad \beta = \frac{-3m+2u-\sqrt{9m^2-4}}{2m}$$

We have:

$$0 < \sqrt{9m^2-4} - 3m + 2u < 2u < 2m < \sqrt{9m^2-4} + 3m - 2u$$

The first inequality: $(\sqrt{9m^2-4}+2u)^2 = 9m^2 - 4 + 4u^2 + \cdots > (3m)^2$.
The second inequality: $\sqrt{9m^2-4} < \sqrt{9m^2} = 3m$.
The third inequality: definition of f_m.
The fourth inequality: $m > 2u$.
So we have: $1 > \alpha > 0$ and $\beta < -1$.
$-\beta$ is a so called a reduced irrational and therefore has a purely periodic continued fraction expansion; $-\beta = [\overline{a_0; a_1, \cdots, a_k}]$.
From this we see that $\frac{-1}{-\alpha} = [\overline{a_k; a_{k-1}, \cdots, a_0}]$.
$\alpha = 0 + \frac{1}{\frac{1}{\alpha}}$, so $\alpha = [0; \overline{a_0, \cdots, a_k}]$.

Continuants

For a finite sequence of integers $a_1, a_2, \cdots, a_n$, the continuant $K(a_1, a_2, \cdots, a_n)$ is defined to be the denominator of the continued fraction $[0; a_1, a_2, \cdots, a_n]$.
So we have, $K(a_1) = a_1$, because $[0; a_1] = 0 + \frac{1}{a_1} = \frac{1}{a_1}$.
$K(a_1, a_2) = a_1 a_2 + 1$, for $[0, a_1, a_2] = 0 + \frac{1}{1+\frac{1}{a_2}} = \frac{a_2}{a_1 a_2 + 1}$.
$K(a_1, \cdots, a_n) = q_n = a_n K(a_1, \cdots, a_{n-1}) + K(a_1, \cdots, a_{n-2})$, for $n \geq 3$.
In some texts, these continuants are also called Euler polynomials. This is because we can obtain $K(a_1, \cdots, a_n)$ by taking the sum of all possible products of $a_1 \cdots, a_n$ in which any number of consecutive terms are deleted. From this we can conclude that reversing the order of a finite sequence does not change the value of its continuant. That is, $K(a_1, \cdots, a_n) = K(a_n \cdots, a_1)$.
So we also have the recursion relation:

$$K(a_1 \cdots, a_n) = a_1 K(a_2, \cdots, a_n) + K(a_3, \cdots, a_n) \quad \text{for } n \geq 3$$

which can be extended to the more general formula

$$K(a_1, \cdots, a_n) = K(a_1, \cdots, a_m) K(a_{m+1}, \cdots, a_n) +$$

$$K(a_1, \cdots, a_{m-1}) K(a_{m+2}, \cdots, a_n) \quad \text{for } 1 \leq m < n.$$

Lemma 6.0.22. *Suppose that the positive integers m, u, v satisfy*
$m \geq 2,\ m > v$ and $mv - u^2 = 1$.
If we expand $\frac{m}{u}$, and $\frac{m}{m-u}$ as continued fractions with an even number of partial quotients, then these continued fractions are symmetric. Moreover, there exist unique positive integers $a_1, \cdots a_n$ such that the following continuant formulas hold:

$$u = K(a_1, \cdots, a_n, a_n, \cdots, a_2) = K(a_2, \cdots, a_n, a_n, \cdots, a_1)$$

$$m = K(a_1, \cdots, a_n, a_n, \cdots, a_1)$$

Proof:
From my earlier lecture *Sums of squares* we know that the continued fraction expansion of $\frac{m}{u}$ is symmetric.
Let's say $\frac{m}{u} = [a_1; \cdots, a_n, a_n, \cdots, a_1]$.
Then by definition, $m = K(a_1, \cdots, a_n, a_n, \cdots, a_1)$.
u is the denominator of $[a_1; \cdots, a_n, a_n, \cdots, a_1]$, which is also the denominator of $[0; a_2, \cdots, a_n, a_n, \cdots, a_1]$. So u is the numerator of $[a_2; a_3, \cdots, a_1]$, which is $K(a_2, \cdots, a_2, a_1) = K(a_1, a_2, \cdots, a_2)$. To prove that $\frac{m}{m-u}$ also has a symmetric continued fraction expansion with an even number of partial quotients, we observe that if we replace v by $(m-2u+v)$ in the prerequisites and u by $m-u$, we get:

$$m > m - 2u + v$$

$$\text{and } m(m-2u+v) - (m-u)^2 =$$

$$m^2 - 2um + mv - m^2 + 2um - u^2 =$$

$$mv - u^2 = 1.$$

Theorem 6.0.23. *Let $m > 2, u,$ and v be the integers in the definition of a Markov form. Then the positive root α_m of $f_m(x, y)$ has a continued fraction expansion whose period has an even number of digits, say $\alpha_m = [0; \overline{a_1, \cdots, a_{2n}}]$.*
Furthermore,

$$m = K(a_1 \cdots, a_{2n-1}), \quad u = K(a_2, \cdots, a_{2n-2})$$

and we also have: $a_1 = a_{2n} = 2,\ a_{2n-2} = a_{2n-1} = 1$, and the sequence $a_2, \cdots a_{2n-3}$ is symmetric.

Proof:
We set $\alpha_m = [0; \overline{a_1, \cdots, a_j}]\,,\ \ d = 9m^2 - 4.$

If m is odd, then f_m is primitive and the least solution of $r^2 - ds^2 = 4$ is $r = 3m,\ s = 1$. If m is even then $\frac{1}{2}f_m(x,y)$ is primitive and the least solution of $r^2 - \frac{1}{4}ds^2 = 4$ is $r = 3m,\ s = 2$.
In both cases we use theorem 3 to get an automorphism of f with:

$$\alpha = \frac{3m - (3m - 2u)}{2} = u \qquad \beta = -(v - 3u) \cdot 1 = 3u - v$$

$$\gamma = ms \qquad \delta = \frac{3m + (3m - 2u)}{2} = 3m - u$$

and also: $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} q_{j-1} & q_j \\ p_{j-1} & p_j \end{pmatrix}$, with $\frac{p_i}{q_i}$ the partial convergents of the continued fraction expansion of $\frac{1}{\alpha_m}$. So we have $(*)$:

$$u = q_{j-1} = K(a_2, \cdots, a_{j-1}) \qquad 3u - v = q_j = K(a_2, \cdots, a_j)$$

$$m = p_{j-1} = K(a_1, \cdots, a_{j-1}) \qquad 3m - u = p_j = K(a_1, \cdots, a_j).$$

Observe that $f_m(0,1) = 3u - v$, so with theorem 4 part *(iii)* we see that $m < 3u$. We already knew that $2u < m$, so now we have $2u < m < 3u$. This fact together with $mv - u^2 = 1$ gives us $2v < u < 3v$.
From $(*)$ we see that

$$\frac{m}{u} = [a_1, \cdots, a_{j-1}],\ \text{so } a_1 = 2$$

$$\frac{3m - u}{m} = \frac{K(a_2, \cdots, a_j)}{K(a_1, \cdots, a_{j-1})} = a_j + \frac{K(a_1, \cdots, a_{j-2})}{K(a_1, \cdots, a_j)} = \cdots = [a_j; a_{j-1}, \cdots, a_1]$$

Together with $\frac{3m-u}{m} = 3 - \frac{u}{m}$, this gives us $a_j = 2$.

$$(**) \qquad \frac{3m - u}{m} = [2; a_{j-1}, \cdots, a_1] = 2 + \frac{1}{[a_{j-1}, \cdots, a_1]}$$

From $(*)$ we have $\frac{3u-v}{u} = [a_j; a_{j-1}, \cdots, a_2] = [2; a_{j-1}, \cdots, a_2]$, hence

$$\frac{3u - v - 2u}{u} = \frac{1}{[a_{j-1}, \cdots, a_2]} \quad \text{so} \quad \frac{u}{u - v} = [a_{j-1} \cdots, a_2].$$

If j were odd, then by lemma 1 and $(**)$ we would have a symmetric continued fraction $[a_{j-1}, \cdots, a_1]$ which would imply $a_{j-1} = 2$. However, $\frac{u}{u-v} = [a_{j-2}, \cdots, a_1]$ leads to contradiction because $u > 2v$. So j must be even, lets say $j = 2n$. From $(**)$ and the fact that $a_1 = 2$ we see that we can write

$$\frac{m}{m - u} = [a_{2n-1}, a_{2n-2}, \cdots, a_2, 1, 1]$$

which has an even number of terms. Applying lemma 1, we conclude that $a_{2n-2} = a_{2n-1} = 1$ and $a_2, \cdots, a_{2n-3}$ is symmetric.

Chapter 7

The nearest integer continued fraction

Willem van Loon

If we calculate the continued fraction expansion of π, we get:

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cdots}}}$$

In the third step, we arrive at 1/0,06...=15,996..., so normally we would take 15 for our expansion. But this fraction lies much closer to 16, so maybe it would make more sense if we choose 16 for our expansion, instead of 15. If we do this consequently, that is taking the nearest integer of the remaining fraction, we get the nearest integer continued fraction expansion.

Definition 7.0.24. The Nearest Integer Continued Fraction (NICF) operator $T_{1/2} : [-\frac{1}{2}, \frac{1}{2}) \to [-\frac{1}{2}, \frac{1}{2})$ is defined by $T_{1/2}(x) := \left|\frac{1}{x}\right| - \left\lfloor \left|\frac{1}{x}\right| + \frac{1}{2} \right\rfloor$, $x \neq 0$, with $x = b_0 + \cfrac{\epsilon_1}{b_1 + \cfrac{\epsilon_2}{b_2 + \cdots}}$, and $\epsilon_n = \operatorname{sgn}\left(T_{1/2}^{n-1}(x - b_0)\right)$

b_0 is hereby chosen such that $x - b_0 \in [-\frac{1}{2}, \frac{1}{2})$.

Lemma 7.0.25. *$[b_0; \epsilon_1 b_1, \epsilon_2 b_2, \cdots]$ is the NICF-expansion of a number $a \in \mathbb{R}$ $\iff$ $b_n \geq 2$ and $b_n + \epsilon_{n+1} \geq 2$ for all $n \geq 1$*

PROOF $-\frac{1}{2} \leq T_{1/2}^{n-1} < \frac{1}{2}$, so $b_n = \left\lfloor \left| \frac{1}{T_{1/2}^{n-1}(x-b_0)} \right| + \frac{1}{2} \right\rfloor \geq \left\lfloor 2 + \frac{1}{2} \right\rfloor = 2$. Furthermore, if $\epsilon_{n+1} = -1$, then

$$T_{1/2}^{n}(x-b_0) = \left| \frac{1}{T_{1/2}^{n-1}(x-b_0)} \right| - \left\lfloor \left| \frac{1}{T_{1/2}^{n-1}(x-b_0)} \right| + \frac{1}{2} \right\rfloor < 0$$

Since $\left| \frac{1}{T_{1/2}^{n-1}(x-b_0)} \right| \geq 2$, it follows that $b_n \geq 3$, so $b_n + \epsilon_{n+1} \geq 2$. The other direction is easy to prove.

Definition 7.0.26. Let x be an irrational number and let $[b_0; \epsilon_1 b_1, \epsilon_2 b_2, \cdots]$ be the continued fraction expansion of x. Suppose $b_{l+1} = 1$ and $\epsilon_{l+1} = \epsilon_{l+2} = 1$. Then the transformation

$$\begin{gathered} [b_0; \epsilon_1 b_1, \cdots, \epsilon_l b_l, \epsilon_{l+1} b_{l+1}, \epsilon_{l+2} b_{l+2}, \epsilon_{l+3} b_{l+3}, \cdots] \\ \Downarrow \\ [b_0; \epsilon_1 b_1, \cdots, \epsilon_l (b_l + 1), -(b_{l+2} + 1), \epsilon_{l+3} b_{l+3}, \cdots] \end{gathered}$$

is called the singularization of b_{l+1}.

For example, if we singularize the expansion of π, we get

$$\begin{gathered} \pi = [3; 7, 15, 1, 292, 1, \cdots] \\ \Downarrow \\ \pi = [3; 7, 16, -293, 1, \cdots] \end{gathered}$$

We haven't shown yet that the new expansion is the same number as the old one. To prove this, define $\frac{r_k}{s_k} = [b_0; \epsilon_1 b_1, \cdots, \epsilon_k b_k]$, the n-th convergent of x before the singularization. In the same way, $\frac{c_k}{d_k}$ is defined to be the n-th convergent of x after singularizing. We now have to prove that $\frac{r_{l+1}}{s_{l+1}} = \frac{c_l}{d_l}$ and $\frac{r_{l+2}}{s_{l+2}} = \frac{c_{l+1}}{d_{l+1}}$. Then the rest would follow, because from here on the expansion stays the same. If you look at the definition of singularizing, it is very easy to see that $\frac{r_{l+1}}{s_{l+1}} = \frac{c_l}{d_l}$, because we have

$$b_{l-1} + \cfrac{\epsilon_l}{b_l + \cfrac{1}{1}} = b_{l-1} + \frac{\epsilon_l}{b_l + 1}$$

For the proof of $\frac{r_{l+2}}{s_{l+2}} = \frac{c_{l+1}}{d_{l+1}}$, we have to show that

$$\frac{\epsilon_l}{b_l + \dfrac{1}{1 + \dfrac{1}{b_{l+2}}}} = \frac{\epsilon_l}{b_l + 1 + \dfrac{-1}{b_{l+2} + 1}},$$

which is easy to see because $\dfrac{1}{1 + \dfrac{1}{b_{l+2}}} = \frac{b_{l+2}}{b_{l+2}+1} = 1 - \frac{1}{b_{l+2}+1}$

Lemma 7.0.27. *Let $x \in [0, 1)$ be some irrational number (without loss of generality), so $x = [0; a_1, a_2, \cdots]$. If you singularize in each block of m consecutive 'ones' in the RCF of x, the first, the third, the fifth,... partial quotiënt, you get the NICF-expansion of x (which is $[b_0; \epsilon_1 b_1, \epsilon_2 b_2, \cdots]$).*

Proof By Lemma 0.0.2 it's sufficiënt to proof $b_n \geq 2$ and $b_n + \epsilon_{n_1} \geq 2$ for all $n \geq 1$, and since it follows directly from the definition of singularization that $b_n \geq 2$ for all $n \geq 1$, we only have to show that $b_n + \epsilon_{n_1} \geq 2$ for all $n \geq 1$.

Suppose $\epsilon_{n+1} = -1$. Then you have singularized a_{n+1}. There are two cases:

- a_{n+1} is the first 1 of a block of ones. Then $a_n \geq 2$, so $a_n + 1 = b_n \geq 3$.
- a_{n+1} is not the first 1 of a block of ones, so $a_{n-1} = a_n = 1$, and thus you have singularized a_{n-1}. Therefore a_n has changed to $-(a_n + 1)$. After that a_{n+1} has been singularized, which changed $-(a_n + 1)$ to $-(a_n + 2) = -b_n$. Because $a_n \geq 1$, it follows that $b_n \geq 3$.

We're now going to introduce two new operators, the 'future'-operator T and the 'past'-operator V.

$$\begin{aligned} T_n &= T^n(x) = [0; a_{n+1}, a_{n+2}, \cdots], & T_0 &= x \\ V_n &= V^n(x) = [0; a_n, a_{n-1}, \cdots, a_1], & V_0 &= 0 \end{aligned}$$

It will be of later importance to note that $V_n = [0; a_n, a_{n-1}, \cdots, a_1] = \frac{q_{n-1}}{q_n}$, with q_n the denominator of the n-th convergent $\frac{p_n}{q_n}$. Furthermore

$$V_{n+1} = \frac{1}{a_{n+1} + V_n} = \frac{1}{a_{n+1} + \dfrac{q_{n-1}}{q_n}} = \frac{q_n}{a_{n+1} q_n + q_{n-1}}$$

and so

$$\frac{q_n}{q_{n+1}} = \frac{q_n}{a_{n+1}q_n + q_{n-1}}$$

Lemma 7.0.28. *Let $x \in [0,1)$ be an irrational number, so x has RCF-expansion $[0; a_1, a_2, \cdots]$. Let $S_{1/2} := [\frac{1}{2}, g) \times [0, g] \cup [g, 1) \times [0, g)$. If you singularize $a_{n+1} = 1$ if and only if $(T_n, V_n) \in S_{1/2}$, then you get the NICF-expansion of x.*

PROOF We want to prove that singularizing the first, third, fifth,... one of each block of ones is the same as singularizing if $(T_n, V_n) \in S_{1/2}$. We assume $a_n \neq 1$, so $a_{n+1} = 1$ is the first one of a block. Then $V_n = [0; a_n, \cdots, a_1] = \frac{1}{a_n + \cdots} < \frac{1}{2}$, and $T_n = [0; a_{n+1}, a_{n+2}, \cdots] = \frac{1}{1 + \cdots} > \frac{1}{2}$, so $(T_n, V_n) \in S_{1/2}$. So we always singularize the first 'one'.

There are now two cases to seperate:

- $a_{n+1} = a_{n+2} = 1$ and $(T_n, V_n) \in S_{1/2}$. In this case $\frac{1}{2} < T_{n+1} < 1$. Furthermore, we have $0 \leq V_n < g \Longrightarrow 0 \leq \frac{q_n}{q_{n+1}} < g \Longrightarrow q_{n-1} < g q_n$. Therefore:

  $$V_{n+1} = \frac{q_n}{q_{n+1}} = \frac{q_n}{a_{n+1}q_n + q_{n-1}} > \frac{q_n}{q_n + g q_n} = \frac{1}{1+g} = g$$

  So $(T_{n+1}, V_{n+1}) \notin S_{1/2}$, and thus we don't singularize the next 'one'.

- $a_{n+1} = a_{n+2} = 1$ and $(T_n, V_n) \notin S_{1/2}$. Again, $\frac{1}{2} < T_{n+1} < 1$. Furthermore, we have $g < V_n \leq 1 \Longrightarrow g < \frac{q_{n-1}}{q_n} \leq 1 \Longrightarrow g q_n < q_{n-1}$. Therefore:

  $$V_{n+1} = \frac{q_n}{q_{n+1}} = \frac{q_n}{a_{n+1}q_n + q_{n-1}} < \frac{q_n}{q_n + g q_n} = \frac{1}{1+g} = g$$

  So $(T_{n+1}, V_{n+1}) \in S_{1/2}$, and so now we do singularize the next 'one'.

Chapter 8

Continued fractions and the LLL algorithm

Geert Popma

Introduction

In 1982 Arjen Lenstra, Hendrik Lenstra and László Lovász published their article on the LLL algorithm. It is a polynomial time algorithm which reduces a basis for an integer lattice. The original application was to give a polynomial time algorithm for factorizing polynomials with rational coefficients into irreducible polynomials[52]. The algorithm can be used for finding simultaneous rational approximations to real numbers and for solving the integer linear programing problem in fixed dimensions.

We first give the basic definition of a lattice, next we give examples of different bases for a lattice. For practical reasons we want to have a *reduced* basis. The celebrated *LLL*-algorithm describes, given a basis, a way to produce a reduced basis. Next in (2) we give a geometrical interpretation of the continued fraction expansion in the language of lattices. This interpetation paves the way to *NICF*, the nearest integer continued fraction. We establish in (3) a connection between *LLL* and *NICF*.

8.1 Lattices and bases

Definition 1 (Lattice)**.** *A lattice is a discrete subset of* $\mathbb{R}^n$ *spanning the enitre vector space.*
Let $(b_1, b_2, \ldots, b_n)$ *be a basis for* $\mathbb{R}^n$*, then* $\mathbb{Z}b_1 + \mathbb{Z}b_2 + \ldots + \mathbb{Z}b_n$ *is a* lattice

in $\mathbb{R}^n$.

Just as for a vector space, the basis of a lattice is not unique. In figure 8.1 we provide two bases for the same lattice.

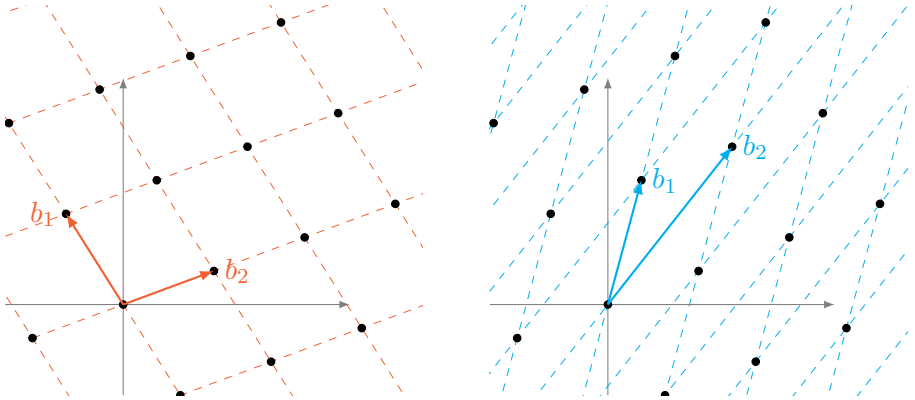


Figure 8.1: The same lattice spanned by different bases.

Since $\mathbb{R}^n$ is a finite-dimensional inner product space, we know it has an orthonormal basis. An arbitrary set of vectors can be turned into an orthonormal basis using the Gram-Schmidt orthogonalization process. However, this does not work for lattices!

In this report we're only concerned with lattices in $\mathbb{R}^2$. As we said above we in general cannot find an orthogonal basis. Neither can we find a 'normal' basis, i.e. a basis with vectors of unit length. If we apply the Gram-Schmidt process we have to be careful, we can only work with integer multiples of the original basis vectors.

The Gram-Schmidt process takes a finite, linearly independent set $S = \{b_1, \ldots, b_k\}$ for $k \leq n$ and generates an orthogonal set $S = \{b_1^*, \ldots, b_k^*\}$ that spans the same k-dimensional subspace of $\mathbb{R}^n$ as S. For $n = 2$ the process comes down to:

$$b_1^* = b_1, \ b_2^* = b_2 - \mu \cdot b_1^*$$

Where $\mu = \frac{\langle b_2, b_1^* \rangle}{\langle b_1^*, b_1^* \rangle}$.

The problem with this process when applied to lattices is that μ has to be an integer, otherwise the *new* basis is no longer a basis. By choosing a suitable integer we can use this process to get a *more* orthogonal basis. We attempt to apply *G-S* to the second basis in figure 8.1.

The new basis (b_1, b_2^*) appears to be a "better" basis for the lattice. To

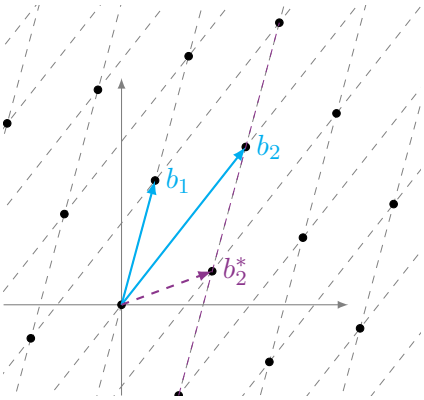


Figure 8.2: *Orthogonalizing* the basis

The *G-S* proces suggests we shift b_2 by a multiple of b_1. These shifts are on the dashed purple line. Our choice of b_2^* should be a point on the line which is also a element of the lattice. Since

$$\mu = \frac{\langle b_2, b_1 \rangle}{\langle b_1, b_1 \rangle} \approx -1.44$$

the best choice for b_2^* according to *G-S* becomes

$$b_2^* = b_2 - b_1.$$

make precise what constitutes a better basis we introduce the notion of a *reduced basis*.

Definition 2 (Reduced basis)**.** *A basis* $(b_1, \ldots, b_n)$ *for a lattice is called* reduced *if*

(i) $|\mu_{ji}| \leq \frac{1}{2}$ *for* $1 \leq i < j \leq n$

(ii) $|b_j^* - \mu_{j\cdot j-1} b_{j-1}^*|^2 \geq \frac{3}{4}|b_{j-1}|^2$ *for* $1 < j \leq n$

Where $\mu_{ji} = \frac{\langle b_j, b_i^* \rangle}{\langle b_i^*, b_i^* \rangle}$ *and* $b_j^* = b_j - \sum_{i=1}^{j-1} \mu_{ji} b_i^*$.

For a two-dimensional lattice we write μ instead of μ_{21}. Being reduced now amounts to

$$|\mu| \leq \frac{1}{2}\ ,\ |b_2| \geq \frac{3}{4}|b_1|.$$

By the first condition a reduced basis has more orthogonal vectors. The basis vectors always span a parallelepiped of the same volume, hence more orthogonal vectors are also shorter. For instance for a parallelogram spanned by vectors of lengths b_1 and b_2 with an angle θ in between the area is $b_1 b_2 \sin\theta$. If the vecotrs become more orthogonal then $\sin\theta$ grows towards 1. The area remains the same thus $b_1 b_2$ grows smaller. This explains why reducing a basis provides short basis vectors. *LLL* does not garantee you get the shortest vector in a lattice, however we obtain the following, which provides short enough vectors in practice.

Propostition 3. *A reduced basis satisfies*

(i) $\det(L) \leq \prod |b_i| \leq 2^{n(n-1)/4} \det(L)$

(ii) for all $x \in L$ we have $|b_i| \leq 2^{(n-1)/2}|x|$

Looking back at the figure 8.1 we get that the first basis is reduced whereas the second one is not. Our "orthogonalized" basis in figure 8.2 *is* reduced, provided we swap b_1 and b_2^* in order to satisfy the second condition. This can be readily checked since the bases are:

$$\begin{pmatrix} \text{-0.63} \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0.37 \end{pmatrix} \qquad \begin{pmatrix} 0.37 \\ 1.37 \end{pmatrix}, \begin{pmatrix} 1.37 \\ 1.74 \end{pmatrix} \qquad \begin{pmatrix} 1 \\ 0.37 \end{pmatrix}, \begin{pmatrix} 0.37 \\ 1.37 \end{pmatrix}$$

The process of *reducing* the basis in figure 8.2 consists of two steps in order to satisfy the reducedness conditions. First we shift b_2 by a multiple of b_1, second we swap b_1 and b_2 . This gives rise to the celebrated LLL-algorithm.

Algorithm (The LLL-algorithm for the case $n = 2$)**.** *Given a basis b_1, b_2 for a lattice do the following. We denote $\lfloor \mu \rceil$ for the nearest integer to μ.*

- *If $|\mu| > \frac{1}{2}$, replace b_2 by $b_2 - \lfloor \mu \rceil b_1$, this also replaces μ with $\mu - \lfloor \mu \rceil$.*
- *If $|b_2|^2 < \frac{3}{4}|b_1|^2$, then swap b_1 and b_2, this also adjusts μ.*

Stop when both conditions are satisfied.

The full algorithm is described in [52].

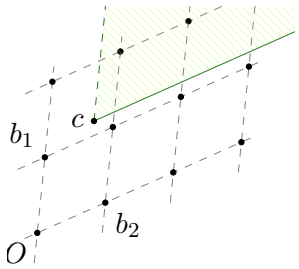
8.2 The geometry of continued fractions

Given a continued fraction $\alpha = [a_0, a_1, a_2, \ldots]$ define numbers p_n, q_n by the recursive relations

$$\begin{aligned} p_{n+2} &= a_{n+2}p_{n+1} + p_n \\ q_{n+2} &= a_{n+2}q_{n+1} + q_n \end{aligned}$$

The rationals $\frac{p_n}{q_n}$ are the *convergents* of α. As n grows larger the convergents become better approximations of α. In general we say that a rational $\frac{a}{b}$ is a *best approximation* of a real number α if for all $\frac{c}{d}$ such that $d \leq b$ we have $|b\alpha - a| \leq |d\alpha - c|$. Lagrange's theorem states that for a rational x which is not an integer, x is a best approximation of α if and only if x is a convergent of α.

The geometrical interpretation of a rational p/q being a best approximation to α is that, amongst all points $(q', p'$ in the integer lattice $\mathbb{Z}^2$ such that $0 \leq q' \leq q$, (q, p) is nearest to the line l given by $y = \alpha x$.



This idea can also be expressed in terms of a *cone*, (q, p) is nearest to $y = \alpha x$ amongst all points outside of the *halfcone with vertex* (q, p).

Definition 4 (Halfcone)**.** *The* (b_1, b_2)*-halfcone with vertex* c *consists of points* $c + t_1 b_1 + t_2 b_2$ *for all numbers* $t_1 \geq 0, t_2 > 0$.

Given a real number number α and lattice points b_1, b_2 we wish to construct the outpoint $\overline{b}$. Let l be the line $y = \alpha x$.

Assume that l meets parallelogram $Ob_1b_*b_2$ in the origin O, and in a point p that is the intersection point between l and the segment b_2b_*. Let q be the intersection point between l and the extension of b_1b_*. We then have $p = \theta b_1 + b_2$, for $0 < \theta < 1$, and it follows that $q = b_1 + \frac{1}{\theta} b_2$. Note that q lies on the lattice edge between the *outpoint* $\overline{b} = b_1 + \lfloor 1/\theta \rfloor b_2$ and $c = \overline{b} + b_2$.

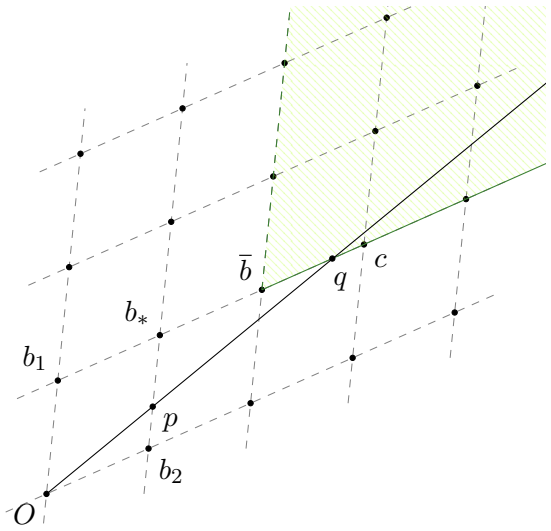


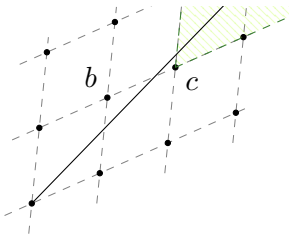
Figure 8.3: The outpoint construction

In the case that l meets $Ob_1b_*b_2$ in the segment b_1b_* we follow a similar construction where the roles of b_1, b_2 are swapped.

The outpoint $\overline{b}$ approximates l in the sense that it is nearer to l than any point outside of the (b_1, b_2)-halfcone with vertex $\overline{b}$. This brings to mind the geometrical interpretation of a best approximation.

We can iterate the outpoint construction. For $n \geq 3$ define b_n as the outpoint of b_{n-2} and b_{n-1}. In fig. 8.3 we get for instance $b_3 = \overline{b}$ and $b_4 = c$. When we start out with the basis for the integer latice, i.e. $b_{-2} = (0, 1)$, $b_{-1} = (1, 0)$, the outpoints become $b_n = (q_n, p_n)$. It turns out the best approximations in $\mathbb{Z}^2$ are precisely the points (q_n, b_n).[36]

The outpoint construction coincides with the continued fraction expansion. This geometrical interpretation also makes clear how the *nearest integer continued fraction*, abbreviated *NICF*, works. We can instead take as oupoint $\overline{b} = b_1 + \lfloor 1/\theta \rceil b_2$, in fig. 8.3 this yields c instead of b. Geometrically this is better since by adding a multiple of b_2 to b_1 the closest we can get to the line l is the point c. We can express this via an *open cone* with vertex c, which consist of all points $c + t_1 b_1 + t_2 b_2$, $t_1, t_2 > 0$. Then c is closest to the line of all points outsde of its open cone.



Compare the *NICF* outpoint construction to the *LLL* algorithm. Both shift a base element by a nearest integer multiple of the other base element. The *LLL* tries to produce an orthogonal basis, whereas the outpoints form a basis with vectors approaching the line $y = \alpha x$, i.e. the iterated outpoints become as least orthogonal as possible. We set out to establish a connection between these two iterations.

8.3 The relation between *LLL* and *NICF*

We set out to show that we can obtain the *NICF* as a special case of the LLL-algorithm.

