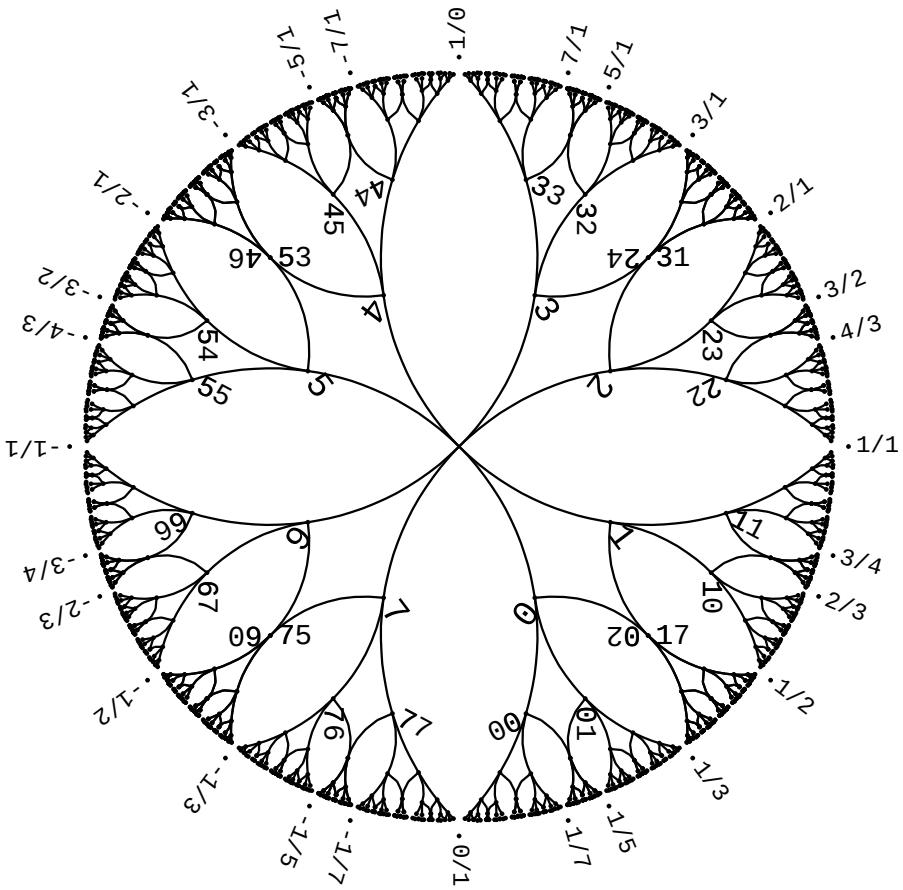
EXACT ARITHMETICAL ALGORITHMS IN REAL NUMBER SYSTEMS

Petr Kůrka



Prague september 2015

Contents

Preface

Current computers work with real numbers in the floating point format and the numbers are rounded up after each arithmetical operation. This usually works quite well but there are cases in which the successive roundings yield wrong results. Exact arithmetical algorithms, on the other hand, work with real numbers specified to an arbitrary precision. The precision of the result depends on the precision of the operands. The theory of exact real computation is based on the concept of on-line algorithms whose inputs and outputs are infinite expansions of real numbers. The algorithms work in a loop in infinite time but each finite prefix of the output is computed in finite time from finite prefixes of the inputs.

The theory of on-line algorithms has been developped by Weihrauch [68]. The idea of on-line arithmetical algorithms has been suggested in an unpublished manuscript of Gosper [21] and developped by Kornerup and Matula [34] and Vuillemin [66]. On-line arithmetical algorithms are treated in the PhD thesis of Potts [55] and in the last chapter of the monograph of Kornerup and Matula [33]. The on-line algorithms do not work in the standard decadic or binary systems but they do work in redundant systems, for example in positional number systems whose number of digits is larger than the base.

The present book is a theoretical treatment of arbitrary precision on-line arithmetical algorithms in general Möbius number systems. To specify a Möbius number system, we start with a finite alphabet A of digits and to each digit we associate a Möbius transformation. This is a mapping of the form $M(x) = \frac{ax+b}{cx+d}$. For example in a positional number system with base $\beta > 1$, the linear transformation $F_a(x) = \frac{x+a}{\beta}$ is associated to the digit a. Then we specify a subshift $\Sigma \subseteq A^\omega$ of admissible infinite sequences of digits and the value mapping $\Phi : \Sigma \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is the extended real line. The value mapping Φ should be surjective and continuous. This means that each number $x \in \overline{\mathbb{R}}$ should have its symbolic representation (an infinite sequence of digits) $u \in \Sigma$ such that $\Phi(u) = x$. Continuity means that the prefixes $u_{[0,n)}$ of u of length n give with increasing n ever better approximations to $\Phi(u) = x$.

The first chapter is introductory and treats classical positional number systems and number systems based on continued fractions. On these examples it is shown how the Möbius transformations are assocoated to the digits, how the value mapping $\Phi : \Sigma \to \overline{\mathbb{R}}$ is constructed and how symbolic representations of real numbers are obtained.

The second chapter treats redundancy as a topological concept and shows that for every compact metric space X (in particular for the space $X = \overline{\mathbb{R}}$) there exists a redundant continuous surjective mapping $\Phi : \Sigma \to X$, where Σ is a symbolic space. The property of redundancy implies that each continuous mapping $G : X \to X$ has a symbolic representation, which is a continuous mapping $F : \Sigma \to \Sigma$ such that $\Phi \circ F = G \circ \Phi$. In arithmetical algorithms, the symbolic space Σ is supposed to be a sofic subshift recognizable by a finite automaton, so the rest of the chapter deals with sofic subshifts.

The third chapter explains basic ideas of projective geometry which gives insight into the spaces connected with a number system. The extended real line $\overline{\mathbb{R}}$ is identified with the one-dimensional projective space $\mathbb{P}(\mathbb{R}^2)$ and the space of Möbius transformations $\mathbb{M}(\mathbb{R})$ is identified

with the three-dimensional space of projective matrices $\mathbb{P}(\mathbb{R}^{2\times 2})$. The geometrical properties of Möbius transformations are exposed with the use of hyperbolic geometry. Then we explain the concept of representation of real numbers by a sequence of transformations: A sequence M_n of real Möbius transformations represents a real number x iff $x = \lim_{n\to\infty} M_n(z)$ for every complex number z with a nonzero imaginary part. In particular, if $\Phi : \Sigma \to \overline{\mathbb{R}}$ is the value mapping of a number system and $u = u_0u_1\cdots \in \Sigma$, then $\Phi(u)$ is represented by a sequence of transformations $F_{u_{[0,n)}} = F_{u_0} \circ \cdots \circ F_{u_{n-1}}$.

The fourth chapter exposes the theory of Möbius number systems and shows several methods how to construct suitable subshifts $\Sigma \subseteq A^\omega$ and suitable value mappings $\Phi : \Sigma \to \overline{\mathbb{R}}$. A special treatment is given to sofic Möbius number systems for which arithmetical algorithms work. Several examples of sofic number systems are given.

The fifth chapter develops the calculus of bilinear tensors which represent binary arithmetical operations. Intervals are represented by projective matrices and operations with tensors and intervals are based on matrix calculus. Based on this calculus we describe the unary algorithm which computes a Möbius transformation and the binary algorithm which computes a bilinear tensor.

The sixth chapter treats number systems whose matrices have integer entries. In particular, modular systems have transformations with unit determinant. We show that if the unary algorithm computes a transformation with integer entries in a modular number system, then the norm of the state matrices is bounded, so the computation can be carried out by a finite state transducer. On the other hand, Möbius transformations are the only rational functions which can be computed by a finite state transducer.

The seventh chapter treats number systems with matrices whose entries are algebraic numbers. We review the theory of algebraic extension fields, algebraic integers and integral bases and give classical results of Parry and Schmidt on positional number systems with algebraic base $\beta > 1$ (so called β-systems introduced by Rényi [58]). Then we treat the algorithms of parallel addition in positional number systems.

The eigth chapter treats algorithms which compute transcendent functions like e^x, $\ln x$, $\tan x$ or $\arctan x$. We review the theory of Padé approximants and the representation of transcendent functions by general continued fractions. We introduce the concept of algebraic tensor $T(x, y)$, which for a fixed y is a rational function of x and for a fixed x is a Möbius transformation of y. We define the transcendent algorithm which works with these algebraic tensors and we show that it computes transcendent functions which can be expressed by general continued fractions. Finally we treat algorithms which compute arithmetical expressions and iterative algorithms which compute stable fixed points of real functions.

The treatment is elementary and self-contained. The basic prerequisite is linear algebra and matrix calculus.

Chapter 1

Basic number systems

Real numbers are defined as cuts of rational numbers or as limits of Cauchy sequences of rational numbers. Alternatively, the space of real numbers is characterized axiomatically by the property of completeness: it is the smallest complete metric space which contains rational numbers. A real number is usually given by its expansion in the decadic number system. But a number should be distiguished from its representation in any number system. The concept of number is geometrical or analytical, the representation of a number is a combinatorial concept. A real number can have many representations in a number system.

1.1 The decadic system

In the **decadic number system**, a real number is represented by an infinite word (a string of letters or characters) $u = su_n u_{n+1} \cdots u_{-1}.u_0 u_1 u_2 \cdots$, where s is either the sign $-$ or empty, $n \leq 0$ is an integer, $u_i \in \{0,1,2,3,4,5,6,7,8,9\}$ are digits and . is the positional decimal point. Such a word u represents the number

$$x = \pm \sum_{i=n}^{\infty} u_i \cdot 10^{-i-1}.$$

We say that x is the **value** of u or that u is an **expansion** of x and we write $\Phi(u) = x$. We admit as expansions also finite words $u = su_n u_{n+1} \cdots u_{-1}.u_0 u_1 \cdots u_k$, which represent the same numbers as infinite words with trailing zeros: $\Phi(u) = \pm \sum_{i=n}^{k} u_i \cdot 10^{-i-1}$, e.g., $\Phi(.2) = \frac{1}{5}$ or $\Phi(-1.5) = \frac{-3}{2}$. A finite prefix $u_{|k} = su_n u_{n+1} \cdots u_{-1}.u_0 u_1 \cdots u_{k-1}$ of an expansion u of x with k decimal places gives an approximation of x:

$$|\Phi(u) - \Phi(u_{|k})| \leq \sum_{i=k}^{\infty} 9 \cdot 10^{-i-1} = \frac{9}{10^{k+1}(1 - \frac{1}{10})} = 10^{-k}.$$

This is essential, since neither people nor computers can handle infinite expansions but only their finite prefixes. To determine a real number, we have to give a rule or an algorithm which generates arbitrarily long prefixes of its expansion. Accordingly, we say that a real number is an **algorithmic number** if there exists an algorithm which computes its expansion to an arbitrary number of decimal places. Algorithmic numbers include all rational numbers, all algebraic numbers, which are solutions of algebraic equations with rational coefficients, and many transcendent numbers like π or e, which can be computed by power series.

The expansions are infinite words in the alphabet $A = \{0,1,2,3,4,5,6,7,8,9,.,-\}$, which contains besides the decimal digits also the positional point . and the negative sign $-$. We

denote by A^* the set of finite words (sequences) of letters of A and by A^ω the set of infinite words. Not every infinite word of A^ω represents a real number: the sign $-$ can appear only at the beginning and the decimal point . must occur exactly once. Thus the expansions must satisfy certain syntactic rules, which may be expressed by a set of **forbidden words**

$$D = \{a- : \ a \in A\} \cup \{.u. : \ u \in A^*\}.$$

This means that an expansion cannot contain as a subword any letter $a \in A$ followed by the minus sign $-$ and it cannot contain two positional points. Denote by Σ_D the set of infinite words which do not contain as a subword any forbidden word. We say that Σ_D is the **subshift** with the forbidden set D. In the subshift Σ_D there are also words which do not contain any positional point at all. We cannot forbid them, since we cannot detect this property in finite prefixes. We assign the value infinity to such words provided they contain at least one nonzero digit, and the value zero otherwise. We therefore extend the real line $\mathbb{R}$ by a point ∞ at infinity and obtain the **extended real line** $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Then the **value mapping** $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is defined on the whole Σ_D. Some arithmetical operations are extended to $\overline{\mathbb{R}}$. We have $\frac{a}{0} = \infty$ for $a \neq 0$ and $a \pm \infty = \infty$, $\frac{a}{\infty} = 0$ for $a \neq \infty$. On the other hand, $\frac{0}{0}$, $\frac{\infty}{\infty}$, $\infty \pm \infty$ are undefined (indeterminate) expressions.

The value mapping Φ is surjective, i.e., each $x \in \overline{\mathbb{R}}$ has an expansion $u \in \Sigma_D$ with $\Phi(u) = x$, but it is not one-to-one. There are infinitely many expansions of ∞ and some finite numbers have two different expansions, for example $0.999\cdots = 1.000\cdots$. In fact a real number has two infinite expansions iff it has a finite expansion:

$$\begin{aligned} u_n \cdots u_{-1}.u_0 \cdots u_{m-1}u_m &= u_n \cdots u_{-1}.u_0 \cdots u_{m-1}u_m 000 \cdots \\ &= u_n \cdots u_{-1}.u_0 \cdots u_{m-1}(u_m - 1)999 \cdots \end{aligned}$$

This duplicity can be felt as an inconvenience but cannot be detected in finite prefixes and cannot be avoided by forbidding finite words. In fact, such a duplicity or **redundancy** is necessary to perform arithmetical operations on the expansions. If we are able to determine real numbers x and y to an arbitrary precision, we would like to determine to an arbitrary precision also their sum $x + y$ or the results of other algebraic operations. This means that the prefix of a length k of (the expansion of) $x + y$ should depend only on the prefixes of some length n_k of (the expansions of) the operands x and y. In the standard decadic system this is not possible since the system is not redundant enough: the carries to the left propagate through arbitrarily long intervals. Imagine that we try to add numbers $\frac{1}{3} = 0.33333\ldots$ and $\frac{2}{3} = 0.66666\ldots$, but we do not know in advance their exact values. We can only inspect arbitrarily long prefixes of their expansions. Then we are unable to determine the first digit of the sum. The first digit would be zero if $u_i + v_i < 9$ for some i or 1 if $u_i + v_i > 9$ for some i. In our case neither alternative ever happens so we are never able to determine the first digit of the result.

1.2 Redundancy

To perform arithmetic operations on the expansions of real numbers, we need redundant positional systems, in which the number of digits is greater than the base. For example, the decadic system can be extended with a digit which represents 10. Another possibility is the **decadic signed system** with digits $A = \{\overline{5}, \overline{4}, \overline{3}, \overline{2}, \overline{1}, 0, 1, 2, 3, 4, 5\}$, where $\overline{n}$ stands for $-n$. This system has 11 digits - one more than the base 10, and it has an additional advantage that

$x =$	$.00102222201111001210222 2\ldots$
$y =$	$.02211222111120002222112 2\ldots$
$u =$	$.02313444312231003432334 4\ldots$
$v =$	$.11123221121111013121232\ldots$
$z =$	$.11212101201111021202 12\ldots$

Table 1.1: Addition in the extended binary system

the negative numbers can be expressed without the $-$ sign. Many numbers have an infinite number of expansions, e.g.,

$$\frac{4}{9} = 0.5\overline{555}\cdots = 0.5\overline{5}^{\omega} = 0.45\overline{5}^{\omega} = 0.445\overline{5}^{\omega} = 0.4445\overline{5}^{\omega} = \cdots$$

In computer arithmetic, positional system with other bases than 10 are frequently used. The **standard binary system** has base $\beta = 2$ and digits $A = \{0, 1\}$. Because the number of digits is the same as the base, there is no properly working addition algorithm either. The **extended binary system** has digits $\{0, 1, 2\}$ and the **binary signed system** has digits $A = \{\overline{1}, 0, 1\}$ representing $-1, 0, 1$. In both these systems arithmetic operations are algorithmic. The result can be evaluated to an arbitrary precision provided we know with sufficient precision the operands.

Denote by $\lfloor a \rfloor \in \mathbb{Z}$ the integer part of a real number $a \in \mathbb{R}$, so $a - 1 < \lfloor a \rfloor \le a$. Denote by $\mathrm{mod}_2(n) \in \{0, 1\}$ the parity of an integer $n \in \mathbb{Z}$, so $\mathrm{mod}_2(n) = 0$ iff n is even. We have $n = 2\lfloor \frac{n}{2} \rfloor + \mathrm{mod}_2(n)$ for each $n \in \mathbb{Z}$. To add two numbers $x = \sum_{i=n}^{\infty} x_i 2^{-i-1}$, $y = \sum_{i=n}^{\infty} y_i 2^{-i-1}$ in the extended binary system, we first add them componentwise, so we obtain $u_i = x_i + y_i \in \{0, 1, 2, 3, 4\}$ for $i \ge n$ and $u_i = 0$ for $i < n$. Then we perform the carries to the left and determine v by

$$v_i = \left\lfloor \frac{u_{i+1}}{2} \right\rfloor + \mathrm{mod}_2(u_i),$$

so $v_i \in \{0, 1, 2, 3\}$ and $v \in \{0, 1, 2, 3\}^{\omega}$ represents the same number as u:

$$\begin{aligned} \sum_{i=n-1}^{\infty} v_i \cdot 2^{-i-1} &= \sum_{i=n-1}^{\infty} \lfloor \tfrac{u_{i+1}}{2} \rfloor \cdot 2^{-i-1} + \sum_{i=n}^{\infty} \mathrm{mod}_2(u_i) \cdot 2^{-i-1} \\ &= \sum_{i=n}^{\infty} \left(2 \cdot \lfloor \tfrac{u_i}{2} \rfloor + \mathrm{mod}_2(u_i)\right) 2^{-i-1} = \sum_{i=n}^{\infty} u_i \cdot 2^{-i-1}. \end{aligned}$$

We perform the carry operation once more and obtain $z_i = \lfloor \frac{v_{i+1}}{2} \rfloor + \mathrm{mod}_2(v_i) \in \{0, 1, 2\}$. Thus $\sum_{i=n-2}^{\infty} z_i \cdot 2^{-i-1} = \sum_{i=n}^{\infty} (x_i + y_i) \cdot 2^{-i-1}$ and z_i depends only on $x_{[i,i+2]} = x_i x_{i+1} x_{i+2}$ and $y_{[i,i+2]} = y_i y_{i+1} y_{i+2}$. The algorithm has an additional advantage that it may be performed in parallel in all positions i simultaneously. This may be much faster than the serial addition. An example can be seen in Table 1.1. Parallel addition is treated in more detail in Section 7.7.

1.3 Symbolic spaces

The principle that finite prefixes of the expansions approximate the expanded numbers can be expressed by the concept of continuity. We regard the set of infinite expansions as a symbolic **metric space**. An **alphabet** A is a finite set with at least two elements, which are referred to as lettres. Words of A are finite or infinite sequences $u = u_0 u_1 \ldots$ of letters of A. We denote by

$$A^n = \{u = u_0 \cdots u_{n-1} : \ u_i \in A\}$$

the set of words of length n. In particular, $A^0 = \{\lambda\}$ consists only of the empty word λ. Denote by $A^* = \bigcup_{n\geq 0} A^n$ the set of finite words, by $A^+ = \bigcup_{n>0} A^n$ the set of nonempty finite words, and by

$$A^\omega = \{u = u_0u_1\cdots :\ u_i \in A\}$$

the set of infinite words. The **length of a word** $u = u_0 \dots u_{n-1} \in A^n$ is denoted by $|u| = n$ and $|u| = \infty$ for $u \in A^\omega$. We say that $v \in A^*$ is a **subword** of $u \in A^* \cup A^\omega$ and write $v \sqsubseteq u$, if $v = u_{[i,j)} = u_i \cdots u_{j-1}$ for some $0 \leq i \leq j \leq |u|$. The **concatenation** of words $u, v \in A^*$ is written as uv, so $(uv)_i = u_i$ for $i < |u|$ and $(uv)_{|u|+i} = v_i$ for $i < |v|$. The concatenation of $u \in A^+$ with itself n times is written as u^n and the infinite concatenation of u with itself as $u^\omega \in A^\omega$. We say that $u \in A^\omega$ is a **periodic word** if $u = vw^\omega$ for some **preperiod** $v \in A^*$ and **period** $w \in A^+$. Given a set of forbidden words $D \subseteq A^+$, we denote by

$$\begin{aligned}\Sigma_D &= \{u \in A^\omega :\ \forall v \sqsubseteq u, v \notin D\}\\ \mathcal{L}_D &= \{u \in A^* :\ \forall v \sqsubseteq u, v \notin D\}\end{aligned}$$

the **subshift** and **language** of D, and by $\mathcal{L}_D^n = \mathcal{L}_D \cap A^n$. The **distance of words** $u, v \in A^\omega$ is defined by

$$d(u, v) = 2^{-n},\ \text{where } n = \min\{k \geq 0 :\ u_k \neq v_k\}.$$

Then d is a **metric** on A^ω. For example, in the binary alphabet we have $d(0100\ldots, 0110\ldots) = 2^{-2} = \frac{1}{4}$ and $d(0\ldots, 1\ldots) = 2^{-0} = 1$. Thus $u, v \in A^\omega$ are close, if they have a long common prefix:

$$d(u, v) \leq 2^{-n} \Leftrightarrow u_{[0,n)} = v_{[0,n)} \Leftrightarrow d(u, v) < 2^{-n+1}$$

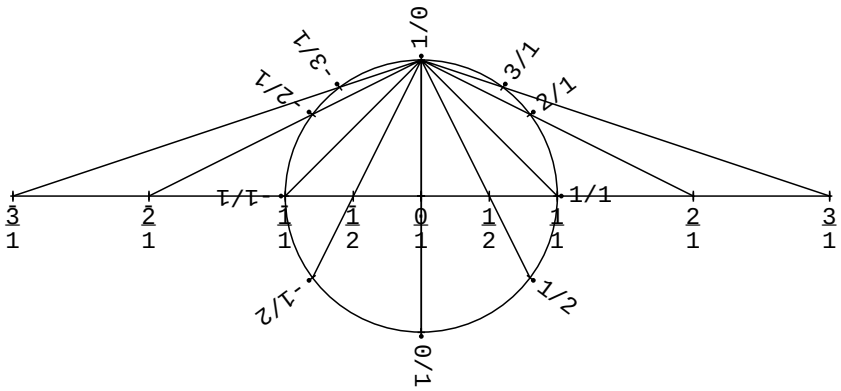


Figure 1.1: The stereographic projection

With the metric d, A^ω and its subspaces Σ_D are turned into metric spaces. A mapping $\Phi : \Sigma_D \to \mathbb{R}$ is **continuous** at $u \in \Sigma_D$, if for every $\varepsilon > 0$ there exists $\delta = 2^{-k}$ such that for every $v \in \Sigma_D$ with $d(u, v) \leq \delta$ we have $|\Phi(u) - \Phi(v)| \leq \varepsilon$. If the range of Φ is the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, then the Euclidean metric $d_e(x, y) = |x - y|$ does not work. Since the extended real line does not distinguish positive and negative infinity, it is topologically equivalent to a circle. Consider the **unit circle**

$$\mathbb{S} = \{z = x + iy \in \mathbb{C} :\ |z| = \sqrt{x^2 + y^2} = 1\}$$

in the **complex plane** $\mathbb{C}$ and project the point $z \in \mathbb{R}$ on the real line to $\mathbb{S}$ by the ray from the **imaginary unit** i (see Figure 1.1). This mapping is called the one-dimensional **stereographic projection**. The line which joins $z \in \mathbb{R}$ with i has parametric equation $x(t) = tz+(1-t)i$. The

equation $1 = |x(t)|^2 = t^2z^2 + (1-t)^2$ gives $t = 0$ and $t = \frac{2}{z^2+1}$, so the stereographic projection $\mathbf{d} : \overline{\mathbb{R}} \to \mathbb{S}$ is given by

$$\mathbf{d}(z) = \frac{2z + (z^2-1)i}{z^2+1},$$

and $\mathbf{d}(\infty) = i$. Zero is projected to $\mathbf{d}(0) = -i$, and $1, -1$ remain fixed: $\mathbf{d}(1) = 1$, $\mathbf{d}(-1) = -1$. The inverse stereographic projection is given by $\mathbf{d}^{-1}(x+iy) = \frac{x}{1-y}$, $\mathbf{d}^{-1}(i) = \infty$.

In the extended real line we have more intervals than in $\mathbb{R}$. Besides standard closed intervals $[a,b] = \{x \in \mathbb{R} : \ a \le x \le b\} \subset \mathbb{R}$, where $a < b$, we consider infinite intervals

$$\begin{aligned} [a,\infty] &= \{x \in \mathbb{R} : \ a \le x\} \cup \{\infty\} \\ [\infty,b] &= \{x \in \mathbb{R} : \ x \le b\} \cup \{\infty\} \\ [a,b] &= \{x \in \mathbb{R} : \ a \le x \text{ or } x \le b\} \cup \{\infty\}, \end{aligned}$$

where $b < a$ are real numbers. We define the **angle length** of these intervals as the length of the counterclockwise arc from $\mathbf{d}(a)$ to $\mathbf{d}(b)$, which is the argument of $\mathbf{d}(b)/\mathbf{d}(a)$. Recall that the **argument** of a nonzero complex number $z = x + iy = r(\cos\varphi + i\sin\varphi)$ is $\arg(x+iy) = \varphi \in [0, 2\pi)$. We have a formula

$$\arg(x+iy) = \begin{cases} 0 & \text{if} \quad x > 0, y = 0, \\ 2\text{arccotg}\frac{y}{\sqrt{x^2+y^2}-x} & \text{otherwise} \end{cases}$$

Since $|\mathbf{d}(b)/\mathbf{d}(a)| = |\mathbf{d}(b)|/|\mathbf{d}(a)| = 1$, the formula simplifies. If $|x+iy| = 1$, then $\arg(x+iy) = 2\text{arccotg}\frac{y}{1-x}$. We have

$$\begin{aligned} \frac{\mathbf{d}(b)}{\mathbf{d}(a)} &= \frac{2b + i(b^2-1)}{b^2+1} \cdot \frac{a^2+1}{2a + i(a^2-1)} \\ &= \frac{a^2+1}{b^2+1} \cdot \frac{(2b + i(b^2-1))(2a - i(b^2-1))}{4a^2 + (a^2-1)^2} \\ &= \frac{4ab + (a^2-1)(b^2-1) + 2i(a(b^2-1) - b(a^2-1))}{(a^2+1)(b^2+1)} \\ &= \frac{(a^2+1)(b^2+1) - 2(b-a)^2 + 2i(b-a)(ab+1)}{(a^2+1)(b^2+1)} \end{aligned}$$

We define the length $|[a,b]|$ of $[a,b] \subseteq \overline{\mathbb{R}}$ by

$$\begin{aligned} |[a,b]| &= \frac{1}{2\pi}\arg\frac{\mathbf{d}(b)}{\mathbf{d}(a)} = \frac{1}{\pi}\text{arccotg}\frac{2(b-a)(ab+1)}{2(b-a)^2} \\ &= \frac{1}{\pi}\text{arccotg}\frac{ab+1}{b-a} \end{aligned}$$

If one of the endpoints is ∞ we get from the limit

$$\begin{aligned} |[a,\infty]| &= \frac{1}{\pi}\text{arccotg}(a), \\ |[\infty,b]| &= \frac{1}{\pi}\text{arccotg}(-b). \end{aligned}$$

Thus for example $|[0,1]| = \frac{1}{4}$, $|[0,\infty]| = \frac{1}{2}$ and $|[0,-1]| = \frac{3}{4}$ (see Figure 1.2).

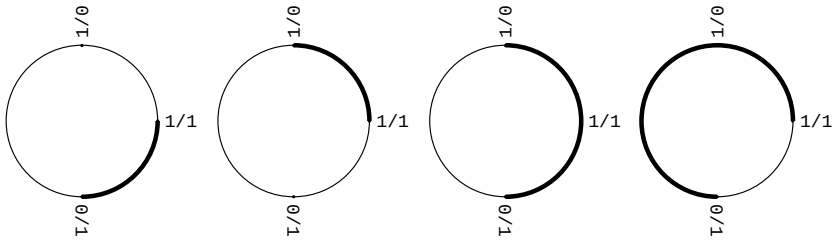


Figure 1.2: Intervals and their length (from left to right): $|[0,1]| = \frac{1}{4}$, $|[1,\infty]| = \frac{1}{4}$, $|[0,\infty]| = \frac{1}{2}$, $|[1,0]| = \frac{3}{4}$.

We define the **angle distance** $d_a(a,b)$ of $a, b \in \overline{\mathbb{R}}$ as the length of the shorter of the two intervals with endpoints a, b:

$$d_a(a,b) \;=\; \min\{|[a,b]|, |[b,a]|\} = \frac{1}{\pi}\mathrm{arccotg}\frac{|ab+1|}{|b-a|}.$$

A mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is **continuous** at $u \in \Sigma_D$, if for every $\varepsilon > 0$ there exists $\delta = 2^{-k}$ such that for every $v \in \Sigma_D$ we have $d_a(\Phi(u), \Phi(v)) \le \varepsilon$ whenever $d(u,v) \le \delta$.

1.4 Positional systems for bounded intervals

We consider a positional system with base $\beta > 1$ and a finite set of digits which form a contiguous interval $A = [r,s] = \{r, r+1, \ldots, s-1, s\} \subset \mathbb{Z}$ of integers. First we consider number systems for bounded intervals. In this case, the positional point is not needed. Thus we have the **value mapping** $\Phi : A^\omega \to \mathbb{R}$ defined by

$$\Phi(u) = \sum_{i=0}^{\infty} u_i \beta^{-i-1} = \frac{u_0}{\beta} + \frac{u_1}{\beta^2} + \frac{u_2}{\beta^3} + \cdots$$

The value mapping is defined also for nonempty finite words by

$$\Phi(u) = \sum_{i=0}^{|u|-1} u_i \beta^{-i-1},\ u \in A^+$$

If $u, v \in A^\omega$ and $u_i \le v_i$ for all i, then $\Phi(u) \le \Phi(v)$, and the inequality is strict if $u_i < v_i$ for some i. Thus the value map is increasing. Define the **cylinder**

$$[u] = \{v \in A^\omega :\ v_{[0,|u|)} = u\}$$

of a finite word $u \in A^*$ as the set of infinite words whose prefix is u. The minimum and maximum of the set $\Phi([u])$ is $\Phi(ur^\omega)$ and $\Phi(us^\omega)$ respectively. Define the closed **cylinder interval** W_u by

$$W_u = [\Phi(ur^\omega), \Phi(us^\omega)] = [\Phi(u) + \tfrac{r}{\beta^n(\beta-1)}, \Phi(u) + \tfrac{s}{\beta^n(\beta-1)}],\ u \in A^n$$

In particular for the empty word we have $W_\lambda = [\Phi(r^\omega), \Phi(s^\omega)] = [\frac{r}{\beta-1}, \frac{s}{\beta-1}]$. We show that the mapping $\Phi : A^\omega \to W_\lambda$ is surjective provided $s - r \ge \beta - 1$:

Lemma 1.1 *If $r - s \ge \beta - 1$ then $W_u = \bigcup_{a \in A} W_{ua}$ for each $u \in A^*$.*

Proof: The condition $W_u = \bigcup_{a\in A} W_{ua}$ is satisfied if the neighbouring intervals W_{ua}, $W_{u,a+1}$ overlap, i.e., if the left endpoint of $W_{u,a+1}$ is smaller or equal than the right endpoint of W_{ua}. Since $\Phi(ua) = \Phi(u) + a\beta^{-n-1}$ for $|u| = n$, this means

$$\Phi(u) + \frac{a+1}{\beta^{n+1}} + \frac{r}{\beta^{n+1}(\beta-1)} \le \Phi(u) + \frac{a}{\beta^{n+1}} + \frac{s}{\beta^{n+1}(\beta-1)},$$

which is equivalent to $s - r \ge \beta - 1$. □

Proposition 1.2 *Let $\beta > 1$ be a real number, $r < s$ integers, and $A = \{a \in \mathbb{Z} :\ r \le a \le s\}$. Then $\Phi : A^\omega \to \mathbb{R}$ defined by $\Phi(u) = \sum_{i=0}^{\infty} u_i\beta^{-i-1}$ is continuous and*

$$\Phi([u]) \subseteq W_u = [\Phi(u) + \tfrac{r}{\beta^n(\beta-1)}, \Phi(u) + \tfrac{s}{\beta^n(\beta-1)}]$$

for any $u \in A^n$. If $s - r \ge \beta - 1$, then $\Phi([u]) = W_u$ and $\Phi : A^\omega \to W_\lambda = [\frac{r}{\beta-1}, \frac{s}{\beta-1}]$ is **surjective***, i.e., any $x \in W_\lambda$ has an expansion $u \in A^\omega$ with $\Phi(u) = x$.*

Proof: If $u \in A^*$ then $\Phi([u]) \subseteq W_u$ and the Euclidean length of W_u is $\Phi(uq^\omega) - \Phi(up^\omega) = \frac{r-s}{\beta^{|u|}(1-\beta)}$, which converges to 0 as $|u| \to \infty$. This shows that Φ is continuous:

$$d(u,v) \le 2^{-n} \ \Rightarrow\ |\Phi(u) - \Phi(v)| \le \frac{s-r}{\beta^n(\beta-1)}.$$

Given $x \in W_\lambda = [\frac{r}{\beta-1}, \frac{s}{\beta-1}]$, we construct its expansion u by induction using Lemma 1.1. Since $W_\lambda = \bigcup_{a\in A} W_a$, there exists u_0 with $x \in W_{u_0}$. If $u_{[0,n)}$ has been constructed and $x \in W_{u_{[0,n)}}$, then there exists u_n such that $x \in W_{u_{[0,n+1)}}$. Since $x \in W_{u_{[0,n)}}$ implies $|x - \Phi(u_{[0,n)})| \le \frac{s-r}{\beta^n(\beta-1)}$, we get $x = \Phi(u)$. □

If $s - r > \beta - 1$ then the system is redundant and a number may have many expansions. If $s - r < \beta - 1$, then $\Phi(A^\omega)$ is a Cantor set included in $[\frac{r}{\beta-1}, \frac{s}{\beta-1}]$, and Φ is one-to-one. If $\beta = 3$ and $A = \{0, 2\}$, then $\Phi(\lambda)$ is the **Cantor middle third set** (see Figure 1.3) obtained from the unit interval $[0, 1]$ by deleting successively the middle thirds of remaining intervals:

$$\Phi(A^\omega) = [0,1] \setminus (\tfrac{1}{3}, \tfrac{2}{3}) \setminus (\tfrac{1}{9}, \tfrac{2}{9}) \setminus (\tfrac{7}{9}, \tfrac{8}{9}) \setminus \cdots$$

The digits in the alphabet $A = \{0, 2\}$ are not contiguous. With contiguous digits $A = \{0, 1\}$ we get $\Phi([\lambda]) \subseteq [0, \frac{1}{2}]$, $\Phi([0]) \subseteq [0, \frac{1}{6}]$, $\Phi([1]) \subseteq [\frac{1}{3}, \frac{1}{2}]$, etc., so

$$\Phi(A^\omega) = [0, \tfrac{1}{2}] \setminus (\tfrac{1}{6}, \tfrac{1}{3}) \setminus (\tfrac{1}{18}, \tfrac{2}{18}) \setminus (\tfrac{7}{18}, \tfrac{8}{18}) \setminus \cdots$$

Figure 1.3: The Cantor middle third set

An expansion of an number $x \in W_\lambda$ can be found by an algorithm which is implicit in Proposition 1.2. There is, however a better algorithm based on an iterative method. For $u \in A^\omega$ and $a \in A$ we have

$$\Phi(au) \ = \ \frac{a}{\beta} + \frac{1}{\beta}\sum_{i=0}^{\infty} u_i\beta^{-i-1} = \frac{a + \Phi(u)}{\beta} = F_a(\Phi(u)),$$

$$\begin{array}{rcl|rcl|rcl}
\frac{2}{7} \in [0,\frac{1}{2}] & = & W_0 & \Rightarrow & u_0 = 0 & \Leftarrow & \frac{2}{7} \in [0,\frac{1}{2}] = W_0 \\
\frac{2}{7} \in [\frac{1}{4},\frac{1}{2}] & = & W_{01} & \Rightarrow & u_1 = 1 & \Leftarrow & F_0^{-1}(\frac{2}{7}) = \frac{4}{7} \in [\frac{1}{2},1] = W_1 \\
\frac{2}{7} \in [\frac{2}{8},\frac{3}{8}] & = & W_{010} & \Rightarrow & u_2 = 0 & \Leftarrow & F_1^{-1}(\frac{4}{7}) = \frac{1}{7} \in [0,\frac{1}{2}] = W_0 \\
\frac{2}{7} \in [\frac{4}{16},\frac{5}{16}] & = & W_{0100} & \Rightarrow & u_3 = 0 & \Leftarrow & F_0^{-1}(\frac{1}{7}) = \frac{2}{7} \in [0,\frac{1}{2}] = W_0
\end{array}$$

Table 1.2: The expansion of $\frac{2}{7}$ in the binary system according to Proposition 1.2 (left) and according to Proposition 1.4 (right).

where $F_a(x) = \frac{x+a}{\beta}$. The value mapping Φ can be derived from the system of real functions $\{F_a : \overline{\mathbb{R}} \to \overline{\mathbb{R}} : \ a \in A\}$. For a finite word $u \in A^n$ we denote by $F_u = F_{u_0} \circ \cdots \circ F_{u_{n-1}}$ the composition of mappings F_{u_i}, and $F_\lambda = \mathrm{Id}$ is the identity mapping. Then $F_{uv} = F_u \circ F_v$ for each $u, v \in A^*$.

Proposition 1.3 *For $A = [r, s]$, $\beta > 1$, $F_a(x) = \frac{x+a}{\beta}$ we have*
1. *$\Phi(uv) = F_u(\Phi(v))$ for $u \in A^*$, $v \in A^* \cup A^\omega$*
2. *$W_{uv} = F_u(W_v)$ for $u, v \in A^*$*
3. *$F_u(x) = \Phi(u) + \frac{x}{\beta^{|u|}}$ for $u \in A^*$, $x \in \mathbb{R}$,*
4. *$\Phi(u) = \lim_{n\to\infty} F_{u_{[0,n)}}(z)$ for $u \in A^\omega$, $z \in \mathbb{R}$*
5. *$\{\Phi(u)\} = \bigcap_{n>0} W_{u_{[0,n)}}$ for $u \in A^\omega$*

Proof: 1. The statement holds trivially for $u = \lambda$. If it holds for u, then $\Phi(auv) = F_a(\Phi(uv)) = F_aF_u(\Phi(v)) = F_{au}(\Phi(v))$.
2. $W_{uv} = [\Phi(uvr^\omega), \Phi(uvs^\omega)] = [F_u\Phi(vr^\omega), F_u\Phi(vs^\omega)] = F_u(W_v)$.
3. The statement holds trivially for $|u| \le 1$. For $|u| > 1$ we use $F_a(x + y) = F_a(x) + \frac{y}{\beta}$ to get $F_{au}(x) = F_a(F_u(x)) = F_a(\Phi(u)) + \frac{x}{\beta^{|u|+1}} = \Phi(au) + \frac{x}{\beta^{|au|}}$.
4. follows from 3.
5. We have $\Phi(u) \in [\Phi(u_{[0,n)})] \subseteq W_{u_{[0,n)}}$. Since the length of these intervals converges to zero, the intersection contains a unique point $\Phi(u)$. □

The mappings F_a are **contracting**, i.e., they contract the Euclidean distance by the factor β: $|F_a(x) - F_a(y)| = |x - y|/\beta$. The inverse mappings $F_a^{-1}(x) = \beta x - a$ are **expanding**, they expand the distances by the factor β.

