Observations on continued fractions from Ford's point of view

Raul Hindov

September 16, 2016

Abstract

This is a project on a geometric method to visualise the calculation of continued fractions, which was developed by Lester R. Ford in the 1930s. The description of the method is followed by the use of the technique of Ford circles on three distinct cases. In contrast to a continued fraction with positive integer coefficients, which is always convergent, a continued fraction with real coefficients might diverge. This project provides an alternative geometric proof for one of the classical convergence theorems on continued fractions.

Among the integer factorization methods there is one that uses continued fractions. We give an alternative geometric proof for the lemma that is the base for the method and for a corollary that reduces the factoring problem by a factor of 2.

Continued fractions are also connected to the subject of Diophantine approximation, and the project presents an alternative geometric proof for the Dirichlet's approximation theorem.

We assume the reader to be a mathematics student who has taken a course on Number Theory and on Analytic Functions, and has certain amount of time on her hands to read the paper.

Acknowledgements

I would like to thank my supervisor, Sandra Pott, for critique and advice while writing the degree project.

Contents

Introduction

We shall study continued fractions considering them as functions of finitely or infinitely many variables. If the variables are positive integers then everything is straightforward, the continued fraction outputs an unique irrational number in the infinite case and a rational number in the finite case. We can also find the unique pre-image for any rational number as well as for any irrational given for example by its decimal expansion.
The calculation of continued fractions gets more complicated if we allow the variables to be not just positive integers but also negative or even any real or complex numbers. Then we face the question of convergence of the continued fraction, as the possibility for divergence arises and the continued fraction might no longer be well-defined.
Similarly the calculation of pre-images is not unique any more if we allow the variables to be in any larger set than that of the positive integers, and we have to make choices.
In order to understand and reason about the convergence and divergence we describe a geometric method developed by Lester R. Ford to visualise the calculation of continued fractions.
In the first chapter the plan is to develop the geometric picture of continued fractions to prepare the ground, paved with numerous examples and illustrations, for proving two classical theorems about convergence of continued fractions. Both proofs are inspired by the work of Ford. In one of them we follow the reasoning given by Beardon and Short, and the other is ours.
At the end of the first chapter we examine the problem of finding the inverse images of continued fractions, i.e. given a real or complex number, constructing its continued fraction representation. By restricting or expanding the domain for the terms in the continued fraction, we have a choice of various algorithms beside the classical simple continued fraction, e.g. nearest integer continued fraction.
One of the applications of the continued fractions in number theory is factoring integers. With the help of the continued fractions we can make the task much easier, reducing the problem of factoring a number with d digits to the problem of factoring a set of numbers with $\frac{d}{2}$ digits.
In the second chapter we attempt to convey the gist of what is involved in the continued fraction method of factorization, and give an alternative proof of the lemma on continued fractions that is the base for the continued fraction factorization method. We prove a corollary that reduces the factorization problem by a factor of 2.
In the third chapter we prove Dirichlet's approximation theorem using Ford and Short circles.

Historical background

Continued fractions have a long and fascinating history, we can find examples throughout mathematics in the last 2000 years.
The Euclidean algorithm for computing the greatest common divisor of two numbers, say a and b, described in his *Elements* (c. 300 BC), actually also produces the continued fraction representations for both $\frac{a}{b}$ and $\frac{b}{a}$.
The Indian mathematician Aryabhata (476–550) attempted to solve linear Diophantine equations $ax + by = c$, where a, b, c are given integers. His technique was somewhat connected to the properties of continued fractions. The key of solving the equation is to consider first $ax + by = 1$ and use the feature of the continued fractions that $\frac{a}{b}$ has its last two convergents $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_n}{q_n} = \frac{a}{b}$ related by:

$$aq_{n-1} - bp_{n-1} = (-1)^{n-1}.$$

Fibonacci describes in his book *Liber Abaci* (1202) how to express fractions in unit fractions in ascending fashion:

$$\frac{13}{30} = \frac{1 + \frac{1+\frac{1}{5}}{4}}{3}.$$

The proper theory of continued fractions began with Rafael Bombelli. In his book *L'Algebra Opera* (1572) there is a method to find square roots by means of infinite continued fractions. He proved for example:

$$\sqrt{13} = 3 + \cfrac{4}{6 + \cfrac{4}{6 + \cfrac{4}{6 + \cfrac{4}{\ddots}}}}.$$

William Brouncker established in 1655 the following identity:

$$\frac{4}{\pi} = 1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{\ddots}}}}.$$

John Wallis introduced in his 1655 book *Arithmetica Infinitorum* the term "continued fraction" for the first time and explained how to calculate them. Before him the term "anthyphairetic ratios" was used.
Christian Huygens (1629–1695) found a way to use continued fractions in a practical application. He used continued fractions for approximating gear ratios in the building of a table top planetarium with the planets Mercury, Venus, Earth, Mars, Jupiter and Saturn.
The theory of continued fractions was developed extensively the 18th and 19th centuries.
Leonhard Euler provided in his *De fractionibus continuis dissertatio* (1737) a then-comprehensive account of the properties of continued fractions, and included the first proof that the number e is irrational, using continued fractions.

Euler describes in *Introductio in analysin infinitorum* (1748) the connection between continued fractions and infinite series, proved that every rational number can be written as a finite continued fraction, and proved that the continued fraction of an irrational number is infinite. He also found the continued fraction representation of e itself which is not periodic but still has a pattern:

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, \ldots, 1, 1, 2n, 1, \ldots].$$

In 1761 Johann Heinrich Lambert gave the first proof of the irrationality of π using a continued fraction for $\tan(x)$.
In 1768 Joseph Louis Lagrange provided the general solution to Pell's equation $x^2 - dy^2 = 1$ using the continued fractions expansion of $\sqrt{d}$.
Lagrange proved in 1770 that the expansion of a real number as a simple continued fraction is (eventually) periodic if and only if it is the real root of a quadratic equation in one variable with rational coefficients. In 1869 Carl Gustav Jacob Jacobi tried to generalize this to cubic irrationals by means of a mixed continued fraction for pairs of real numbers but did not succeed.
In 1798 Adrien-Marie Legendre proved that if the rational number $\frac{p}{q}$ satisfies

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

then $\frac{p}{q}$ is one of the convergents in the continued fraction representation of an irrational α and Karl Theodor Vahlen proved in 1895 that at least one of every two consecutive convergents to α satisfies the above inequality.
Carl Friedrich Gauss derived in *Werke* (1813) a class of complex-valued continued fractions using the hypergeometric function. An application of the technique gives for example:

$$\frac{\sqrt{\pi}}{2} e^{z^2} \operatorname{erf}(z) = \cfrac{z}{1 - \cfrac{z^2}{\frac{3}{2} + \cfrac{z^2}{\frac{5}{2} - \cfrac{\frac{3}{2} z^2}{\ddots}}}}.$$

In 1829 Évariste Galois proved that if a quadratic equation in one variable with rational coefficients has a root x_1 whose continued fraction representation is purely periodic, then the other root x_2 is such that $\frac{-1}{x_2}$ is also purely periodic given by the inverse of the period of x_1.
Peter Gustav Lejeune Dirichlet proved in 1842 that for any real number α and positive integer N, there exist integers p, q such that $1 \leq q < N$ and $|\alpha q - p| \leq \frac{1}{N}$. His proof is based on the pigeonhole principle.
In 1844 Joseph Liouville proved the existence of transcendental numbers by a construction using continued fractions.
Adolf Hurwitz proved in 1891 that for every irrational number α there are infinitely many distinct rational numbers $\frac{p}{q}$ with:

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

and Émile Borel proved in 1903 that at least one of every three consecutive convergents to α satisfies the above inequality.

Hermann Minkowski defined in 1904 the question mark function

$$?(x) = b_0 + 2\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2^{b_1+\ldots+b_n}}$$

where $[b_0; b_1, b_2, \ldots]$ is the continued fraction representation of an irrational number x.
Srinivasa Ramanujan had considerable interest in evaluating infinite continued fractions and expanding functions in continued fractions, his notebooks contain about 200 results on continued fractions. One example of his continued fraction expansions in his letter to G. H. Hardy in 1913:

$$e^{\frac{2\pi}{\sqrt{5}}} \left(\frac{\sqrt{5}}{1 + \sqrt[5]{5^{\frac{3}{4}}(\frac{\sqrt{5}-1}{2})^{\frac{5}{2}} - 1}} - \frac{\sqrt{5}+1}{2} \right) = \cfrac{1}{1 + \cfrac{e^{-2\pi\sqrt{5}}}{1 + \cfrac{e^{-4\pi\sqrt{5}}}{1 + \cfrac{e^{-6\pi\sqrt{5}}}{1 + \ddots}}}}.$$

Ramanujan also found continued fraction representations for the Riemann zeta function $\zeta(z)$ at $z = 2$ and $z = 3$:

$$\zeta(2) = 1 + \cfrac{1}{1 + \cfrac{1^2}{1 + \cfrac{1 \cdot 2}{1 + \cfrac{2^2}{1 + \cfrac{2 \cdot 3}{1 + \cfrac{3^2}{1 + \ddots}}}}}} \text{ and } \zeta(3) = 1 + \cfrac{1}{4 + \cfrac{1^3}{1 + \cfrac{1^3}{12 + \cfrac{2^3}{1 + \cfrac{2^3}{20 + \cfrac{3^3}{\ddots}}}}}}$$

Aleksandr Yakovlevich Khinchin proved in 1935 that the geometric mean of the partial quotients of continued fractions:

$$\lim_{n\to\infty} (b_1 b_2 \ldots b_n)^{1/n} = \prod_{m=1}^{\infty} \left(1 + \frac{1}{m(m+2)}\right)^{\frac{\log m}{\log 2}} = \text{ constant } \approx 2.685452001\ldots$$

is a constant for almost all real numbers. The exceptions are rational numbers, quadratic irrationals and some other irrational numbers, like for example e. The set of exceptions is of Lebesgue measure zero.
Continued fraction arithmetic for performing addition, subtraction, multiplication and division was developed by Bill Gosper in 1972.
Continued fractions keep playing role today in number theory and Diophantine approximation.

Chapter 1

Geometric picture of continued fractions

1.1 Continued fractions and Möbius transformations

We shall describe the functions (**K** for *Kettenbruch*):

$$\mathbf{K}(b_n) = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{\ddots + \cfrac{1}{b_N}}}} \text{ and } \mathbf{K}(b_n) = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots}}}}$$

of the $N+1$ variables $b_0, b_1, \ldots, b_N$, as a *finite continued fraction* and of the infinite number of variables $(b_n)_{n=0}^{\infty}$ as an *infinite continued fraction*, respectively. We call the variables $(b_n)_{n=0}^{\infty}$ *partial quotients* (also known as *continued fraction digits or coefficients*) and use less voluminous notation for the continued fraction by putting the partial quotients into square brackets $[b_0; b_1, b_2, \ldots]$ with the risk of losing some expressiveness.
If we restrict the domain for the partial quotients such that $b_0 \in \mathbb{Z}$ and $b_n \in \mathbb{Z}^+$ for $n > 0$, then we have *simple* continued fractions. We shall also consider the cases when $b_n \in \mathbb{Z}$, $b_n \in \mathbb{R}$ and $b_n \in \mathbb{C}$.
In the case of finite continued fraction the value of $[b_0; b_1, b_2, \ldots, b_N]$ is calculated in the obvious way. If there are infinitely many partial quotients then we define the value of

$$[b_0; b_1, b_2, \ldots] := \lim_{n\to\infty} [b_0; b_1, b_2, \ldots, b_n].$$

We call the fraction

$$\frac{p_n}{q_n} = [b_0; b_1, b_2, \ldots, b_n]$$

that ignores the partial quotients after b_n, nth-order *convergent* (also known as *approximant* or *truncation*). The existence of the limit and various ways of

divergence will be studied in this project.
The nominator and denominator of a convergent can be calculated separately using Wallis-Euler recursion:

$$p_{-2} = 0, p_{-1} = 1 \text{ and } q_{-2} = 1, q_{-1} = 0,$$

$$p_k = b_k p_{k-1} + p_{k-2} \text{ and } q_k = b_k q_{k-1} + q_{k-2}.$$

Note that denominators q_k do not depend on b_0.
For example $\sqrt[12]{2} = [1; 16, 1, 4, 2, \ldots]$ has its first 5 convergents calculated with Wallis-Euler recursion in Table 1.1, where $\frac{196}{185} = 1.059459\ldots$ and $\sqrt[12]{2} = 1.059463\ldots$.

Table 1.1: Calculation of first 5 convergents of $\sqrt[12]{2} = [1; 16, 1, 4, 2, \ldots]$.

k	-2	-1	0	1	2	3	4	...
b_k			1	16	1	4	2	
p_k	0	1	1	17	18	89	196	
q_k	1	0	1	16	17	84	185	
k-th convergent			1	17/16	18/17	89/84	196/185	...

Following the reasoning of Beardon and Short in [1] we express the above recurrence in matrix form:

$$\begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix} = \begin{pmatrix} b_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} b_k & 1 \\ 1 & 0 \end{pmatrix}. \tag{1.1}$$

We see that all of the matrices involved have determinant ± 1, therefore are all invertible and so we can associate each matrix with a non-singular Möbius transformation of the complex plane. The association has the form:

$$\text{matrix } \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \text{ multiplication} \rightleftharpoons \text{ Möbius transform } \frac{az+b}{cz+d}, \text{ composition.}$$

In particular, each matrix on the right side of 1.1 is connected with a Möbius transformation $t_k(z) = b_k + 1/z$ for each k, which is a composition of a complex inversion and a translation. Then the product on the right hand side of 1.1 corresponds to the composition that is again a Möbius map:

$$t_0 \circ t_1 \circ \ldots \circ t_k(z) =: T_k(z)$$

and using the left hand side of 1.1 we can express the map as:

$$T_k(z) = \frac{p_k z + p_{k-1}}{q_k z + q_{k-1}}.$$

Evaluating $T_k(z)$ at ∞ and 0 give:

$$T_k(\infty) = \frac{p_k}{q_k} \text{ and } T_k(0) = \frac{p_{k-1}}{q_{k-1}}$$

and so we can not only recover all the convergents from the Möbius maps T_k but also the partial quotients b_k:

$$b_k = t_k(\infty) = T_{k-1}^{-1} \circ T_k(\infty).$$

1.2 Horocycles and horospheres

The possibility of studying continued fractions in terms of Möbius maps opens up a way to see the continued fractions geometrically as chains of horocycles and horospheres in cases of $b_n \in \mathbb{R}$ and $b_n \in \mathbb{C}$ respectively, an alluring visualisation developed by Ford in his paper [2].
By a horocycle we mean either a circle in the upper half of the complex plane $\mathbb{H}$ that is tangent to the real axis, or a horizontal line in $\mathbb{H}$. Each horocycle has a *base point* that is the point of tangency with the real line in the former case and ∞ in the latter case.
For each rational point $x = \frac{p}{q}$, where p and q are relatively prime integers, we construct a *Ford circle* of radius

$$R_{\frac{p}{q}} = \frac{1}{2q^2},$$

tangent to the x-axis at the given point and lying in the upper half-plane. Ford circles are are either tangent or disjoint, in no case do they overlap. Thus the set of Ford circles is a subset of all horocycles.
For each Gaussian rational point $x = \frac{p}{q}$ on the complex plane, where p and q are relatively prime Gaussian integers, we construct a *Ford sphere* of radius

$$R_{\frac{p}{q}} = \frac{1}{2q\bar{q}},$$

tangent to the complex plane at the given point and lying in the upper half-space. Similar to the Ford circles the Ford spheres can be tangent or disjoint but never overlapping.
We cannot construct Ford circles for irrationals but we can construct a *Short circle* for each real point α relative to any rational $\frac{p}{q}$, that is a circle of radius

$$R_{\frac{p}{q}}(\alpha) = \frac{1}{2}|q\alpha - p|^2, \tag{1.2}$$

tangent to the x-axis at α and tangent to the Ford circle of $\frac{p}{q}$. The construction is described in Short's paper [4]. We see the construction of the Short circle of α relative to $\frac{p}{q}$ in Figure 1.1.