Consider the lattice spanned by the vectors

$$b_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, b_2 = \begin{pmatrix} \alpha \\ \frac{\varepsilon}{\sqrt{2}} \end{pmatrix}$$

LLL yields p, q such that $|p - q\alpha| \leq \varepsilon$ and $1 \leq q \leq \sqrt{2}/\varepsilon$, see propostion 1.39 in [52]. The smaller we take ε the better our approximation of α by p/q gets. What happens if we set $\varepsilon = 0$? This would mean we apply *LLL* to $(1, 0)$ and $(\alpha, 0)$. Then this is simply 1-dimensional so we're not dealing

with inproducts. We get

$$b_1 = 1,\ b_2 = \alpha,\ \mu = \tfrac{b_2 b_1}{b_1 b_1} = \tfrac{b_2}{b_1}.$$

Write $\mu = q_0 + r_0$ where $q_0 \in \mathbb{Z}$ and $-\frac{1}{2} < r_0 \leq \frac{1}{2}$. Using the fact that $\frac{b_2}{b_1} = \mu = q_0 + r_0$ implies $b_2 = b_1 q_0 + b_1 r_0$, we get

$$b_2 - q_0 b_1 = b_1 r_0.$$

But this is exactly what the *LLL* algorithm tells us to replace b_2 with since $q_0 = \lfloor \mu \rceil$. Now

$$|b_2|^2 = |r_0 b_1|^2 \leq \tfrac{1}{4}|b_1|^2$$

so by *LLL* we have to swap b_1 and b_2. So our current situation becomes

$$b_1' = r_0 b_1,\ b_2' = b_1,\ \mu = \tfrac{1}{r_0}.$$

Next write $1/r_0 = q_1 + r_1$ where $q_1 \in \mathbb{Z}$ and $-\frac{1}{2} < r_1 \leq \frac{1}{2}$. In the same way as above we get

$$b_2' - q_1 b_1' = b_1' r_1$$

so we replace b_2' with $b_1' r_1 = r_0 r_1 b_1$. Now

$$|b_2'|^2 = |r_0 r_1 b_1|^2 \leq \tfrac{1}{4}|r_0 b_1|^2 \leq \tfrac{1}{4}|b_1'|^2$$

so we interchange the vectors. This brings us to the situation

$$b_1'' = r_0 r_1 b_1,\ b_2'' = r_0 b_1,\ \mu = \tfrac{1}{r_1}.$$

Repeating this we get

$$\begin{aligned} \alpha &= q_0 + r_0 \\ \tfrac{1}{r_0} &= q_1 + r_1 \\ \tfrac{1}{r_1} &= q_2 + r_2 \\ &\vdots \end{aligned}$$

or

$$\alpha = q_0 + \cfrac{1}{q_1 + \cfrac{1}{\ddots}}$$

where $q_i \in \mathbb{Z}$ and $-\frac{1}{2} < r_i \leq \frac{1}{2}$.

If we insert the vectors $(1, 0)$ and $(\alpha, 0)$ this yields the *NICF* expansion of α. This special case is however *not* an application of the algorithm as our input is not a basis of a lattice. Reasoning above we set $\varepsilon = 0$, depending on a choice of ε we find an n such that

$$\alpha \approx q_0 + \cfrac{1}{q_1 + \cfrac{1}{\ddots + \frac{1}{q_n}}}$$

Conclusions

We have seen the *NICF* expansion as an application of *LLL*, however it is not an application since we do not provide an appropiate input for the algorithm. This result is remarkable in the sense that *LLL* produces an orthogonal basis, whereas the geometrical interpretation of *NICF* is about producing a basis approaching the line $y = \alpha x$, the oppisote of orthogonal.

Chapter 9

Continued fractions and Ford circles

Geert Popma

Introduction

Recall the following notions from continued fractions.

Definition 5 (Convergents)**.** *Given a continued fraction* $\alpha = [a_0, a_1, a_2, \ldots]$ *define numbers* p_n, q_n *by the recursive relations*

$$p_{n+2} = a_{n+2}p_{n+1} + p_n$$

$$q_{n+2} = a_{n+2}q_{n+1} + q_n$$

The rationals $\frac{p_n}{q_n}$ *are the convergents of* α*.*

Definition 6 (Best approximation)**.** *Let* α *be a real number. A rational* $\frac{a}{b}$ *is a best approximation of* α *if for all* $\frac{c}{d}$ *such that* $d \leq b$*:* $|b\alpha - a| \leq |d\alpha - c|$*.*

We have seen the following result:

Theorem 7 (Lagrange)**.** *Let* x *be a rational which is not an integer. Then* x *is a best approximation of* α *if and only if* x *is a convergent of* α*.*

In this essay there will be an alternative proof of this theorem using Ford Circles. We follow a proof by Short (2011)[86].

Ford Circles

Definition 8 (Ford Circle)**.** *The Ford circle $\mathbb{C}[a,b]$ corresponding to a reduced fraction $\frac{a}{b}$ is a circle in $\mathbb{R}^2$ in the upper half-plane tangent to the x-axis. Its point of tangency to the x-axis is $\frac{a}{b}$ and its radius is $\frac{1}{2b^2}$.*

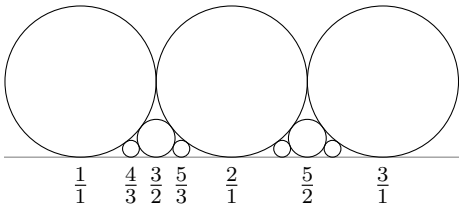


Figure 9.1: Some Ford circles

Propostition 9. *Different Ford circles $\mathbb{C}[a,b]$ and $\mathbb{C}[c,d]$ do not overlap eachother. They are tangent if and only if $|ad-bc|=1$.*

Proof. Denote d the distance between the centers of the circles, $(\frac{a}{b}, \frac{1}{2b^2})$ and $(\frac{c}{d}, \frac{1}{2d^2})$. By the pythagorean theorem we have $d^2 = \left(\frac{c}{d}-\frac{a}{b}\right)^2 + \left(\frac{1}{2d^2}-\frac{1}{2b^2}\right)^2$. Denote s the sum of the radii, $s = \frac{1}{2d^2}+\frac{1}{2b^2}$. A little calculation gives:

$$\begin{aligned} d^2 - s^2 &= \left(\frac{c}{d}-\frac{a}{b}\right)^2 + \left(\frac{1}{2d^2}-\frac{1}{2b^2}\right)^2 - \left(\frac{1}{2d^2}+\frac{1}{2b^2}\right)^2 \\ &= \left(\frac{c}{d}-\frac{a}{b}\right)^2 - 4\frac{1}{2d^2}\cdot\frac{1}{2b^2} \\ &= \frac{(ad-bc)^2-1}{b^2d^2} \end{aligned}$$

Here $(ad-bc)^2 \geq 1$ since they are different rationals. Hence $d^2-s^2 \geq 0$ and $d \geq s$ and the circles do not overlap. The circles are tangent if $d^2 = s^2$, this is the case if $|ad-bc|=1$. □

What are the tangent circles to a given circle $\mathbb{C}[a,b]$? We can find c,d such that $|ad-bc|=1$. Suppose we have another tangent circle $\mathbb{C}[c',d']$. We then have $|ad-bc| = 1 = |ad'-bc'|$. If $ad-bc$ and $ad'-bc'$ have the same sign, then $a(d-d')-b(c-c')=0$. So we get $a(d-d') = b(c-c')$, hence $a|c-c'$. Set $c-c' = n\cdot a$, or $c' = c-na$. Then $a(d-d') = b(na)$, so $d' = d-nb$. If $ad-bc$ and $ad'-bc'$ have opposite signs, then $a(d+d') = b(c+c')$. Set $c' = -c+na$, then $a(d+d') = b(na)$ hence $d' = -d+nb$. In both cases $\frac{c'}{d'} = \frac{c-na}{d-nb}$. We see that all the adjacents are of the form $\mathbb{C}[c-na, d-nb], n \in \mathbb{Z}$.

Definition 10 (Continued fraction chain)**.** *Let α be a real number, denote $\mathbb{C}_n := \mathbb{C}[p_n, q_n]$ for the Ford circle corresponding to the convergent $\frac{p_n}{q_n}$ of α. The continued fraction chain of α is the sequence of Ford circles $\mathbb{C}_0, \mathbb{C}_1, \mathbb{C}_2, \ldots$.*

We have the equality $|p_n q_{n-1} - p_{n-1} q_n| = 1$ so two consecutive circles in the chain are tangent. Since the numbers q_n are increasing the radii of the circles $\mathbb{C}_n$ are decreasing.
Given a rational $x = \frac{a}{b}$ and a real number α define $R_x(\alpha) = \frac{1}{2}|b\alpha - a|^2 = \frac{b^2}{2}|\alpha - x|^2$. From the previous proposition we see that $\mathbb{C}[a, b]$ is tangent to the circle at α with radius $R_x(\alpha)$.

Theorem 11. *Let α be a real number and $x = \frac{a}{b}$ a rational which isn't an integer.*

(i) *x is a convergent of α*

(ii) *$\mathbb{C}_x$ is a member of the continued fraction chain of α.*

(iii) *x is a best approximation of α*

(iv) *if z is a rational such that $\mathbb{C}_z$ has larger radius than $\mathbb{C}_x$, then $R_z(\alpha) \geq R_x(\alpha)$.*

Statement (ii) is merely a geometric reformulation of statement (i) just like (iv) is of (iii).

Propostition 12. *Let $x = \frac{a}{b}$ be a rational.*

(i) *if $|\alpha - x| < |\beta - x|$ then $R_x(\alpha) < R_x(\beta)$*

(ii) *if z is a rational distinct from x then $\mathrm{rad}(\mathbb{C}_z) \leq R_x(z)$ with equality if and only if $\mathbb{C}_x$ and $\mathbb{C}_z$ are tangent.*

Proof. (i) follows immediately from the definition.
For (ii) let $z = \frac{c}{d}$, then $\frac{1}{2d^2} \leq \frac{1}{2d^2}|ad - bc|^2 = \frac{b^2}{2}|\frac{a}{b} - \frac{c}{d}|^2 = R_x(z)$. They are equal iff $|bc - ad|^2 = 1$ which is equivalent to $\mathbb{C}_x$ and $\mathbb{C}_z$ being tangent. □

Propostition 13. *Let $\mathbb{C}_x$ and $\mathbb{C}_y$ be tangent Ford circles. If z is a rational that lies strictly in between x and y, then $\mathbb{C}_z$ has smaller radius than both $\mathbb{C}_x$ and $\mathbb{C}_y$.*

Proof. We have $|z - x| < |y - x|$ so $R_x(z) < R_x(y) = \mathrm{rad}(\mathbb{C}_y)$. Furthermore $\mathrm{rad}(\mathbb{C}_z) \leq R_x(z)$ so $\mathrm{rad}(\mathbb{C}_z) \leq \mathrm{rad}(\mathbb{C}_y)$. By a similar argument we see $\mathrm{rad}(\mathbb{C}_z) \leq \mathrm{rad}(\mathbb{C}_x)$. □

Propostition 14. *Let* $\mathbb{C}_x$ *and* $\mathbb{C}_y$ *be tangent Ford circles such that* $\text{rad}(\mathbb{C}_x) > \text{rad}(\mathbb{C}_y)$. *Suppose* α *lies strictly between* x *and* y *and a rational* z *lies strictly outside the interval bounded by* x *and* y. *Then* $R_x(\alpha) < R_z(\alpha)$.

Proof. We have $|x - \alpha| < |x - y|$ so $R_x(\alpha) < R_x(y) = \text{rad}(\mathbb{C}_y)$.
If y lies between z and α then $|z - y| < |z - \alpha|$ so $R_z(y) < R_z(\alpha)$. From $\text{rad}(\mathbb{C}_y) \leq R_z(y)$ we conclude $R_x(\alpha) < R_z(\alpha)$.
If x lies between z and α then $|z - x| < |z - \alpha|$ so $R_z(x) < R_z(\alpha)$. From $\text{rad}(\mathbb{C}_x) \leq R_z(x)$ we conclude $R_x(\alpha) < \text{rad}(\mathbb{C}_y) \leq \text{rad}(\mathbb{C}_x) < R_z(\alpha)$. □

Propostition 15. *Let* $\mathbb{C}_x$ *and* $\mathbb{C}_y$ *be tangent Ford circles such that* $\text{rad}(\mathbb{C}_x) > \text{rad}(\mathbb{C}_y)$. *Suppose* α *lies strictly between* x *and* y. *If* z *is a rational such that* $\text{rad}(\mathbb{C}_z) \geq \text{rad}(\mathbb{C}_x)$ *then* $R_x(\alpha) \leq R_z(\alpha)$, *with equality if and only if* $z = x$.

Proof. If $z \neq x$ then z lies outside the interval bounded by x and y, hence $R_x(\alpha) < R_z(\alpha)$. □

Proof of the theorem. $(i) \Rightarrow (iv)$ First the case if α irrational. Suppose $x = \frac{p_n}{q_n}$, define $y = \frac{p_{n+1}}{q_{n+1}}$. Then $\mathbb{C}_x$ and $\mathbb{C}_y$ are tangent and α lies strictly between x and y. Also $\text{rad}(\mathbb{C}_x) < \text{rad}(\mathbb{C}_y)$.
Suppose z is a ratonal such that $\text{rad}(\mathbb{C}_z) \geq \text{rad}(\mathbb{C}_x)$. Then by Prop 15 $R_x(\alpha) \leq R_z(\alpha)$.

Now for the rational case, suppose $\alpha = \frac{p_N}{q_N}$. Then for $n < N - 1$ we can apply the same argument to $x = \frac{p_n}{q_n}$ as above. If $x = \frac{p_N}{q_N}$, then $R_x(\alpha) = 0$ and statement (iv) is trivially true. If $x = \frac{p_{N-1}}{q_{N-1}}$ define $y = \frac{u}{v} = \frac{p_N - p_{N-1}}{q_N - q_{N-1}}$. I claim that y is a reduced fraction and $v > 0$, hence $\mathbb{C}_y$ exists and I further claim that $\mathbb{C}_x$ is tangent to $\mathbb{C}_y$ and $\text{rad}(\mathbb{C}_x) > \text{rad}(\mathbb{C}_y)$. Statement (iv) then follows by Prop 15 since α lies between x and y. Proof of the claim: if $k|gcd(u, v)$, then $k|\,|uq_{N-1} - vp_{N-1}| = 1$, hence $gcd(u, v) = 1$ and y is reduced; the q_n are increasing hence $v > 0$; the circles are tangent since $|uq_{N-1} - vp_{N-1}| = |p_N q_{N-1} - p_{N-1} q_N| = 1$; the radius is larger since $q_{N-1} < q_N - q_{N-1} = (a_N - 1)q_{N-1} + q_{N-2}$.
$(iv) \Rightarrow (i)$ by contradiction. Suppose x is not a convergent of α. Denote $r_n = \frac{1}{q_n^2}$ for the radius of $\mathbb{C}_n$. The r_n are strictly decreasing. If α is irrational then $r_n \to 0$, if α is rational then the sequence ends at $r_N = \text{rad}(\mathbb{C}_n)$. If α is rational we may assume $\text{rad}(\mathbb{C}_x) > \text{rad}(\mathbb{C}_\alpha) = r_N$, otherwise $R_x(\alpha) \leq R_\alpha(\alpha) = 0$ and then statement (iv) fails for $x \neq \alpha$ and $R_x(\alpha) = R_\alpha(\alpha)$ So there is a unique integer n such that $r_n \geq \text{rad}(\mathbb{C}_x) r_{n+1}$.

α lies strictly between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$, unless $\alpha = \frac{p_{n+1}}{q_{n+1}}$. We have $\text{rad}(\mathbb{C}_x) > \text{rad}(\mathbb{C}_{n+1})$ so x lies outside the interval bounded by $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$, hence

$R_{p_n/q_n}(\alpha) < R_x(\alpha)$. Now statement (iv) fails because for $z = \frac{p_n}{q_n}$ we have $\text{rad}(\mathbb{C}_z) \geq \text{rad}(\mathbb{C}_x)$ and $R_z(\alpha) < R_x(\alpha)$. If $\alpha = \frac{p_{n+1}}{q_{n+1}}$ then α lies strictly between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}-p_n}{q_{n+1}-q_n}$ so again statement (iv) fails. □

Hurwitz inequality

Ford circles were originally introduced to prove Hurwitz inequality. This concerns inequalities of the form $|\omega - y/x| < \alpha/x^2$. Given an irrational ω, for which $\alpha > 0$ is the above inequality satisfied for infinitely many fractions x/y. In 1891 Adolf Hurwitz found that $\alpha = 1/\sqrt{5}$ is the best value, i.e. the smallest. This can be derived using Ford circles.

Consider a vertical line $x = \gamma$, descending from the upper half-plane to the x-axis. If γ is rational then the line passes through the x-axis directly from a Ford circle. If γ is irrational then the line must leave every circle which it enters. The line then passes through infinitely many circles. To see this, recall given a circle $\mathbb{C}[a, b]$ and a circle $\mathbb{C}[c, d]$ tangent to it, all the adjacents of $\mathbb{C}[a, b]$ are of the form $\mathbb{C}[c-na, d-nb], n \in \mathbb{Z}$. As n grows large, positively or negatively, we see that the base point of the adjacents coverge to a/b. So any Ford circle is surrounded by a chain of adjacents(this is not the continued fraction chain). So whenever the line leaves a circle it must enter a new one. If the line passes through a circle $\mathbb{C}[a, b]$ we have

$$\left|\gamma - \frac{a}{b}\right| < \frac{1}{2b^2}$$

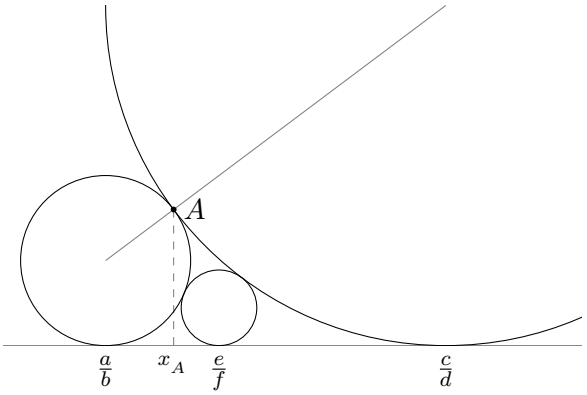
the line passes through infinitely many circles hence we have

Propostition 16. *The inequality $|\omega - y/x| < \alpha/x^2$ is satisfied infinitely often for $\alpha = \frac{1}{2}$.* □

So far we have considered the circles, but they do not cover the entire plane. Three mutually tangent circles leave the triangle in between uncovered. This area is called the *mesh*. A mesh has three corners which are the points of tangency among the circles. Let A be the point of tangency of two circles $\mathbb{C}[a, b], \mathbb{C}[c, d]$ which are both tangent to a third circle.

The point A divides the line connecting the centers of the circles in the ratio $\frac{1}{2b^2} : \frac{1}{2d^2} = d^2 : b^2$. Also $x_{_A}$, the first coordinate of A, divides the line between the basepoints in the same ratio. Hence

$$x_{_A} = \frac{b^2(a/b) + d^2(c/d)}{b^2 + d^2} = \frac{ab + cd}{b^2 + d^2}.$$



Thus for an irrational ω the line $x = \omega$ cannot pass through the corner of a mesh. To be specific we have drawn the figure for the case $a/b < c/d$ and $0 < d < b$. The argument will be entirely analogous no matter how the fractions are disposed in the inequalities.

Let B,C be the other corners of the mesh, the corresponding abscissas are

$$x_{_B} = \frac{cd + ef}{d^2 + f^2} \qquad x_{_C} = \frac{ab + ef}{b^2 + f^2}$$

We know that the line $x = \omega$ cannot pass through any of the corners, so it must pass thourgh the interior of a mesh. Then it passes through the side of the mesh which has the longest projection onto the x-axis. Suppose $\mathbb{C}[\mathcal{A}, \mathcal{B}]$ forms this side of the mesh and y is the vertex of the mesh on this circle which lies furthest away from the line $x = \mathcal{A}/\mathcal{B}$. Then we can form the inequality:

$$\left|\omega - \frac{\mathcal{A}}{\mathcal{B}}\right| < y - \frac{\mathcal{A}}{\mathcal{B}}$$

We would now like to estimate $y - \mathcal{A}/\mathcal{B}$ so that it will be less than $1/2\mathcal{B}^2$. We first consider the relations between the vertices of the mesh triangle.

What can we say about $x_{_B} - x_{_A}$? To simplify computation we find:

$$x_{_A} - \frac{a}{b} = \frac{ab + cd}{b^2 + d^2} - \frac{a}{b} = \frac{bcd - ad^2}{b(b^2 + d^2)} = \frac{d}{b} \cdot \frac{bc - ad}{b^2 + d^2} = \frac{d}{b(b^2 + d^2)}.$$

Similarly we find

$$x_{_B} - \frac{a}{b} = \frac{f}{b(b^2 + f^2)}.$$

Substracting these and recalling $f = b + d$

$$\begin{aligned} x_{_B} - x_{_A} &= \frac{f(b^2+d^2) - d(b^2+f^2)}{b(b^2+d^2)(b^2+f^2)} \\ &= \frac{(b+d)(b^2+d^2) - d(b^2+(b+d)^2)}{b(b^2+d^2)(b^2+f^2)} \\ &= \frac{b^3 - db^2 - bd^2}{b(b^2+d^2)(b^2+f^2)} = \frac{b^2 - bd - d^2}{(b^2+d^2)(b^2+f^2)} \end{aligned}$$

Substuting $s = b/d$ we get

$$x_{_B} - x_{_A} = \frac{d^2(s^2 - s - 1)}{(b^2+d^2)(b^2+f^2)}.$$

If the difference were zero we would have $s = (1 \pm \sqrt{5})/2$ but s is by definition rational, hence $x_{_B}, x_{_A}$ can't coincide. We can use the roots of $s^2 - s - 1$ to characterize different cases.

$$x_{_B} - x_{_A} = \frac{d^2\left(s - (1+\sqrt{5})/2\right)\left(s - (1-\sqrt{5})/2\right)}{(b^2+d^2)(b^2+f^2)}$$

so the sign of $x_{_B} - x_{_A}$ only depends on the sign of $s - (1+\sqrt{5})/2$, i.e. we have $x_{_A} < x_{_B}$ if $s > (1+\sqrt{5})/2$ and $x_{_A} > x_{_B}$ if $s < (1+\sqrt{5})/2$.

Propostition 17 (Hurwitz Inequality)**.** *For any irrational ω there are infinitely many fractions such that $|\omega - a/b| < 1/\sqrt{5}b^2$.*

Proof. We will distinguish two cases $x_{_A} < x_{_B}$ and $x_{_A} > x_{_B}$.

In the case $x_{_A} < x_{_B}$ we know that the line $x_{_A}x_{_C}$ is the longest projection of a side of the mesh, hence $x_{_A} < \omega < x_{_C}$. Then the $x = \omega$ passes through $\mathbb{C}[c, d]$. We have:

$$\begin{aligned} \left|\omega - \frac{b}{d}\right| = \frac{b}{d} - \omega &< \frac{b}{d} - x_{_A} \\ &= \frac{b}{d} - \frac{ab+cd}{b^2+d^2} = \frac{cb^2 - abd}{d(b^2+d^2)} \\ &= \frac{b}{d} \cdot \frac{1}{b^2+d^2}, \end{aligned}$$

since $\frac{a}{b}, \frac{c}{d}$ are adjacent fractions. Introduce $s = b/d$, we have

$$\left|\omega - \frac{b}{d}\right| < \frac{s}{s^2+1} \cdot \frac{1}{d^2} = \frac{\xi(s)}{d^2}$$

This should be better than our previous result. If so we ought to have the following inequality:

$$0 < \xi = \frac{s}{s^2+1} \leq \frac{1}{2}$$

This is equivalent to

$$0 \leq s^2 - 2s + 1 = (s-1)^2$$

which is true. We consider the behaviour of $\xi(s)$ for $s > (1+\sqrt{5})/2$. Can we find an upper bound for ξ less than $1/2$? Let us choose a second value of s', say $s' > s$, then

$$\frac{s}{s^2+1} - \frac{s'}{s'^2+1} = \frac{ss' + s - s's^2 - s'}{(s^2+1)(s'^2+1)} = \frac{(s-s')(1-ss')}{(s^2+1)(s'^2+1)}.$$

Since $s \geq 1, 1 - ss' < 0$. Hence the difference between the values of ξ is positive, i.e. the function is decreasing. So the function attains its maximum for the smallest value of s. So in the case $x_{_A} < x_{_B}$ this leads to

$$\xi = \frac{s}{s^2+1} < \frac{(\sqrt{5}+1)/2}{\big(\sqrt{5}+1)/2\big)^2 + 1} = \frac{1}{\sqrt{5}}$$

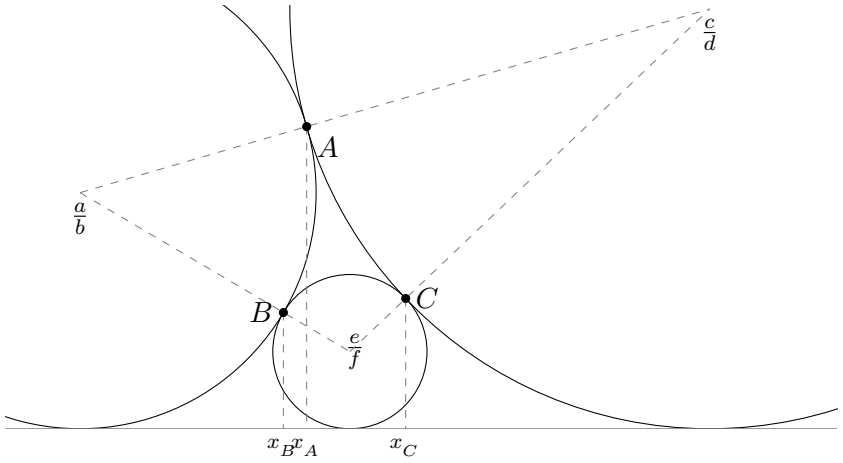
So in this case we have find the inequality

$$\left|\omega - \frac{b}{d}\right| < \frac{\xi(s)}{d^2} \leq \frac{1}{\sqrt{5}d^2}.$$

Let's consider the case $x_{_A} > x_{_B}$, remeber that we are still in the case $a/b < c/d$ and $0 < d < b$.

Now BC is the side of the mesh with the longest projection onto the x-axis, i.e. $x_{_B} < \omega < x_{_C}$. Hence the line $x = \omega$ passes through $\mathbb{C}[e, f]$ and we consider $|\omega - e/f|$. We can compare this to $|x_{_B} - e/f|$ and $|x_{_C} - e/f|$, of which the first one is the larger one, since C lies higher than B on the circle by our assumption $d > b$. We can estimate the follwoing, using a/b and e/f are adjacent.

$$\begin{aligned}\left|\omega - \frac{e}{f}\right| < \left|x_{_B} - \frac{e}{f}\right| &= \frac{e}{f} - x_{_B} = \frac{e}{f} - \frac{ab+ef}{b^2+f^2} = \frac{eb^2+abf}{f(b^2+f^2)} \\ &= \frac{b(eb+af)}{f(b^2+f^2)} = \frac{b}{f(b^2+f^2)}\end{aligned}$$



We are now in the case $1 \leq s < (\sqrt{5}+1)/2$, we would like to reduce the estiamtion to something of the form $\xi(s)/f^2$. We can form the following, recalling $f = b + d$:

$$\begin{aligned}\frac{b}{f(b^2+f^2)} &= \frac{bf}{f^2(b^2+f^2)} = \frac{1}{f^2}\left(\frac{b(b+d)}{b^2+(b+d)^2}\right) \\ &= \frac{1}{f^2}\left(\frac{s(s+1)}{s^2+(s+1)^2}\right) = \xi(s)\cdot\frac{1}{f^2}\end{aligned}$$

Thus we have $|\omega - e/f| < \xi/f^2$. What can we say about $\xi(s)$ is this case? Take s' such that $s < s' < (\sqrt{5}+1)/2$ and consider the difference $\xi(s)-\xi(s')$:

$$\begin{aligned}\frac{s(s+1)}{s^2+(s+1)^2} - \frac{s'(s'+1)}{s'^2+(s'+1)^2} &= \frac{s(s+1)}{2s(s+1)+1} - \frac{s'(s'+1)}{2s'(s'+1)+1} \\ &= \frac{s(s+1)-s'(s'+1)}{(2s(s+1)+1)\,(2s'(s'+1)+1)} \\ &= \frac{(s-s')(s+s'+1)}{(2s(s+1)+1)\,(2s'(s'+1)+1)}\end{aligned}$$

Clearly $s + s' + 1 > 0$, so this difference is negative. Thus the function is increasing and attains it maximum value at $s = (\sqrt{5}+1)/2$. Hence

$$\begin{aligned}\frac{s(s+1)}{2s(s+1)+1} &< \frac{\big((\sqrt{5}+1)/2\big)\big((\sqrt{5}+3)/2\big)}{2\big((\sqrt{5}+1)/2\big)\big((\sqrt{5}+3)/2\big)+1} \\ &= \frac{2+\sqrt{5}}{2\big(2+\sqrt{5}\big)+1} = \frac{2+\sqrt{5}}{2\sqrt{5}+5} = \frac{1}{\sqrt{5}}\end{aligned}$$

We can again conclude

$$\left|\omega - \frac{e}{f}\right| < \frac{\xi}{f^2} < \frac{1}{\sqrt{5}f^2}.$$

We have now shown Hurwitz' inequality, for any irrational ω there are infinitely many fractions such that $|\omega - a/b| < 1/\sqrt{5}b^2$. □

We now need to show that $1/\sqrt{5}$ is the best value, i.e. we must give an irrational number for which the inequality $|\omega - a/b| < \eta/b^2$ is satisfied for only finitely many fractions, $0 < \eta < 1/\sqrt{5}$. It turns out the golden mean $\omega = \big(\sqrt{5}+1\big)/2$ will do the trick.[69]

Ford also considered complex continued fractions, this gives rise to Ford spheres. There is an analogue of Hurwitz's theorem in the complex case, proven by Ford himself.[27]

Chapter 10

Decimals vs. continued fractions

Sandra Hommersom

By now we know that we can represent a real number in different ways. We already knew about the decimal expansion and during this course, we learned how to represent a real number by its continued fraction expansion. In this chapter, we are going to compare these two expansions. Our main goal is to prove the theorem of the German mathematician Gustav Lochs. This theorem says that for almost every real number x, if we are given the first n decimals of x, then in average we can obtain about the same number of partial quotients of x. To prove this theorem, we need a proposition of the French mathematician Paul Lévy. First we do some preliminaries.

Let us start with a few examples: we consider the number π. At first we assume that we are given the first 10 decimals of $\pi = 3,14159265358\ldots$. We define rational numbers $x = 3,1415926535$ and $y = 3,1415926536$, then we know $x < \pi < y$. Using this, we can get information about the partial quotients of π. We can work out the continued fraction expansions of x and y by hand. We then obtain:

$$x = [3; 7, 15, 1, 292, 1, 1, 6, 2, 13, 3, 1, 12, 3], \quad y = [3; 7, 15, 1, 292, 1, 1, 1, 4, 1, 1, 1, 45, 1, 1, 8].$$

If we forget about the integer part 3, we see that the first six partial quotients of the continued fraction expansions of x and y coincide. Since π lies in between these two numbers, its first six partial quotients must be the same. Thus: we obtain six partial quotients of the continued fraction expansion of π from its first ten decimals.

The second example is from Gustav Lochs himself. In 1963 he showed that if we are given the first 1000 decimals of π, then we will obtain 968 partial quotients of the continued fraction expansion. After we have proved the theorem of Lochs, we will know that this result holds 'in general' (that is, for almost every real number).

10.1 Preliminaries

We will now derive some properties related to the continued fraction expansion of a real number x. Without loss of generality we assume that $x \in [0,1)$. We denote the n^{th} partial quotient of x by $a_n(x)$. Furthermore, for $a_n(x) = a_n$ we denote the n^{th} convergent of x by $\frac{p_n(x)}{q_n(x)} = [0; a_1, \ldots, a_n]$. These convergents satisfy the following well-known relations:

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

and we have the conventions $p_{-1} = q_0 = 1$, $q_{-1} = 0$ and $p_0 = a_0 = 0$.
We will also use the continued fraction map

$$T : [0,1) \to [0,1),\ x \mapsto \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor.$$

For $x = [0; a_1, a_2, \ldots]$, we have $Tx = [0; a_2, a_3, \ldots]$, i.e. $a_n(Tx) = a_{n+1}(x)$ for every $x \in [0,1)$.

Definition 10.1.1. For $a_1, \ldots, a_n \in \mathbb{Z}_{\geq 1}$, the **cylinder** of $a_1, \ldots, a_n$ is the set

$$\Delta(a_1, \ldots, a_n) = \{x \in [0,1) \mid a_1(x) = a_1 \wedge \ldots \wedge a_n(x) = a_n\},$$

or simply Δ_n if it is clear which integers we are taking the cylinder of.

For now, we consider cylinders $\Delta_n = \Delta(a_1, \ldots, a_n)$ containing real numbers $x = [0; a_1, a_2, \ldots]$. The following lemma gives more information about what these cylinders look like. Notice that by definition, the first n convergents of two elements in Δ_n will always be the same.

Lemma 10.1.2. *Δ_n is an interval. If $\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}$ are the first n convergents belonging to elements of Δ_n, then*

$$\Delta_n = \begin{cases} \left[\frac{p_n}{q_n}, \frac{p_n + p_{n-1}}{q_n + q_{n-1}}\right) & \text{if } n \text{ even} \\ \left(\frac{p_n + p_{n-1}}{q_n + q_{n-1}}, \frac{p_n}{q_n}\right] & \text{if } n \text{ odd} \end{cases}.$$

Furthermore, the length of this interval is $\frac{1}{q_n(q_n+q_{n-1})}$.