Proposition 1.4 *Assume that $r - s \ge \beta - 1 > 0$. A word $u \in A^\omega$ is an expansion of $x = x_0 \in W_\lambda$ iff there exists a sequence of numbers $x_i \in W_{u_i}$ such that $x_{i+1} = F_{u_i}^{-1}(x_i)$.*

Proof: If $x = \Phi(u)$, then $x \in W_{u_{[0,n]}}$ for each n. If $x = x_0 \in W_{u_{[0,n]}}$, then $x_1 = F_{u_0}^{-1}(x_0) \in F_{u_0}^{-1}(W_{u_{[0,n]}}) = W_{u_{[1,n]}}$, so by induction $x_i \in W_{u_{[i,n]}} \subseteq W_{u_i}$ for every $i \le n$. Conversely, if $x_n \in W_{u_n}$, then $x_{n-1} = F_{u_n}(x_n) \in F_{u_n}(W_{u_n}) = W_{u_{[n-1,n]}}$ and by induction $x_i \in W_{u_{[i,n]}}$ for every $i \le n$, in particular $x = x_0 \in W_{u_{[0,n]}}$. It follows $x = \Phi(u)$. □

An example of an expansion process according to both methods of Propositions 1.2 and 1.4 is in Table 1.2. The iterative algorithm of Proposition 1.4 is better, since the inequalities involve rational numbers with smaller numerators and denominators. Moreover, we see immediately that the expansion process is periodic so the expansion of $\frac{2}{7}$ is the periodic word $(010)^\omega$. The iterative expansion process is illustrated in Figure 1.4 which shows the graphs of mappings F_a^{-1}. Given x_0, we draw the vertical line from $(x_0, 0)$ to $(x_0, x_1) = (x_0, F_{u_0}^{-1}(x_0))$, the horizontal

line to (x_1, x_1) on the diagonal $y = x$, the vertical line to (x_1, x_2), etc. In the standard binary systems, the intervals W_a intersect only in their endpoints, so most of the times, the expansions are unique: we have two possibilities only if $x_n = \frac{1}{2}$. In the binary signed system, on the other hand, the neighbouring intervals W_a overlap, so the expansion algorithm is nondeterministic. When $x_n \in [-\frac{1}{2}, \frac{1}{2}]$, then we have two or three choices for x_{n+1}. It follows that each number (except 0 and 1) has an infinite number of expansions. There exist also deterministic expansion algorithms with smaller expansion intervals. For example, the **greedy expansion** algorithm takes always the largest possible letter. This is accomplished with the iterative algorithm which uses semi-closed expansion intervals $W_{\overline{1}} = [-1, -\frac{1}{2})$, $W_0 = [-\frac{1}{2}, 0)$, $W_1 = [0, 1]$. Since these intervals W_a are pairwise disjoint and their union is the whole W_λ, each $x \in W_\lambda$ has a unique expansion $\mathcal{E}(x) \in A^\omega$ with $\Phi(\mathcal{E}(x)) = x$. However, the mapping $\mathcal{E} : W_\lambda \to A^\omega$ is not continuous.

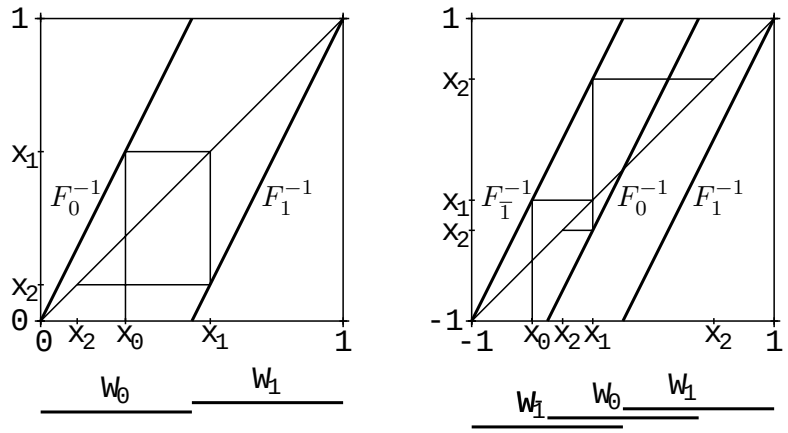


Figure 1.4: Expansions of real numbers in the standard binary system (left) and in the binary signed system (right)

1.5 Positional systems for the extended real line

To obtain number system for the whole extended real line $\overline{\mathbb{R}}$, we extend the alphabet with a digit $\overline{0}$ (which stands for ∞) and associate to $\overline{0}$ the real function $F_{\overline{0}}(x) = \beta x$. For a word $u \in [r, s]^n$ and $m \geq 0$ we define

$$\Phi(\overline{0}^m u) = F_{\overline{0}^m}(\Phi(u)) = \sum_{i=0}^{n-1} u_i \beta^{m-i-1}$$

and $W_{\overline{0}^m u} = F_{\overline{0}^m}(W_u)$. For an infinite word $u \in \{r, \ldots, s\}^\omega$ we get

$$\Phi(\overline{0}^m u) = \lim_{n\to\infty} \Phi(\overline{0}^m u_{[0,n)}) = \sum_{i\geq 0} u_i \beta^{m-i-1}.$$

Thus the value of $\overline{0}^m u$ is the same as the value of the word $u_0 \cdots u_{m-1}.u_m u_{m+1} \ldots$ with the positional point before u_m. With the extended alphabet $A = \{r, \ldots, s, \overline{0}\}$, every real number has an expansion provided $s - r \geq \beta - 1$ and $r < 0 < s$. The mapping Φ, however, cannot be defined on all words of A^ω but only on the words of the form $\overline{0}^m u$, where $u \in [r, s]^\omega$. These words form the subshift Σ_D with the set of forbidden words $D = \{a\overline{0} : \ a \in [r, s]\}$.

Proposition 1.5 *Let $s - r \geq \beta - 1 > 0$, $r < 0 < s$, $A = \{r, \ldots, s, \overline{0}\}$, $D = \{a\overline{0} : \ r \leq a \leq s\}$, $F_{\overline{0}}(x) = \beta x$ and $F_a(x) = (x + a)/\beta$ for $r \leq a \leq s$. Define $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ by $\Phi(\overline{0}^m u) = \sum_{i \geq 0} u_i \beta^{m-i-1}$, $\Phi(\overline{0}^\omega) = \infty$. Then $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective.*

Proof: From Proposition 1.2 we get

$$\Phi(\{\overline{0}^m u : \ u \in \{r, \ldots, s\}^\omega\}) = F_{\overline{0}}^m([\Phi(r^\omega), \Phi(s^\omega)]) = [\tfrac{r\beta^m}{\beta-1}, \tfrac{s\beta^m}{\beta-1}]$$

and these intervals cover whole $\mathbb{R}$. Since $\Phi(\overline{0}^\omega) = \infty$, the mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective. □

The mapping Φ from Proposition 1.5 is not a reasonable number system, since it is not continuous. The words $\overline{0}^n 0^\omega$ converge in A^ω to $\overline{0}^\omega$ with value $\Phi(\overline{0}^\omega) = \infty$, but $\Phi(\overline{0}^n 0^\omega) = 0$. To make Φ continuous, we must forbid words which contain $\overline{0}0$, and possibly some other words as well. Suitable subshifts depend on β.

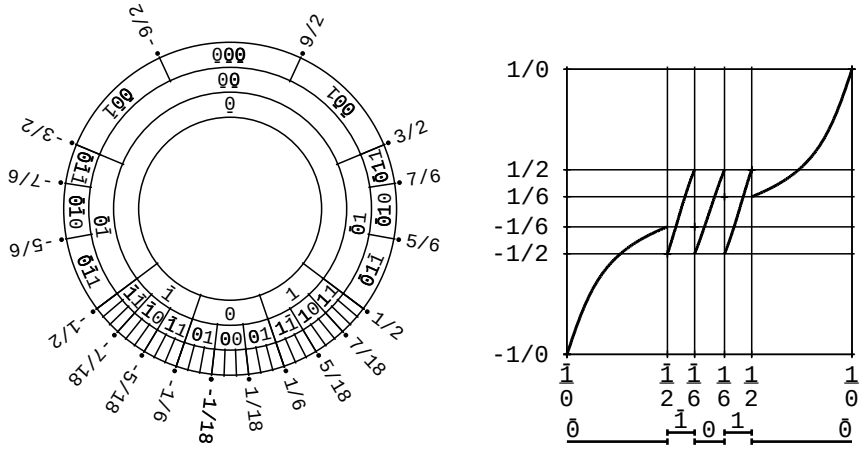


Figure 1.5: The ternary signed system: the cylinder intervals (left), the graphs of the inverse mappings F_a^{-1} (right) and the intervals W_a (bottom right).

Proposition 1.6 *For the* **ternary signed system** *(Figure 1.5) with base $\beta = 3$, alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$ and forbidden words $D = \{\overline{1}\overline{0}, 0\overline{0}, 1\overline{0}, \overline{0}\overline{0}\}$, the map $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective and continuous.*

Proof: By Proposition 1.2, $\Phi([\overline{1}]) = [-\frac{1}{2}, -\frac{1}{6}]$, $\Phi([0]) = [-\frac{1}{6}, \frac{1}{6}]$, $\Phi([1]) = [\frac{1}{6}, \frac{1}{2}]$, so

$$\Phi([\overline{0}^m\overline{1}]) = [-\tfrac{3^m}{2}, -\tfrac{3^{m-1}}{2}], \Phi([\overline{0}^m 1]) = [\tfrac{3^{m-1}}{2}, \tfrac{3^m}{2}].$$

Since $\Phi(\overline{0}^\omega) = \infty$, we get

$$\begin{aligned}
\Phi([\overline{0}^m]) &= \bigcup_{n \geq m} \Phi([\overline{0}^n 1]) \cup \{\Phi(\overline{0}^\omega)\} \cup \bigcup_{n \geq m} \Phi([\overline{0}^n\overline{1}]) \\
&= [\tfrac{3^m}{6}, \tfrac{3^m}{2}] \cup [\tfrac{3^{m+1}}{6}, \tfrac{3^{m+1}}{2}] \cup \cdots \cup [\tfrac{-3^{m+1}}{2}, -\tfrac{3^{m+1}}{6}] \cup [\tfrac{-3^m}{2}, -\tfrac{3^m}{6}] \cup \{\infty\} \\
&= [\tfrac{3^{m-1}}{2}, -\tfrac{3^{m-1}}{2}]
\end{aligned}$$

The Euclidean length of $\Phi([\overline{0}^m])$ is infinite but its angle length is $|\Phi([\overline{0}^m])| = \frac{1}{\pi}\mathrm{arccotg}\frac{3^{2m-2}-1}{4 \cdot 3^{m-1}}$, which converges to zero as $m \to \infty$. This shows that Φ is continuous at $\overline{0}^\omega$. At any other point, Φ is continuous by Proposition 1.2. To show that $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective, set

$$W_{\overline{1}} = [-\tfrac{1}{2}, -\tfrac{1}{6}], \ W_0 = [-\tfrac{1}{6}, \tfrac{1}{6}], \ W_1 = [\tfrac{1}{6}, \tfrac{1}{2}], \ W_{\overline{0}} = [\tfrac{1}{2}, -\tfrac{1}{2}]$$

Then

$$\begin{aligned} F_a^{-1}(W_a) &= W_{\overline{1}} \cup W_0 \cup W_1,\ a \in \{\overline{1}, 0, 1\} \\ F_{\overline{0}}^{-1}(W_{\overline{0}}) &= W_1 \cup W_{\overline{0}} \cup W_{\overline{1}} \end{aligned}$$

Thus for each $a \in A$ we have $F_a^{-1}(W_a) = \cup\{W_b :\ ab \in \mathcal{L}_D\}$, and $\overline{\mathbb{R}} = \cup\{W_a :\ a \in A\}$. The expansion algorithm of Proposition 1.4 works with a small modification. At each step we check, whether the constructed word belongs to Σ_D. Given $x = x_0 \in \overline{\mathbb{R}}$ we construct a sequence $x_n \in \overline{\mathbb{R}}$ and $u_n \in A$ as follows. Find u_0 with $x_0 \in W_{u_0}$ and set $x_1 = F_{u_0}^{-1}(x_0)$. If u_{i-1}, x_i have been already constructed, find u_i with $u_{i-1}u_i \in \mathcal{L}_D$, $x_i \in W_{u_i}$ and set $x_{i+1} = F_{u_i}^{-1}(x_i)$. Then $u \in \Sigma_D$ and $x = F_{u_{[0,n)}}(x_n) \in F_{u_{[0,n)}}(W_{u_n}) = W_{u_{[0,n]}}$ and the diameter of these sets converges to zero, so $\Phi(u) = x$. □

In Figure 1.5 we see the cylinder intervals $\Phi[u]$ (left) and the graphs of mappings F_a^{-1} of the ternary signed system. We now generalize Proposition 1.6.

Proposition 1.7 *Let A be a finite alphabet and $D \subset A^2$ a set of forbidden words. For each $a \in A$, let $F_a : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ be a one-to-one continuous mappings and $W_a \subset \overline{\mathbb{R}}$ a closed interval. Assume that $\bigcup_{a\in A} W_a = \overline{\mathbb{R}}$, $F_a^{-1}(W_a) = \bigcup\{W_b :\ ab \in \mathcal{L}_D\}$ and that the angle length of intervals $F_u(W_a)$ converges to zero as the length of words $ua \in \mathcal{L}_D$ converges to infinity. Then there exists a continuous surjective function $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ such that $\{\Phi(u)\} = \bigcap_{n>0} F_{u_{[0,n)}}(W_{u_n})$.*

Proof: If $ab \in \mathcal{L}_D$ then $W_b \subseteq F_a^{-1}(W_a)$, so $F_a(W_b) \subseteq W_a$. For any $u \in \Sigma_D$ we get by induction

$$\cdots \subseteq F_{u_{[0,3)}}(W_{u_3}) \subseteq F_{u_{[0,2)}}(W_{u_2}) \subseteq F_{u_0}(W_{u_1}) \subseteq W_{u_0}$$

Since the length of these intervals converges to zero, they have a nonempty intersection which contains a unique point $\Phi(u)$, and the mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous. We show that it is surjective. For $x = x_0 \in \overline{\mathbb{R}}$ there exists u_0 with $x_0 \in W_{u_0}$. If u_n with $x_n \in W_{u_n}$ has been constructed, there exists u_{n+1} such that $u_nu_{n+1} \in \mathcal{L}_D$ and $x_{n+1} = F_{u_n}^{-1}(x_n) \in W_{u_{n+1}}$. Thus $x_0 \in F_{u_{[0,n)}}(W_{u_n})$ for each n and therefore $\Phi(u) = x$. □

In the binary signed system with $\beta = 2$, $r = \overline{1} = -1$, $s = 1$, the subshift Σ_D of Proposition 1.6 does not work, since $\Phi(\overline{0}^n 1\overline{1}^\omega) = 0$ while $\overline{0}^n 1\overline{1}^\omega$ converge to $\overline{0}^\omega$ with value $\Phi(\overline{0}^\omega) = \infty$. This means that Φ is not continuous at $\overline{0}^\omega$. To make Φ continuous, we forbid words $01\overline{1}^\omega$ and $0\overline{1}1^\omega$. One possibility is to forbid $01\overline{1}$ and $0\overline{1}1$. To get a subshift with forbidden words of length 2, we forbid $\overline{1}1$ and $1\overline{1}$.

Proposition 1.8 *In the binary signed system with alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$ and forbidden words $D = \{\overline{1}\overline{0}, 0\overline{0}, 1\overline{0}, \overline{0}0, 1\overline{1}, \overline{1}1\}$, the map $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective.*

Proof: The smallest number in $\Phi([\overline{1}])$ is $\Phi(\overline{1}^\omega) = \frac{-1}{2} + \frac{-1}{4} + \cdots = -1$, and the largest is $\Phi(\overline{1}01^\omega) = \frac{-1}{2} + \frac{1}{8} + \frac{1}{16} + \cdots = \frac{-1}{4}$. We set $W_{\overline{1}} = [-1, -\frac{1}{4}]$ and similarly define other intervals W_a with $\Phi([a]) \subseteq W_a$:

$$\begin{aligned} W_{\overline{1}} &= [\Phi(\overline{1}^\omega), \Phi(\overline{1}01^\omega)] = [-1, \tfrac{-1}{4}] \\ W_0 &= [\Phi(0\overline{1}^\omega), \Phi(01^\omega)] = [\tfrac{-1}{2}, \tfrac{1}{2}] \\ W_1 &= [\Phi(10\overline{1}^\omega), \Phi(1^\omega)] = [\tfrac{1}{4}, 1] \\ W_{\overline{0}} &= [\Phi(\overline{0}10\overline{1}^\omega), \Phi(\overline{0}\overline{1}01^\omega)] = [\tfrac{1}{2}, \tfrac{1}{-2}] \end{aligned}$$

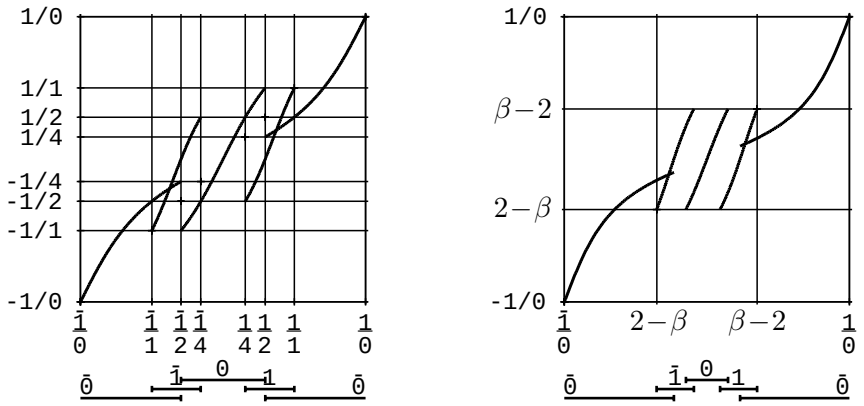


Figure 1.6: The binary signed system (left) and the system with an algeraic base $\beta = \frac{\sqrt{5}+3}{2} \doteq 2.618$ (right).

We have $\Phi[\overline{0}^n] \subseteq W_{\overline{0}^n} = [2^{n-2}, -2^{n-2}]$, and the angle length of this interval converges to zero as $n \to \infty$. To show that $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective, consider the inverse images of intervals W_a:

$$\begin{aligned} F_{\overline{1}}^{-1}(W_{\overline{1}}) &= [-1, \tfrac{1}{2}] = W_{\overline{1}} \cup W_0 \\ F_0^{-1}(W_0) &= [-1, 1] = W_{\overline{1}} \cup W_0 \cup W_1 \\ F_1^{-1}(W_1) &= [-\tfrac{1}{2}, 1] = W_0 \cup W_1 \\ F_{\overline{0}}^{-1}(W_{\overline{0}}) &= [\tfrac{1}{4}, -\tfrac{1}{4}] = W_1 \cup W_{\overline{0}} \cup W_{\overline{1}} \end{aligned}$$

Thus $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ continuous and surjective by Proposition 1.7. □

The base of a positional system need not be an integer, it may be any real number $\beta > 1$. Taking $\beta = \frac{\sqrt{5}+3}{2} \doteq 2.618\cdots$, we get a redundant system with alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$ and forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0\}$, so we get the same subshift Σ_D as in the case of the ternary signed system. Using the equation $\beta^2 = 3\beta - 1$ we can evaluate intervals W_a according to Proposition 1.2:

$$\begin{aligned} W_{\overline{1}} &= [2-\beta, 3\beta-8] \doteq [-0.618, -0.146], \\ W_0 &= [5-2\beta, 2\beta-5] \doteq [-0.236, 0.236], \\ W_1 &= [8-3\beta, \beta-2] \doteq [0.146, 0.618], \\ W_{\overline{0}} &= [3-\beta, \beta-3] \doteq [0.382, -0.382], \\ F_a^{-1}(W_a) &= [-\beta+2, \beta-2] = W_{\overline{1}} \cup W_0 \cup W_1,\ a \in \{\overline{1}, 0, 1\} \\ F_{\overline{0}}^{-1}(W_{\overline{0}}) &= [-3\beta+8, 3\beta-8] = W_1 \cup W_{\overline{0}} \cup W_{\overline{1}}. \end{aligned}$$

Thus $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective by Proposition 1.7.

Positional number systems can have also negative base $\beta < -1$. Let $r < s$ and $A = [r, s] \subset \mathbb{Z}$ be an alphabet, and consider a number system without positional point with value mapping $\Phi(u) = \sum_{i \geq 0} u_i \beta^{-i-1}$. If $u, v \in A^\omega$ are such that $u_{2i} \geq v_{2i}$ and $u_{2i+1} \leq v_{2i+1}$ for all i, then $\Phi(u) \leq \Phi(v)$. It follows that the minimum of $\Phi(A^\omega)$ is

$$\Phi((sr)^\omega) = \left(\frac{s}{\beta} + \frac{r}{\beta^2}\right) \cdot \left(1 + \frac{1}{\beta^2} + \frac{1}{\beta^4} + \cdots\right) = \frac{s\beta + r}{\beta^2 - 1}.$$

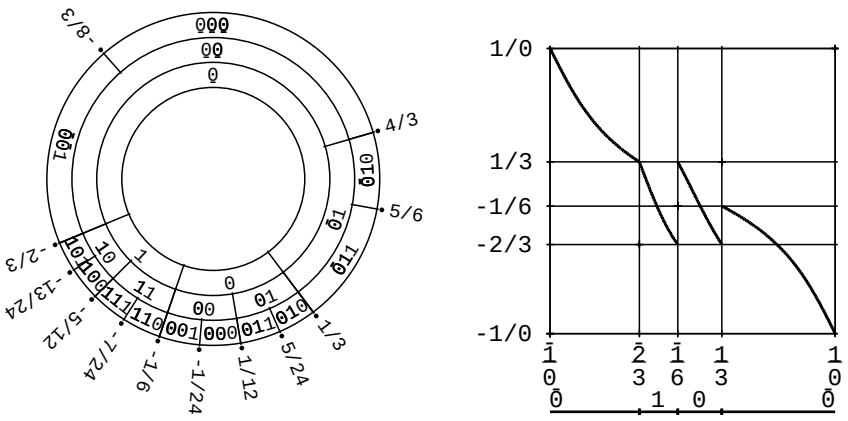


Figure 1.7: The negative binary system with $\beta = -2$, $A = \{0, 1, \overline{0}\}$

Similarly we compute the maximum of $\Phi(A^\omega)$, so $\Phi(A^\omega) \subseteq W_\lambda = [\frac{s\beta+r}{\beta^2-1}, \frac{r\beta+s}{\beta^2-1}]$. For $a \in A$ we get $W_a = F_a W_\lambda = [\frac{r\beta+s}{\beta(\beta^2-1)} + \frac{a}{\beta}, \frac{s\beta+r}{\beta(\beta^2-1)} + \frac{a}{\beta}]$. Then W_λ is covered by W_a if the left endpoint of W_a is smaller than the right endpoint of W_{a+1}, i.e., if

$$\frac{r\beta+s}{\beta(\beta^2-1)} + \frac{a}{\beta} \leq \frac{s\beta+r}{\beta(\beta^2-1)} + \frac{a+1}{\beta},$$

which holds provided $s - r \geq -\beta - 1$.

Proposition 1.9 *If $s - r \geq -\beta - 1 > 0$ and $A = \{r, \ldots, s\}$, then $\Phi : A^\omega \to W_\lambda$ is continuous and surjective.*

To obtain a number system for $\overline{\mathbb{R}}$, we add digit $\overline{0}$ with mapping $F_{\overline{0}}(x) = \beta x$. Negative base allows to express negative real numbers with nonnegative digits.

Proposition 1.10 *For the* **negative binary system** *with base $\beta = -2$, alphabet $A = \{0, 1, \overline{0}\}$ and forbidden set $D = \{0\overline{0}, 1\overline{0}, \overline{0}0\}$, the value mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective.*

Proof: We get $\Phi([0]) = W_0 = [-\frac{1}{6}, \frac{1}{3}]$, $\Phi([1]) = W_1 = [-\frac{2}{3}, -\frac{1}{6}]$. Using $\Phi([\overline{0}^n 1]) = F_{\overline{0}^n}(W_1) = W_{\overline{0}^n 1}$ we get $W_{\overline{0}1} = [\frac{1}{3}, \frac{4}{3}]$, $W_{\overline{00}1} = [-\frac{8}{3}, -\frac{2}{3}]$, $W_{\overline{0}^{2n}1} = [-\frac{2^{2n+1}}{3}, -\frac{2^{2n-1}}{3}]$, $W_{\overline{0}^{2n-1}1} = [\frac{2^{2n-2}}{3}, \frac{2n}{3}]$. It follows $W_{\overline{0}} = \bigcup_{n>0} W_{\overline{0}^n 1} = [\frac{1}{3}, -\frac{2}{3}]$, $W_{\overline{00}} = [\frac{4}{3}, -\frac{2}{3}]$, so $W_{\overline{0}^{2n}} = \left[\frac{2^{2n}}{3}, -\frac{2^{2n-1}}{3}\right]$, $W_{\overline{0}^{2n+1}} = \left[\frac{2^{2n}}{3}, -\frac{2^{2n+1}}{3}\right]$. The angle length of $W_{\overline{0}^n}$ converges to zero as $n \to \infty$, so $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous. To prove surjectivity, we consider inverse images:

$$\begin{aligned} F_0^{-1}(W_0) = F_1^{-1}(W_1) &= [-\tfrac{2}{3}, \tfrac{1}{3}] = W_0 \cup W_1 \\ F_{\overline{0}}^{-1}(W_{\overline{0}}) &= [\tfrac{1}{3}, -\tfrac{1}{6}] = W_{\overline{0}} \cup W_1 \end{aligned}$$

so $F_a^{-1}(W_a) = \cup\{W_b : \ ab \in \mathcal{L}_D\}$, and $\overline{\mathbb{R}} = \cup\{W_a : \ a \in A\}$. Thus $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective and $\Phi([u]) = W_u$ for each $u \in \mathcal{L}_D$. □

1.6 Continued fractions

A quite different number system is based on continued fractions. A **finite simple continued fraction** is an expression

$$u_0 + \frac{1}{u_1 + \frac{1}{u_2 + \cdots + \frac{1}{u_n}}} = u_0 + \cfrac{1}{u_1 + \cfrac{1}{u_2 + \ddots + \cfrac{1}{u_n}}}$$

where $u_n \in \mathbb{Z}$ and $u_n > 0$ for $n > 0$. Infinite simple continued fractions are limits of finite simple continued fractions. We can conceive simple continued fractions as a number system with the infinite alphabet $A = \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ of all integers. Denote by

$$\begin{aligned} \mathcal{Z} &= \{u \in \mathbb{Z}^\omega : \forall n > 0, u_n > 0\} \\ \mathcal{L}(\mathcal{Z}) &= \{u \in \mathbb{Z}^* : \forall n > 0, u_n > 0\} \\ \mathcal{L}^n(\mathcal{Z}) &= \mathcal{L}(\mathcal{Z}) \cap \mathbb{Z}^n \end{aligned}$$

the sets of infinite and finite words of $\mathbb{Z}$ with all but the first element positive. For $a \in \mathbb{Z}$ take the mapping $F_a(x) = a + \frac{1}{x}$. For $u \in \mathcal{L}^n(\mathbb{Z})$ we get

$$F_u(x) = F_{u_0} \circ \cdots \circ F_{u_{n-1}}(x) = u_0 + \frac{1}{u_1 + \frac{1}{u_2 + \cdots + \frac{1}{u_{n-1} + \frac{1}{x}}}}$$

so $u_0 + \frac{1}{u_1 + \frac{1}{u_2 + \cdots + \frac{1}{u_{n-1}}}} = F_u(\infty)$. Define the **convergents** $p_n = p_n(u)$, $q_n = q_n(u)$ of an infinite sequence $u \in \mathcal{Z}$ by $u_0 + \frac{1}{u_1 + \frac{1}{u_2 + \cdots + \frac{1}{u_{n-1}}}} = \frac{p_n(u)}{q_n(u)}$. Then

$$\begin{aligned} p_0 = u_0, \quad p_1 = 1 + u_1 u_0, \quad \ldots \quad p_n = p_{n-2} + u_n p_{n-1} \\ q_0 = 1, \quad q_1 = u_1, \quad \ldots \quad q_n = q_{n-2} + u_n q_{n-1} \end{aligned}$$

We extend this definition with $p_{-1} = 1$, $q_{-1} = 0$ to get the recurrent formula for all $n > 0$.

Proposition 1.11 *Let $u \in \mathcal{Z}$ be an infinite word and $p_n = p_n(u)$, $q_n = q_n(u)$ its convergents. Then for $n > 0$ we have*
1. $p_{n-1}q_{n-2} - p_{n-2}q_{n-1} = (-1)^n$.
2. $p_n q_{n-2} - p_{n-2}q_n = (-1)^n u_n$.
3. $F_{u_{[0,n)}}(x) = (p_{n-1}x + p_{n-2})/(q_{n-1}x + q_{n-2})$.
In particular $F_{u_{[0,n)}}(\infty) = \frac{p_{n-1}}{q_{n-1}}$, $F_{u_{[0,n)}}(0) = \frac{p_{n-2}}{q_{n-2}}$.

Proof: For $n = 1$ we have $p_0 q_{-1} - p_{-1}q_0 = -1$, $p_1 q_{-1} - p_{-1}q_1 = -u_1$, $F_{u_0}(x) = \frac{u_0 x + 1}{x} = \frac{p_0 x + p_{-1}}{q_0 x + q_{-1}}$. Assume that the statement holds for n. Then

$$\begin{aligned} p_n q_{n-1} - p_{n-1}q_n &= (p_{n-2} + u_n p_{n-1})q_{n-1} - p_{n-1}(q_{n-2} + u_n q_{n-1}) \\ &= -(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = (-1)^{n+1} \\ p_{n+1}q_{n-1} - p_{n-1}q_{n+1} &= (p_{n-1} + u_{n+1}p_n)q_{n-1} - p_{n-1}(q_{n-1} + u_{n+1}q_n) \\ &= u_{n+1}(p_n q_{n-1} - p_{n-1}q_n) = (-1)^{n+1}u_{n+1} \\ F_{u_{[0,n+1)}}(x) &= F_{u_{[0,n)}}(u_n + \tfrac{1}{x}) = \tfrac{p_{n-1}(u_n + \frac{1}{x}) + p_{n-2}}{q_{n-1}(u_n + \frac{1}{x}) + q_{n-2}} = \tfrac{(u_n p_{n-1} + p_{n-2})x + p_{n-1}}{(u_n q_{n-1} + q_{n-2})x + q_{n-1}} \\ &= \frac{p_n x + p_{n-1}}{q_n x + q_{n-1}} \quad \square \end{aligned}$$

Proposition 1.12 *If $u \in \mathcal{Z}$, then for every nonnegative real number $z \geq 0$ there exist the limits*

$$\Phi(u) = \lim_{n\to\infty} F_{u_{[0,n)}}(z) = \lim_{n\to\infty} \frac{p_n(u)}{q_n(u)}.$$

If $a \in \mathbb{Z}$ and $u_0 > 0$ then $F_a(\Phi(u)) = \Phi(au)$.

Proof: We have $\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n}{q_{n-1}q_{n-2}}$, $\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^{n+1}u_n}{q_n q_{n-2}}$, so

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \ldots < \frac{p_5}{q_5} < \frac{p_3}{q_3} < \frac{p_1}{q_1}.$$

Since $q_n \to \infty$ as $n \to \infty$, p_n/q_n is a converging sequence. For each $z \geq 0$ we have

$$\left| F_{u_{[0,n+1)}}(z) - \frac{p_n}{q_n} \right| = \frac{1}{q_n(q_n z + q_{n-1})}$$

which converges to zero as $n \to \infty$. □

To expand a real number into a simple continued fraction, consider intervals $W_a = [a, a+1]$ for $a \in \mathbb{Z}$. Then

$$F_a^{-1}(W_a) = [1, \infty] = W_1 \cup W_2 \cup \cdots \cup \{\infty\}$$

Given $x \in \mathbb{R}$ we find its expansion as follows. Set $x_0 = x$ and construct sequences u_n, x_n by induction: $u_n = \lfloor x_n \rfloor$, $x_{n+1} = F_{u_n}^{-1}(x_n) = 1/(x_n - u_n)$. If $x_n > 0$ for all $n > 0$, then we get an infinite $u \in \mathcal{Z}$. If $x_n = 0$ for some n, then we get a finite $u \in \mathcal{L}(\mathcal{Z})$. This happens iff x is a rational number.

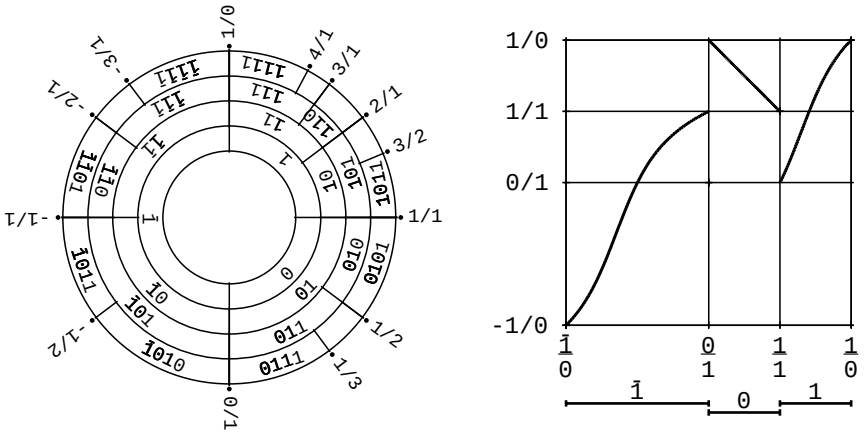


Figure 1.8: The number system of simple continued fractions with $A = \{\overline{1}, 0, 1\}$, $F_{\overline{1}}(x) = x - 1$, $F_0(x) = 1/x$, $F_1(x) = x + 1$, $D = \{\overline{1}1, 0\overline{1}, 00, 1\overline{1}\}$.

To get a number system with a finite alphabet, we decompose the expansion process into elementary steps. To subtract the integer part, we add or subtract repeatedly one till the result is in the unit interval $(0, 1)$. Thus we consider the alphabet $A = \{\overline{1}, 0, 1\}$, mappings F_a and intervals W_a given by

$$\begin{array}{rclrclrcl} F_{\overline{1}}(x) & = & x - 1, & F_0(x) & = & 1/x, & F_1(x) & = & x + 1, \\ W_{\overline{1}} & = & [\infty, 0], & W_0 & = & [0, 1], & W_1 & = & [1, \infty]. \end{array}$$

Then $F_{\overline{1}}^{-1}(W_{\overline{1}}) = W_{\overline{1}} \cup W_0$, $F_0^{-1}(W_0) = W_1$, $F_1^{-1}(W_1) = W_0 \cup W_1$. If we take forbidden words $D = \{\overline{1}1, 0\overline{1}, 00, 1\overline{1}\}$ then $F_a^{-1}(W_a) = \cup\{W_b : \ ab \in \mathcal{L}_D\}$, and $\overline{\mathbb{R}} = \cup\{W_a : \ a \in A\}$, so we can apply Proposition 1.7.

Definition 1.13 *The number system of* **simple continued fractions** *has the alphabet* $A = \{\overline{1}, 0, 1\}$, *transformations* $F_{\overline{1}}(x) = x - 1$, $F_0(x) = 1/x$, $F_1(x) = x + 1$, *forbidden words* $D = \{\overline{1}1, 0\overline{1}, 00, 1\overline{1}\}$ *and the value mapping* $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ *given by*

$$\begin{aligned} \Phi(1^{a_0}01^{a_1}01^{a_2}0\cdots) &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}} \\ \Phi(1^{a_0}01^{a_1}01^{a_2}0\cdots 01^{a_{n-1}}01^{\omega}) &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_{n-1}}}} \\ \Phi(\overline{1}^{\omega}) = \Phi(1^{\omega}) &= \infty \end{aligned}$$

Then $\Phi(uv) = F_u(\Phi(v))$ for each $uv \in \Sigma_D$. The letter $\overline{1}$ can appear only at the begining of a word $u \in \Sigma_D$ and any such word can be written as $u = 1^{a_0}01^{a_1}01^{a_2}0\cdots$, where $a_0 \in \mathbb{Z}$ and $a_n > 0$ for $n > 0$. If $a_0 < 0$ then 1^{a_0} stands for $\overline{1}^{-a_0}$. The sequence of a_i may be finite with last element $a_n = \infty$. Thus the sequences $u \in \Sigma_D$ are in one-to-one correspondence with elements of $\mathcal{Z} \cup \mathcal{L}(\mathcal{Z})$. Since $F_0(x) = 1/x$ is a decreasing function, a continued fraction is increasing in its even entries and decreasing in its odd entries. If $a_{2i} \le b_{2i}$ and $a_{2i+1} \ge b_{2i+1}$ for all i, then $\Phi(1^{a_0}01^{a_1}01^{a_2}0\cdots) \le \Phi(1^{a_0}01^{a_1}01^{a_2}0\cdots)$ Using this fact we obtain the images of the value function on cylinders (see Figure 1.8).

$$\begin{aligned} \Phi([0]) &= [0, 1], \\ \Phi([1^{a_0}]) &= [a_0, \infty], \text{ for } a_0 > 0 \\ \Phi([1^{a_0}]) &= [\infty, a_0 + 1], \text{ for } a_0 < 0 \\ \Phi([1^{a_0}0]) &= [a_0, a_0 + 1] \\ \Phi([1^{a_0}0\cdots 1^{a_{n-1}}01^{a_n}]) &= [\tfrac{p_{n-1}}{q_{n-1}}, \tfrac{p_n}{q_n}], \text{ for } n \text{ odd} \\ \Phi([1^{a_0}0\cdots 1^{a_{n-1}}01^{a_n}]) &= [\tfrac{p_n}{q_n}, \tfrac{p_{n-1}}{q_{n-1}}], \text{ for } n \text{ even} \\ \Phi([1^{a_0}0\cdots 1^{a_{n-1}}01^{a_n}0]) &= \Phi([1^{a_0}0\cdots 1^{a_{n-1}}01^{a_n}01]) \end{aligned}$$

Since the angle length of these intervals converges to zero with the increasing length of words, the value mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous.

There is another number system based on continued fractions. Using the fact that F_{00} is the identity, we replace a word $u = 1^{a_0}01^{a_1}01^{a_2}\cdots$ of Σ_D by $v = 1^{a_0}(010)^{a_1}1^{a_2}(010)^{a_3}\cdots$. We replace now F_0 by $F_{010}(x) = x/(x+1)$, which maps the unit interval $[0, 1]$ to $[0, \infty]$. To make the system symmetric, we take also $F_{0\overline{1}0}(x) = x/(-x+1)$ and apply it to the interval $[-1, 0]$.

Definition 1.14 *The number system of* **symmetric continued fractions** *has alphabet* $A = \{\overline{1}, \overline{0}, 0, 1\}$, *transformations and intervals*

$$\begin{array}{rclrclrclrcl} F_{\overline{1}}(x) &=& x-1, & F_{\overline{0}}(x) &=& \frac{x}{-x+1}, & F_0(x) &=& \frac{x}{x+1}, & F_1(x) &=& x+1, \\ W_0 &=& [\infty, -1], & W_{\overline{0}} &=& [-1, 0], & W_0 &=& [0, 1], & W_1 &=& [1, \infty], \end{array}$$

and forbidden words $D = \{\overline{0}0, \overline{0}1, \overline{1}0, \overline{1}1, 0\overline{1}, 0\overline{0}, 1\overline{1}, 1\overline{0}\}$.

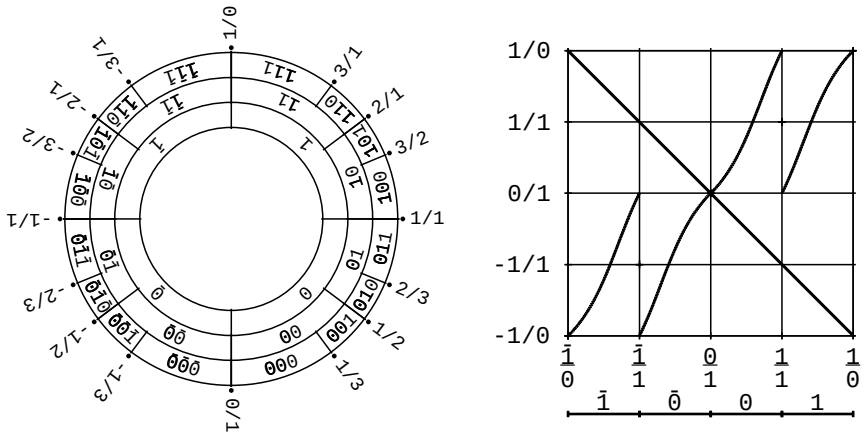


Figure 1.9: The number system of symmetric continued fractions with $A = \{\overline{1}, \overline{0}, 0, 1\}$, $F_{\overline{1}}(x) = x - 1$, $F_{\overline{0}}(x) = \frac{x}{1-x}$, $F_0(x) = \frac{x}{x+1}$, $F_1(x) = x + 1$, $D = \{\overline{0}0, \overline{0}1, \overline{1}0, \overline{1}1, 0\overline{1}, 0\overline{0}, 1\overline{1}, 1\overline{0}\}$.