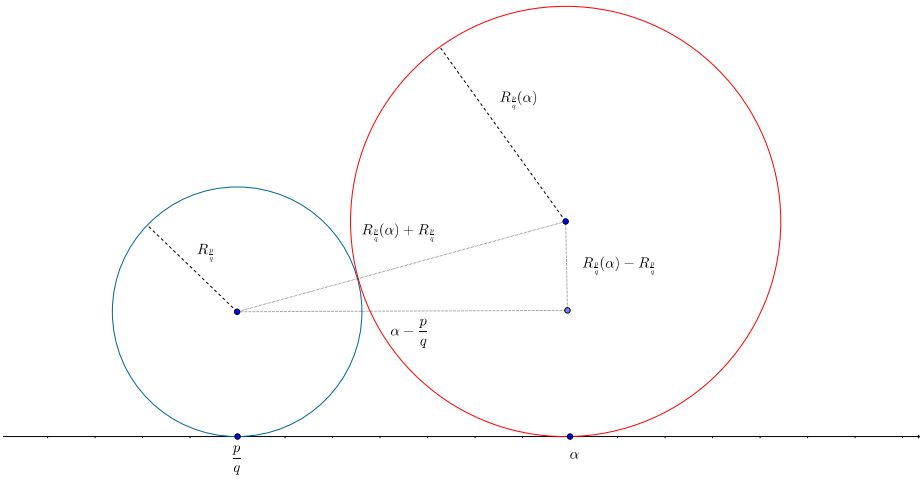

Figure 1.1:
Ford circle of $\frac{p}{q}$ with radius $R_{\frac{p}{q}} = \frac{1}{q^2}$ (blue) and Short circle of α relative to $\frac{p}{q}$ with radius $R_{\frac{p}{q}}(\alpha) = \frac{1}{2}|q\alpha - p|^2$ (red).

Applying Pythagoras's theorem to the drawn triangle:

$$\left|\alpha - \frac{p}{q}\right|^2 + \left(R_{\frac{p}{q}}(\alpha) - R_{\frac{p}{q}}\right)^2 = \left(R_{\frac{p}{q}}(\alpha) + R_{\frac{p}{q}}\right)^2$$

we get the radius of the Short circle

$$R_{\frac{p}{q}}(\alpha) = \frac{1}{4R_{\frac{p}{q}}}\left|\alpha - \frac{p}{q}\right|^2 = \frac{q^2}{2}\left|\alpha - \frac{p}{q}\right|^2,$$

since $R_{\frac{p}{q}} = \frac{1}{2q^2}$.
Thus the set of all Short circles is another subset of all horocycles.
Similarly, we construct a *Short sphere* for each complex number relative to any Gaussian rational, that is a sphere of radius

$$R_{\frac{p}{q}}(\alpha) = \frac{1}{4R_{\frac{p}{q}}}\left|\alpha - \frac{p}{q}\right|^2 = \frac{q\bar{q}}{2}\left|\alpha - \frac{p}{q}\right|^2$$

tangent to the complex plane at the given complex number α and to the Ford sphere of the Gaussian rational $\frac{p}{q}$.

1.2.1 Ford's clockwork

Now we put the idea of visualising real and complex numbers as horocycles and horospheres into work, and unravel the essence of Ford's clockwork that assembles the chains of horocycles and horospheres. The construction of the clockwork is based on Ford's paper [2].
Successive convergents, $\frac{p_{k-1}}{q_{k-1}}$ and $\frac{p_k}{q_k}$, have their Ford circles tangent to each

other[1], i.e.

$$p_k q_{k-1} - q_k p_{k-1} = (-1)^{k-1}. \tag{1.3}$$

This can be seen by taking determinants from the both sides of Eq.1.1 and considering the triangle ABC in Figure 1.2, where

$$\begin{aligned} AB^2 &= \left(\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k}\right)^2 + \left(\frac{1}{2q_{k-1}^2} - \frac{1}{2q_k^2}\right)^2 \\ &= \frac{(q_k p_{k-1} - p_k q_{k-1})^2}{q_{k-1}^2 q_k^2} + \frac{1}{4q_{k-1}^4} - \frac{1}{2q_{k-1}^2 q_k^2} + \frac{1}{4q_k^4} \\ &= \frac{(q_k p_{k-1} - p_k q_{k-1})^2 - 1}{q_{k-1}^2 q_k^2} + \left(\frac{1}{2q_{k-1}^2} + \frac{1}{2q_k^2}\right)^2 \\ &= \frac{(q_k p_{k-1} - p_k q_{k-1})^2 - 1}{q_{k-1}^2 q_k^2} + (AD + EB)^2. \end{aligned}$$

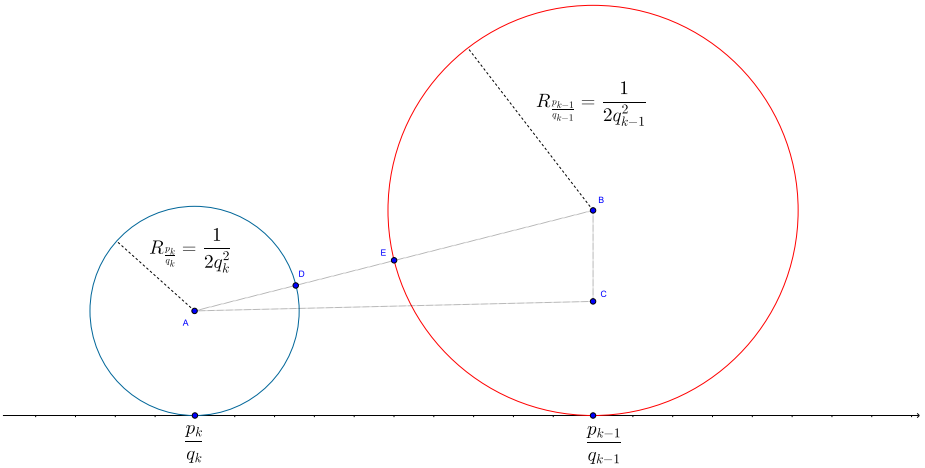

Figure 1.2:
Ford circles of $\frac{p_k}{q_k}$ and $\frac{p_{k-1}}{q_{k-1}}$ with radii $\frac{1}{q_k^2}$ and $\frac{1}{q_{k-1}^2}$.

Consequently the circles are tangent if $(q_k p_{k-1} - p_k q_{k-1})^2 = 1$ since then $AB = AD + EB$.
If we have the tangent Ford circles of $\frac{p_{k-1}}{q_{k-1}}$ and $\frac{p_k}{q_k}$, then finding the next Ford circle of $\frac{p_{k+1}}{q_{k+1}}$ could be visualised by Ford's clockwork as seen in Figure 1.3.
In the beginning there is the line i, i.e. the horocycle Π_{-1}. The next horocycle has its base at b_0 and it touches Π_{-1} at $b_0 + i$. The touching point fixes 0 for the current circle and all possible touching points with neighbouring Ford circles determine the clock-face so that $+1$ is the highest on the right and all positive integers are on the right tighter and tighter approaching infinity near the base. Similarly -1 is the highest on the left and all negative integers are on the left. Thus b_1 points the hand on the integer on the dial and the next horocycle is

[1]In passing, note a nice property of two rational numbers. If $\frac{p}{q}$ and $\frac{r}{s}$ have their Ford circles tangent to each other, then the cross products, ps and rq, are consecutive integers.

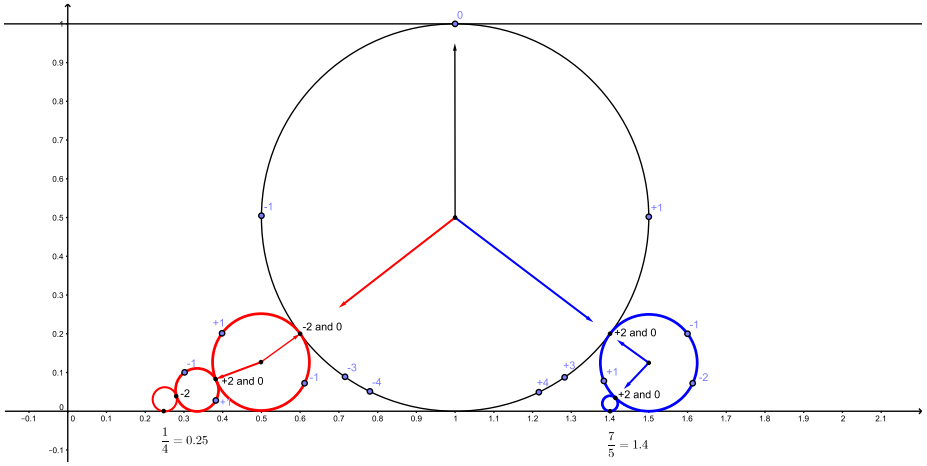

Figure 1.3:
Ford's clockwork. Two exmaples: $[1; 2, 2] = \frac{7}{5} = 1.4$ (blue) and $[1; -2, 2, -2] = \frac{1}{4} = 0.25$ (red).

chosen.
This new touching point fixes 0 for the new horocycle and the process is repeated, while keeping in mind that the positive and negative sides swap each time. So that if we have b_n all positive or all negative then the horocycles alternate like pendulum around the final limit point but if there is a change in sign then we jump to the other side.
There are examples of the continued fractions $[1; 2, 2] = \frac{7}{5} = 1.4$ and $[1; -2, 2] = \frac{1}{3}$ in Figure 1.3.
We can also use the clockwork for $b_n \in \mathbb{R}$ but instead of Ford circles we must use Short circles. Now the next horocycle can touch the current one at any point and clearly all the Short circles that are touching at points between -1 and +1 on the clock dial are larger and so the door to divergence is open.
The clockwork visualisation is extendible to the $b_n \in \mathbb{C}$ as well. Instead of the dial numbers on the circle we have the dial numbers on the sphere, the modulus $|b_n|$ gives the latitude and the argument of b_n gives the longitude.
The touching point with the previous horosphere fixes the 0 as the North Pole and the Western and Eastern Hemispheres will be alternating similarly to the alternating of the positive and negative sides of horocycles.

1.2.2 Chains of horocycles

If $b_n \in \mathbb{Z}^+$ for $n > 0$ then all the horocycles in the chain are Ford circles and so their properties are the following:

- The circles are either tangent or disjoint, they never overlap.
- The circles in the chain are strictly decreasing after the first two circles, which can be of equal size if $b_1 = 1$.
- Even and odd convergents are alternately on the both sides of the final limit.

- If there are ones among partial quotients then the horocycle corresponding to the 1 is not only tangent to the previous horocycle but also to the one before the previous.
- The pre-images are unique in the sense that for any given rational or irrational number we can construct its continued fraction representation. We shall see some constructions in §1.4.

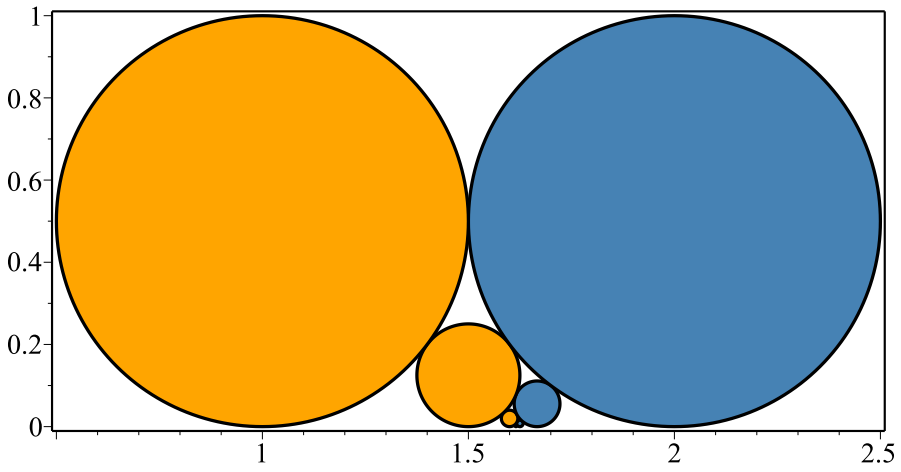

Figure 1.4:
Example of a continued fraction with $b_n \in \mathbb{Z}^+$. First 6 Ford circles in the chain of horocycles of the golden ratio $\phi = \frac{1+\sqrt{5}}{2} = [1;\bar{1}] = 1.6180339887\ldots$. Even convergents have orange Ford circles and odd convergents have blue.

The examples of $b_n \in \mathbb{Z}^+$ are the beginning of the chain of horocycles of the golden ratio in Figure 1.4 and the chain of horocycles of e in Figure 1.5. Note that all horocycles are tangent to previous two in case of the golden ratio and some in case of e.

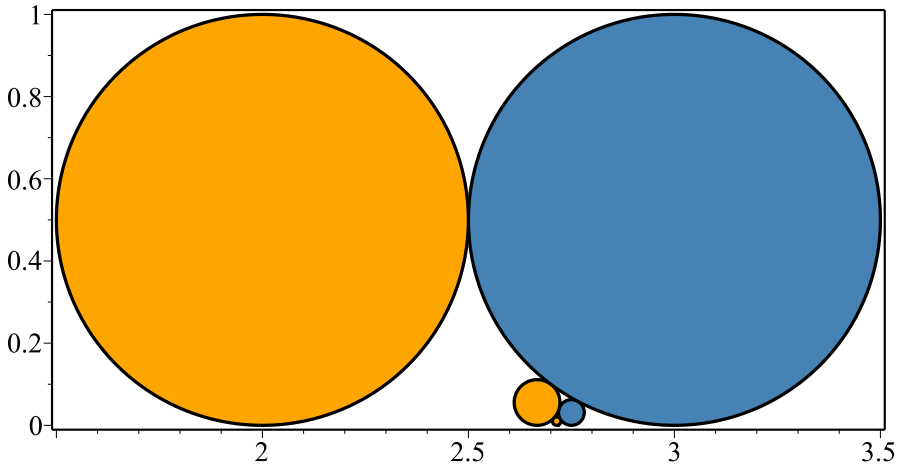

Figure 1.5:
Example of a continued fraction with $b_n \in \mathbb{Z}^+$. First 5 Ford circles in the chain of horocycles of $e = [2; 1, 2, 1, 1, 4, 1, \ldots] = 2.71828\ldots$.