Proof. One should give this proof by induction. Here we only consider the cases $n = 1, 2$ to give the intuition and leave the rest to the reader.
Case $n = 1$. We have $x \in \Delta_1$ if and only if $a_1(x) = a_1$. It holds that $a_1 = \lfloor \frac{1}{x-a_0} \rfloor = \lfloor \frac{1}{x} \rfloor$, so $a_1 \leq \frac{1}{x} < a_1 + 1$ and therefore $\frac{1}{a_1+1} < x \leq \frac{1}{a_1}$. On one hand, we have

$$\frac{1}{a_1} = \frac{a_1 \cdot 0 + 1}{a_1 \cdot 1 + 0} = \frac{a_1 p_0 + p_{-1}}{a_1 q_0 + q_{-1}} = \frac{p_1}{q_1}.$$

On the other hand, the natural number $a_1 + 1$ would have been the lower bound for $\frac{1}{x}$ if the integer part of x was $a_1 + 1$. Then the first convergent would have been $\frac{1}{a_1+1}$. If we write this convergent as $\frac{r_1}{s_1}$, we can again write down the relations:

$$\frac{1}{a_1 + 1} = \frac{r_1}{s_1} = \frac{(a_1 + 1)p_0 + p_{-1}}{(a_1 + 1)q_0 + q_{-1}} = \frac{p_1 + p_0}{q_1 + q_0}.$$

So we now have $x \in \Delta_1$ if and only if $\frac{p_1+p_0}{q_1+q_0} < x \leq \frac{p_1}{q_1}$. Therefore, we can conclude

$$\Delta_1 = \left(\frac{p_1 + p_0}{q_1 + q_0}, \frac{p_1}{q_1} \right].$$

Case $n = 2$. We have $x \in \Delta_2$ if and only if $a_1(x) = a_1$ and $a_2(x) = a_2$. But we also know that $a_2 = a_1(Tx)$, where $Tx = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor = \frac{1}{x} - a_1$. So similarly as in the case $n = 1$, we have the inequalities $\frac{1}{a_2+1} < Tx \leq \frac{1}{a_2}$, which we can rewrite as

$$\frac{a_2}{a_1 a_2 + 1} \leq x < \frac{a_2 + 1}{a_1 a_2 + a_1 + 1}.$$

On one hand, we have by definition

$$\frac{a_2}{a_1 a_2 + 1} = \frac{1}{a_1 + \frac{1}{a_2}} = \frac{p_2}{q_2}.$$

On the other hand, we have $\frac{a_2+1}{a_1 a_2+a_1+1} = \frac{a_2+1}{a_1(a_2+1)+1} = \frac{r_1}{s_1}$, where $\frac{r_1}{s_1}$ is the second convergent if the second partial quotient would have been $a_2 + 1$ in stead of a_2. Therefore, we also have

$$\frac{a_2 + 1}{a_1 a_2 + a_1 + 1} = \frac{r_1}{s_1} = \frac{(a_2 + 1)p_1 + p_0}{(a_2 + 1)q_1 + q_0} = \frac{p_2 + p_1}{q_2 + q_1}.$$

So we have $x \in \Delta_2$ if and only if $\frac{p_2}{q_2} \leq x < \frac{p_2+p_1}{q_2+q_1}$. Therefore we can conclude

$$\Delta_2 = \left[\frac{p_2}{q_2}, \frac{p_2 + p_1}{q_2 + q_1} \right).$$

Finally, for the length of the interval Δ_n, we have

$$\left|\frac{p_n}{q_n} - \frac{p_n + p_{n-1}}{q_n + q_{n-1}}\right| = \left|\frac{p_n q_n + p_n q_{n-1} - p_n q_n - p_{n-1} q_n}{q_n(q_n + q_{n-1})}\right| = \frac{|p_n q_{n-1} - p_{n-1} q_n|}{q_n(q_n + q_{n-1})} = \frac{1}{q_n(q_n + q_{n-1})}.$$

□

Now we will prove two more lemmas, which we both need later on to prove Lévy's proposition. The first lemma gives a relation between a real number and its convergents, while the second lemma gives an interesting relation between convergents when we apply the continued fraction map T.

Lemma 10.1.3. *Let $x \in [0, 1)$. Then for $n \geq 1$ we have*

$$0 \leq \log x - \log \tfrac{p_n(x)}{q_n(x)} \leq \tfrac{1}{q_n(x)} \qquad \text{if } n \text{ even},$$

$$0 \geq \log x - \log \tfrac{p_n(x)}{q_n(x)} \geq -\tfrac{1}{q_n(x)} \qquad \text{if } n \text{ odd}.$$

Proof. First suppose that n is even. Then by the previous lemma we have $x \geq \frac{p_n(x)}{q_n(x)}$. Therefore we have $\log x \geq \log \frac{p_n(x)}{q_n(x)}$. We can write down the following (in)equalities:

$$0 \leq \log x - \log \frac{p_n(x)}{q_n(x)} = \left(x - \frac{p_n(x)}{q_n(x)}\right) \frac{1}{\zeta} \leq \frac{1}{q_n(q_n + q_{n-1})} \frac{q_n}{p_n} = \frac{1}{p_n(q_n + q_{n-1})} \leq \frac{1}{q_n}.$$

Here we applied the Mean Value Theorem to the map log to find $\frac{p_n}{q_n} \leq \zeta \leq x$. Furthermore, we used the previous lemma to bound from above by the length of Δ_n and we used that $(q_i)_{i \geq 0}$ is an increasing sequence of natural numbers. Now suppose that n is odd. Then by the previous lemma we have $\frac{p_n + p_{n-1}}{q_n + q_{n-1}} < x \leq \frac{p_n}{q_n}$, so $\log \frac{p_n}{q_n} \geq \log x$. We again apply the Mean Value Theorem to the map log and now we find $x \leq \zeta \leq \frac{p_n}{q_n}$. Then in particular we have $\frac{p_n + p_{n-1}}{q_n + q_{n-1}} \leq \zeta$. Then we can write down the following (in)equalities:

$$0 \leq \log \frac{p_n(x)}{q_n(x)} - \log x = \left(\frac{p_n(x)}{q_n(x)} - x\right) \frac{1}{\zeta} \leq \frac{1}{q_n(q_n + q_{n-1})} \frac{q_n + q_{n-1}}{p_n + p_{n-1}} = \frac{1}{q_n(p_n + p_{n-1})} \leq \frac{1}{q_n},$$

which completes the proof in case n is odd. □

Remark 10.1.4. We can actually see that the inequalities are strict inequalities, if we don't allow x to be a rational number. But in the proof of Lévy's proposition it is enough to have this slightly weaker result. That is why we stated it like this here.

Lemma 10.1.5. *Let $x \in [0,1)$, then $p_n(x) = q_{n-1}(Tx)$.*

Proof. For $x = [0; a_1, a_2, \ldots]$, write $Tx = [0; b_1, b_2, \ldots]$. Then the relation $b_n = a_{n+1}$ holds. We prove lemma this by induction. For $n = 0$, we have $p_0(x) = a_0 = 0$ and $q_{-1}(Tx) = 0$ by convention.
Now suppose we know the relation holds up to some natural number n. Then we have

$$\begin{aligned} p_{n+1}(x) &= a_{n+1}p_n(x) + p_{n-1}(x) \\ &= a_{n+1}q_{n-1}(Tx) + q_{n-2}(Tx) \\ &= b_n q_{n-1}(Tx) + q_{n-2}(Tx) \\ &= q_n(Tx). \end{aligned}$$

So the relation holds for $n+1$. This completes the proof by induction. □

10.2 Results of Lévy and Lochs

The aim of this section is to prove the theorem of Lochs. We can do this partly by basic mathematics, but on our way we will need another result. This result is due to the mathematician Paul Lévy. Therefore we first prove the following proposition, which Paul Lévy proved in 1929.

Proposition 10.2.1 (Paul Lévy)**.** *For almost every $x \in [0,1)$, we have*

$$\lim_{n\to\infty} \frac{\log q_n(x)}{n} = \frac{\pi^2}{12\log 2}.$$

Proof. Let $x = [0; a_1, a_2, \ldots] \in [0,1)$ and let n be a natural number. Using lemma 10.1.5, we have

$$\begin{aligned} \tfrac{1}{q_n(x)} &= \tfrac{1}{q_n(x)} \tfrac{p_n(x)}{q_{n-1}(Tx)} \tfrac{p_{n-1}(Tx)}{q_{n-2}(T^2x)} \cdots \tfrac{p_2(T^{n-2}x)}{q_1(T^{n-1}x)} \\ &= \tfrac{p_n(x)}{q_n(x)} \tfrac{p_{n-1}(Tx)}{q_{n-1}(Tx)} \cdots \tfrac{p_1(T^{n-1}x)}{q_1(T^{n-1}x)}. \end{aligned}$$

Here we are allowed to write down the second equality, since $T^{n-1}x = [0; a_n, a_{n+1}, \ldots]$, so therefore $p_1(T^{n-1}x) = a_1(T^{n-1}x)p_0(T^{n-1}x) + p_{-1}(T^{n-1}x) = a_n \cdot 0 + 1 = 1$. Taking log on both sides of the equation, one sees

$$-\log q_n(x) = \log \frac{p_n(x)}{q_n(x)} + \log \frac{p_{n-1}(Tx)}{q_{n-1}(Tx)} + \ldots + \log \frac{p_1(T^{n-1}x)}{q_1(T^{n-1}x)}.$$

We know that for every m, the convergents $\frac{p_k(T^m x)}{q_k(T^m x)}$ approximate $T^m x$. Therefore, it makes sense to compare the right-hand side of the equation

to $\log x + \log Tx + \ldots + \log T^{n-1}x$. Up to some error term $R(n,x)$ we then obtain another equation for $-\log q_n(x)$, namely

$$-\log q_n(x) = \log x + \log Tx + \ldots + \log T^{n-1}x + R(n,x).$$

Combining the two equations we have for $-\log q_n(x)$, we get

$$\begin{aligned} R(n,x) &= \log \tfrac{p_n(x)}{q_n(x)} + \log \tfrac{p_{n-1}(Tx)}{q_{n-1}(Tx)} + \ldots + \log \tfrac{p_1(T^{n-1}x)}{q_1(T^{n-1}x)} - \log x - \log Tx - \ldots - \log T^{n-1}x \\ &= \left(\log \tfrac{p_n(x)}{q_n(x)} - \log x\right) + \left(\log \tfrac{p_{n-1}(Tx)}{q_{n-1}(Tx)} - \log Tx\right) + \ldots + \left(\log \tfrac{p_1(T^{n-1}x)}{q_1(T^{n-1}x)} - \log T^{n-1}x\right). \end{aligned}$$

Now we first observe that from lemma 10.1.3 for every $y \in [0,1)$ and for every $k \geq 1$ we have $|\log \frac{p_k(y)}{q_k(y)} - \log y| \leq \frac{1}{q_k(y)}$. Let further $F_1, F_2, \ldots$ be the Fibonacci sequence $1, 1, 2, 3, 5, \ldots$, then by our recursion relation for the q_k's we know that for every $k \geq 1$ the inequality $q_k \geq F_k$ holds. Using this, we are able to bound the error term from above:

$$\begin{aligned} |R(n,x)| &\leq \left|\log \tfrac{p_n(x)}{q_n(x)} - \log x\right| + \ldots + \left|\log \tfrac{p_1(T^{n-1}x)}{q_1(T^{n-1}x)} - \log T^{n-1}x\right| \\ &\leq \tfrac{1}{q_n(x)} + \ldots + \tfrac{1}{q_1(T^{n-1}x)} \\ &\leq \tfrac{1}{F_n} + \ldots + \tfrac{1}{F_1} \end{aligned}$$

So we have $|R(n,x)| \leq \sum_{i=1}^{n} \frac{1}{F_i}$, which we want to bound from above by a constant. For that aim, let $G := \frac{1+\sqrt{5}}{2} > 1$ and $g := \frac{1}{G} = \frac{1-\sqrt{5}}{2} < 1$ be the golden means. Then one has the equality

$$F_k = \frac{G^k - g^k}{\sqrt{5}}$$

and therefore F_k 'grows like' $\frac{G^k}{\sqrt{5}}$ at infinity. Since $\sum_{i=0}^{\infty} \sqrt{5} G^{-i}$ is a geometric series, we have that $\sum_{i=1}^{n} \frac{1}{F_i}$ is the n^{th} partial sum of the convergent series $\sum_{i=1}^{\infty} \frac{1}{F_i} =: C$. Thus, we have found an upper bound for the error term, namely $|R(n,x)| \leq C$ for every n, x.
We now have obtained that $-\lim_{n\to\infty} \frac{\log q_n(x)}{n}$ exists if and only if $\lim_{n\to\infty} \frac{1}{n}(\log x + \log Tx + \ldots + \log T^{n-1}x)$ exists, and these limits are equal. But from earlier lectures we know that $([0,1), \mathcal{L}, \mu, T)$ is an ergodic system, so we can apply the Ergodic Theorem to see that the second limit exists for almost every x. So by the Ergodic Theorem we have for almost every x:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log T^i x = \int_{[0,1)} \log d\mu = \frac{1}{\log 2} \int_0^1 \frac{\log x}{1+x} dx.$$

Now the integral on the right is not an easy one and we will not compute it here, because that is part of another course. But one can check that

$$\int_0^1 \frac{\log x}{1+x} dx = -\frac{1}{2} \sum_{i=1}^{\infty} \frac{1}{i^2} = -\frac{1}{2} \frac{\pi^2}{6} = -\frac{\pi^2}{12}$$

and from that we obtain $\int_{[0,1)} \log d\mu = -\frac{\pi^2}{12 \log 2}$. We know that this is also the limit of $-\frac{\log q_n(x)}{n}$ for $n \to \infty$ for almost every x, so that we can finally conclude:

$$\text{for almost every } x \in [0,1) : \lim_{n \to \infty} \frac{\log q_n(x)}{n} = \frac{\pi^2}{12 \log 2}.$$

□

We are now able to prove the theorom of Lochs. The proof we are following here is actually the one given by Lochs himself in 1964. Our task is to understand his proof and fill in the gaps that Lochs left to the reader.

Theorem 10.2.2 (Gustav Lochs)**.** *For $x \in [0,1)$, let m denote the number of partial quotients $b_k(x)$ of the continued fraction expansion of x that can be obtained from the first n decimals in the decimal expansion of x. Then for almost every $x \in [0,1)$ we have*

$$\lim_{n \to \infty} \frac{m}{n} = \frac{6 \log 2 \log 10}{\pi^2} \approx 0,9703.$$

Before giving the proof we recall the definition of the **small-o notation**, that is: $f(x) = \mathrm{o}(g(x))$ if and only if $\lim_{n \to \infty} \frac{f(n)}{g(n)} = 0$. Intuitively it says that the growth of f is nothing compared to that of g.
The first part of the proof describes exactly the method used in the first example to obtain partial quotients of π.

Proof. Write $x = 0, a_1 a_2 \ldots$ for the decimal expansion of x. We define two rational numbers $y := 10^{-n} \lfloor 10^n x \rfloor = 0, a_1 a_2 \ldots a_n$ and $z := y + 10^{-n}$, such that $y < x < z$. Let now m be the natural number such that the first m partial quotients of y and z coincide and the $(m+1)^{\text{st}}$ partial quotients of y and z are different. Then we can write $y = [0; b_1, \ldots, b_m, y_{m+1}]$ and $z = [0; b_1, \ldots, b_m, z_{m+1}]$ and also we know that the first m partial quotients of x are $b_1, \ldots, b_m$ (so that this natural number m is actually the number m mentioned in the theorem).

Now notice that we know $y = \frac{p_m y_{m+1}+p_{m-1}}{q_m y_{m+1}+q_{m-1}}$, since this relation holds for every $k \geq 0$. But then we can write down the following equations:

$$\begin{aligned} y &= \frac{p_m q_m y_{m+1}+p_{m-1}q_m+p_m q_{m-1}-p_m q_{m-1}}{q_m(q_m y_{m+1}+q_{m-1})} \\ &= \frac{p_m(q_m y_{m+1}+q_{m-1})+(-1)^m}{q_m(q_m y_{m+1}+q_{m-1})} \\ &= \frac{p_m}{q_m} + \frac{(-1)^m}{q_m(q_m y_{m+1}+q_{m-1})}. \end{aligned}$$

Similarly, this holds for z: $z = \frac{p_m}{q_m} + \frac{(-1)^m}{q_m(q_m z_{m+1}+q_{m-1})}$, so then we have

$$10^{-n} = z - y = \frac{|y_{m+1} - z_{m+1}|}{(q_m y_{m+1} + q_{m-1})(q_m z_{m+1} + q_{m-1})}.$$

We are now going to study this last expression. Therefore we define a few numbers, these are $u := \max(y_{m+1}, z_{m+1})$, $v := |y_{m+1} - z_{m+1}|$ and $t := \frac{q_{m-1}}{q_m}$. Note that then holds $\{u, u+v\} = \{y_{m+1}, z_{m+1}\}$. The denominator of the expression for $z - y$ contains similar formulas for y_{m+1} and z_{m+1}, so we can replace them by u and $u + v$ without having any problem. Then we have the following equations:

$$\begin{aligned} 10^n &= \frac{(q_m y_{m+1}+q_{m-1})(q_m z_{m+1}+q_{m-1})}{v} \\ &= \frac{(q_m u+q_{m-1})(q_m(u+v)+q_{m-1})}{v} \\ &= \frac{(q_m u+q_m t)(q_m(u+v)+q_m t)}{v} \\ &= \frac{q_m(u+t)q_m(u+v+t)}{v} \\ &= q_m^2(u+t)\frac{u+v+t}{v} \end{aligned}$$

Now we are going to say something about the size of the factors that appear in the last equation. First notice that $0 < t < 1 < u$, so that $u + t < 2u$. For considering the right-hand factor, we distinguish two cases:

- If $v < 1$, then $\frac{u+v+t}{v} < \frac{u+1+t}{v} < \frac{3u}{v} = \frac{3u}{\min(1,v)}$.
- If $v \geq 1$, then $\frac{u+v+t}{v} = 1 + \frac{u+t}{v} \leq 1 + u + t < 3u = \frac{3u}{\min(1,v)}$.

On the other hand, we also have $u + t > 1$ and $\frac{u+v+t}{v} \geq 1$, so that

$$1 < (u+t)\frac{u+v+t}{v} \leq 2u\frac{3u}{\min(1,v)} = \frac{6u^2}{\min(1,v)}.$$

Multiplying this by q_m^2 and using the above deduced equation, we obtain

$$q_m^2 < 10^n \leq q_m^2 \frac{6u^2}{\min(1,v)}.$$

Since log is an increasing function, we have $\log(\min(1,v)) = \min(\log 1, \log v) = \min(0, \log v)$. Applying this to the inequality, we obtain the following:

$$\begin{array}{rcccl} 2\log q_m & < & n\log 10 & \leq & 2\log q_m + \log 6u^2 - \min(\log v, 0) \\ & & & \leq & 2\log q_m + \log 6 + 2\log u - \log v. \end{array}$$

Clearly we have $\log 6 = \mathrm{o}(m)$. To complete the proof of the theorem, we assume for the moment that for almost every x both $\log u$ and $\log v$ are equal to $\mathrm{o}(m)$ as well. Then the whole term $\log 6 + 2\log u - \log v$ reduces to a term $\mathrm{o}(m)$. Then we find the useful inequality

$$2 \cdot \frac{\log q_m}{m} < \frac{n}{m}\log 10 \leq 2 \cdot \frac{\log q_m}{m} + \frac{\mathrm{o}(m)}{m}.$$

Now dividing by $\log 10$, taking limits and applying the proposition of Lévy we get the result

$$\lim_{n\to\infty} \frac{n}{m} = \frac{2}{\log 10}\frac{\pi^2}{12\log 2} = \frac{\pi^2}{6\log 2\log 10}.$$

But then also $\lim_{n\to\infty} \frac{m}{n}$ exists and is equal to $\frac{6\log 2\log 10}{\pi^2}$, which completes the proof of the theorom.

We still have to prove $\log u = \mathrm{o}(m)$ and $\log v = \mathrm{o}(m)$. Let us start with $\log u$. Suppose that $\log u = \mathrm{o}(m)$ does not hold for some $x \in [0,1)$, that is $\lim_{m\to\infty} \frac{\log u}{m} = 0$ does not hold. Then there exists $\varepsilon > 0$ such that for all $N \in \mathbb{N}$ there exists $m > N$ such that $|\frac{\log u}{m}| > \varepsilon$. Since $u > 1$, we find thus infinitely many m such that $\log u > \varepsilon m$. For these m, we have $b_{m+1} \geq \lfloor u \rfloor \geq \lfloor e^{\varepsilon m} \rfloor$, so therefore we can also find $\delta > 0$ such that $b_{m+1} > e^{\delta m}$. Since $q_{m+1} = b_{m+1}q_m + q_{m-1}$, we obtain for these m

$$\log q_{m+1} > \log b_{m+1}q_m = \log b_{m+1} + \log q_m > \delta m + \log q_m. \tag{10.1}$$

On the other hand, from the proposition of Lévy we have $\frac{\log q_m}{m} = \frac{\pi^2}{12\log 2} + \alpha_m$ where $\alpha_m \to 0$ if $m \to \infty$. The latter implies that $m\alpha_m = \mathrm{o}(m)$, so that we obtain for all m

$$\log q_m = \frac{\pi^2}{12\log 2}m + \mathrm{o}(m) \tag{10.2}$$

Since $\log q_m \neq \mathrm{o}(m)$, the equations 10.1 and 10.2 contradict each other. Thus we can conclude that if $\log u \neq \mathrm{o}(m)$ for some $x \in [0,1)$, then that x does not satisfy the proposition of Lévy. That is exactly: for almost every $x \in [0,1)$ we have $\log u = \mathrm{o}(m)$.

Now we consider the case of $\log v$. Notice that $\log v$ only appears in the equation if $\min(\log v, 0) = \log v$, that is if $v < 1$. Suppose that $\log v = \mathrm{o}(m)$ does not hold for some $x \in [0,1)$, then there exists $\varepsilon > 0$ such that for all $N \in \mathbb{N}$ there exists $m > N$ such that $|\frac{\log v}{m}| > \varepsilon$. Since we may assume that $\log v < 0$, we find infinitely many m such that $-\log v > \varepsilon m$. For these m, we thus have $|a_{m+1}(y) - a_{m+1}(z)| < e^{-\varepsilon m} < 1$, which means that we can find a natural number K that is in between $a_{m+1}(y)$ and $a_{m+1}(z)$. Otherwise we have $a_{m+1}(y) = a_{m+1}(z)$, which contradicts the definition of m. Since x_{m+1} lies in between y_{m+1} an z_{m+1}, we have $a_{m+1}(x) \in \{K, K-1\}$ and certainly $|x_{m+1} - K| < e^{-\varepsilon m}$.

Write $x_{m+1} = a_{m+1}(x) + \frac{1}{x_{m+2}}$. If $a_{m+1}(x) = K$, then we obtain $x_{m+2} > e^{\varepsilon m}$, which means $a_{m+2}(x) \geq e^{\varepsilon m}$. If $a_{m+1}(x) = K - 1$, then we obtain $|\frac{1}{x_{m+2}} - 1| < e^{-\varepsilon m}$, so for m large enough $a_{m+2} = 1$ holds. Writing $x_{m+2} = 1 + \frac{1}{x_{m+3}}$ and $|\frac{1}{x_{m+2}} - 1| = \frac{x_{m+2}-1}{x_{m+2}}$, we then obtain $x_{m+3} + 1 > e^{\varepsilon m}$. Taking both cases together, we can always find $\delta > 0$ such that infinitely many partial quotients b_m of x satisfy $b_m > e^{\delta m}$ This brings us back to what we obtained in the case of $\log u$ and we will again have the contradiction between the equations 10.1 and 10.2. So therefore we also have $\log v = \mathrm{o}(m)$ for almost every $x \in [0,1)$. □

Remark 10.2.3. In the above proof, we used the following unproven fact: for almost every $x \in [0,1)$: if $n \to \infty$, then $m \to \infty$. It is not the intention to give a proof of this here, but one could intuitively understand. Also, one can think of examples for which the result does not hold. The easiest examples are rational numbers.

References

During my research on the theorem of Lochs I used the book [21] and articles [54] and [55] in particular.

Chapter 11

Entropy and the theorem of Lochs

Sandra Hommersom

In my previous lecture I gave a proof of Lochs's theorem, which gives us information about the relation between the decimal expansion and the continued fraction expansion of a real number. In this lecture we again consider this theorem, but now our goal is to prove it from a completely different area of mathematics, namely the theory of entropy. Since entropy is an unknown subject to many mathematicians, we will first give an introduction to it. The proof of Lochs's theorem we give here uses a famous theorem of Shannon, McMillan and Breiman. We will of course mention this theorem, but unfortunately not give a proof.

11.1 Introduction to entropy

Entropy is a notion of uncertainty or randomness. It was first developed by the American mathetician Claude Shannon to study the amount of information one can get from a transmitted message. This idea can be extended to the amount of information one can get from an occurring event. We can intuitively understand what entropy is. Therefore we look at the example of rolling a die.
First suppose we have a fair die, so every side occurs with probability $\frac{1}{6}$ after rolling. Then the best way to predict the outcome of dicing is just randomly guess a value. Since we have no clue about the outcome, then acually seeing the outcome gives us much information about this event. This is just another way of saying that the entropy of rolling a fair die is large.

Now suppose we have a die which rolls a 6 with probability $\frac{9}{10}$ and probability $\frac{1}{50}$ to roll each of the other values. Then the outcome of rolling this die does not give us much information, because we could already predict with pretty large probability that the outcome will be 6. This is just saying that the entropy of this die should be small.

Now it must be clear that entropy has something to do with probability theory. We can now define what entropy is in this area of mathematics.

Definition 11.1.1. Let e be an event occurring with probability p. Then we define the amount of **information** $I(e)$ of e as $I(e) := -\log p$.

We can see why this is a suitable definition: if $p = 1$, then $I(e) = 0$. That is: the event e gives us no useful information. On the other hand, when we are considering a discrete random variable with n possible values, uniformly distributed, then $I(e)$ is maximal, namely $I(e) = -\log \frac{1}{n} = \log n$.

Definition 11.1.2. Let $X = (x_1, \ldots, x_n)$ be a discrete random variable with probability distribution $P = (p_1, \ldots, p_n)$. Then we define the **entropy of** X as

$$H(X) := E(I(X)) = -\sum_{i=1}^{n} p_i \log p_i,$$

together with the notion $0 \log 0 = 0$.

Later we would like to consider the ergodic system $([0,1), \mathcal{L}, \lambda, T)$, where T is the continued fraction map. This means that now we must think of a way to define entropy on dynamical systems. Let us therefore go back to the information of a transmitted message. We view a message as a string of symbols $\ldots x_{-1}x_0x_1 \ldots$ from an alphabet $\{a_1, \ldots, a_n\}$, where every a_i has probability p_i to be received and symbols are sent independently from each other. Of course the p_i's satisfy $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$.
In ergodic theory we see this situation as a dynamical system $(X, \mathcal{F}, \mu, T)$, where:

- $X = \{a_1, \ldots, a_n\}^{\mathbb{Z}}$,
- $\mathcal{F}$ is the σ-algebra generated by sets

$$\Delta(a_{i_0}, \ldots, a_{i_{m-1}}) := \{x \in X \mid x_0 = a_{i_0}, \ldots, x_{m-1} = a_{i_{m-1}}\},$$

- μ is the product measure, that is $\mu(\Delta(a_{i_0}, \ldots, a_{i_{m-1}})) = p_{i_0} \cdot \ldots \cdot p_{i_{m-1}}$,

- T is the left shift.

Then we define the entropy of this system as $H(X, \mathcal{F}, \mu, T) = -\sum_{i=1}^{n} p_i \log p_i$. We can think of H as the average amount of information per symbol.

We are going to extend this definition to define entropy on a arbitrary measure preserving system $(X, \mathcal{F}, \mu, T)$. To the alphabet $\{a_1, \ldots, a_n\}$ we now relate a partition $\alpha = \{A_1, \ldots, A_n\}$ of X. Then with $x \in X$ we associate an infinite sequence $\ldots x_{-1}x_0x_1 \ldots$ where $x_i = a_j$ if and only if $T^i x \in A_j$.

Definition 11.1.3. Given a partition α of X, we define the **entropy of the partition** α by

$$H(\alpha) := -\sum_{i=1}^{n} \mu(A_i) \log \mu(A_i).$$

In this definition T does not appear yet. As we will see later, the entropy of the system is actually defined by the entropy of the transformation T. Therefore we need a definition that is independent of the choice of the partition α.

First we numerate some facts about partitions and properties of entropy of partitions.

Lemma 11.1.4. *Let $\alpha = \{A_1, \ldots, A_n\}$ and $\beta = \{B_1, \ldots, B_m\}$ be partitions of X.*

(i) $T^{-1}\alpha := \{T^{-1}(A_1), \ldots, T^{-1}(A_n)\}$ *is a partition of X,*

(ii) $\alpha \vee \beta := \{A_i \cap B_j \mid i = 1, \ldots, n,\ j = 1, \ldots, m\}$ *is a partition of X.*

The proof of this lemma is very straightforward and is therefore omitted. The set $\alpha \vee \beta$ is called the **common refinement** of α and β. More general, the partition β is called a **refinement** of α, written $\alpha \leq \beta$, if for all $j \leq m$ there exists $i \leq n$ such that $B_j \subset A_i$.

Lemma 11.1.5. *Let $(X, \mathcal{F}, \mu, T)$ be a measure preserving system and let α, β be partitions of X.*

(i) $H(T^{-1}\alpha) = H(\alpha)$,

(ii) If $\alpha \leq \beta$, *then* $H(\alpha) \leq H(\beta)$,

(iii) $H(\alpha \vee \beta) \leq H(\alpha) + H(\beta)$,

(iv) If α, β are independent, then $H(\alpha \vee \beta) = H(\alpha) + H(\beta)$.

This proof is also omitted, but intuitively it is easily seen why the statements could be true. The first part follows directly from the fact that T is measure preserving. This lemma needs one more definition:

Definition 11.1.6. Two partitions α, β are called **independent** if $\mu(A \cap B) = \mu(A)\mu(B)$ for all $A \in \alpha$, $B \in \beta$.

Now given a partition α, we consider the partition $\vee_{i=0}^{n-1} T^{-i}\alpha$. The elements of this partition look like $A_{i_0} \cap T^{-1}(A_{i_1}) \cap \ldots \cap T^{-n+1}(A_{i_{n-1}})$, which contains elements $x \in X$ satisfying $x \in A_0$, $Tx \in A_{i_1}$, $\ldots$, $T^{n-1}x \in A_{i_{n-1}}$.

Definition 11.1.7. The **entropy of T w.r.t.** α is given by

$$h(\alpha, T) := \lim_{n\to\infty} \frac{1}{n} H(\vee_{i=0}^{n-1} T^{-i}\alpha).$$

By definition, we have $H(\vee_{i=0}^{n-1} T^{-i}\alpha) = -\sum_{A \in \vee_{i=0}^{n-1} T^{-i}\alpha} \mu(A) \log \mu(A)$.

Remark 11.1.8. We don't say anything about the existence of the limit in the above definition. One could prove that the limit exists using that $(H(\vee_{i=0}^{n-1} T^{-i}\alpha))_{n\in\mathbb{N}}$ is a subadditive sequence.

We end this section by defining entropy on the measure preserving system, which was our goal. This is done by defining entropy on the transformation T. We choose an easy way to get rid of the dependence on the chosen partition, namely we just take the supremum over all partitions.

Definition 11.1.9. Define $P_{\text{fin}} := \{\alpha \mid \alpha \text{ is a partition of } X \text{ and } H(\alpha) < \infty\}$. Then the **entropy of** T is given by

$$h(T) := \sup\{h(\alpha, T) \mid \alpha \in P_{\text{fin}}\}.$$

Note that we use H to denote the entropy of a partition and h to denote the entropy of a transformation.

Remark 11.1.10. In this section we only considered finite partitions. It turns out that exactly the same definitions can be given if we allow countable partitions. We will need this generalization in the other sections.

11.2 Calculation of entropy

Practically, calculating the entropy from the definition is impossible, because we have to take a supremum over possibly an infinite number of partitions. If we were given a partition, then we could use the properties of that partition to actually do a calculation, which seems much easier. Therefore, the question arises wether we could find a partition α of X satisfying $h(T) = h(\alpha, T)$. In some cases this can be done, but first we need another definition: let α be a finite or countable partition of X and let $m, k \in \mathbb{Z}$ such that $k < m$. Then we define:

- $\alpha_k^m := \vee_{i=k}^m T^{-i}\alpha$,
- $\sigma(\alpha, T)$ is the smallest σ-algebra containing all elements of α_k^m for all $m, k \in \mathbb{Z}$ with $k < m$.

Definition 11.2.1. A partition α of X is called a **generator of** T if $\sigma(\alpha, T) = \mathcal{F}$ (up to some sets of measure 0).

As an example we will compute the entropy of the first dynamical system we considered, namely the one where T is the left shift. In this way we see that the generalized definition of entropy coincides with the one for that particular dynamical system. We will use the following theorem.

Theorem 11.2.2 (Kolmogorov-Sinai, 1958)**.** *If α is a finite or countable generator for T satisfying $H(\alpha) < \infty$, then $h(T) = h(\alpha, T)$.*

For the alphabet we take the set $\{1, 2, \ldots, n\}$ together with a given probability distribution $(p_1, \ldots, p_n)$ and we may assume $-\sum_{i=1}^n p_i \log p_i < \infty$. We can do this because we could just leave out the symbols which have occurring probability 0. As earlier, we thus consider:

- $X = \{1, \ldots, n\}^{\mathbb{Z}}$,
- $\mathcal{F}$ is the σ-algebra generated by cylinder sets

$$\Delta(i_0, \ldots, i_{m-1}) := \{x \in X \mid x_0 = i_0, \ldots, x_{m-1} = i_{m-1}\},$$

- μ is the product measure, that is $\mu(\Delta(i_0, \ldots, i_{m-1})) = p_{i_0} \cdot \ldots \cdot p_{i_{m-1}}$,
- T is the left shift.