Then $\Sigma_D = \{\overline{1}, \overline{0}\}^\omega \cup \{0, 1\}^\omega$,

$$\begin{aligned} F_{\overline{1}}^{-1}(W_{\overline{1}}) = F_{\overline{0}}^{-1}(W_{\overline{0}}) &= [-\infty, 0] = W_{\overline{1}} \cup W_{\overline{0}}, \\ F_0^{-1}(W_0) = F_1^{-1}(W_1) &= [0, \infty] = W_0 \cup W_1. \end{aligned}$$

Thus $F_a^{-1}(W_a) = \cup\{W_b : \ ab \in \mathcal{L}_D\}$, and $\overline{\mathbb{R}} = \cup\{W_a : \ a \in A\}$. A word $u \in \{0,1\}^\omega$ can be written as $u = 1^{a_0}0^{a_1}1^{a_2}\cdots$, where $a_0 \geq 0$ and $a_i > 0$ for $i > 0$. The sequence of a_i may be finite if its last element a_n is infinite.

Proposition 1.15 *The value mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ of the system of symmetric continued fractions defined by*

$$\begin{aligned} \Phi(1^{a_0}0^{a_1}1^{a_2}\cdots) &= a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}} \\ \Phi(\overline{1}^{a_0}\overline{0}^{a_1}\overline{1}^{a_2}\cdots) &= -a_0 - \frac{1}{a_1 - \frac{1}{a_2 - \cdots}} \\ &= -\left(a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}\right) \end{aligned}$$

is continuous and surjective.

Proof: For the cylinder intervals we get (see Figure 1.9)

$$\begin{aligned} F_u(x) &= a_0 + \frac{1}{a_1 + \cdots + \frac{1}{a_{2n} + x}}, \ u = 1^{a_0}0^{a_1}\cdots 1^{a_{2n}} \\ F_u(x) &= a_0 + \frac{1}{a_1 + \cdots + \frac{1}{a_{2n+1} + \frac{1}{x}}}, \ u = 1^{a_0}1^{a_1}\cdots 0^{a_{2n+1}} \\ \Phi[u] &= [\Phi(u0^\omega), \Phi(u1^\omega)] = [\tfrac{p_{2n}}{q_{2n}}, \tfrac{p_{2n-1}}{q_{2n-1}}], \ u = 1^{a_0}0^{a_1}\cdots 1^{a_{2n}} \\ \Phi[u] &= [\Phi(u0^\omega), \Phi(u1^\omega)] = [\tfrac{p_{2n}}{q_{2n}}, \tfrac{p_{2n+1}}{q_{2n+1}}], \ u = 1^{a_0}1^{a_1}\cdots 0^{a_{2n+1}} \end{aligned}$$

Thus $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective by Proposition 1.7. □

Chapter 2

Symbolic dynamics

A number system consists of a continuous value mapping whose domain is a symbolic space of infinite words and whose range is the extended real line. We say that the value mapping is a **symbolic extension** of $\overline{\mathbb{R}}$. The properties of symbolic spaces and symbolic extensions are treated in symbolic dynamics, which is based on the theory of compact metric spaces. See e.g., Hocking and Young [24] for an introduction to the theory of metric spaces.

2.1 Metric spaces

Definition 2.1 *A* **metric space** (X, d) *consists of a set* X *and a metric* $d : X \times X \to [0, \infty)$ *which gives the distance* $d(x, y)$ *of points* $x, y \in X$*. The following properties are assumed:*

1. $d(x, y) = 0 \Leftrightarrow x = y$,
2. $d(x, y) = d(y, x)$: *symmetry,*
3. $d(x, z) \leq d(x, y) + d(y, z)$: *triangle inequality.*

We refer to elements of X as points. A classical example of a metric space is the n-dimensional **Euclidean space** $\mathbb{R}^n = \{x = (x_1, \ldots, x_n) : \ x_i \in \mathbb{R}\}$ with metric

$$d_e(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

In particular, the set $\mathbb{R}$ of real numbers is a metric space with metric $d_e(x, y) = |x - y|$. The extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is a metric space with the angle metric (see Section 1.3)

$$d_a(x, y) = \frac{1}{\pi}\text{arccotg}\ \frac{|xy + 1|}{|y - x|}, \ d_a(x, \infty) = \frac{1}{\pi}\text{arccotg}|x|.$$

If (X, d) is a metric space and $Y \subseteq X$, then d restricted to $Y \times Y$ is a metric on Y and we say that (Y, d) is a **subspace** of (X, d). The **ball** with center $x \in X$ and radius $r > 0$ is the set

$$B_r(x) = \{y \in X : \ d(y, x) < r\}.$$

In $\mathbb{R}$, balls are open intervals $B_r(x) = (x - r, x + r)$. The **interior** Y° and **closure** $\overline{Y}$ of a set $Y \subseteq X$ are defined by

$$\begin{aligned} Y^\circ &= \{x \in X : \ \exists r > 0, B_r(x) \subseteq Y\}, \\ \overline{Y} &= \{x \in X : \ \forall r > 0, B_r(x) \cap Y \neq \emptyset\}, \end{aligned}$$

so $Y^\circ \subseteq Y \subseteq \overline{Y}$, $\overline{X \setminus Y} = X \setminus Y^\circ$, and $(X \setminus Y)^\circ = X \setminus \overline{Y}$, where $X \setminus Y = \{x \in X : \ x \notin Y\}$ is the **set difference** of Y from X. For example, if $Y = [0,1) \subset \mathbb{R}$ is a semiclosed interval, then $Y^\circ = (0,1)$ and $\overline{Y} = [0,1]$. If $Y, Z \subseteq X$, then

$$\begin{aligned} (Y \cap Z)^\circ &= Y^\circ \cap Z^\circ, \\ (Y \cup Z)^\circ &\supseteq Y^\circ \cup Z^\circ, \\ \overline{Y \cap Z} &\subseteq \overline{Y} \cap \overline{Z}, \\ \overline{Y \cup Z} &= \overline{Y} \cup \overline{Z}. \end{aligned}$$

A set $Y \subseteq X$ is **open**, if $Y = Y^\circ$, and **closed** if $\overline{Y} = Y$. It follows that $Y \subseteq X$ is closed iff $X \setminus Y$ is open. The interior of a set Y is the largest open set included in Y and the closure of Y is the smallest closed set which includes Y. It follows from the triangle inequality that every ball $B_r(x)$ is an open set. A semi-open (or semi-closed) interval $[a,b) = \{x \in \mathbb{R} : \ a \leq x < b\}$ is neither closed nor open in $\mathbb{R}$. A set is **clopen** if it is both closed and open. The sets $\emptyset$ and X are clopen in any metric space. If they are the only clopen sets, then we say that X is a **connected space**. The Euclidean space $\mathbb{R}^n$ is connected. The union of two intervals $[0,1] \cup [2,3]$ is not connected, since $[0,1]$ and $[2,3]$ are its clopen sets.

A sequence $\{x_n \in X : \ n \geq 0\}$ of points of X converges to a point $x \in X$ if for every $\varepsilon > 0$ there exists n_0 such that $d(x_n, x) < \varepsilon$ for every $n \geq n_0$. A sequence cannot converge to two distinct points, so we write $\lim_{n\to\infty} x_n = x$ if x_n converge to x and say that $\{x_n : \ n \geq 0\}$ is a **convergent sequence**. A **subsequence** of $\{x_n : \ n \geq 0\}$ is any sequence $\{x_{n_i} : \ i \geq 0\}$, where $\{n_i : \ i \geq 0\}$ is an increasing sequence of indices.

Definition 2.2 *A metric space is* **compact** *if any its sequence has a converging subsequence. A subset of a metric space is compact, if it is compact as a subspace.*

The real line $\mathbb{R}$ is not compact, since the sequence $x_n = n$ has no converging subsequence. The open interval $(0,1)$ is not compact either since the sequence $x_n = 1/n$ has in $(0,1)$ no converging subsequence: all its subsequences converge to zero, which is not in the space $(0,1)$. A closed bounded interval $[a,b]$ is compact in $\mathbb{R}$. We show that a set $Y \subseteq \mathbb{R}^n$ is compact iff it is closed and bounded. We say that a set $Y \subseteq X$ is **bounded**, if $Y \subseteq B_r(x)$ for some $x \in X$ and $r > 0$. This happens iff the set has a finite **diameter** $\mathrm{diam}(Y) = \sup\{d(y,y') : \ y, y' \in Y\}$.

Proposition 2.3

1. A compact subset of a metric space is closed and bounded.

2. A closed subset of a compact space is compact.

3. A subset of an Euclidean space $\mathbb{R}^n$ is compact iff it is closed and bounded.

Proof: 1. Let $Y \subseteq X$ be compact and assume by contradiction that it is not closed, so there exists $y \in \overline{Y} \setminus Y$. For each $n > 0$ there exists $y_n \in Y$ such that $d(y_n, y) < 1/n$, so $\lim_{n\to\infty} y_n = y \in X \setminus Y$. Each subsequence of $\{Y_n : \ n \geq 0\}$ has the same limit y. This means that no its subsequence has a limit in Y. This is a contradiction. Assume that Y is not bounded. Take any $y_0 \in Y$. There exist points $y_n \in Y$ such that $d(y_n, y_0) > n$, and the sequence $\{y_n : \ n \geq 0\}$ has no converging subsequence. This is a contradiction.
2. Let X be compact and let $Y \subseteq X$ be closed. A sequence $\{y_n \in Y : \ n \geq 0\}$ has a subsequence which converges to some $y \in X$. Since X is closed, $y \in Y$, so Y is compact.
3. Let $Y \subseteq \mathbb{R}$ be closed and bounded and $x_n \in Y$. There exists an interval $[a_0, b_0] \supseteq Y$. Denote by $c_0 = \frac{a_0+b_0}{2}$. An infinite number of x_n belong either to $[a_0, c_0]$ or to $[c_0, b_0]$. In the former case set $[a_1, b_1] = [a_0, c_0]$ and in the latter case set $[a_1, b_1] = [c_0, b_0]$. Let n_1 be the first index

with $x_{n_1} \in [a_1, b_1]$. We continue by induction. At each step k the interval $[a_k, b_k]$ is one half of the interval $[a_{k-1}, b_{k-1}]$ and contains an infinite number of x_n. Let n_k be the smallest integer greater than n_{k-1} such that $x_{n_k} \in [a_k, b_k]$. Then $\{x_{n_k} : \ k \geq 1\}$ converges to the common limit of a_k and b_k. Since Y is closed, this limit belongs to Y, so Y is compact. If $Y \subseteq \mathbb{R}^n$ is closed and bounded, and $\{x_m = (x_{m,1}, \ldots, x_{m,n}) : \ m \geq 0\}$ is a sequence in Y, then for each coordinate $i \leq n$, $\{x_{m,i} : \ m \geq 0\}$ is a bounded sequence. There exists a subsequence whose first coordinate converges, a subsequence of this subsequence whose second coordinate converges, etc. Thus there exists a subsequence of $\{x_m : \ m \geq 0\}$ which converges in each coordinate. Since Y is closed, the limit belongs to Y. □

A **cover** of a space X is any collection $\mathcal{U} = \{U_i : \ i \in I\}$ of sets $U_i \subseteq X$ whose union is X. The index set I may be finite or infinite with arbitrary cardinality. If all U_i are open, we say that $\mathcal{U}$ is an **open cover**. If $J \subseteq I$ and $\bigcup_{i \in J} U_i = X$, then we say that $\{U_i : \ i \in J\}$ is a **subcover** of $\mathcal{U}$. The **diameter of a cover** is the supremum of the diameters of its elements.

Proposition 2.4 *Let X be a metric space. The following three conditions are equivalent.*
1. *X is compact.*
2. *Every open cover of X has a finite subcover.*
3. *If $\{V_n \subseteq X : \ n \geq 0\}$ is a sequence of closed nonempty sets such that $V_{n+1} \subseteq V_n$, then the intersection $\bigcap_{n \geq 0} V_n$ is nonempy.*

Proof: 1 $\Rightarrow$ 2: Assume that $\mathcal{U} = \{U_n \subseteq X : \ n \geq 0\}$ is a countable cover which does not have a finite subcover. Then there exist points $x_n \in U_n \setminus (U_0 \cup \cdots \cup U_{n-1})$. The sequence $\{x_n : \ n \geq 0\}$ has a converging subsequence $\lim_{k \to \infty} x_{n_k} = x$. Since $\mathcal{U}$ is a cover, $x \in U_n$ for some n. Since U_n is open, $x_{n_k} \in U_n$ for each sufficiently large n_k and this is a contradiction. If $\mathcal{U}$ is an uncountable cover, then its countable cover shoud be first found using the concept of countable open basis (see e.g., Hocking and Young [24]).
2 $\Rightarrow$ 3: Let $\emptyset \neq V_{n+1} \subseteq V_n \subseteq X$ be nonempty closed sets and assume that their intersection is empty. Then $\{U_n = X \setminus V_n : \ n \geq 0\}$ is an open cover of X and has a finite subcover, so there exists n such that $X = U_0 \cup \cdots \cup U_n = X \setminus V_n$. This implies $V_n = \emptyset$ which is a contradiction.
3 $\Rightarrow$ 1. Let $\{x_n \in X : \ n \geq 0\}$ be any sequence of points and set $V_n = \overline{\{x_i : i \geq n\}}$. Then $V_{n+1} \subseteq V_n$ are nonempty and closed, so there exists $x \in \bigcap_n V_n$. Since V_1 is closed, $B_1(x) \cap V_1 \neq \emptyset$, so there exists n_1 such that $x_{n_0} \in B_1(x)$. In a similar way we show that there exists $n_2 > n_1$ such that $x_{n_2} \in B_{1/2}(x)$. By induction we get a subsequence $\{x_{n_k} : \ k \geq 0\}$ with $x_{n_k} \in B_{1/k}(x)$, so $\lim_{k \to \infty} x_{n_k} = x$. □

A mapping $F : X \to Y$ from a set X to a set Y assigns to elements $x \in X$ elements $F(x) \in Y$. If $G : Y \to Z$ is another mapping, then the composition $G \circ F : X \to Z$ is defined by $(G \circ F)(x) = G(F(x))$. A mapping $F : X \to Y$ is **injective**, if $x \neq x' \in X$ implies $F(x) \neq F(x')$. It is **surjective**, if for each $y \in Y$ there exists $x \in X$ with $y = F(x)$. It is **bijective**, if it is one-to-one and surjective. A bijective mapping $F : X \to Y$ has the inverse mapping $F^{-1} : Y \to X$ such that $F^{-1}(F(x)) = x$ for every $x \in X$, so the compositions $F^{-1} \circ F = \mathrm{Id}_X$, $F \circ F^{-1} = \mathrm{Id}_Y$ are the identity mappings on X and Y. If (X, d_X) and (Y, d_Y) are metric spaces, then we say that $F : X \to Y$ is **continuous** at $x \in X$, if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x' \in X, (d_X(x, x') < \delta \ \Rightarrow \ d_Y(F(x), F(x')) < \varepsilon).$$

We say that F is continuous, if it is continuous at every point $x \in X$. We say that F is a **homeomorphism** if it is bijective and both F and F^{-1} are continuous. Metric spaces X, Y are **homeomorphic**, if there exists a homeomorphism from X to Y. For example, the function $F(x) = 1/x$ is a homeomorphism between the intervals $X = (0, 1)$ and $Y = (0, \infty)$.

Proposition 2.5 *A mapping $F : X \to Y$ beween metric spaces is continuous iff for every open set $U \subseteq Y$, the preimage $F^{-1}(U) = \{x \in X : \ F(x) \in U\}$ is an open set in X iff the preimage $F^{-1}(V)$ of every closed set $V \subseteq Y$ is a closed set.*

Proof: Assume that F is continuous and let $U \subseteq Y$ be an open set. If $x \in F^{-1}(U)$, then $F(x) \in U$, so there exists $\varepsilon > 0$ such that $B_\varepsilon(F(x)) \subseteq U$. By the continuity of F in x there exists $\delta > 0$ such that if $y \in B_\delta(x)$ then $F(y) \in B_\varepsilon(F(x)) \subseteq U$. This means that $B_\delta(x) \subseteq F^{-1}(U)$, so $F^{-1}(U)$ is open in X. Conversely assume that the preimage of any open set is open. Given $x \in X$ and $\varepsilon > 0$, the ball $U = B_\varepsilon(F(x))$ is an open set, so its preimage $F^{-1}(U)$ is open in X. Since $x \in F^{-1}(U)$ there exists $\delta > 0$ such that $B_\delta(x) \subseteq F^{-1}(U)$ and this is just the condition of continuity. If $V \subseteq Y$ is a closed set, then $F^{-1}(Y \setminus V) = X \setminus F^{-1}(V)$ is an open set so $F^{-1}(V)$ is a closed set. □

Proposition 2.6 *If X is a compact space and $F : X \to Y$ is continuous and surjective, then Y is compact. If F is also injective (and therefore bijective), then $F^{-1} : Y \to X$ is continuous, so F is a homeomorphism.*

Proof: Let $\{U_i : \ i \in I\}$ be an open cover of Y. Then $\{F^{-1}(U_i) : \ i \in I\}$ is an open cover of X so it has a finite subcover $\{F^{-1}(U_i) : \ i \in K\}$, and $\{U_i : \ i \in K\}$ is an open cover of Y. Thus Y is compact. Assume that F is bijective. We show that for each closed set $V \subseteq X$, $(F^{-1})^{-1}(V) \subseteq Y$ is a closed. Since V is a closed subset of a compact space, it is compact, so by the preceding proof, $(F^{-1})^{-1}(V) = F(V)$ is a compact set and therefore closed. □

The stereographic projection $\mathbf{d}(x) = \frac{2x+i(x^2-1)}{x^2+1}$ is a bijective mapping $\mathbf{d} : \overline{\mathbb{R}} \to \mathbb{S}$. With the angle metric on $\overline{\mathbb{R}}$ and the Euclidean metric on $\mathbb{S} \subset \mathbb{C}$, $\mathbf{d}$ is a homeomorphism. Since $\mathbb{S}$ is a closed and bounded subset of $\mathbb{C} \approx \mathbb{R}^2$, it is compact and $\overline{\mathbb{R}}$ is compact too.

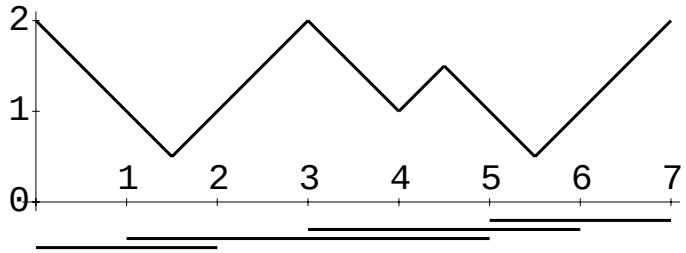


Figure 2.1: The function $f(x) = \sup\{r > 0 : \ \exists a \in A, B_r(x) \subseteq U_a\}$ for the cover $\mathcal{U} = \{[0,2),(1,5),(3,6),(5,7]\}$ of $X = [0,7]$.

Theorem 2.7 *Any open cover $\mathcal{U} = \{U_a : \ a \in A\}$ of a compact space X has a* **Lebesgue number** *$L > 0$ such that $\forall x \in X, \exists a \in A, B_L(x) \subseteq U_a$.*

Proof: Let $\mathcal{U} = \{U_a : \ a \in A\}$ be an open cover of X. If $U_a = X$ for some $a \in A$, then any $L > 0$ is a Lebesgue number of $\mathcal{U}$. Assume therefore that $U_a \neq X$ for each $a \in A$. Define a function $f : X \to (0, \infty)$ by

$$f(x) = \sup\{r > 0 : \ \exists a \in A, B_r(x) \subseteq U_a\} < \infty$$

We show that f is continuous: If $d(x,y) < \delta$ and $0 < r < f(x)$, then there exists $a \in A$ such that $B_r(x) \subseteq U_a$, $B_{r-\delta}(y) \subseteq U_a$, so $f(y) > r - \delta$. Since this holds for any $r < f(x)$, we get $f(y) \geq f(x) - \delta$. Interchanging x and y we get $f(x) \geq f(y) - \delta$, so $|f(x) - f(y)| \leq \delta$ and this proves the continuity of f (see Figure 2.1). By Proposition 2.6, a continuous image of a compact space is compact, so $f(X) \subseteq (0,\infty)$ is compact and therefore closed. Since $f(X)$ does not contain zero, its minimum $L_0 = \min f(X)$ is positive. If $0 < L < L_0$, then L is a Lebesgue number of $\mathcal{U}$. □

We say that a mapping $F : X \to Y$ is **uniformly continuous** if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x, x' \in X (d(x,x') < \delta \ \Rightarrow \ d(F(x), F(x')) < \varepsilon)$$

A uniformly continuous map is continuous. The map $f : (0,1) \to (0,\infty)$ defined by $f(x) = 1/x$ is continuous but not uniformly continuous.

Proposition 2.8 *If $F : X \to Y$ is a continuous map and X is compact, then F is uniformly continuous.*

Proof: Pick $\varepsilon > 0$. For each $x \in X$ there exists $\delta_x > 0$ such that if $d_X(y,x) < \delta_x$, then $d_Y(F(x), F(y)) < \frac{\varepsilon}{2}$. Let $\delta > 0$ be a Lebesgue number of an open cover $\mathcal{U} = \{B_{\delta_x}(x) : \ x \in X\}$. If $y, z \in X$ and $d_X(y,z) < \delta$, then there exists $x \in X$ such that $B_\delta(y) \subseteq B_{\delta_x}(x)$, so both y, z belong to $B_{\delta_x}(x)$ and therefore $d_Y(F(y), F(z)) \leq d_Y(F(y), F(x)) + d_Y(F(x), F(z)) < \varepsilon$. □

2.2 The Cantor space

Recall that if A is an alphabet (a finite set with at least two elements), then the **power space** A^ω is a metric space with metric

$$d(u,v) = 2^{-n}, \text{ where } n = \min\{k \geq 0 : \ u_k \neq v_k\}$$

Clearly d is symmetric, $d(u,v) = d(v,u)$ and $d(u,v) = 0$ iff $u = v$. To show that d satisfies the triangle inequality, let $d(u,v) = 2^{-n}$, $d(v,w) = 2^{-m}$ and $p = \min\{m,n\}$. Then $u_{[0,p)} = v_{[0,p)} = w_{[0,p)}$, so $d(u,w) \leq 2^{-p} \leq \max\{d(u,v), d(v,w)\} \leq d(u,v) + d(v,w)$.

To get insight to the topology of the power spaces A^ω, we show that these spaces are homeomorphic to the Cantor middle third set

$$C = [0,1] \setminus (\tfrac{1}{3}, \tfrac{2}{3}) \setminus (\tfrac{1}{9}, \tfrac{2}{9}) \setminus (\tfrac{7}{9}, \tfrac{8}{9}) \setminus (\tfrac{1}{27}, \tfrac{2}{27}) \setminus \cdots$$

The set C is obtained from the closed unit interval $[0,1]$ by deleting the open middle third interval $(\frac{1}{3}, \frac{2}{3})$ and repeating this deleting procedure indefinitely with the remaining closed intervals (see Figure 1.3). If we express the numbers $x \in [0,1]$ in the ternary system $x = \sum_{n \geq 0} u_n 3^{-n-1}$, where $u_n \in \{0,1,2\}$, then the interval $(\frac{1}{3}, \frac{2}{3})$ consists of points whose first digit is $u_0 = 1$. The endpoints of this intervals have two expansions: $\frac{1}{3} = .10^\omega = .02^\omega$, $\frac{2}{3} = .20^\omega = .12^\omega$, so $[0,1] \setminus (\frac{1}{3}, \frac{2}{3})$ consists of points which have ternary expansions with $u_0 \neq 1$. By induction, we show that C consists of points which have ternary expansions with digits $u_i \in \{0,2\}$.

Proposition 2.9 *The Cantor middle third set $C = [0,1] \setminus (\frac{1}{3}, \frac{2}{3}) \setminus (\frac{1}{9}, \frac{2}{9}) \setminus (\frac{7}{9}, \frac{8}{9}) \cdots$ is homeomorphic to $\{0,1\}^\omega$*

Proof: Define $\Phi_3 : \{0,1\}^\omega \to C$ by $\Phi_3(u) = \sum_{i\geq 0} 2u_i \cdot 3^{-i-1}$. If $d(u,v) = 2^{-n}$, then $u_{[0,n)} = v_{[0,n)}$, $u_n \neq v_n$, so

$$|\Phi_3(u) - \Phi_3(v)| = \left|\sum_{i=n}^{\infty} 2(u_i - v_i)3^{-i-1}\right| \leq 2\sum_{i=n}^{\infty} 3^{-i-1} = \frac{2\cdot 3^{-n-1}}{1-\frac{1}{3}} = 3^{-n}$$

$$|\Phi_3(u) - \Phi_3(v)| \geq 2\cdot 3^{-n-1} - 2\sum_{i=n+1}^{\infty} 3^{-i-1} = 3^{-n-1}$$

This shows that Φ_3 is bijective. If $d(u,v) < 2^{-n+1}$ then $|\Phi_3(u) - \Phi_3(v)| \leq 3^{-n}$ and if $|x-y| < 3^{-n-1}$ then $d(\Phi_3^{-1}(x), \Phi_3^{-1}(y)) < 2^{-n}$. This means that Φ_3 is a homeomorphism. □

While the Cantor middle third set C is obtained from the closed unit interval by deleting the middle thirds, the unit interval is obtained from the Cantor middle third set by gluing the endpoints of its cylinders. This is done by the mapping $\Phi_2 \circ \Phi_3^{-1} : C \to [0,1]$ (see Figure 2.2 left), where $\Phi_3 : \{0,1\}^\omega \to C$ is the homeomorphism from the proof of Proposition 2.9 and $\Phi_2 : \{0,1\}^\omega \to [0,1]$ is defined by $\Phi_2(u) = \sum_{i=0}^{\infty} u_i \cdot 2^{-i-1}$, The mapping $\Phi_2 \circ \Phi_3^{-1}$ defined on C can be extended to a continuous mapping $f : [0,1] \to [0,1]$ which is constant on the intervals deleted from the Cantor middle third set. This mapping is known as the Devil's staircase (see Figure 2.2 right).

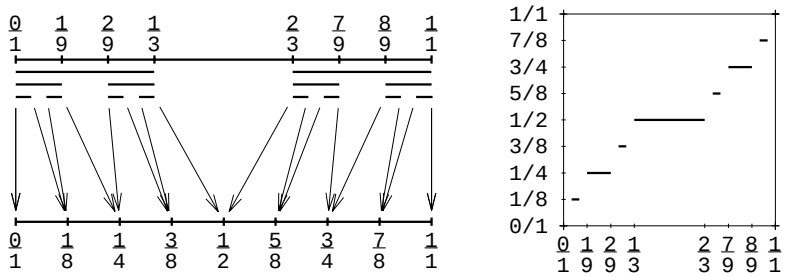


Figure 2.2: The mapping $\Phi_2 \circ \Phi_3^{-1} : C \to [0,1]$ (left) and the Devil's staircase (right)

Proposition 2.10 *If A is an alphabet, then A^ω is homeomorphic to $\{0,1\}^\omega$.*

Proof: For $A = \{0,1,\ldots,k\}$, $k > 2$ define a bijective map $\psi : A^\omega \to \{0,1\}^\omega$ by $\psi(a) = 1^a 0$ for $a < k$ and $\psi(k) = 1^k$. If $d(x,y) \leq 2^{-n}$ then $d(\psi(x), \psi(y)) \leq 2^{-n}$ since the length of each $\psi(a)$ is at least 1. If $d(\psi(x), \psi(y)) \leq 2^{-kn}$ then $d(x,y) \leq 2^{-n}$ since the length if each $\psi(a)$ is at most k. Thus both ψ and ψ^{-1} are continuous. □

Proposition 2.11 *If A is an alphabet and $u \in A^*$, then the* **cylinder**

$$[u] = \{w \in A^\omega : \ w_{[0,n)} = u\}$$

of u is a clopen (closed and open) set.

Proof: If $w \in [u]$ then $[u] = B_{2^{-n+1}}(w)$ is an open ball (whose center is any its element), so $[u]$ is an open set. The complement $A^\omega \setminus [u] = \bigcup\{[v] : \ v \in A^n \setminus \{u\}\}$ is a union of open sets so it is open and therefore $[u]$ is closed. □

We characterize the power spaces A^ω by three topological properties.

Definition 2.12
1. A metric space X is **perfect** *if it has no isolated points, i.e., if*

$$\forall x \in X, \forall \varepsilon > 0, \exists y \in X, 0 < d(y,x) < \varepsilon$$

2. A metric space X is **totally disconnected** *if points can be separated by clopen sets, i.e., if*

$$x \neq y \;\Rightarrow\; \exists W \text{ clopen}, x \in W, y \in X \setminus W$$

3. A metric space is a **Cantor space** *if it is compact, perfect, and totally disconnected.*

Theorem 2.13 *A metric space is a Cantor space iff it is homeomorphic to a power space A^ω.*

Proof: 1. We show that A^ω is compact. Let $w_n \in A^\omega$ be a sequence of points and denote by $w_{n,k} \in A$ the k-th letter of w_n. There exists $z_0 \in A$ such that the set $N_0 = \{n \in \mathbb{N} :\; w_{n,0} = z_0\}$ is infinite. Choose $n_0 \in N_0$. There exists $z_1 \in A$ such that the set $N_1 = \{n \in N_0 :\; w_{n,1} = z_1\}$ is infinite. Choose $n_1 \in N_1$ with $n_1 > n_0$ and continue by induction. If $n_k \in N_k$ has been already constructed then there exists z_{k+1} such that the set $N_{k+1} = \{n \in N_k :\; w_{n,k+1} = z_{k+1}\}$ is infinite and we take $n_{k+1} \in N_{k+1}$ with $n_{k+1} > n_k$. Then $(w_{n_k})_{[0,k]} = z_{[0,k]}$, so $\lim_{k\to\infty} w_{n_k} = z$.
2. We show that A^ω is perfect: For $w \in A^\omega$ there exists $z \in A^\omega$ with $z_{[0,n)} = w_{[0,n)}$, $z_n \neq w_n$, so $d(w,z) = 2^{-n}$
3. We show that A^ω is totally disconnected: For $w \neq z$ there exists n such that $w_n \neq z_n$, $w \in W = [w_{[0,n]}]$, $z \in A^\omega \setminus W$. The converse proof that each Cantor space is homeomorphic to $\{0,1\}^\omega$ can be found e.g., in Hockinkg and Young [24] or Kůrka [35]. □

We say that a metric space X is a **symbolic space** if it is homeomorphic to a closed subspace of A^ω. Symbolic spaces are compact and totally disconnected but not necessarily perfect. For example, every finite metric space is a symbolic space. Continuous mappings between symbolic spaces can be characterized combinatorially:

Proposition 2.14 *A mapping $F : A^\omega \to B^*$ between symbolic spaces is continuous iff there exists a sequence of mappings $\{f_n : A^{k_n} \to A$ such that $F(u)_n = f_n(u_{0,k_n})$.*

Proof: By definition, F is continuous iff for every $\varepsilon = 2^{-n}$ there exists $\delta = 2^{-k_n}$ such that

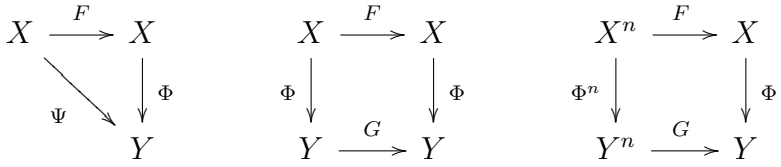
$$\begin{aligned} d(u,v) < \delta \quad &\Rightarrow \quad d(F(u),F(v)) < \varepsilon \\ u_{0,k_n} = v_{0,k_n} \quad &\Rightarrow \quad F(u)_{[0,n)} = F(v)_{[0,n)} \end{aligned}$$

Thus $F(u)_n$ depends only on $u_{[0,n)}$ and this dependence defines f_n. □

2.3 Redundant symbolic extensions

If we have a symbolic extension $\Phi : X \to \overline{\mathbb{R}}$, we want to perform arithmetical operations on symbolic representations of real numbers. A unary arithmetical operation like a linear function $g(x) = ax + b$ is a continuous mapping on $\overline{\mathbb{R}}$ (with $g(\infty) = \infty$). Its symbolic extension is a mapping $f : X \to X$ such that $g(\Phi(x)) = \Phi(f(x))$ for each $x \in X$. Symbolic extensions of continuous mappings exist provided Φ is redundant, i.e., if the images $\Phi([u])$ of cylinders overlap in $\overline{\mathbb{R}}$. The redundancy encountered in Section 1.2 is thus a topological concept.

Definition 2.15 *We say that a continuous surjective mapping $\Phi : X \to Y$ is a* **symbolic extension***, if X is a symbolic space. We say that a continuous mapping $\Phi : X \to Y$ is* **redundant***, if for each continuous mapping $\Psi : X \to Y$ there exists a continuous mapping $F : X \to X$ such that $\Phi \circ F = \Psi$.*



If $\Phi : X \to Y$ is a redundant symbolic extension, then continuous self maps of Y can be lifted to X. If $G : Y \to Y$ is a continuous mapping, then for $G \circ \Phi : X \to Y$ there exists a continuous mapping $F : X \to X$ such that $\Phi \circ F = G \circ \Phi$. We say that F is an **extension** of G by Φ. This can be generalized to mappings of several variables:

Proposition 2.16 *Let $\Phi : X \to Y$ be a redundant symbolic extension. Then for each continuous mapping $G : Y^n \to Y$ there exists a continuous mapping $F : X^n \to X$ such that $\Phi \circ F = G \circ \Phi^n$ (see the diagram).*

Proof: If X is a Cantor space, then X^n is also a Cantor space and therefore it is homeomorphic to X. Let $H : X^n \to X$ be the homeomorphism. For $G : Y^n \to Y$ we have a continuous mapping $g = G \circ \Phi^n \circ H^{-1} : X \to Y$, so there exists a continuous mapping $f : X \to X$ with $\Phi \circ f = g$. For $F = f \circ H : X^n \to X$ we get $\Phi \circ F = \Phi \circ f \circ H = g \circ H = G \circ \Phi^n \circ H^{-1} \circ H = G \circ \Phi^n$. □

The redundancy implies surjectivity: If $\Phi : X \to Y$ is redundant and $y \in Y$, then for the constant mapping $\Psi : X \to Y$ given by $\Psi(x) = y$ there exists a mapping $F : X \to X$ with $\Phi \circ F = \Psi$, so for any $x \in X$, $\Phi(F(x)) = \Psi(x) = y$. Since the continuous image of a compact space is compact, only compact spaces can have symbolic extensions. In particular, the real line $\mathbb{R}$ has no symbolic extension.

Example 2.17 *1. The binary value map $\Phi_2 : \{0,1\}^\omega \to [0,1]$ defined by $\Phi_2(u) = \sum_{i \geq 0} u_i \cdot 2^{-i-1}$ is a symbolic extension which is not redundant.*

Proof: The mapping Φ_2 is clearly continuous and surjective. We show that it is not redundant. Let $c \in (0,1)$ be an irrational number and consider the mapping $g(x) = \frac{x}{2c}$. Since c is irrational, there exists a unique $u \in \{0,1\}^\omega$ with $\Phi_2(u) = c$. Assume that $f : \{0,1\}^\omega \to \{0,1\}^\omega$ is an extension of g by Φ_2 and denote by $a = f(u)_0 \in \{0,1\}$. Since f is continuous at u, there exists $n > 0$ such that $f([u_{[0,n)}]) \subseteq [a]$, so $g\Phi_2([u_{[0,n)}]) = \Phi_2 f([u_{[0,n)}]) \subseteq \Phi_2([a])$. However, c is an inner point of $\Phi_2([u_{[0,n)}])$ and $g(c) = \frac{1}{2}$, so $g\Phi_2([u_{[0,n)}])$ is included neither in $\Phi_2([0]) = [0, \frac{1}{2}]$ nor in $\Phi_2([1]) = [\frac{1}{2}, 1]$. This is a contradiction. □

Theorem 2.18 *If X is a Cantor space and Y is compact metric space, then there exists a symbolic redundant extension $\Phi : X \to Y$.*

Proof: We can assume $X = \{0,1\}^\omega$. There exists a finite open cover of Y of diameter at most $2^0 = 1$. Repeating some of the sets if necessary, we can assume that its number of elements is a power of 2. Thus there exists $n_0 > 0$, and an open cover $\mathcal{V}_0 = \{V_u : u \in \{0,1\}^{n_0}\}$ of X of diameter at most 1. Let $\lambda_0 > 0$ be its Lebesgue number. We continue by induction. Assume that we have constructed an open cover $\mathcal{V}_k = \{V_u : u \in \{0,1\}^{n_k}\}$ of diameter at most

2^{-k} and Lebesgue number $\lambda_k > 0$. There exists $n_{k+1} > n_k$, such that for each $u \in \{0,1\}^{n_k}$ there exists an open cover $\mathcal{W}(u) = \{W_{uv} : \ v \in \{0,1\}^{n_{k+1}-n_k}\}$ of $\overline{V_u}$ with diameter at most 2^{-k-1}. There exists $\lambda_{k+1} < \lambda_k$ which is a Lebesgue number of each $\mathcal{W}(u)$. Set $V_{uv} = V_u \cap W_{uv}$. Then $\mathcal{V}_{k+1} = \{V_{uv} : \ uv \in \{0,1\}^{n_{k+1}}\}$ is an open cover of Y with diameter at most 2^{-k-1} and Lebesgue number $\lambda_{k+1} > 0$. If $u \in \{0,1\}^{n_k}$ and $v \in \{0,1\}^{n_{k+1}-n_k}$, then $V_{uv} \subseteq V_u$. For $u \in \{0,1\}^\omega$, $\bigcap_{k\geq 0} \overline{V_{u_{[0,n_k)}}} \neq \emptyset$ has zero diameter and therefore contains a unique element

$$\Phi(u) \in \bigcap_{k\geq 0} \overline{V_{u_{[0,n_k)}}}.$$

Then $\Phi : \{0,1\}^\omega \to Y$ is continuous and surjective. We show that Φ is redundant. Let $\Psi : \{0,1\}^\omega \to Y$ be a continuous mapping. Then Ψ is uniformly continuous and there exists an increasing integer sequence $\{m_k : \ k \geq 0\}$ such that

$$d(x,y) < 2^{-m_k} \ \Rightarrow \ d(\Psi(x), \Psi(y)) < \lambda_k.$$

We construct a sequence of mappings $f_k : \{0,1\}^{m_k} \to \{0,1\}^{n_k}$ such that $\Psi([u]) \subseteq V_{f_k(u)}$ for $u \in \{0,1\}^{m_k}$. For $u \in \{0,1\}^{m_0}$ choose a point $x \in [u]$. Then $\Psi([u]) \subseteq B_{\lambda_0}(\Psi(x))$ by uniform continuity of Ψ. Since $\mathcal{V}_0$ has Lebesgue number λ_0, there exists $f_0(u) \in \{0,1\}^{n_0}$ such that $B_{\lambda_0}(\Psi(x)) \subseteq V_{f_0(u)}$. Thus $\Psi([u]) \subseteq V_{f_0(u)}$. Assume we have constructed $f_k : \{0,1\}^{m_k} \to \{0,1\}^{n_k}$. For $u \in \{0,1\}^{m_k}$, $v \in \{0,1\}^{m_{k+1}-m_k}$ we have $\Psi([uv]) \subseteq \Psi([u]) \subseteq V_{f_k(u)}$. Choose $x \in [uv]$. There exists $w \in \{0,1\}^{n_{k+1}-n_k}$ such that $\Psi([uv]) \subseteq B_{\lambda_{k+1}}(\Psi(x)) \subseteq V_{f_k(u)w}$ and we set $f_{k+1}(uv) = f_k(u)w$. Define $F : \{0,1\}^\omega \to \{0,1\}^\omega$ by $F(u)_{[0,n_k)} = f_k(u_{[0,m_k)})$. Then F is continuous. For each $u \in \{0,1\}^\omega$ we have $\Psi(u) \in \Psi([u_{[0,m_k)}]) \subseteq V_{f_k(u_{[0,m_k)})} = V_{F(u)_{[0,n_k)}}$. Since $\Phi F(u) \in \overline{V_{F(u)_{[0,n_k)}}}$, we get $\Psi(u) = \Phi F(u)$. □

If X is a metric space and $Y \subseteq X$, then Y is a metric space with the metric of X restricted to Y. The closure and interior of a set $V \subseteq Y$ in Y usually differs from its closure and interior in X. The closure of V in Y is $\{y \in Y : \ \forall r > 0, B_r(y) \cap V \neq 0\} = \overline{V} \cap Y$, where $\overline{V}$ is the closure of V in X. For the interior of V in Y we get

$$\mathrm{int}_Y(V) = \{y \in Y : \ \exists r > 0, B_r(y) \cap Y \subseteq V\} = Y \setminus \overline{Y \setminus V}$$

For example, $\mathrm{int}_{[0,2]}([0,1]) = [0,1)$: the point 0 is an inner point of $[0,1]$ regarded as a subspace of $[0,2]$.