Now if we consider also positive real numbers that are not necessarily integers, the picture changes. With $b_n \in \mathbb{R}^+$ the properties of the chains of horocycles are the following:

- Only successive horocycles are always tangent, others can be tangent, disjoint or overlapping.
- The horocycles in the chain can increase and decrease. But the even and odd sub-sequences are still decreasing.
- Even and odd convergents still alternate like a pendulum clock.
- The horocycles in the chain are Short circles, not necessarily Ford circles.
- A completely new feature is the possibility to diverge. But not to infinity, just divergence with two limit points.

The example of non-successive overlap and when every even circle is succeeded by a bigger odd circle but still both even and odd sub-sequences decrease in radius is seen in Figure 1.6 for the continued fraction $[1; \frac{1}{2}, \frac{1}{3}, \ldots]$. In this example the partial quotients sum up to the harmonic series which diverges and therefore the continued fraction converges but extremely slowly. It takes about 70 partial quotients to fix the second significant digit 1.7. The convergence of the

continued fraction follows from the criteria we will prove in Theorem 1 below in §1.3.

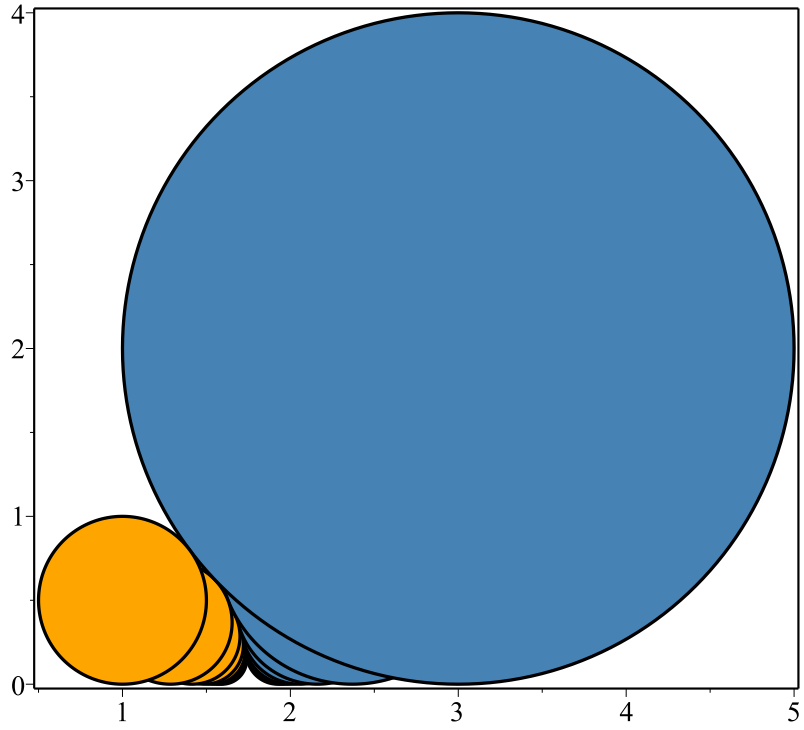

Figure 1.6:
Example of a continued fraction with $b_n \in \mathbb{R}^+$, that converges. First 15 horocycles of the continued fraction $[1; \frac{1}{2}, \frac{1}{3}, \ldots]$ that has partial quotients that sum up to the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$.

The example of divergence is seen in Figure 1.7. In this example the partial quotients sum up to a series that converges and therefore the continued fraction diverges, again by Theorem 1 below in §1.3.
Although the continued fraction $[1; 1, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \frac{1}{6^2} \ldots]$ diverges, even and odd sub-sequences of convergents do converge to the two separate finite limits.

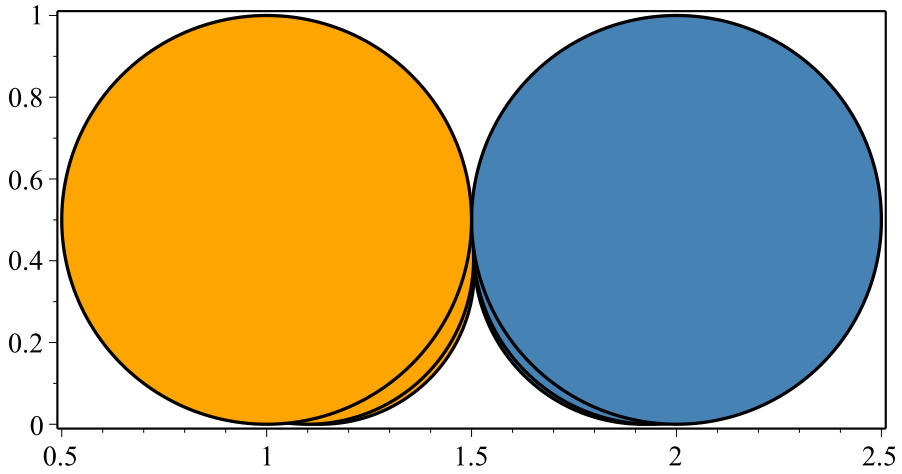

Figure 1.7:
Example of a continued fraction with $b_n \in \mathbb{R}^+$, that diverges to two limit points. First 6 horocycles of the continued fraction $[1; 1, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \frac{1}{6^2} \ldots]$ that has partial quotients that sum up to the series $\frac{3}{4} + \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + 1 + \frac{1}{3^2} + \frac{1}{4^2} + \ldots = \frac{3}{4} + \zeta(2) = \frac{3}{4} + \frac{\pi^2}{6}$.

Next we consider again integers, but now we allow also negative integers, so that $b_n \in \mathbb{Z}$. Compared with $b_n \in \mathbb{Z}^+$ we have again new features:

- Sign change in the sequence of partial quotients causes a disruption in the alternating behaviour of the convergents.
- Minus one turns back time in a sense that it causes the next horocycle to be bigger.
- The pre-images are not unique. We shall see some examples of nearest integer and even integer continued fractions in §1.4.

For example, the sequence of 1,-1 in the beginning of the partial quotients takes us back to the primeval home $\Pi_{-1} = \{z : \Im(z) = 1\} \cup \{\infty\}$.
Alternating signs in the sequence of the partial quotients result in the chain of horocycles that evolve to one side only, as seen for example in Figure 1.8 and 1.9. Both of these examples converge due to the criteria we will prove in Theorem 2 below in §1.3.

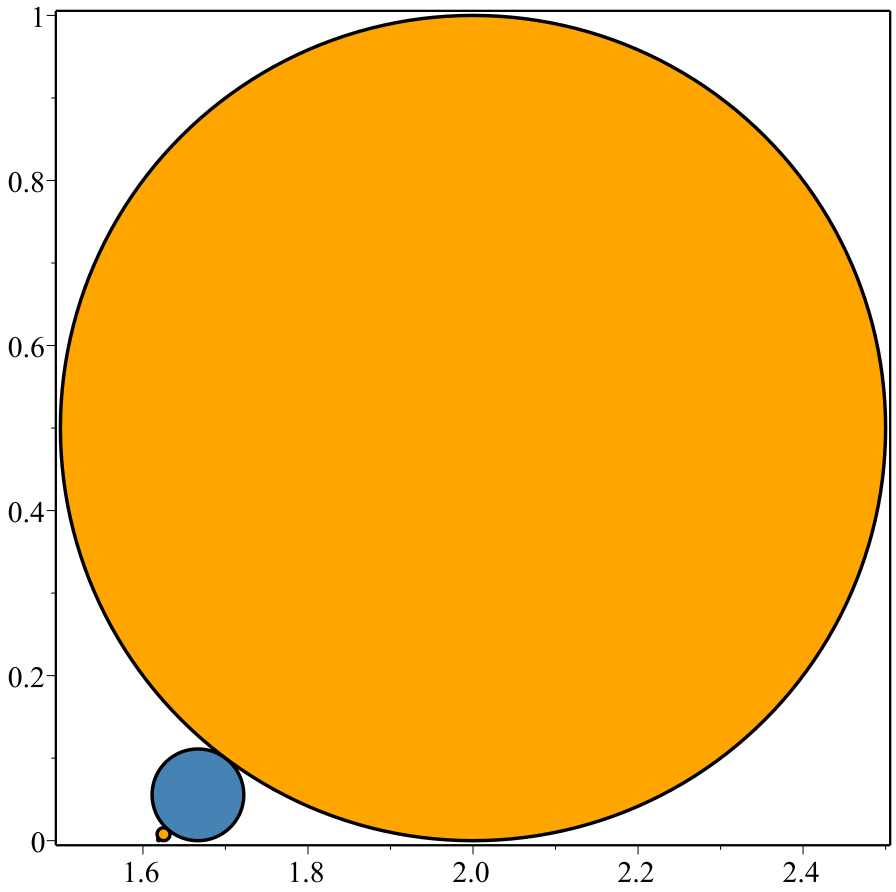

Figure 1.8:
Example of a continued fraction with $b_n \in \mathbb{Z}$, that converges. First 4 Ford circles in the nearest integer chain of horocycles of the golden ratio $\phi = \frac{1+\sqrt{5}}{2} = \text{NI}[2; \overline{-3, 3}] = 1.6180339887\ldots$.
Even and odd convergents both evolve to the left side only.
See the contrast in the speed of convergence compared with figure 1.4 that illustrates much faster convergence of the nearest integer continued fractions.

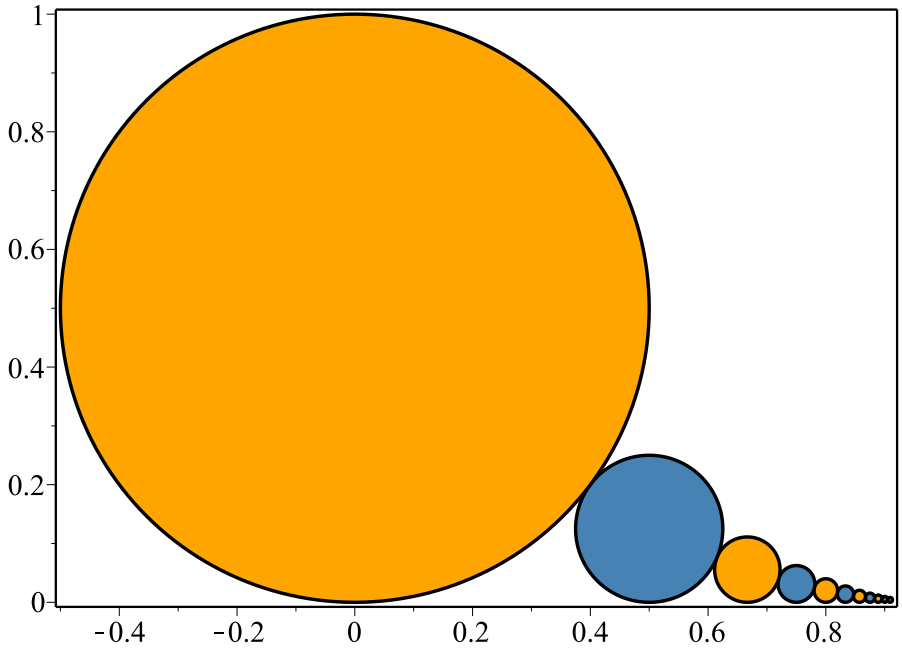

Figure 1.9:
Example of a continued fraction with $b_n \in \mathbb{Z}$, that converges. The beginning of the chain of horocycles of $[0; \overline{2, -2}] = 1$. Even and odd convergents both evolve to the right side only.

Summarizing, allowing $b_n \in \mathbb{R}$ we can get all the above mentioned features in a chain of horocycles, and in addition divergence to positive or negative infinity. An example of divergence such that even convergents diverge to infinity and odd convergents converge to some finite limit is seen in Figure 1.10.

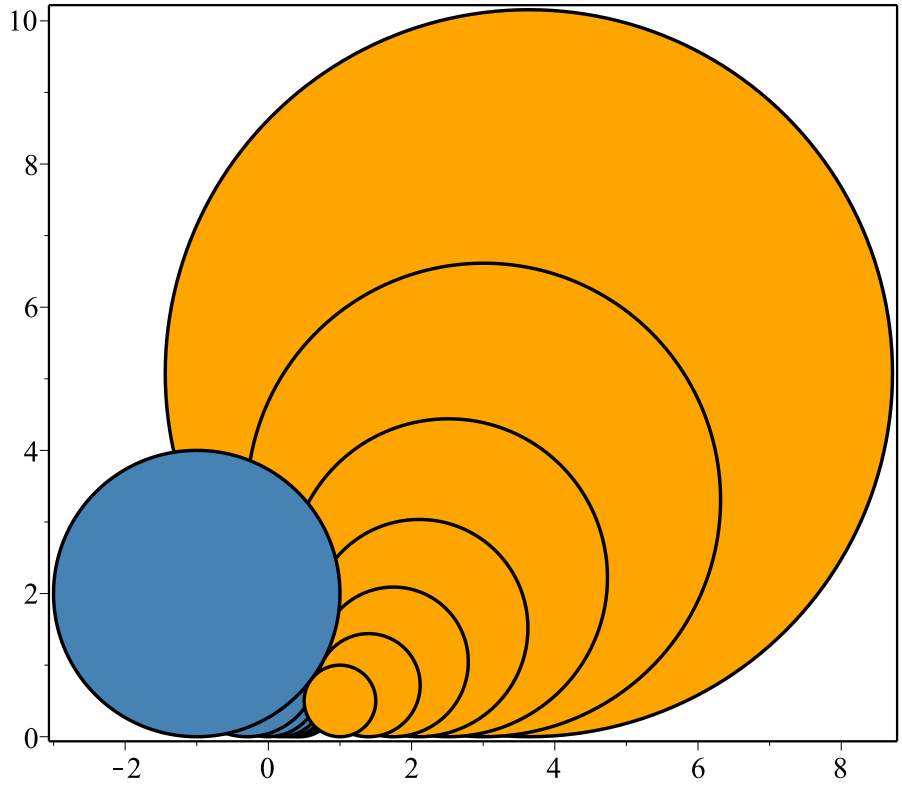

Figure 1.10:
Example of a continued fraction with $b_n \in \mathbb{R}$, that diverges. First 15 horocycles of the continued fraction $[1; -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4} \ldots]$ that has partial quotients that sum up to the alternating harmonic series $\sum_{n=1}^{\infty} \frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots = \log 2$. Even convergents diverge to infinity.