For α we choose the **time-zero partition**, which means $\alpha = \{A_1, \ldots, A_n\}$, where $A_i := \Delta(i)$, that is $A_i = \{x \in X \mid x_0 = i\}$. Since $\mu(\Delta(i)) = p_i$, we then have $H(\alpha) = -\sum_{i=1}^{n} \mu(A_i) \log \mu(A_i) = -\sum_{i=1}^{n} p_i \log p_i < \infty$. Further we see that for every $m \in \mathbb{N}$: $T^{-m} A_i = \{x \in X \mid x_m = i\}$, so that

$$A_{i_0} \cap T^{-1}(A_{i_1}) \cap \ldots \cap T^{-m+1}(A_{i_{m-1}}) = \{x \in X \mid x_0 = i_0, \ldots, x_{m-1} = i_{m-1}\}.$$

From this we can conclude that $\vee_{i=0}^{m-1} T^{-i}\alpha$ is precisely the set of all cylinders of length m. In other words: taking the $\vee_{i=0}^{m-1} T^{-i}\alpha$ for all m together, then we obtain exactly the generators for $\mathcal{F}$. Therefore α is a (finite) generator of T. Then by the Kolmogorov-Sinai theorem we know $h(t) = h(\alpha, T) = \lim_{m\to\infty} \frac{1}{m} H(\vee_{i=0}^{m-1} T^{-i}\alpha)$. So we must compute $h(\alpha, T)$.
Now since each of α, $T^{-1}\alpha$, $\ldots$, $T^{-m+1}\alpha$ determines a different coordinate and since we are using the product measure, we know that all these partitions are independent. Then by lemma 11.1.5$(i), (iv)$ we have:

$$\begin{aligned} H(\alpha \vee T^{-1}\alpha \vee \ldots \vee T^{-m+1}\alpha) &= H(\alpha) + H(T^{-1}\alpha) + \ldots + H(T^{-m+1}\alpha) \\ &= m \cdot H(\alpha) \\ &= m \cdot -\sum_{i=1}^{n} p_i \log p_i \end{aligned}$$

So now we can conclude: $h(T) = \lim_{n\to\infty} \frac{1}{m} \cdot -m \sum_{i=1}^{n} p_i \log p_i = -\sum_{i=1}^{n} p_i \log p_i$. This indeed coincides with our earlier definition.

11.3 The theorom of Lochs

Until now we have seen two very technical sections containing a lot of definitions. In this section we will see how we could use these definitions to give a alternative proof of Lochs's theorem. First we mention the following famous theorem, which we will use to prove the theorem of Lochs.

Theorem 11.3.1 (Shannon-McMillan-Breiman)**.** *Let $(X, \mathcal{F}, \mu, T)$ be an ergodic measure preserving system and let α be a finite or countable partition of X satisfying $H(\alpha) < \infty$. Let $A_n(x)$ denote the unique element $A_n \in \vee_{i=0}^{n-1} T^{-i}\alpha$ such that $x \in A_n$. Then for almost every $x \in X$ we have:*

$$\lim_{n\to\infty} -\frac{1}{n} \log \mu(A_n(x)) = h(\alpha, T).$$

11.3.1 Computation of $h(T)$

First of all, our goal is to compute $h(T)$, where T is the continued fraction map. We will need this calculation to prove the Theorem of Lochs.

Lemma 11.3.2. *For almost every $x \in [0,1)$:*

$$\lim_{n\to\infty} \frac{\log \lambda(\Delta_n(x))}{\log \mu(\Delta_n(x))} = 1.$$

Proof. We use a well-known correspondence between the Lebesgue measure λ and the Gauss measure μ: for all $A \in \mathcal{L}$:

$$\frac{1}{2\log 2}\lambda(A) \leq \mu(A) \leq \frac{1}{\log 2}\lambda(A).$$

For A we take the cylinders $\Delta_n(x) \in \mathcal{L}$. Then the correspondence is equivalent to:

$$\log 2 \leq \frac{\lambda(\Delta_n(x))}{\mu(\Delta_n(x))} \leq 2\log 2.$$

By multiplying by $\mu(\Delta_n(x))$ and taking logarithms, we obtain:

$$\log(\log 2) + \log \mu(\Delta_n(x)) \leq \log \lambda(\Delta_n(x)) \leq \log(2\log 2) + \log \mu(\Delta_n(x)).$$

For n large enough we have $\mu(\Delta_n(x)) < 1$, so that $\log \mu(\Delta_n(x)) < 0$. So dividing by $\log \mu(\Delta_n(x))$ flips the inequality symbol for these n:

$$\frac{\log(2\log 2) + \log \mu(\Delta_n(x))}{\log \mu(\Delta_n(x))} \leq \frac{\log \lambda(\Delta_n(x))}{\log \mu(\Delta_n(x))} \leq \frac{\log(\log 2) + \log \mu(\Delta_n(x))}{\log \mu(\Delta_n(x))}.$$

Now we take a look at the denominator of the most left- and right-hand side of the inequality. Observe that for almost every x we have $\lim_{n\to\infty} \mu(\Delta_n(x)) = 0$, so that $\lim_{n\to\infty} \log \mu(\Delta_n(x)) = -\infty$. Using the Squeeze Test, this gives us the wanted result:

$$\lim_{n\to\infty} \frac{\log \lambda(\Delta_n(x))}{\log \mu(\Delta_n(x))} \text{ exists and is equal to } 1.$$

□

Notice that we could also prove this lemma for cylinders D_n belonging to the decimal map S, since we only used that the measure of the cylinders is 0 at infinity. A direct consequence of the lemma is the following equality for almost every x:

$$\lim_{n\to\infty} -\frac{1}{n}\lambda(\Delta_n(x)) = \lim_{n\to\infty} -\frac{1}{n}\mu(\Delta_n(x)),$$

that is, if one of both limits exists, then so does the other and these limits are equal. Soon we will see that the left-hand limit exists.

Now we will compute $h(T)$ using the proposition of Lévy, which we proved in the previous lecture. Recall that this proposition says the following: for almost every $x \in [0,1)$ we have $\lim_{n\to\infty} \frac{1}{n} \log q_n = \frac{\pi^2}{6 \log 2}$. The following lemma is a consequence of Lévy's result.

Lemma 11.3.3. *For almost every $x \in [0,1)$:*

$$\lim_{n\to\infty} \frac{1}{n} \log \lambda(\Delta_n(x)) = -\frac{\pi^2}{6 \log 2}.$$

Proof. From the previous lecture we know $\lambda(\Delta_n(x)) = \frac{1}{q_n(q_n+q_{n-1})}$, so therefore

$$\log \lambda(\Delta_n(x)) = \log 1 - \log q_n(q_n + q_{n-1}) = -\log q_n - \log(q_n + q_{n-1}).$$

On one hand, we have $q_n+q_{n-1} > q_n$, so that $\log q_n+\log(q_n+q_{n-1}) > 2 \log q_n$. From this we obtain thus $\log \lambda(\Delta_n(x)) < -2 \log q_n$. On the other hand, we have $q_n + q_{n-1} < 2q_n$, so that $\log q_n + \log(q_n + q_{n-1}) < \log q_n + \log 2q_n = \log 2 + 2 \log q_n$. From this we thus obtain $\log \lambda(\Delta_n(x)) > -\log 2 - 2 \log q_n$. Taking this two observations together we have the following result:

$$-\log 2 - 2 \log q_n \leq \log \lambda(\Delta_n(x)) \leq -2 \log q_n.$$

Using the proposition of Lévy and again the squeeze test we now can conclude that for almost every x:

$$\lim_{n\to\infty} \frac{1}{n} \log \lambda(\Delta_n(x)) \text{ exists and is equal to } -2 \lim_{n\to\infty} \frac{1}{n} \log q_n = -\frac{\pi^2}{6 \log 2}.$$

□

With all these results it is easy to compute $h(T)$. For the partition α we take a time-zero partition, which was introduced in the previous section. This time we choose this partition with respect to the cylinders Δ_n belonging to the continued fraction map T. We do not discuss the details here that α is a generator of T and that it has finite entropy. Using the theorems 11.2.2 and 11.3.1 and the previous two lemmas, we have for almost every x:

$$h(T) = h(\alpha, T) = \lim_{n\to\infty} -\frac{1}{n} \log \mu(\Delta_n(x)) = \lim_{n\to\infty} -\frac{1}{n} \log \lambda(\Delta_n(x)) = \frac{\pi^2}{6 \log 2}.$$

Remark 11.3.4. One could also compute entropy by the Rohlin Entropy Formula, which for the continued fraction map has the form

$$h(T) = \int_{[0,1)} \log |T'x| d\mu(x).$$

This formula holds for a larger class of functions, including n-ary transformations $T_n : x \mapsto nx \mod 1$. For example, for the decimal map $S : x \mapsto 10x \mod 1$ one obtains $h(S) = \log 10$.

11.3.2 Proof of the theorem

Let us now discuss just a few more results before we get to the theorem of Lochs.

Lemma 11.3.5. *For a cylinder $\Delta_n = \Delta(a_1, \ldots, a_n)$, let us define $\Delta_n^+ := \Delta(a_1, \ldots a_{n-1}, a_n + 1)$. Then $\lambda(\Delta_n) \leq 3\lambda(\Delta_n^+)$. The cylinders Δ_n and Δ_n^+ are called **adjacent**.*

Proof. Notice that for $i = 1, \ldots, n-1$ we have $q_i = q_i^+$. Furthermore $q_n^+ = (a_n + 1)q_{n-1}^+ + q_{n-2}^+ = (a_n + 1)q_{n-1} + q_{n-2} = q_n + q_{n-1}$. Therefore we obtain

$$\lambda(\Delta_n^+) = \frac{1}{q_n^+(q_n^+ + q_{n-1}^+)} = \frac{1}{(q_n + q_{n-1})(q_n + 2q_{n-1})}.$$

The claim $\frac{1}{3}\lambda(\Delta_n) \leq \lambda(\Delta_n^+)$ is equivalent to $3q_n(q_n+q_{n-1}) \geq (q_n+q_{n-1})(q_n+2q_{n-1})$. Here the left-hand side is equal to $3q_n^2+3q_nq_{n-1}$, while the right-hand side equals $q_n^2 + 3q_nq_{n-1} + 2q_{n-1}^2$. Since $q_{n-1} \leq q_n$, the wanted inequality thus holds. This proves the lemma. □

Now we are going to take a closer look at how Δ_{n+1} is obtained from Δ_n. From my previous lecture, we know what the Δ_n look like: they are half open intervals and the precise form depends on if n is even or odd. We see that if n is even, then $\Delta_n(1), \Delta_n(2), \ldots$ is a sequence of disjoint cylinders of length $n+1$ in Δ_n and they are ordered from left to right. Here $\Delta_n(j)$ is the set of real numbers in Δ_n with the $(n+1)^{\text{st}}$ partial quotient equal to j. Similarly, if n is odd, then we have this sequence in Δ_n ordered from right to left. In other words: to obtain Δ_{n+1} from Δ_n, one refines Δ_n from left to right if n is even, and from right to left if n is odd.
Suppose we are given such Δ_n. When we repeat the described proces for cylinders of length $n+1$, $n+2$, …, we alternately refine from left to right and from right to left. Everytime when we consider a next level, the intervals are shrinking. Now if we consider a sufficient level (which means cylinders

of length $n+j$ for j large enough), there exists an interval which is small enough in the following sense: suppose I is any interval in $[0,1)$ and Δ_n is the smallest cylinder containing I, then for amost every $x \in I$ we have either $\Delta_{n+j}(x) \subset I$, or its adjacent cylinder $(\Delta_{n+j}(x))^+ \subset I$. The important thing here is that j is bounded. Actually, it turns out that j is at most 3.

Using this knowledge, we are now able to give a proof of Lochs's theorem. The above discussion appears in the proof to compare the number of partial quotients to the number of decimals.

Theorem 11.3.6 (Gustav Lochs)**.** *For $x \in [0,1)$, let m denote the number of partial quotients $b_k(x)$ of the continued fraction expansion of x that can be obtained from the first n decimals in the decimal expansion of x. Then for almost every $x \in [0,1)$ we have*

$$\lim_{n\to\infty} \frac{m}{n} = \frac{6 \log 2 \log 10}{\pi^2} \approx 0,9703.$$

Proof. Let S denote the decimal map $x \mapsto 10x \mod 1$ and D_n the cylinders belonging to S. Suppose $x \in [0,1)$ and we have fixed some n. Given the interval $D_n(x) \subset [0,1)$, then by definition of m, the smallest cylinder containing $D_n(x)$ is $\Delta_m(x)$. By the above discussion, we find $j \leq 3$ such that either $\Delta_{m+j}(x) \subset D_n(x)$ or $(\Delta_{m+j}(x))^+ \subset D_n(x)$. If we denote this cylinder we find by Δ_{m+j}, we thus have

$$\Delta_{m+j} \subset D_n(x) \subset \Delta_m(x).$$

Using lemma 11.3.5, we see that without regard on the precise form of Δ_{m+j}, we always have $\frac{1}{3}\lambda(\Delta_{m+j}(x)) \leq \lambda(\Delta_{m+j})$. So we have

$$\frac{1}{3}\lambda(\Delta_{m+j}(x)) \leq \lambda(D_n(x)) \leq \lambda(\Delta_m(x)),$$

which implies

$$-\frac{1}{n}\log 3 + \frac{1}{n}\log \lambda(\Delta_{m+j}(x)) \leq \frac{1}{n}\log \lambda(D_n(x)) \leq \frac{1}{n}\log \lambda(\Delta_m(x)) \quad (11.1)$$

by taking logarithms and dividing by n.
We write the right-hand inequality as $\frac{1}{n}\log \lambda(D_n(x)) \leq \frac{m}{n}\frac{1}{m}\log \lambda(\Delta_m(x))$, which we can rewrite as

$$\begin{aligned} \frac{m}{n} &\leq \frac{\frac{1}{n}\log \lambda(D_n(x))}{\frac{1}{m}\log \lambda(\Delta_m(x))} \\ &= \frac{\frac{1}{n}\log \lambda(D_n(x))}{\frac{1}{m}\log \mu(\Delta_m(x))\frac{\log \lambda(\Delta_m(x))}{\log \mu(\Delta_m(x))}}. \end{aligned}$$

Here one must notice $\log \lambda(\Delta_m(x)) < 0$, so that we indeed have a $\leq$-symbol. Taking limits, we have from lemma 11.3.2 for almost every $x \in [0,1)$:

$$\lim_{n\to\infty} \frac{\log \lambda(\Delta_m(x))}{\log \mu(\Delta_m(x))} = 1.$$

Furthermore, using our computation of $h(T)$, and similarly one for $h(S)$, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \lambda(D_n(x)) = -h(S) \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{m} \log \mu(D_m(x)) = -h(T).$$

Therefore we have

$$\limsup_{n\to\infty} \frac{m}{n} \leq \frac{-h(S)}{-h(T)} = \frac{h(S)}{h(T)}.$$

So far, we have only used the right-hand inequality of equation 11.1. Considering the left-hand inequality, we obtain a similar result, namely

$$\liminf_{n\to\infty} \frac{m}{n} \geq \frac{h(S)}{h(T)},$$

so that

$$\lim_{n\to\infty} \frac{m}{n} = \frac{h(S)}{h(T)}.$$

As we already mentioned once, a calculation on the Rohlin Entropy Formula for S results in $h(S) = \log 10$. Therefore we can finally conclude

$$\lim_{n\to\infty} \frac{m}{n} = \frac{\log 10}{\frac{\pi^2}{6 \log 2}} = \frac{6 \log 2 \log 10}{\pi^2}.$$

□

References

During my research on entropy and its relation to the theorem of Lochs, I used the book [21].

One could have observed that these lecture notes contain a lot af gaps, which are left to the reader. For example the discussion, following the lemma about adjacent cylinders, is not very precise, but also the proofs of two famous theorems are ommitted. Therefore, this project provides an opportunity for further research. Completing this research will lead to full understanding of what is happening here.
For now I chose to not going through these gaps, because I think I would pass the limit of this course 'Continued Fractions'. But I certainly think these are interesting points to have a look at in the future.

Chapter 12

Complex continued fractions

Ewelina Omiljan

12.1 Greatest common divisor of two Gaussian integers

Take

$$u, v \in \{a + bi | a, b \in \mathbb{Z}\}$$

such that

$$|u| \geqslant |v|.$$

Then

$$(u_{k+1}, v_{k=1}) = (v_k, u_k - g v_k),$$

where g is a Gaussian integer nearest $\frac{u_k}{v_k}$ and rounding down real and imaginary part. We halt when some $v_r = 0$.

Example 18.

$$(u, v) = (u_0, v_0) = (77 + 190i, 20 + 204i)$$

$$|u_0| \geqslant |v_0|$$

$$g_1 = 1 - 0 \cdot i$$

$$(u_1, v_1) = (20 + 204i, 77 + 190i - 1 \cdot 20 + 1 \cdot 204i) = (20 + 204i, 57 - 14i)$$

$$g_2 = 0 + 31$$

$$(u_2, v_2) = (57 - 14i, -22 + 33i)$$

$$g_3 = -1 - i$$

$$(u_3, v_3) = (-22 + 33i, 2 - 3i)$$
$$g_4 = 11$$
$$(u_4, v_4) = (2 - 3i, 0)$$

12.2 Generalized circles

Definition 19. *A generalized circle, or g-circle is the set of complex solutions to an equation of the form:*

$$Aw\overline{w} + Bw + \overline{B}\overline{w} + D = 0,$$

where $\overline{w}$ denotes complex conjugation of w, A and D are real coefficients, B is complex coefficient. They satisfy: $B\overline{B} - AD \geqslant 0$

We can denote a g-circle by the matrix $\begin{pmatrix} A & \overline{B} \\ B & D \end{pmatrix}$. Then the equation has the form:

$$Aw\overline{w} + Bw + \overline{B}\overline{w} + D = \begin{pmatrix} \overline{w} & 1 \end{pmatrix} \begin{pmatrix} A & \overline{B} \\ B & D \end{pmatrix} \begin{pmatrix} w \\ 1 \end{pmatrix}.$$

We use that definition because the set of solutions in the complex $w = x+yi$-palne form ordinary circle with centre in $-\frac{\overline{B}}{A}$ and radius $\frac{\sqrt{|B|^2 - AD}}{|A|}$ when $A \neq 0$. When $A = 0$ they form a line $ax - by = -\frac{D}{2}$. It's from:

$Bw + \overline{B}\overline{w} + D = 0$ when $B = a + bi, w = x + yi$

$$(a + bi)(x + yi) + (a - bi)(x - yi) = -D$$
$$ax + by = -\frac{D}{2}.$$

If $A = 0$ then $ax + by = -\frac{D}{2}$ and when $D = 0$ it is of the form $y = \frac{a}{b}x$ so it passes through the origin. If $A \neq 0$ then g-circle has centre in $-\frac{\overline{B}}{A}$ and radius $|\frac{B}{A}|$.

12.3 Hurwitz mapping

Now we will show that the map

$$w \to \frac{1}{w}$$

maps g-circles to g-circles. To show that we take g-circle

$$C = \begin{pmatrix} A & \overline{B} \\ B & D \end{pmatrix}.$$

Under the map we get:

$$A \cdot \frac{1}{w} \cdot \frac{1}{\overline{w}} + B \cdot \frac{1}{w} + \overline{B} \cdot \frac{1}{\overline{w}} + D = 0$$

$$A + B\overline{w} + \overline{B}w + Dw\overline{w} = 0$$

So our new circle has the form

$$C = \begin{pmatrix} D & B \\ \overline{B} & A \end{pmatrix}.$$

And it is a g-circle as well.
If we take any translation of the complex plane it also maps g-circle to g-circle. Let's take

$$Hw = \frac{1}{w} - \alpha$$

and g-circle:

$$C = \begin{pmatrix} A & \overline{B} \\ B & D \end{pmatrix}.$$

We get:

$$A(\frac{1}{w} - \alpha)(\frac{1}{\overline{w}} - \overline{\alpha}) + B(\frac{1}{w} - \alpha) + \overline{B}(\frac{1}{\overline{w}} - \overline{\alpha}) + D = 0$$

$$w\overline{w}(A\alpha\overline{\alpha} - B\alpha - \overline{B}\overline{\alpha} + D) + w(-\alpha A + \overline{B}) + \overline{w}(-A\overline{\alpha} + B) + A = 0$$

$$C = \begin{pmatrix} A\alpha\overline{\alpha} - B\alpha - \overline{B}\overline{\alpha} + D & -\overline{\alpha}A + B \\ -\alpha A + \overline{B} & A \end{pmatrix}.$$

And after a lot of calculations we can write C as:

$$C = \begin{pmatrix} \overline{\alpha} & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} A & \overline{B} \\ B & D \end{pmatrix} \begin{pmatrix} \alpha & -1 \\ -1 & 0 \end{pmatrix}.$$

We should notice that H leaves determinant of the matrix C invariant. It is indeed:

$$A^2\alpha\overline{\alpha} - BA\alpha - \overline{B}A\overline{\alpha} + DA - (\overline{\alpha}\alpha A^2 - \alpha AB - \overline{\alpha}A\overline{B} + B\overline{B}) = DA - B\overline{B}$$

Like in C before applying H map.

We take z - complex number and then

$$\alpha_0 = \lfloor z \rceil$$

$$z_0 = z - \alpha_0$$

Let the circle C has centre $-\alpha_0$ and radius $|z|$.

$$||w + \alpha_0|| = (w + \alpha_0)(\overline{w + \alpha_0}) = |z|^2$$

$$(w + \alpha_0)(\overline{w} + \overline{\alpha}_0) - |z|^2 = 0$$

$$w\overline{w} + w\overline{\alpha}_0 + \overline{w}\alpha_0 + |\alpha_0|^2 - |z|^2 = 0$$

From there we have:

$$C_0 = \begin{pmatrix} 1 & \alpha_0 \\ \overline{\alpha}_0 & |\alpha_0|^2 - |z|^2 \end{pmatrix}.$$

For $n \geqslant 1$ we define $\alpha_n = \lfloor \frac{1}{z_n} \rceil$ and $z_n = \frac{1}{z_{n-1}} - \alpha_n$. Then $[\alpha_0, \alpha_1, \ldots]$ is the Hurwitz continued fraction expansion of z. By the definition z_0 lies on g-circle C_0 and also in the unit B, where:

$$B = \{z \in \mathbb{C} | -\frac{1}{2} \leqslant \Im w, \Re w \leqslant \frac{1}{2}\}$$

If we apply $H_1 : w \to \frac{1}{w} - \alpha_1$ to z_0 we obtain $z_1 \in B$. Repeating this we find g-circles $C_0, C_1, C_2, \ldots$ with corresponding matrices $C_j = \begin{pmatrix} A_j & \overline{B}_j \\ B_j & D_j \end{pmatrix}$ for $j \geqslant 0$ and complex numbers $z_j \in C_j \cap B$.
Moreover, $A_j D_j - B_j \overline{B}_j = A_0 D_0 - B_0 \overline{B}_0 = -|z|^2, j \geqslant 1$. We call C_j the sequence of g-circles corresponding to the Hurwitz expansion $z = [\alpha_0, \alpha_1, \ldots]$.

Lemma 1. *If $|z|^2 = n \in \mathbb{Z}$ then for the g-circles $C_j = \begin{pmatrix} A_j & \overline{B}_j \\ B_j & D_j \end{pmatrix}$ corresponding to the Hurwitz continued fraction expansion of z it holds that $A_j, D_j \in \mathbb{Z}$ and $B_j \in \mathbb{Z}[i]$ and $B_j\overline{B}_j - A_j D_j = n$.*

Proof. It is true for $j = 0$ then $C_0 = \begin{pmatrix} 1 & \alpha_0 \\ \overline{\alpha}_0 & |\alpha_0|^2 - |z|^2 \end{pmatrix}$
where $\alpha_0 \in \mathbb{Z}[i], \alpha_0$ is the nearest Gaussian integer to z. For $j > 0$ we use induction and Hurwitz map: $H_j : w \to \frac{1}{w} - \alpha_j$ □

12.4 Finite number of g-circles

Theorem 12.4.1. *Let z be a complex number. If $n = |z|^2 \in \mathbb{Z}_{>0}$ then the sequence $C_0, C_1, \ldots$ of g- circles corresponding to the Hurwitz expansion of z consists of finitely many different g-circles.*

Proof. If $A_j = 0$ the g-circle is a line $r_j x - i_j y = -\frac{D}{2}$ where $r_j = \Re B_j, i_j = \Im B_j$, and r_j, i_j are rational integers satysfaying ${r_j}^2 + {i_j}^2 = n$

$$0 - B_j\overline{B}_j = -n \Rightarrow -({r_j}^2 + {i_j}^2) = -n$$

From that we have that there are only finitely many solutions for B_j. For the line to intersects the unit box one needs $D_j \leqslant |r_j| + |i_j|$. For the case $A_j \neq 0$ we use induction on j. We want to show thet radius R_j satisfies ${R_j}^2 > \frac{1}{8}$ for all j.

For $j = 0$: ${R_0}^2 = n$ because $R_0 = \frac{\sqrt{|\alpha_0|^2 - |\alpha_0|^2 + |z|^2}}{1} = |z|$ and $n \in \mathbb{Z}_{>0}$.
The induction hypothesis is that if C_{j-1} is a proper circle, then the radius ${R_{j-1}}^2 > \frac{1}{8}$.
Suppose taht g-circle C_j pass throug the origin for some $j \geqslant 1$. It means that $D_j = 0$. It has to intersect the unit box B, so the point P on it has to be at the distance less than $\frac{1}{\sqrt{2}}$ from the origin.
Under H point P of C_{j-1} gets mapped to the point opposed from the origin on C_j and will be at distance at least $\sqrt{2}$. From there ${R_j}^2$of C_j will be at least $\frac{1}{2}$.
In the rest of cases A_j and D_j non-zero integers and A_{j-1},D_{j-1} as well.
Suppose that A_{j-1} and D_{j-1} have opposite signs. This means that the origin is the interior point of the g-circle $C_{j-1}.z_{j-1} \in C_{j-1} \cap B$ is at distance at most $\frac{1}{\sqrt{2}}$ from the origin. The image of C_{j-1} under H_0 is a g-circle that also has origin as an interior point. That contains $\frac{1}{z_{j-1}}$ which is at distance at least $\sqrt{2}$ from the origin. From that: ${R_j}^2 > \frac{1}{\sqrt{2}} \Rightarrow {R_j}^2 > \frac{1}{2}$.
Now if A_{j-1} and D_{j-1} have the same sign, then origin is an exterior point of C_{j-1} and C_j. The point P on C_{j-1} nearest to the origin is at distance $p < \frac{1}{\sqrt{2}}$. The diametrically opposed point Q on the same g-circle is at distance $p+$ diameter of C_{j-1} from the origin. Using induction hypothesis: $d > 2R_{j-1} > \frac{1}{\sqrt{2}}$. The diameter of the image of C_{j-1} under H_0, and from there we have diameter of C_j:

$$\frac{1}{p} - \frac{1}{d+p} = \frac{d}{(p+d)p} > \frac{\frac{1}{\sqrt{2}}}{p(p+\frac{1}{\sqrt{2}})} > \frac{p}{p(p+\frac{1}{\sqrt{2}})} > \frac{1}{\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}} = \frac{1}{\sqrt{2}}$$

so ${R_j}^2 > \frac{1}{8}$ In any case we see that ${R_j}^2 > \frac{1}{8}$. From that and lemma we have:

$$ {R_j}^2 = \frac{B_j\overline{B}_j - A_jD_j}{{A_j}^2} = \frac{n}{{A_j}^2}. $$

It leaves only finitely many posibilities for integer A_j. For C_j to intersect the unit box, its center$(-\frac{\overline{B}}{A})$ can not be too far from the origin:

$$ |-\frac{\overline{B}_j}{A_j}| \leqslant \frac{1}{2\sqrt{2}} + \frac{\sqrt{n}}{|A_j|} $$

And this leaves only finitely many possibilities for B_j for each A_j. D_j is determined by A_j and B_j so we have the same situation for D_j. □

12.5 Bounded partial quotients

Corollary 12.5.1. *Let $z \in \mathbb{C}$ be such that its norm $n = |z|^2 \in \mathbb{Z}_{>0}$ is not the sum of two squares of integers. Then the partial quotients in the Hurwitz continued fraction of z are bounded.*

Proof. According to the Theorem 0.1 the remainders z_i of the Hurwitz continued fraction operator all lie on a finite number of different g-circles. If a g-circle C_j passes through the origin then we know that $D_j = 0$ and our g-circle is $\begin{cases} if A = 0 : y = \frac{a}{b}x, where B = a + bi, w = x + yi \\ if A \neq 0 : D(-\frac{\overline{B}}{A}; |\frac{B}{A}|) \end{cases}$
and B_j is a Gaussian integer from the previous lemma and

$$ B_j\overline{B}_j - A_jD_j = n \Rightarrow B_j\overline{B}_j = n $$

$$ B_j\overline{B}_j = (a + bi)(a - bi) = a^2 + b^2 $$

This is a contradiction with assumption that n is not a sum of two integer squares. So none of the g-circles passes through the origin. It means that there exists a positive constant $c > 0$ which is the shortest distance from any of the g-circles to the origin such that $|z_j| \geqslant c$ and then:

$$ |\alpha_{j+1}| = \lfloor \frac{1}{z_j} \rceil \leqslant \lceil \frac{1}{c} \rceil $$

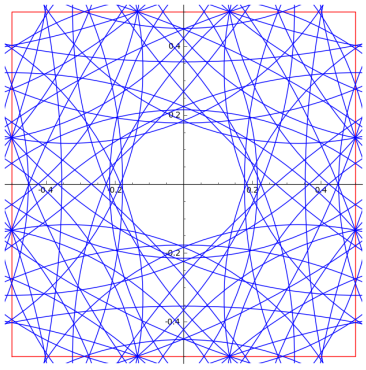
So we have our bound. □

Theorem 12.5.2. *For every even integer d there exists algebraic element $\alpha \in \mathbb{C}\backslash\mathbb{R}$ of degree d over $\mathbb{Q}$ for which the Hurwitz continued fraction expansion has bounded partial quotients.*

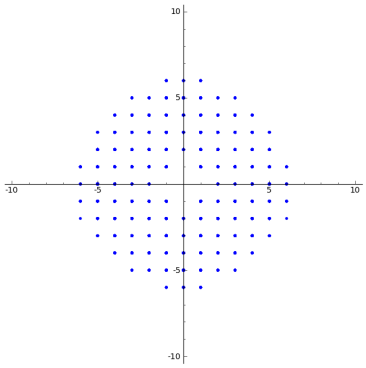
Examples 12.5.3. Let's start with $z = \sqrt{2}+i\sqrt{5}$ which norm $|z|^2 = n = 7$.
The minimal polynomial for z is $z^4 + 6z^2 + 49$.
The Hurwitz continued fraction expansion of z reads:

$$z = [2i + 1, -i + 2, i - 5, -i - 2, -4, i - 2, -4, -2, i - 1, -2i, \ldots]$$

And Doug Hensley calculated that there are probably 72 g-circles for this z.
If we'll draw it on the picture, it looks as follow:

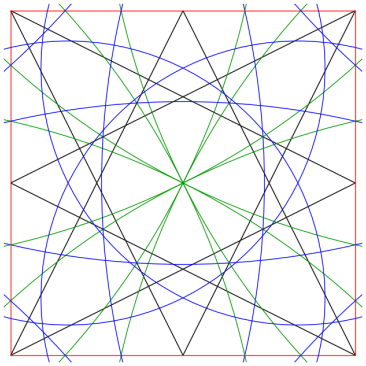


We can see there the intersection of g-circles of $z = \sqrt{2} + i\sqrt{5}$ with the unit box B.

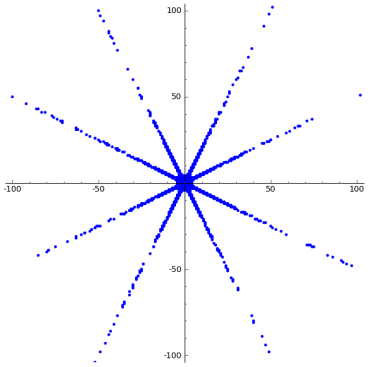


On this picture we can see the partial quotients of $z = \sqrt{2} + i\sqrt{5}$.
Now let's take an example for transcendental number. $z = \sqrt{\pi} + i\sqrt{7 - \pi}$.
The picture of g-circles and partial quotients is similar to the previous one so I will not put the picture again. It is hard to find out when some number is transcendental or not. To prove it for this z we will probably need that π is transcendental.

An example for $z = \sqrt{2} + i\sqrt{3}$ where $|z|^2 = n = 5 = 1^2 + 2^2$, so n is the sum of two squares of integers.



We see there g-circles of this z, some of them pass through the origin.



On this picture we can see partial quotients of $z = \sqrt{2} + i\sqrt{3}$. They are unbounded.

Conjecture 12.5.4. *Let $z \in \mathbb{C}$ be such that its norm $n = |z|^2 \in \mathbb{Z}_{>0}$ is the sum of two squares of integers. Then the partial quotients in the Hurwitz continued fraction of z are unbounded, unless z is in $\mathbb{Q}(i)$ or quadratic over $\mathbb{Q}(i)$.*

In my presentation I used the work [13].

Chapter 13

Geodesics

Willem van Loon

This chapter will be about geodesics. There is a very interesting connection between geodesics (on the modular surface M, the quotient of the hyperbolic plane by the modular group $\mathrm{SL}(2,\mathbb{Z})$) and continued fractions.