Proposition 2.19 *Let $\Phi : A^\omega \to Y$ be a symbolic extension and assume that for every $u \in A^*$, $\{\mathrm{int}_{\Phi([u])}(\Phi([ua])) : \ a \in A\}$ is a cover of $\Phi([u])$. Then $\Phi : A^\omega \to Y$ is redundant.*

Proof: For each integer k there exists $\lambda_k > 0$ such that for each $u \in A^k$, the open cover $\{\mathrm{int}_{\Phi([u])}(\Phi([ua])) : \ a \in A\}$ of $\Phi([u])$ has a Lebesgue number λ_k. We can assume that $\lambda_{k+1} < \lambda_k$. If $\Psi : A^\omega \to Y$ is continuous, then it is uniformly continuous and there exists n_k such that if $d(u,v) < 2^{-n_k}$ then $d(\Psi(u), \Psi(v)) < \lambda_k$. We can assume that $n_{k+1} > n_k$. Similarly as in the proof of Theorem 2.18 we construct a continuous $F : A^\omega \to A^\omega$ with $\Phi \circ F = \Psi$. □

Positional number systems for bounded intervals studied in Section 1.4 can be obtained from contractive iterative systems. Recall that the diameter of a set $Y \subseteq X$ is $\mathrm{diam}(Y) = \sup\{d(x,y) : \ x,y \in Y\}$.

Definition 2.20 *Let X be a metric space.*

1. *We say that a mapping $F : X \to X$ is a* **contraction** *if there exists an increasing continuous function $\psi : [0,\infty) \to [0,\infty)$ such that $\psi(0) = 0$, $\psi(t) < t$ for $t > 0$ and $\mathrm{diam}(F(V)) \leq \psi(\mathrm{diam}(V))$ for every set $V \subseteq X$.*
2. *A* **contractive iterative system** *over an alphabet A is a pair (X, F), where X is a compact metric space and $F = \{F_a : X \to X : \ a \in A\}$ is a system of contractions indexed by the letters of A.*
3. *For a finite word $u \in A^n$ set $F_u = F_{u_0} \circ \cdots \circ F_{u_{n-1}}$. For the empty word set $F_\lambda = \mathrm{Id}_X$,*

Any contraction is continuous. We have $F_{uv} = F_u \circ F_v$ for any $u, v \in A^*$.

Theorem 2.21 *Let (X, F) be a contractive iterative system over A. There exists a continuous* **value mapping** *$\Phi : A^\omega \to X$ such that*

1. *$\{\Phi(u)\} = \bigcap_{n>0} F_{u_{[0,n)}}(X)$ for $u \in A^\omega$.*
2. *$F_u(\Phi(v)) = \Phi(uv)$ for $u \in A^*$, $v \in A^\omega$.*
3. *If $u \in A^*$ then $\Phi([u]) \subseteq F_u(X)$.*
4. *$\Phi(u) = \lim_{n\to\infty} F_{u_{[0,n)}}(z)$ for any $z \in X$.*
5. *$\Phi : A^\omega \to X$ is surjective iff $\bigcup_{a\in A} F_a(X) = X$.*
6. *If $\Phi : A^\omega \to X$ is surjective, then $\Phi([u]) = F_u(X)$ for each $u \in A^*$.*
7. *If every F_a is injective and $X = \bigcup_{a\in A} F_a(X)^\circ$, then $\Phi : A^\omega \to X$ is redundant.*

Proof: 1. Since $F_{u_{[0,n+1)}}(X) \subseteq F_{u_{[0,n)}}(X)$ are nonempty closed sets, their intersection is nonempty. We have

$$\mathrm{diam}(F_{u_{[0,n)}}(X)) \leq \psi(\mathrm{diam}(F_{u_{[1,n)}}(X))) \leq \psi^2(\mathrm{diam}(F_{u_{[2,n)}}(X))) \leq \cdots \leq \psi^n(\mathrm{diam}(X)).$$

Since $\lim_{n\to\infty} \psi^n(\mathrm{diam}(X)) = 0$, the intersection $\bigcap_{n>0} F_{u_{[0,n)}}(X)$ has zero diameter and contains a unique point which is by definition $\Phi(u)$.
2. Both $F_u(\Phi(v))$ and $\Phi(uv)$ belong to all $F_{uv_{[0,n)}}(X)$, so they are equal.
3. If $uv \in [u]$ then $\Phi(uv) = F_u(\Phi(v)) \in F_u(X)$, so $\Phi([u]) \subseteq F_u(X)$. Since $\mathrm{diam}(\Phi([u])) \leq \mathrm{diam}(F_u(X)) \leq \psi^{|u|}(\mathrm{diam}(X))$, $\Phi : A^\omega \to X$ is continuous.
4. Since $\Phi(u), F_{u_{[0,n)}}(z) \in F_{u_{[0,n)}}(X)$, we get $d(\Phi(u), F_{u_{[0,n)}}(z)) \leq \psi^n(\mathrm{diam}(X))$. It follows $\lim_{n\to\infty} F_{u_{[0,n)}}(z) = \Phi(u)$.
5. For each $u \in A^\omega$ we have $\Phi(u) \in F_{u_0}(X)$, so $\Phi(A^\omega) \subseteq \bigcup_{a\in A} F_a(X)$. If $\bigcup_{a\in A} F_a(X) \neq X$, then Φ is not surjective. Conversely, assume that $\bigcup_{a\in A} F_a(X) = X$. Then for every $u \in A^*$ we have $\bigcup_{a\in A} F_{ua}(X) = F_u\left(\bigcup_{a\in A} F_a(X)\right) = F_u(X)$. Given $x \in X$, there exists u_0 such that $x \in F_{u_0}(X)$, there exists u_1 such that $x \in F_{u_{[0,1]}}(X)$ and by induction we construct $u \in A^\omega$ such that $x \in F_{u_{[0,n)}}(X)$ for each n, so $x = \Phi(u)$.
6. If $x \in F_u(X)$ then $x = F_u(y)$ for some $y \in X$ and there exists $v \in A^\omega$ with $y = \Phi(v)$, so $x = \Phi(uv)$ and $x \in \Phi([u])$.
7. Since $\Phi([u]) = F_u(X)$, by Theorem 2.19 it suffices to show that $\{\mathrm{int}_{F_u(X)}(F_{ua}(X)) : \ a \in A\}$ is a cover of $F_u(X)$ for each $u \in A^*$. Let $x \in F_u(X)$, so $x = F_u(y)$ for some $y \in X$. By the assumption there exists $a \in A$ and $\varepsilon > 0$ such that $B_\varepsilon(y) \subseteq F_a(X)$. Since $F_u^{-1} : F_u(X) \to X$ is a homeomorphism, there exists $\delta > 0$ such that $F_u^{-1}(B_\delta(x)) \subseteq B_\varepsilon(y) \subseteq F_a(X)$, so $B_\delta(x) \subseteq F_{ua}(X)$ and $x \in \mathrm{int}_{F_u(X)}(F_{ua}(X))$. □

Thus for example $\Phi_2 : \{0,1\}^\omega \to [0,1]$ is the value mapping of the contractive iterative system $F_a(x) = \frac{x+a}{2}$ on alphabet $A = \{0,1\}$ while the mapping $\Phi : \{0,1,2\} \to [0,2]$ defined by $\Phi(u) = \sum_i u_i 2^{-i-1}$ is the value mapping of $F_a(x) = \frac{x+a}{2}$ on the alphabet $A = \{0,1,2\}$.

2.4 Subshifts

The value mappings of number systems for the whole $\overline{\mathbb{R}}$ are usually not defined on a whole symbolic space A^ω but on some its subshift. Subshifts are treated in symbolic dynamic (see e.g., Lind and Marcus [46] or Kůrka [35]).

Definition 2.22 *For an alphabet A and a set $D \subseteq A^*$ of* **forbidden words**, *denote by*

$$\Sigma_D = \{u \in A^\omega : \forall v \in D : v \not\sqsubseteq u\}.$$

We say that a nonempty set $\Sigma \subseteq A^\omega$ is a **subshift**, *if $\Sigma = \Sigma_D$ for some $D \subseteq A^*$. If $D \subset A^*$ is a finite set then we say that Σ_D is a* **subshift of finite type** *(SFT). The* **order** *of a SFT Σ is the smallest $p \geq 2$ such that there exists $D \subseteq A^p$ with $\Sigma = \Sigma_D$.*

To forbid a word $u \in A^*$ is equivalent to forbidding words ua for all $a \in A$. Thus any SFT has an order. For example the SFT $\Sigma_{\{00,111\}} = \Sigma_{\{000,001,111\}}$ in $A = \{0,1\}$ has order 3. Some examples of SFT of order 2 in the alphabet $A = \{0,1\}$ are

$$\begin{aligned} \Sigma_{\{00,11\}} &= \{(01)^\omega, (10)^\omega\}, \\ \Sigma_{\{10\}} &= \{0^n 1^\omega : n \geq 0\} \cup \{0^\omega\} \\ \Sigma_{\{11\}} &= \{0, 10\}^\omega \end{aligned}$$

The subshift $\Sigma_{\{00,11\}}$ is finite, $\Sigma_{\{10\}}$ is countable and $\Sigma_{\{11\}}$ is uncountable: any concatenation of 10 with 0 belongs to $\Sigma_{\{11\}}$. An example of a subshift which is not SFT is the **occurrence one subshift** of words which contain at most one occurrence of 1. Its forbidden set is $D = \{10^n 1 : n \geq 0\}$. The **shift map** $\sigma : A^\omega \to A^\omega$ is defined by $\sigma(u)_i = u_{i+1}$. Thus $\sigma(u)$ is obtained form u by forgeting the first letter u_0. The shift map is continuous since $d(\sigma(u), \sigma(v)) \leq 2d(u,v)$.

Proposition 2.23 *A nonempty set $\Sigma \subseteq A^\omega$ is a subshift iff it is closed and shift-invariant, i.e., if $\sigma(w) \in \Sigma$ whenever $w \in \Sigma$.*

Proof: If forbidden words do not occur in w then they do not occur in $\sigma(w)$, so Σ_D is shift-invariant. To show that Σ_D is closed, we show that its complement is open. If $u \in A^\omega \setminus \Sigma_D$, then for some $i < j$, $u_{[i,j)} \in D$, and no $w \in A^\omega$ with $w_{[0,j)} = u_{[0,j)}$ belongs to Σ_D, so $[u_{[0,j)}] \subseteq A^\omega \setminus \Sigma_D$. This means that $A^\omega \setminus \Sigma_D$ is open and therefore Σ_D is closed. Conversely assume that $\Sigma \subseteq A^\omega$ is closed and shift-invariant and set

$$D = \{v \in A^* : \forall u \in \Sigma, v \not\sqsubseteq u\}.$$

If $u \in \Sigma$ and $v \in D$ then $v \not\sqsubseteq u$, so $u \in \Sigma_D$. Thus we have proved $\Sigma \subseteq \Sigma_D$. If $u \in A^\omega \setminus \Sigma$, then, since $A^\omega \setminus \Sigma$ is open, there exists $v = u_{[0,n)}$ such that $[v] \subseteq A^\omega \setminus \Sigma$. Assume by contradiction that v occurrs in some $w \in \Sigma$, so $v = w_{[i,i+n)}$. Then $\sigma^i(w) \in \Sigma$, but $\sigma^i(w) \in [v] \subseteq A^\omega \setminus \Sigma$ and this is a contradiction. It follows that $v \in D$ and therefore $u \in A^\omega \setminus \Sigma_D$. Thus we have shown $A^\omega \setminus \Sigma \subseteq A^\omega \setminus \Sigma_D$, so $\Sigma = \Sigma_D$. □

Definition 2.24 *The* **language** *of a subshift $\Sigma \subseteq A^\omega$ is the set of finite words which occurr as subwords of infinite words of Σ:*

$$\mathcal{L}(\Sigma) = \{u \in A^* : \exists x \in \Sigma, u \sqsubseteq x\}.$$

We denote by $\mathcal{L}^n(\Sigma) = \mathcal{L}(\Sigma) \cap A^n$. If $\Sigma = \Sigma_D$ then we denote by $\mathcal{L}_D = \mathcal{L}(\Sigma_D)$, $\mathcal{L}_D^n = \mathcal{L}_D \cap A^n$.

Some examples are

$$\begin{aligned}\mathcal{L}_{\{00,11\}} &= \{\lambda, 0, 1, 01, 10, 010, 101, 0101, 1010, \ldots\},\\ \mathcal{L}_{\{10\}} &= \{\lambda, 0, 1, 00, 01, 11, 000, 001, 011, 111, \ldots\},\\ \mathcal{L}_{\{11\}} &= \{\lambda, 0, 1, 00, 01, 10, 000, 001, 010, 100, 101, \ldots\}.\end{aligned}$$

Definition 2.25 *A nonempty language $L \subseteq A^*$ is an* **extendable language**, *if*
1. $v \in L$ for any $v \sqsubseteq u \in L$,
2. for any $u \in L$ there exists $a \in A$ such that $ua \in L$.
The subshift of an extendable language $L \subseteq A^*$ is

$$\mathcal{S}(L) = \{x \in A^\omega : \ \forall n \geq 0, x_{[0,n)} \in L\}.$$

Proposition 2.26
1. If $L \subseteq A^*$ is an extendable language, then $\mathcal{S}(L)$ is a subshift and $\mathcal{L}(\mathcal{S}(L)) = L$.
2. If $\Sigma \subseteq A^\omega$ is a subshift, then $\mathcal{L}(\Sigma)$ is an extendable language and $\mathcal{S}(\mathcal{L}(\Sigma)) = \Sigma$.

Proof: 1. Let $L \subseteq A^*$ be an extendable language. For $n > 0$ set $X_n = \{x \in A^\omega : \ x_{[0,n)} \in L\}$, so $\mathcal{S}(L) = \bigcap_{n>0} X_n$. Since L contains words of any length, X_n is nonempty. Since X_n is a finite union of cylinders, it is closed. Since $X_{n+1} \subseteq X_n$, their intersection $\mathcal{S}(L)$ is nonempty and closed. Clearly, $\mathcal{S}(L)$ is invariant, so it is a subshift. We show $\mathcal{L}(\mathcal{S}(L)) = L$. If $u \in L$, $|u| = n$, then there exists $u_n \in A$ such that $u_{[0,n]} \in L$. Repeating this infinitely many times we extend u to a point $x \in A^\omega$ such that for any m, $x_{[0,m)} \in L$. Thus $x \in \mathcal{S}(L)$ and $u \in \mathcal{L}(\mathcal{S}(L))$, so $L \subseteq \mathcal{L}(\mathcal{S}(L))$. If $u \in \mathcal{L}(\mathcal{S}(L))$, then there exists $x \in \mathcal{S}(L)$ with $u = x_{[i,j)}$ for some $i < j$. Since $x_{[0,j)} \in L$ and u is its subword, $u \in L$. Thus $\mathcal{L}(\mathcal{S}(L)) \subseteq L$.
2. Let $\Sigma \subseteq A^\omega$ be a subshift. If $v \sqsubseteq u \in \mathcal{L}(\Sigma)$, then $u \sqsubseteq x$ for some $x \in \Sigma$ and therefore $v \sqsubseteq x$. If $u = x_{[i,i+|u|)}$, then $ux_{i+|u|} \sqsubseteq x$, so $ux_{i+|u|} \in \mathcal{L}(x)$. Thus we have proved that $\mathcal{L}(\Sigma)$ is extendable. We show $\mathcal{S}(\mathcal{L}(\Sigma)) = \Sigma$. If $x \in \Sigma$, then for any n, $x_{[0,n)} \in \mathcal{L}(\Sigma)$, so $x \in \mathcal{S}(\mathcal{L}(\Sigma))$. Thus $\Sigma \subseteq \mathcal{S}(\mathcal{L}(\Sigma))$. Suppose that $x \in \mathcal{S}(\mathcal{L}(\Sigma))$ and $x \notin \Sigma$. Since $A^\omega \setminus \Sigma$ is open, there exists n such that $[x_{[0,n)}] \subseteq A^\omega \setminus \Sigma$. Since $x \in \mathcal{S}(\mathcal{L}(\Sigma))$, $x_{[0,n)} \in \mathcal{L}(\Sigma)$ and there exists $y \in \Sigma$ such that $y_{[j,j+n)} = x_{[0,n)}$. Thus $\sigma^j(y) \in [x_{[0,n)}]$ and this is a contradiction. Thus $\mathcal{S}(\mathcal{L}(\Sigma)) \subseteq \Sigma$. □

If Σ is a subshift and $u \in \mathcal{L}(\Sigma)$ then we denote by

$$[u]_\Sigma = [u] \cap \Sigma = \{w \in \Sigma : \ w_{[0,|u|)} = u\}$$

For a fixed subshift Σ we often drop the index and write $[u]$ instead of $[u]_\Sigma$. We often consider symbolic extensions $\Phi : \Sigma \to \overline{\mathbb{R}}$ and in this case we have a generalization of the redundancy test, whose proof is the same as that of Theorem 2.19.

Theorem 2.27 *Let $\Sigma \subseteq A^\omega$ be a subshift and $\Phi : \Sigma \to Y$ a surjective continuous map such that for each $u \in \mathcal{L}(\Sigma)$, $\{\mathrm{int}_{\Phi([u])}(\Phi([ua])) : \ a \in A, ua \in \mathcal{L}(\Sigma)\}$ is a cover of $\Phi([u])$. Then Φ is redundant.*

2.5 Sofic subshifts

When we work with a subshift, we want to know whether an infinite word belongs to the subshift or not. Since we can work only with finite prefixes of infinite words, we need a device which reads successively letters of a word and stops (or signals an error) if the word read does not belong to the language of the subshift. In the case of an SFT (and in a more general class of sofic subshifts) such a test can be performed by a **finite automaton**. A finite automaton is a device with a finite set B of inner states. When the automaton reads a letter $a \in A$, it changes its inner state according to a mapping $\delta_a : B \to B$. The change of state upon reading a word $u \in A^2$ is $\delta_u(p) = \delta_{u_1}(\delta_{u_0}(q))$, so $\delta_{u_0u_1} = \delta_{u_1} \circ \delta_{u_0}$. For $u \in A^n$ we get analogously $\delta_u = \delta_{u_{n-1}} \circ \cdots \circ \delta_{u_0}$. If we set $\delta_\lambda = \mathrm{Id}_B$, then $\delta_{uv} = \delta_v \circ \delta_u$. Thus $\delta_a : B \to B$ form an iterative systems, but in contrast to iterative systems of Section 2.3, the mappings are composed in the reverse order. We asuume that the automaton has an initial state $\mathbf{i} \in B$ and a set of final (accepting) states $F \subseteq B$. A word $u \in A^*$ is accepted if $\delta_u(\mathbf{i}) \in F$. We say that $L \subseteq A^*$ is a **regular language**, if there exists a finite automaton $(B, \delta, \mathbf{i}, F)$ such that $u \in L$ iff $\delta_u(\mathbf{i}) \in F$.

If L is an extendable language and $\delta_u(\mathbf{i}) \in F$, then $\delta_v(\mathbf{i}) \in F$ for each prefix v of u: A word can be accepted only if all its prefixes have been accepted. This property leads to a simplification of the automaton since the rejecting states in $B \setminus F$ are not needed. We can remove them and leave $\delta_a(p)$ undefined whenever $\delta_a(p) \in B \setminus F$. Thus we get partial mappings $\delta_a : B \to B$ and we write $\exists\delta_a(p)$ when δ_a is defined at p. The compositions $\delta_u : B \to B$ are also partial mappings which are defined on $p \in B$ provided all δ_{u_i} are defined on $\delta_{u_{[0,i)}}(p)$.

Definition 2.28 *An accepting automaton over an alphabet A is a triple $\mathcal{A} = (B, \delta, \mathbf{i})$, where B is a finite set of states, $\delta_a : B \to B$ are partial mappings and $\mathbf{i} \in B$ is an initial state. The language accepted by $\mathcal{A}$ is $\mathcal{L}_\mathcal{A} = \{u \in A^* : \exists\delta_u(\mathbf{i})\}$. A subshift $\Sigma \subseteq A^\omega$ is* **sofic** *iff $\mathcal{L}(\Sigma)$ is a regular language iff there exists an accepting automaton $\mathcal{A}$ such that $\mathcal{L}(\Sigma) = \mathcal{L}_\mathcal{A}$.*

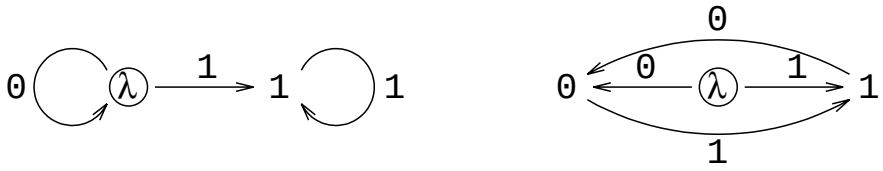


Figure 2.3: Accepting automata for SFT

We represent accepting automata by oriented labelled graphs whose vertices are states of B and whose edges are labelled by letters of A. The initial state is enclosed in a circle. There is an edge $p \xrightarrow{a} q$ from p to q with label a, if $\delta_a(p) = q$. The SFT $\Sigma_{\{10\}} = \{0^n1^\omega : n \geq 0\} \cup \{0^\omega\}$ has an accepting automaton with $B = \{\lambda, 1\}$, $\delta_0(\lambda) = \lambda$, $\delta_1(\lambda) = 1$, $\delta_1(1) = 1$ and initial state λ (Figure 2.3 left). The SFT $\Sigma_{\{00,11\}} = \{(01)^\omega, (10)^\omega\}$ has an accepting automaton with $B = \{\lambda, 0, 1\}$, initial state $\mathbf{i} = \lambda$, and transition function $\delta_a(\lambda) = a$, $\delta_a(a) = 1 - a$ for $a \in \{0, 1\}$ (Figure 2.3 right).

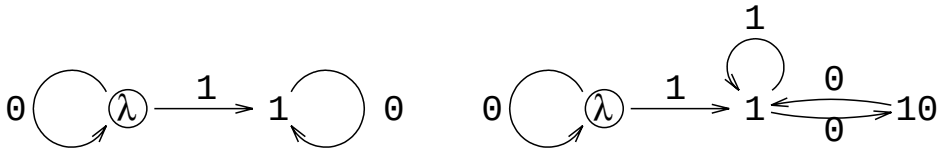


Figure 2.4: Accepting automata for sofic subshifts

We give examples of sofic subshifts which are not SFT. The **occurrence one subshift** in the binary alphabet $A = \{0, 1\}$ consists of words which contain at most one occurrence of the letter 1, so its forbidden set is $D = \{10^n1 : \ n \geq 0\}$. Its language is accepted by the automaton with states $B = \{\lambda, 1\}$, initial state λ and transition function $\delta_0(\lambda) = \lambda$, $\delta_1(\lambda) = 1$, $\delta_0(1) = 1$ (see Figure 2.4 left). The **even subshift** in the binary alphabet $A = \{0, 1\}$ consists of words which do not contain an odd number of zeros between two ones, so its forbidden set is $D = \{10^{2n+1}1 : \ n \geq 0\}$. Its language is accepted by the automaton with states $B = \{\lambda, 1, 10\}$, initial state λ and transition function $\delta_0(\lambda) = \lambda$, $\delta_1(\lambda) = 1$, $\delta_1(1) = 1$, $\delta_0(1) = 10$, $\delta_0(10) = 1$ (see Figure 2.4 right).

Definition 2.29 *Given a subshift $\Sigma \subseteq A^\omega$, the* **follower set** *of $u \in A^*$ is*

$$\mathcal{F}_u = \{v \in A^\omega : \ uv \in \Sigma\}.$$

Given an accepting automaton $\mathcal{A} = (B, \delta, \mathbf{i})$, the **follower set** *of $p \in B$ is*

$$\mathcal{F}_p = \{v \in A^\omega : \ \forall n, \exists \delta_{v_{[0,n)}}(p)\}.$$

Clearly $\mathcal{F}_u \neq \emptyset$ iff $u \in \mathcal{L}(\Sigma)$. For the empty word we have $\mathcal{F}_\lambda = \Sigma$. For the subshift $\Sigma_{\{11\}}$ there are just two follower sets: for each word $u \in \{0, 1\}^*$ we get $\mathcal{F}_{u0} = \Sigma$, $\mathcal{F}_{u1} = \{0u : \ u \in \Sigma\}$. For the occurrence one subshift we have also two follower sets: $\mathcal{F}_u = \Sigma$ provided $1 \not\sqsubseteq u$ and $\mathcal{F}_u = \{0^\omega\}$ otherwise.

Proposition 2.30 *If $u, v \in A^*$, $a \in A$ and $\mathcal{F}_u = \mathcal{F}_v$, then $\mathcal{F}_{ua} = \mathcal{F}_{va}$.*

Proof: Assume If $w \in \mathcal{F}_{ua}$ then $uaw \in \mathcal{L}(\Sigma)$, so $aw \in \mathcal{F}_u$, $aw \in \mathcal{F}_v$, and $w \in \mathcal{F}_{va}$. □

Theorem 2.31 *Σ is a sofic subshift iff the set $\{\mathcal{F}_u : \ u \in A^*\}$ of its follower sets is finite.*

Proof: If $\Sigma = \Sigma_\mathcal{A}$ with $\mathcal{A} = (B, \delta, \mathbf{i})$ and $u \in \mathcal{L}(\Sigma)$, then $\mathcal{F}_u = \mathcal{F}_p$ where $p = \delta_u(\mathbf{i}) \in B$. Since B is a finite set, $\{\mathcal{F}_u : \ u \in \mathcal{L}(\Sigma)\}$ is finite too. Conversely assume that $B = \{\mathcal{F}_u : \ u \in \mathcal{L}(\Sigma)\}$ is a finite set. We construct an accepting automaton $\mathcal{A} = (B, \delta, \mathcal{F}_\lambda)$ with initial state $\mathbf{i} = \mathcal{F}_\lambda = \Sigma$. Define the transition function by $\delta_a(\mathcal{F}_u) = \mathcal{F}_{ua}$ provided $ua \in \mathcal{L}(\Sigma)$, otherwise $\delta_a(\mathcal{F}_u)$ is undefined. By Proposition 2.30, this definition is correct. If $u \in \Sigma$ then $\delta_{u_{[0,n)}}(\mathcal{F}_\lambda) = \mathcal{F}_{u_{[0,n)}}$, so $u \in \Sigma_\mathcal{A}$. Conversely, if $\exists \delta_{u_{[0,n)}}(\mathcal{F}_\lambda)$ for each n, then $u_{[0,n)} \in \mathcal{L}(\Sigma)$, so $u \in \Sigma$. □

The construction of an accepting automaton is particularly simple for subshifts of finite type. If $\Sigma \subseteq A^\omega$ is a SFT of order $p \geq 2$, then an infinite word $u \in A^\omega$ belongs to Σ iff $u_{[n,n+p)} \in \mathcal{L}(\Sigma)$ for each n. It follows that for $u, v \in A^*$ with $|v| \geq p - 1$ we have $\mathcal{F}_{uv} = \mathcal{F}_v$, so $\{\mathcal{F}_v : \ |v| \leq p - 1\}$ is the set of all follower sets. Some of these sets, however, may coincide. This can be tested by a simple criterion: $\mathcal{F}_u = \mathcal{F}_v$ iff for all $w \in A^*$ with $|w| < p$, $uw \in \mathcal{L}(\Sigma)$ iff $vw \in \mathcal{L}(\Sigma)$.

2.6 Labelled graphs

Let $\mathcal{A} = (B, \delta, \mathbf{i})$ be an accepting automaton. We say that a state $p \in B$ is reachable, if $\delta_u(\mathbf{i}) = p$ for some u. In an accepting computation, only the reachable states appear, so we can remove all nonreachable states without changing the accepted language:

Proposition 2.32 *Let $\mathcal{A} = (B, \delta, \mathbf{i})$ be an accepting automaton whose every state is reachable. Then for each $u \in A^*$ we have $\exists \delta_u(\mathbf{i})$ iff $\exists p \in B, \exists \delta_u(p)$.*

Proof: If $\delta_u(p) = q$, and $\delta_v(\mathbf{i}) = p$, then $\delta_{vu}(\mathbf{i}) = q$ so $vu \in \mathcal{L}(\Sigma)$ and $u \in \mathcal{L}(\Sigma)$. □

In an accepting automaton whose only states are reachable, the initial state state need not be distinguished, since an accepting process can start at any state of B. The automaton is thus reduced to a partial iterative system $\delta_a : B \to B$. The accepted language of δ is $\mathcal{L}_\delta = \{u \in A^* : \exists p, \exists \delta_u(p)\}$. Since the computation may start at any state, we say that such an automaton is **nondeterministic**. A nondeterministic automaton may have fewer states that the deterministic one. For example if we remove from the accepting deterministic automaton of the even shift the initial state λ, we get a nondeterminsitic automaton which accepts the same language. Its states are $B = \{1, 10\}$, and transition function is given by $\delta_1(1) = 1$, $\delta_0(1) = 10$, $\delta_0(10) = 1$ (see Figure 2.4 right). We show that conversely, a language accepted by a nondeterministic finite automaton is accepted also by a deterministic automaton (Theorem 2.35), but its number of states may be much (exponentially) larger. A nondeterministic finite automaton can be equivalently described by a finite labelled graph.

Definition 2.33

1. *A* **labelled graph** *over an alphabet A is a pair $G = (B, E)$, where B is a finite set of vertices and $E \subseteq B \times A \times B$ is a set of labelled edges.*
2. *The source and target maps $s, t : E \to B$ are the projections $s(p, a, q) = p$, $t(p, a, q) = q$. We assume that $\forall p \in B, \exists e \in E, s(e) = p$. The labelling map $\ell : E \to A$ is the projection $\ell(p, a, q) = a$.*
3. *The* **edge subshift** *$\Sigma_{|G|}$ of G is $\Sigma_{|G|} = \{u \in E^\omega : \forall i \geq 0, t(u_i) = s(u_{i+1})\} \subseteq E^\omega$.*
3. *The* **subshift** *of G is $\Sigma_G = \{\ell(u) : u \in \Sigma_{|G|}\} \subseteq A^\omega$.*
4. *The* **language** *of G is $\mathcal{L}_G = \mathcal{L}(\Sigma_G)$.*

Note that $\Sigma_{|G|}$ is a SFT of order 2. A path is a finite or infinite word $u \in E^* \cup E^\omega$ such that $t(u_i) = s(u_{i+1})$. A finite path is equivalently described by a pair $(p, u) \in B^* \times A^*$ such that $|p| = |u| + 1$ and $(p_i, u_i, p_{i+1}) \in E$ for all $i < |u|$. An infinite path is a pair $(p, u) \in B^\omega \times A^\omega \approx (B \times A)^\omega$ such that $(p_i, u_i, p_{i+1}) \in E$ for all i. Thus the edge subshift may be equivalently defined as a subset of $(B \times A)^\omega$. The labelling map ℓ can be extended to the continuous mapping $\ell : E^\omega \to A^\omega$ defined by $\ell(u)_i = \ell(u_i)$. It follows that $\Sigma_G = \ell(\Sigma_{|G|})$ is compact and therefore it is a closed subset of A^ω. Since Σ_G is also shift-invariant, it is a subshift. Thus we have

Proposition 2.34 *If $\Sigma \subseteq A^\omega$ is a sofic subshift, then there exists a labelled graph G such that $\Sigma = \Sigma_G$.*

Proof: Given an accepting automaton $\mathcal{A} = (B, \delta, \mathbf{i})$, we construct the labelled graph $G = (B_0, E)$, where $B_0 = \{\delta_u(\mathbf{i}) : u \in A^*\}$ is the set of reachable states and $E = \{(p, a, q) \in B_0 \times A \times B_0 : \delta_a(p) = q\}$. □

Proposition 2.35 *Any subshift of any labelled graph is sofic.*

Proof: Let $G = (B, E)$ be a labelled graph, let $Q = \mathcal{P}(B) \setminus \{\emptyset\}$ be the set of nonempty subsets of B. Define transition functions $\delta_a : Q \to Q$ by

$$\delta_a(M) = \{q \in Q : \exists e \in E, s(e) \in M, t(e) = q, \ell(e) = a\},$$

provided $\delta_a(M)$ is not empty, otherwise $\delta_a(M)$ is undefined. The initial state is $B \in Q$. We show that (Q, δ, B) accepts $\mathcal{L}(\Sigma_G)$. If $q_0 \xrightarrow{u_0} q_1 \cdots \xrightarrow{u_{n-2}} q_{n-1} \xrightarrow{u_{n-1}} q_n$ is a path in G, then $q_n \in \delta_u(V)$, so $\delta_u(V) \neq \emptyset$ and u is accepted. Conversely, if $\delta_u(V) \neq \emptyset$, then pick some $q_n \in \delta_u(V)$. There exists $q_{n-1} \in \delta_{u_{[0,n-2]}}(V)$ such that $q_{n-1} \xrightarrow{u_{n-1}} q_n$ is a labelled edge in G. Continuing backwards, we obtain a path in G with label u. □

Definition 2.36 *A* **morphism** *from a subshift $\Sigma \subseteq A^\omega$ to a subshift $\Theta \subseteq B^\omega$ is a continuous mapping $F : \Sigma \to \Theta$ such that for every $u \in \Sigma$, $\sigma(F(u)) = F(\sigma(u))$. If F is surjective, we say that Θ is a* **factor** *of Σ.*

$$\begin{array}{ccc} \Sigma & \xrightarrow{\sigma} & \Sigma \\ {\scriptstyle F}\downarrow & & \downarrow{\scriptstyle F} \\ \Theta & \xrightarrow{\sigma} & \Theta \end{array}$$

Proposition 2.37 *Any morphism $F : \Sigma \to \Theta \subseteq B^\omega$ is a* **sliding block code**. *This means that there exists $r \geq 0$ and a* **local rule** *$f : \mathcal{L}^r(\Sigma) \to B$ such that $F(x)_i = f(x_{[i,i+r)})$ for every $x \in \Sigma$.*

Proof: Since F is uniformly continuous, for $\varepsilon = 1$ there exists $\delta > 0$ such that if $d(x, y) < \delta$, then $d(F(x), F(y)) < 1$. Take $r > 0$ with $2^{-r} < \delta$. Then

$$\begin{aligned} x_{[0,r)} = y_{[0,r)} &\Rightarrow d(x, y) \leq 2^{-r} < \delta \Rightarrow d(F(x), F(y)) < 1 \\ &\Rightarrow F(x)_0 = F(y)_0. \end{aligned}$$

Thus $F(x)_0$ depends only on the first r letters of x, and there exists a local rule $f : \mathcal{L}^r(\Sigma) \to B$ such that $f(x_{[0,r)}) = F(x)_0$. Since F is a morphism, we get

$$F(x)_n = \sigma^n(F(x))_0 = F(\sigma^n(x))_0 = f(\sigma^n(x)_{[0,r)}) = f(x_{[n,n+r)}). \quad \square$$

Theorem 2.38 (Weiss [69]) *A subshift is sofic iff it is a factor of an SFT.*

Proof: If Σ is sofic, then $\Sigma = \Sigma_G$ for some labelled graph G and $\ell : (\Sigma_{|G|}, \sigma) \to (\Sigma_G, \sigma)$ is a factor map with SFT $\Sigma_{|G|}$. Conversely, let $F : (\Sigma, \sigma) \to (\Theta, \sigma)$ be a factor map, $\Sigma \subseteq A^\omega$ an SFT and $\Theta \subseteq B^{\mathbb{N}}$. Let p be the order of Σ, so $u \in \Sigma$ iff $u_{[i,i+p)} \in \mathcal{L}(\Sigma)$ for all i. By Proposition 2.37, there exists a local rule $f : \mathcal{L}^r(\Sigma) \to B$ such that $F(x)_i = f(x_{[i,i+r)})$. We can assume $r \geq p$. Define a labelled graph $G = (V, E)$, where $V = \mathcal{L}^{r-1}(\Sigma)$,

$$E = \{(au, f(aub), ub) \in V \times B \times V : a, b \in A, aub \in \mathcal{L}^r(\Sigma)\}$$

We show that $\Sigma_G = \Theta$. If $v = F(u) \in \Theta$ then we have a path

$$u_{[0,r-1)} \xrightarrow{f(u_{[0,r)})} u_{[1,r)} \xrightarrow{f(u_{[1,r+1)})} u_{[2,r+1)} \cdots$$

with label v. Conversely, if we have such a path in (V, E) with label v, then $u_{[i,i+r-1)} \in \mathcal{L}(\Sigma)$, so $u \in \Sigma$ and $v = F(u)$. □

Definition 2.39 *Let $G = (B, E)$ be a labelled graph over A.*

1. *We say that G is* **initialized**, *if there exists $\mathbf{i} \in B$ such that $\mathcal{F}_{\mathbf{i}} = \Sigma_G$, there is no edge with target $\mathbf{i}$ and for each $p \in B \setminus \{\mathbf{i}\}$ there exists a path $\mathbf{i} \xrightarrow{u} p$.*
2. *We say that G is* **right-resolving** *if $(p, a, q), (p, b, r) \in E$ and $a = b$ implies $q = r$, i.e., if the edges with the same source carry different labels.*
3. *We say that G is* **deterministic**, *if it is initialized and right-resolving.*

For any graph G there exists an initialized graph with the same language. We just add to G a new vertex $\mathbf{i}$ and for any edge $p \xrightarrow{a} q$ we add a new edge $\mathbf{i} \xrightarrow{a} q$. Alternatively, if we allow edges with label λ, we may add edges $\mathbf{i} \xrightarrow{\lambda} p$ for each vertex p of G. The deterministic graphs are exactly graphs of deterministic finite automata, so each sofic subshift is a subshift of a deterministic graph. If G is an deterministic graph then there exists a continuous mapping $\nu : \Sigma_G \to \Sigma_{|G|}$ such that $\ell(\nu(u)) = u$ for each $u \in \Sigma_G$. For $u \in \Sigma_G$, $\nu(u)$ is the unique path with source $\mathbf{i}$ and label u. Note that ν is continuous but does not commute with the shift map, so it is not a morphism.

Chapter 3

Matrices and transformations

As we have seen in Chapter 1, an essential ingredient of a number system are its transformations $F_a : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$. These transformations are in all cases Möbius transformations of the form $M(x) = \frac{ax+b}{cx+d}$. Their geometrical structure can be understood in the context of projective geometry. The extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ can be regarded as the one-dimensional **projective space**. Möbius transformations are projective transformations of $\overline{\mathbb{R}}$ and form a three-dimensional projective space.

3.1 Projective geometry

Projective geometry (see e.g., Coxeter [9]) studies transformations which map lines to lines but do not necessarily preserve distances or angles. While the Euclidean geometry studies geometrical constructions with the compass and ruler, the projective geometry studies constructions with the ruler alone.

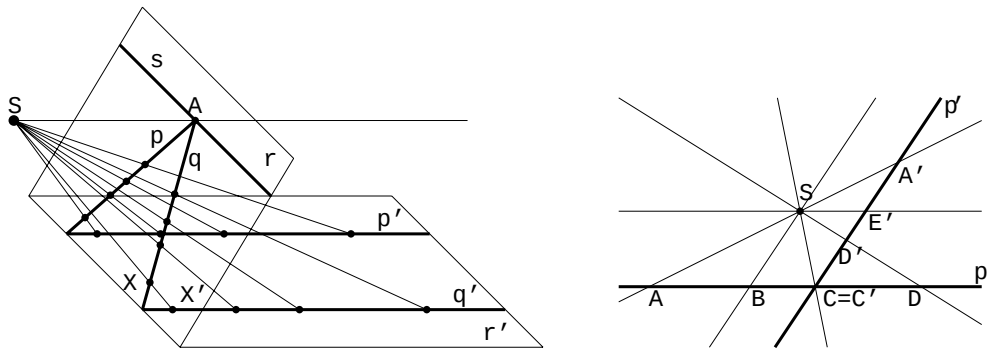


Figure 3.1: Perspectivities between planes(left) and lines (right).