We could even construct such a chain of horocycles that visits each rational number infinitely often, hinted at by Lester R. Ford in [2].
The construction is seen in Figure 1.11. We take a zigzag line that has turning points at the coordinates $((-1)^n n, \frac{1}{n})$, and observe what Ford circles it crosses in succession. We see that each Ford circle is visited infinitely many times and thus each rational number is hit infinitely often.
For the construction of the corresponding continued fraction we will use Ford's clockwork.
The first horocycle has base at -1 and radius $\frac{1}{2}$, thus $b_0 = -1$.
The second horocycle has base at 0 and radius $\frac{1}{2}$, thus $b_1 = 1$.
The third horocycle has base at 1 and radius $\frac{1}{2}$, thus $b_2 = -2$.
Note that we needed the negative sign to get to the other side of the horocycle.
The fourth horocycle has base at 2 and radius $\frac{1}{2}$, thus $b_3 = 2$.
The change of sign again in order to get to the other side of the horocycle.

The zigzag line turns now back, so the fifth horocycle has base at 1 and radius $\frac{1}{2}$, thus we need to go back to the horocycle where we game from and this can be achieved by $b_4 = 0$.
The sixth horocycle has base at 0 and radius $\frac{1}{2}$, thus $b_5 = -2$.
We can continue in the same fashion until we hit the 17th horocycle that is smaller, and so we need $b_{16} = 1$.
The beginning of the continued fraction until the 17th partial quotient is:

$$[-1; 1, -2, 2, 0, -2, 2, -2, 2, 0, -2, 2, -2, 2, -2, 2, 1, \ldots].$$

This beginning takes us to the 17th convergent, $\frac{p_{16}}{q_{16}} = 3.5$, that has the horocycle with base at 3.5 and radius $\frac{1}{8}$.
By following the zigzag line we can therefore construct the continued fraction that diverges and its convergents visit every rational number infinitely often.

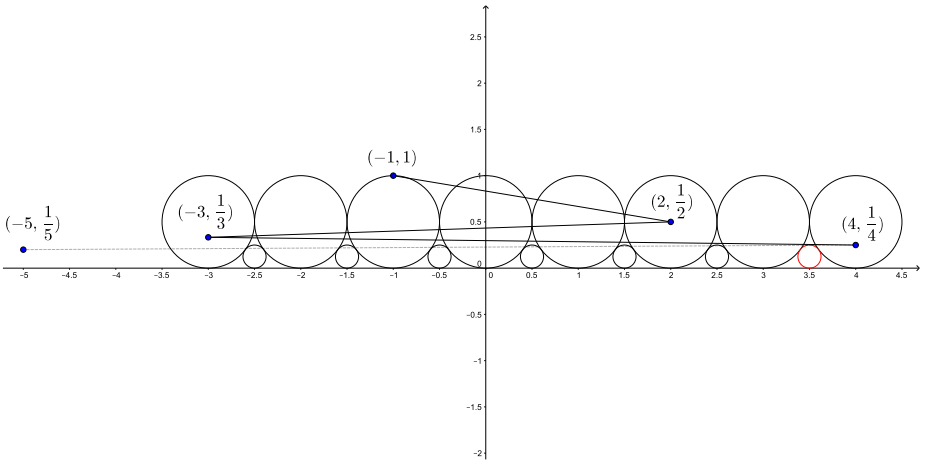

Figure 1.11:
The construction of the chain of horocycles that visits each rational number infinitely often. The zigzag line has turning points at the coordinates $((-1)^n n, \frac{1}{n})$ and it visits each Ford circle infinitely many times. The 17th Ford circle that is visited by the zigzag line is marked red.

1.2.3 Chains of horospheres

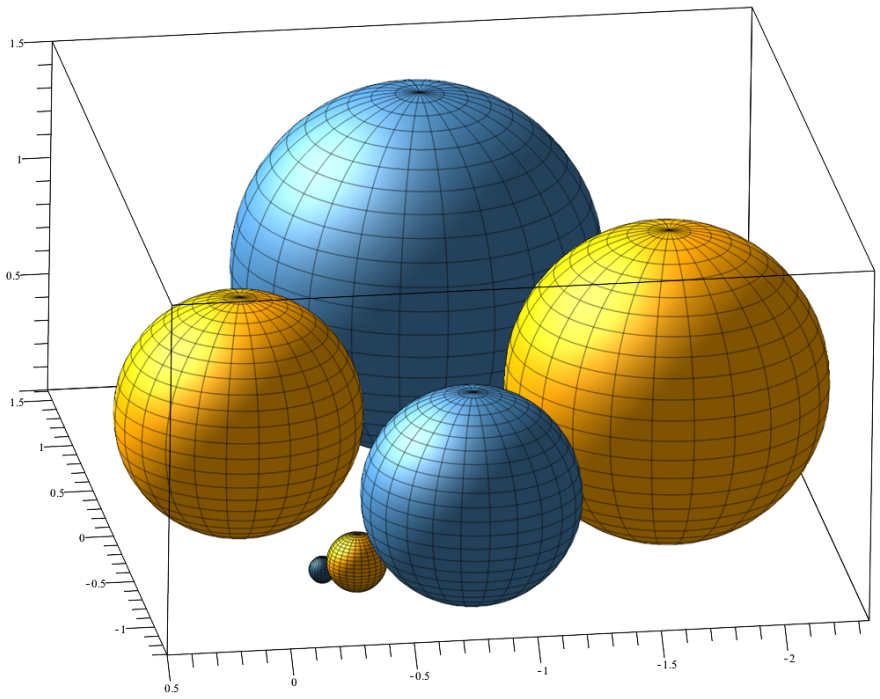

Figure 1.12:
Chain of horospheres of the continued fraction
$[0; \frac{1+i}{\sqrt{3}}, -1.5, 1.5, -2.5, 2] \approx -0.6 - 0.2i$.

In the construction of the chain of horospheres we follow Beardon and Short in [1].
Allowing $b_n \in \mathbb{C}$ takes us to the world of horospheres. However we can still use Möbius transforms to visualise the evolution of the continued fractions, we just have to use the Poincaré extension of Möbius transformation in terms of quaternions.
We shall use quaternions of the form:

$$x + yi + tj + wk \text{ in } \mathbb{H}^3 = \{w = 0, (x, y, t) \in \mathbb{R}^3 : t > 0\} = \{z + tj : z \in \mathbb{C}, t > 0\},$$

where the boundary of $\mathbb{H}^3$ is $\mathbb{C}$.
The Poincaré extension to $\mathbb{H}^3$ of the Möbius transformation $g(z) = \frac{az+b}{cz+d}$ in $\mathbb{C}$ is:

$$g(z + tj) = \frac{(az + b)(\overline{cz + d}) + a\bar{c}t^2 + |ad - bc|tj}{|cz + d|^2 + |c|^2 t^2}. \tag{1.4}$$

Notice that if $t = 0$ then we get back the complex plane Möbius transformation.
If $g(z) = z + b$ is a translation on $\mathbb{C}$ then $g(z + tj) = z + b + tj$ is a translation by b on $\mathbb{H}^3$.
If $g(z) = \frac{1}{z}$ is a complex inversion on $\mathbb{C}$ then $g(z + tj) = \frac{\bar{z}+tj}{|z|^2+t^2}$ is a geometric

inversion in the unit half-sphere followed by reflection in the Euclidean plane $y = 0$ on $\mathbb{H}^3$ and so $g(z + tj)$ preserves $\mathbb{H}^3$.
Now when we have $b_n \in \mathbb{C}$, all the continued fractions start their horosphere evolution from the horosphere $\Sigma_{-1} = \{z + j : z \in \mathbb{C}\} \cup \{\infty\}$, that is the horizontal plane through j.
The main mechanism for the evolution of the horospheres is the Möbius transformation $t_k(z) = b_k + 1/z$. Since the coefficients of a Möbius map are not unique in the sense that we can always normalise the map, we will use $t_k(z) = \frac{ib_k z+i}{iz+0}$ in order to simplify the use of 1.4:

$$t_k(z + tj) = \frac{(ib_k z + i)(\overline{iz + 0}) + ib_k \bar{i} t^2 + |ib_k 0 - ii|tj}{|iz + 0|^2 + |i|^2 t^2} = b_k + \frac{\bar{z} + tj}{|z|^2 + t^2}.$$

Thus the Möbius map t_k is the composition of geometric inversion in the unit half-sphere followed by reflection in the Euclidean plane $y = 0$ and translation by b_k on $\mathbb{H}^3$. All these transformations map $\mathbb{H}^3$ back to itself.
The first horosphere Σ_{-1} has its base point at ∞ and so the next horosphere Σ_0 has its base point at $t_0(\infty) = b_0$.
We see from 1.4 that the highest point of the new horosphere will be at $z = -\frac{d}{c}$ and it is at height $\frac{1}{|c|^2} j$. Therefore the radius of the new horosphere is:

$$\frac{1}{2|c|^2}$$

and in particular the horosphere Σ_0 has always radius $\frac{1}{2}$ (since $q_0 = 1$), Σ_1 has radius $\frac{1}{2|b_1|^2}$, Σ_2 has radius $\frac{1}{2|b_1 b_2+1|^2}$, etc.
The horospheres Σ_{-1} and Σ_0 are tangent at the point $b_0 + j = t_0(j)$. In general, every time we start to calculate parameters of the new horosphere by $T_k(z) = t_0 \circ t_1 \circ \ldots \circ t_k(z)$ we see that Σ_{-1} and $t_k(\Sigma_{-1})$ are tangent at the point $t_k(j)$.
Consequently the horospheres $T_{k-1}(\Sigma_{-1})$ and $T_k(\Sigma_{-1}) = T_{k-1}(t_k(\Sigma_{-1}))$ are tangent at the point $T_k(j)$. So the successive horospheres are always tangent in the chain but non-successive horospheres might overlap.
The chain of horospheres is generated by $T_k(\Sigma_{-1})$ for $k = 0, 1, 2. \ldots$ as spheres are mapped to spheres by the extended Möbius transforms. The chain is calculated from the continued fraction $[b_0; b_1, b_2, \ldots]$, where $b_k \in \mathbb{C}$. An example of the chain of horospheres is in Figure 1.12.
To complete the correspondence between complex continued fractions and chains of horospheres we need a way to calculate the continued fraction from given chain of horospheres. If we are given a chain of horospheres by its base points $\infty, z_0, z_1, z_2, \ldots$ we can recover the continued fraction recursively:

$$b_0 = z_0 \text{ and } b_k = T_{k-1}^{-1}(z_k).$$

Given a Gaussian rational, for example $\frac{5}{11} + i$, we calculate its continued fraction representation $[i; 2, 5]$ and then we can visualise it as a chain of horocycles as seen in Figure 1.13.
The same procedure with $e + \pi i$ produces $[3 + 3i; 1 + 3i, 2 + i, \ldots]$ and the first 3 horospheres are seen in Figure 1.14.

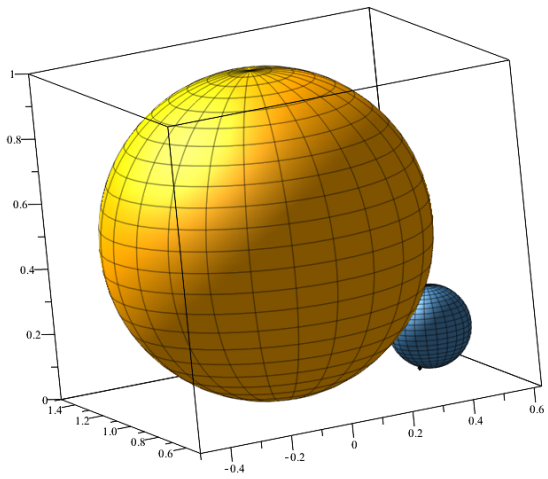

Figure 1.13:
All 3 horospheres of the chain of the continued fraction of $\frac{5}{11} + i = [i; 2, 5]$.

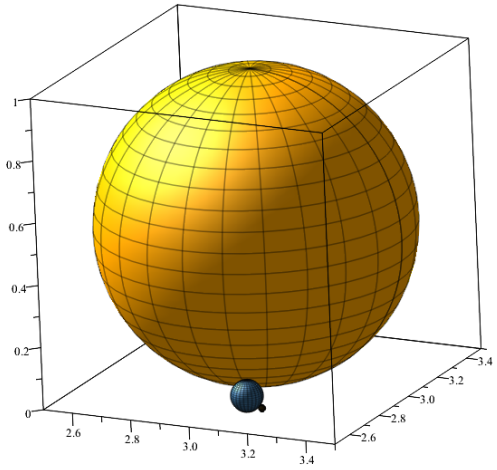

Figure 1.14:
First 3 horospheres of the chain of the continued fraction of
$e + \pi i = [3 + 3i; 1 + 3i, 2 + i, \ldots]$.

1.3 Convergence and divergence of continued fractions

Now we want look at the criterion for convergence, which was already used in examples in §1.2.2.
Given a sequence of numbers $(b_n)_{n=0}^{\infty}$ then similarly to the series $\sum_{n=0}^{\infty} b_n$ and infinite products $\prod_{n=0}^{\infty} b_n$ for which the convergence means convergence of the sequence of partial sums and partial products, the convergence of continued fraction $\mathbf{K}_{n=0}^{\infty}(b_n)$ means convergence of sequence of convergents $(\frac{p_n}{q_n})_{n=0}^{\infty}$.
In the case of $b_n > 0$ for all $n \geq 1$, there is a useful test for convergence that is due to Philipp Ludwig von Seidel and Moritz Abraham Stern in the 1840s. The converse is due to Moritz Abraham Stern and Otto Stolz in the 1860s.

Theorem 1. *For the continued fraction* $\mathbf{K}_{n=0}^{\infty}(b_n) = [b_0; b_1, b_2, b_3, \ldots]$ *with* $b_n \in \mathbb{R}^+$ *to converge, it is necessary and sufficient that the series*

$$\sum_{n=1}^{\infty} b_n$$

is divergent.