Let $\mathbb{H} = \{z \in \mathbb{C} | \Im(z) > 0\}$ be the upper half-plane with the Poincaré metric

$$ds^2 = \frac{dx^2 + dy^2}{y^2}$$

Furthermore, let $\mathrm{SL}(2,\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \middle| ad - bc = 1, a, b, c, d \in \mathbb{Z} \right\}$. This modular group acts on the upper complex plane as a group of fractional linear transformations via the correspondence

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

Proposition 13.0.5. *Let* $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$. *Then* γ *maps* $\mathbb{H}$ *to itself bijectively.*

Proof.

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

The inverse function of γ is $\gamma^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ □

It will be conveniënt to add infinity into the definition. Suppose $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then we define:

$$M(\infty) = \begin{cases} \frac{a}{c} & \text{if } c \neq 0 \\ \infty & \text{if } c = 0 \end{cases}$$
$$M\left(-\frac{d}{c}\right) = \infty$$

Circles and lines

Now we're going to look at circles and lines in the complex plane. We know that the equation of a line L in $\mathbb{R}^2$ has the form

$$ax + by + c = 0$$

for some $a, b, c \in \mathbb{R}$. In the complex plane we can write x and y in terms of z as $x = \frac{1}{2}(z + \overline{z}), y = \frac{1}{2i}(z - \overline{z})$. If we substitute these expressions in our equation for the line, we get

$$\frac{1}{2}(a - ib)z + \frac{1}{2}(a + ib)\overline{z} + c = 0$$

So if we let $\beta = (a - ib)/2$, then the equation of L is

$$\beta z + \overline{\beta}\overline{z} + c = 0$$

Now we do the same thing for the circle. The equation for a circle in $\mathbb{R}^2$ is

$$(x - x_0)^2 + (y - y_0)^2 = r^2.$$

Let $z = x + iy$ and $z_0 = x_0 + iy_0$. Then we get $|z - z_0|^2 = r^2$, which we can write as

$$(z - z_0)(\overline{z - z_0}) = r^2$$

If we write this out further, and let $\beta = -\overline{z_0}$ and $\gamma = z_0\overline{z_0} - r^2$, the equation for the circle becomes

$$z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$$

We can combine these two results to get the following proposition:

Proposition 13.0.6. *If A is either a circle or a line in the complex plane, then A has the equation*

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$$

where $\alpha, \gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$ ($\alpha = 0$ stands for a line).

It is not hard to see that if $\beta \in \mathbb{R}$, then we either have a circle with its on the real axis, or a vertical straight line. If we only look at $\mathbb{H}$, the upper half-plane, then these circles will become semi-circles that meets the $\mathbb{R}$-axis orthogonally. We denote this set of semi-circles orthogonal to $\mathbb{R}$ and vertical lines in the upper half-plane $\mathbb{H}$ by $\mathcal{H}$.

Proposition 13.0.7. *Let H be either (i) a semi-circle orthogonal to the real axis, or (ii) a vertical straight line. Let $\gamma \in \mathrm{SL}(2,\mathbb{Z})$. Then $\gamma(H)$ is either a semi-circle orthogonal to the real axis or a vertical straight line.*

Proof By Proposition 0.0.1 we know that γ maps the upper half-plane to itself bijectively. Hence it is sufficient to show that γ maps vertical straight lines in $\mathbb{C}$ and circles in $\mathbb{C}$ with real centres to vertical straight lines and circles with real centres.
Let L be a vertical line or circle with real centre in $\mathbb{C}$. Then L is given by an equation of the form

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$$

for some $\alpha, \beta, \gamma \in \mathbb{R}$.

Let $\omega = \gamma(z) = \frac{az+b}{cz+d}$. Then $z = \frac{d\omega - b}{-c\omega + a}$, and if we substitute this in our equation for L, we get:

$$\alpha\left(\frac{d\omega-b}{-c\omega+a}\right)\left(\frac{d\overline{\omega}-b}{-c\overline{\omega}+a}\right) + \beta\left(\frac{d\omega-b}{-c\omega+a}\right) + \beta\left(\frac{d\overline{\omega}-b}{-c\overline{\omega}+a}\right) + \gamma = 0$$

Therefore

$$\begin{aligned}&\alpha(d\omega-b)(d\overline{\omega}-b) + \beta(d\omega-b)(-c\overline{\omega}+a) + \\ &\beta(d\overline{\omega}-b)(-c\omega+a) + \gamma(-c\omega+a)(-c\overline{\omega}+a) = 0\end{aligned}$$

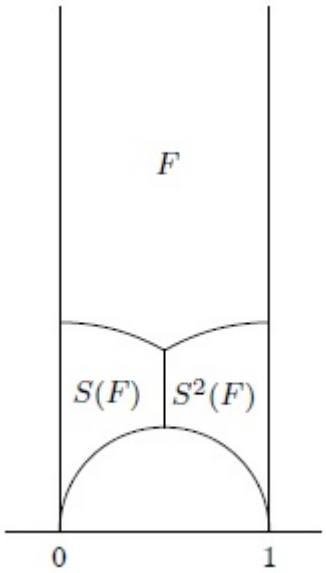
and simplifying this further, we get:

$$\begin{aligned}&(\alpha d^2 - 2\beta cd + \gamma c^2)\omega\overline{\omega} + (-\alpha bd + \beta ad + \beta bc - \gamma ac)\omega + \\ &(-\alpha bd + \beta ad + \beta bc - \gamma ac)\overline{\omega} + (\alpha b^2 - 2\beta ab + \gamma a^2) = 0\end{aligned}$$

which is exactly the equation for a circle or line.

Farey Tesselation

With this in mind we can now discuss the Farey tesselation. This is a tesselation (a covering by triangles) of $\mathbb{H}$ with the Poincaré metric by ideal triangles, which means triangles whose vertices all lie on $\mathbb{R} \cup \{\infty\}$. One way to achieve this Farey tesselation is by looking at the standard fundamental region, which is the region $F = \{z; |\Re(z)| \leq \frac{1}{2}, |z| \geq 1\}$. If we move the left half of this region one unit to the right and glue the pieces together, we get a new fundamental region, a quadrilateral with vertices i, $i+1$, $\rho = \frac{1}{2} + \sqrt{\frac{3}{4}}i$ and ∞. If we look at the images of this quadrilateral under the maps I, $S = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ and $S^2 = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$, we get three quadrilaterals, one of which is F. The next is $S(F)$ with vertices $\frac{1}{2} + \frac{1}{2}i$, i, ρ and 0. Because of Proposition 0.0.3, we know that S sends geodesics to geodesics, so the sides



of $S(F)$ must look like in the figure below.
We see that $S^2(F)$ is also a quadrilateral with vertices $i+1$, $\frac{1}{2} + \frac{1}{2}i$, ρ and 1. Now if we take D to be the union of these images, then $D = F \cup S(F) \cup S^2(F)$ is a triangle with vertices 0, 1, and ∞. Now the images of D under $\mathrm{SL}(2, \mathbb{Z})$ make up a tesselation of $\mathbb{H}$ in ideal triangles (triangles which vertices all lie either on the real line or at ∞). This I won't proof here.

Chapter 14

Hall's theorem

Roy Loos

We will look at a theorem of the mathematician Marshall Hall. He published his result in 1947. In his paper he askes himself which real numbers can be constructed if we give restrictions to the partial quotients. His theorem is about continued fractions with partial quotients not exceeding four. It tells us, more or less, that every real number can be written as a sum of two continued fractions, whose partial quotients do not exceed four.

14.1 Cantor Set

In order to formulate Hall's Theorem we will at first have a look at the construction of the Cantor Set. Cantor constructed the Cantor as an example of a perfect set. He introduced the notion of a perfect set in order to prove a weaker form of the Continuum Hypothesis. One can define the Cantor Set by means of a step by step construction. We will define an infinite sequence of closed sets $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, ...$ as follows:

We define the Cantor Set $\mathcal{C}$ as follows:

- Stage 0: we start with the unit interval, so $\mathcal{C}_0 = [0, 1]$.
- Stage 1: we remove the open middle-third-interval, that is $(\frac{1}{3}, \frac{2}{3})$. Define $\mathcal{C}_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$.
- Stage 2: remove the middle-thirds of the two closed intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. Define $\mathcal{C}_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$.
- Stage 3: now remove the middle thirds of the closed interval which appear in C_2.

- And so on

After infinite many stages we may collect the leftovers of the above construction. That is, we define the Cantor Set $\mathcal{C}$ as follows:

$$\mathcal{C} = \bigcap \{\mathcal{C}_n \mid n \in \mathbb{N}\}$$

We see that on each stage n we take the union of 2^n closed intervals. To speak about one of those 2^n intervals, we may speak simply about a closed interval which appears on the n'th stage. For example, the interval $[0, \frac{1}{9}]$ appears on the second stage. Furthermore we introduce the notion of a "successor". An interval appearing on the n'th stage, contains precisely two intervals which appear on the n+1'th stage, we call those two intervals the successors. We immediately take this terminology into practice in order to show that the Cantor Set is non-empty.

Proposition 14.1.1. *For every n in $\mathbb{N}$, for every q in $\mathbb{Q}$, if there exists a closed interval A which appears on the n'th stage such that q is an endpoint of A, then q is in $\mathcal{C}$.*

Proof. Have a close look at the construction of $\mathcal{C}$, we cannot get rid of endpoints. □

So the Cantor Set is non-empty, we also see that that the Cantor Set is a closed set, because it is an intersection of closed sets. At this point, one may think that the endpoints of the closed intervals which appear on the stages are the only points of the Cantor Set. Actually, those endpoints form just a very small part of the Cantor Set to some extent. There are many more elements of the Cantor Set, as we may deduce from the next proposition.

Proposition 14.1.2. *Let $\langle B_n \mid n \in \mathbb{N}\rangle$ be a sequence such that for every n, B_n is a closed interval which appears on the n'th stage of the Cantor Set. If for every n, $B_{n+1} \subseteq B_n$, then there exists a real number x such that $\bigcap\{B_n \mid n \in \mathbb{N}\} = \{x\}$ and x is in $\mathcal{C}$.*

Proof. At first we define a sequence $\langle x_n\rangle_{n\in\mathbb{N}}$ of real numbers, such that, for all n, x_m is the left endpoint of the closed interval B_n. Observe that the sequence $\langle x_n\rangle$ is a increasing sequence which is bounded. This means that the sequence converges, so let us say that $\lim\langle x_n\rangle = x$, for some real number x. It is clear that for every k, for every $n \geq k$, we have b_k in B_n. We conclude: for all k, for all $n \geq k$, $\lim\langle b_i\rangle_{i\geq k} \in B_n$. Thus, for all n, x in B_n. To finish the theorem we need to prove that x is the only element which exists in $\bigcap B_n$. But this is obvious, since for all n, there exists N, such that for all $i \geq N$, $|B_i| \leq \frac{1}{n}$. □

The following proposition tells us that the Cantor Set has the cardinality of the continuum.

Proposition 14.1.3. *There exists an injection φ from $2^{\mathbb{N}}$ to $\mathcal{C}$.*

Proof. We will construct the injection φ as follows. Let α be a function from $\mathbb{N}$ to 2. We will define a sequence B_n of closed intervals. At first we define $B_0 = [0, 1]$. Now, we let n in $\mathbb{N}$ and suppose that $n \geq 1$, we proceed as follows:

- if $\alpha(n-1) = 0$, then we define B_n to be the left successor of B_{n-1}
- if $\alpha(n-1) = 1$, then we define B_n to be the right successor of B_{n-1}.

Thanks to the previous lemma we can calculate x such that $\{x\} = \bigcap\{B_n \mid n \in \mathbb{N}\}$, so we define $\varphi(\alpha) = x$. □

We can also show that $\mathcal{C}$ is a perfect set. A perfect set is a closed set without isolated points. One could also say that a perfect set is a neat set.

Proposition 14.1.4. *The Cantor Set is a perfect set.*

Proof. We are done if we show that $\mathcal{C}$ has no isolated points, so we have to prove:

$$\forall x \in \mathcal{C} \forall n \exists y \in \mathcal{C}[|x - y| < \frac{1}{n}]$$

Let x in $\mathcal{C}$. Let n be a natural number. We start an infinite search for the real number x. Let $\langle B_n \rangle_{n \in \mathbb{N}}$ be a sequence such that for every n, B_n appears on the n'th stage of the Cantor Set, x in B_n and B_{n+1} is a successor of B_n. Calculate N such that $|B_N| < \frac{1}{n}$. Now define a new sequence $\langle C_n \rangle_{n \in \mathbb{N}}$ such that for all i,

- if $i \leq N$, then $C_i = B_i$
- if $i = N + 1$, then we let C_i to be the unique successor of B_N such that x is not in C_i.
- if $i > N + 1$, then we let C_i to be the left successor of C_{i-1}.

To finish the proof we let y such that $\{y\} = \bigcap\{C_n \mid n \in \mathbb{N}\}$. We conclude: $|x - y| < \frac{1}{n}$. □

The idea of the Cantor Set was to remove consecutive the middle-thirds. It is possible to generalize this idea slightly. In a Generalized Cantor Set you start just as in the normal Cantor Set with an interval A, on the first stage we are allowed to choose which piece of A we delete, so for example we are allowed to get rid of the middle-fifth of A. On each stage one has more freedom to choose the successors, but there are restrictions. The following definition slightly sloppy, but hopefully the idea is clear.

Definition 14.1.5. Let A be a closed and bounded interval. $L(A)$ is a Generalized Cantor Set, in case, there exists "stages" $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, ...$, such that $L(A) = \bigcap\{\mathcal{A}_n \mid n \in \mathbb{N}\}$ and for every n, for every C which appears on stage n, we have the following:

- C has two successors X and Y.
- $\min(X \cup Y) = \min(X)$
- $\max(X \cup Y) = \max(X)$
- $X \cap Y = \emptyset$

Definition 14.1.6. For every z in $\mathbb{Z}$, for every n in $\mathbb{N}$ we define

$$F(z, n) = \{[a_0, a_1, ...] \mid a_0 = z \wedge \forall i(i > 0 \to 1 \leq a_i \leq n)\}$$

Definition 14.1.7. For every set X, for every set Y, we define:

$$X + Y = \{x + y \mid x \in X \wedge y \in Y\}$$

We are now ready to formulate Hall's Theorem.

Theorem 14.1.8 (Hall's Theorem)**.** $\mathbb{R} \subseteq \bigcup\{F(z, 4) \mid z \in \mathbb{Z}\} + F(0, 4)$

Before we can actually prove this, we need two theorems. At first we want to find a closed and bounded interval A. The first theorem is about turning A it into some kind of Generalized Cantor Set $L(A)$ such that $L(A) = F(0, 4)$. Secondly we will prove a theorem which ensures us that $L(A)+L(A) = A+A$. If we have these two theorems then the prove is as follows.

Proof Sketch. Let x be a real number. The particular A of above will have the property that $|A + A| > 1$. Which means, that we are able to find z in $\mathbb{Z}$ and a in $A+A$, such that $x = z+a$. Because $A+A = F(0, 4)+F(0, 4)$, we determine $[a_0, a_1, ...]$ and $[b_0, b_1, ...]$ in $F(0, 4)$ such that $a = [0, a_1, ...]+[0, b_1, ...]$.

Conclude: $x = [z, a_1, a_2, ...] + [0, b_1, b_2, ...]$.

A suitable choice for A seems to be the interval $[\min F(0,4), \max F(0,4)]$. So in particular $F(0,4) \subseteq A$. We will start by thinking about the minimum and maximum of $F(0,4)$.

Lemma 14.1.9. $\max F(0,4) = [0,4,1,4,1,4,...]$ *and*
$\min F(0,4) = [0,1,4,1,4,...]$

Proof. We will construct the maximum $[a_0, a_1, ...]$ step by step. Of course we want a_0 to be as big as possible, unfortunately we are only allowed to let a_0 be zero. The second step we want $0 + \frac{1}{[a_1,...]}$ to be as big as possible. But this means we want to minimize $[a_1, a_2, ...]$, thus $a_1 = 1$. We continue with this procedure, and we find out that $a_2 = 4$, $a_3 = 1$, etcetra. The other proof goes mutatis mutandis. □

Lemma 14.1.10. *For every a in $\mathbb{N}$, for every b in $\mathbb{N}$, if $a \geq 1$ and $b \geq 1$, then*

$$[0, a, b, a, b, ...] = -\frac{b}{2} + \sqrt{\frac{b}{2} + \frac{b}{a}}$$

Proof. Observe: $[0, a, b, ...] = \frac{1}{[a,b,...]} = \frac{1}{a + \frac{1}{b+[0,a,b,...}}$. Define $x = [0, a, b, ...]$. Then the latter equation becomes: $x = \frac{1}{a+\frac{1}{b+x}}$. Some calculus transforms this equation into: $ax^2 + abx - b = 0$. You can use the abc-formula to obtain that: $x = -\frac{b}{2} \pm \sqrt{\frac{b}{2} + \frac{b}{a}}$. One of the two solutions is negative, which is impossible, hence we have proven the lemma. □

The lemma tells us a little bit more about $F(0,4)$. We are now able to calculate the minimum and maximum of $F(0,4)$, that is $A = [\frac{\sqrt{2}-1}{2}, 2\sqrt{2}-2]$. Furthermore notice that we have indeed $|A + A| > 1$.

Definition 14.1.11 $(*)$**.** Let $L(A)$ be a GCS. $L(A)$ satisfies $*$ in case for every closed interval $C = [x, x + c]$ such that C appears on some stage of $L(A)$, the length's of the two successors $C_1 = [x, x + c_1]$ and $C_2 = [x + c_1 + c_{12}, x + c]$ of C in $L(A)$ satisfy $c_{12} \leq c_1$ and $c_{12} \leq c_2$.

As notices in the "proof" of Hall's Theorem, we need to show that $L(A) + L(A) = A + A$. To show this we will proof a more or less related theorem. But before so, we need one more definition.

Definition 14.1.12. Let $C = [x, x+c]$ and $D = [y, y+d]$ be closed intervals. Calculate $e = min\{c, d\}$. We define the left associative $(C, D)^- = [x+y, x+y+2e]$ and the right associative $(C, D)^+ = [x+y+c+d-2e, x+y+c+d]$. Furthermore we define: $As(C, D) = (C, D)^- \cup (C, D)^+$.

Theorem 14.1.13. *Let A and B be closed and bounded intervals. Let $L(A)$ and $L(B)$ be Generalized Cantor Sets. If $L(A)$ and $L(B)$ both satisfy $*$, then $As(A, B) \subseteq L(A) + L(B)$.*

Proof. We will start the proof with a claim which is, as we will see later, very useful in order to prove the theorem.
Claim: Let $C = [x, x + c]$ and $D = [y, y + d]$ such that C appears on some stage of $L(A)$ and D appears on some stage of $L(B)$. Furthermore, we let $C_1 = [x, x + c_1]$ and $C_2 = [x + c_1 + c_{12}, x + c]$ be the two successors of C and in the same way we let D_1 and D_2 be the successors of D. For every real number γ such that γ in $As(C, D)$,

- if $c \leq d$, then γ in $As(C, D_1)$ or γ in $Ass(C, D_2)$.
- if $d \leq c$, then γ in $As(C_1, D)$ or γ in $Ass(C_2, D)$.

Proof of the claim:
In order to prove the claim, we let γ in $Ass(C, D)$. Observe that it is enough to prove the claim under the assumption $c \leq d$. So, we assume: $c \leq d$. We have to treat four cases.

- Case 1: $c \leq d_1$ and $c \leq d_2$. We observe the following: $(C, D)^- = [x + y, x + y + 2c] = (C, D_1)^-$ and $(C, D)^+ = (C, D_2)^+$.
- Case 2: $c \leq d_1$ and $c > d_2$. Just as in case 1 one can calculate that $As(C, D) \subseteq As(C, D_1) \cup As(C, D_2)$.
- Case 3: $c > d_1$ and $c \leq d_2$. The same trick as above.
- Case 4: $c > d_1$ and $c > d_2$. Let us write down the associatives: $(C, D_1)^-[x+y, x+y+2d_1]$, $(C, D_2)^- = [x+y+d_1+d_{12}, x+y+d_1+d_{12}+2d_2]$, $(C, D_1)^+ = [x+y+c-d_1, x+y+c+d_1]$ and $(C, D_2)^+ = [x+y+c+d-2d_2, x+y+c+d]$. We will show that these four interval will contain the set $C+D$, then we are done, because $Ass(C, D) \subseteq C+D$. Observe that $\min(C + D) = \min((C, D_1)^-$. Remember that $*$ tell us $d_{12} \leq d_1$, thus we conclude: $\min((C, D_2)^-) \leq \max((C, D_1)^-)$. Furthermore we have $\min((C, D_1)^+) \leq \max((C, D_2)^-)$, because $x+y+c \leq x+y+d = x+y+d_1+d_{12}+d_2$. We prove that $\min((C, D_2)^+) \leq \max((C, D_1)^+)$ as

follows: $*$ says that $d_{12} \leq d_2$, thus $d_{12} - d_2 \leq 0$, thus $d_1 + d_{12} - d_2 \leq d_1$, but that is just $d - 2d_2 \leq d_1$, which proves the statement.

Let γ be in $As(A, B)$. We will show that γ is in $L(A) + L(B)$. By using the above claim we can construct, step-by-step, an infinite sequence $\langle A_n, B_n \rangle_{n \in \mathbb{N}}$ such that, $A_0 = A$, $B_0 = B$, for every n, A_{n+1} is a successor of A_n in $L(A)$ and B_{n+1} is a successor of B_n in $L(B)$ and for every n, γ in $As(A_n, B_n)$. So, at each step we compare the length a_n of A_n and the length b_n of B_n and then apply the claim. Observe that the sequences $\langle a_n \rangle$ and $\langle b_n \rangle$ are decreasing in a closed interval, so the limit of these sequences will exist. Once again, we have to distinguish cases about the a_n and b_n.

- Case 1: $\lim_{n \to \infty}(\max\{a_n, b_n\}) = 0$. So if this is the case, we have in particular: $\lim a_n = 0$ and $\lim b_n = 0$, thus we are able to define real numbers α and β such that $\bigcap A_n = \{\alpha\}$ and $\bigcap B_n = \{\beta\}$. Furthermore observe that for all n, we have γ in $A_n + B_n$. In order to prove $\alpha + \beta = \gamma$, we show: for all n, $|(\alpha + \beta) - \gamma| < \frac{1}{n}$. Let n in $\mathbb{N}$. Calculate N such that $|A_N| < \frac{1}{2n}$ and $|B_N| < \frac{1}{2n}$. Conclude: $|A_n + B_n| < \frac{1}{n}$. This finishes the proof.

- Case 2: $\lim_{n \to \infty}(\max\{a_n, b_n\}) = t > 0$. In this case, there are three possibilities, which are proven in case 2a and case 2b below.

- Case 2a: $\lim b_n = t$ and $t > \lim a_n$ (this case also deals with the case $\lim a_n = t$ and $\lim b_n < t$). Define $s = \lim a_n$. Because $s < t$, we calculate N such that $a_N < t$. Conclude: $\forall n \geq N(a_n < b_n)$. Have a look at the lemma, this means, that the interval A_N is unharmed, that is, for all $n \geq N$, $a_n = A_N$. But this means also that for all $n \geq N$, $B_{n+1} \subsetneq B_n$. So, if we write $L(B)$ as an intersection of stages: $L(B) = \bigcap \mathcal{B}_n$, just as with the normal Cantor Set. Then we see: $\forall k(\bigcap\{B_n \mid n \in \mathbb{N}\} \subseteq \mathcal{B}_k)$. Conclude: $\bigcap B_n \subseteq \bigcap \mathcal{B}_n = L(B)$. Let us find suitable real numbers u, s, v and t such that: $A_N = [u, u + s]$ and $\bigcap B_n = [v, v + t]$. We have for all n, γ in $A_n + B_n$, but also for all n, γ in $A_N + B_n$. I claim the following: $\forall n[u + v - \frac{1}{n} \leq \gamma \leq u + v + \frac{1}{n}]$. To show this, we let n in $\mathbb{N}$. Calculate $M \geq N$ such that $|B_M| < \frac{1}{n}$. Because $\bigcap B_n \subseteq B_M$, it follows that $B_M \subseteq [v - \frac{1}{n}, v + t + \frac{1}{n}]$. Since $\gamma \in A_N + B_M$, we conclude $\gamma \in [u, u + s] + [v - \frac{1}{n}, v + t + \frac{1}{n}]$. Hence we have proven the claim. If we let n go to infinity, we immediately have: $\gamma \in A_N + \bigcap B_n = [u + v, u + v + s + t]$. If $u + v \leq \gamma \leq u + v + t$, then we define $\alpha = u$ and $\beta = \gamma - u$, we see that α is an endpoint of an interval appearing in $L(A)$, so we have α in $L(A)$, on the other

hand we have that β in $[v, v+t]$, so also β in $L(B)$. If it happens that $u+v+t \leq \gamma \leq u+v+s+t$, then we define $\alpha = u+s$ and $\beta = \gamma - u - s$, with the same arguments we see: α in $L(A)$ and β in $L(B)$. So in both cases we can find suitable α and β. Thus γ in $L(A) + L(B)$.

- Case 2b: $\lim a_n = t$ and $\lim b_n = t$. If it is the case that: $\exists N \forall n \geq N(A_n = A_N)$, then we can easily adapt the proof of case 2a. So let us assume that for all n, there exists $m > n$ such that $A_m \neq A_n$. We assume this also for the other sequence: for all n, there exists $m > n$, such that $B_m \neq B_n$. Let us write once again the A and B as an intersection of stages: $A = \bigcap \mathcal{A}_n$ and $B = \bigcap \mathcal{B}_n$. Now observe that: $\forall n \exists m(A_m \in \mathcal{A}_n)$ and $\forall n \exists m(B_m \in \mathcal{B}_n)$. Thus we have, $\bigcap A_n \subseteq L(A)$ and $\bigcap B_n \subseteq L(B)$. We can show γ in $\bigcap A_n + \bigcap B_n$. Now we may find real numbers u, v and t such that: $\bigcap A_n = [u, u+t]$ and $\bigcap B_n = [v, v+t]$. If $u+v \leq \gamma \leq u+v+t$, then we choose $\alpha = u$ and $\beta = \gamma - u$. If $u+v+t \leq \gamma \leq u+v+2t$, then we choose $\alpha = u+t$ and $\beta = \gamma - u - t$. In both cases we have that α in $L(A)$, β in $L(B)$ and $\alpha + \beta = \gamma$. This finishes the proof.

□

We will now turn $F(0,4)$ into a Generalized Cantor Set $L(A)$. As mentioned earlier, $L(A)$ can be seen as a intersection of stages, just as the normal Cantor Set. On the first stage we will only have the interval A. But how do we proceed? The next paragraph will answer that question. Somehow we have to make sure that $F(0,4)$ will be a subset of $L(A)$, we don't want to spill any real number which is in $F(0,4)$. We have to keep that in mind, when constructing the sages of $L(A)$. So on stage zero, we are still doing fine. In order to keep track of an element $[a_0, a_1, ...]$ in $F(0,4)$, we will look at an initial segment of $[a_0, a_1, ...]$. We will try to formalize this idea.

Definition 14.1.14. We define a very special set Γ of finite sequences as follows:

$$\begin{aligned} \langle a_0, a_1, ..., a_{n-1}, A_n \rangle \in \Gamma \quad \rightleftarrows \quad & a_0 = 0 \ \wedge \ \forall\, 0 < i < n (1 \leq a_i \leq 4) \\ & \wedge \ (\exists m [A_n = \{m\} \ \wedge \ 1 \leq m \leq 4] \\ & \vee A_n = \{2,3,4\} \vee A_n = \{3,4\}) \end{aligned}$$

Definition 14.1.15. Define the function $T : \Gamma \to \wp(F(0,4))$ as follows, for every $\langle a_0, a_1, ..., A_n \rangle$ in Γ:

$$T(\langle a_0, ..., A_n \rangle) = \{[\alpha_0, \alpha_1, ...] \mid \forall i < n [\alpha_i = a_i] \wedge \alpha_n \in A_n \wedge \forall i > n [1 \leq a_i \leq 4]\}$$

We also define a function $F : \Gamma \to \{X \mid X \subseteq A \mid X \text{ is a closed interval}\}$ by for every $\bar{a}$ in Γ, $F(\bar{a}) = [\min T(\bar{a}), \max T(\bar{a})]$.

We would like to see $L(A)$ as an intersection of stages: $L(A) = \bigcap\{\mathcal{A}_n \mid n \in \mathbb{N}\}$. The elements in $Im(F)$ are supposed to live in some stage of $L(A)$. To assign a stage to each of the elements in $Im(F)$ we just describe for every C in $Im(F)$ the two successors of C. One could formalize this proces as creating a binary tree, altough we will not do that.

Definition 14.1.16. Let n be in $\mathbb{N}$. Let $\langle a_0, ..., A_n\rangle$ be in Γ. In order to define the two successors of $F(\langle a_0, ..., A_n\rangle)$ we distinguish three cases.

- If $A_n = m$, then the two successors are: $F(\langle a_0, ..., m, 1\rangle)$ and $F(\langle a_0, ..., m, \{2, 3, 4\}\rangle)$.
- If $A_n = \{2, 3, 4\}$, then the two successors are: $F(\langle a_0, ..., a_{n-1}, 2\rangle)$ and $F(\langle a_0, ..., a_{n-1}, \{3, 4\}\rangle)$.
- If $A_n = \{3, 4\}$, then the two successors are: $F(\langle a_0, ..., a_{n-1}, 3\rangle)$ and $F(\langle a_0, ..., a_{n-1}, 4\rangle)$.

Proposition 14.1.17. *$L(A)$ is a Generalized Cantor Set, that is to say, for every $\bar{a}$ in Γ, which appears on some stage in $L(A)$, we let X and Y be the successors of $F(\bar{a})$, then:* $\min(X \cup Y) = \min F(\bar{a})$, $\max(X \cup Y) = \max F(\bar{a})$ *and* $X \cap Y = \emptyset$.

Proof. Let $\bar{a} = \langle a_0, ..., a_N\rangle$ be in Γ. We will distinguish three cases, $A_N = m$ for some m, $A_N = \{2, 3, 4\}$ and $A_N = \{3, 4\}$. We will look at the first case, $A_N = m$. However in each case we also need to distinguish whether n is even or odd, so let us assume that n is even. The two successors of $F(\bar{a})$ are $F(\bar{a}, 1)$ and $F(\bar{a}, \{2, 3, 4\})$. Let us move on to the calculations of the maximums and minimums of the intervals, we use lemma 3.1.9 for this. $\max F(\bar{a}) = [\bar{a}, 1, 4, 1, 4, ...]$, $\min F(\bar{a}) = [\bar{a}, 4, 1, 4, ...]$, $\min F(\bar{a}, 1) = [\bar{a}, 1, 1, 4, 1, 4, ...]$, $\max F(\bar{a}, \{2, 3, 4\}) = [\bar{a}, 2, 4, 1, 4, ...]$. From the above calculations we conclude: $\max F(\bar{a}, 1) = \max F(\bar{a})$, $\min F(\bar{a}, \{2, 3, 4\}) = \min F(\bar{a})$ and $\max F(\bar{a}, \{2, 3, 4\}) < \min F(\bar{a}, 1)$. This handles this specific case. The other cases are similar.

□

It is easy to see that we indeed have $L(A) = F(0, 4)$. We now want to prove that $L(A)+L(A) = A+A$, because then we have $|F(0, 4)+F(0, 4)| > 1$. Our task is to prove that $L(A)$ satisfies $*$, then we shall apply a previous theorem and conclude $As(A, A) = A + A \subseteq L(A) + L(A)$. But also $L(A) + L(A) \subseteq A + A$, thus $L(A) + L(A) = A + A$.

Lemma 14.1.18. *Let n be a natural number. Let $[a_0, ..., a_n]$ be a finite continued fraction. For every real number μ, for every real number ν, we have: $|[a_0, ..., a_n, \mu] - [a_0, ..., a_n, \nu]| = |\frac{\mu-\nu}{q_n(\mu+\epsilon)(\nu+\epsilon)}|$, where $\epsilon = \frac{q_{n-1}}{q_n}$.*

Proof. Use Theorem 2.1.5. to conclude that $[a_0, ..., a_n, \mu] = \frac{\mu p_n + p_{n-1}}{\mu q_n + q_{n-1}}$. Now observe: $|[a_0, ..., a_n, \mu] - [a_0, ..., a_n, \nu]| = |\frac{\mu p_n + p_{n-1}}{\mu q_n + q_{n-1}} - \frac{\nu p_n + p_{n-1}}{\nu q_n + q_{n-1}}| = |\frac{\mu(p_n q_{n-1} - p_{n-1} q_n) + \nu(p_{n-1} q_n - p_n q_{n-1})}{q_n(\mu+\epsilon)(\nu+\epsilon)}| = |\frac{\mu-\nu}{q_n(\mu+\epsilon)(\nu+\epsilon)}|$. In the last step we use Lemma 2.1.7, which states that $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1}$.