A paradigmatic example is a **central perspectivity** (see Figure 3.1 left). We have two planes r and r' in a three-dimensional space and a center of perspectivity S which lies neither in r nor in r'. A point X of r is mapped to the intersection X' of the ray SX with the plane r'. A line q of r is mapped to the intersection of the plane Sq with the plane r'. This correspondence, however, is not defined everywhere. Some points in one plane do not have any image in the other plane. For example the lines p and q which intersect at A map to paralel lines p', q' which have no intersection in r'. The point A of r has no image in r'. To make the correspondence one-to-one, projective geometry extends the Euclidean plane by **ideal points at infinity**. The parallel lines p' and q' intersect at an ideal point A' of r'. Moreover, every

line parallel to p' intersects it at A' since every such line is mapped to a line which passes through A. The **projective plane** is obtained from the Euclidean plane by adding an ideal point in every direction (determined by a set of mutually parallel lines). Each (ordinary) line is extended by an ideal point. There is also an ideal line which consists of all ideal points. The line s of the plane r is mapped to the ideal line of r', since the plane Ss is parallel with r'. Thus any two different lines (ordinary or ideal) intersect in a unique point and any two different points determine a unique line on which these two points lie. The axiomatic of the projective geometry is thus simpler and more symmetric than the axiomatic of the Euclidean geometry, in which parallel lines do not intersect.

In a similar way we obtain the **projective line** - the projective space of dimension one. Consider a plane with two distinct lines p and p' and a point S of the plane, which lies neither on p nor on p' (Figure 3.1 right). The projectivity with the center S maps a point X of p to the intersection X' of p' with the line SX. To make the correspondence one-to-one, both lines p and p' are extended by a single ideal point at infinity. The point B of p projects to the ideal point B' of p' and the ideal point E of p projects to the point E' of p'.

There is another way to conceive a projective space without the cumbersome distinction between the ordinary and ideal points. Each point of a projective line (ordinary or ideal) is determined by a unique ray passing through S. If the ambient two-dimensional space is the Euclidean vector space $\mathbb{R}^2$, we can assume that the center of perspectivity is the zero point $S = 0 = (0,0)$. A ray passing through 0 is then just a one-dimensional subspace of $\mathbb{R}^2$. Similarly, points of a projective plane can be conceived as rays passing through a point S of a three-dimensional Euclidean space, or as one-dimensional subspaces of the three-dimensional vector space $\mathbb{R}^3$. The concept readily generalizes to any dimension.

Definition 3.1 *The* **projective space** $\mathbb{P}(\mathbb{R}^{n+1})$ *of dimension n consists of all one-dimensional subspaces of the vector space $\mathbb{R}^{n+1}$. The elements of $\mathbb{P}(\mathbb{R}^{n+1})$ are called* **projective points**. *A projective line in $\mathbb{P}(\mathbb{R}^{n+1})$ (for $n \geq 2$) is a linear subspace of $\mathbb{R}^{n+1}$ of dimension 2. The one-dimensional projective space is called the* **extended real line** $\mathbb{P}(\mathbb{R}^2) = \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

3.2 The extended real line

A one-dimensional subspace of $\mathbb{R}^2$ is determined by any its nonzero point $z = (z_0, z_1) \neq (0,0)$. We say that z is a **homogeneous coordinate** of the subspace $\{\lambda z : \ \lambda \in \mathbb{R}\}$. Two nonzero points z, w determine the same subspace, if one is a nonzero multiple of the other iff the matrix with columns z, w has zero determinant. We obtain an equivalence $\sim$ on $\mathbb{R}^2 \setminus \{(0,0)\}$ given by $z \sim w$ iff $\exists \lambda \neq 0, z = \lambda w$ iff $\det(z,w) = z_0 w_1 - z_1 w_0 = 0$. Thus we may conceive $\mathbb{P}(\mathbb{R}^2) = \overline{\mathbb{R}}$ as the factor space $\overline{\mathbb{R}} = (\mathbb{R}^2 \setminus \{(0,0)\})/\sim$. If we represent $\mathbb{R}$ by the line $z_1 = 1$ parallel to the z_0 axis, then the ray through a point $z = (z_0, z_1)$ with $z_1 \neq 0$ intersects the real line at $(\frac{z_0}{z_1}, 1)$, so it represents the number $\frac{z_0}{z_1} \in \mathbb{R}$. We write conventionally the homogeneous coordinate (z_0, z_1) as $\frac{z_0}{z_1}$, so the ideal point ∞ at infinity has homogeneous coordinate $\frac{z_0}{0}$, where $z_0 \neq 0$ (see Figure 3.2 top).

Of all homogeneous coordinates of a point $z = \frac{z_0}{z_1} \in \overline{\mathbb{R}}$ there are two which lie at the unit circle

$$\mathbb{S} = \{z \in \mathbb{R}^2 : \ z_0^2 + z_1^2 = 1\}.$$

They are $(\frac{z_0}{||z||}, \frac{z_1}{||z||})$, and $(\frac{-z_0}{||z||}, \frac{-z_1}{||z||})$, where $||z|| = \sqrt{z_0^2 + z_1^2}$ is the **norm** of z. The projective line is thus obtained from the unit circle by the identification of its opposite points. If $z \neq \infty$, then z has a unique homogeneous coordinate which lies at the upper semi-circle $\{z \in \mathbb{S} : \ z_1 \geq 0\}$. Both its endpoints $(-1,0)$ and $(1,0)$ represent ∞ (see Figure 3.2 bottom). If we stretch the

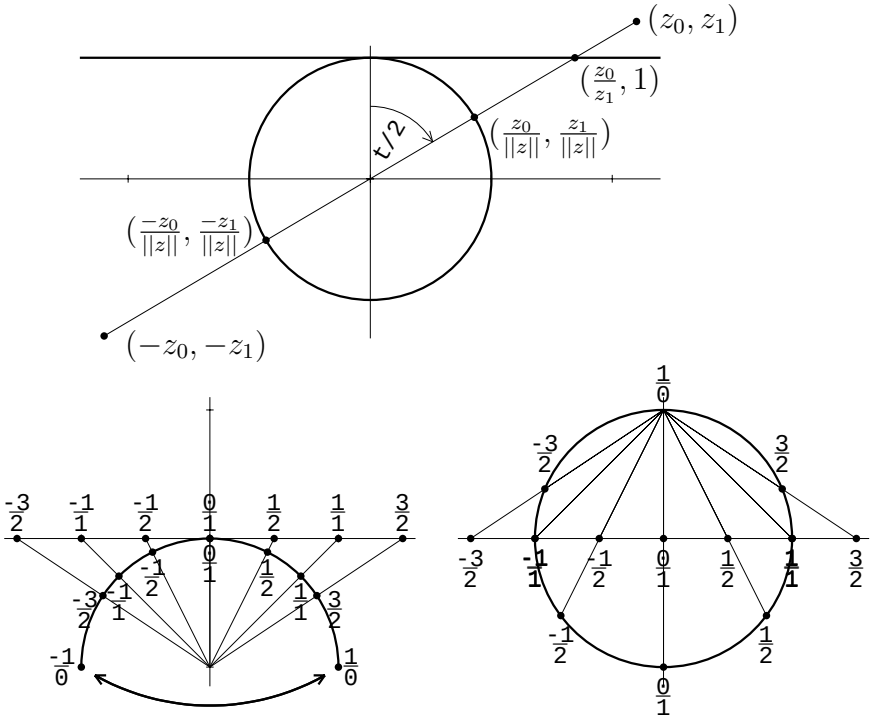


Figure 3.2: The homogeneous coordinates (top). The stereographic projection (bottom) doubles the angles and reverses the orientation.

upper semicircle twice and glue its two endpoints, we get the full unit circle. This stretching and gluing operation is realized by the **stereographic projection** introduced in Section 1.3. We look now into geometrical properties of this transformation. The transformation takes a point at the upper semicircle, projects it to the real line as in Figure 3.2 bottom left and then to the unit circle as in Figure 3.2 bottom right. In this way we get a projection which doubles the angles, reverses the orientation and maps the upper semicircle to the full circle.

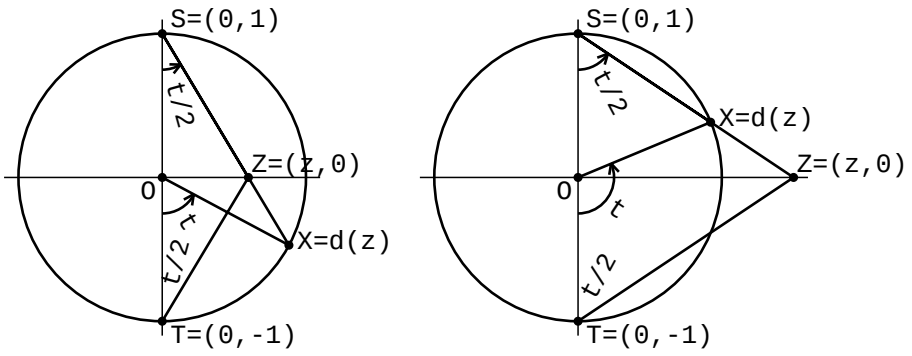


Figure 3.3: The stereographic projection in the Cartesian plane

We consider the stereographic projection in the plane with the cartesian coordinates (x_0, x_1). The real line is now identified with the x_0-axis with equation $x_1 = 0$. A point $Z = (z, 0)$ on the x_0-axis is projected to the intersection $\mathtt{d}(z) = X = (x_0, x_1)$ of the unit circle with the line SZ. Here $S = (0, 1)$ is the north pole (see Figure 3.3). If $t \in [-\pi, \pi]$ is the angle $\angle TOX$, then $X = (\cos(t - \frac{\pi}{2}), \sin(t - \frac{\pi}{2})) = (\sin t, -\cos t)$. The triangle OSX is equilateral with angle $\pi - t$ at O and angles $\frac{t}{2}$ at S and X. The triangle STZ is also equilateral with angles $\frac{t}{2}$ at S and T. In fact, $\angle OTZ = \frac{t}{2} \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ is just the angle of the homogeneous coordinate $\frac{z_0}{z_1}$ with the

z_1-axis (see Figure 3.2 top), so $\tan\frac{t}{2} = z$, $\sin\frac{t}{2} = \frac{z}{\sqrt{z^2+1}}$, $\cos\frac{t}{2} = \frac{1}{\sqrt{z^2+1}}$. It follows $\sin t = \frac{2z}{z^2+1}$, $\cos t = \frac{1-z^2}{z^2+1}$, so $\mathbf{d}(z) = (\sin t, -\cos t) = (\frac{2z}{z^2+1}, \frac{z^2-1}{z^2+1})$, and $\mathbf{d}(\infty) = (0,1)$. In homogeneous coordinates we get

$$\mathbf{d}\begin{pmatrix} z_0 \\ z_1 \end{pmatrix} = \left(\frac{2z_0z_1}{z_0^2+z_1^2}, \frac{z_0^2-z_1^2}{z_0^2+z_1^2}\right).$$

From the similarity of triangles we get $z : 1 = x : (1-y)$, so the inverse stereographic projection is given by $\mathbf{d}^{-1}(x,y) = \frac{x}{1-y}$. The parametrization of the unit circle by the variable t is the map $t \mapsto e^{i(t-\frac{\pi}{2})} = (\sin t, -\cos t)$. This yields the parametrization $\mathbf{t} : \mathbb{R} \to \overline{\mathbb{R}}$ of $\overline{\mathbb{R}}$ given by

$$\mathbf{t}(t) = \mathbf{d}^{-1}(\sin t, -\cos t) = \frac{\sin t}{1+\cos t} = \tan\frac{t}{2} = \frac{\sin\frac{t}{2}}{\cos\frac{t}{2}}$$

Here $\frac{\sin\frac{t}{2}}{\cos\frac{t}{2}}$ should be regarded as a homogeneous coordinate of $\mathbf{t}(t)$. The projection $\mathbf{t}$ is bijective on every semiclosed interval $[t, t+2\pi)$. In particular, $\mathbf{t}$ has the inverse $\mathbf{t}^{-1}(x) = 2\arctan x$ on the semiclosed interval $[-\pi, \pi)$.

An **interval** is a connected subset of $\overline{\mathbb{R}}$. We say that $I \subseteq \overline{\mathbb{R}}$ is a **proper interval**, if it has two distinct endpoints $a, b \in \overline{I} \setminus I^\circ$. If $a, b \in I$ then I is closed and if $a, b \in \overline{\mathbb{R}} \setminus I$ then I is open. **Improper intervals** are the empty set, singletons, their complements and the full interval $\overline{\mathbb{R}}$. Given two distinct points $a, b \in \overline{\mathbb{R}}$, there exist two proper open intervals I, J with endpoints $a, b \in \overline{\mathbb{R}}$ which satisfy $I \cap J = \emptyset$, $\overline{I} \cup \overline{J} = \overline{\mathbb{R}}$. We distinguish these intervals by the order of a, b and write them conventionally as $I = (a,b)$, $J = (b,a)$. A point $x \in \overline{\mathbb{R}}$ belongs to (a,b), if the triple a, x, b is **positively oriented**, i.e., if $\det(a,x) \cdot \det(x,b) \cdot \det(b,a) > 0$. This means that $\mathbf{d}(x)$ belongs to the counterclockwise arc from $\mathbf{d}(a)$ to $\mathbf{d}(b)$.

Definition 3.2 *The* **open interval** *and the* **closed interval** *with distinct endpoints $a, b \in \overline{\mathbb{R}}$ are*

$$\begin{aligned} (a,b) &= \{x \in \overline{\mathbb{R}} : \det(a,x) \cdot \det(x,b) \cdot \det(b,a) > 0\}, \\ [a,b] &= \{x \in \overline{\mathbb{R}} : \det(a,x) \cdot \det(x,b) \cdot \det(b,a) \geq 0\}. \end{aligned}$$

The **size** *of an interval $I = [a,b]$ or $I = (a,b)$ is defined by*

$$\operatorname{sz}(I) = \frac{a \cdot b}{\det(b,a)} = \frac{a_0b_0 + a_1b_1}{a_1b_0 - a_0b_1}.$$

Note that the property $x \in (a,b)$ does not depend on the representation of a, x, b by homogeneous coordinates. For example we have $\frac{1}{1} \in (\frac{0}{1}, \frac{1}{0})$, $\frac{1}{1} \in (\frac{0}{1}, \frac{-1}{0})$, or $\frac{-1}{-1} \in (\frac{0}{1}, \frac{1}{0})$. Definition 3.2 is compatible with the usage of Section 1.3. We have $\frac{x}{1} \in [\frac{a}{1}, \frac{b}{1}]$ iff $(a-x)(x-b)(b-a) \geq 0$. If $a < b$, this is equivalent to $a \leq x \leq b$. If $b < a$, this is equivalent to $a \leq x$ or $x \leq b$. The length of an interval $I = [a,b]$ defined in Section 1.3 can be written in homogeneous coordinates as

$$|I| = \frac{1}{\pi}\operatorname{arccotg}\frac{a_0b_0 + a_1b_1}{a_1b_0 - a_0b_1} = \frac{1}{\pi}\operatorname{arccotg}\operatorname{sz}(I).$$

The length of small intervals can be estimated by their size.

Proposition 3.3 *The length of an interval $I \subseteq \overline{\mathbb{R}}$ is $|I| = \frac{1}{2} - \frac{1}{\pi}\arctan\operatorname{sz}(I)$. We have $\operatorname{sz}(I) \geq 1$ iff $|I| \leq \frac{1}{4}$ and in this case*

$$\frac{1}{4 \cdot \operatorname{sz}(I)} \leq |I| \leq \frac{1}{\pi \cdot \operatorname{sz}(I)}.$$

Proof: We have $\operatorname{arccotg} x = \frac{\pi}{2} - \arctan x$ for each $x \in \mathbb{R}$, so $|I| = \frac{1}{2} - \frac{1}{\pi}\arctan \mathrm{sz}(I)$. For $0 \le y \le 1$ we have $0 \le \arctan(y) \le \frac{\pi}{4}$, $\arctan'(y) = \frac{1}{y^2+1} \le 1$, so $\frac{\pi y}{4} \le \arctan y \le y$ and therefore $\frac{y}{4} \le \frac{1}{\pi}\arctan y \le \frac{y}{\pi}$. For $x = \frac{1}{y}$ we have $x \ge 1$ and $\arctan y = \operatorname{arccotg} x$, so $\frac{1}{4x} \le \frac{1}{\pi}\operatorname{arccotg} x \le \frac{1}{\pi x}$. This implies the estimate for $|I|$. □

Alternatively we obtain the length of an interval from the parametrization $\mathbf{t} : \mathbb{R} \to \overline{\mathbb{R}}$ of $\overline{\mathbb{R}}$.

Proposition 3.4 *If $I = [a, b] \subseteq \overline{\mathbb{R}}$ is a proper interval, $a = \mathbf{t}(t)$, $b = \mathbf{t}(s)$ and $0 < s - t < 2\pi$, then $|I| = \frac{s-t}{2\pi}$.*

The proof is a simple verification.

3.3 Projective metrics

The angle $0 \le \varphi(x, y) \le \pi$ between two nonzero vectors $x, y \in \mathbb{R}^{n+1}$ can be obtained by the cosine rule as $\varphi(x, y) = \arccos \frac{x \cdot y}{||x|| \cdot ||y||}$, where $x \cdot y = \sum_i x_i y_i$ is the scalar product and $||x|| = \sqrt{x \cdot x}$ is the Euclidean norm. The angle between $-x$ and y is $\pi - \varphi(x, y) = \arccos \frac{-x \cdot y}{||x|| \cdot ||y||}$. Taking the smaller of these two angles we define the **angle metric** in the projective space $\mathbb{P}(\mathbb{R}^{n+1})$ by

$$d_a(x, y) = \frac{1}{\pi} \min\{\varphi(x, y), \pi - \varphi(x, y)\} = \frac{1}{\pi} \arccos \frac{|x \cdot y|}{||x|| \cdot ||y||} \in [0, \tfrac{1}{2}]$$

The formula does not depend on the choice of representing vectors: $d_a(\lambda x, \mu y) = d_a(x, y)$ for every nonzero λ, μ. In $\mathbb{P}(\mathbb{R}^2) = \overline{\mathbb{R}}$ we use the formula $\arccos x = \operatorname{arccotg} \frac{x}{\sqrt{1-x^2}}$ to get

$$\begin{aligned} d_a(x, y) &= \frac{1}{\pi} \arccos \frac{|xy + 1|}{\sqrt{(x^2 + 1)(y^2 + 1)}} = \frac{1}{\pi} \operatorname{arccotg} \frac{|xy + 1|}{|x - y|} \\ d_a(x, \infty) &= \frac{1}{\pi} \operatorname{arccotg} |x| \end{aligned}$$

Alternatively, we consider the **projective metric** which is based on the approximation $\varphi \approx 2 \sin \frac{\varphi}{2}$. It is the distance of the normalized homogeneous coordinate $x/||x||$ from $y/||y||$ or from $-y/||y||$:

$$\begin{aligned} d_p(x, y) &= \min\left\{2 \sin \frac{\varphi(x, y)}{2}, 2 \sin \frac{\pi - \varphi(x, y)}{2}\right\} \\ &= \min\{\sqrt{2(1 - \cos \varphi(x, y))}, \sqrt{2(1 + \cos \varphi(x, y))}\} \\ &= \min\left\{\sqrt{2\left(1 - \frac{x \cdot y}{||x|| \cdot ||y||}\right)}, \sqrt{2\left(1 + \frac{x \cdot y}{||x|| \cdot ||y||}\right)}\right\} \\ &= \min\left\{\left|\left|\frac{x}{||x||} - \frac{y}{||y||}\right|\right|, \left|\left|\frac{x}{||x||} + \frac{y}{||y||}\right|\right|\right\} \in [0, \sqrt{2}], \end{aligned}$$

The last equality follows from

$$\begin{aligned} ||x \cdot ||y|| \pm ||x|| \cdot y||^2 &= 2 \cdot ||x||^2 \cdot ||y||^2 \pm 2 \cdot ||x|| \cdot ||y|| \cdot (x \cdot y) \\ &= 2 \cdot ||x|| \cdot ||y|| \cdot (||x|| \cdot ||y|| \pm (x \cdot y)) . \end{aligned}$$

A simpler metric is obtained from the approximation $\varphi \approx \sin\varphi$. Since $\sin\varphi = \sin(\pi - \varphi)$, we define the **chord metric** by

$$d_c(x,y) = \sin\varphi(x,y) = \frac{\sqrt{||x||^2 \cdot ||y||^2 - (x \cdot y)^2}}{||x|| \cdot ||y||} \in [0,1]$$

In homogeneous coordinates in $\overline{\mathbb{R}}$, we get

$$d_c(x,y) = \frac{|\det(x,y)|}{||x|| \cdot ||y||} = \frac{1}{2}||\mathbf{d}(x) - \mathbf{d}(y)||.$$

This follows from $\sin\varphi(x,y) = \sin\frac{\varphi(\mathbf{d}(x),\mathbf{d}(y))}{2} = \frac{1}{2}||\mathbf{d}(x) - \mathbf{d}(y)||$. For $x, y \in \mathbb{R}$ we get

$$\begin{aligned} d_c(x,y) &= \frac{|x-y|}{\sqrt{(x^2+1)(y^2+1)}}, \\ d_c(x,\infty) &= \frac{1}{\sqrt{x^2+1}}. \end{aligned}$$

Proposition 3.5 *The three projective metrics are equivalent. We have* $d_c(x,y) \le d_p(x,y) \le \pi d_a(x,y) \le \frac{\pi}{2} d_c(x,y)$ *and*

$$\lim_{y\to x} \frac{\pi d_a(x,y)}{d_p(x,y)} = \lim_{y\to x} \frac{d_p(x,y)}{d_c(x,y)} = 1.$$

Proof: For $0 \le \alpha \le \frac{\pi}{2}$ we have $\sin\alpha \le 2\sin\frac{\alpha}{2} \le \alpha \le \frac{\pi}{2} \cdot \sin\alpha$ and $\lim_{\alpha\to 0} \frac{2\sin\frac{\alpha}{2}}{\alpha} = \lim_{\alpha\to 0} \frac{\sin\alpha}{\alpha} = 1$. □

3.4 Transformations

A linear transformation of the vector space $\mathbb{R}^2$ is determined by a (2×2)-matrix $M = \left[\begin{smallmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{smallmatrix}\right]$. The M-image of a column vector $x \in \mathbb{R}^2$ is $Mx \in \mathbb{R}^2$ defined by $(Mx)_i = \sum_{j=0}^{1} M_{ij}x_j$:

$$Mx = \begin{bmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} M_{00}x_0 + M_{01}x_1 \\ M_{10}x_0 + M_{11}x_1 \end{bmatrix}$$

As a vector space, the space $\mathbb{R}^{2\times 2}$ of (2×2)-matrices is isomorphic to $\mathbb{R}^4$, but $\mathbb{R}^{2\times 2}$ has an additional structure of matrix multiplication $(MP)_{ik} = \sum_{j=0}^{1} M_{ij} \cdot P_{jk}$. If $\det(M) = M_{00}M_{11} - M_{01}M_{10} \neq 0$, then $M : \mathbb{R}^2 \to \mathbb{R}^2$ is bijective and the M-image of a one-dimensional subspace of $\mathbb{R}^2$ is a one-dimensional subspace of $\mathbb{R}^2$. This means that M determines a transformation of the projective space $\mathbb{P}(\mathbb{R}^2) = \overline{\mathbb{R}}$ which is called a **Möbius transformation**. A nonzero multiple λM of M determines the same transformation as M, so a Möbius transformation is determined by a **projective matrix**, i.e., by a one-dimensional subspace of $\mathbb{R}^{2\times 2}$ which is a point of the projective space $\mathbb{P}(\mathbb{R}^{2\times 2})$. We do not distinguish between a projective matrix and its transformation. The determinant of a projective matrix is not a well-defined concept, since $\det(\lambda M) = \lambda^2 \det(M)$. However, the sign of the determinant does not depend on λ so we can classify transformations according to the sign of their determinant:

$$\begin{aligned} \mathbb{M}(\mathbb{R}) &= \{M \in \mathbb{P}(\mathbb{R}^{2\times 2}) :\ \det(M) \neq 0\} :\ \textbf{regular transformations} \\ \mathbb{M}^+(\mathbb{R}) &= \{M \in \mathbb{M}(\mathbb{R}) :\ \det(M) > 0\} :\ \textbf{increasing transformations} \\ \mathbb{M}^-(\mathbb{R}) &= \{M \in \mathbb{M}(\mathbb{R}) :\ \det(M) < 0\} :\ \textbf{decreasing transformations} \end{aligned}$$

If $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$, then $M : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is bijective and has an inverse $M^{-1} = \left[\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right] = \left[\begin{smallmatrix} -d & b \\ c & -a \end{smallmatrix}\right]$. The composition of two transformations is again a transformation whose matrix is obtained by matrix multiplication. Regular Möbius transformations thus form a group. If $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$, $x \in \overline{\mathbb{R}}$, then $M(x) = \frac{ax_0+bx_1}{cx_0+dx_1}$, in particular $M(\frac{-d}{c}) = \infty$, $M(\infty) = \frac{a}{c}$. A transformation can be lifted by the parametrization $\mathbf{t} : \mathbb{R} \to \overline{\mathbb{R}}$ to a continuous function $\widetilde{M} : \mathbb{R} \to \mathbb{R}$ which commutes in the diagram $M \circ \mathbf{t} = \mathbf{t} \circ \widetilde{M}$:

$$\begin{array}{ccc} \mathbb{R} & \xrightarrow{\widetilde{M}} & \mathbb{R} \\ {\scriptstyle \mathbf{t}}\downarrow & & \downarrow{\scriptstyle \mathbf{t}} \\ \overline{\mathbb{R}} & \xrightarrow{M} & \overline{\mathbb{R}} \end{array}$$

If $M \in \mathbb{M}^+(\mathbb{R})$ is increasing then $\widetilde{M}(t + 2\pi) = \widetilde{M}(t) + 2\pi$. If $M \in \mathbb{M}^-(\mathbb{R})$ is decreasing then $\widetilde{M}(t + 2\pi) = \widetilde{M}(t) - 2\pi$. The graphs of some lifts $\widetilde{M}$ can be seen in Figures 3.4 and 3.5.

The derivation of M in $x \in \mathbb{R}$ is readily computed as $M'(x) = (ad - bc)/(cx + d)^2$. If $|M'(x)| < 1$, then, in a neighbourhood of x, M is contracting with respect to the Euclidean metric $d_e(x, y) = |x - y|$. If we work in $\overline{\mathbb{R}}$, we are rather interested in the derivation of M with respect to the projective metrics.

Definition 3.6 *The* **circle derivation** *of $M \in \mathbb{M}(\mathbb{R})$ in $x \in \overline{\mathbb{R}}$ is defined by*

$$M^\bullet(x) = \frac{\det(M) \cdot ||x||^2}{||Mx||^2}$$

Note that while the norm $||x||$ depends on a particular homogeneous representation of x, the ratio $||x||/||M(x)||$ does not. For $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, $x \in \mathbb{R}$ we get

$$\begin{aligned} M^\bullet(x) &= \frac{(ad - bc)(x^2 + 1)}{(ax + b)^2 + (cx + d)^2}, \\ M^\bullet(\infty) &= \frac{ad - bc}{a^2 + c^2} \end{aligned}$$

Proposition 3.7 *If $M \in \mathbb{M}(\mathbb{R})$ is a transformation then*

$$\begin{aligned} |M^\bullet(x)| &= \lim_{y \to x} \frac{d_c(M(y), M(x))}{d_c(y, x)} \\ M^\bullet(x) &= \widetilde{M}'(\mathbf{t}^{-1}(x)) \end{aligned}$$

Proof: From $\det(M(y), M(x)) = \det(M) \cdot \det(x, y)$ we get

$$\begin{aligned} \lim_{y \to x} \frac{d_c(M(y), M(x))}{d_c(y, x)} &= \lim_{y \to x} \frac{|\det(M(y), M(x))|}{|\det(y, x)|} \cdot \frac{||y|| \cdot ||x||}{||M(y)|| \cdot ||M(x)||} \\ &= \frac{|\det(M)| \cdot ||x||^2}{||M(x)||^2} = |M^\bullet(x)| \end{aligned}$$

For $\mathbf{t}^{-1}(x) = 2\arctan x$ we have $(\mathbf{t}^{-1})'(x) = \frac{2}{x^2+1}$ and $\mathbf{t}'(\mathbf{t}^{-1}(x)) = (1/\mathbf{t}^{-1})'(x)$. From $\widetilde{M} = \mathbf{t}^{-1} \circ M \circ \mathbf{t}$ we get

$$\begin{aligned}\widetilde{M}'(\mathbf{t}^{-1}(x)) &= \frac{(\mathbf{t}^{-1})'(M(x)) \cdot M'(x)}{(\mathbf{t}^{-1})'(x)} \\ &= \frac{2}{M^2(x)+1} \cdot \frac{\det(M)}{(cx+d)^2} \cdot \frac{x^2+1}{2} \\ &= \frac{\det(M)(x^2+1)}{(ax+b)^2+(cx+d)^2} = M^\bullet(x)\end{aligned}$$

□

Using $\det(MP) = \det(M) \cdot \det(P)$, we immediately get the chain rule:

$$\begin{aligned}(MP)^\bullet(x) &= \frac{\det(M) \cdot ||Px||^2}{||MPx||^2} \cdot \frac{\det(P) \cdot ||x||^2}{||Px||^2} \\ &= M^\bullet(Px) \cdot P^\bullet(x)\end{aligned}$$

Proposition 3.8 *If $M \in \mathbb{M}(\mathbb{R})$, $I \subset \overline{\mathbb{R}}$ is an interval and $q_0 \le M^\bullet(x) \le q_1$ for every $x \in I$, then $q_0|I| \le |M(I)| \le q_1|I|$.*

Proof: We use Proposition 3.4. Let $I = [a, b]$, $a = \mathbf{t}(t)$, $b = \mathbf{t}(s)$ and $0 < s - t < 2\pi$. By the mean value theorem, there exists $t \le x \le s$ such that $\widetilde{M}'(x) = \frac{\widetilde{M}(s)-\widetilde{M}(t)}{s-t} = \frac{|M(I)|}{|I|}$. If $q_0|I| > |M(I)|$, or $q_1|I| < |M(I)|$ then $|\widetilde{M}'(x)| < q_0$ or $|\widetilde{M}'(x)| > q_1$ which is a contradiction. □

Definition 3.9 *The* **expanding interval** *and the* **contracting interval** *of a transformation $M \in \mathbb{M}(\mathbb{R})$ are defined by*

$$\begin{aligned}\mathbf{U}(M) &= \{x \in \overline{\mathbb{R}} : \ |M^\bullet(x)| < 1\}, \\ \mathbf{V}(M) &= \{x \in \overline{\mathbb{R}} : \ |(M^{-1})^\bullet(x)| > 1\}.\end{aligned}$$

The **trace of a matrix** $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is $\mathrm{tr}(M) = a + d$. Define the **trace of a projective matrix** $M \in \mathbb{M}(\mathbb{R})$ by

$$\mathrm{trc}(M) = \frac{\mathrm{tr}(M)^2}{\det(M)} = \frac{(a+d)^2}{ad-bc}$$

If M is decreasing then $\mathrm{trc}(M) \le 0$, otherwise $\mathrm{trc}(M) \ge 0$. Increasing transformations are classified into three kinds according to the number of their fixed points. We say that $x \in \overline{\mathbb{R}}$ is a **fixed point** of M, if $M(x) = x$. Every $x \in \overline{\mathbb{R}}$ is a fixed point of the identity $\mathrm{Id} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$.

Proposition 3.10 *A decreasing transformation $M \in \mathbb{M}^-(\mathbb{R})$ has two fixed points. If $M \in \mathbb{M}^+(\mathbb{R})$ is a nonidentical increasing transformation, there are three cases:*

1. *If $\mathrm{trc}(M) < 4$, then M has no fixed point. We say that M is* **elliptic**.
2. *If $\mathrm{trc}(M) = 4$, then M has one fixed point. We say that M is* **parabolic**.
3. *If $\mathrm{trc}(M) > 4$, then M has two fixed points. We say that M is* **hyperbolic**.

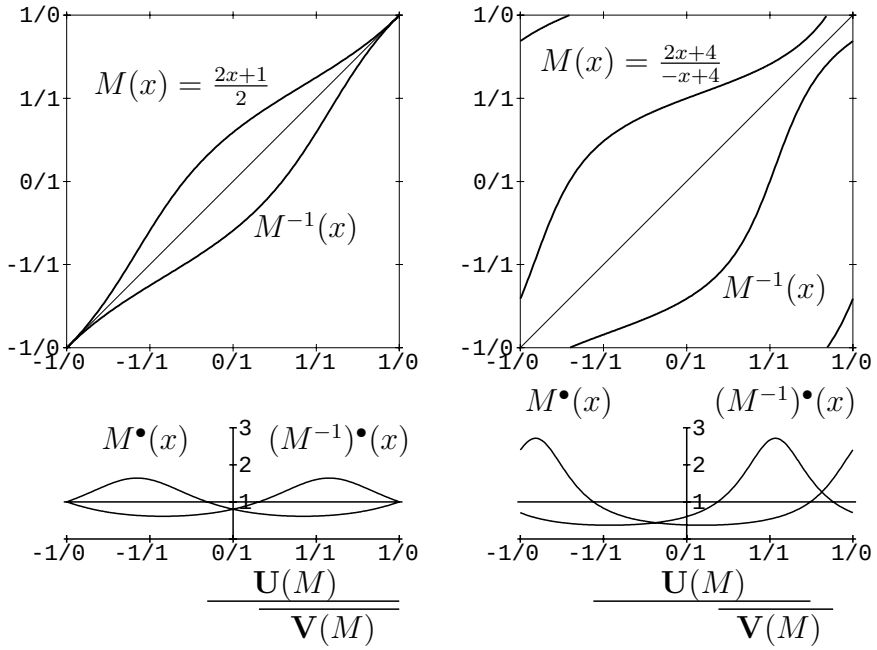


Figure 3.4: Möbius transformations and their circle derivations: $M(x) = \frac{2x+1}{2}$ is parabolic with $\mathbf{U}(M) = (\frac{-1}{4}, \frac{1}{0})$, $\mathbf{V}(M) = (\frac{1}{4}, \frac{1}{0})$ (left), $M(x) = \frac{2x+4}{-x+4}$ is elliptic (right).

Proof: If $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is not the identity, then x is a fixed point of M iff $cx_0^2+(d-a)x_0x_1-bx_1^2 = 0$. If $c \neq 0$, this is a quadratic equation with discriminant

$$D = (a-d)^2 + 4bc = \mathrm{tr}(M)^2 - 4\det(M),$$

so $D \geq 0$ iff either $\det(M) < 0$ (and then $\mathrm{trc}(M) \leq 0$) or $\det(M) > 0$ and $\mathrm{trc}(M) \geq 4$. If $c = 0$ then we have one solution $x = \frac{1}{0}$ and the other $x = \frac{b}{d-a}$. If $d \neq a$ then M has two fixed points and either $\det(M) < 0$ or $\det(M) > 0$ and $\mathrm{trc}(M) = \frac{(a+d)^2}{ad} > 4$. If $d = a$, $b \neq 0$, then M has a unique fixed point ∞ and $\mathrm{trc}(M) = \frac{4a^2}{a^2} = 4$. If $d = a$, $b = 0$, then M is the identical transformation. □

Some graphs of transformations and their circle derivations can be seen in Figures 3.4 and 3.5. The extended real line is displayed in the arc metric as a finite interval from $\frac{-1}{0}$ to $\frac{1}{0}$. In other words, we use the function $\mathbf{t}(x) = \tan\frac{x}{2}$ which maps $\mathbb{R}$ bijectively to $(-\pi, \pi)$ and the graphs show the real functions $\widetilde{M} = \mathbf{t}^{-1} \circ M \circ \mathbf{t} : (-\pi, \pi) \to (-\pi, \pi)$. The fixed points are the intersections of the graphs with the diagonal $y = x$.

3.5 Conjugated transformations

Definition 3.11 *We say that transformations $P, Q \in \mathbb{M}(\mathbb{R})$ are* **conjugated** *if there exists $M \in \mathbb{M}(\mathbb{R})$ such that $Q = M^{-1}PM$.*

Conjugated transformations have the same dynamical properties. If $Q = M^{-1}PM$, then $Q^n = M^{-1}P^nM$ for any $n \in \mathbb{Z}$. If x is a fixed point of P, then $y = M^{-1}x$ is a fixed point of Q and

$$\begin{aligned} Q^{\bullet}(y) &= (M^{-1})^{\bullet}(PM(y)) \cdot P^{\bullet}(M(y)) \cdot M^{\bullet}(y) = (M^{-1})^{\bullet}(M(y)) \cdot P^{\bullet}(x) \cdot M^{\bullet}(y) \\ &= P^{\bullet}(x). \end{aligned}$$

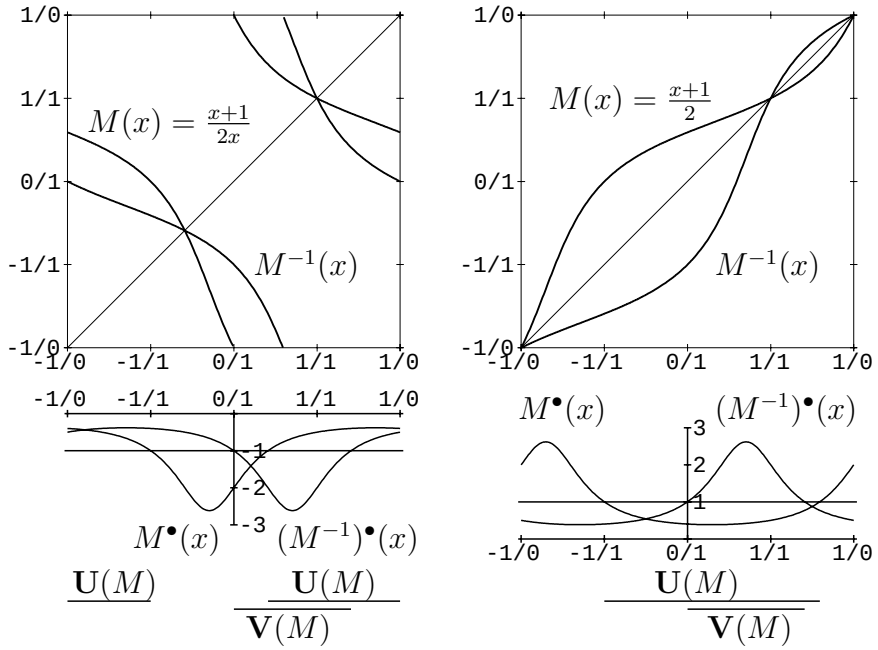


Figure 3.5: Möbius transformations and their circle derivations: $M(x) = \frac{x+1}{2x}$ is decreasing with $\mathbf{U}(M) = (\frac{1}{3}, \frac{1}{-1})$, $\mathbf{V}(M) = (0, 2)$ (left), $M(x) = \frac{x+1}{2}$ is hyperbolic with $\mathbf{U}(M) = (-1, 3)$, $\mathbf{V}(M) = (0, 2)$.

Conjugated transformations have the same trace. A direct computation shows that $\mathrm{tr}(PQ) = \sum_{i,j} P_{ij}Q_{ji} = \mathrm{tr}(QP)$. If $Q = M^{-1}PM$, then $\mathrm{tr}(Q) = \mathrm{tr}(PMM^{-1}) = \mathrm{tr}(P)$. Since $\det(Q) = \det(P)$, we get $\mathrm{trc}(Q) = \mathrm{trc}(P)$. We are going to show that two transformations of the same orientation (increasing or decreasing) are conjugated iff they have the same trace by showing that each transformation is conjugated to a canonical form which is either a **similarity**, **translation** or **rotation**.