Proof. We follow the geometrical reasoning given by Beardon and Short in[1].
We have seen that the fate of the sequence of convergents $(\frac{p_n}{q_n})_{n=0}^{\infty}$ is connected to the action of the Möbius maps $T_n(z) = t_0 \circ t_1 \circ \ldots \circ t_n(z)$ and we can follow the evolution of the bases of the horocycles by calculating consecutive

$$T_n(\infty) = \frac{p_n}{q_n} = T_{n+1}(0).$$

Since $b_n > 0$ each T_n maps $[0, \infty]$ to itself such that the point $T_n(b_{n+1}) = T_{n+1}(\infty)$ lies between $T_n(0)$ and $T_n(\infty)$. Thus each base T_{n+1} lies between two previous bases $T_{n-1}(\infty)$ and $T_n(\infty)$.
As $b_0 = T_0(\infty) < T_1(\infty) = t_0(t_1(\infty)) = t_0(b_1) = b_0 + \frac{1}{b_1}$ we see that

$$T_0(\infty) < T_2(\infty) < \ldots < T_{2n}(\infty) < T_{2n+1}(\infty) < \ldots < T_3(\infty) < T_1(\infty)$$

Therefore the even subsequence is increasing and is bounded. Similarly, the odd subsequence is decreasing and is bounded. So both of the limits exist and now the question is whether the limits are the same, in which case the continued fraction converges, or are different so that the continued fraction diverges by oscillation.
Let $\alpha = \lim_{n\to\infty} T_{2n}(\infty)$ and $\beta = \lim_{n\to\infty} T_{2n+1}(\infty)$ be the two limits.
Suppose that $\mathbf{K}_{n=0}^{\infty}(b_n)$ diverges and so the limits α and β are distinct.
Consecutive horocycles are tangent and their bases and radii are related by

$$|T_n(\infty) - T_{n-1}(\infty)|^2 = (R_n + R_{n-1})^2 - (R_n - R_{n-1})^2 = 4R_nR_{n-1}$$

and as the distances between the bases are decreasing during the evolution

$$4R_nR_{n-1} = |T_n(\infty) - T_{n-1}(\infty)|^2 < |T_{n-1}(\infty) - T_{n-2}(\infty)|^2 = 4R_{n-1}R_{n-2}$$

the even and odd sub-sequences of radii both decrease, $R_n < R_{n-2}$.
Both must have positive limits since

$$4R_nR_{n-1} = |T_n(\infty) - T_{n-1}(\infty)|^2 > |\alpha - \beta|^2.$$

We can recover the partial quotients b_n from the Möbius maps:

$$b_n = t_n(\infty) = T_{n-1}^{-1} T_n(\infty) = T_{n-1}^{-1}\left(\frac{p_n}{q_n}\right) =$$
$$= \left| \frac{q_{n-2}\frac{p_n}{q_n} - p_{n-2}}{-q_{n-1}\frac{p_n}{q_n} + p_{n-1}} \right| = |q_{n-2}p_n - p_{n-2}q_n| =$$
$$= q_n q_{n-2} \left| \frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} \right| = \frac{|T_n(\infty) - T_{n-2}(\infty)|}{2\sqrt{R_n R_{n-2}}}$$

By summing both sides, we see that $\sum b_n$ converges since $\sum |T_n(\infty) - T_{n-2}(\infty)|$ converges and R_n are bounded away from zero.
We have proved that, if $\sum b_n$ diverges then $\mathbf{K}_{n=0}^{\infty}(b_n)$ converges, by proving its contrapositive, if $\mathbf{K}_{n=0}^{\infty}(b_n)$ diverges then $\sum b_n$ converges.
Now we are left to prove the converse, if $\mathbf{K}_{k=0}^{\infty}(b_k)$ converges then $\sum b_k$ diverges. This part of the theorem applies not only for positive partial quotients but for any complex b_k. In the complex case, if $\mathbf{K}_{n=0}^{\infty}(b_n)$ converges then $\sum |b_k|$ diverges.
We saw that the main mechanism for the evolution of the horospheres, the extended Möbius transformation $t_k(z) = b_k + 1/z$, takes the point j to $t_k(j) = b_k + j$ which is the point of tangency of Σ_{-1} and $t_k(\Sigma_{-1})$.
Now we consider all smooth paths in $\mathbb{H}^3$ from j to $b_k + j$ as a collection Γ, among those there is the straight line, say $\delta \in \Gamma$, with the Euclidean length $|b_k|$.
We can also equip $\mathbb{H}^3 = \{(x, y, t) \in \mathbb{R}^3 : t > 0\}$ with the hyperbolic metric:

$$d\hat{s} = \frac{ds}{t} = \frac{\sqrt{dx^2 + dy^2 + dt^2}}{t}$$

and so we can define the corresponding distance function:

$$\rho(a, b) = \inf_{\eta} \int_{\eta} \frac{\sqrt{dx^2 + dy^2 + dt^2}}{t}$$

where the infimum is taken over all smooth paths η in $\mathbb{H}^3$ from a to b.
The extended Möbius transformations and in particular, t_k, map $\mathbb{H}^3$ to itself and if we use the hyperbolic metric then distance between any two points a and b is equal to the distance between their images. Thus:

$$\rho(t_k(a), t_k(b)) = \rho(a, b).$$

Considering again the points j and $b_k + j$ but now with the hyperbolic distance, we get:

$$\rho(j, b_k + j) = \inf_{\gamma \in \Gamma} \int_{\gamma} \frac{ds}{t} \leq \int_{\delta} \frac{ds}{t} = |b_k|,$$

since $t = 1$ along δ.
The extended Möbius transformation T_{k-1} preserves hyperbolic distances as well and so:

$$\rho(j, b_k + j) = \rho(j, t_k(j)) = \rho(T_{k-1}(j), T_{k-1}(t_k(j))) = \rho(T_{k-1}(j), T_k(j)).$$

By the triangle inequality:

$$\rho(j, T_k(j)) \leq \rho(j, T_1(j)) + \rho(T_1(j), T_2(j)) + \ldots + \rho(T_{k-1}(j), T_k(j))$$
$$\leq |b_1| + |b_2| + \ldots + |b_k|.$$

If the continued fraction $\mathbf{K}_{n=0}^{\infty}(b_n)$ converges, the adjacent horospheres have their bases $T_k(0)$ and $T_k(\infty)$ approaching the same point as $k \to \infty$ all of them being at the horizon of the hyperbolic space $\mathbb{H}^3$. Since the point $T_k(j)$ lies on a hyperbolic line (Euclidean half-circle, seen in Figure 1.15) between $T_k(0)$ and $T_k(\infty)$, it must approach the hyperbolic horizon as $k \to \infty$.
That means:

$$\rho(j, T_k(j)) \to \infty \text{ as } k \to \infty.$$

Consequently, $\sum b_k$ diverges. ■

Remark. *The result shows that continued fractions with positive integer partial quotients always converge.*

Remark. *The condition $b_n > 0$ for all $n \geq 1$ in the test is necessary since for example $[1; -1, 1, -1, 1, \ldots]$ has $\sum_{n=1}^{\infty} b_n$ divergent but the continued fraction doesn't converge and instead oscillates between ∞ and 1.*

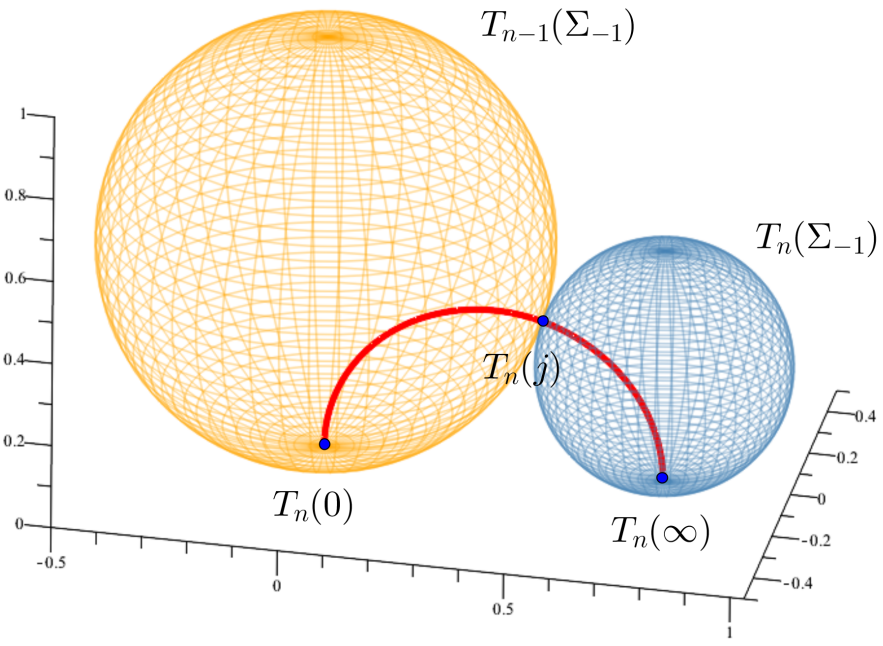

Figure 1.15:
Successive horospheres $T_{n-1}(\Sigma_{-1})$ and $T_n(\Sigma_{-1})$ are tangent at the point $T_n(j)$. Hyperbolic line (Euclidean semicircle) between $T_n(0)$ and $T_n(\infty)$ is the image of the line between ∞ and 0 in $\mathbb{H}^3$ under the extended Möbius transformation T_n.

Another test for convergence is due to Ivan Śleszyński and Alfred Pringsheim in the 1890s.

Theorem 2. *The continued fraction $\mathbf{K}_{n=0}^{\infty}(b_n) = [b_0; b_1, b_2, b_3, \ldots]$ with $b_n \in \mathbb{C}$ converges, if for all n*

$$|b_n| \geq 2.$$

Moreover, all convergents are in the open unit disk centered at b_0 and the final limit value is in the closed unit disk centered at b_0.

Proof. Lorentzen and Waadeland in [3] (page 30-31) have a proof of the theorem. We use the Ford circle method to give an alternative geometric proof. We call two fractions *adjacent* if their Ford circles are tangent. If we have got one fraction $\frac{P}{Q}$ that is adjacent to $\frac{p}{q}$ then we can calculate all the others by:

$$\frac{P_n}{Q_n} = \frac{P + np}{Q + nq} \text{ for all } n \in \mathbb{Z}. \tag{1.5}$$

These Ford circles form a ring of circles around the Ford circle of $\frac{p}{q}$ as seen in Figure 1.16. We can also see that each Ford circle with $q > 1$ has exactly two tangent circles that are larger than it, with all the others being smaller. All integers ($q = 1$) have just one larger tangent Ford circle, line $y = 1$, that can also be considered as a circle with infinite radius. Obviously the integers have two tangent circles of the same size too, with all the others being smaller.

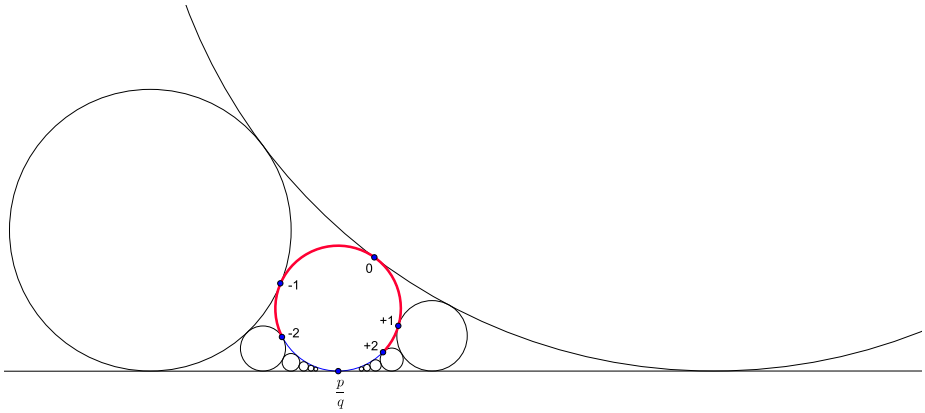

Figure 1.16:
Ring of circles around the Ford circle of $\frac{p}{q}$. The condition $|b_n| \geq 2$ is marked as red prohibited zone for tangent Ford and Short circles.

Using the Ford clockwork we see that the condition $|b_n| \geq 2$ eliminates the possibility that any Ford circle is followed by one of its 3 biggest tangent Ford circles (those can only appear for $b_n = -1, 0, 1$) in case of $b_n \in \mathbb{Z}$ as seen in Figure 1.16.
In the case of $b_n \in \mathbb{R}$ the prohibited zone for tangent Short circles is marked with red in the same Figure 1.16. Similarly for $b_n \in \mathbb{C}$, the Ford's clockwork on spheres as described in the end of §1.2.1 excludes the upper hemisphere for the next sphere to be tangent with, if we start with an upright or up to a 90° tilted sphere. Two examples, an upright and a 90° tilted sphere is seen in Figure 1.17.
We have seen that all continued fractions start their horocycle evolution from the horocycle $\Pi_{-1} = \{z : \Im(z) = 1\} \cup \{\infty\}$ or from the horosphere $\Sigma_{-1} = \{z + j : z \in \mathbb{C}\} \cup \{\infty\}$ and the next horosphere Σ_0 has base at b_0.

In terms of Ford's clockwork, Σ_0 is an upright sphere with radius $\frac{1}{2}$ and the next sphere Σ_1 has maximum tilt and radius if b_1 has smallest possible modulus (argument can be arbitrary). If $|b_1| = 2$ then the radius of Σ_1 is $R_1 = \frac{1}{2|b_1|^2} = \frac{1}{8}$ and the tilt of Σ_1 is $\arccos \frac{R_0 - R_1}{R_0 + R_1} = \arccos \frac{\frac{1}{2} - \frac{1}{8}}{\frac{1}{2} + \frac{1}{8}} = \arccos \frac{3}{5} = 53.13°$.

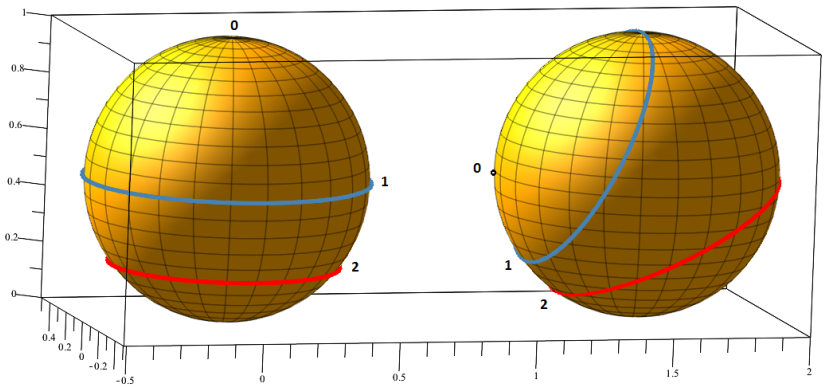

Figure 1.17:
The condition $|b_n| \geq 2$ imposes that any Ford or Short sphere can only be followed by a sphere that is tangent at the red $|b_n| = 2$ line or below. On the left there is an example of an upright sphere and on the right a 90° tilted sphere.

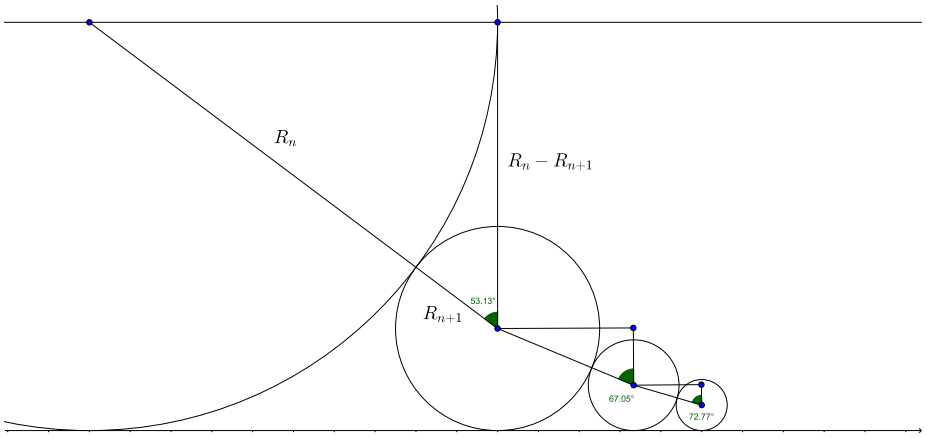

Figure 1.18:
Angles of tilt, $\arccos \frac{R_n - R_{n+1}}{R_n + R_{n+1}}$, in case of $[b_0, 2, -2, 2]$.