□

Theorem 14.1.19. *The Generalized Cantor Set $L(A)$ satisfies $*$.*

Proof. At first we want have a better understanding of the length of the intervals which appear in $L(A)$. Some agreements: $\frac{p_n}{q_n}$ is the n'th convergent of a certain continued fraction and $\epsilon = \frac{q_{n-1}}{q_n}$. Furthermore we need to define a special irrational: $\zeta = [1, 4, 1, 4, ...] = \frac{1}{2}(\sqrt{2} + 1)$. Observe that: $\frac{1}{\zeta} = [0, 1, 4, 1, 4, ...] = 4\zeta - 4$.
We now let $\bar{a} = [a_0, ..., A_n]$ be in Γ. We have to distinguish the cases whether n is odd or even and whether $A_n = m$ or $A_n = \{2, 3, 4, \}$ or $A_n = \{3, 4\}$.

Case 1: n is even and $A_n = m$.
We let $C_1 = F(\bar{a}, \{2, 3, 4\})$, $C_2 = F(\bar{a}, 1)$ and $C_{12} = F(\bar{a}) \setminus (C_1 \cup C_2)$ be as usual. We are done if we prove: $\frac{c_{12}}{c_1} \leq 1$ and $\frac{c_{12}}{c_2} \leq 1$, where the small c's are the lengths of the intervals just as before.

Have a look at the two successors of $F(\bar{a})$ and observe that $\min F(\bar{a}, 1) = [\bar{a}, 1, 1, 4, 1, ...]$,$\max F(\bar{a}, 1) = [\bar{a}, 1, 4, 1, 4, ...]$, $\min F(\bar{a}, \{2, 3, 4\}) = [\bar{a}, 4, 1, 4, ...]$ and $\max F(\bar{a}, \{2, 3, 4\}) = [\bar{a}, 2, 4, 1, 4, ...]$. We will continue by calculating the lengths of the intervals C_1, C_{12} and C_2.

The length of C_{12}. We define: $\mu = [1, 1, 4, 1, 4, ...] = 2\sqrt{2} - 1 = \frac{1}{\zeta} + 1$. Define $\nu = [2, 4, 1, 4, ...] = 1 + \zeta$. Thus: $c_{12} = \min F(\bar{a}, 1) - \max F(\bar{a}, \{2, 3, 4\}) = [\bar{a}, \mu] - [\bar{a}, \nu] = \frac{\mu-\nu}{q_n(\mu+\epsilon)(\nu+\epsilon)} = \frac{1+\frac{1}{\zeta}-\zeta}{q_n(\frac{1}{\zeta}+1+\epsilon)(1+\zeta+\epsilon}$).

The length of C_1. The minimum and maximum of C_1 are $[\bar{a}, 4, 1, 4, ...]$ respectively $[\bar{a}, 2, 4, 1, 4, ...]$. So we have: $c_1 = [\bar{a}, 2, 4, ...] - [\bar{a}, 4, 1, 4, ...] = \frac{1-3\zeta}{q_n(1+\zeta+\epsilon)(4\zeta+\epsilon)}$, because $[4, 1, 4, ...] = 4\zeta$ and $[2, 4, 1, 4, ...] = 1 + \zeta$.

The length of C_2. The minimum and maximum of C_2 are $[\bar{a}, 1, 1, 4, 1, 4, ...]$ respectively $[\bar{a}, 1, 4, 1, 4, ...]$. So we have: $c_2 = [\bar{a}, 1, 4, ...] - [\bar{a}, 1, 1, 4, ...] =] = \frac{\zeta-\frac{1}{\zeta}-1}{q_n(\zeta+\epsilon)(\frac{1}{\zeta}+1+\epsilon)}$, because $[1, 1, 4, 1, 4, ...] = \frac{1}{\zeta} + 1$ and $[1, 4, 1, 4, ...] = \zeta$.

We now calculate the fraction $\frac{c_{12}}{c_1} = \frac{\frac{1}{\zeta}-\zeta(4\zeta+\epsilon)}{(\frac{1}{\zeta}+1+\epsilon)(1-3\zeta)}$. We can see this as a function $f : \mathbb{Q} \to \mathbb{R}$ in the variable ϵ. Recall lemma 2.1.7, which states that $q_n > q_{n-1}$. So we deduce that $\epsilon \in \mathbb{Q} \cap [0, 1]$. You can use calculus (or a computer) to convince yourself that $\{f(\epsilon) \mid \epsilon \in \mathbb{Q} \cap [0, 1]\} \subseteq (0, 1]$.

If you look at the fraction $\frac{c_{12}}{c_2} = \frac{(1-\zeta)(\zeta+\epsilon)}{(1+\zeta+\epsilon)(\zeta-\frac{1}{\zeta}-1)}$, then we can do more or less the same trick. So this fraction gives rise to a function $g : \mathbb{Q} \to \mathbb{R}$ and we can conclude that $\{g(\epsilon) \mid \epsilon \in \mathbb{Q} \cap [0, 1]\} \subseteq (0, 1]$.

The above discussion threats one of the six cases. However all the other cases are just, like this case, a lot of calculations. Nothing really happens there. So I will finish the proof of the theorem here.

□

Now that we have convinced ourselves that $L(A)$ satisfies $*$, we are able to perform the *Proof Sketch*, which I wrote down earlier. So this means we have proven Hall's Theorem. Let us define for every natural number N, the set $F(N) = \{[a_0, a_1, ...] \mid a_0 \in \mathbb{Z} \wedge \forall n > 0(1 \leq a_n \leq N)\}$. In the literature one formulates Hall's Theorem simply as $F(4) + F(4) = \mathbb{R}$. Although the proof of Hall is about a slightly different statement. Hall also proved, in the same paper (1947), that every real number γ is the product of two real numbers in $F(4)$. After this, one discovered related results. Divis and Cusick showed that $F(3) + F(3) \neq \mathbb{R}$ and $3F(2) \neq \mathbb{R}$, while $3F(2) = \mathbb{R}$ and $4F(2) = \mathbb{R}$. Hlvaka showed that $F(4) + F(3) = \mathbb{R}$, $F(4) + F(2) + F(2) = \mathbb{R}$, $F(3) + F(3) + F(2) = \mathbb{R}$, $F(7) + F(2) = \mathbb{R}$, but $F(4) + F(2) \neq \mathbb{R}$ and $F(3) + F(2) + F(2) \neq \mathbb{R}$. More recently Astels showed that $F(5) \pm F(2) = \mathbb{R}$, $F(3) \pm F(4) = \mathbb{R}$, $F(3) - F(3) = \mathbb{R}$, $F(3) \pm F(2) \pm F(2) = \mathbb{R}$ and $\bigcap\{F(n) \mid n \text{ is odd}\} + \bigcap\{F(n) \mid n \text{ is odd}\} = \mathbb{R}$. So, the theorem of Hall has inspired a lot of other interesting results.

Chapter 15

Bounded complex partial quotients

Ewelina Omiljan

We take complex number $z \in C$, $z = z_0$ and sequence of approximations to z:
$(\frac{p_n}{q_n})$ where $p_n, q_n \in G = \{a + bi | a, b \in Z\}$,G - Gaussian integers and p_n, q_n - relatively prime.
We denote by $\lfloor z \rceil$ the Gaussian integer nearest z, rounding down in the both the real and imaginary part.
We take the domain $B = \{x + iy | -\frac{1}{2} \leqslant x, y < \frac{1}{2}\}$.
The Hurwitz complex continued fraction algorithm proceeds by steps of the form:

$$z_{n+1} = \frac{1}{z_n} - \lfloor \frac{1}{z_n} \rceil.$$

If $z + z_0 \in Q(i)$ then algorithm terminates when $z_n = 0$ and final finite-depth continued fraction gives a reduced fraction $\frac{p_n}{q_n}$ equal to z.
If $z \notin Q(i)$ then algorithm continues indefinitely.
As in the classical algorithm we have:

$$\begin{vmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{vmatrix} = p_{n-1}q_n - p_n q_{n-1} = (-1)^n$$

We introduce some notation:

$$z = z_0$$

is given

$$z_0 \in B$$

$$p_{-1} = q_0 = 1, p_0 = q_{-1} = 0$$

For $n \geqslant 1$ let $a_n = \lfloor \frac{1}{z_{n-1}} \rceil$.

$$p_n = a_n p_{n-1} + p_{n-2}$$

$$q_n = a_n q_{n-1} + q_{n-2}$$

Let $x_n = \frac{1}{z_{n-1}}$ and $w_n = \frac{q_{n-1}}{q_n}$. Then:

$$z_{n+1} = \frac{1}{z_n} - \lfloor \frac{1}{z_n} \rceil = \frac{1}{z_n} - a_{n+1} = Hz_n$$

And it is called the Hurwitz continued fraction operator.

$$w_{n+1} = \frac{q_n}{q_{n+1}} = \frac{q_n}{a_{n+1}q_n + q_{n-1}} = \frac{1}{a_{n+1} + \frac{q_{n-1}}{q_n}} = \frac{1}{a_{n+1} + w_n}$$

Let $\frac{1}{B}$ denote the set of reciprocals of the nonzero elements of B.

$\frac{1}{B}$ is bounded by arcs of circles of radius 1 about ± 1 and $\pm i$, and these arcs pass through $\pm 1 \pm i$ and $\pm 2 \pm 2i$.
Let G' denote $G \backslash \{0, \pm 1, \pm i\}$.

Lemma 15.0.20. *Suppose $z \in B, n \geqslant 1$ and z has a Hurwitz continued fraction to depth $n+2$. If $|w_n| \geqslant \frac{2}{3}$ then $|w_{n+1}| < \frac{2}{3}$. Furthermore, either $|w_n| < \frac{2}{3}$ or $\frac{2}{3} \leqslant |w_n| < 1$ and one of the following, or it's negative or complex-conjugate counterpart, holds: $|w_n - \frac{9}{14}(1-i)| < \frac{3}{7}$ and $a_n = 1+i$, $|w_n - \frac{9}{16}| < \frac{3}{16}$ and $a_n = 2$.*

Proof. First we start with observation that:

$$w_0 = \frac{q_{-1}}{q_0} = \frac{0}{1} = 0.$$

We define $D(s,r) := \{z \in C : |z - s| < r\}$ as a disc in C with radius r about s. If $|z| > r$ then the reciprocals of the disc have form:

$$\frac{1}{D(s,r)} = D(\frac{\overline{s}}{|s|^2 - r^2}, \frac{r}{|s|^2 - r^2})$$

(the reciprocal of $z \in C$ is $\frac{1}{z} = \frac{\overline{z}}{|z|^2}$).
Now let D_0 be the disc about 0 with radius $\frac{2}{3}$. And let $D_1 = \frac{1}{1+i+D_0}$. Using the formula $\frac{1}{D(s,r)}$ we calculate D_1.

$$D_1(\frac{\overline{0+1+i}}{(\sqrt{2})^2 - (\frac{2}{3})^2}; \frac{\frac{2}{3}}{2 - \frac{4}{9}}) = D_1(\frac{9}{14}(1-i); \frac{3}{7}).$$

In the same way we calculate other discs:

$$D_2 = \frac{1}{-1+i+D_0}$$

$$D_3 = \frac{-1}{1+i+D_0}$$

$$D_4 = \frac{1}{1-i+D_0}$$

$$D_5 = \frac{1}{2+D_0}$$

$$D_6 = \frac{1}{2i+D_0}$$

$$D_7 = \frac{-1}{2+D_0}$$

$$D_8 = \frac{1}{-2i+D_0}$$

Assuming the claim to be true for $k \leqslant n$ we have either that $|w_n| \in D_0$ or that w_n lies in the intersection of one of 8 other discs with the open unit disc and a_n takes a value either $\pm 1 \pm i$or ± 2 or $\pm 2i$. Now if $|a_n| \geqslant \sqrt{5}$ then $\frac{1}{a+D_0} \subseteq D_0$. If a is one of the 8 Gaussian integers from G' nearest the origin then $\frac{1}{a+D_0}$ is one of the $D_1, D_2, \ldots, D_8$.
For $1 \leqslant k \leqslant 8$ for any successor a to $a_n : \frac{1}{a+D_k} \subseteq D_0$. So for example when $w_n \in D_1 \backslash D_0$, we have $a_n = 1+i$ so that $a_{n+1} \in G_{1+i} = \{u = iv : u \geqslant 0, v \geqslant 0, u+v \geqslant 2\}$ where G_k denote the set of possitive value in G'.
For $a \in G_{1+i}$ though if $|a| \geqslant 2\sqrt{2}$ and $|w| < 1$ then $|\frac{1}{a+w}| < \frac{2}{3}$. It is from:

$$\frac{1}{|a+w|} < \frac{1}{|a|-|w|} < \frac{1}{2\sqrt{2}-1} \cdot \frac{2\sqrt{2}+1}{2\sqrt{2}+1} = \frac{2\sqrt{2}+1}{7} < \frac{2}{3}.$$

While if $a \in \{2, 2-i, 1-i, 1-2i, -2i\}$ then $\frac{1}{a+D_1} \subset D_0$.
So $|w_{n+1}| < \frac{2}{3}$. It shows that if $|w_n| \geqslant \frac{2}{3} \to |w_{n+1}| < \frac{2}{3}$. □

Theorem 15.0.21. *If $z \in B$ has a Hurwitz continued fraction expansion to depth $n+2$, then $|\frac{q_{n+2}}{q_n}|$.*

Proof. $\frac{1}{|w_n||w_{n+1|}} = |\frac{q_n}{q_{n-1}} \cdot \frac{q_{n+1}}{q_n}| = |\frac{q_{n+1}}{q_{n-1}}|$
And we know that $|w_n||w_{n+1}| < 1 \cdot \frac{2}{3} = \frac{2}{3}$. So $\frac{1}{|w_n||w_{n+1}|} < \frac{3}{2}$ □

Theorem 15.0.22. *Suppose $\alpha \in C$ has a Hurwitz algorithm sequence of convergents to depth at least n, and suppose (p_{n-1}, q_{n-1}) and (p_n, q_n) are the numerators and denominators of the $(n-1)$th and nth convergents. Suppose $q \in G$ with $|q_{n-1}| < |q| \leqslant |q_n|, p \in G$ and $\frac{p}{q} \neq \frac{p_n}{q_n}$. Then $|\frac{p}{q} - \alpha| \geqslant \frac{1}{5}|\frac{p_n}{q_n} - \alpha| \cdot |\frac{q_n}{q}|$.*

Proof. We first write:

$$(p, q) = s(p_{n-1}, q_{n-1}) + t(p_n, q_n)$$

From that:

$$p = sp_{n-1} + tp_n$$

$$q = sq_{n-1} + tq_n$$

If $s = 0$ then $(p, q) = t(p_n, q_n) \Rightarrow p = tp_n$ and $q = tq_n$.

$$|\frac{p}{q} - \alpha| = |\frac{tp_n}{tq_n} - \alpha| = |\frac{p_n}{q_n} - \alpha|$$

Equivalently:

$$1 \geqslant \frac{1}{5}|\frac{q_n}{q}| = \frac{1}{5}|\frac{q_n}{tq_n}| = \frac{1}{5}|\frac{1}{t}| \Rightarrow |t| \geqslant \frac{1}{5}$$

It's true because $t \neq 0$ and $t \in Z$. So the estimate is true.

If $|s| = 1$, we can consider only $s = 1$ because we can multiply s and t by the same unit.

$$|\alpha - \frac{p_n}{q_n}| = |\frac{p_n + z_n p_{n-1}}{q_n + z_n q_{n-1}} - \frac{p_n}{q_n}| = |\frac{z_n(-1)^n}{{q_n}^2(1 + z_n w_n)}| = \frac{|z_n|}{|q_n||q_n + z_n q_{n-1}|}$$

$$s + 1 \Rightarrow p = p_{n-1} + tp_n; q = q_{n-1} + tq_n$$

$$|\alpha - \frac{p}{q}| = |\frac{p_n + z_n p_{n-1}}{q_n + z_n q_{n-1}} - \frac{tp_n + p_{n-1}}{tq_n + q_{n-1}}| = |\frac{(1 - tz)(p_n q_{n-1} - p_{n-1} q_n)}{(q_n + z_n q_{n-1})(tq_n + q_{n-1})}| =$$

$$= |\frac{tz_n - 1}{q(q_n + z_n q_{n-1})}| = \frac{|z_n||t - \frac{1}{z_n}|}{|q||q_n + z_n q_{n-1}|} \geqslant \frac{1}{5}\frac{|z_n|}{|q||q_n + z_n q_{n-1}|} \cdot \frac{|q_n|}{|q_n|} = \frac{1}{5} \cdot \frac{|q_n|}{|q|} \cdot \frac{|z_n|}{|q_n + z_n q_{n-1}|} =$$

$$= \frac{1}{5} \cdot |\alpha - \frac{p_n}{q_n}| \cdot \frac{|q_n|}{|q|}$$

$q = q_{n+1} \Leftrightarrow t = \lfloor \frac{1}{z_n} \rceil$, but $|q| \leqslant |q_n| < |q_{n+1}|$, so $t \neq \lfloor \frac{1}{z_n} \rceil$ and $|t - \frac{1}{z_n}| \geqslant \frac{1}{2} \geqslant \frac{1}{5}$.

The case $|s| > 1$ we can break down into subcases depending on the value of

a_n. Because some of rotations, symmetry and reflections these cases reduce to the following:

$$|a_n| \geqslant 3, a_n = 2 + 2i, a_n = 2 + i.$$

The value of a_n constrains both w_n and z_n.
w_n because $w_n = \frac{1}{w_{n-1}+a_n}, |w_{n-1}| < 1$ so $w_n \in D(\frac{\overline{a}_n}{|a_n|^2-1}; \frac{1}{|a_n|^2-1})$.
z_n because $z_n \in D(a_n, 1) \cap B$. □

In my presentation I used [35].

Chapter 16

Binary quadratic forms

Merlijn Keune

A binary quadratic form is a map $f(x,y) = ax^2 + bxy + cy^2$ with $a, b, c \in \mathbb{Z}$. In this chapter the word form will often be used, always meaning a binary quadratic form. We first need some basic terminology.

Definitions 16.0.23. *Given a form f and an integer n, we say n is represented by f if there are $x, y \in \mathbb{Z}$ such that $f(x,y) = n$. One may wonder which integers are represented by a given form. For example, we already saw which integers can be represented by the form $f(x,y) = x^2 + y^2$; the integers that are a sum of squares.*

A form is called positive/negative definite if it only represents positive/negative integers apart from $f(0,0) = 0$, and indefinite if it represents both negative and positive integers. Note that not every form has to be in one of these three categories.

A form is called degenerate if it can be factorized into lineair parts. We are not interested in these forms, since they are easy to solve. For example, $4x^2 + 12xy + 9y^2 = (2x + 3y)^2$ and represents exactly the perfect squares, since $\gcd(2,3) = 1$.

Definition 16.0.24. Let $f(x,y) = ax^2 + bxy + cy^2$ be a form. The *matrix* of f is

$$M = \begin{pmatrix} 2a & b \\ b & 2c \end{pmatrix}.$$

The *discriminant* of f is $d = b^2 - 4ac = -\det(M)$. Note that we can now write $f(x,y) = \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix} M \begin{pmatrix} x \\ y \end{pmatrix}$.

From now on we will assume $|d|$ not to be a square, since otherwise the form would be degenerate. The following lemma then becomes an easy calculation.

Lemma 16.0.25. *Let f be a form. If $d > 0$ then f is indefinite, if $d < 0$ then f is definite.*

Lemma 16.0.26. *An integer d is the discriminant of a form if and only if $d \equiv 0, 1 \pmod 4$.*

Proof Obviously $d = b^2 - 4ac \equiv b^2 \equiv 0, 1 \pmod 4$. Conversely, if $d \equiv 0 \pmod 4$, then $x^2 - \frac{d}{4}y^2$ has discriminant d and if $d \equiv 1 \pmod 4$, then $x^2 + xy - \frac{d-1}{4}y^2$ has discriminant d.

The question we will be considering is: given an integer d, which forms have discriminant d? Since there are infinitely many of such forms, we need an equivalence relation to classify them.

Definition 16.0.27. Two forms f and f' with matrices M and M' are said to be equivalent if there is a matrix $P \in \mathrm{SL}_2(\mathbb{Z})$ such that $M' = P^\top MP$. This is a proper equivalence relation, which follows easily by looking at the identity matrix, P^{-1} and matrix multiplication. It can also be seen as the group $\mathrm{SL}_2(\mathbb{Z})$ acting on the set of forms, with the orbits as equivalence classes.

Proposition 16.0.28. *Equivalent forms have the same discriminant and represent the same integers.*

Proof Since $P \in \mathrm{SL}_2(\mathbb{Z})$ we have $\det(P) = 1$, so $\det(M') = \det(M)$. Thus $d = d'$. Now let $n \in \mathbb{Z}$ be represented by f, then there are $x, y \in \mathbb{Z}$ such that $\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix} P^\top MP \begin{pmatrix} x \\ y \end{pmatrix} = n$. Putting $\begin{pmatrix} x' \\ y' \end{pmatrix} = P \begin{pmatrix} x \\ y \end{pmatrix}$ gives $\frac{1}{2}\begin{pmatrix} x' & y' \end{pmatrix} M \begin{pmatrix} x' \\ y' \end{pmatrix} = n$. The other direction follows from symmetry.

Remark 16.0.29. If M is the matrix of $f(x, y) = ax^2 + bxy + cy^2$ and $M' = P^\top MP$ with $P = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, then M' is the matrix of $f'(x, y) = a'x^2 + b'xy + c'y^2$ with

$$\begin{aligned} a' &= f(p, r), \\ b' &= \begin{pmatrix} p & r \end{pmatrix} M \begin{pmatrix} q \\ s \end{pmatrix}, \\ c' &= f(q, s). \end{aligned}$$

This can be seen by simply filling in the equations.

16.1 Positive definite forms

The next step in this theory would normally be to look at positive definite forms. There are finitely many equivalence classes of forms for a given discriminant d, which can be canonically represented by a special sort of forms. This however has nothing to do with continued fractions, so it won't be treated here. The steps that should be taken are as follows.

- Define when a form is reduced.
- Show that the number of reduced forms for a given d is finite.
- Give an algorithm to determine a reduced form equivalent to a given form.
- Proof that no distinct reduced forms are equivalent.

Having done this, we know there is a unique reduced form in every equivalence class.

16.2 Indefinite forms

We could try to solve the problem in the indefinite case similarly, but there will be a complication. This is where continued fractions provide a solution. In this section all forms are assumed to be indefinite, thus with discriminant $d > 0$.

Definition 16.2.1. Let $f(x, y) = ax^2 + bxy + cy^2$ be a form. We define $t = \frac{-b+\sqrt{d}}{2a}$ to be the *first root* of f. Note that this is a solution of $f(x, 1) = 0$.

Lemma 16.2.2. *We have* $\frac{1}{|t|} = \left|\frac{b+\sqrt{d}}{2c}\right|$.

Proof

$$\frac{1}{t} = \frac{2a}{\sqrt{d}-b} = \frac{2a(\sqrt{d}+b)}{d-b^2} = \frac{2a(\sqrt{d}+b)}{-4ac} = -\frac{b+\sqrt{d}}{2c}.$$

Taking absolute value on both sides proves the lemma.

Definition 16.2.3. A form f is said to be *reduced* if $\frac{1}{|t|}$ is a reduced quadratic irrational, so if $1 < \frac{1}{|t|}$ and $-1 < \frac{1}{|t|}' < 0$, where $\frac{1}{|t|}'$ is the conjugate of $\frac{1}{|t|}$.

Lemma 16.2.4. *If $f(x,y) = ax^2+bxy+cy^2$ with discriminant d is reduced, then we have $0 < b < \sqrt{d}$ and $0 < |c| < \sqrt{d}$.*

PROOF We write

$$1 < \frac{1}{|t|} = \left| \frac{b+\sqrt{d}}{2c} \right| = \frac{b+\sqrt{d}}{2\varepsilon c}$$

where $\varepsilon = \pm 1$. Suppose $\varepsilon c > 0$, then since $\frac{1}{|t|}$ is reduced we have

$$b + \sqrt{d} > 2\varepsilon c = 2|c|, \quad -2|c| < b - \sqrt{d} < 0,$$

where the last implies $b < \sqrt{d}$. This gives us

$$\left.\begin{array}{r} b+\sqrt{d} > 2|c| \\ b-\sqrt{d} > -2|c| \end{array}\right\} b > 0,$$

$$\left.\begin{array}{r} b+\sqrt{d} > 2|c| \\ 0 > b-\sqrt{d} \end{array}\right\} \sqrt{d} > |c|.$$

Corollary 16.2.5. *There are only finitely many reduced indefinite forms of given discriminant d: b and c are bounded and a is determined by b, c and d.*

Definition 16.2.6. Let f be a form with matrix M. Let

$$P = \begin{pmatrix} 0 & 1 \\ -1 & k \end{pmatrix} \quad k \in \mathbb{Z},$$

then the form with matrix $P^\top MP$ is called the *right neighbour* of f *over* k. In particular:

$$f'(x,y) = f(y, ky-x) = cx^2 - (b+2ck)xy + (a+bk+ck^2)y^2.$$

Similarly we obtain the *left neighbour* of f *over* k from the matrix P^{-1}.

Proposition 16.2.7. *Let f be an indefinite form with discriminant d and first root t. If g is the right neighbour of f over k, then g has first root $k - \frac{1}{t}$.*

PROOF By our previous result for $\frac{1}{t}$:

$$\frac{b+2ck+\sqrt{d}}{2c} = k + \frac{b+\sqrt{d}}{2c} = k - \frac{1}{t}.$$

Note that d remains invariant since the forms are equivalent.

At this point we should introduce an algorithm to determine an equivalent reduced form given any indefinite form. This however is rather technical and requires some long calculations, so we will omit this and just suppose this can be done.

Corollary 16.2.8. *Every equivalence class contains at least one reduced form.*

In the definite case we would have proved that the reduced form in an equivalence class is unique. In the indefinite case this is not true.

Proposition 16.2.9. *Let $f(x,y) = ax^2 + bxy + cy^2$ be a reduced indefinite form with first root t, $|t| = \varepsilon t$ and let $\frac{1}{|t|} = [a_0; a_1, a_2, \ldots]$. Then, if g is the right neighbour of f over εa_0 and g has first root T, g is reduced, $|T| = -\varepsilon T$ and $\frac{1}{|T|} = [a_1; a_2, \ldots]$.*

Proof

$$T = \varepsilon a_0 - \frac{1}{t} = \varepsilon\left(a_0 - \frac{1}{|t|}\right) = \varepsilon(a_0 - [a_0; a_1, a_2, \ldots]) = -\varepsilon\frac{1}{[a_1; a_2, \ldots]}.$$

So $|T| = -\varepsilon T$ and

$$\frac{1}{|T|} = [a_1; a_2, \ldots].$$

Since f is reduced, $\frac{1}{|t|}$ has a purely periodic continued fraction expansion, so also $\frac{1}{|T|}$ had a purely periodic continued fraction expansion. Therefore g is reduced.

Lemma 16.2.10. *The form g from the previous proposition is the only reduced right neighbour of f.*

Proof For reduced forms we have

$$\sqrt{d} - b < 2|c| < \sqrt{d} + b$$

and

$$(\sqrt{d} - b)(\sqrt{d} + b) = d - b^2 = -4ac = 2|a| \cdot 2|c|,$$

because both $\sqrt{d} - b$ and $\sqrt{d} + b$ are positive. From this it follows that

$$\sqrt{d} - b < 2|a| < \sqrt{d} + b.$$

If $h(x,y) = a'x^2 + b'xy + c'y^2$ is the right neighbour of f over k, then $b + b' = b - b - 2ck$, so $b + b' \equiv 0 \pmod{2|c|}$. Together with the inequality $0 < \sqrt{d} - b' < 2|a'| = 2|c|$ this ensures that $k = \varepsilon a_0$ is unique. (No multiples of $2|c|$ can be added to or subtracted from b'.)

Remark 16.2.11. Also the left reduced neighbour is unique. This follows from a similar calculation.

Now for every reduced form, we have an associated sign ε and a purely periodic continued fraction $\frac{1}{|t|} = [\overline{a_0; a_1, \ldots, a_{m-1}}]$ of period m. We construct a chain of equivalent forms, beginning at some reduced form f and taking right neighbours over εa_0. As we've seen, this has the following effect:

$$\varepsilon \mapsto -\varepsilon, \quad [\overline{a_0; a_1, \ldots, a_{m-1}}] \mapsto [\overline{a_1; a_2, \ldots, a_{m-1}, a_0}].$$

If m is odd, after m steps we return at the same continued fraction, but instead of ε we have $-\varepsilon$. So the cykel is back at the beginning after $\operatorname{lcm}(2, k)$ steps. So now we have proven most of the main theorem:

Theorem 16.2.12. *Let $f(x, y)$ be a reduced indefinite form with discriminant d and first root t, where $\varepsilon t = |t|$. Suppose that $\frac{1}{|t|} = [\overline{a_0; a_1, \ldots, a_{m-1}}]$ with m even. (Take twice the smallest period if needed.) Let $f_0 = f$ and let f_i be the right neighbour of f_{i-1} by $(-1)^{i-1}\varepsilon a_{i-1}$. Then $f_0, f_1, \ldots, f_{m-1}$ are all the reduced forms equivalent to f, and $f_m = f$.*

The only thing left to prove is that there are no to f equivalent reduced forms outside this chain. Unfortunately that proof is rather hard and requires more knowledge about continued fractions then we have at this point, so this won't be proven. For instance, a nice proof can be given using so called minus-continued fractions, expressions of the form

$$a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots}}}$$

with $a_0, a_1, a_2, \ldots \geq 2$.

The number of equivalence classes is called the class number of d. This corresponds with the class number h_d we know from number theory, where $d = D_m$, the discriminant of the quadratic number field K_m.

Example 16.2.13. To calculate the class number h_{17}, we first need to determine all reduced forms $f(x, y) = ax^2 + bxy + cy^2$ with discriminant 17. We have $0 < b < \sqrt{17}$ and $0 < |c| < \sqrt{17}$. Also b is odd, since $b^2 - 4ac = 17$. So $b = 1$ or $b = 3$.

If $b = 1$, then $-4ac = 16$, so $ac = -4$. So $c \in \{\pm 1, \pm 2, \pm 4\}$. Because

$$\frac{1}{|t|} = \left| \frac{b + \sqrt{d}}{2c} \right| = \frac{1 + \sqrt{17}}{2|c|}$$

is a reduced quadratic irrational, we have $1+\sqrt{17} > 2|c|$ and $1-\sqrt{17} > -2|c|$, so $c = \pm 2$. This gives us two forms:

$$2x^2 + xy - 2y^2,\ -2x^2 + xy + 2y^2.$$

Similar reasoning gives another four reduced forms if $b = 3$:

$$x^2 + 3xy - 2y^2,\ -x^2 + 3xy + 2y^2,\ 2x^2 + 3xy - y^2,\ -2x^2 + 3xy + y^2.$$

Taking $f_0 = 2x^2 + xy - 2y^2$ with first root $t = \frac{-1+\sqrt{17}}{4}$ we get

$$\frac{1}{t} = \overline{[1; 3, 1]} = \overline{[1; 3, 1, 1, 3, 1]}\,.$$

So all reduced forms with discriminant 17 are equivalent, which tells us $h_{17} = 1$. Knowing a bit about number theory, we have now proved that $\mathbb{Z}\left[\frac{1+\sqrt{17}}{2}\right]$ is a principal ideal domain.

Chapter 17

CFs in power series fields

David Venhoek

17.1 Introduction

Instead of working from the reals, we can also make continued fraction expansions of elements in a power series field. Let k be a field, then the role of $\mathbb{Z}$ is played by $k[X]$, that of $\mathbb{Q}$ by $k(X)$ and that of $\mathbb{R}$ by $k((X^{-1}))$.[80]

Throughout this piece I will use the following notational conventions. Lower case symbols such as a, x are elements of k. Upper case symbols (with the obvious exception of X) such as A, D are elements of $k[X]$, and Greek letters such as α, ϕ are elements of $k((X^{-1}))$.

We can define a norm on the elements of $k((X^{-1}))$. Let $\alpha = \sum_{i=-t}^{\inf} a_{-i}x^{-i}$ with $a_t \neq 0$. Then the we define the norm of α to be $|\alpha| = 2^t$. This norm is special in the sense that it is a non-archimedean norm. This means that instead of the normal triangle inequality $|\alpha+\beta| \leq |\alpha|+|\beta|$ we have $|\alpha+\beta| \leq \max(|\alpha|, |\beta|)$. In this particular case we have the additional property that $|\alpha + \beta| = \max(|\alpha|, |\beta|)$ when $|\alpha| \neq |\beta|$.

The regular continued fraction expansion for elements of $k((X^{-1}))$ is defined as the expansion

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots}}}$$

where $\forall i > 0 : |a_i| > 1$. We will take the same shorthand notation used for

the continued fractions over the reals:[80]

$$[A_0] = A_0$$
$$[A_0; A_1] = A_0 + \frac{1}{A_1}$$
$$[A_0; A_1, ..., A_n] = \left[A_0; A_1, ..., A_{n-2}, A_{n-1} + \frac{1}{A_n}\right]$$

We can now introduce again the notion of convergents and complete quotients. Let $\alpha = [A_0; A_1, A_2, ...]$. Then the partial quotients are:

$$\frac{P_m}{Q_m} = [A_0; A_1, ..., A_m]$$

and the complete quotients are $\alpha_m = [A_m; A_{m+1}, ...]$. Furthermore we have $\alpha = [A_0; A_1, ..., A_{m-1}, \alpha_m]$. From this it follows immediately that $|A_i| = |\alpha_i|$.[80]

We also have a recursion relation for the partial quotients.