Definition 3.12 *A* **similarity** *is a transformation $Q_r(x) = rx$, where $0 \neq r \neq 1$.*

Thus $Q_r = [\frac{r}{0}, \frac{0}{1}]$, $\det(Q_r) = r$ and $\mathrm{trc}(Q_r) = (r+1)^2/r$. The fixed points are 0 and ∞ with circle derivations $Q_r^\bullet(0) = r$, $Q_r^\bullet(\infty) = \frac{1}{r}$. The composition of similarities is again a similarity: $Q_{rt} = Q_r \circ Q_t$. If $r < 0$ then Q_r is decreasing, in particular $Q_{-1}(x) = -x$. If $0 < r \neq 1$ then Q_r is hyperbolic.

Proposition 3.13 *A decreasing transformation $M \in \mathbb{M}^-(\mathbb{R})$ is conjugated to a similarity with quotient $-1 \leq r < 0$. A hyperbolic transformation $M \in \mathbb{M}^+(\mathbb{R})$ is conjugated to a similarity with quotient $0 < r < 1$. If $0 < |r| < 1$, then M has un unstable fixed point $\mathbf{u}(M)$ and a stable fixed point $\mathbf{s}(M)$ such that $\lim_{n\to\infty} M^n(x) = \mathbf{s}(M)$ for each $x \neq \mathbf{u}(M)$. Moreover, $M^\bullet(\mathbf{u}(M)) > 1$, $M^\bullet(\mathbf{s}(M)) < 1$ and $M^\bullet(\mathbf{u}(M)) \cdot M^\bullet(\mathbf{s}(M)) = 1$. If $r = -1$ then $M^2 = \mathrm{Id}$.*

Proof: Let $a, b \in \overline{\mathbb{R}}$ be the two fixed points of M and set $P = [a, b] = [\frac{a_0}{a_1}, \frac{b_0}{b_1}]$. Then $P(0) = b$ and $P(\infty) = a$, so $P^{-1}MP$ has fixed points 0 and ∞. It follows $P^{-1}MP = Q_r$ with $0 \neq r \neq 1$. From $M = PQ_rP^{-1}$ we get $M^\bullet(b) = Q_r^\bullet(0) = r$, $M^\bullet(a) = Q_r^\bullet(\infty) = 1/r$, so $M^\bullet(a) \cdot M^\bullet(b) = 1$. If $|r| < 1$ then $M^\bullet(b) < 1$ and we have $\mathbf{s}(M) = b$, $\mathbf{u}(M) = a$. Since $\lim_{n\to\infty} Q_{r^n}(x) = 0$ for every $x \neq \infty$, we get $\lim_{n\to\infty} M^n(x) = \mathbf{s}(M)$ for every $x \neq \mathbf{u}(M)$. If $|r| > 1$ then $M^\bullet(b) > 1$ and $\mathbf{s}(M) = a$, $\mathbf{u}(M) = b$. We get again $\lim_{n\to\infty} M^n(x) = \mathbf{s}(M)$ for every $x \neq \mathbf{u}(M)$. A similarity Q_r with $|r| > 1$ is conjugated to $Q_{1/r}$, since for $P(x) = -1/x$ we have $P^{-1}Q_rP = Q_{1/r}$. Thus M is conjugated to a similarity Q_r with $-1 \leq r < 1$. If $r = -1$ then $\mathrm{tr}(M) = 0$ so $M = [\begin{smallmatrix} a & b \\ c & -a \end{smallmatrix}]$

and M^2 is the identical transformation. □

Definition 3.14 *The* **translation** *and* **rotation** *with parameter* $t \in \mathbb{R} \setminus \{0\}$ *are transformations with matrices*

$$T^t = \left[\frac{1}{0}, \frac{t}{1}\right], R^t = \left[\frac{\cos\frac{t}{2}}{-\sin\frac{t}{2}}, \frac{\sin\frac{t}{2}}{\cos\frac{t}{2}}\right]$$

We have $T^{t+s} = T^t \circ T^s$, $R^{t+s} = R^t \circ R^s$, and $T^0 = R^0 = \mathrm{Id}$. Moreover, $R^{t+2\pi} = R^t$ (as transformations, not as matrices). For $x = \frac{r\sin\frac{s}{2}}{r\cos\frac{s}{2}}$ we have $R^t x = \frac{r\sin\frac{s+t}{2}}{r\cos\frac{s+t}{2}}$, $\mathbf{d}(x) = (\sin s, -\cos s)$, $\mathbf{d}R^t(x) = R^{-2t}\mathbf{d}(x)$. A translation is parabolic and has a unique fixed point ∞ with circle derivation $(T^t)^\bullet(\infty) = 1$. A rotation is elliptic and has no fixed point and the unit circle derivation everywhere: $(R^t)^\bullet(x) = 1$. It follows that its contraction and expansion intervals are empty $\mathbf{U}(R^t) = \mathbf{V}(R^t) = \emptyset$. A parabolic transformation is a translation iff its fixed point is ∞. An elliptic transformation has no real fixed point but it has two complex fixed points. It is a rotation iff its fixed points are i and $-i$.

Proposition 3.15 *A parabolic transformation* M *is conjugated to the translation* $T^1(x) = x+1$. M *has a unique fixed point* $\mathbf{s}(M)$ *such that* $\lim_{n\to\infty} M^n(x) = \mathbf{s}(M)$ *for each* $x \in \overline{\mathbb{R}}$, *and* $M^\bullet(\mathbf{s}(M)) = 1$.

Proof: Let $s = \mathbf{s}(M)$ be the unique fixed point of M. We take a transformation P with the first column s and positive determinant. Then $P(\infty) = s$, and $P^{-1}MP$ is a parabolic transformation with fixed point ∞, so $P^{-1}MP = T^r$ for some $r \neq 0$. From $x + r = r(\frac{x}{r} + 1)$ we get $T^r = Q_r T^1 Q_r^{-1}$, so T^r is conjugated to T^1. □

Proposition 3.16 *An elliptic transformation is conjugated to a rotation* R^t *with* $0 < t \leq \pi$.

Proof: If $M = [\frac{a}{c}, \frac{b}{d}]$ is elliptic, then $c \neq 0$ and we can assume $c > 0$. The transformation has no fixed point in $\overline{\mathbb{R}}$ but it has two complex fixed points $s = \frac{a-d+i\sqrt{-D}}{2c}$, $u = \frac{a-d-i\sqrt{-D}}{2c}$, where $D = (a+d)^2 - 4(ad-bc)$. The transformation $P = [\frac{\sqrt{-D}}{0}, \frac{a-d}{2c}]$ satisfies $\det(P) > 0$, $P(i) = s$, $P(-i) = u$, so $P^{-1}MP$ has fixed points i and $-i$. It follows that it is a rotation with $0 < t < 2\pi$. Since R^t is conjugated to R^{-t} via $Q(x) = -x$, M is conjugated to a rotation with angle $0 < t \leq \pi$. □

Theorem 3.17 *Two transformations from* $\mathbb{M}^+(\mathbb{R})$ *are conjugated iff they have the same trace. Two transformations from* $\mathbb{M}^-(\mathbb{R})$ *are conjugated iff they have the same trace.*

Proof: We have $\mathrm{trc}(Q_r) = \frac{(r+1)^2}{r}$. If $0 < |r|, |s| < 1$ and $r \neq s$, then $\mathrm{trc}(Q_r) \neq \mathrm{trc}(Q_s)$, so Q_r is not conjugated to Q_s. We have $\mathrm{trc}(R^t) = 4\cos^2\frac{t}{2}$. If $0 < t < s \leq \pi$ then $\mathrm{trc}(R^t) \neq \mathrm{trc}(R^s)$, so R^t and R^s are not conjugated. □

A similarity can be written as $Q_r = S^t = [\frac{e^{t/2}}{0}, \frac{0}{e^{-t/2}}]$, where $t = \ln r$. Then $S^{t+s} = S^t \circ S^s$ and $\mathrm{trc}(S^t) = 2\cosh\frac{t}{2}$. The transformation S^t is conjugated to $[\frac{\cosh\frac{t}{2}}{\sinh\frac{t}{2}}, \frac{\sinh\frac{t}{2}}{\cosh\frac{t}{2}}]$ with fixed points -1,1 and the same trace $2\cosh\frac{t}{2}$. These formulas reveal a formal analogy of hyperbolic and elliptic transformations.

Definition 3.18 *The* **similarity quotient** $\mathrm{sim}(M) > 0$ *of a hyperbolic transformation and the* **rotation angle** $\mathrm{rot}(M) \in (0, \pi]$ *of an elliptic transformation are defined by*

$$\mathrm{sim}(M) = 2\,\mathrm{argcosh}\,\frac{\mathrm{trc}(M)}{2},\ \mathrm{rot}(M) = 2\arccos\frac{\mathrm{trc}(M)}{2}.$$

Thus $\mathrm{sim}(S^t) = t$, $\mathrm{rot}(R^t) = t$.

Proposition 3.19 *For every increasing transformation $M \in \mathbb{M}^+(\mathbb{R})$ there exists a system of transformations $(M^t)_{t\in\mathbb{R}}$ such that M^0 is the identity, $M^1 = M$ and $M^{t+s} = M^t \circ M^t$ for every $t, s \in \mathbb{R}$.*

Proof: If $M = P^{-1}S^rP$ is hyperbolic, then $M^t = P^{-1}S^{rt}P$. If $M = P^{-1}T^1P$ is parabolic, then $M^t = P^{-1}T^tP$. If $M = P^{-1}R^rP$ is elliptic, then $M^t = P^{-1}R^{rt}P$. □

3.6 Complex transformations

Möbius transformations can be aplied not only to real numbers but to complex numbers as well and their geometric and dynamic properties are more apparent in this setting. The real and imaginary parts of a complex number $z = x+iy$ is denoted by $\Re(z) = x$, $\Im(z) = y$, the complex conjugate of z is $\overline{z} = x-iy$ and its absolute value $|z| = \sqrt{z\cdot\overline{z}} = \sqrt{x^2+y^2}$. We consider general Möbius transformations on the **complex sphere** (i.e., extended complex plane) $\overline{\mathbb{C}} = \mathbb{C}\cup\{\infty\}$ given by

$$M(z) = \frac{az+b}{cz+d},\ M(-d/c) = \infty,\ M(\infty) = a/c,$$

where $a, b, c, d \in \mathbb{C}$ are complex numbers with $ad - bc \neq 0$. A complex transformation is determined by a complex matrix $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and if $\lambda \neq 0$ is a complex number, then M and λM determine the same transformation. Thus we have the space of complex projective matrices $\mathbb{P}(\mathbb{C}^{2\times2})$ and the space of regular complex projective matrices

$$\mathbb{M}(\mathbb{C}) = \{M \in \mathbb{P}(\mathbb{C}^{2\times2}) :\ \det(M) \neq 0\}$$

For the special case of **linear transformations** $M(z) = az + b$ we have $|M(z) - M(w)| = |a|\cdot|z-w|$, so a linear transformation is a similarity with respect to the Euclidean metric and therefore preserves all shapes. In particular, the image of a line is a line and the image of a circle is a circle. In a general complex transformation, the image of a line is either a line or a circle and the image of a circle is either a circle or a line. Thus the group of complex transformations creates a geometry, in which lines and circles cannot be distinguished. We show this property first for the transformation $M(z) = 1/z$.

Proposition 3.20 *The transformation $M(z) = 1/z$ transforms lines and circles to either lines or circles.*

Proof: 1. If $c \neq 0$ then the line $\{ct :\ t \in \mathbb{R}\}$ which joins 0 and c is transformed by M to the line $\{t/c :\ t \in \mathbb{R}\}$ which joins 0 and $1/c$ (see Figure 3.6 left).
2. If $c \neq 0$, then the line $\{c(1+it) :\ t \in \mathbb{R}\}$ which passes through c and is perpendicular to $0c$ is transformed to the circle with center $1/2c$ and radius $|1/2c|$ which passes through 0 and $1/c$. Indeed we have

$$\left|\frac{1}{c(1+it)} - \frac{1}{2c}\right| = \frac{|1-it|}{2|c|\cdot|1+it|} = \frac{1}{2|c|}$$

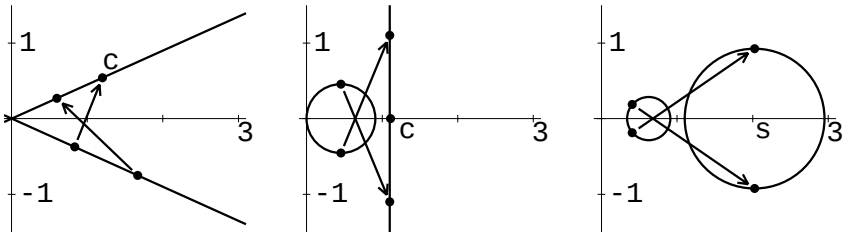


Figure 3.6: Transformation $1/z$ in the complex plane

since $|1-it|^2 = 1+t^2 = |1+it|^2$ (see Figure 3.6 center). Conversely a circle which passes through 0 is transformed to a line.
3. If $s \in \mathbb{C} \setminus \{0\}$ and $0 < r \neq 1$, then $\{s(1+r\alpha) : \ |\alpha| = 1\}$ is the circle with the center s and radius $r|s|$ which does not pass through 0. Its image is the circle with center $1/s(1-r^2)$ and radius $r/|s(1-r^2)|$. Indeed

$$\left| \frac{1}{s(1+r\alpha)} - \frac{1}{s(1-r^2)} \right| = \frac{r|r+\alpha|}{|s(1-r^2)| \cdot |1+r\alpha|} = \frac{r}{|s(1-r^2)|}$$

since $|r+\alpha| = r^2+1+r(\alpha+\overline{\alpha}) = |1+r\alpha|$ (see Figure 3.6 right). If $r > 0$ then the image of the circle $\{r\alpha : \ |\alpha| = 1\}$ is the circle $\{\frac{1}{r}\alpha : \ |\alpha| = 1\}$. □

Proposition 3.21 *Any complex transformation transforms lines and circles to either lines or circles.*

Proof: Let $M(z) = \frac{az+b}{cz+d}$. If $c = 0$ then M is a linear transformation which transforms lines to lines and circles to circles. If $c \neq 0$ then

$$M(z) = \frac{a}{c} + \frac{b-ad/c}{cz+d} = F_0F_1F_2(z)$$

where $F_0(z) = \frac{a}{c} + (b-ad/c)z$, $F_1(z) = 1/z$, $F_2(z) = cz+d$ and all F_i transform lines and circles to either lines or circles. □

Another important geometrical property of Möbius transformations is that they are **conformal**, i.e., they preserve angles. If two curves meet at angle α then the M-images of these curves meet at the same angle α. The conformality is a general property of **holomorphic functions**, (i.e., functions which have derivative - see e.g., Silverman [62]) at points c where their derivation $f'(c)$ is nonzero. In the neighbourhood of c we get an approximation $f(c+z) \approx f(c) + f'(c)z$ and the mapping $z \mapsto f(c) + f'(c)z$ is a similarity.

An example of a complex transformation is $\mathtt{d}(z) = \frac{iz+1}{z+i}$, which extends the stereographic projection to the extended complex plane (see Figure 3.7). Indeed for $x \in \mathbb{R}$ we get our original formula

$$\mathtt{d}(x) = \frac{ix+1}{x+i} \cdot \frac{x-i}{x-i} = \frac{2x+i(x^2-1)}{x^2+1}$$

Thus $\mathtt{d}$ maps the extended real line $\overline{\mathbb{R}} = \{z \in \mathbb{C} : \ \Im(z) = 0\} \cup \{\infty\}$ to the unit circle $\mathbb{S} = \{z \in \mathbb{C} : \ |z| = 1\}$. Since $\mathtt{d}(i) = 0$, the **upper half-plane** $\mathbb{U} = \{z \in \mathbb{C} : \ \Im(z) > 0\}$ is mapped to the **unit disc** $\mathbb{D} = \{z \in \mathbb{C} : \ |z| < 1\}$.

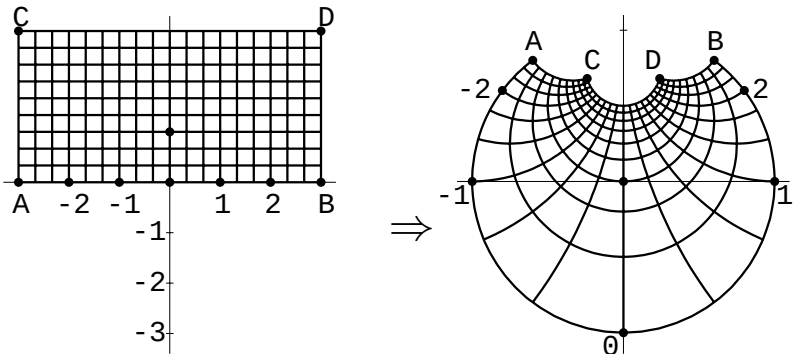


Figure 3.7: The stereographic projection in the complex plane

3.7 Hyperbolic geometry

Increasing transformations $M \in \mathbb{M}^+(\mathbb{R})$ map the **upper half-plane** $\mathbb{U} = \{z \in \mathbb{C} : \Im(z) > 0\}$ onto itself and preserve the hyperbolic (noneuclidean) metric in $\mathbb{U}$ (see e.g., Beardon [4]).

Proposition 3.22 *If $M \in \mathbb{M}^+(\mathbb{R})$, $M(z) = \frac{az+b}{cz+d}$, and $z \in \mathbb{U}$, then $M(z) \in \mathbb{U}$ and*

$$\Im(M(z)) = \frac{(ad-bc)\cdot\Im(z)}{|cz+d|^2}$$

Proof: We have

$$M(z) = \frac{(az+b)(c\overline{z}+d)}{(cz+d)(c\overline{z}+d)} = \frac{ac|z|^2+bd+adz+bc\overline{z}}{|cz+d|^2}$$

so if $\Im(z) > 0$ then $\Im(M(z)) = (ad-bc)\Im(z)/|cz+d|^2 > 0$. □

Definition 3.23 *The hyperbolic metric on $\mathbb{U}$ is defined by the differential form*

$$ds = \frac{|dz|}{\Im(z)} = \frac{\sqrt{dx^2+dy^2}}{y}, \text{ where } z = x+iy.$$

The hyperbolic metric is a special case of a **Riemannian metric** which is determined by a positive definite differential form. With a Riemannian metric, we can compute length of curves. In the case of the hyperbolic metric, if $z : [t_0, t_1] \to \mathbb{U}$ is a **differentiable curve** $z(t) = x(t) + iy(t)$, then the length of z is

$$L(z) = \int_{t_0}^{t_1} \frac{\sqrt{x'(t)^2+y'(t)^2}}{y(t)}\,dt.$$

Thus for example the curve $z(t) = t + ic$ maps $\mathbb{R}$ to the horizontal line through ic, so the length of a horizontal line from $a+ci$ to $b+ci$, where $a < b$ is

$$L(a+ci, b+ci) = \int_a^b \frac{dt}{c} = \frac{t}{c}\Big|_a^b = \frac{b-a}{c}.$$

The curve $z(t) = c + it$ maps $\mathbb{R}$ to the vertical line through c, so the length of a vertical line from $c+ai$ to $c+bi$, where $0 < a < b$ is

$$L(c+ai, c+bi) = \int_a^b \frac{dt}{t} = \ln(t)|_a^b = \ln\frac{b}{a}.$$

Proposition 3.24 *Transformations $M \in \mathbb{M}^+(\mathbb{R})$ preserve the hyperbolic metric. If $z : [t_0, t_1] \to \mathbb{U}$ is a differentiable curve, then $L(M \circ z) = L(z)$.*

Proof: For $M(z) = \frac{az+b}{cz+d}$ we have $M'(z) = \frac{ad-bc}{(cz+d)^2}$. If $w = M(z)$, then by Proposition 3.22 we get $\Im(w) = \frac{(ad-bc)\Im(z)}{|cz+d|^2}$, $dw = \frac{(ad-bc)dz}{(cz+d)^2}$, and

$$\frac{|dw|}{\Im(w)} = \frac{(ad-bc)\cdot|dz|}{|cz+d|^2} \cdot \frac{|cz+d|^2}{(ad-bc)\cdot\Im(z)} = \frac{|dz|}{\Im(z)}. \quad \square$$

Definition 3.25 *We say that a differentiable curve $z : [t_0, t_1] \to \mathbb{U}$ is a* **geodesic***, if its length is shorter than the length of any other differential curve from $z(t_0)$ to $z(t_1)$. We say that $z : \mathbb{R} \to \mathbb{U}$ is a geodesic if each its restriction to a finite interval $[t_0, t_1]$ is a geodesic.*

Proposition 3.26 *The vertical lines perpendicular to the real axis $\mathbb{R} = \{z \in \mathbb{C} : \ \Re(z) = 0\}$ are geodesics of the hyperbolic metric.*

Proof: Let $z : [t_0, t_1] \to \mathbb{U}$ be a differentiable curve with $\Re(z(t_0)) = \Re(z(t_1))$ and $\Im(z(t_0)) < \Im(z(t_1))$. Then

$$\begin{aligned} L(z) &= \int_{t_0}^{t_1} \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)} dt \geq \int_{t_0}^{t_1} \frac{|y'(t)|}{y(t)}\, dt \\ &\geq \int_{t_0}^{t_1} \frac{y'(t)}{y(t)}\, dt = \ln \frac{y(t_1)}{y(t_0)}. \end{aligned}$$

This is exactly the length of the vertical line joining $z(t_0)$ and $z(t_1)$. $\square$

Since the transformations of $\mathbb{M}^+(\mathbb{R})$ preseve hyperbolic metric, they map geodesics to geodesics. Since they are conformal, the image of a line perpendicular to $\mathbb{R}$ is a line or circle perpendicular to $\mathbb{R}$. Thus we have

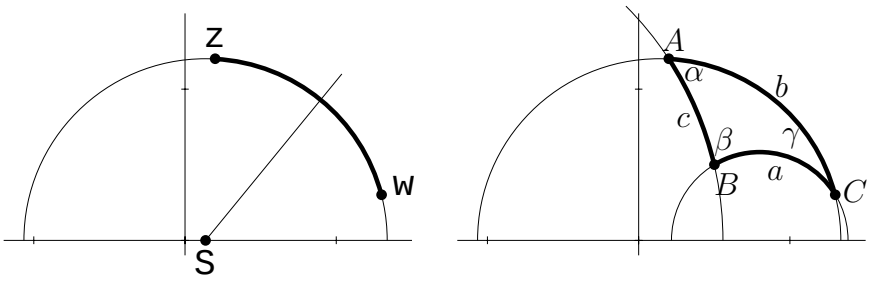


Figure 3.8: The geodesic which joins z and w(left) and a hyperbolic triangle(right)

Theorem 3.27 *The geodesics of the hyperbolic metric in $\mathbb{U}$ are either half-lines or semi-circles perpendicular to the real line $\mathbb{R}$.*

There exists a unique geodesic which joins two different $z, w \in \mathbb{U}$. If $\Re(z) = \Re(w)$ then it is the vertical line. If $\Re(z) \neq \Re(w)$ then the geodesic is the arc whose center S lies on the real

line and has the same Euclidean distance from z and w (Figure 3.8 left). The length of this geodesic, or the **hyperbolic distance** of z, w is given by

$$\varrho(z,w) = \ln \frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}$$

(see Beardon [4] for a proof). In particular we get $\varrho(c+ia, c+ib) = |\ln \frac{a}{b}|$. Three distinct points $A, B, C \in \mathbb{U}$ determine a unique **hyperbolic triangle** with vertices A, B, C. Its angles α, β, γ and the lengths of their sides a, b, c satisfy the relations of **hyperbolic trigonometry**. In the Euclidean geometry we have the sine and cosine rules which read

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c},\ \cos\gamma = \frac{a^2+b^2-c^2}{2ab}$$

In hyperbolic geometry we have

$$\begin{aligned} \frac{\sinh a}{\sin\alpha} &= \frac{\sinh b}{\sin\beta} = \frac{\sinh c}{\sin\gamma}, \\ \cos\gamma &= \frac{\cosh a\cdot\cosh b - \cosh c}{\sinh a\cdot\sinh b}, \\ \cosh c &= \frac{\cos\alpha\cdot\cos\beta + \cos\gamma}{\sin\alpha\cdot\sin\beta}, \end{aligned}$$

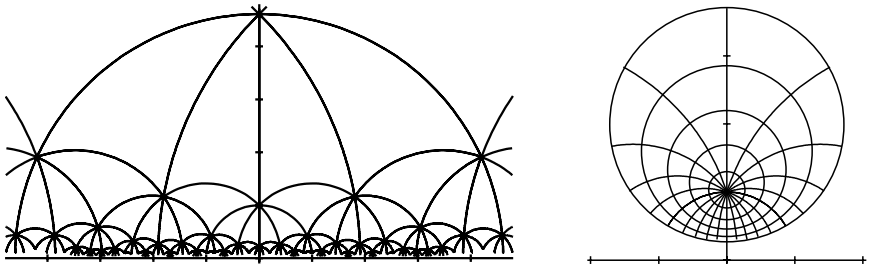


Figure 3.9: A tessellation of the hyperbolic plane by equilateral triangles with angles $\pi/8$ (left) and concentric circles and radii in the hyperbolic upper half-plane (right).

Since $\sinh x = \frac{e^x - e^{-x}}{2} \approx x$, the sine rule of the hyperbolic geometry approximates for small traingles the sine rule of the Euclidean geometry. We have two cosine rules. The first one is an analogue of the cosine rule of the Euclidean geometry obtained from the approximation $\cosh x = \frac{e^x + e^{-x}}{2} \approx 1 + \frac{x^2}{2}$. The second cosine rule computes angles from the sides. This is impossible in the Euclidean geometry, since there exist similar triangles with the same angles but differerent sides. In hyperbolic geometry there are no similar triangles. The sum of angles of a hyperbolic triangle is always less then π and larger triangles have smaller sum of angles. In fact we have the formula $\alpha + \beta + \gamma = \pi - P$ where P is the hyperbolic area of the triangle. Thus for example there exist equilateral triangles with angles π/n for each $n \geq 7$, and they tessellate the hyperbolic plane. One such tessellation can be seen in Figure 3.9 left. As another visualization of the hyperbolic plane, Figure 3.9 right shows concentric circles with center i and hyperbolic radii which form an arithmetic sequence. Hyperbolic circles are Euclidean circles but their hyperbolic center need not coincide with their hyperbolic center.

3.8 Disc transformations

The stereographic projection $\mathbf{d} : \mathbb{U} \to \mathbb{D} = \{z \in \mathbb{C} : \ |z| < 1\}$ maps the upper half-plane to the unit disc, and the lower halfplane $\mathbb{C} \setminus \overline{\mathbb{U}}$ to the exterior $\overline{\mathbb{C}} \setminus \overline{\mathbb{D}}$ of the unit disc. To each real transformation $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$ there corresponds a conjugated **disc transformation**

$$\widehat{M}(z) = \mathbf{d} \circ M \circ \mathbf{d}^{-1}(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}},$$

where $\alpha = (a+d) + (b-c)i$, $\beta = (b+c) + (a-d)i$. A disc transformation preserves the unit circle: if $z \in \mathbb{S}$ then $\widehat{M}(z) \in \mathbb{S}$. If $\det(M) > 0$, then

$$\det(\widehat{M}) = |\alpha|^2 - |\beta|^2 = (a+d)^2 - (a-d)^2 + (b-c)^2 - (b+c)^2 = 4(ad - bc) > 0$$

and $\widehat{M}$ preserves the unit disc. If $\det(M) < 0$ then $\det(\widehat{M}) < 0$ and $\widehat{M}$ maps the unit disc to its exterior and the exterior of $\mathbb{D}$ to $\mathbb{D}$. Conversely, any complex transformation of the form $F(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}}$ with $|\beta| \neq |\alpha|$ preserves the unit circle since

$$|F(e^{it})| = \frac{|\alpha e^{it} + \beta|}{|\overline{\beta} e^{it} + \overline{\alpha}|} = \frac{|\alpha e^{it} + \beta|}{|\overline{\beta + \alpha e^{it}} \cdot e^{it}|} = 1.$$

If $|F(0)| = \frac{|\beta|}{|\alpha|} < 1$ then F preserves the unit disc, otherwise it maps the unit disc to its exterior. The transformation $M = \mathbf{d}^{-1} \circ F \circ \mathbf{d} = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}^+(\mathbb{R})$ has real coefficients $a = \frac{\Re(\alpha)+\Im(\beta)}{2}$, $b = \frac{\Re(\beta)+\Im(\alpha)}{2}$, $c = \frac{\Re(\beta)-\Im(\alpha)}{2}$, $d = \frac{\Re(\alpha)-\Im(\beta)}{2}$. The hyperbolic metric on the upper half-plane is mapped by the stereographic projection $\mathbf{d} : \mathbb{U} \to \mathbb{D} = \{z \in \mathbb{C} : \ |z| < 1\}$ to a hyperbolic metric on the unit disc. A circle perpendicular to $\mathbb{R}$ is mapped to a circle perpendicular to the unit circle. Thus the geodesics of the **hyperbolic unit disc** are arcs or lines (diameters) perpendicular to the unit circle. Some tesselations of the hyperbolic disc are shown in Figure 3.10.

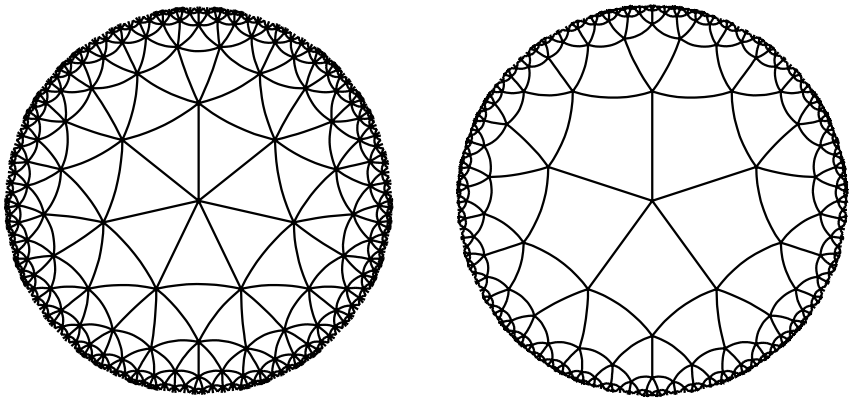


Figure 3.10: Tesselations of the hyperbolic unit disc by equilateral triangles with angles $2\pi/7$ (left) and by squares with angles $2\pi/5$ (right).

Proposition 3.28 *The stereographic projection transforms the metric $ds = |dz|/\Im z$ on $\mathbb{U}$ to the hyperbolic metric on the unit disc $\mathbb{D}$ given by*

$$\begin{aligned} ds &= \frac{2|dz|}{1-|z|^2} = \frac{2\sqrt{dx^2+dy^2}}{1-x^2-y^2}, \\ \varrho(z,w) &= 2 \operatorname{arg\,cosh} \frac{|1 - z\overline{w}|}{\sqrt{(1-|z|^2)(1-|w|^2)}} \end{aligned}$$

Proof: If $w = \mathbf{d}(z) = \frac{iz+1}{z+i}$, then $z = \mathbf{d}^{-1}(w) = \frac{-iw+1}{w-i} \cdot \frac{\overline{w}+i}{\overline{w}+i} = \frac{w+\overline{w}+i(1-|w|^2)}{|w-i|^2}$, so $\Im(z) = \frac{1-|w|^2}{|w-i|^2}$. By differentiation we get $dz = \frac{(i^2-1)dw}{(w-i)^2}$, so $|dz| = \frac{2|dw|}{|w-i|^2}$, and $\frac{|dz|}{\Im(z)} = \frac{2|dw|}{1-|w|^2}$. For the proof of the formula for $\varrho(z,w)$ se Beardon [4]. □

Since real Möbius transformations preserve the hyperbolic metric on $\mathbb{U}$, the circle transformations preserve the hyperbolic metric on $\mathbb{D}$. This can be verified directly. If $w = \frac{\alpha z+\beta}{\overline{\beta}z+\overline{\alpha}}$ then $|dw| = \frac{(|\alpha|^2-|\beta|^2)|dz|}{|\overline{\beta}z+\overline{\alpha}|^2}$, $1-|w|^2 = \frac{(|\alpha|^2-|\beta|^2)(1-|z|^2)}{|\overline{\beta}z+\overline{\alpha}|^2}$, so $\frac{2|dw|}{1-|w|^2} = \frac{2|dz|}{1-|z|^2}$.

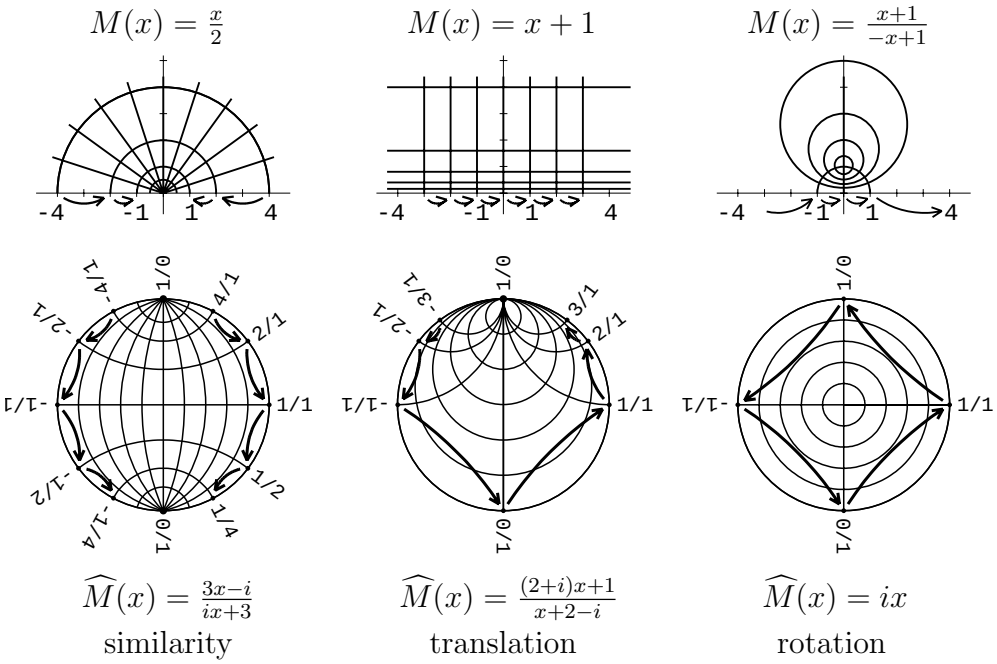


Figure 3.11: The similarity $Q_{\frac{1}{2}}(x) = \frac{x}{2}$ (left), the translation $T^1(x) = x+1$ (center) and the rotation $R^{\pi/2}(x) = \frac{1+x}{1-x}$ (right) in the upper half-plane (top) and in the unit disc (bottom)

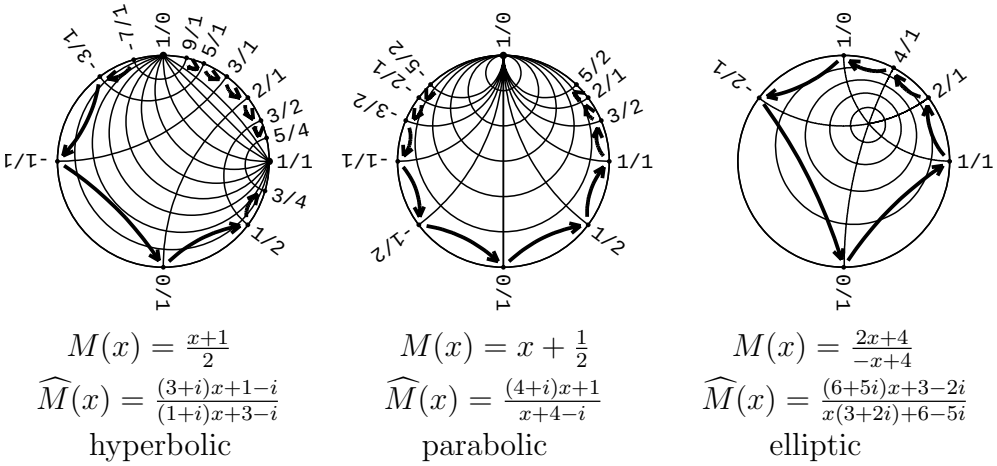


Figure 3.12: Disc transformations

The dynamics of a similarity, translation and rotation can be seen in Figure 3.11. The upper row shows the dynamics in the upper half-plane and the bottom row shows that in the unit disc. In the upper half-plane, the similarity $Q_{\frac{1}{2}}(x) = \frac{x}{2}$ maps a semicircle with center 0

and radius r (i.e., the geodesics joining $-r$ to r) to the circle with center 0 and radius $\frac{r}{2}$ and leaves invariant every line passing through 0. The translation $T^1(x) = x + 1$ maps a vertical line $\{x+it:\ t>0\}$ to the vertical line $\{x+1+it:\ t>0\}$ and leaves invariant every horizontal line. The rotation $R^{\pi/2}(x) = \frac{1+x}{1-x}$ maps a geodesic which passes through i to a perpendicular geodesic through i. In the unit disc, the similarity maps the geodesic which joins $\mathbf{d}(-x)$ with $\mathbf{d}(x)$ to the geodesic which joins $\mathbf{d}(-x/2)$ with $\mathbf{d}(x/2)$. It leaves invariant every arc which joins ∞ with 0. The translation maps a geodesic which joins $\mathbf{d}(x)$ with $\mathbf{d}(\infty)$ to the geodesic which joins $\mathbf{d}(x+1)$ with $\mathbf{d}(\infty)$ and leaves invariant every circle which passes through ∞. The rotation maps each diameter to a perpendicular diameter. It leaves invariant every circle with center 0. The dynamics of the transformation from Figures 3.4, 3.5 can be seen in Figure 3.12. Since these transformations are conjugated either to a similarity or to a translation or to a rotation, we have in each case a family of geodesics mapped to one another and a system of invariant curves perpendicular to these geodesics.

3.9 Isometric circles

For disc transformations we have an analogue of Proposition 3.7.