The base of Σ_1 is at the distance of $|b_0 - \frac{p_1}{q_1}| = |b_0 - \frac{b_1 b_0 + 1}{b_1}| = \frac{1}{2}$ from b_0. Without loss of generality we may assume that $b_0 = 0$, since it doesn't affect the radii of the horospheres in the chain nor the distance travelled by the bases

of the horospheres.
For the following spheres in the chain, each of them will have the next sphere with maximum tilt and radius in the case of minimal $|b_n|$ with argument chosen properly so that the next sphere is at the antipodal side from the previous. Thus an extreme scenario is achieved with $b_n = \pm 2$ with alternating signs that give the radii of the horospheres in the chain:

$$R_n = \frac{1}{2q_n^2} = \frac{1}{2(n+1)^2}, \tag{1.6}$$

where $q_n = 2(-1)^{n+1}q_{n-1} + q_{n-2}$, $q_{-2} = 1$, $q_{-1} = 0$ by Wallis-Euler recursion and by induction

$$q_n = i^{n^2-n}(n+1).$$

The cases $n = -2$ and $n = -1$ are obvious and we assume that the formula in question is also true for all $k \leq n$. Then

$$\begin{aligned}
q_{n+1} &= 2(-1)^{n+2}i^{n^2-n}(n+1) + i^{(n-1)^2-n+1}n \\
&= n\left(2(-1)^n i^{n^2-n} + i^{n^2-3n+2}\right) + 2(-1)^n i^{n^2-n} \\
&= ni^{n^2+n}\left(2(-1)^n i^{-2n} + i^{-4n+2}\right) + 2i^{n^2+n}(-1)^n i^{-2n} \\
&= ni^{n^2+n}\left(2(-1)^n(-1)^n + 1(-1)\right) + 2i^{n^2+n}(-1)^n(-1)^n \\
&= ni^{n^2+n} + 2i^{n^2+n} \\
&= i^{(n+1)^2-(n+1)}(n+1+1)
\end{aligned}$$

and so the formula holds for $n+1$, whenever it holds for n and $n-1$. It follows by induction that it is valid for all $n \geq -2$.
Similarly, we get the bases of the horospheres in the chain:

$$\frac{p_n}{q_n} = \frac{n}{n+1}.$$

Using the formula for radii Eq.1.6, we get the maximum possible tilt at n steps:

$$\arccos\frac{R_n - R_{n+1}}{R_n + R_{n+1}} = \arccos\frac{\frac{1}{2n^2} - \frac{1}{2(n+1)^2}}{\frac{1}{2n^2} + \frac{1}{2(n+1)^2}} = \arccos\frac{2n+1}{2n^2+2n+1}.$$

In Figure 1.18 there are shown maximum possible angles of tilt for the beginning of the chain of horospheres.
Letting $n \to \infty$ the radii of the horospheres approach 0, the tilt approaches 90° and the distance travelled by the bases approaches 1, i.e. the continued fraction converges, $[0, 2, -2, 2, -2, \ldots] = 1$. Alternatively, this can be also seen from:

$$x = \frac{1}{2 + \frac{1}{-2+x}} = \frac{x-2}{2x-3} \implies (x-1)^2 = 0 \implies x = 1.$$

Thus the supremum of the distance the chain of horospheres can travel is approached by using the lower bound $|b_n| = 2$ and alternating signs in case of real b_n to get to the antipodal sides. See figure 1.9 which illustrates that such a chain travels a unit distance from its starting point.

■

Remark. *Letting just one 1 or -1 in the continued fraction allows us construct a continued fraction that diverges to infinity, for example:*

$$[2; 1, \overline{-2, 2}]$$

diverges to $+\infty$.

Remark. *The bound* $|b_n| \geq 2$ *is sharp for* $b_n \in \mathbb{R}$ *as for example* $[0, 2-\epsilon, -2+\epsilon, 2-\epsilon, -2+\epsilon, \ldots]$ *diverges to* $+\infty$ *for any* ϵ. *An example of the beginning of the chain of horocycles for the case* $\epsilon = 0.1$ *is seen in Figure 1.19.*

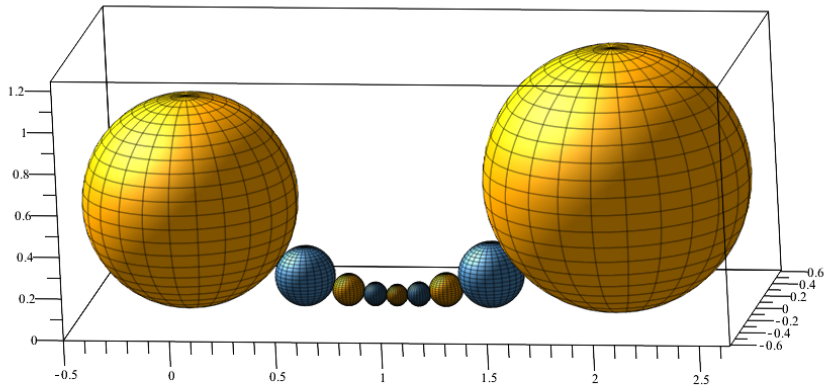

Figure 1.19:
First 9 horospheres of the chain for $[0, 2-\epsilon, -2+\epsilon, 2-\epsilon, -2+\epsilon, \ldots]$ with $\epsilon = 0.1$. The chain diverges to $+\infty$.

Remark. *The lower bound of the speed of convergence of continued fraction* $\mathbf{K}_{n=0}^{\infty}(b_n)$ *with* $|b_n| \geq 2$ *is logarithmic convergence. It follows from the limits:*

$$\lim_{k\to\infty} \frac{\frac{p_{k+1}}{q_{k+1}} - 1}{\frac{p_k}{q_k} - 1} = \lim_{k\to\infty} \frac{\frac{k+1}{k+2} - 1}{\frac{k}{k+1} - 1} = \lim_{k\to\infty} \frac{k+1}{k+2} = 1 \text{ } \textit{(sublinearity)},$$

and

$$\lim_{k\to\infty} \frac{\frac{p_{k+2}}{q_{k+2}} - \frac{p_{k+1}}{q_{k+1}}}{\frac{p_{k+1}}{q_{k+1}} - \frac{p_k}{q_k}} = \lim_{k\to\infty} \frac{\frac{k+2}{k+3} - \frac{k+1}{k+2}}{\frac{k+1}{k+2} - \frac{k}{k+1}} = \lim_{k\to\infty} \frac{k+1}{k+3} = 1 \text{ } \textit{(logarithmic)}.$$

1.4 Inverse images of continued fractions

The inverse problem is also compelling, i.e. given a real or complex number we construct its continued fraction representation. It turns out that there are several procedures how to construct these inverse images.
The Euclidean algorithm leads to a simple continued fraction expression for every rational number. We could also imagine the process of construction of

the pre-image as repeated process of subtracting integer part that would be the partial quotient then taking the reciprocal of the remainder. Then subtracting integer part again giving the next partial quotient, etc. In case of a rational the process stops at some stage and in case of an irrational it goes on forever.
This method is easily implemented when we know the decimal expansion of the irrational, for example $e = 2.718281828\ldots$. Then $b_0 = 2$, $b_1 = \left[\frac{1}{0.718281828\ldots}\right] = 1$, etc.
In case of quadratic irrationals we don't need to know the decimal expansion, just the integer part is enough. For example $x = \sqrt{23} \approx 4.8$.
Then $x_0 = \sqrt{23} = 4 + (\sqrt{23} - 4) \implies b_0 = 4$.
$x_1 = \frac{1}{x_0 - [x_0]} = \frac{1}{\sqrt{23}-4} = \frac{\sqrt{23}+4}{7} = 1 + \frac{\sqrt{23}-3}{7} \implies b_1 = 1$.
$x_2 = \frac{1}{x_1 - [x_1]} = \frac{7}{\sqrt{23}-3} = \frac{\sqrt{23}+3}{2} = 3 + \frac{\sqrt{23}-3}{2} \implies b_2 = 3$, etc.
The continued fraction of a quadratic irrational is (eventually) periodic but pure square roots of non-square numbers have a particular pattern:

$$[b_0; \overline{b_1, b_2, b_3, b_4, b_5, \ldots, b_5, b_4, b_3, b_2, 2b_0}],$$

where the periodic part starts right after b_0 and the periodic part is almost a palindrome with the exception at the end where we have always $2b_0$.
For example $\sqrt{23} = [4; \overline{1, 3, 1, 8}]$, $\sqrt{26} = [5; \overline{10}]$. But for those who are lazy there is an almost effortless formula:

$$\sqrt{z} = \sqrt{x^2 + y} = x + \cfrac{y}{2x + \cfrac{y}{2x + \cfrac{y}{2x + \cfrac{y}{2x + \cfrac{y}{\ddots}}}}}.$$

The formula above follows from the identity:

$$\sqrt{z} = x + \frac{z - x^2}{x + \sqrt{z}}$$

and denoting $y := z - x^2$.
For the irrationals that are not quadratic and we don't have their decimal expansion at hand, we can still calculate the continued fraction representation if we could recognise the number as an alternating series of the form:

$$b_0 + \sum_{n=0}^{\infty} \frac{(-1)^n}{q_n q_{n+1}} = [b_0; b_1, b_2, \ldots].$$

This representation follows from summing up the terms in Eq.1.3.
Another version on the same theme is:

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{b_n} = \cfrac{1}{b_1 + \cfrac{b_1^2}{b_2 - b_1 + \cfrac{b_2^2}{b_3 - b_2 + \cfrac{b_3^2}{b_4 - b_3 + \cfrac{b_4^2}{\ddots}}}}}.$$

For example:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \cfrac{1}{1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \ddots}}}}},$$

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots = \cfrac{1}{1 + \cfrac{1^2}{1 + \cfrac{2^2}{1 + \cfrac{3^2}{1 + \cfrac{4^2}{1 + \ddots}}}}}.$$

If we allow $b_n \in \mathbb{Z}$ then we have multiple of choices for pre-images. For example we could instead of subtracting the integer part subtract the nearest integer. Then clearly the partial quotients can also be negative, but we can get much faster convergence since none of the partial quotients have absolute value less than 2. In other words, nearest integer continued fractions are devoid of numbers ± 1 and so by Theorem 2 they always converge, whereas Theorem 1 is not applicable since some b_n can be negative.
Or if we want to make life harder we could always subtract the faraway integer. This leads to the pre-images that have the partial quotients that have values only in the double binary system $\pm 1, \pm 2$.
Or we could always subtract the even integer leading to the pre-image that has only even partial quotients.
For example the standard and nearest integer continued fraction of $\frac{7}{11}$:

$$\frac{7}{11} = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{3}}}} = [0; 1, 1, 1, 3] = \cfrac{1}{1 - \cfrac{1}{3 - \cfrac{1}{4}}} = \text{NI}[1; -3, 4]$$

and e:

$$e = [2; 1, 2, 1, 1, 4, 1, \ldots] = \text{NI}[3; -4, 2, -5, 2, -7, 2, -9, 2, -11, 2, -13, 2, \ldots].$$

The even integer continued fraction of $\frac{7}{11}$:

$$\frac{7}{11} = \cfrac{1}{2 - \cfrac{1}{2 + \cfrac{1}{4 - \cfrac{1}{2 - \cfrac{1}{2 + \cfrac{1}{\ddots}}}}}} = \text{EI}[0; 2, -2, -4, 2, -2, 2, -2, \ldots].$$

In case of complex numbers we are not as free when it comes to choosing which Gaussian integer to subtract since we need for the complex inversion a complex number that has modulus less than 1. As seen in Figure 1.20, in the center region we are free to choose but at the four border regions we have only a choice of two. Note also that the circular arcs with radius $\frac{1}{2}$ distinguish the regions for each corner where the inversion gives at least 2 as integer part and in the center region it is always 1, no matter which corner we choose.
For complex numbers the algorithm for the nearest Gaussian integer continued fraction is the standard. Other possibilities would be to choose always even-even/odd-odd Gaussian integer or always mixed Gaussian integers.
Examples of continued fractions for complex numbers $\frac{7}{11} + \frac{11}{7}i$ and $\pi + e \cdot i$:

$$\frac{7}{11} + \frac{11}{7}i = 1 + 2i - \cfrac{i}{1 + i + \cfrac{i}{1 + 2i + \cfrac{i}{3 - \cfrac{1}{3 - \cfrac{1}{1 + 2i}}}}},$$

$$\pi + e \cdot i = 3 + 3i + \cfrac{1}{1 + 3i + \cfrac{1}{2 + i + \cfrac{1}{1 + 5i + \cfrac{1}{1 + 4i - \cfrac{1}{1 + 2i - \cfrac{1}{1 + 2i - \cfrac{1}{\ddots}}}}}}}.$$

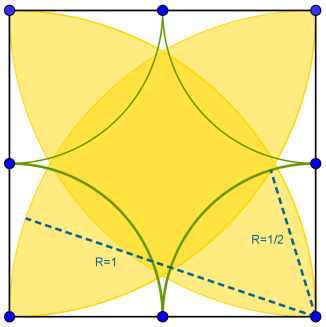

Figure 1.20:
Unit in the Gaussian lattice. If the complex number is in the dark yellow region then we are free to choose the corner relative to which we can invert. In the white border regions we have only choice of two corners. The circular arcs with radius $\frac{1}{2}$ distinguish the regions for each corner where the inversion gives at least 2 as integer part.

Chapter 2

Continued fraction method of factoring integers

Decomposition of a composite number into a product of smaller integers has been of interest through the centuries. The method of using continued fractions for factorization was described by Legendre in his *Théorie des Nombres* of 1798, developed further by Maurice Kraïtchik, Derrick Henry Lehmer and Ralph Ernest Powers in the 1920s. Translation of the method into a computer algorithm by John Brillhart and Michael A. Morrison made factoring of 50-digit numbers commonplace in the 1970s. Their heyday started with factoring the seventh Fermat number $F_7 = 2^{128} + 1$, a 39 digit number, on September 13, 1970.
For all currently known integer factorization methods the hardest numbers to factor are semi-primes, the product of two prime numbers. As of December 12, 2009 the largest semi-prime factored is RSA-768, a 232-digit number, using the general number field sieve. The effort took about 3 calendar years for a six-institution research team. As of today, the next RSA Factoring Challenge semi-prime RSA-896 with 270 decimal digits, is unsolved.
We shall start the description of the continued fraction method of factorization by recalling the method of Fermat. Then turn to prove the lemma on continued fractions that is the base for the continued fraction factorization method and prove a corollary that reduces the factorization problem by a factor of 2. Then we give details of the continued fraction factorization algorithm and finish with an example calculation using both the original and modified methods.

2.1 Fermat's factorization method

We assume that n is an odd integer, not a perfect square, that we are trying to factorize. Each odd number has a representation as a difference of two squares:

$$n = x^2 - y^2 \iff n = (x - y)(x + y)$$

and unless $x - y = 1$, we have a proper factorization if we could find these x, y. This representation exists, since if $n = ab$ then $n = (\frac{a+b}{2})^2 - (\frac{a-b}{2})^2$ and divisions by 2 give integers since both a, b are odd.