17.1.1 Lemma 1

Let $\frac{P_i}{Q_i}$ be the convergents of $\alpha = [A_0; A_1, ...]$. Then with $P_{-2} = 0$, $Q_{-2} = 1$, $P_{-1} = 1$ and $Q_{-1} = 0$ we have $P_i = A_iP_{i-1} + P_{i-2}$ and $Q_i = A_iQ_{i-1} + Q_{i-2}$.

Proof: Verifying for $i = 0$ is trivial. Suppose the lemma is true for all $i < n$. Then

$$\frac{P_n}{Q_n} = [A_0; A_1, ..., A_n] = \left[A_0; A_1, ..., A_{n-1} + \frac{1}{A_n}\right]$$
$$= \frac{\left(A_{n-1} + \frac{1}{A_n}\right) P_{n-2} + P_{n-3}}{\left(A_{n-1} + \frac{1}{A_n}\right) Q_{n-2} + Q_{n-3}} = \frac{P_{n-1} + \frac{P_{n-2}}{A_n}}{Q_{n-1} + \frac{Q_{n-2}}{A_n}}$$
$$= \frac{A_nP_{n-1} + P_{n-2}}{A_nQ_{n-1} + Q_{n-2}}$$

and the lemma is thus also true for n. Then by induction the lemma is true for all $i \in \mathbb{N}$.[80]

By induction it is now also easy to prove that the relationship $Q_nP_{n-1} - P_nQ_{n-1} = (-1)^n$ holds for $n \geq -1$.

Finally we need to define some useful subsets of k. $k^\times$ are all the units (invertible elements) in k. We define $(k^\times)^2$ to be the set of all squares in k, and for $a \in k^\times$, $a(k^\times)^2$ is the coset of $(k^\times)^2$ containing a.[80]

17.2 Properties of convergents

We now have all the definitions in place to prove some facts about the convergents of the continued fraction of an element α.

17.2.1 Lemma 2

If $\frac{P_n}{Q_n}$ *is a convergent of* α *then* $\left|\alpha - \frac{P_n}{Q_n}\right| = \frac{1}{|Q_{n+1}||Q_n|} < \frac{1}{|Q_n|^2}$.

Proof: Using the definitions and lemma 1 we get:

$$\left|\alpha - \frac{P_n}{Q_n}\right| = \left|\frac{\alpha_{n+1}P_n + P_{n-1}}{\alpha_{n+1}Q_n + Q_{n-1}} - \frac{P_n}{Q_n}\right|$$
$$= \left|\frac{Q_n\left(\alpha_{n+1}P_n + P_{n-1}\right) - P_n\left(\alpha_{n+1}Q_n + Q_{n-1}\right)}{Q_n\left(\alpha_{n+1}Q_n + Q_{n-1}\right)}\right|$$
$$= \left|\frac{(-1)^n}{Q_n\left(\alpha_{n+1}Q_n + Q_{n-1}\right)}\right| = \frac{1}{|Q_{n+1}|\,|Q_n|}$$

Because the degrees of the Q_n are strictly increasing, the inequality also holds.[80]

17.2.2 Lemma 3

If $\alpha \in k((X^{-1}))\backslash k(X)$, $\left|\alpha - \frac{P}{Q}\right| < \frac{1}{|Q|^2}$, *then there exists an* n *such that* $\frac{P}{Q} = \frac{P_n}{Q_n}$

Proof: There exists an n such that $|Q_n| \leq |Q| < |Q_{n+1}|$. We now have

$$\left|\alpha - \frac{P}{Q}\right| < \frac{1}{|Q|^2} < \frac{1}{|Q||Q_n|}$$
$$\left|\alpha - \frac{P_n}{Q_n}\right| = \frac{1}{|Q_n||Q_{n+1}|} < \frac{1}{|Q||Q_n|}$$

From this we can now determine $\left|\frac{P}{Q} - \frac{P_n}{Q_n}\right|$.

$$\left|\frac{P}{Q} - \frac{P_n}{Q_n}\right| = \left|\alpha - \frac{P_n}{Q_n} - \left(\alpha - \frac{P}{Q}\right)\right| \leq \max\left(\left|\alpha - \frac{P_n}{Q_n}\right|, \left|\alpha - \frac{P}{Q}\right|\right) < \frac{1}{|Q||Q_n|}$$

and thus $\frac{P}{Q} = \frac{P_n}{Q_n}$.[80]

17.3 Relations between continued fraction expansions

We will now look at an interesting relation between elements of $k((X^{-1}))$, that indicates a relationship between the continued fraction expansions of elements.

We define the relation $\approx$ as follows:[80]

$$\alpha \approx \beta \iff \exists R, S, T, U \in k[X] : RU - ST \in k^\times \wedge \beta = \frac{R\alpha + S}{T\alpha + U}$$

We will not prove that this is indeed an equivalence relation, this is very easy to verify for yourself, though we will elaborate on some of the basic facts about it in the next section.

We have two simple consequences of this definition. First $\alpha_n \approx \alpha_m \forall n, m \in \mathbb{N}$. From this it follows that $\alpha \approx \beta \forall \alpha, \beta \in K(X)$.

To analyze the consequences of this relationship further it is easier to first look at a slightly different relationship.

17.3.1 Lemma 4

Let $\alpha = \frac{A\beta+B}{C\beta+D}$ *where* $|D| < |C|$, $AD - BC = a \in k^\times$, $\alpha, \beta \notin k(X)$ *and* $|\beta| > 1$. *Let* $\frac{P_n}{Q_n}$ *be the convergents of* α. *Then for some* n *we have*

$$\frac{A}{C} = \frac{P_n}{Q_n}$$
$$\frac{B}{D} = \frac{P_{n-1}}{Q_{n-1}}$$

and $\beta = b\alpha_{n+1}$ *for some* $b \in (-1)^{n+1}a(k^\times)^2$.

Proof: We can write $\frac{A}{C}$ as a continued fraction: $\frac{A}{C} = [A_0; A_1, ..., A_n] = \frac{P_n^*}{Q_n^*}$, where the star indicates it is a convergent of $\frac{A}{C}$. A and C are coprime and thus we have $A = c^{-1}P_n^*$, $B = c^{-1}Q_n^*$ with $c \in k^\times$. Thus:

$$P_n^*D - Q_n^*B = c(AD - BC) = ac = ac(-1)^n(Q_n^*P_{n-1}^* - P_n^*Q_{n-1}^*)$$
$$P_n^*(D + (-1)^n acQ_{n-1}^* = Q_n^*(B + (-1)^n acP_{n-1}^*)$$

P_n^* and Q_n^* are coprime, and thus $Q_n^*|(D + (-1)^n acP_{n-1}^*)$. $|D| < |C| = |Q_n^*|$, and thus we have $D = (-1)^{n+1}acQ_{n-1}^*$, $B = (-1)^{n+1}acP_{n-1}^*$. Thus

$$\alpha = \frac{((-1)^{n+1}c^{-2}a^{-1}\beta)P_n^* + P_{n-1}^*}{((-1)^{n+1}c^{-2}a^{-1}\beta)Q_n^* + Q_{n-1}^*}$$

and we have $\frac{A}{C} = \frac{P_n}{Q_n}$, $\frac{B}{D} = \frac{P_{n-1}}{Q_{n-1}}$ and $\beta = (-1)^{n+1}ac^2\alpha_{n+1}$.[80]

17.3.2 Theorem 2

Let $\alpha, \beta \in K((X^{-1}))\backslash K(X)$. Then $\alpha \approx \beta$ if and only if we have $n, m \in \mathbb{N}$, $a \in k^\times$ such that $\beta_m = a\alpha_n$

Proof: $\beta_m = a\alpha_n = \frac{a\alpha_n + 0}{0\alpha_n + 1}$ gives $\alpha_n \approx \beta_m$, and thus $\alpha \approx \beta$. The other directions requires a bit more work. $\alpha \approx \beta$ gives $\frac{R\alpha+S}{S\alpha+T} = \beta$. Let $\frac{P_i}{Q_i}$ be the convergents of α. We can write:

$$\beta = \frac{R\left(P_{n-1}\alpha + P_{n-2}\right) + S\left(Q_{n-1}\alpha + Q_{n-2}\right)}{T\left(P_{n-1}\alpha + P_{n-2}\right) + U\left(Q_{n-1}\alpha + Q_{n-2}\right)} \tag{17.1}$$

$$= \frac{\left(RP_{n-1} + SQ_{n-1}\right)\alpha + \left(RP_{n-2} + SQ_{n-2}\right)}{\left(TP_{n-1} + UQ_{n-1}\right)\alpha + \left(TP_{n-2} + UQ_{n-2}\right)} \tag{17.2}$$

$$= \frac{A\alpha + B}{C\alpha + D} \tag{17.3}$$

with

$$\begin{aligned} A &= RP_{n-1} + SQ_{n-1} \\ B &= RP_{n-2} + SQ_{n-2} \\ C &= TP_{n-1} + UQ_{n-1} \\ D &= TP_{n-2} + UQ_{n-2} \end{aligned}$$

We have $\left|\alpha - \frac{P_{n-1}}{Q_{n-1}}\right| < \frac{1}{|Q_{n-1}|^2}$. And thus we can write $P_{n-1} = \alpha Q_{n-1} + \delta$ with $|\delta| < \frac{1}{|Q_{n-1}|}$. We can thus write:

$$\begin{aligned} C &= \left(T\alpha + U\right)Q_{n-1} + \delta T \\ D &= \left(T\alpha + U\right)Q_{n-2} + \delta T \end{aligned}$$

If we take n big enough we have $|C| = |T\alpha + U|\,|Q_{n-1}|$ and $|D| = |T\alpha + U|\,|Q_{n-1}|$. Thus for large enough n we have $|C| > |D|$. Now we can apply lemma 4 to get an $a \in k^\times$ and $m \in \mathbb{N}$ such that $\beta_m = a\alpha_n$, proving the theorem.[80]

17.4 Möbius transformations and matrix notation

The relation in the previous section can be seen as a requirement on the existence of a certain kind of Möbius transformation. A Möbius transformation is a function of the form

$$f(x) = \frac{Ax + B}{Cx + D}$$

We limit ourselves here to those transformations for which we have $AD - BC \in k^\times$. These form a group with a left action on $k((X^{-1}))$. We claim that this group behaves like the group of 2x2 matrices over $k[X]$ with determinant in $k^\times$, from now on we will call this group $\mathrm{SL}(2, k[X])$. First we need to show that it is indeed a group. For this we need to verify only two things, the rest follows from the fact that we are working with matrices. First, let $f, g \in \mathrm{SL}(2, k[X])$. Then $\det(f\dot{g}) = (\det f)\dot{(}\det g)$. Since $k^\times$ is a group, we thus have $\det(f\dot{g}) \in k^\times$ and thus $f\dot{g} \in \mathrm{SL}(2, k[X])$. Now let $f \in \mathrm{SL}(2, k[X])$. We have

$$f^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \frac{D}{AD-BC} & -\frac{B}{AD-BC} \\ -\frac{C}{AD-BC} & \frac{A}{AD-BC} \end{pmatrix}$$

And since we have $\det f = AD - BC \in k^\times$ and $k^\times$ a group, we have $f^{-1} \in \mathrm{SL}(2, k[X])$. And thus we have $\mathrm{SL}(2, k[X])$ a group.

We now define the left action of $\mathrm{SL}(2, k[X])$ on $k((X^{-1}))$. Let $f \in \mathrm{SL}(2, k[X])$ and $\alpha \in k((X^{-1}))$. Then if

$$f = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

We define $f\alpha$ to be

$$f\alpha = \frac{A\alpha + B}{C\alpha + D}$$

This is indeed a left action because we have:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} AA' + BC' & AB' + BD' \\ CA' + DC' & CB' + DD' \end{pmatrix}$$

and

$$f(f'\alpha) = \frac{A\frac{A'\alpha+B'}{C'\alpha+D'} + B}{C\frac{A'\alpha+B'}{C'\alpha+D'} + D} = \frac{A\left(A'\alpha + B'\right) + B\left(C'\alpha + D'\right)}{C\left(A'\alpha + B'\right) + D\left(C'\alpha + D'\right)} = \frac{\left(AA' + BC'\right)\alpha + \left(AB' + BD'\right)}{\left(CA' + DC'\right)\alpha + \left(CB' + DD'\right)}$$

17.5 Pseudoperiodic continued fractions

We can extend the notion of a periodic continued fraction in the real case to the Power Field Case. We use the notation $\left[A_0, A_1, A_2, A_3, \overline{A_4, A_5, A_6}\right] = \left[A_0, A_1, A_2, A_3, A_4, A_5, A_6, \overline{A_4, A_5, A_6}\right]$. However, it turns out to be useful to slightly extend the notion of periodicity, by allowing the period to be

repeated with a factor $a \in k^\times$. We can write:
$\left[A_0, A_1, A_2, A_3, \overline{A_4, A_5}^a\right] = \left[A_0, A_1, A_2, A_3, A_4, A_5, \overline{aA_4, a^{-1}A_5}^a\right]$
In which we require the period to always be even.[80]

We can now formulate and proof the following theorem:

17.5.1 Theorem 3

Let $\alpha \in k((X^{-1}))\backslash k(X)$. Then α has a pseudoperiodic continued fraction expansion if and only if we have

$$\alpha = \frac{R\alpha + S}{T\alpha + U}$$

where

$$\begin{pmatrix} R & S \\ T & U \end{pmatrix}$$

has determinant in $k^\times$ and is not a multiple of the identity matrix.

Proof: Let α have a pseudoperiodic continued fraction. Then we can write $\alpha = \left[A_0, .., A_n, \overline{A_{n+1}, .., A_k}^a\right]$ with $n \neq k$. And thus we have $\alpha_{n+1} = \left[\overline{A_n + 1, .., A_k}^a\right]$, and $\alpha_{k+1} = a\alpha_{n+1}$. Letting $\frac{P_n}{Q_n}$ be the convergents of α we now define

$$M_l = \begin{pmatrix} P_{l-1} & P_{l-2} \\ Q_{l-1} & Q_{l-2} \end{pmatrix}$$

We then have $\alpha = M_l\alpha_l$ and thus:

$$M_{k+1}^{-1}\alpha = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} M_{n+1}^{-1}\alpha\alpha = M_{k+1}\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} M_{n+1}^{-1}\alpha$$

And we thus have a relation $\alpha = \frac{R\alpha+S}{T\alpha+U}$ with

$$\begin{pmatrix} R & S \\ T & U \end{pmatrix} = M_{k+1}\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} M_{n+1}^{-1}$$

If this matrix where a multiple of the identity matrix we would have a $b \in k^\times$ such that:

$$M_{k+1}\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} = M_{n+1}\begin{pmatrix} b & 0 \\ 0 & b \end{pmatrix}$$

However, the lower left entry of the left side is aQ_k and is thus of greater degree than the corresponding entry bQ_n on the right side of the equation, and thus this equality can hold for no $b \in k^\times$, and thus we have a relation as required by the theorem.

Now suppose that we have an α such that we have

$$\alpha = \begin{pmatrix} R & S \\ T & U \end{pmatrix} \alpha$$

with $RU - ST \in k^\times$ and

$$\begin{pmatrix} R & S \\ T & U \end{pmatrix}$$

not a multiple of the identity matrix. Then by theorem 2 we have $\alpha_n = b\alpha_m$ for some $m, n \in \mathbb{N}$ and $b \in k^\times$. We only need to show now that $m \neq n$. But the result was reached through lemma 4, which also gives $\frac{A}{C} = \frac{P_{m-1}}{Q_{m-1}}$ and $\frac{B}{D} = \frac{P_{m-2}}{Q_{m-2}}$. Suppose $m = n$, then we have $A = uP_{n-1}$, $C = uQ_{n-1}$, $B = vP_{n-2}$ and $D = vP_{n-2}$ with $u, v \in k^\times$. Substituting these into equation 17.1 gives

$$\alpha = \frac{uP_{n-1}\alpha_n + vP_{n-2}}{uQ_{n-1}\alpha_n + vQ_{n-2}}$$

since also $\alpha = M_n\alpha_n$ we have $v = u$, and thus

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = uM_n$$

and

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} R & S \\ T & U \end{pmatrix} M_n$$

But since the matrix with RSTU is not a multiple of the identity matrix, this is impossible, and thus $m \neq n$. From this it follows that α has a pseudoperiodic continued fraction expansion.[80]

17.6 Calculating continued fraction

Up until now we have only looked at abstract properties of continued fractions. But we might also be interested in calculating the continued fraction

expansion of a zero of an equation. There are two strategies that will be treated here.

First of all, we can calculate the continued fraction expansion from its power series expansion. Suppose we have the element $\alpha = \sum_{i=0}^{\infty} x^{-i}$ of $GF(2)((X^{-1}))$. We get it's continued fraction expansion by repeatedly taking the part consisting of positive powers of α_i, which is the next partial quotient, and using that to obtain α_{i+1}. In this case this gives $A_0 = 0$ and $\alpha_1 = 1 + X$, and the continued fraction expansion is thus $[0, 1 + X]$.

However, it is not always easy to obtain the power series expansion of an element, and it can be rather hard to manipulate it in certain situations. There are sometimes alternative approaches.

Suppose $\alpha, \beta \in GF(p)((X^{-1}))$, and suppose we know the continued fraction expansion of α. If we know have $\beta = \frac{R\alpha^p+S}{T\alpha^p+U}$, then we can use a method by Mills and Robbins [59].

The method of Mills and Robbins consists of an iterated process, in which during each iteration one executes on of two possible steps. These steps either produce a new element of β, or consume an element of α. The state of the algorithm after each iteration is a relation $\beta_i = \frac{R\alpha_j^p+S}{T\alpha_j^p+U}$, and we can for the rest of this piece assume that the greatest common divisor of R, S, T and U is 1.

We first need some notational conventions. We define $D = RU - ST$, and let r, s, t, u, d be the degree of R, S, T, U, D resp, and these are $-\infty$ if their corresponding polynomial is 0. When we write $[\alpha]$ this means we take the integer part of α (the positive powers of X). We can now go on to look at the actual steps of the algorithm.

17.6.1 Step of type I

Steps of type I produce an element of the continued fraction expansion of β, and can thus only be executed in the situation that no further element of the continued fraction expansion of α influences it.

Let $\beta_i = \frac{A\alpha_j^p+B}{C\alpha_j^p+D}$. We can execute a step of type I when:

- $\deg(\alpha_j) > 1$
- $t + p > u$
- $2t + p > d$

Then let $B_i = \left[\frac{R}{T}\right]$, and we have $\beta_{i+1} = \frac{T\alpha_j^p+U}{(R-B_iT)\alpha_j^p+S-B_iU}$.

17.6.2 Lemma 5

A step of type I is valid under the given conditions

Proof: $D = RU - ST$, which gives

$$\beta_i - \frac{R}{T} = \frac{R\alpha_j^p + S}{T\alpha_j^p + U} - \frac{R}{T} = \frac{-D}{T\left(T\alpha_j^p + U\right)}$$

$t + p > u$ and thus we have that $deg\left(T\left(T\alpha_j^p + U\right)\right) \geq 2t + p$. $2t + p > d$ and thus the right hand side of the equation has integer part 0. It follows that the next partial coefficient of β is $\left[\frac{R}{T}\right]$.

We now have $\beta_i = B_i + \frac{1}{\beta_i + 1}$, and thus $\beta_{i+1} = \frac{1}{\beta_i - B_i}$. Substituting for β_i we get:

$$\beta_{i+1} = \frac{1}{\frac{R\alpha_j^p + S}{T\alpha_j^p + U} - B_i} = \frac{T\alpha_j^p + U}{R\alpha_j^p + S - B_i\left(T\alpha_j^p + U\right)} = \frac{T\alpha_j^p + U}{\left(R - B_iT\right)\alpha_j^p + S - B_iU}$$

Which proves the lemma.

17.6.3 Step of type II

Steps of type II consumes an element of the continued fraction expansion α, and can always be done. Because in practical calculations one wants to calculate the required number of partial quotients as fast as possible, one will usually only want to do steps of type II if one cannot do a step of type I, though this is by no means necessary.

Let the next partial quotient of α be A_j. Then the result of a step of type II is

$$\beta_i = \frac{\left(RA_j^p + S\right)\alpha_{j+1}^p + R}{\left(TA_j^p + U\right)\alpha_{j+1}^p + T}$$

17.6.4 Lemma 6

A step of type II is valid

Proof: We have $\alpha_j = A_j + \frac{1}{\alpha_{j+1}}$. Since we are working in $GF(p)$ we have

$\alpha_j^p = A_j^p + \frac{1}{\alpha_{j+1}^p}$. Putting this in the original relation gives:

$$\beta_i = \frac{R\left(A_j^p + \frac{1}{\alpha_{j+1}^p}\right) + S}{T\left(A_j^p + \frac{1}{\alpha_{j+1}^p}\right) + U} = \frac{\left(S + RA_j^p\right)\alpha_{j+1}^p + R}{\left(U + TA_j^p\right)\alpha_{j+1}^p + T}$$

Both types of steps can easily be shown to preserve up to sign the determinant.

Now that we have both types of steps, one last thing that is useful to prove, is that it actually can only take a finite number of steps of type II to make the next step of type I possible. In other words, given enough steps, one will always get the next partial quotient. One cannot get stuck.

17.6.5 Lemma 7

It only takes a finite number of steps of type II to make a step of type I possible

Proof: Suppose that α_i is of degree 0. The definition of the complete quotients then gives that $i = 0$. A single step of type II will thus make it so that this property is satisfied, and further steps of type I and type II will preserve this property.

Suppose $t_i + p \leq u_i$. Then after a step of type II we have $u_{i+1} = t_i < u_i$. As we cannot have $u_i = -\infty \wedge t_i = -\infty$, and u_i is an integer, we can only have a finite number of steps before $t_{i+k} + p > u_{i+k}$. If we would know have another step of type II we have $t_i + p > u_i$ and thus $t_{i+1} \geq t_i + p > t_i = u_{i+1}$. Thus further steps will preserve the property $t + p > u$.

Suppose $t_i + p > u_i$ but $2t_i + p \leq d$. The previous also gave that if $t_i + p > u_i$ then $t_{i+1} > t_i$. Since t_i is an integer it takes only a finite number of steps to satisfy the property $2t_i + p > d$. By the same argument steps of type II preserve the property once it is satisfied.

There is thus need for only a finite number of steps of type II between steps of type I.

17.6.6 Calculating a continued fraction from a relation with itself

Suppose we have $\alpha \in GF(p)((X^{-1}))$, which satisfies the equation

$$\alpha = \frac{R\alpha^p + S}{T\alpha^p + U}$$

Then, as long as the method of Mills and Robbins produces every partial quotient before it is needed as input for the procedure, we can use it to calculate the continued fraction expansion of α.

For example: Let $\alpha \in GF(2)((X^{-1}))$ satisfy the equations

$$\alpha = \frac{\left(x^2 + x\right)\alpha^2 + 1}{x\alpha^2 + 1} \text{degree}\left(\alpha\right) \geq 1$$

This has determinant $D = x^2$. The degree of the determinant is thus 2. We can verify that the conditions for a type I step hold, so doing that gives $A_0 = x + 1$ and

$$\alpha_1 = \frac{\left(x\right)\alpha^2 + 1}{x}$$

Now we need to do a step of type II, giving

$$\alpha_1 = \frac{\left(x^3 + x + 1\right)\alpha_1^2 + x}{\left(x\right)\alpha_1^2}$$

We can now once again do a step of type I, giving $A_1 = x^2 + 1$ and

$$\alpha_2 = \frac{x\alpha_1^2}{\alpha_1^2 + x}$$

Again, we now need a step of type II, giving

$$\alpha_2 = \frac{\left(x^5 + x\right)\alpha_2^2 + x}{\left(x^4 + x + 1\right)\alpha_2^2 + 1}$$

Continuing this way will give us further partial quotients of α.

Chapter 18

Computing Möbius transformations

18.1 Introduction

One of the difficulties of the continued fraction of a real number is that it is rather hard to do calculations with it. This appendix will look at a technique, first described by G.N. Raney in [70], for doing Möbius transformations on the Continued Fraction Expansion of a real number.

A Möbius transformation is a function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(\alpha) = \frac{A\alpha + B}{C\alpha + D}$$

with $A, B, C, D \in \mathbb{Z}$. In order to look at a technique for applying these to the Continued Fraction Expansion of a number we need to get some preliminaries out of the way

18.2 Words

Words are finite or infinite sequences of symbols from an alphabet Σ. An alphabet is simply a set of symbols. So for example we can have $\Sigma = 1, 2, 3$ and then we can have the words 123, 1212121212121212..., etc. We also can have the empty word, that is a word consisting of 0 symbols. We denote this as ϵ. The length of a word w is denoted by $|w|$.

The set of all possible finite words over an alphabet Σ is denoted as Σ^*. Oftentimes it is usefull to consider all finite words excluding the empty one.

This set is denoted by Σ^+. Last, the set of all infinite words is denoted as Σ^ω.

If we have two words, w and v then wv denotes the concatenation of these two words, that is, the word formed by first writing down all symbols from the first word (w) and then all symbols from the second word (v). Using this notation we can now make the following definitions. If we have $x, w \in \Sigma^*$, $\mathbf{v} \in \Sigma^\omega$, then:

- x is a prefix of w iff there exists an $u \in \Sigma^*$ such that $xu = w$.
- x is a strict prefix of w iff there exists an $u \in \Sigma^+$ such that $xu = w$.
- x is a suffix of w iff there exists an $u \in \Sigma^*$ such that $ux = w$.
- x is a strict suffix of w iff there exists an $u \in \Sigma^+$ such that $ux = w$.
- x is a prefix of $\mathbf{v}$ iff there exists an $\mathbf{u} \in \Sigma^\omega$ such that $x\mathbf{u} = \mathbf{v}$.

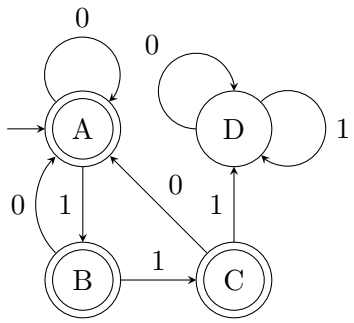
Sometimes, we don't want to talk about all words over an alphabet, but over a subset of them. Such a subset is called a Language. A language L is defined as being a set L such that $L \subset \Sigma^*$.

There is also a useful type of map on words. A morphism ϕ is a function $\phi : \Sigma^* \cap \Sigma^\omega \to \Delta^* \cap \Delta^\omega$, where Σ, Δ two alphabets, with the property that for all $u \in \Sigma^*$, $v \in \Sigma^* \cap \Sigma^\omega$ we have $\phi(uv) = \phi(u)\phi(v)$. From this definition it follows that a morphism is completely defined by its action on the symbols of Σ [2].

18.3 Finite automata

A finite automaton is the simplest model for calculations. For some languages, we can create a finite automaton that determines for a word whether or not it is an element of the language.

A finite automaton can be seen as a set of positions, called states, with instructions on where to go next from a state and the first unused symbol of a word. We can draw these in the form of diagrams:



We start in state A (given by the arrow without source on this node). Suppose we have the word 011010. We then walk through the automaton in the following way:

- We have as first unused symbol a 0. The arrow with 0 besides it going away from state A points to state A, so we are now in state A.
- We have as first unused symbol a 1. The arrow with 1 besides it going away from state A points to state B, so we are now in state B.
- We have as first unused symbol a 1. The arrow with 1 besides it going away from state B points to state C, so we are now in state C.
- We have as first unused symbol a 0. The arrow with 0 besides it going away from state C points to state A, so we are now in state A.
- We have as first unused symbol a 1. The arrow with 1 besides it going away from state A points to state B, so we are now in state B.
- We have as first unused symbol a 0. The arrow with 0 besides it going away from state B points to state A, so we are now in state A.
- We have no unused symbols left, so we are done.

We say the finite automaton accepts a word if the state in which it ends after proccesing all the symbols in that word is an accepting state. Such a state is drawn with a double edge. So the word in the example is accepted by the automaton, since state A is an accepting state. The set of all words accepted by an automaton is called the corresponding language. For this automaton, it is the set of words with no 3 subsequent 1's.

We can formalize these notions in the following way. A finite automaton M is a tuple $(Q, \Sigma, \delta, q_0, F)$ where

- Q is a finite set of states (the circles in the diagram).

- Σ is the finite alphabet of the input words.
- $\delta : Q \times \Sigma \to Q$ is the transition function (the arrows in the diagram).
- $q_0 \in Q$ is the initial state.
- $F \subset Q$ is the set of accepting states.

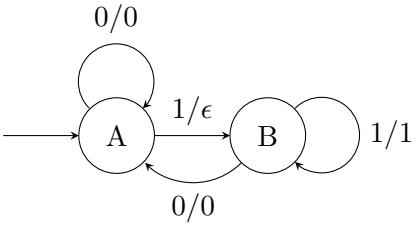
We start defining the notion of accepted words and corresponding language by extending δ. We define $\delta^* : Q \times \Sigma^* \to Q$ with the following property: Let $w \in \Sigma^*$, $a \in \Sigma$ and $q \in Q$, then:

$$\delta^*(q, \epsilon) = q,$$
$$\delta^*(q, aw) = \delta^*(\delta(q, a), w).$$

Using this, we say the automaton M accepts a word w iff $\delta^*(q_0, w) \in F$, and the corresponding language $L(M)$ is defined by $L(M) = \{w \in \Sigma^* | \delta^*(q_0, w) \in F\}$ [2].

18.4 Transducers

A finite automaton is not that practical for our purposes, since it produces only one piece of information as its output. However, we can make a variation on the finite automaton that also produces output. We take a regular finite automaton, and to every edge, we assign a word that gets concatenated to the end of the current output when that transition is taken. For example, take the following transducer:



Suppose we feed it with the word 011110110101110. Then the output will be 01110100110. Notice that we have no more accepting states. We also have no need for them anymore, since the output of the automaton is generated during its execution, not afterwards. This means that we can also execute the automaton on infinite words. This then produces a possibly infinitely long output word, depending on the automaton.

We can formalize this in the following way. A transducer T is a tuple $(Q, \Sigma, \delta, q_0, \Delta, \lambda)$ where

- Q is a finite set of states.
- Σ is the finite alphabet of the input word.
- $\delta : Q \times \Sigma \to Q$ is the transition function.
- $q_0 \in Q$ is the initial state.
- Δ is the finite alphabet of the output word.
- $\lambda : Q \times \Sigma \to \Delta^*$ is the output function.

We define the function $\lambda^* : Q \times (\Sigma^\omega \cap \Sigma^*) \to \Delta^\omega \cap \Delta^*$ with:

$$\lambda^*(q, \epsilon) = \epsilon,$$
$$\lambda^*(q, aw) = \lambda(q, a)\lambda^*(\delta(q, a), w)$$

with $q \in Q$, $a \in \Sigma$, $w \in \Sigma^\omega \cap \Sigma^*$. The output of the transducer on a word $w \in \Sigma^\omega \cap \Sigma^*$ is defined as being $\lambda^*(q_0, w)$ [2].

18.4.1 Multi-symbol input

For our purposes it is usefull to make a further generalization of the Transducer, by allowing the automaton to consume multiple symbols from the input in one transition. To make this generalization, we first assume that all our input words are of infinite length. We can then define some of the necessary concepts to formalize this.

A set of words $B \subset \Sigma^*$ is called a base iff we have that for every word $\mathbf{w} \in \Sigma^\omega$ there exists an unique element $b \in B$ that is a prefix of $\mathbf{w}$. Using this we can now formally define a Transducer with Multi-symbol input.

A Multi-symbol input Transducer T is a tuple $(Q, \Sigma, \Delta, P, q_0)$, where:

- Q is the finite collection of states.
- Σ is the input alphabet.
- Δ is the output alphabet.
- $P \subset Q \times \Sigma^* \times Q \times \Delta^*$ is the transition table.
- q_0 is the initial state.

To make this a valid Multi-symbol input Transducer, we require that $B_q = \{l|(q, l, .., ..) \in P\}$ is a base for all $q \in Q$. We now define a function ϕ', a transformation function, as the function satisfying the following equation:

$$\forall(p, u, p', v) \in P, w \in \Sigma^\omega : \phi(p, uw) = v\phi(p', w).$$

It follows from the requirement on P that this is a unique well defined function, and the output of the transducer on an input word w is defined to be $\phi(q_0, w)$ [70].

18.5 LR representation of the continued fraction expansion

Now that we have these definitions out of the way, there is only one piece missing before we can talk about Möbius transformations. The definitions of Transducers assume finite input and output alphabets. However, the default way of writing down a continued fraction does not lend itself wel to a finite alphabet. However, there is a representation of the same information that is more useful for this purpose, the LR representation.

We define two matrices L and R as:

$$L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$
$$R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

An LR word is now a sequence of L's and R's, and represents the matrix obtained by calculating the matrix product of the sequence [70].

We say that a vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $x_1, x_2 \in \mathbb{R}_{\geq 0}$, accepts a word W iff there exists a vector $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $y_1, y_2 \in \mathbb{R}_{\geq 0}$, such that $x = Wy$. We now say that a number $z \in R_{\geq 0}$ is represented by x iff $z = \frac{x_1}{x_2}$. A number $z \in R_{\geq 0}$ accepts a word W iff it's representative accepts W. This is a sensible definition because if x, x' both represent z, then $x = ax'$ with $a \in R_{\geq 0}$, and thus if one accepts W, then so does the other [70].

We now want to prove a correspondence between the LR words a number accepts, and it's continued fraction expansion.