Proposition 3.29 *If $M \in \mathbb{M}(\mathbb{R})$ is a real MT, then $|M^\bullet(x)| = |\widehat{M}'(\mathbf{d}(x))|$*

Proof: Since $\widehat{M}(z) = \mathbf{d} \circ M \circ \mathbf{d}^{-1}(z)$ and $(\mathbf{d}^{-1})'(\mathbf{d}(x)) = 1/\mathbf{d}'(x)$, we get

$$\begin{aligned}
\widehat{M}'(\mathbf{d}(x)) &= \frac{\mathbf{d}'(M(x)) \cdot M'(x)}{\mathbf{d}'(x)} = \frac{-2 \cdot M'(x)}{(M(x)+i)^2} \cdot \frac{(x-i)^2}{-2} \\
|\widehat{M}'(\mathbf{d}(x))| &= \frac{|x-i|^2}{|M(x)+i|^2} \cdot |M'(x)| = \frac{x^2+1}{M^2(x)+1} \cdot \frac{|ad-bc|}{(cx+d)^2} \\
&= \frac{|ad-bc| \cdot (x^2+1)}{(ax+b)^2+(cx+d)^2} = |M^\bullet(x)|. \quad \square
\end{aligned}$$

Consider a real transformation $M \in \mathbb{M}(\mathbb{R})$, its disc conjugate $\widehat{M}(z) = \frac{\alpha z+\beta}{\overline{\beta} z+\overline{\alpha}}$, its inverse $\widehat{M}^{-1}(z) = \frac{\overline{\alpha} z-\beta}{-\overline{\beta} z+\alpha}$ and their derivations

$$\begin{aligned}
\widehat{M}'(z) &= \frac{|\alpha|^2-|\beta|^2}{(\overline{\beta} z+\overline{\alpha})^2} = \frac{|\beta|^2}{\overline{\beta}^2} \cdot \frac{|\frac{\alpha}{\beta}|^2-1}{(z+\frac{\overline{\alpha}}{\overline{\beta}})^2} \\
(\widehat{M}^{-1})'(z) &= \frac{|\alpha|^2-|\beta|^2}{(\overline{\beta} z-\alpha)^2} = \frac{|\beta|^2}{\overline{\beta}^2} \cdot \frac{|\frac{\alpha}{\beta}|^2-1}{(z-\frac{\alpha}{\beta})^2}
\end{aligned}$$

Note that $|\widehat{M}(0)| = |\widehat{M}^{-1}(0)| = |\frac{\beta}{\alpha}|$, $|\widehat{M}(\infty)| = |\widehat{M}^{-1}(\infty)| = |\frac{\alpha}{\beta}|$. $\overline{\widehat{M}(\infty)} = 1/\widehat{M}(0) = \frac{\overline{\alpha}}{\beta}$. If $\beta \neq 0$ then we have **isometric circles** K_M, $K_{M^{-1}}$ and **expanding discs** D_M, $D_{M^{-1}}$ defined

by

$$\begin{aligned} K_M &= \{z:\ |\widehat{M}'(z)|=1\} = \{z\in\mathbb{C}:\ |z+\tfrac{\overline{\alpha}}{\overline{\beta}}| = \sqrt{|1-|\tfrac{\alpha}{\beta}|^2|}\ \}, \\ D_M &= \{z:\ |\widehat{M}'(z)|>1\} = \{z\in\mathbb{C}:\ |z+\tfrac{\overline{\alpha}}{\overline{\beta}}| < \sqrt{|1-|\tfrac{\alpha}{\beta}|^2|}\ \}, \\ K_{M^{-1}} &= \{z:\ |(\widehat{M}^{-1})'(z)|=1\} = \{z\in\mathbb{C}:\ |z+\tfrac{\alpha}{\overline{\beta}}| = \sqrt{|1-|\tfrac{\alpha}{\beta}|^2|}\ \}, \\ D_{M^{-1}} &= \{z:\ |(\widehat{M}^{-1})'(z)|>1\} = \{z\in\mathbb{C}:\ |z+\tfrac{\alpha}{\overline{\beta}}| < \sqrt{|1-|\tfrac{\alpha}{\beta}|^2|}\ \}. \end{aligned}$$

All these circles and discs have the same radius $\mathbf{r}(M)=\sqrt{1-|\widehat{M}(\infty)|^2}$. If $\alpha=0$ then both K_M and $K_{M^{-1}}$ are the unit circles and D_M and $D_{M^{-1}}$ are the unit discs. For the **expanding interval** and the **contracting interval** of a transformation $M\in\mathbb{M}(\mathbb{R})$ we get by Proposition 3.29

$$\begin{aligned} \mathbf{U}(M) &= \{x\in\overline{\mathbb{R}}:\ |\widehat{M}'(\mathbf{d}(x))|<1\}, \\ \mathbf{V}(M) &= \{x\in\overline{\mathbb{R}}:\ |(\widehat{M}^{-1})'(\mathbf{d}(x))|>1\}. \end{aligned}$$

If either $\beta=0$ or $\alpha=0$, then $|\widehat{M}'(z)|=1$ for every z in the unit circle, so $\mathbf{U}(M)$ and $\mathbf{V}(M)$ are empty. Otherwise they are proper intervals and $\mathbf{d}(\mathbf{U}(M))=\mathbb{S}\setminus D_M$, $\mathbf{d}(\mathbf{V}(M))=\mathbb{S}\cap D_{M^{-1}}$.

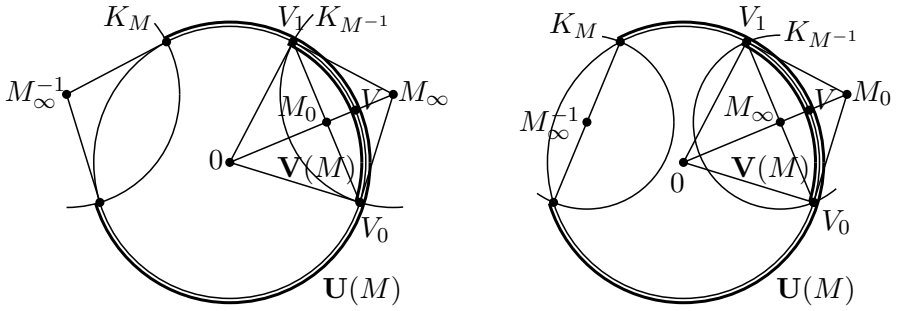


Figure 3.13: Isometric circles of increasing transformations with $0<|\beta|<|\alpha|$ (left) and decreasing transformations with $0<|\alpha|<|\beta|$ (right). Here $M_0=\widehat{M}(0)=\frac{\beta}{\alpha}$, $M_\infty=\widehat{M}(\infty)=\frac{\alpha}{\overline{\beta}}$, $V=\frac{\widehat{M}(0)}{|\widehat{M}(0)|}$, $M_\infty^{-1}=\widehat{M}^{-1}(\infty)=-\frac{\overline{\alpha}}{\overline{\beta}}$.

Proposition 3.30 *Let $M\in\mathbb{M}(\mathbb{R})$ and $\widehat{M}(0)\neq 0\neq\widehat{M}(\infty)$, so $\alpha\neq 0\neq\beta$. Then*
1. $\widehat{M}(K_M)=K_{M^{-1}}$,
2. $\widehat{M}(D_M)=\mathbb{C}\setminus\overline{D_{M^{-1}}}$,
3. $\widehat{M}(\mathbb{C}\setminus D_M)=\overline{D_{M^{-1}}}$,
4. $M(\mathbf{U}(M))=\mathbf{V}(M)$,
5. $|\mathbf{V}(M)|<\frac{1}{2}<|\mathbf{U}(M)|$,
6. $|\mathbf{U}(M)|+|\mathbf{V}(M)|=1$.

Proof: 1. We have $z\in K_M$ iff $|\widehat{M}'(z)|=1$ iff $|(\widehat{M}^{-1})'(\widehat{M}(z))|=1$ iff $\widehat{M}(z)\in K_{M^{-1}}$.
2,3. We have $z\in D_M$ iff $|\widehat{M}'(z)|>1$ iff $|(\widehat{M}^{-1})'(\widehat{M}(z))|<1$ iff $\widehat{M}(z)\notin\overline{D_{M^{-1}}}$.
4. We have $x\in\mathbf{U}(M)$ iff $|M^\bullet(x)|<1$ iff $|(M^{-1})^\bullet(M(x))|>1$ iff $M(x)\in\mathbf{V}(M)$.
5,6. Since the radii of D_M and $D_{M^{-1}}$ are the same, we see immediately $|\mathbf{V}(M)|<\frac{1}{2}<|\mathbf{U}(M)|$, $|\mathbf{U}(M)|+|\mathbf{V}(M)|=1$. □

Proposition 3.31 *Let $M \in \mathbb{M}(\mathbb{R})$, $\widehat{M}(0) \neq 0 \neq \widehat{M}(\infty)$ and denote by V_0, V_1 the $\mathbf{d}$-images of the endpoints of the expanding interval $\mathbf{V}(M) = (\mathbf{d}^{-1}(V_0), \mathbf{d}^{-1}(V_1))$. Then*

$$\begin{aligned} M \in \mathbb{M}^+(\mathbb{R}) \quad &\Rightarrow \quad \widehat{M}(0) = \tfrac{V_0+V_1}{2},\ |\mathbf{V}(M)| = \tfrac{1}{\pi}\arccos|\widehat{M}(0)| \\ M \in \mathbb{M}^-(\mathbb{R}) \quad &\Rightarrow \quad \widehat{M}(\infty) = \tfrac{V_0+V_1}{2}, |\mathbf{V}(M)| = \tfrac{1}{\pi}\arccos|\widehat{M}(\infty)|. \end{aligned}$$

Proof: Since $\widehat{M}(0)/\widehat{M}(\infty) = \frac{|\beta|^2}{|\alpha|^2} \in \mathbb{R}$, the points $0, \widehat{M}(0), \widehat{M}(\infty)$ lie on the same line. The triangles $(0, \widehat{M}(0), V_1)$ and $(0, V_1, \widehat{M}(\infty))$ are similar since they have the same angle at 0 and $|\widehat{M}(0)| : |V_1| = |V_1| : |\widehat{M}(\infty)|$. We distinguish two cases. If $M \in \mathbb{M}^+(\mathbb{R})$, then the triangle $(0, V_1, \widehat{M}(\infty)$ has the right angle at V_1 since $|V_1|^2 + |\widehat{M}(\infty) - V_1|^2 = |\widehat{M}(\infty)|^2$. It follows that the triangle $(0, \widehat{M}(0), V_1)$ has the right angle at $\widehat{M}(0)$. Thus V_0, $\widehat{M}(0)$, V_1 lie on the same line, $\widehat{M}(0) = (V_0 + V_1)/2$ and $|\mathbf{V}(M)| = \frac{1}{\pi}\arccos|\widehat{M}(0)|$ (see Figure 3.13 left). If $M \in \mathbb{M}^-(\mathbb{R})$ then $|\widehat{M}(\infty)| < 1$ and $(0, V_1, \widehat{M}(\infty)$ has the right angle at V_1 since $|V_1|^2 = |\widehat{M}(\infty) - V_1|^2 + |\widehat{M}(\infty)|^2$. Thus V_0, $\widehat{M}(\infty)$, V_1 lie on the same line, $\widehat{M}(\infty) = (V_0 + V_1)/2$ and $|\mathbf{V}(M)| = \frac{1}{\pi}\arccos|\widehat{M}(\infty)|$. □

Proposition 3.32 *Let $M \in \mathbb{M}(\mathbb{R})$. Then*

$$\begin{aligned} \min\{|M^\bullet(x)| :\ x \in \overline{\mathbb{R}}\} &= \left|\frac{1-|\widehat{M}(0)|}{1+|\widehat{M}(0)|}\right|, \\ \max\{|M^\bullet(x)| :\ x \in \overline{\mathbb{R}}\} &= \left|\frac{1+|\widehat{M}(0)|}{1-|\widehat{M}(0)|}\right| \end{aligned}$$

so $\min\{M^\bullet(x) :\ x \in \overline{\mathbb{R}}\} \cdot \max\{M^\bullet(x) :\ x \in \overline{\mathbb{R}}\} = 1$.

Proof: If either $\widehat{M}(0) = 0$ or $\widehat{M}(0) = \infty$, then $|\widehat{M^\bullet}(x)| = 1$ for all x and the claim holds. Assume $\widehat{M}(0) \neq 0 \neq \widehat{M}(\infty)$. Since $V = \frac{\widehat{M}(0)}{|\widehat{M}(0)|} = \frac{\beta|\alpha|}{\overline{\alpha}|\beta|} = \frac{\widehat{M}(\infty)}{|\widehat{M}(\infty)|} = \frac{\alpha|\beta|}{\overline{\beta}|\alpha|}$ is the closest point of $\mathbb{S}$ to $\widehat{M}(\infty)$, $|(\widehat{M^{-1}})'(x)|$ attains its smallest value in $\mathbb{S}$ at V. Since the centres of K_M and $K_{M^{-1}}$ have the same absolute value $|\widehat{M^{-1}}(\infty)| = |\widehat{M}(\infty)| = \frac{|\alpha|}{|\beta|}$, by Proposition 3.29 we get

$$\begin{aligned} \max\{|M^\bullet(x)| :\ x \in \overline{\mathbb{R}}\} &= \max\{|(\widehat{M^{-1}})'(x)| :\ x \in \mathbb{S}\} \\ &= |(\widehat{M^{-1}})'(V)| = \left|\frac{|\alpha|^2-|\beta|^2}{(\frac{\alpha|\beta|}{|\alpha|}-\alpha)^2}\right| = \left|\frac{|\alpha|^2-|\beta|^2}{(|\alpha|-|\beta|)^2}\right| \\ &= \left|\frac{|\alpha|+|\beta|}{|\alpha|-|\beta|}\right| = \left|\frac{1+|\widehat{M}(0)|}{1-|\widehat{M}(0)|}\right| \end{aligned}$$

Similarly

$$\begin{aligned} \min\{|M^\bullet(x)| :\ x \in \overline{\mathbb{R}}\} &= \min\{|(\widehat{M^{-1}})'(x)| :\ x \in \mathbb{S}\} \\ &= |(\widehat{M^{-1}})'(-V)| = \left|\frac{|\alpha|^2-|\beta|^2}{(\frac{\alpha|\beta|}{|\alpha|}+\alpha)^2}\right| = \left|\frac{|\alpha|^2-|\beta|^2}{(|\alpha|+|\beta|)^2}\right| \\ &= \left|\frac{|\alpha|-|\beta|}{|\alpha|+|\beta|}\right| = \left|\frac{1-|\widehat{M}(0)|}{1+|\widehat{M}(0)|}\right| \quad \square \end{aligned}$$

□

Proposition 3.33 *If* $\mathbf{V}(M) \neq \emptyset$*, then* $M : \mathbf{U}(M) \to \mathbf{V}(M)$ *is a contraction (see Definition 2.20).*

Proof: For $t \leq |\mathbf{U}(M)|$ set $\psi_M(t) = \sup\{|M(I)| : \ I \subseteq \mathbf{U}(M), |I| = t\}$. Then $\psi_M(t) < t$ and ψ_M is continuous. The function ψ_M can be extended arbitrarily to a decreasing function on $[0, 1]$ such that $\psi_M(t) < t$ for all t. □

We have seen that $|\widehat{M}(0)|$ characterizes the maximal and minimal contraction (circle derivation) of a transformation. An alternative characteristic is the norm of a transformation. The **norm of a matrix** $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is $||M|| = \sqrt{a^2 + b^2 + c^2 + d^2}$.

Definition 3.34 *Define the* **norm of a projective matrix** $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$ *by*

$$\mathrm{nrm}(M) = \frac{||M||^2}{\det(M)} = \frac{a^2 + b^2 + c^2 + d^2}{ad - bc}$$

Thus the norm of a decreasing transformation is negative.

Proposition 3.35 *If* $M \in \mathbb{M}(\mathbb{R})$ *then* $|\mathrm{nrm}(M)| \geq 2$*,* $|\mathrm{nrm}(M)| = 2$ *iff either* $\widehat{M}(0) = 0$ *or* $\widehat{M}(0) = \infty$*, and*

$$\begin{aligned}
|\widehat{M}(0)|^2 &= \frac{\mathrm{nrm}(M) - 2}{\mathrm{nrm}(M) + 2}, \\
\mathrm{nrm}(M) &= 2 \cdot \frac{1 + |\widehat{M}(0)|^2}{1 - |\widehat{M}(0)|^2}, \\
\min\{|M^\bullet(x)| : \ x \in \overline{\mathbb{R}}\} &= \frac{1}{2}(|\mathrm{nrm}(M)| - \sqrt{\mathrm{nrm}^2(M) - 4}\,) \\
\max\{|M^\bullet(x)| : \ x \in \overline{\mathbb{R}}\} &= \frac{1}{2}(|\mathrm{nrm}(M)| + \sqrt{\mathrm{nrm}^2(M) - 4}\,) \\
\mathbf{r}(M) &= \frac{2}{\sqrt{|\mathrm{nrm}(M) - 2|}} \ \textit{if}\ \widehat{M}(0) \neq 0 \\
|\mathbf{V}(M)| &= \frac{1}{\pi} \arcsin \frac{2}{\sqrt{|\mathrm{nrm}(M)| + 2}} \ \textit{if}\ 0 \neq \widehat{M}(0) \neq \infty
\end{aligned}$$

Thus $|\widehat{M}(0)| < 1$ *iff* $\mathrm{nrm}(M) \geq 2$ *and* $|\widehat{M}(0)| > 1$ *iff* $\mathrm{nrm}(M) \leq -2$.

Proof:

$$|\widehat{M}(0)|^2 = \left|\frac{\beta}{\alpha}\right|^2 = \frac{(b+c)^2 + (a-d)^2}{(a+d)^2 - (b-c)^2} = \frac{||M||^2 - 2\det(M)}{||M||^2 + 2\det(M)} = \frac{\mathrm{nrm}(M) - 2}{\mathrm{nrm}(M) + 2}$$

The other formulas follow from Proposition 3.32 by a simple algebra with the use of the formula $\arccos x = \arcsin \sqrt{1 - x^2}$. □

Proposition 3.36
1. If M is hyperbolic then $\overline{\mathbf{V}(M)} \subset \mathbf{U}(M)$.
2. If M is parabolic, then $\mathbf{V}(M) \subset \mathbf{U}(M)$ have a common endpoint.
3. If M is elliptic and $\mathbf{V}(M) \neq \emptyset$, then $\mathbf{V}(M) \not\subseteq \mathbf{U}(M)$.

Proof: 1. If M is hyperbolic then $\mathbf{u}(M) \in \overline{\mathbb{R}} \setminus \mathbf{U}(M)$ and $\lim_{n\to\infty} M^{-n}(x) = \mathbf{s}(M^{-1}) = \mathbf{u}(M) \in \mathbf{V}(M)$ so $\mathbf{V}(M) = M^{-1}(\mathbf{U}(M)) \subset \mathbf{U}(M)$ and therefore $\overline{\mathbf{V}(M)} \subset \mathbf{U}(M)$.
2. If M is parabolic, then its fixed point $\mathbf{s}(M)$ is an endpoint of both $\mathbf{V}(M)$ and $\mathbf{U}(M)$. Since M is orientation-preserving, we get $\mathbf{V}(M) \subset \mathbf{U}(M)$.
3. Suppose by contradiction that M is elliptic and $M^{-1}(\mathbf{U}(M)) = \mathbf{V}(M) \subseteq \mathbf{U}(M)$. Then M^{-1} has a fixed point in $\mathbf{U}(M)$ and this is impossible. □

3.10 Singular transformations

Besides **regular transformations** with nonzero determinant we consider **singular transformations** with zero determinant and the **zero transformation** $\mathbf{0} = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right] \in \mathbb{R}^{2\times 2}$ which does not belong to $\mathbb{P}(\mathbb{R}^{2\times 2})$. Denote by

$$\begin{aligned} \mathbb{M}^0(\mathbb{R}) &= \{M \in \mathbb{P}(\mathbb{R}^{2\times 2}) : \det(M) = 0\}, \\ \overline{\mathbb{M}}(\mathbb{R}) &= \mathbb{P}(\mathbb{R}^{2\times 2}) \cup \{\mathbf{0}\}. \end{aligned}$$

Thus $\overline{\mathbb{M}}(\mathbb{R})$ is the set of all subspaces of $\mathbb{R}^{2\times 2}$ of dimension at most 1. If $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is a singular transformation, then one of its rows is a multiple of the other, say $a = sc$, $b = sd$, so $M(x) = s$ whenever $cx + d \neq 0$. We say that $s = \mathbf{s}(M)$ is the **stable point** of M. We have $\mathbf{s}(M) = \frac{a}{c}$ provided $\frac{a}{c} \neq \frac{0}{0}$, otherwise $\mathbf{s}(M) = \frac{b}{d}$. For $x = \frac{-d}{c}$, we get $M(x) = \frac{0}{0} \notin \overline{\mathbb{R}}$ and we say that $\frac{-d}{c} = \mathbf{u}(M)$ is the **unstable point** of M provided $\frac{-d}{c} \neq \frac{0}{0}$, otherwise $\mathbf{u}(M) = \frac{-b}{a}$. For example for $M = \left[\begin{smallmatrix} a & 0 \\ c & 0 \end{smallmatrix}\right]$ we have $\mathbf{s}(M) = \frac{a}{c}$, $\mathbf{u}(M) = \frac{0}{1}$. For $M = \left[\begin{smallmatrix} 0 & b \\ 0 & d \end{smallmatrix}\right]$ we have $\mathbf{s}(M) = \frac{b}{d}$, $\mathbf{u}(M) = \frac{1}{0}$. If $\mathbf{s}(M) = s$ and $\mathbf{u}(M) = u$, then $M = \left[\begin{smallmatrix} s_0u_1 & -s_0u_0 \\ s_1u_1 & -s_1u_0 \end{smallmatrix}\right]$. The stable and unstable point of the zero transformation is defined by $\mathbf{u}(\mathbf{0}) = \mathbf{s}(\mathbf{0}) = \frac{0}{0}$. The operation of inversion $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right]$ is applied to singular or zero transformations as well and $(x, y) \in \Gamma(M)$ iff $(y, x) \in \Gamma(M^{-1})$. If M is singular, then $\mathbf{s}(M^{-1}) = \mathbf{u}(M)$, $\mathbf{u}(M^{-1}) = \mathbf{s}(M)$, and MM^{-1} is the zero transformation.

Proposition 3.37 *Let $P, Q \in \overline{\mathbb{M}}(\mathbb{R})$. Then $PQ = \mathbf{0}$ iff either $P = \mathbf{0}$ or $Q = \mathbf{0}$ or $P, Q \in \mathbb{M}^0(\mathbb{R})$ and $\mathbf{u}(P) = \mathbf{s}(Q)$. Otherwise PQ is singular provided either P or Q is singular and*

$$\begin{array}{lll} \mathbf{s}(PQ) = P(\mathbf{s}(Q)), & \mathbf{u}(PQ) = \mathbf{u}(Q) & \text{if } P \in \mathbb{M}(\mathbb{R}), Q \in \mathbb{M}^0(\mathbb{R}) \\ \mathbf{s}(PQ) = \mathbf{s}(P), & \mathbf{u}(PQ) = Q^{-1}(\mathbf{u}(P)) & \text{if } P \in \mathbb{M}^0(\mathbb{R}), Q \in \mathbb{M}(\mathbb{R}) \\ \mathbf{s}(PQ) = \mathbf{s}(P), & \mathbf{u}(PQ) = \mathbf{u}(Q) & \text{if } P, Q \in \mathbb{M}^0(\mathbb{R}), \mathbf{u}(P) \neq \mathbf{s}(Q) \end{array}$$

Proof: 1. Let $P \in \mathbb{M}(\mathbb{R}), Q \in \mathbb{M}^0(\mathbb{R})$. For each $x \neq \mathbf{u}(Q)$ we have $PQ(x) = P(\mathbf{s}(Q))$.
2. Let $P \in \mathbb{M}^0(\mathbb{R}), Q \in \mathbb{M}(\mathbb{R})$. For each $x \neq Q^{-1}(\mathbf{u}(Q))$ we have $PQ(x) = P(Q(x)) = \mathbf{s}(P)$.
3. Let $P, Q \in \mathbb{M}^0(\mathbb{R})$. For each $x \neq \mathbf{u}(Q)$ we have $PQ(x) = P(Q(x)) = \mathbf{s}(P)$. □

The projective space $\mathbb{P}(\mathbb{R}^{2\times 2})$ is a metric space with one of the equivalent projective metrics d_a, d_p, d_c (see Section 3.3) and singular transformations appear as limits of regular transformations. Note that $\mathbb{M}(\mathbb{R})$ is an open set in $\mathbb{P}(\mathbb{R}^{2\times 2})$, so its complement $\mathbb{M}^0(\mathbb{R})$ is a closed set.

Proposition 3.38
1. *If $M \in \mathbb{M}(\mathbb{R})$ is a hyperbolic transformation, then $\lim_{n\to\infty} M^n = Q \in \mathbb{M}^0(\mathbb{R})$ is a singular transformation with $\mathbf{s}(Q) = \mathbf{s}(M)$, $\mathbf{u}(Q) = \mathbf{u}(M)$.*
2. *If $M \in \mathbb{M}(\mathbb{R})$ is a parabolic transformation, then $\lim_{n\to\infty} M^n = Q \in \mathbb{M}^0(\mathbb{R})$ is a singular transformation with $\mathbf{s}(Q) = \mathbf{u}(Q) = \mathbf{s}(M)$.*
3. *If M is an elliptic transformation, then $\{M^n : n \geq 0\}$ has no limit in $\mathbb{P}(\mathbb{R}^{2\times 2})$.*

Proof: 1.A hyperbolic transformation is conjugated to a similarity, so there exists $P \in \mathbb{M}(\mathbb{R})$ and $0 < r < 1$ such that $M = PQ_rP^{-1}$. We have $\lim_{n\to\infty} Q_r^n = Q_0 = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ which has the stable point 0 and the unstable point ∞. It follows that $\lim_{n\to\infty} M^n = PQ_0P^{-1}$ has the stable point $\mathbf{s}(M)$ and unstable point $\mathbf{u}(M)$.
2. A parabolic transformation is conjugated to the translation $T(x) = x + 1$, so there exists $P \in \mathbb{M}(\mathbb{R})$ such that $M = PTP^{-1}$. We have $\lim_{n\to\infty} T^n = T_0 = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ with $\mathbf{s}(T_0) = \mathbf{u}(T_0) = \frac{1}{0}$. It follows that $\lim_{n\to\infty} M^n = PT_0P^{-1}$ has the stable and unstable point $\mathbf{s}(M)$. □

3.11 Representing sequences

If $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ is a sequence of regular transformations which has a singular limit $M \in \mathbb{M}^0(\mathbb{R})$ then we may say that $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ represents $\mathbf{s}(M)$. There is a more general concept of representation. Consider a sequence of hyperbolic transformations

$$M_{2n} = \begin{bmatrix} \varepsilon_n & 0 \\ 1-\varepsilon_n & 1 \end{bmatrix},\ M_{2n+1} = \begin{bmatrix} \varepsilon_n & 0 \\ \varepsilon_n - 1 & 1 \end{bmatrix},$$

where $\varepsilon_n > 0$ and $\lim_{n\to\infty} \varepsilon_n = 0$. Then $\lim_{n\to\infty} M_{2n} = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 1 \end{smallmatrix}\right]$, $\lim_{n\to\infty} M_{2n+1} = \left[\begin{smallmatrix} 0 & 0 \\ -1 & 1 \end{smallmatrix}\right]$, so $\lim_{n\to\infty} M_n$ does not exist. However, $\lim_{n\to\infty} M_n(z) = 0$ for each $z \in \overline{\mathbb{R}} \setminus \{-1, 1\}$. If we consider also complex z, then we find that $\lim_{n\to\infty} M_n(z) = 0$ for each $z \in \mathbb{C}$ with nonzero imaginary part. It turns out that this leads to a fruitful concept of representation which is based on Proposition 3.39.

Proposition 3.39 *Let $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ be a sequence of transformations and $x \in \overline{\mathbb{R}}$ such that $\lim_{n\to\infty} \widehat{M_n}(0) = \mathbf{d}(x)$. Then $\lim_{n\to\infty} \widehat{M_n}(z) = \mathbf{d}(x)$ for each $z \in \mathbb{C}$ with $|z| \neq 1$.*

Proof: See Figure 3.13. Denote by S_n the center of $K_{M_n^{-1}}$, S_n' the center of K_{M_n}, and r_n their radius. Since $\lim_{n\to\infty} \widehat{M_n}(0) = \mathbf{d}(x) \in \mathbb{S}$, we get $\lim_{n\to\infty} r_n = 0$. Given $z \in \overline{\mathbb{C}} \setminus \mathbb{S}$ there exists n_0 such that for every $n > n_0$ we have $z \in \mathbb{D} \setminus D_{M_n}$, so $\widehat{M_n}(z) \in D_{M_n^{-1}}$. Since $\widehat{M_n}(0) \in D_{M_n^{-1}}$, we get $|\widehat{M_n}(z) - \widehat{M_n}(0)| < r_n \to 0$, so $\lim_{n\to\infty} \widehat{M_n}(z) = \mathbf{d}(x)$. □

Since $\mathbf{d}(i) = 0$, we have $\lim_{n\to\infty} M_n(i) = x$ iff $\lim_{n\to\infty} M_n(z) = x$ for all $z \in \mathbb{C}$ with $\Im(z) \neq 0$. Here we use the convergence in $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. If $z_n \in \mathbb{C}$, then $\lim_{n\to\infty} z_n = \infty$ means $\lim_{n\to\infty} |z_n| = \infty$. If $z \in \mathbb{C}$, then $\lim_{n\to\infty} z_n = z$ is the convergence in the Euclidean metric.

Definition 3.40 *We say that a sequence of transformations $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ bfinrepresents $x \in \overline{\mathbb{R}}$ if $\lim_{n\to\infty} M_n(i) = x$.*

Theorem 3.41 *Given a sequence $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ and $x \in \overline{\mathbb{R}}$, the following conditions are equivalent:*

1. $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ *represents* $x \in \overline{\mathbb{R}}$.
2. $\exists z \in \mathbb{C} \setminus \mathbb{R}, \lim_{n\to\infty} M_n(z) = x$.
3. $\forall z \in \mathbb{C} \setminus \mathbb{R}, \lim_{n\to\infty} M_n(z) = x$.
4. $\exists z \in \overline{\mathbb{C}} \setminus \mathbb{S}, \lim_{n\to\infty} \widehat{M}_n(z) = \mathbf{d}(x)$.
5. $\forall z \in \overline{\mathbb{C}} \setminus \mathbb{S}, \lim_{n\to\infty} \widehat{M}_n(z) = \mathbf{d}(x)$.
6. *For each open interval* $I \subset \overline{\mathbb{R}}$ *with* $x \in I$ *we have* $\lim_{n\to\infty} |M_n^{-1}(I)| = 1$.
7. *There exists* $c > 0$ *and a sequence of closed intervals* I_m *such that* $x \in I_m$, $\lim_{m\to\infty} |I_m| = 0$, *and* $\liminf_{n\to\infty} |M_n^{-1}(I_m)| > c$ *for each* m.
8. *There exists a sequence* $\{x_n \in \overline{\mathbb{R}} : n \geq 0\}$ *with* $\lim_{n\to\infty} x_n = x$ *and* $\lim_{n\to\infty}(M_n^{-1})^\bullet(x_n) = \infty$.

Proof: 1 $\Rightarrow$ 2 is trivial.
2 $\Leftrightarrow$ 4 and 3 $\Leftrightarrow$ 5 follow from $\mathbf{d}(\overline{\mathbb{R}}) = \mathbb{S}$.
4 $\Rightarrow$ 5: Assume that $w \in \overline{\mathbb{C}} \setminus \mathbb{S}$ and $\lim_{n\to\infty} \widehat{M}_n(w) = \mathbf{d}(x)$. There exists a disc transformation F such that $\widehat{F}(0) = w$, so $\lim_{n\to\infty} \widehat{M}_n\widehat{F}(0) = \mathbf{d}(x)$ and by Proposition 3.39, $\lim_{n\to\infty} \widehat{M}_n\widehat{F}(y) = \mathbf{d}(x)$ for each $y \in \overline{\mathbb{C}}\setminus\mathbb{S}$. For each $z \in \mathbb{C}\setminus\mathbb{R}$ we get $\lim_{n\to\infty} \widehat{M}_n(z) = \lim_{n\to\infty} \widehat{M}_n\widehat{F}\widehat{F}^{-1}(z) = \mathbf{d}(x)$.
5 $\Rightarrow$ 6: By Proposition 3.35, $\lim_{n\to\infty} \mathrm{nrm}(M_n) = \infty$, $\lim_{n\to\infty} |\mathbf{V}(M_n)| = 0$, and $\lim_{n\to\infty} |\mathbf{U}(M_n)| = 1$. There exists n_0 such that for every $n > n_0$ we have $\mathbf{V}(M_n) \subseteq I$, so $\mathbf{U}(M_n) \subseteq M_n^{-1}(I)$. and $\lim_{n\to\infty} |M_n^{-1}(I)| = 1$.
6 $\Rightarrow$ 7 is trivial: We can take for I_m any intervals which contain x in their interior.
7 $\Rightarrow$ 8: For each m there exists n_m and $x_m \in I_m$ such that $|M_{n_m}^{-1})^\bullet(x_m)| \geq c/|I_m|$. It follows that the radii of the isometric circles converge to zero: $\lim_{m\to\infty} \mathbf{r}(M_{n_m}) = 0$. If $c/|I_m| > 1$ then $x_m \in \mathbf{V}(M_{n_m}) \cap I_m \neq \emptyset$ and $|\mathbf{d}(x_m) - \widehat{M}_{n_m}(0)| \leq \mathbf{r}(M_{n_m})$. It follows $\lim_{m\to\infty} \mathbf{d}(x_m) = \mathbf{d}(x)$, $\lim_{m\to\infty} x_m = x$, and $\lim_{m\to\infty} M_{n_m}^\bullet(x_m) = \infty$.
8 $\Rightarrow$ 1: From $\lim_{n\to\infty} \max\{|(M_n^{-1})^\bullet(x)| : x \in \overline{\mathbb{R}}\} = \infty$, we get $\lim_{n\to\infty} \mathrm{nrm}(M_n) = \infty$, $x_n \in \mathbf{V}(M_n)$, $\lim_{n\to\infty} |\mathbf{V}(M_n)| = 0$. Thus $|\mathbf{d}(x_n) - \widehat{M}_n(0)| \leq \mathbf{r}(M_n)$ and therefore $\lim_{n\to\infty} \widehat{M}_n(0) = \mathbf{d}(x)$, $\lim_{n\to\infty} M_n(i) = x$. □

Proposition 3.42 *Assume that* $\{M_n \in \mathbb{M}(\mathbb{R}) : n \geq 0\}$ *is a sequence of transformations,* $z, w \in \overline{\mathbb{R}}$, $z \neq w$ *and* $\lim_{n\to\infty} M_n(z) = \lim_{n\to\infty} M_n(w) = x \in \overline{\mathbb{R}}$. *Then* $\lim_{n\to\infty} M_n(i) = x$.

Proof: Let $M_n = [\frac{a_n}{c_n}, \frac{b_n}{d_n}]$. We can assume that the matrices M_n and vectors z, w are normed, i.e., $a_n^2 + b_n^2 + c_n^1 + d_n^2 = 1$, $z_0^2 + z_1^2 = w_0^2 + w_1^2 = 1$. Assume first $x \neq \infty$, so $x \in \mathbb{R}$. Then

$$0 = \lim_{n\to\infty} (M_n(z) - M_n(w)) = \lim_{n\to\infty} \frac{(a_nd_n - b_nc_n)\cdot(z_0w_1 - z_1w_0)}{(c_nz_0 + d_nz_1)\cdot(c_nw_0 + d_nw_1)}.$$

Since $z \neq w$, either $z \neq \infty$ or $w \neq \infty$. Assume $w \neq \infty$ and take $v = \frac{w_0+i}{w_1}$. Since $|c_nv_0 + d_nv_1| \geq |\Re(c_nv_0 + d_nv_1)| = |c_nw_0 + d_nw_1|$, we get

$$\lim_{n\to\infty} (M_n(z) - M_n(v)) = \lim_{n\to\infty} \frac{(a_nd_n - b_nc_n)\cdot(z_0v_1 - z_1v_0)}{(c_nz_0 + d_nz_1)\cdot(c_nv_0 + d_nv_1)} = 0.$$

Thus $\lim_{n\to\infty} M_n(v) = x$, and since $\Im(v) \neq 0$, we get $\lim_{n\to\infty} M_n(i) = x$ by Theorem 3.41.2. If $x = \infty$, then for the transformations $P_n = 1/M_n$ we have $\lim_{n\to\infty} P_n(z) = \lim_{n\to\infty} P_n(w) = 0$, so $\lim_{n\to\infty} P_n(i) = 0$ and therefore $\lim_{n\to\infty} M_n(i) = \infty$. □

Proposition 3.43 *Assume that $I \subset \overline{\mathbb{R}}$ is a proper closed interval and $\{M_n : n \geq 0\}$ is a sequence of transformations such that $M_n(I) \subseteq I$ for each n. If there exists a limit $\lim_{n\to\infty} M_0 M_1 \cdots M_n(i) = x$, then $x \in I$.*

Proof: Assume by contradiction that $x \in J = \overline{\mathbb{R}} \setminus I$ and denote by $P_n = M_0 \cdots M_n$. Then $P_n(I) \subseteq I$, $J = \overline{\mathbb{R}} \setminus I \subseteq \overline{\mathbb{R}} \setminus P_n(I) = P_n(J)$ and $P_n^{-1}(J) \subseteq J$. By Theorem 3.41.6, $\lim_{n\to\infty} |P_n^{-1}(J)| = 1$ and this is a contradiction with $P_n^{-1}(J) \subseteq J$. □

A special case of a representation involve **general continued fractions**. Let $\{a_n \in \mathbb{R}\setminus\{0\} : n \geq 1\}$, $\{b_n \in \mathbb{R} : n \geq 0\}$ be sequences of real numbers, The continued fraction

$$b_0 + \frac{a_1}{b_1 + \frac{a_2}{b_2 + \frac{a_3}{b_3 + \cdots}}} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdots}}}$$

represents an infinite product of regular transformations $\left[\begin{smallmatrix} 1 & b_0 \\ 0 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & a_1 \\ 1 & b_1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & a_2 \\ 1 & b_2 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & a_3 \\ 1 & b_3 \end{smallmatrix}\right] \cdots$. The n-th **convergents** p_n, q_n are defined by $p_{-1} = 1$, $q_{-1} = 0$, $p_0 = b_0$, $q_0 = 1$, $p_1 = a_1 + b_0 b_1$, $q_1 = b_1, \ldots$ $p_n = a_n p_{n-2} + b_n p_{n-1}$, $q_n = a_n q_{n-2} + b_n q_{n-1}$. Thus

$$\begin{bmatrix} p_{-1} & p_0 \\ q_{-1} & q_0 \end{bmatrix} = \begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} p_{n-2} & p_{n-1} \\ q_{n-2} & q_{n-1} \end{bmatrix} \cdot \begin{bmatrix} 0 & a_n \\ 1 & b_n \end{bmatrix} = \begin{bmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{bmatrix}$$

$$\begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_1 \\ 1 & b_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2 \\ 1 & b_2 \end{bmatrix} \cdots \begin{bmatrix} 0 & a_n \\ 1 & b_n \end{bmatrix} = \begin{bmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{bmatrix}$$

Definition 3.44 *We say that a general continued fraction $b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \ldots$ converges to $x \in \overline{\mathbb{R}}$ and write $b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \ldots = x$, if $\lim_{n\to\infty} \left[\begin{smallmatrix} 1 & b_0 \\ 0 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & a_1 \\ 1 & b_1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & a_2 \\ 1 & b_2 \end{smallmatrix}\right] \cdots \left[\begin{smallmatrix} 0 & a_n \\ 1 & b_n \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right] = x$.*

Definition 3.44 is more general than the classical definition of convergence which requires that p_n/q_n converge to x. If $\lim_{n\to\infty} \frac{p_n}{q_n} = x$, then by Proposition 3.42, $b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \ldots = x$, since the sequence converges to x in $z = 0$ and $z = \infty$. The converse implication, however, is not always satisfied. A counterexample is a periodic continued fraction

$$\frac{2}{1} + \frac{1}{0} + \frac{2}{1} + \frac{1}{0} + \frac{2}{1} + \cdots = 1.$$

The transformation $M = \left[\begin{smallmatrix} 0 & 2 \\ 1 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ is hyperbolic, has the stable fixed point 1 and the unstable fixed point 0, so $\lim_{n\to\infty} M^n(i) = \lim_{n\to\infty} \left[\begin{smallmatrix} 2^n & 0 \\ 2^n - 1 & 1 \end{smallmatrix}\right] \left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right] = 1$. However, p_n/q_n do not converge since $\frac{p_{2n}}{q_{2n}} = \frac{0}{1}$, $\frac{p_{2n+1}}{q_{2n+1}} = \frac{2^n}{2^n - 1} \to 1$, Even for this generalized convergence concept we have a classical result on equivalence of continued fractions:

Proposition 3.45 *Assume that $b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \ldots = x$ is a convergent continued fraction and let $\{r_i : i \geq 1\}$ be nonzero real numbers. Then*

$$b_0 + \frac{r_1 a_1}{r_1 b_1 +} \; \frac{r_1 r_2 a_2}{r_2 b_2 +} \; \frac{r_2 r_3 a_3}{r_3 b_3 +} \cdots = x.$$

Proof:

$$\begin{bmatrix} 0 & r_1 a_1 \\ 1 & r_1 b_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & r_1 a_2 \\ 1 & b_2 \end{bmatrix} = \begin{bmatrix} r_1 a_1 & r_1 a_1 b_2 \\ r_1 b_1 & r_1 a_2 + r_1 b_1 b_2 \end{bmatrix} = \begin{bmatrix} 0 & a_1 \\ 1 & b_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2 \\ 1 & b_2 \end{bmatrix} \quad \square$$

Chapter 4

Möbius number systems

A number system specifies the representation of real numbers by symbolic sequences, so its key element is the value mapping $\Phi : \Sigma \to \overline{\mathbb{R}}$. Möbius number systems are based on representations of real numbers by sequences of Möbius transformations, so the alphabet of the subshift Σ consists of the symbols of the transformations. We have several means how to define suitable subshift Σ and suitable value mapping Φ.