As $x > \sqrt{n}$ we start the trial and error with $x = [\sqrt{n}] + 1$, check whether $x^2 - n$ is a perfect square. If yes, then we are done, if not, then we add one to x and repeat the check. The search will end sooner or later, at the latest at $x = \frac{n+1}{2}$. One way to speed up the search is instead of looking for x, y such that $x^2 - y^2 = n$ is to look for x, y such that $x^2 \equiv y^2 \pmod{n}$. For a random such pair, x, y, the probability is at least $\frac{1}{2}$ that n is factored by writing it as a product $\gcd(x - y, n) \cdot \gcd(x + y, n)$.
Fermat's factorization method is the foundation on which the continued fraction factorization is built besides the quadratic sieve and general number field sieve, the currently best-known factoring algorithms.

2.2 A lemma on continued fractions

Lorentzen and Waadeland in [3] (page 427-428) have a proof of the lemma below, we give an alternative geometric proof using Ford circles.

Lemma 3. *Let $\xi > 1$ be an irrational number, and $(\frac{p_k}{q_k})_{k=0}^{\infty}$ its sequence of simple continued fraction convergents. Then:*

$$|p_k^2 - \xi^2 q_k^2| < 2\xi \text{ for all } k \in \mathbb{N}. \tag{2.1}$$

Proof. Let us choose one of the convergents of ξ, say $\frac{p}{q}$. Then

$$|p^2 - \xi^2 q^2| = |p + \xi q||p - \xi q| = q^2 \left|\frac{p}{q} + \xi\right| \left|\frac{p}{q} - \xi\right| = \frac{1}{2R}\left(\frac{p}{q}\left|\frac{p}{q} - \xi\right| + \xi\left|\frac{p}{q} - \xi\right|\right)$$

where $R = \frac{1}{2q^2}$ is the radius of Ford circle of $\frac{p}{q}$.

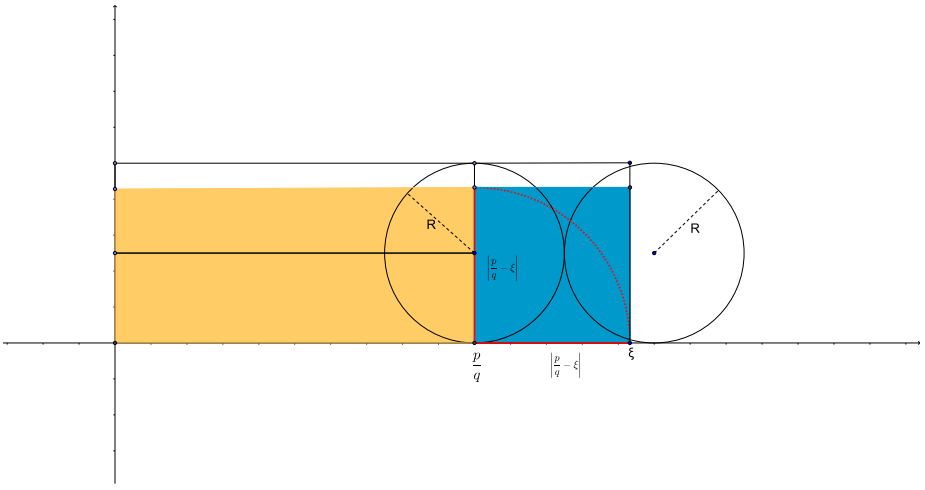

Figure 2.1:
Real ξ is to the right of its convergent $\frac{p}{q}$. The product $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ is represented as the yellow area and it is dominated by the area $\frac{p}{q} \cdot 2R$. The area $\xi \cdot \left|\frac{p}{q} - \xi\right|$ is the sum of yellow and blue areas and it is dominated by the area $\xi \cdot 2R$.

If ξ is to the right of its convergent $\frac{p}{q}$ as shown in Figure 2.1, then

$$|p^2 - \xi^2 q^2| = \frac{1}{2R}\left(\frac{p}{q}\left|\frac{p}{q} - \xi\right| + \xi\left|\frac{p}{q} - \xi\right|\right) < \frac{1}{2R}\left(\frac{p}{q}2R + \xi 2R\right) < 2\xi$$

since $\left|\frac{p}{q} - \xi\right| < 2R$ and $\frac{p}{q} < \xi$.
The dominance of $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ by $\frac{p}{q} \cdot 2R$ and $\xi \cdot \left|\frac{p}{q} - \xi\right|$ by $\xi \cdot 2R$ is illustrated in Figure 2.1. The distance from $\frac{p}{q}$ to ξ is dominated by $2R$ since this is the maximal distance between consecutive convergents in case of simple continued fractions. This proves our statement for the case ξ is to the right of $\frac{p}{q}$.
To see that no better bound is possible notice that the maximal distance $2R$ can be approached with the following configuration of partial quotients $b_{2n} \to \infty$, $b_{2n+1} = 1$, and $b_{2n+2} \to \infty$. In terms of Ford's clockwork it is the case when an upright even horosphere is followed by a horosphere of the same size and then by another upright horosphere.

If ξ is to the left of its convergent $\frac{p}{q}$ as shown in Figure 2.2, then $\left|\frac{p}{q} - \xi\right| < 2R$ and $\frac{p}{q} < \xi + 2R$.

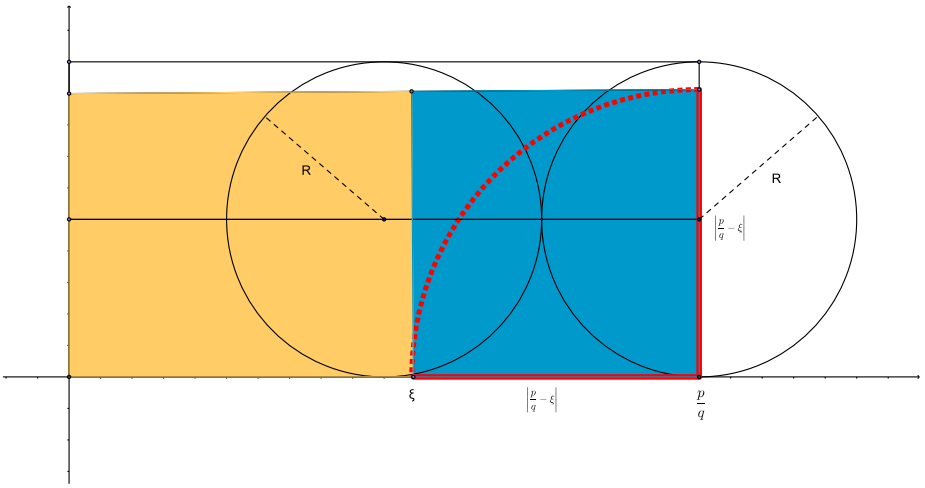

Figure 2.2:
Real ξ is to the left of its convergent. The product $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ is represented as the sum of yellow and blue areas and it is dominated by the area $(\xi + 2R) \cdot 2R$.
The area $\xi \cdot \left|\frac{p}{q} - \xi\right|$ is painted yellow and it is dominated by the area $\xi \cdot 2R$.

The dominance of $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ by $(\xi + 2R) \cdot 2R$ and $\xi \cdot \left|\frac{p}{q} - \xi\right|$ by $\xi \cdot 2R$ is illustrated in Figure 2.2. The distance from $\frac{p}{q}$ to ξ is again dominated by $2R$ since this is the maximal distance between consecutive convergents in case of simple continued fractions. This proves our statement for the case ξ is to the left of $\frac{p}{q}$ with $R \to \infty$.
To see that no better bound is possible notice that the maximal distance $2R$

can be approached this time with the following configuration of partial quotients $b_{2n+1} \to \infty$, $b_{2n+2} = 1$, and $b_{2n+3} \to \infty$. In terms of Ford's clockwork it is the case when an upright odd horosphere is followed by a horosphere of the same size and then by another upright horosphere.

But whenever a partial quotient goes to ∞, the radius of the corresponding horosphere approaches 0. Therefore we must take the limit $R \to 0$ to get the bound:

$$|p^2 - \xi^2 q^2| = \frac{1}{2R}\left(\frac{p}{q}\left|\frac{p}{q} - \xi\right| + \xi\left|\frac{p}{q} - \xi\right|\right) < \lim_{R\to 0} \frac{1}{2R}\left((\xi + 2R)2R + \xi 2R\right) = 2\xi.$$

■

If we choose nearest integer continued fraction instead, then we get a better result.

Corollary 3.1. *Let $\xi > 1$ be an irrational number, and $(\frac{p_k}{q_k})_{k=0}^{\infty}$ its sequence of nearest integer continued fraction convergents. Then:*

$$|p_k^2 - \xi^2 q_k^2| < \xi \text{ for all } k \in \mathbb{N}. \tag{2.2}$$

Proof. If ξ is to the right of its convergent $\frac{p}{q}$ as shown in Figure 2.3, then

$$|p_k^2 - \xi^2 q_k^2| = \frac{1}{2R}\left(\frac{p}{q}\left|\frac{p}{q} - \xi\right| + \xi\left|\frac{p}{q} - \xi\right|\right) < \frac{1}{2R}\left(\frac{p}{q}R + \xi R\right) < \xi.$$

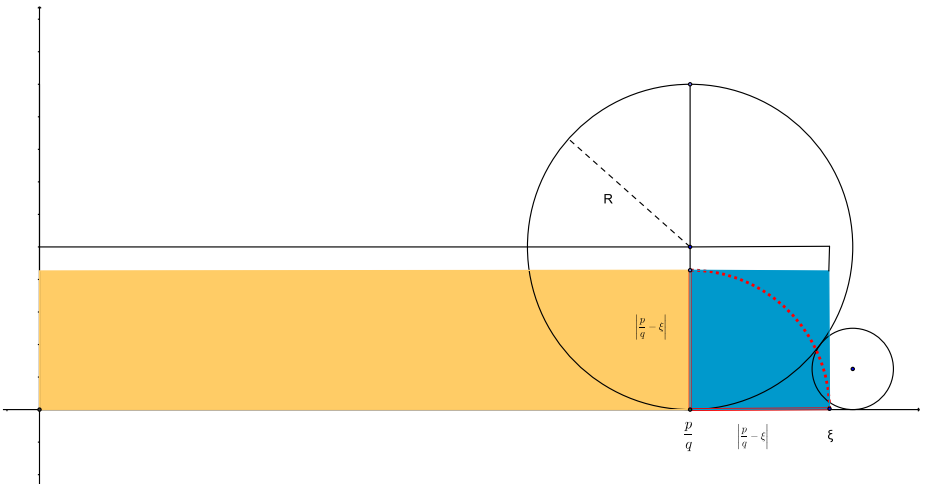

Figure 2.3:
Nearest integer continued fraction. Real ξ is to the right of its convergent. The product $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ is represented as the yellow painted area and it is dominated by the area $\frac{p}{q} \cdot R$. The area $\xi \cdot \left|\frac{p}{q} - \xi\right|$ is the sum of yellow and blue areas and it is dominated by the area $\xi \cdot R$.

The dominance of $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ by $\frac{p}{q} \cdot R$ and $\xi \cdot \left|\frac{p}{q} - \xi\right|$ by $\xi \cdot R$ is illustrated in Figure 2.3. The distance from $\frac{p}{q}$ to ξ is dominated by R since this is the maximal

distance between consecutive convergents in case of nearest integer continued fractions. This proves our statement for the case ξ is to the right of $\frac{p}{q}$.
To see that no better bound is possible notice that the maximal distance R can be approached with the following configuration of partial quotients $b_{2n} \to \infty$, $b_{2n+1} = 2$, and $b_{2n+2} \to \infty$. In terms of Ford's clockwork it is the case when an upright even horosphere is followed by a horosphere with 4 times smaller radius and then by another upright horosphere.

If ξ is to the left of its convergent $\frac{p}{q}$ as shown in Figure 2.4, then $\left|\frac{p}{q} - \xi\right| < R$ and $\frac{p}{q} < \xi + R$.

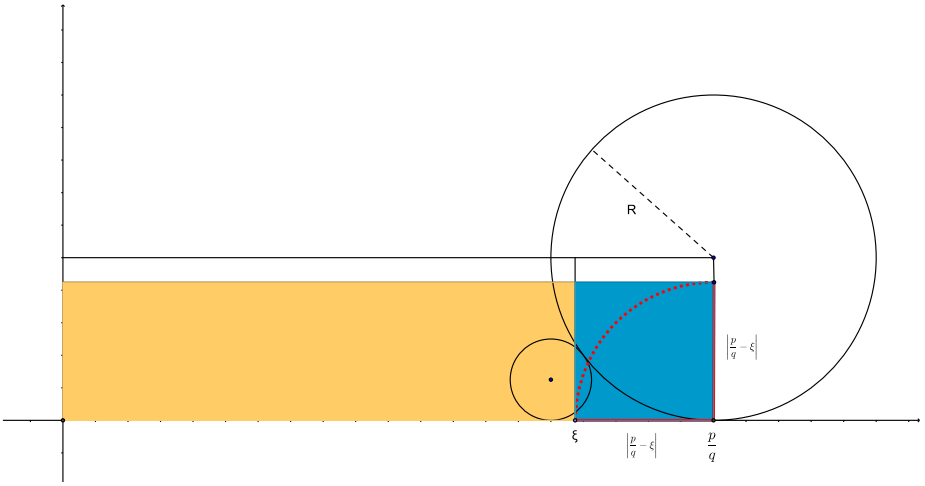

Figure 2.4:
Nearest integer continued fraction. Real ξ is to the right of its convergent. The product $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ is represented as the sum of yellow and blue painted areas and it is dominated by the area $(\xi + R) \cdot R$. The area $\xi \cdot \left|\frac{p}{q} - \xi\right|$ is painted yellow and it is dominated by the area $\xi \cdot R$.