Lemma 18.5.1. *Let x represent $\alpha \neq 1$. Then x accepts exactly one of L, R.*

Proof. First we analyse the effects of L^{-1} and R^{-1}. We have

$$L^{-1}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}x_1\\x_2-x_1\end{pmatrix}$$
$$R^{-1}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}x_1-x_2\\x_2\end{pmatrix}$$

In other words, x accepts L iff $x_2 - x_1 \geq 0$, and x accepts R iff $x_1 - x_2 \geq 0$. We now have two cases.

Case 1: $\alpha < 1$. We have $\frac{x_1}{x_2} = \alpha < 1$. Then $x_2 - x_1 > 0$, and thus x accepts L but not R.

Case 2: $\alpha > 2$. We have $\frac{x_1}{x_2} = \alpha > 1$. Then $x_1 - x_2 > 0$, and thus x accepts R but not L. □

Lemma 18.5.2. *If α accepts the LR word w, then α^{-1} accepts the LR word $\phi(w)$, where ϕ is the morphism with $\phi(L) = R$ and $\phi(R) = L$.*

Proof. We look at the basis transformation V with the matrix:

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

Now we have $V = V^{-1}$, $VLV^{-1} = R$ and $VRV^{-1} = L$. But we also have that every x accepts V, and that if x represents α, then Vx represents α^{-1}.

Thus if x represents α^{-1}, then x accepts VwV. And because of the identities above we have $VwV = \phi(w)$. Thus α^{-1} accepts $\phi(w)$. □

Lemma 18.5.3. *Let $\alpha \in \mathbb{R}\backslash\mathbb{Q}$. Then the infinitely long LR word accepted by α is unique and equal to $R^{A_0}L^{A_1}R^{A_2}L^{A_3}....$*

Proof. We know $A_0 = \lfloor\alpha\rfloor$, and also

$$R^{-n}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}x_1-nx_2\\x_2\end{pmatrix}.$$

Let x be a representative of α. Because α is irrational we know that there exists no $m, n \in \mathbb{Z}$ such that $mx_1 = nx_2$. R^{-1} and L^{-1} only create linear combinations of x_1 and x_2, and thus it follows that for any word w $w^{-1}\alpha$ represents another irrational number. Since 1 is rational it follows from Lemma 18.5.1 that if α accepts w and v and $|v| \leq |w|$, then v a prefix of w.

We have $\frac{x_1}{x_2} = \alpha \geq A_0$, from which we can conclude $x_1 \geq A_0x_2$. We also have $\frac{x_1}{x_2} = \alpha < A_0 + 1$, from which we can conclude $x_1 < (A_0 + 1)x_2$.

From these facts combined it follows that the word R^{A_0} is a prefix of the infinite word that α accepts, but R^{A_0+1} is not. We also have that R^{-A_0} represents $\frac{1}{\alpha_1}$, and we know α_1 is irrational. Lemma 18.5.2 now gives that the accepting word of this is the accepting word of α_1 with L and R switched. Using induction we now have proven the lemma. □

Lemma 18.5.4. *Let* $x = \begin{pmatrix}1\\1\end{pmatrix}$. *Then the only infinite words* x *accepts are* RL^∞ *and* LR^∞.

Proof. x accepts R with $x' = \begin{pmatrix}0\\1\end{pmatrix}$. Then by Lemma 18.5.1 x' accepts only one of R, L. Since $Lx' = x'$ x accepts the infinite word LR^∞, and no other infinite word starting with L. x also accepts L with $x'' = \begin{pmatrix}1\\0\end{pmatrix}$. Again by Lemma 18.5.1 x'' accepts only one of R, L. Since $Rx'' = x''$ x accepts the inifinite word RL^∞, and no other infinite word starting with R. Since every infinite word should start with either R or L, these are the only two options, proving the lemma. □

Lemma 18.5.5. *Let* $\alpha \in \mathbb{Q}$. *Then* $\alpha = [A_0, A_1, ..., A_n]$ *iff we have:*

- *If* n *even then* α *accepts* $R^{A_0}L^{A_1}...L^{A_n}R^\infty$.
- *If* n *odd then* α *accepts* $R^{A_0}L^{A_1}...R^{A_n}L^\infty$.

Proof. Let $[A'_0, A'_1, ..., A'_{n'}]$ be the unique continued fraction expansion of α with $A'_{n'} = 1$. We know from the introduction lectures that we have only two options for the continued fraction expansion of α, the one we constructed and $[A'_0, A'_1, ..., A'_{n'-1} + A_{n'}]$. We can use the procedure of Lemma 18.5.3 up to the point where we are left with a number that represents 1. Depending on n' we then have that every word that α accept is either a prefix of, or has as a prefix, the word $R^{A'_0}L^{A'_1}...L^{A'_{n'}}$ or $R^{A'_0}L^{A'_1}...R^{A'_{n'-1}}$. Then by Lemma 18.5.4 we have that the two unique words that are $R^{A'_0}L^{A'_1}...L^{A'_{n'-1}+1}R^\infty$ and $R^{A'_0}L^{A'_1}...L^{A'_{n'-1}}R^1L^\infty$, or $R^{A'_0}L^{A'_1}...R^{A'_{n'-1}+1}L^\infty$ and $R^{A'_0}L^{A'_1}...R^{A'_{n'-1}}L^1R^\infty$. These correspond each with one of the two continued fraction expansions of α as required by the lemma, thus proving it. □

18.6 Other 2×2 matrices over $\mathbb{N}$

In order to construct the automata to calculate the Möbius transformation it is usefull to look at and classify the other 2×2 matrices over $\mathbb{N}$. We first

split these matrices into categories depending on their determinant. We let $\mathcal{D}_n$ denote all 2×2 matrices over $\mathbb{N}$ with determinant n.

We now introduce the concept of dominance. A row or column is said to be dominant over the other if we have that it's values are all greater than or equal to those of the other row or column. A 2×2 matrix is called row-balanced (resp. column-balanced) if neither of its rows (resp columns) is dominant. We denote the set of these matrices with $\mathcal{RB}_n$ (resp $\mathcal{CB}_n$), where n denotes the determinant of all the matrices in that set. If a matrix is both row-balanced and column-balanced it is called doubly-balanced. The set of all doubly balanced matrices is denoted with $\mathcal{DB}_n$ [70].

The following facts are usefull in proving the more complicated theorems later and proving them is left as an exercise to the reader. These originate from [70].

Lemma 18.6.1. *1. $\mathcal{RB}_n$ and $\mathcal{RB}_n$ are finite sets.*

2. *$\mathcal{RB}_1 = \mathcal{CB}_1 = I$.*

3. *For $M \in \mathcal{D}_n$ the first (second) row of M is dominant iff $M \in R \cdot \mathcal{D}_n$ ($M \in L \cdot \mathcal{D}_n$).*

4. *The sets $R \cdot \mathcal{D}_n$, $L \cdot \mathcal{D}_n$ and $\mathcal{RB}_n$ are pairwise disjoint and their union is $\mathcal{D}_n$.*

5. *For $M \in \mathcal{D}_n$ the first (second) column of M is dominant iff $M \in L \cdot \mathcal{D}_n$ ($M \in R \cdot \mathcal{D}_n$).*

6. *The sets $L \cdot \mathcal{D}_n$, $L \cdot \mathcal{D}_n$ and $\mathcal{CB}_n$ are pairwise disjoint and their union is $\mathcal{D}_n$.*

7. *For every $M \in \mathcal{D}_1$ there exists exactly one word w such that the matrix represented by w equals M.*

8. *Each matrix $M \in \mathcal{D}_n$ has a unique decomposition of the form PQ with $P \in \mathcal{D}_1$ and $Q \in \mathcal{RB}_n$.*

9. *Each matrix $M \in \mathcal{D}_n$ has a unique decomposition of the form QP with $P \in \mathcal{D}_1$ and $Q \in \mathcal{CB}_n$.*

10. *If $M \in \mathcal{CB}_n$ and $M = PQ$ with $P \in \mathcal{D}_1$ and $Q \in \mathcal{RB}_n$, then $Q \in \mathcal{DB}_n$.*

11. *If $M \in \mathcal{RB}_n$ and $M = QP$ with $P \in \mathcal{D}_1$ and $Q \in \mathcal{CB}_n$, then $Q \in \mathcal{DB}_n$.*

18.7 Enumerating matrices in $\mathcal{RB}_n$, $\mathcal{CB}_n$ and $\mathcal{DB}_n$.

We now introduce a formalism useful in enumerating balanced matrices, and showing that they remain balanced after certain transformations. Let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathcal{D}_n$. We define $r(M) = \begin{pmatrix} p \\ q \end{pmatrix}$, where $p = d - b$ and $q = a - c$. When M is row-balanced both p and q are positive. In this case we denote the generating word of $r(M)$ with W_M [70].

We then have the following identities:

Lemma 18.7.1. *For every $M \in \mathcal{D}_n$ we have $r(ML) = L^{-1}r(M)$ and $r(MR) = R^{-1}r(M)$.*

Proof. A straightforward calculation shows this. □

Lemma 18.7.2. *[70] For every $M \in \mathcal{D}_n$ and LR-word W we have: $r(Mw) = (w)^{-1}r(M)$.*

Proof. From linear algebra we know $(AB)^{-1} = B^{-1}A^{-1}$. The lemma follows from repeatedly applying Lemma 18.7.1. □

From the definition we also know that if $r(M) = \begin{pmatrix} g \\ g \end{pmatrix}$, that the matrix is of the form $M = \begin{pmatrix} s' + g & s \\ s' & s + g \end{pmatrix}$, with $g(s' + s + g) = n$. We now introduce two more definitions. A triple (g, s, s') of non-negative integers will be called a (*)-triple. A (*)-triple is associated with a matrix M if and only if there exists a word W such that $MW = \begin{pmatrix} s' + g & s \\ s' & s + g \end{pmatrix}$.

Theorem 20. *[70] For every $M \in \mathcal{RB}_n$ there is exactly one (*)-triple (g, s, s') associated with M and for this triple the equation $MW_M = \begin{pmatrix} g + s' & s \\ s' & g + s \end{pmatrix}$ holds.*

Proof. We have $r(M) = W_m \begin{pmatrix} g \\ g \end{pmatrix}$. By Lemma 18.7.2 we then get $r(MW_M) = \begin{pmatrix} g \\ g \end{pmatrix}$. It now follows that there exists a (*)-triple (g, s, s') with $MW_M = \begin{pmatrix} g + s' & s \\ s & g + s \end{pmatrix}$. This (*)-triple is associated with M. Suppose that a (*)-triple (g_1, s_1, s_1') is associated with M. Then by Lemma 18.7.2 we know

that there exists a word W_1 such that $W_1 \begin{pmatrix} g_1 \\ g_1 \end{pmatrix} = r(M) = W_M \begin{pmatrix} g \\ g \end{pmatrix}$. From this equality and the properties of LR-words it follows that $W_1 = W_M$ and $g_1 = g$. But then it also follows that $s_1 = s$ and $s'_1 = s'$. □

Theorem 21. *[70] For every (*)-triple (g, s, s') there is exactly one matrix $Q \in \mathcal{DB}_n$ whose associated (*)-triple is (g, s, s'). The matrices $M \in \mathcal{RB}_n$ having (g, s, s') as their associated (*)-triple are precisely those of the form $M = QU$ with U an LR-word that is a prefix of W_Q. Furthermore we have $W_Q = UW_M$.*

Proof. Let (g, s, s') be a (*)-triple. We have $\begin{pmatrix} g+s' & s \\ s' & g+s \end{pmatrix} \in \mathcal{RB}_n$ and thus (by Lemma 18.6.1 parts 7,9 and 11) there exist $Q \in \mathcal{DB}_N$ and an LR-word U such that $\begin{pmatrix} g+s' & s \\ s' & g+s \end{pmatrix} = QW$. Because of this equation we know that Q has (g, s, s') as it's associated (*)-triple, and by Theorem 20 that $W = W_Q$.

Now let $M \in \mathcal{RB}_n$ with (g, s, s') as it's associated (*)-triple. Then $MW_M = \begin{pmatrix} g+s' & s \\ s' & g+s \end{pmatrix} = QW_Q$. By Lemma 18.6.1 parts 7,9 and 11 we have $M = Q'U$ where U is an LR-word. This gives $Q'UW_M = QW_Q$ and this implies that $Q' = Q$ and $W_Q = UW_M$. If M was in fact an element of $\mathcal{DB}_n$ then this gives, in combination with 18.6.1 part 9, that $M = Q$. Thus Q is the only matrix in $\mathcal{DB}_n$ with (g, s, s') as it's associated triple. □

Lemma 18.7.3. *[70] If $M \in \mathcal{RB}_n$ and V an LR-word, then $MV \in \mathcal{RB}_n$ iff V is a prefix of W_M.*

Proof. From Theorem 21 we know that we have $Q \in \mathcal{DB}_n$ associated to the same (*)-triple as M, and that $MW_M = QUW_M = QW_Q$. If V is a prefix of W_M, the theorem gives immediately that $MV \in \mathcal{RB}_n$. If $MV \in \mathcal{RB}_n$, then we know $MVW_{MV} = QW_Q$, and since $W_Q = UW_M$ we get that V is a prefix of W_M, which completes the proof. □

18.8 Transformations on row balanced matrices

We now have the tools to go towards transformations on row balanced matrices. We need one final definition. Let $M \in \mathcal{RB}_n$. Define the base B_M to be the base formed by taking all words w such that every strict prefix of w is a prefix of W_M, and w itself is not a prefix of W_M. This means that the only difference between W_M and w is in the last symbol of w [70].

Theorem 22. *[70] Let $M_1 \in \mathcal{RB}_n$ and $V_1 \in B_{M_1}$. Then if V_1 ends on an L, we have $M_1V_1 \in L \cdot \mathcal{CB}_n$, and if V_1 ends on an R, we have $M_1V_1 \in R \cdot \mathcal{CB}_n$. Furthermore, there exists also an LR-word V_2 and a matrix $M_2 \in \mathcal{DB}_n$ such that $M_1V_1 = V_2M_2$.*

Proof. We start by proving the first conclusion. For this, we seperate the proof into 4 distinct cases.

Case 1: $V_1 = W_{M_1}L$.
In this case we have

$$\begin{aligned} M_1V_1 &= M_1W_{M_1}L \\ &= \begin{pmatrix} g+s' & s \\ s' & g+s \end{pmatrix} L \\ &= \begin{pmatrix} s'+g+s & s \\ s'+s+g & s+g \end{pmatrix} \\ &= L \begin{pmatrix} s'+g+s & s \\ 0 & g \end{pmatrix} \in L \cdot \mathcal{CB}_n. \end{aligned}$$

Case 2: $V_1 = W_{M_1}R$.
In this case we have

$$\begin{aligned} M_1V_1 &= M_1W_{M_1}R \\ &= \begin{pmatrix} g+s' & s \\ s' & g+s \end{pmatrix} R \\ &= \begin{pmatrix} s'+g & s'+g+s \\ s' & s'+s+g \end{pmatrix} \\ &= R \begin{pmatrix} g & 0 \\ s' & s'+s+g \end{pmatrix} \in R \cdot \mathcal{CB}_n. \end{aligned}$$

Case 3: $W_{M_1} = URZ$ for some LR-words U and Z, and $V_1 = UL$.
First let $Z = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$. Since $Z \in \mathcal{D}_1$ we have $Z^{-1} = \begin{pmatrix} w & -y \\ -z & x \end{pmatrix}$. It now follows that

$$\begin{aligned} M_1U &= M_1W_{M_1}Z^{-1}R^{-1} \\ &= \begin{pmatrix} s'+g & s \\ s' & s+g \end{pmatrix} \begin{pmatrix} w & -y \\ -z & x \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} s'w+gw-sz & -s'w-s'y-gw-gy+sz+sx \\ s'w-sz-gz & -s'w-s'y+sz+sx+gz+gx \end{pmatrix}. \end{aligned}$$

Since the elements of M_1U are nonnegative we get $s'w - sz - gz \geq 0$ and $s'w + gw - sz > 0$. These facts can now be used to obtain

$$\begin{aligned} M_1V_1 &= M_1UL \\ &= \begin{pmatrix} -s'y - gy + sx & -s'w - s'y - gw - gy + sz + sx \\ -s'y + sx + gx & -s'w - s'y + sz + sx + gz + gx \end{pmatrix} \\ &= L \begin{pmatrix} -s'y - gy + sx & -s'w - s'y - gw - gy + sz + sx \\ g(x+y) & g(z+x+w+y) \end{pmatrix} \in L\mathcal{CB}_n. \end{aligned}$$

Case 4: $W_{M_1} = ULZ$ for some LR-words U and Z, and $V_1 = UR$. Again take $Z = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$. We get:

$$\begin{aligned} M_1U &= M_1W_{M_1}Z^{-1}L^{-1} \\ &= \begin{pmatrix} s'w + s'y + gw + gy - sz - sx & -s'y - gy + sx \\ s'w + s'y - sz - sx - gz - gx & -s'y + sx + gx \end{pmatrix}. \end{aligned}$$

Again, this matrix has no negative elements, and thus $-s'y - gy + sx \geq 0$ and $-s'y + sx + gx > 0$. We then get

$$\begin{aligned} M_1V_1 &= M_1UR \\ &= R \begin{pmatrix} g(w+y+z+x) & g(w+z) \\ s'w + s'y - sz - sx - gz - gx & s'w - sz - gz \end{pmatrix} \in R\mathcal{CB}_n. \end{aligned}$$

We now only need to prove the second part. However, it follows from 18.6.1 parts 7, 9 and 11 that $R\mathcal{CB}_n \in (\mathcal{D}_1\mathcal{DB}_n)$, and that there thus exists an unique LR-word V_2 and a matrix $M_2 \in \mathcal{DB}_n$ such that $M_1V_1 = V_2M_2$. □

18.9 Transducers for Möbius transformations

Using the results from the previous section we can now build transducers that calculate certain Möbius transformations.

Let Q be the set of all matrices $M \in \mathcal{DB}n$ which have as greatest divisor of their elements ν. Then let P be the set of all tuples (M_1, V_1, M_2, V_2) with $M_1, M_2 \in Q$, $V_1 \in B_{M_1}$ and $M_1V_1 = V_2M_2$ [70].

Theorem 23. *[70] $T = (Q, \{L, R\}, \{L, R\}, P, M)$ is a transducer $\forall M \in Q$, which produces the result of the Möbius transformation specified by M on it's input.*

Proof. First since all matrices in $\mathcal{D}_1$ are either the unit matrix or can be written as an LR word, we get that if $M_1V_1 = V_2M_2$ for $M_1, M_2 \in \mathcal{D}_n$ and $V_1, V_2 \in \mathcal{D}_1$ that the greatest common divisor of the elements of M_1 is equal to the greatest common divisor of the elements of M_2.

With this fact it follows from Lemma 18.6.1 and Theorem 22 that P does satisfy the requirements from our definition of a multi-symbol input transducer. Let $x_1, x_2 \in R_{\geq 0}$ be such that $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ accepts the LR word given as input to the transducer. It then follows from the definition of P that Mx accepts the output word of the transducer. But Mx represents the result of the Möbius transformation on the number represented by x. Thus our transducer performs the Möbius transformation M, proving the theorem. □

18.10 Conclusion

The result from the previous section gives a method for calculating certain Möbius transformations, on continued fraction expansions representing positive numbers. In the original work by G.N. Raney [70] he also presents a set of steps by which this method can be extended to continued fraction expansions of arbitrary numbers and arbitrary Möbius transformations. These results show that Möbius transformations can be relativly easily calculated on the continued fraction expansion of a number.

Bibliography

[1] Adler, R., M.S. Keane and M. Smorodinsky – *A construction of a Normal Number for the Continued Fraction Transformation,* J. of Number Th. **13** (1981), 95-105.

[2] Allouche, Jean-Paul and Jeffrey Shallit, *Automatic sequences: theory, applications, generalizations*, Cambridge University Press 2003.

[3] Austin, D. – *Trees, Teeth, and Time: The mathematics of clock making*, `http://www.ams.org/samplings/feature-column/fcarc-stern-brocot`, Accessed 29 January 2013.

[4] Bagemihl, F. and J.R. McLauglin – *Generalization of some classical theorems concerning triples of consecutive convergents to simple continued fractions*, J. reine Angew. Math. **221** (1966), 146-149.

[5] Barbolosi, D. and H. Jager – *On a theorem of Legendre in the theory of continued fractions*, Sém. Th. Nombres Bordeaux **6** (1994), 81-94.

[6] Barrionuevo, Jose, Robert M. Burton, Karma Dajani and Cor Kraaikamp – *Ergodic Properties of Generalized Lüroth Series*, Acta Arithm., **LXXIV** (4) (1996), 311-327.

[7] Billingley, P. – *Ergodic Theory and Information*, John Wiley and Sons, 1965.

[8] Billingley, P. – *Probability and Measure,* John Wiley and Sons, 2nd Ed. 1986.

[9] Blanchard, F. – *β-Expansions and Symbolic Dynamics,* Theoretical Comp. Sc., **65** (1989), 131-141.

[10] Bogomolny, A. – *Stern-Brocot Tree. Introduction* from Interactive Mathematics Miscellany and Puzzles,

http://www.cut-the-knot.org/blue/Stern.shtml, Accessed 29 January 2013

[11] Bogomolny, A. – *Stern-Brocot Tree, a second look at the binary encoding* from Interactive Mathematics Miscellany and Puzzles http://www.cut-the-knot.org/blue/chaos_game.shtml#tree, Accessed 27 January 2013

[12] Borel, É. – *Contribution à l'analyse arithmétique du continu*, J. Math. Pures Appl. (5) **9** (1903), 329-375.

[13] Bosma, W. and D. Gruenewald – *Complex numbers with bounded partial quotients*, to appear in Journal of the Australian Mathematical Society.

[14] Bosma, W., H. Jager and F. Wiedijk – *Some metrical observations on the approximation by continued fractions*, Indag. Math. **45** (1983), 281-299.

[15] Boyd, David W. – *On the beta expanson for Salem numbers of degree 6,* Math. Comp. **65** (1996), 861-875, S29-S31.

[16] Brauer, A. – *On algebraic equations with all but one root in the interior of the unit circle,* Math. Nachr. **4** (1951), 250-257.

[17] Bressoud, D.M. – *Factorization and Primality Testing*, Springer UTM, Spriner Verlag, Berlin, New York, 1989.

[18] Brown, James R. – *Ergodic Theory and Topological Dynamics*, Academic Press, New York, San Francisco, London, 1976.

[19] Champernowne, D.G. – *The construction of decimal normal in the scale of ten,* J. London Math. Soc., **8** (1933), 254-260.

[20] Cornfeld, I.P., S.V. Fomin and Ya.G. Sinai – *Ergodic Theory,* Grundlehren der math. Wiss. **245**, Springer-Verlag New York, Heidelberg, Berlin (1982).

[21] Dajani, K., C. Kraaikamp – *Ergodic Theory of Numbers*, Mathematical Association of America, 2002.

[22] Dajani, K., C. Kraaikamp and B. Solomyak – *The natural extension of the β-transformation,* Acta Math. Hungar., **73** (1-2) (1996), 97-109.

[23] Davenport, H. – *The higher arithmetic. An introduction to the theory of numbers,* Sixth edition. Cambridge University Press, Cambridge, 1992.

[24] Davenport, H. and Erdös, P. – *Note on normal decimals,* Canadian J. Math. **4** (1952). 58-63. MR 13,825g

[25] Erdös, P., I. Joo and V. Komornik – *Characterization of the unique expansions* $1 = \sum_{i=1}^{\infty} q^{-n_i}$ *and related problems,* Bull. Soc. Math. France **118** (1990), (3), 377–390. MR 91j:11006

[26] William Feller – *An Introduction to Probability Theory and Its Applications, II,* John Wiley and Sons, 1966.

[27] Ford, L.R. – *Fractions* The American Mathematical Monthly, Vol. 45, No. 9, pp. 586-601. Mathematical Association of america, November 1938.

[28] Friedman, N.A. and D.S. Ornstein – *On isomorphisms of weak Bernoulli transformations,* Adv. in Math. **5** (1970), 365-390.

[29] Frougny, C. and B. Solomyak – *Finite beta-expansions,* Ergod. Th. & Dynam. Sys. **12** (1992), 713-723.

[30] Galambos, J. – *Representations of Real numbers by Infinite Series,* Springer LNM **502**, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[31] Gauss, C.F. – *Mathematisches Tagebuch 1796-1814,* Akademische Verlagsgesellschaft Geest & Portig K.G., Leipzig 1976.

[32] Hardy, G. H. and E.M. Wright – *An introduction to the theory of numbers,* Fifth edition. The Clarendon Press, Oxford University Press, New York, 1979. MR 81i:10002

[33] Havinga, E. and W.E. van Wijk and J.F.M.G. d'Aumerie – *Planetariumboek Eise Eisinga*, Arnhem, 1928, 382–410. Also accessible via `http://adcs.home.xs4all.nl/Huygens/21/plan.v.html` (Accessed 3 February 2013)

[34] Hayes, B. – *On the Teeth of Wheels*, American Scientist, July-August 2000, **88**-4, 296–300

[35] Hensley, D. – *The Hurwitz complex continued fraction*, preprint, January 2006.

[36] Irwin, M.C. – *Geometry of Continued Fractions* The American Mathematical Monthly, Vol. 96, No. 8, pp. 696-703. Mathematical Association of america, October 1989.

[37] J1 Jager, H. – *The distribution of certain sequences connected with the continued fraction,* Indag. Math. **48** (1986), no. 1, 61–69. MR 87g:11092

[38] Jager, H. – *Continued Fractions and Ergodic Theory,* Trancendental Numbers and related Topics, RIMS Kokyuroku **599**, Kyoto University, Kyoto, Japan (1986), 55-59.

[39] Jager, H. and C. Kraaikamp – *On the approximation by continued fractions*, Indag. Math. **51** (1989), 289-307. MR 90k:11084

[40] Jager, H. and C. de Vroedt – *Lüroth series and their ergodic properties,* Indag. Math. **31** (1968), 31-42. MR 39 #157

[41] Kakutani, Shizuo – *Induced measure preserving transformations,* Proc. Imp. Acad. Tokyo **19**, (1943), 635–641. MR 7,255f

[42] Kamae, T. – *A simple proof of the ergodic theorem using non-standard analysis,* Israel J. Math. **42** (1982), 284-290.

[43] Katznelson, Y. and B. Weiss – *A simple proof of some ergodic theorems,* Israel J. Math. **42** (1982), 291-296.

[44] Kesseböhmer and Stratmann – *Multifractal analysis for Stern-Brocot intervals*, J. reine angew. Math. **605** (2007), 133–163

[45] Khintchine, A.Ya. – *Continued Fractions,* Groningen: Noordhoff, 1963.

[46] Kitchens, Bruce P. – *Symbolic Dynamics,* Springer Universitext, Springer-Verlag Berlin Heidelberg New York, 1998.

[47] Kingman, J.F.C. and S.J. Taylor – *Introduction to measure and probability,* Cambridge University Press, Cambridge, 1966.

[48] Kolmogorov, A.N. – *A new metric invariant of transitive dynamical systems and Lebesgue space automorphisms,* Dokl. Acad. Sc. USSR **119**, no. 5, (1958), 861-864.

[49] Kraaikamp, C. – *On the approximation by continued fractions, II*, Indag. Math. New Series **1** (1990), 63-75.

[50] Kraaikamp, C. – *A new class of continued fraction expansions*, Acta Arithm., **LVII** (1) (1991), 1-39.

[51] Krengel, Ulrich – *Ergodic theorems,* de Gruyter Studies in Mathematics, 6. Walter de Gruyter & Co., Berlin-New York, 1985. MR 87i:28001

[52] Lenstra, A.K., H.W. Lenstra and L. Lovász – *Factoring Polynomials with Rational Coefficients*, Mathematische Annalen 1982, pp. 515-534. Springer-Verlag

[53] Lind, D. and B. Marcus – *Symbolic dynamics and coding,* Cambridge University Press 1995.

[54] Lochs, G. - *Vergleich der Genauigkeit von Dezimalbruch und Kettenbruch*, Abh. Math. Sem. Hamburg, **27** (1964), 142-144.

[55] Lochs, G. - *Die ersten 968 Kettenbruchnenner von* π, Monatsh. Math. **67** (1963), 311-316.

[56] Lüroth, J. – *Ueber eine eindeutige Entwickelung von Zahlen in eine unendliche Reihe,* Math. Annalen **21** (1883), 411-423.

[57] Mané, Ricardo – *Ergodic theory and differentiable dynamics,* Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 8. Springer-Verlag, Berlin-New York, 1987. MR 88c:58040

[58] Martin, Nathaniel F.G. and James W. England – *Mathematical Theory of Entropy,* Encyclopedia of Mathematics and its Applications, 12. Addison-Wesley Publishing Co., Reading, Mass., 1981. MR 83k:28019

[59] Mills, W.H. and D.P. Robbins – *Continued fractions for certain algebraic power series*, Journal of Number Theory **23**-3 (1986), 388–404.

[60] Nakada, H. – *Metrical Theory for a Class of Continued Fraction Expansions and their Natural Extensions*, Tokyo J. Math. **4** (1981), 399-426.

[61] Nakada, H., Sh. Ito and S. Tanaka – *On the invriant measure for the transformations associated with some real continued fractions*, Keio Engeneering Reports **30** (1977), 159-175.

[62] Danny Oorburg – *Een onderzoek naar het LLL- en Kettingbreukalgoritme*, Groningen 1997.

[63] Oxtoby, J.C. – *Measure and Category,* Springer GTM **2**, Springer-Verlag New York Heidelberg Berlin, 1971.

[64] Parry, W. – *On the* β*-expansion of real numbers,* Acta Math. Acad. Sci. Hungary **11** (1960), 401-416.

[65] Perron, O. – *Die Lehre von den Kettenbrüchen, Band I*, B.G. Teubner, Stuttgart, 3. verb. u. erw. Aufl.

[66] Perron, O. – *Irrationalzahlen,* Walter de Gruyter & Co., Berlin, 1960.

[67] Petersen, Karl – *Ergodic Theory,* Corrected reprint of the 1983 original. Cambridge Studies in Advanced Mathematics, 2. Cambridge University Press, Cambridge, 1989. MR 92c:28010

[68] Pollicott, Mark and Michiko Yuri – *Dynamical Systems and Ergodic Theory,* London Mathematical Society Student Texts **40**, Cambridge University Press 1998.

[69] Hans Rademacher, *Higher Mathematics from an Elementary Point of View*. Birkhauser, 1983. Chapter 8: Ford Circles

[70] Raney, G.N. – *On continued fractions and finite automata*, Mathematische Annalen, **206**-4 (1973), 265–283.

[71] Rényi, A. – *Representations for real numbers and their ergodic properties,* Acta Math. Acad. Sci. Hungary **8** (1957), 401-416.

[72] Rockett, A.M. and P. Szüsz – *Continued Fractions,* Singapore: World Scientific, 1992.

[73] Rohlin, V.A. – *Exact endomorphisms of a Lebesgue space,* Izv. Akad. Naik SSSR, Ser. Mat., **24** (1960); English AMS translation, Series **2**, **39** (1969), 1-36.

[74] Royden, H.L. – *Real Analysis,* Collier MacMillan International Editions, 2nd Ed., 1968.

[75] Rudin, Walter – *Real and Complex Analysis,* McGraw-Hill Book Company, 3rd Ed., 1986.

[76] Rudolph, Daniel J – *Fundamentals of measurable dynamics. Ergodic theory on Lebesgue spaces,* Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1990. MR 92e:28006

[77] Salem, R. – *Algebraic Numbers and Fourier Analysis,* Duke Math. LJ. **12** (1945), 153-172.

[78] Schmidt, K. – *On periodic expansions of Pisot numbers and Salem numbers,* Bull. London Math. Soc. **12** (1980), 269-278.

[79] Schmidt, W.M. – *Diophantine Approximation*, Springer LNM **785**, 1980.

[80] Schmidt, W.M. – *On continued fractions and diophantine approximation in power series fields*, Acta Arithmetica **95**-2 (2000), 139–166.

[81] Fritz Schweiger – *Ergodic Theory of Fibered Systems and Metric Number Theory,* Clarendon Press, Oxford 1995.

[82] Series, Caroline – *Non-Euclidean geometry, continued fractions, and ergodic theory,* Math. Intelligencer 4 (1982), no. 1, 24–31. MR 84h:58086

[83] Series, Caroline – *The geometry of Markoff numbers,* Math. Intelligencer **7** (1985), 20–29. MR 86j:11069

[84] Segre, B. – *Lattice points in infinite domains and asymmetric Diophantine approximation*, Duke J. Math. **12** (1945), 337-365.

[85] Shannon, C. – *A mathematical theory of communication,* Bell Syst. Tech. J. **27** (1948), 379-423, 623-656.

[86] Short, Ian – *Ford circles, Continued Fractions, and Rational Approximation* The American Mathematical Monthly, Vol. 118, No. 2, pp. 130-135. Mathematical Association of america, February 2011.

[87] Swinden, J.H. van – *Beschrijving van het Eisinga planetarium*, uitgeverij van Wijnen Franeker, 1994. ISBN: 90 5194 105 6

[88] Tong, Jingcheng – *Approximation by nearest integer continued fractions,* Math. Scand. **71** (1992), 161-166.

[89] Vitányi, Paul – *Randomness,* CWI Quarterly **8** (1995), 67–82. MR 97c:01032

[90] Walters, P. – *An Introduction to Ergodic Theory,* GTM **79**, Springer-Verlag New York, Heidelberg, Berlin (1982).