The dominance of $\frac{p}{q} \cdot \left|\frac{p}{q} - \xi\right|$ by $(\xi + R) \cdot R$ and $\xi \cdot \left|\frac{p}{q} - \xi\right|$ by $\xi \cdot R$ is illustrated in Figure 2.4. The distance from $\frac{p}{q}$ to ξ is again dominated by R since this is the maximal distance between consecutive convergents in case of nearest integer continued fractions. This proves our statement for the case ξ is to the left of $\frac{p}{q}$ with $R \to \infty$.
To see that no better bound is possible notice that the maximal distance R can be approached this time with the following configuration of partial quotients $b_{2n+1} \to \infty$, $b_{2n+2} = 2$, and $b_{2n+3} \to \infty$. In terms of Ford's clockwork it is the case when an upright odd horosphere is followed by a horosphere with 4 times smaller radius and then by another upright horosphere.
But whenever a partial quotient approaches ∞, the radius of the corresponding horosphere approaches 0. Therefore we can take the limit $R \to 0$ to get a tighter

bound:

$$|p_k^2 - \xi^2 q_k^2| = \frac{1}{2R}\left(\frac{p}{q}\left|\frac{p}{q} - \xi\right| + \xi\left|\frac{p}{q} - \xi\right|\right) < \lim_{R\to 0}\frac{1}{2R}\left((\xi + R)R + \xi R\right) = \xi.$$

■

2.3 Continued fraction factorization

In the description of the factorisation method we follow Lorentzen and Waadeland in [3] §IX.3.
The main effort of the Fermat's method is to find integers x, y such that $x^2 \equiv y^2 \pmod{n}$.
If we go and calculate the simple continued fraction pre-image of $\sqrt{n}$ and from it some nominators of convergents p_k, raise them to the second power and reduce them $\pmod{n}$, then without knowing the lemma we would expect them to be:

$$-\frac{n}{2} < p_k^2 \pmod{n} < \frac{n}{2}.$$

But using the lemma on the assumption $\xi^2 = n$, in which case we have

$$|p_k^2 - nq_k^2| < 2\sqrt{n},$$

we see a miracle happening:

$$-2\sqrt{n} < p_k^2 \pmod{n} < 2\sqrt{n}.$$

Hence with the help of continued fractions we can make factoring much easier reducing the problem of factoring number with d digits to the problem of factoring a set of numbers with $\frac{d}{2}$ digits.
As a next step we need to factor fully these $p_k^2 \pmod{n}$ and then concentrate on the exponents in the prime factorization. There is too much information in these exponents, so we reduce them $\pmod 2$. Now every $p_k^2 \pmod{n}$ is associated to a vector of zeros and ones.
If there is among them a linearly dependent set of vectors then the product of corresponding set of $p_k^2 \pmod{n}$ is a square, say y^2. Let x^2 be the square of the product of corresponding p_k. Unless $x \equiv \pm y \pmod{n}$ we are done, since $\gcd(x + y, n$ or $\gcd(x - y, n)$ is a proper factor in n, which is then easily found by the Euclidean algorithm.
If we didn't find a linearly dependent set of the binary exponent vectors or if we reached the situation $x \equiv \pm y \pmod{n}$ then we have to continue with the continued fraction expansion and run through the previous steps again.
Note that using the nearest integer continued fractions and the Corollary 3.1 instead of the lemma we get smaller bounds for $p_k^2 \pmod{n}$:

$$-\sqrt{n} < p_k^2 \pmod{n} < \sqrt{n}$$

making the factorization problem slightly easier to solve.

2.4 Example of continued fraction factorization

Let $n = 9073$. Then $\sqrt{n} = [95; 3, 1, 26, 2, \ldots]$ and preparatory calculations are in Table 2.1.

Table 2.1: Calculations for factoring 9073 using the simple continued fraction of $\sqrt{9073} = [95; 3, 1, 26, 2, \ldots]$.

k	-2	-1	0	1	2	3	4	...
b_k			95	3	1	26	2	...
p_k	0	1	95	286	381	1119	2619	...
$p_k^2 \pmod n$			-48	139	-7	87	-27	...
factorization			$(-1)2^4 3$	139	(-1)7	$3 \cdot 29$	$(-1)3^3$	...
interesting			$\star$				$\star$	...

The nominators of convergents are calculated using Wallis-Euler recursion:

$$p_{-2} = 0, p_{-1} = 1 \text{ and } p_k = b_k p_{k-1} + p_{k-2}.$$

After squaring and reducing $\pmod n$ we calculated the factorizations of $p_k^2 \pmod n$. Note that the bounds for $p_k^2 \pmod{9073}$ are $\pm\left\lfloor 2\sqrt{9073}\right\rfloor = \pm 190$ instead of $\pm\left\lfloor \frac{9073}{2}\right\rfloor = \pm 4536$.
Next we need to compare all the factorizations of $p_k^2 \pmod n$ and if we find factors that occur more than once then we mark the corresponding factorizations interesting. All the different factors in these interesting factorizations form the factor base, in our case it is $\{-1, 2, 3\}$.
The following step is to reduce the powers of factors $\pmod 2$ and form the exponent vectors of the interesting factorizations with respect to the factor base, in our case as follows:

$$95 \to -48 = (-1)2^4 3 \to \text{ exponent vector } \langle 1, 0, 1\rangle,$$

$$2619 \to -27 = (-1)3^3 \to \text{ exponent vector } \langle 1, 0, 1\rangle.$$

Luckily, the vectors are linearly dependent as their sum is the zero vector $\langle 0, 0, 0\rangle$. This implies that the product of interesting factorizations is a square:

$$y^2 = (-48)(-27) = (-1)^2 2^4 3^4,$$

and taking the square root:

$$y = (-1)2^2 3^2 = -36.$$

Next we multiply the corresponding nominators p_k and reduce $\pmod n$:

$$x \equiv 95 \cdot 2619 \pmod n \implies x = 3834.$$

Luckily, $x \not\equiv \pm y \pmod n$, and so we have found a proper factor:

$$\gcd(x - y, n) = \gcd(3870, 9073) = 43.$$

The full factorization follows easily:

$$9073 = 43 \cdot 211.$$

In the factorization process we were two times lucky, but if it had not been the case then we would have needed to go back to calculating the next nominator of the convergent, p_k, in the continued fraction of $\sqrt{9073}$ and repeat the subsequent steps of the algorithm.
Let's try the same method but now with the nearest integer continued fraction $\sqrt{n} = \text{NI}[95; 4, -27, -2, -7, \ldots]$. The preparatory calculations are in Table 2.2.

Table 2.2: Calculations for factoring 9073 using the nearest integer continued fraction of $\sqrt{9073} = \text{NI}[95; 4, -27, -2, -7, \ldots]$.

k	-2	-1	0	1	2	3	4	...
b_k			95	4	-27	-2	-7	...
p_k	0	1	95	381	-10192	20765	-155547	...
$p_k^2 \pmod n$			-48	-7	87	-27	-88	...
interesting			$\star$			$\star$		...

This time the bounds for $p_k^2 \pmod{9073}$ are $\pm \left\lfloor \sqrt{9073} \right\rceil = \pm 95$ compared to the bounds ± 190 we had with simple continued fraction.
By the same steps and calculations as before we get the same result and even faster with the nearest integer continued fraction, suggesting that the modification might be propitious.

Chapter 3

Dirichlet's approximation theorem

The distinction between numbers being rational or irrational, algebraic or transcendental have been studied by means of Diophantine approximation, that is the investigation how closely, or with what degree of accuracy, can a given number be approximated by rational numbers or algebraic numbers.
Dirichlet's approximation theorem is one of the cornerstone results in Diophantine approximation which tells us that rational numbers can approximate arbitrary real numbers quite well in terms of the size of the denominator of the approximate rational numbers.
First, if we consider all rational numbers with a fixed denominator and try to approximate a given irrational α with them, then clearly α lies between two such rational numbers, say $\frac{r}{q} < \alpha < \frac{r+1}{q}$ and so:

$$\left|\alpha - \frac{r}{q}\right| < \frac{1}{q},$$

or if we choose properly $p = r$ or $p = r + 1$, whichever is closer, then:

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q}.$$

We say that $\frac{p}{q}$ is a *good approximation* to α if for all $\frac{a}{b}$ with $0 < b \leq q$ we have:

$$\left|\alpha - \frac{p}{q}\right| < \left|\alpha - \frac{a}{b}\right|$$

and if $\frac{p}{q}$ can satisfy even a stronger requirement:

$$q\left|\alpha - \frac{p}{q}\right| < b\left|\alpha - \frac{a}{b}\right|$$

then we call $\frac{p}{q}$ the *best approximation* to α. It turns out that the convergents of the continued fraction representation of α are precisely the best approximations to α.

Next, we will prove the Dirichlet's approximation theorem using Ford and Short circles. The original proof by Dirichlet from 1842 uses the pigeon-hole principle.

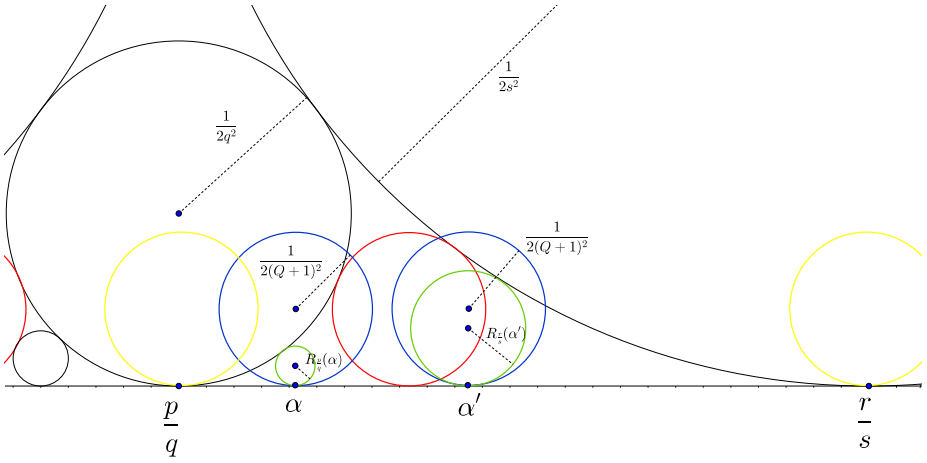

Figure 3.1:
Red circle is a Ford circle of $\frac{n}{Q+1}$ for some n. Blue circles are examples where the Ford circle is translated to α and to α' and green circles are Short circles. Yellow circles are examples where the Ford circle of $\frac{1}{Q+1}$ is inside a larger Ford circle.

Theorem 4. *Let α be a real number and Q be a positive integer. Then there exist integers p, q such that $1 \leq q \leq Q$, and*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q(Q+1)}. \tag{3.1}$$

Proof. In the beginning of §1.2 (Eq.1.2) we saw that the Short circle of α relative to $\frac{p}{q}$ has radius:

$$R_{\frac{p}{q}}(\alpha) = \frac{q^2}{2}\left|\alpha - \frac{p}{q}\right|^2$$

and so we can combine the two inequalities and restate the theorem as an existence of a rational number $\frac{p}{q}$ for the given α and Q such that:

$$R_{\frac{p}{q}}(\alpha) \leq \frac{1}{2(Q+1)^2} < \frac{1}{2q^2}.$$

It means that we have to find a rational number $\frac{p}{q}$ such that its Ford circle is strictly bigger than the Ford circle of $\frac{1}{Q+1}$ and the Short circle of α relative to $\frac{p}{q}$ is smaller or equal to the Ford circle of $\frac{1}{Q+1}$.
Every Ford circle has infinitely many adjacent Ford circles and if we consider a rational number $\frac{a}{b}$ with $b > 1$ then we see from the Eq.1.5 that the Ford circle of $\frac{a}{b}$ has exactly two adjacent Ford circles larger than itself. Namely, $|B+nb| < |b|$, i.e. $|n + \frac{B}{b}| < 1$ for exactly two values of n, the integers between which $-\frac{B}{b}$ lies. We have given Q which fixes a Ford circle with radius $\frac{1}{2(Q+1)^2}$. If we translate this circle along the x-axis then there are exactly three possibilities:

- the circle is inside a larger Ford circle (yellow circles in Figure 3.1),
- the circle is tangent to two larger Ford circles (red circle in Figure 3.1),
- the circle is overlapping with some larger Ford circle (two examples in Figure 3.1 with blue circles).

If we have that after the translation of the Ford circle of $\frac{1}{Q+1}$ is inside a larger Ford circle, then α satisfies itself Dirichlet's theorem. This is the case when α is of the form $\frac{n}{Q-m}$ for any $n > 0$ and $0 \leq m < Q$.
In the second case, if we have that after the translation of the Ford circle with radius $\frac{1}{2(Q+1)^2}$, its base is at α and it is tangent to two larger Ford circles, then in respect of both of them we have:

$$R_{\frac{p}{q}}(\alpha) = \frac{1}{2(Q+1)^2} < \frac{1}{2q^2}.$$

This is the case when α is a rational number of the form $\frac{n}{Q+1}$ for any $n > 0$ and $(Q+1) \nmid n$. Such $\alpha = \frac{n}{Q+1}$ is adjacent to two Ford circles, $\frac{k}{Q}$ and $\frac{l}{Q-m}$, where $1 \leq m < Q$ and k, l some positive integers. With the former:

$$R_{\frac{k}{Q}}(\frac{n}{Q+1}) = \frac{Q^2}{2}\left|\frac{n}{Q+1} - \frac{k}{Q}\right|^2 = \frac{1}{2(Q+1)^2} < \frac{1}{2Q^2},$$

and similarly with the latter:

$$R_{\frac{l}{Q-m}}(\frac{n}{Q+1}) = \frac{(Q-m)^2}{2}\left|\frac{n}{Q+1} - \frac{l}{Q-m}\right|^2 = \frac{1}{2(Q+1)^2} < \frac{1}{2Q^2}.$$

Thus we have choice two possibilities of which the rational with smaller Ford circle approximates clearly more closely then the other.
Finally, if we have that after the translation of the Ford circle with radius $\frac{1}{2(Q+1)^2}$ is overlapping some larger Ford circle, then we choose the rational number $\frac{p}{q}$ that gives rise to this larger Ford circle and have our Short circle relative to that:

$$R_{\frac{p}{q}}(\alpha) < \frac{1}{2(Q+1)^2} < \frac{1}{2q^2}.$$

This is the case when α is not of the form $\frac{n}{Q+1}$ for any $n > 0$ nor of the form $\frac{n}{Q-m}$ for any $n > 0$ and $0 \leq m < Q$. That means α is either an irrational or a rational $\frac{a}{b}$ with $b > Q+1$.

■

Two examples are shown in Figure 3.1. For given α and Q we have the left blue circle which is the Ford circle of $\frac{1}{Q+1}$ that is translated to α and the left green circle which is the Short circle of α relative to $\frac{p}{q}$. This Short circle is tangent to x-axis at α and to the Ford circle of $\frac{p}{q}$. Similarly, for given α' and Q we have the right blue circle which is the Ford circle of $\frac{1}{Q+1}$ that is translated to α' and the right green circle which is the Short circle of α' relative to $\frac{r}{s}$. This Short circle is tangent to x-axis at α' and to the Ford circle of $\frac{r}{s}$.
How, then, can we find a rational number $\frac{p}{q}$ for the given α and Q that satisfies Dirichlet's theorem?

One possible way is to use the convergents of the continued fraction representation of α. If n is the index satisfying $q_n \leq Q < q_{n+1}$, then:

$$\left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{q_n(Q+1)}.$$

This can be seen using Eq.1.3 that implies:

$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \leq \frac{1}{q_n q_{n+1}}.$$

References

[1] Alan Frank Beardon, Ian Short,
A Geometric Representation of Continued Fractions, *The American Mathematical Monthly* **121** (2014) 391-402.

[2] Lester Randolph Ford,
Fractions, *The American Mathematical Monthly* **45** (1938) 586-601.

[3] Lisa Lorentzen, Haakon Waadeland,
Continued Fractions with Applications, Elsevier Science Publishers, Amsterdam, 1992.

[4] Ian Short,
Ford Circles, Continued Fractions, and Rational Approximation, *The American Mathematical Monthly* **118** (2011) 130-135.