4.1 Iterative systems

An **iterative system** over an alphabet A is a system $F = \{F_a \in \mathbb{M}(\mathbb{R}) : \ a \in A\}$ of Möbius transformations indexed by letters of A. For a finite word $u \in A^n$, we denote by $F_u = F_{u_0} \circ \cdots \circ F_{u_{n-1}}$, the composition of F_{u_i} and by $F_\lambda = \mathrm{Id}_{\overline{\mathbb{R}}}$ the identity. An iterative system can be regarded as a mapping $F : A^* \times \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ which satisfies $F_{uv} = F_u \circ F_v$. Using the concept of representation from Definition 3.40, we define the **convergence space** $\mathbb{X}_F \subseteq A^\omega$ and the **value mapping** $\Phi : \mathbb{X}_F \to \overline{\mathbb{R}}$ by

$$\mathbb{X}_F = \{u \in A^\omega : \ \lim_{n\to\infty} F_{u_{[0,n)}}(i) \in \overline{\mathbb{R}}\}, \quad \Phi(u) = \lim_{n\to\infty} F_{u_{[0,n)}}(i).$$

Here i is the imaginary unit. Thus $u \in A^\omega$ belongs to $\mathbb{X}_F$ if the limit $\lim_{n\to\infty} F_{u_{[0,n)}}(i)$ exists and belongs to $\overline{\mathbb{R}}$.

Proposition 4.1 *Let F be an iterative system over A.*
1. *For $v \in A^+$, $u \in A^\omega$ we have $vu \in \mathbb{X}_F$ iff $u \in \mathbb{X}_F$, and then $\Phi(vu) = F_v(\Phi(u))$.*
2. *For $v \in A^+$ we have $v^\omega \in \mathbb{X}_F$ iff F_v is either parabolic or hyperbolic or decreasing with $F_v^2 \neq \mathrm{Id}$. In this case $\Phi(v^\omega) = \mathbf{s}(F_v)$ is the stable fixed point of F_v.*

Proof: 1 follows from the continuity of F_v.
2: If F_v is elliptic, then all $F_{v^k}(i)$ lie on a closed curve in $\mathbb{U}$, so $F_{v^k}(i)$ cannot converge to a real number. Let F_v be hyperbolic or parabolic, $|v| = p$. For each $0 \leq m < p$ we have $\lim_{k\to\infty} F_{(v^\omega_{[0,kp+m)})}(i) = \lim_{k\to\infty} F_v^k F_{v_{[0,m)}}(i) = \mathbf{s}(F_v)$, since the stable fixed point $\mathbf{s}(F_v)$ attracts all points of $\mathbb{U}$. Thus $\Phi(v^\omega) = \mathbf{s}(F_v)$. If F_v is decreasing and $F_v^2 \neq \mathrm{Id}$, then F_v^2 is a hyperbolic transformation and $\Phi(v^\omega) = \mathbf{s}(F_v^2) = \mathbf{s}(F_v)$. □

Note that the set $\mathbb{X}_F$ need not be closed, so it need not to be a subshift. Moreover, the value mapping $\Phi : \mathbb{X}_F \to \overline{\mathbb{R}}$ can be neither continuous nor surjective.

Definition 4.2 *We say that (F, Σ) is a* **number system***, if F is an iterative system and $\Sigma \subseteq \mathbb{X}_F$ is a subshift such that $\Phi : \Sigma \to \overline{\mathbb{R}}$ is continuous and surjective.*

Occasionally as in Section 1.4 we consider number systems for proper closed intervals $I \subset \overline{\mathbb{R}}$. In this case we require that $\Phi : \Sigma \to I$ is continuous and surjective.

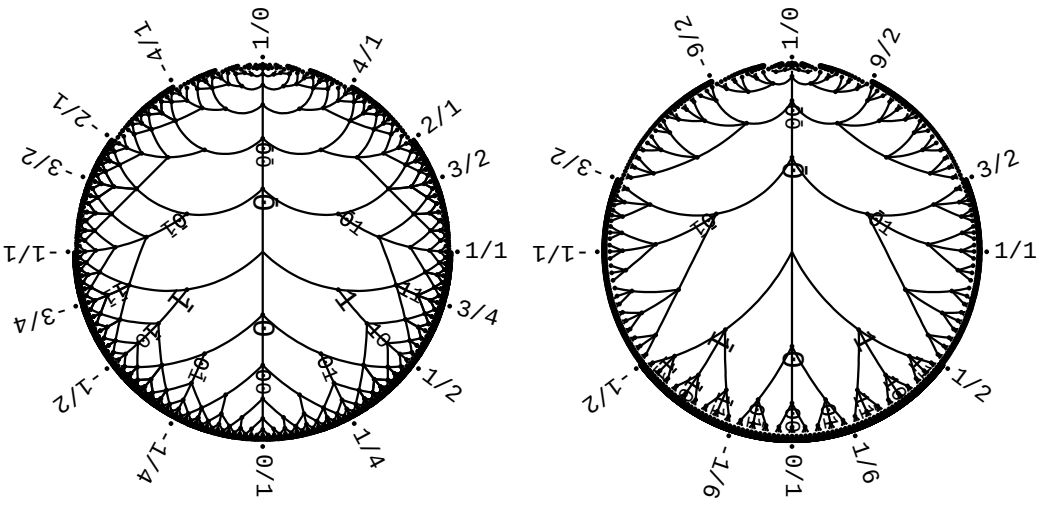


Figure 4.1: The binary signed system with forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0, 1\overline{1}, \overline{1}1\}$ (left) and the ternary signed system with forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0\}$ (right).

Example 4.3 *The binary signed system (F, Σ_D) from Proposition 1.8 has alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$, transformations $F_{\overline{1}}(x) = \frac{x-1}{2}$, $F_0(x) = \frac{x}{2}$, $F_1(x) = \frac{x+1}{2}$, $F_{\overline{0}}(x) = 2x$, and the subshift Σ_D with forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0, 1\overline{1}, \overline{1}1\}$.*

A finite word of $\mathcal{L}_D$ can be written as $\overline{0}^m u$, where $m \geq 0$ and $u \in \{\overline{1}, 0, 1\}^*$. If $|u| = n$ then $F_{\overline{0}^m u}(x) = 2^m \left(\frac{u_0}{2} + \cdots + \frac{u_{n-1}}{2^n} + \frac{x}{2^n}\right)$, so for $u \in \{\overline{1}, 0, 1\}^\omega$ we get

$$\Phi(\overline{0}^m u) = \lim_{n\to\infty} F_{\overline{0}^m u_{[0,n)}}(i) = \sum_{n\geq 0} u_n \cdot 2^{m-n-1}$$

Thus $\Sigma_D \subseteq \mathbb{X}_F$ and $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective. Figure 4.1 left shows the values of the disc transformations $\widehat{F}_u(0)$ in the complex unit disc $\mathbb{D} = \{z \in \mathbb{C} : \ |z| < 1\}$. The labels $u \in A^+$ at $\widehat{F}_u(0)$ are written in the direction of the tangent vectors $\widehat{F}'_u(0)$. Recall that for an increasing transformation $M \in \mathbb{M}^+(\mathbb{R})$ there exists a family of transformations $(M^t)_{t\in\mathbb{R}}$ such that $M^0 = \mathrm{Id}$, $M^1 = M$, and $M^{t+s} = M^t \circ M^s$ (Proposition 3.19). In Figure 4.1, a point $\widehat{F}_u(0)$ is joined to $\widehat{F}_{ua}(0)$ by the curve $\{\widehat{F}_u \widehat{F}_a^t(0) : \ 0 \leq t \leq 1\}$.

Example 4.4 *The ternary signed system (F, Σ_D) from Proposition 1.6 has alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$, transformations $F_{\overline{1}}(x) = \frac{x-1}{3}$, $F_0(x) = \frac{x}{3}$, $F_1(x) = \frac{x+1}{3}$ $F_{\overline{0}}(x) = 3x$ and the subshift Σ_D with forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0\}$.*

For $u \in \{\overline{1}, 0, 1\}^\omega$ we get

$$\Phi(\overline{0}^m u) = \lim_{n\to\infty} F_{\overline{0}^m u_{[0,n)}}(i) = \sum_{n\geq 0} u_n \cdot 3^{m-n-1}$$

Thus $\Sigma_D \subseteq \mathbb{X}_F$ and as proved in Chapter 1, $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is continuous and surjective (see Figure 4.1 right).

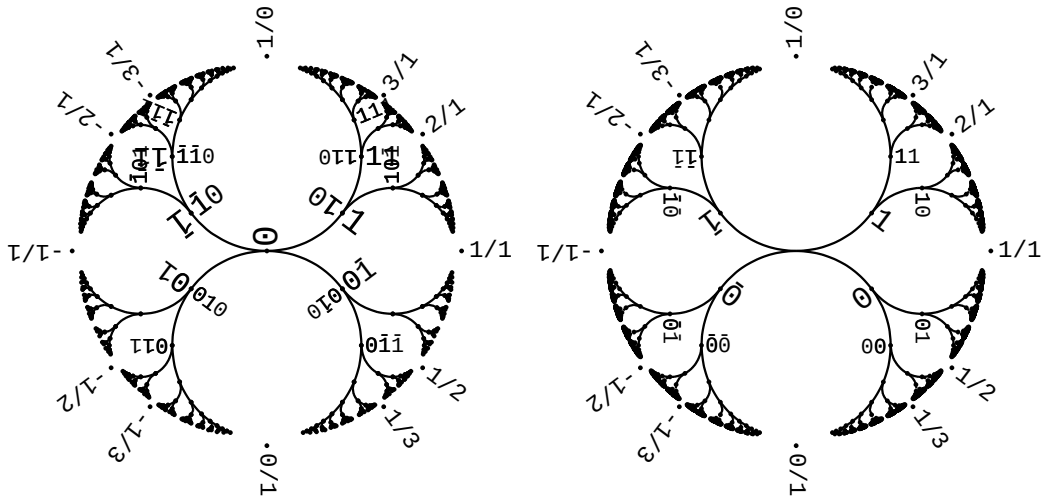


Figure 4.2: The system of signed continued fractions from Example 4.5 (left) and the system of symmetric continued fractions from Definition 1.14 and Example 4.6 (right).

We modify the number system of simple continued fractions of Definition 1.13. Instead of the decreasing transformation $1/x$ we take the increasing transformation $F_0(x) = -1/x$. When we expand a number $x > 1$, we subtract 1 (apply $F_1^{-1}(x) = x - 1$) till we get into the interval $[0, 1)$.The we apply $F_0^{-1}(x) = -1/x$, so we get a negative number smaller than -1. Then we apply $F_{\overline{1}}^{-1}(x) = x + 1$ till we get into the interval $(-1, 0]$. The words 101 and $\overline{1}0\overline{1}$ do not occur in this expansion process.

Example 4.5 *The system* (F, Σ_D) *of* **signed continued fractions** *consists of the alphabet* $A = \{\overline{1}, 0, 1\}$*, transformations* $F_{\overline{1}}(x) = x - 1$*,* $F_0(x) = -1/x$*,* $F_1(x) = x + 1$*, and the subshift with forbidden words* $D = \{00, \overline{1}1, 1\overline{1}, \overline{1}0\overline{1}, 101\}$*.*

Each word $u \in \mathcal{L}_D$ can be written as $u = 1^{a_0}01^{a_1}0\cdots01^{a_n}$, where $a_i \in \mathbb{Z}$, $a_0a_1 \leq 0$ and $a_ia_{i+1} < 0$ for $i > 0$ (so $a_i \neq 0$ for $i > 0$). If $a < 0$ then 1^a means $\overline{1}^{-a}$. For the transformation F_u we get

$$\begin{aligned} F_u(x) &= \begin{bmatrix} 1 & a_0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & -1 \\ 1 & a_1 \end{bmatrix} \cdots \begin{bmatrix} 0 & -1 \\ 1 & a_n \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} \\ &= a_0 - \frac{1}{a_1 -} \frac{1}{a_2 -} \cdots - \frac{1}{a_{n-1} -} \frac{1}{(a_n + x)}. \end{aligned}$$

This is equivalent to a simple continued fraction which has either positive entries $a_1, -a_2, a_3, -a_4, \ldots$ or positive entries $-a_1, a_2, -a_3, \ldots$. For an infinite word $u = 1^{a_0}01^{a_1}01^{a_2}0\cdots \in \Sigma_D$ we get a converging sequence

$$\Phi(u) = \lim_{n\to\infty} F^{-1}_{u_{[0,n)}}(i) = a_0 - \frac{1}{a_1 -} \frac{1}{a_2 -} \frac{1}{a_3 -} \cdots$$

The sequence $\{a_n : n \geq 0\}$ may be finite if its last member is infinite. In this case we get

$$\begin{aligned} \Phi(1^{a_0}0\cdots01^{a_n}01^\omega) &= a_0 - \frac{1}{a_1 -} \frac{1}{a_2 -} \cdots - \frac{1}{a_{n-1} -} \frac{1}{a_n}, \\ \Phi(1^\omega) &= \Phi(\overline{1}^\omega) = \infty. \end{aligned}$$

Example 4.6 *The number system of* **symmetric continued fractions** *of Definition 1.14 with the alphabet* $A = \{\overline{1}, \overline{0}, 0, 1\}$ *consists of transformations* $F_{\overline{1}}(x) = x - 1$, $F_{\overline{0}}(x) = \frac{x}{1-x}$, $F_0(x) = \frac{x}{x+1}$, $F_1(x) = x + 1$, *and the subshift* $\Sigma_D = \{\overline{1}, \overline{0}\}^\omega \cup \{0, 1\}^\omega$ *with forbidden words* $D = \{\overline{0}0, \overline{0}1, \overline{1}0, \overline{1}1, 0\overline{1}, 0\overline{0}, 1\overline{1}, 1\overline{0}\}$.

As proved in Chapter 1, the value mapping Φ is continuous and surjective

$$\begin{aligned} \Phi(1^{a_0}0^{a_1}1^{a_2}\cdots) &= a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}} \\ \Phi(\overline{1}^{a_0}\overline{0}^{a_1}\overline{1}^{a_2}\cdots) &= -a_0 - \frac{1}{a_1 - \frac{1}{a_2 - \cdots}} \end{aligned}$$

The transformations of the system are parabolic. $F_{\overline{1}}$, F_1 have the fixed point ∞, $F_{\overline{0}}$, F_0 have the fixed point 0. The system has two symmetries. The transformation $-x$ conjugates $F_{\overline{1}}$ to F_1 and $F_{\overline{0}}$ to F_0. The transformation $1/x$ conjugates $F_{\overline{0}}$ to $F_{\overline{1}}$ and F_0 to F_1.

4.2 Interval number systems

In Chapter 1 we define several number systems by means of the expansion process. In all cases we have a SFT Σ of order two and a system of closed intervals $\{W_a : a \in A\}$ such that $F_a^{-1}(W_a) = \bigcup\{W_b : ab \in \mathcal{L}_D^2\}$. Let us generalize this approach. Given an iterative system F over A and a system of intervals $\{W_a : a \in A\}$, we may consider the subshift of all expansions. A word $u \in A^\omega$ is an expansion of $x \in \overline{\mathbb{R}}$ iff $x_i = F^{-1}_{u_{[0,i)}}(x) \in W_{u_i}$ for all i. It turns out, however, that this does not always work properly. For example in the system of simple continued fractions from Definition 1.13 with $F_0(x) = 1/x$, $W_0 = [0, 1]$, we get the expansion 0^ω of 1, but 0^ω belongs neither to Σ_D nor to $\mathbb{X}_F$: the sequence $F_0^n(i)$ does not converge. A remedy is to take for W_a the open intervals $W_{\overline{1}} = (\infty, 0)$, $W_0 = (0, 1)$, $W_1 = (1, \infty)$. Since $F_0^{-1}(W_0) \cap W_0 = \emptyset$, the word 00 is a forbidden and for a similar reason, the words $\overline{1}1$, $0\overline{1}$ and $1\overline{1}$ are forbidden as well. Although W_a do not cover $\overline{\mathbb{R}}$, the numbers $0, 1, \infty$, which are not covered by W_a have expansions which are the limits of expansions of points in W_a.

Definition 4.7 *We say that* $W = \{W_a \subset \overline{\mathbb{R}} : a \in A\}$ *is an* **open almost-cover** *if* W_a *are proper open intervals and* $\bigcup_{a\in A} \overline{W_a} = \overline{\mathbb{R}}$. *Let* $F = \{F_a \in \mathbb{M}(\mathbb{R}) : a \in A\}$ *be an iterative system and let* $W = \{W_a \subset \overline{\mathbb{R}} : a \in A\}$ *be an open almost-cover. A finite or infinite sequence* $u \in A^* \cup A^\omega$ *is an* **expansion** *of* $x \in \overline{\mathbb{R}}$ *if* $x_n = F^{-1}_{u_{[0,n)}}(x) \in W_{u_n}$ *for each* $n < |u|$. *The sequence* $\{x_n : n \geq 0\}$ *is called the* **trajectory** *of* x. *We denote by* W_u *the set of points with the expansion* $u \in A^*$.

Thus $x \in W_u$ if for each $i \leq n$ we have $x_i = F^{-1}_{u_{[0,i)}}(x) \in W_{u_i}$ iff $x \in F_{u_{[0,i)}}(W_{u_i})$. For $u \in A^{n+1}$ we get

$$W_u = W_{u_0} \cap F_{u_0}(W_{u_1}) \cap F_{u_{[0,2)}}(W_{u_2}) \cap \cdots \cap F_{u_{[0,n)}}(W_{u_n}).$$

The **expansion subshift** $\mathcal{S}_{F,W}$ with the **expansion language** $\mathcal{L}_{F,W} = \mathcal{L}(\mathcal{S}_{F,W})$ is defined by

$$\begin{aligned} \mathcal{L}_{F,W} &= \{u \in A^* : W_u \neq \emptyset\}, \\ \mathcal{S}_{F,W} &= \{u \in A^\omega : \forall n, W_{u_{[0,n)}} \neq \emptyset\}. \end{aligned}$$

As a convention we set $W_\lambda = \overline{\mathbb{R}}$. For $u, v \in A^*$ we have $W_{uv} = W_u \cap F_u(W_v)$.

Example 4.8 *For the ternary signed system (F, Σ_D) from Example 4.4 we have transformations and intervals*

$$\begin{array}{rclrclrclrcl} F_{\overline{1}}(x) & = & (x-1)/3, & F_0(x) & = & x/3, & F_1(x) & = & (x+1)/3, & F_{\overline{0}}(x) & = & 3x \\ W_{\overline{1}} & = & (-\frac{1}{2}, -\frac{1}{6}), & W_0 & = & (-\frac{1}{6}, \frac{1}{6}), & W_1 & = & (\frac{1}{6}, -\frac{1}{2}), & W_{\overline{0}} & = & (\frac{1}{2}, -\frac{1}{2}) \end{array}$$

and we get $\mathcal{S}_{F,W} = \Sigma_{\{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0\}}$.

Example 4.9 *For the system of simple continued fractions of Definition 1.13 we have transformations and intervals*

$$\begin{array}{rclrclrcl} F_{\overline{1}}(x) & = & x-1, & F_0(x) & = & 1/x, & F_1(x) & = & x+1, \\ W_{\overline{1}} & = & (\infty, 0), & W_0 & = & (0,1), & W_1 & = & (1, \infty). \end{array}$$

and we get $\mathcal{S}_{F,W} = \Sigma_{\{\overline{1}1, 0\overline{1}, 00, 1\overline{1}\}}$.

Example 4.10 *For the system of signed continued fractions from Definition 4.5 we have transformations and intervals*

$$\begin{array}{rclrclrcl} F_{\overline{1}}(x) & = & x-1, & F_0(x) & = & -1/x, & F_1(x) & = & x+1, \\ W_{\overline{1}} & = & (\infty, -1), & W_0 & = & (-1,1). & W_1 & = & (1, \infty). \end{array}$$

and we get $\mathcal{S}_{F,W} = \Sigma_{\{00, \overline{1}1, 1\overline{1}, \overline{1}0\overline{1}, 101\}}$.

Example 4.11 *For the system of symmetric continued fractions from Definition 1.14 we have transformations and intervals*

$$\begin{array}{rclrclrclrcl} F_{\overline{1}}(x) & = & x-1, & F_{\overline{0}}(x) & = & \frac{x}{1-x}, & F_0(x) & = & \frac{x}{x+1}, & F_1(x) & = & x+1, \\ W_{\overline{1}} & = & (\infty, -1), & W_{\overline{0}} & = & (-1, 0), & W_0 & = & (0,1), & W_1 & = & (1, \infty). \end{array}$$

and we get $\mathcal{S}_{F,W} = \{\overline{1}, \overline{0}\}^\omega \cup \{0,1\}^\omega$.

In the next Theorem 4.12 we give conditions which imply that $(F, \mathcal{S}_{F,W})$ is a number system. In the proof we work with the lengths of sets W_u which are not necessarily intervals. Each W_u is either a proper interval or a finite union of proper intervals. Define the length of a set $Y \subseteq \overline{\mathbb{R}}$ as the length of the shortest interval I such that $Y \subseteq I$.

Theorem 4.12 (Kůrka and Kazda [44]) *Let $F = \{F_a \in \mathbb{M}(\mathbb{R}) : \ a \in A\}$ be an iterative system and $W = \{W_a \subset \overline{\mathbb{R}} : \ a \in A\}$ an open almost-cover such that $W_a \subseteq \mathbf{V}(F_a)$ for each $a \in A$. Then*

1. *$(F, \mathcal{S}_{F,W})$ is a number system, so $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is continuous and surjective.*
2. *$\{\Phi(u)\} = \bigcap_{n>0} \overline{W_{u_{[0,n)}}}$ for each $u \in \mathcal{S}_{F,W}$.*
3. *$\Phi([u]) = \overline{W_u}$ for each $u \in \mathcal{L}_{F,W}$.*
4. *If $\{W_a : \ a \in A\}$ is a cover of $\overline{\mathbb{R}}$, then $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is redundant.*

Proof: We use the angle metric d_a, so if $I = (\mathfrak{l}(I), \mathbf{r}(I))$ is an interval with length $|I| < \frac{1}{2}$, then its length is the distance of its endpoints. For a proper interval $I \subseteq \overline{\mathbb{R}}$ and $0 < \varepsilon < |I|/2$ denote by

$$I^{\varepsilon-} = \overline{I} \cap B_\varepsilon(\mathfrak{l}(I)), \ I^{\varepsilon+} = \overline{I} \cap B_\varepsilon(\mathbf{r}(I)), \ I^{\varepsilon} = \overline{I} \setminus (I^{\varepsilon-} \cup I^{\varepsilon+}).$$

Denote by $l_a = \mathfrak{l}(W_a)$, $r_a = \mathbf{r}(W_a)$ the left and right endpoints of W_a. Since F_a are contractions on $F_a^{-1}(W_a) \subseteq \mathbf{U}(F_a)$, there exists an increasing continuous function $\psi : [0,1] \to [0,1]$ such

that $\psi(0) = 0$, $0 < \psi(t) < t$ for $t > 0$, and $|F_a(Y)| \le \psi(|Y|)$ for each $a \in A$ and any set $Y \subseteq F_a^{-1}(W_a)$. Given $u \in \mathcal{S}_{(F,W)}$ and $m \le n$ we have

$$F^{-1}_{u_{[0,m]}}(W_{u_{[0,n]}}) \subseteq F^{-1}_{u_{[0,m]}} F_{u_{[0,m)}}(W_{u_m}) = F^{-1}_{u_m}(W_{u_m}) \subseteq \mathbf{U}(F_{u_m}).$$

For each $n > 0$ we get

$$\begin{aligned} |W_{u_{[0,n]}}| &= |F_{u_0} F^{-1}_{u_0}(W_{u_{[0,n]}})| \le \psi(|F^{-1}_{u_0}(W_{u_{[0,n]}})|) = \psi(|F_{u_1} F^{-1}_{u_{[0,1]}}(W_{u_{[0,n]}})|) \\ &\le \psi^2(|F^{-1}_{u_{[0,1]}}(W_{u_{[0,n]}})|) \le \cdots \le \psi^n(|F^{-1}_{u_{[0,n)}}(W_{u_{[0,n]}})|) \le \psi^n(|W_{u_n}|) \le \psi^n(1). \end{aligned}$$

Since $\psi(t) < t$ and the only fixed point of ψ is zero, we get $\lim_{n\to\infty} |W_{u_{[0,n]}}| = 0$, so there exists a unique point

$$x \in \bigcap_{n\ge 0} \overline{W_{u_{[0,n]}}} \subseteq \overline{W_{u_0}} \cap F_{u_0}(\overline{W_{u_1}}) \cap \cdots \cap F_{u_{[0,n)}}(\overline{W_{u_n}}).$$

We show that $u \in \mathbb{X}_F$ and $\Phi(u) = x$. If $a, b \in A$ and $F_a^{-1}(l_a) \in W_b$, then $F_a^{-1}(I) \subseteq W_b$ for some open interval $I \ni a$. Thus there exists $\varepsilon > 0$ such that for any $a, b \in A$,

$$\begin{aligned} F_a^{-1}(l_a) \in W_b &\quad\Rightarrow\quad F_a^{-1}(W_a^{\varepsilon-}) \subseteq W_b^{\varepsilon} \\ F_a^{-1}(r_a) \in W_b &\quad\Rightarrow\quad F_a^{-1}(W_a^{\varepsilon+}) \subseteq W_b^{\varepsilon} \end{aligned}$$

Denote by $x_n = F^{-1}_{u_{[0,n)}}(x)$, $x_0 = x$. Since $x \in \overline{W_{u_{[0,n]}}}$, we get $x_n \in F^{-1}_{u_{[0,n)}}(\overline{W_{u_{[0,n]}}}) \subseteq \overline{W_{u_n}}$. For the circle derivation we get

$$(F^{-1}_{u_{[0,n)}})^\bullet(x) = (F_{u_0}^{-1})^\bullet(x_0) \cdot (F_{u_1}^{-1})^\bullet(x_1) \cdots (F^{-1}_{u_{n-1}})^\bullet(x_{n-1}),$$

and each factor in this product is at least 1. If $x_n \in W^{\varepsilon}_{u_n}$ for an infinite number of n, then $\lim_{n\to\infty}(F^{-1}_{u_{[0,n)}})^\bullet(x) = \infty$ and $\Phi(u) = x$ by Theorem 3.41. Assume therefore that there exists n_0 such that $x_n \in \overline{W_{u_n}} \setminus W^{\varepsilon}_{u_n} = W^{\varepsilon-}_{u_n} \cup W^{\varepsilon+}_{u_n}$ for each $n \ge n_0$. Let $x_n \in W^{\varepsilon-}_{u_n}$. Since $x_{n+1} = F^{-1}_{u_n}(x_n) \notin W^{\varepsilon}_{u_{n+1}}$, we get $F^{-1}_{u_n}(l_{u_n}) \notin W_{u_{n+1}}$, so $F^{-1}_{u_n}(l_{u_n}) \in \{l_{u_{n+1}}, r_{u_{n+1}}\}$. Since $F^{-1}_{u_n}(W_{u_n}) \cap W_{u_{n+1}}$ is nonempty, we get $F^{-1}_{u_n}(l_{u_n}) = l_{u_{n+1}}$ provided F_{u_n} is increasing and $F^{-1}_{u_n}(l_{u_n}) = r_{u_{n+1}}$ provided F_{u_n} is decreasing. Similarly, if $x_n \in W^{\varepsilon+}_{u_n}$, then $F^{-1}_{u_n}(r_{u_n}) = r_{u_{n+1}}$ provided F_{u_n} is increasing and $F^{-1}_{u_n}(r_{u_n}) = l_{u_{n+1}}$ provided F_{u_n} is decreasing. It follows that there exists an open interval I whose one endpoint is x, such that $F^{-1}_{u_{[0,n)}}(I) \cap W_{u_n}$ is a nonempty interval for each n. If $F^{-1}_{u_{[0,n)}}(I) \subseteq W^o_{u_n}$, then $|F^{-1}_{u_{[0,n+1)}}(I)| \ge \psi^{-1}(|F^{-1}_{u_{[0,n)}}(I)|)$, so there exists $c > 0$ such that $|F^{-1}_{u_{[0,n)}}(I)| > c$ for all sufficiently large n. By Theorem 3.41, $\Phi(u) = x$, so we have proved $\mathcal{S}_{F,W} \subseteq \mathbb{X}_F$ and $\{\Phi(u)\} = \bigcap_{n>0} \overline{W_u}$. For each $u \in \mathcal{S}_{F,W}$ and $n > 0$ we have $\Phi(u) \in \overline{W_{u_{[0,n)}}}$, so for each $u \in \mathcal{L}_{F,W}$ we have $\Phi([u]) \subseteq \overline{W_u}$. Conversely, if $x \in \overline{W_u}$ then there exists $a \in A$ such that $F_u^{-1}(x) \in \overline{W_a}$ and $F_u^{-1}(W_u) \cap W_a \ne \emptyset$, so $W_{ua} \ne \emptyset$ and $x \in \overline{W_{ua}}$. It follows that we can extend u to an infinite word $v \in [u]$ such that $x \in \overline{W_{v_{[0,m)}}}$ for each m, so $x = \Phi(v)$. Thus we have proved $\Phi([u]) = \overline{W_u}$. This works also for $W_\lambda = \overline{\mathbb{R}}$, so $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is surjective. Since $\lim_{n\to\infty} |\Phi([u_{[0,n)}]| = 0$, $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is continuous, so $(F, \mathcal{S}_{F,W})$ is a number system. If $\{W_a : a \in A\}$ is a cover of $\overline{\mathbb{R}}$, then $\{\mathrm{int}_{\overline{W_u}}(\overline{W_{ua}}) : ua \in \mathcal{L}_{F,W}\}$ is a cover of $\overline{W_u}$ for every $u \in \mathcal{L}_{F,W}$: If $x \in \overline{W_u}$, then there exists $a \in A$ such that $F_u^{-1}(x) \in W_a$, so $x \in W_u \cap F_u(W_a) = W_{ua}$ and $ua \in \mathcal{L}_{F,W}$. By Theorem 2.27, $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is redundant. □

The system of **symmetric continued fractions** of Definition 1.14 is a number system according to Theorem 4.12 since $W_a \subseteq \mathbf{V}(F_a)$ (see Figure 4.3). The circle derivations of the inverse transformations F_a^{-1} can be seen in Figure 4.3 left. Since $F_{\overline{1}}^{-1}(W_{\overline{1}}) = F_{\overline{0}}^{-1}(W_{\overline{0}}) = (\infty, 0)$,

a	F_a	W_a	$\mathbf{V}(F_a)$	$F_a^{-1}(W_a)$
$\overline{1}$	$\left[\begin{smallmatrix}1 & -1\\ 0 & 1\end{smallmatrix}\right]$	$(\infty,-1)$	$(\infty,-\frac{1}{2})$	$(\infty,0)$
$\overline{0}$	$\left[\begin{smallmatrix}1 & 0\\ -1 & 1\end{smallmatrix}\right]$	$(-1,0)$	$(-2,0)$	$(\infty,0)$
0	$\left[\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right]$	$(0,1)$	$(0,2)$	$(0,\infty)$
1	$\left[\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right]$	$(1,\infty)$	$(\frac{1}{2},\infty)$	$(0,\infty)$

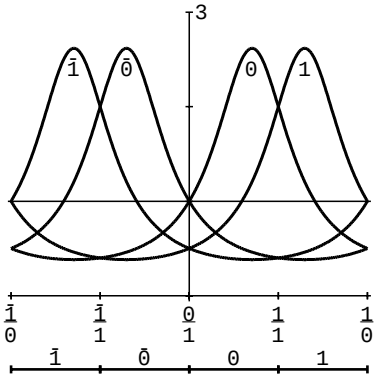


Figure 4.3: The system of symmetric continued fractions from Definition 1.14 and Example 4.6 (left) and the circle derivations of is inverse transformations $(F_a^{-1})^\bullet$ (right).

$F_0^{-1}(W_0) = F_1^{-1}(W_1) = (0,\infty)$, $\mathcal{S}_{F,W} = \{0,1\}^\omega \cup \{\overline{0},\overline{1}\}^\omega$ is a SFT Σ_D with forbidden words $D = \{\overline{0}0, \overline{0}1, \overline{1}0, \overline{1}1, 0\overline{1}, 0\overline{0}, 1\overline{1}, 1\overline{0}\}$. Note that $0 \in \overline{W_{\overline{0}}} \cap \overline{W_0}$ is a fixed point of both $F_{\overline{0}}$ and F_0. If the intervals W_a were assumed closed, any sequence in $\{\overline{0},0\}^\omega$ would be an expansion of 0. With open W_a, the only expansions of 0 are 0^ω and $\overline{0}^\omega$.

For the system of signed continued fractions from Example 4.5, Theorem 4.12 cannot be applied since $F_0(x) = -1/x$ is a rotation, and $\mathbf{V}(F_0) = \emptyset$. However, for the words of length 2 we get $W_u \subseteq \mathbf{V}(F_u)$ (see Figure 4.4). In the next Theorem 4.13 we show that a number system is obtained in this case also.

Theorem 4.13 *Let $F = \{F_a \in \mathbb{M}(\mathbb{R}) : a \in A\}$ be an iterative system and $W = \{W_a \subset \overline{\mathbb{R}} : a \in A\}$ an open almost-cover. Assume that there exists $n \geq 1$ such that $W_u \subseteq \mathbf{V}(F_u)$ for each $u \in \mathcal{L}_{F,W}^n$. Then $(F, \mathcal{S}_{F,W})$ is a number system and $\Phi([u]) = \overline{W_u}$ for each $u \in \mathcal{L}_{F,W}$. If $\{W_a : a \in A\}$ is a cover of $\overline{\mathbb{R}}$ then $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is redundant.*

Proof: Consider the alphabet $B = \mathcal{L}_{F,W}^n$ and the iterative system G over B given by $G_u = F_u$. Then $V = \{W_u : u \in B\}$ is an open almost-cover, so $(G, \mathcal{S}_{G,V})$ is a number system by Theorem 4.12. Given $u \in A^\omega$, define $\tilde{u} \in B^{\mathbb{N}}$ by $\tilde{u}_k = u_{[kn,(k+1)n)}$. If $u \in \mathcal{S}_{F,W}$, then $\tilde{u} \in \mathcal{S}_{G,V}$, so $\lim_{k\to\infty} F_{u_{[0,kn)}}(z) = \Phi_G(\tilde{u})$ for any $z \in \mathbb{U}$. In particular the condition is satisfied for each $z = F_v(i)$, where $v \in A^+$, $|v| < n$. If $kn \leq j < k(n+1)$, then $F_{u_{[0,j)}}(i) = F_{u_{[0,kn)}} F_{u_{[kn,j)}}(i)$, so $\lim_{j\to\infty} F_{u_{[0,j)}}(i) = \Phi_G(\tilde{u})$, and $\Phi_F(u) = \Phi_G(\tilde{u})$. Thus $\mathcal{S}_{F,W} \subseteq \mathbb{X}_F$ and $\Phi : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is continuous, since $\lim_{n\to\infty} |W_{u_{[0,n)}}| = 0$. Since V is an almost-cover, $\Phi_F : \mathcal{S}_{F,W} \to \overline{\mathbb{R}}$ is surjective. By Theorem 4.12 we get $\Phi_F([u]) = \Phi_G(\tilde{u}) = \overline{V_{\tilde{u}}} = \overline{W_u}$ for each $u \in \mathcal{L}_{F,W}$. If $\{W_a : a \in A\}$ is a cover of $\overline{\mathbb{R}}$, then $\{\mathrm{int}_{\overline{W_u}}(\overline{W_{ua}}) : ua \in \mathcal{L}_{F,W}\}$ is a cover of $\overline{W_u}$ for each $u \in \mathcal{L}_{F,W}$, so Φ_F is redundant. □

Definition 4.14 *We say that (F,W) is an* **interval number system** *of order $n \geq 1$ over an alphabet A, if $F = \{F_a \in \mathbb{M}(\mathbb{R}) : a \in A\}$ is an iterative system, $W = \{W_a \subset \overline{\mathbb{R}} : a \in A\}$ is an open almost-cover and $W_u \subseteq \mathbf{V}(F_u)$ for each $u \in \mathcal{L}_{F,W}^n$. We say that (F,W) is redundant, if $\{W_a : a \in A\}$ is a cover of $\overline{\mathbb{R}}$.*

The system of signed continued fractions is an interval number system of order 2 with intervals $W_{\overline{1}} = (\infty,-1)$, $W_0 = (-1,1)$, $W_1 = (1,\infty)$. The ternary signed system from Proposition 1.6 and Example 4.8 is an interval number system of order 4. The following Theorem 4.15 is a partial answer to the question whether for a given iterative system there exists a subshift which forms with it a number system.

b	u	F_u	W_u	$\mathbf{V}(F_u)$
$\overline{2}$	$\overline{11}$	$\left[\begin{smallmatrix}1 & -2\\ 0 & 1\end{smallmatrix}\right]$	$(\infty,-2)$	$(\infty,-1)$
$\overline{1}$	$\overline{1}0$	$\left[\begin{smallmatrix}1 & 1\\ -1 & 0\end{smallmatrix}\right]$	$(-2,-1)$	$(\infty,-\frac{1}{2})$
$\overline{0}$	01	$\left[\begin{smallmatrix}0 & 1\\ -1 & -1\end{smallmatrix}\right]$	$(-1,0)$	$(-2,0)$
0	$0\overline{1}$	$\left[\begin{smallmatrix}0 & 1\\ -1 & 1\end{smallmatrix}\right]$	$(0,1)$	$(0,2)$
1	10	$\left[\begin{smallmatrix}-1 & 1\\ -1 & 0\end{smallmatrix}\right]$	$(1,2)$	$(\frac{1}{2},\infty)$
2	11	$\left[\begin{smallmatrix}1 & 2\\ 0 & 1\end{smallmatrix}\right]$	$(2,\infty)$	$(1,\infty)$

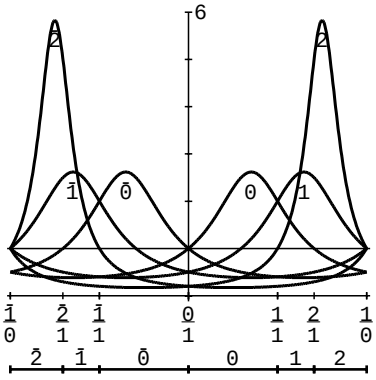


Figure 4.4: The second iteration of the system of signed continued fractions of Example 4.5 with alphabet $B = \{\overline{2},\overline{1},\overline{0},0,1,2\} = \{\overline{11},\overline{1}0,01,0\overline{1},10,11\}$.

Theorem 4.15 (Kůrka [37]) *Let $F = \{F_a \in \mathbb{M}(\mathbb{R}) : a \in A\}$ be an iterative system.*

1. *If there exists a finite set $B \subseteq A^+$ such that $\{\overline{\mathbf{V}(F_u)} : u \in B\}$ is a cover of $\overline{\mathbb{R}}$, then $\Phi(\mathbb{X}_F) = \overline{\mathbb{R}}$ and there exists a subshift $\Sigma \subseteq A^\omega$ such that (F,Σ) is a number system.*
2. *If $\overline{\bigcup_{u\in A^+} \mathbf{V}(F_u)} \neq \overline{\mathbb{R}}$ then $\Phi(\mathbb{X}_F) \neq \overline{\mathbb{R}}$, so there exists no number system with the iterative system F.*

Proof: Item 1 is a consequence of Theorem 4.13. If x does not belong to the closure of the union of all $\mathbf{V}(F_u)$, then there exists an open interval I which contains x and is disjoint from all $\mathbf{V}(F_u)$. Given $u \in A^\omega$, then for each n we have $|F^{-1}_{u_{[0,n)}}(I)| \leq |I|$, so $F_{u_{[0,n)}}(i)$ cannot converge to x. Thus $x \notin \Phi(\mathbb{X}_F)$. □

To find an expansion of $x \in \overline{\mathbb{R}}$ in an interval number system (F,W), we find u_0 with $x \in \overline{W_{u_0}}$ and repeat the procedure with $x_1 = F^{-1}_{u_0}(x)$. However, if some $x_n = F^{-1}_{u_{[0,n)}}(x)$ is an endpoint of W_{u_n}, then we are constrained in the choice of further u_m with $m > n$: if $F_{u_{[n,m)}}$ is increasing and $x_n = \mathfrak{l}(W_{u_n})$ then x_m cannot be $\mathbf{r}(W_{u_m})$ since we would get $W_{u_{[n,m)}} = \emptyset$. This is why during the expansion process we should keep information whether an endpoint of W_{u_i} has been visited. For $u \in A^*$ denote by $\mathbf{o}(u) \in \{-1,1\}$ the **orientation** of F_u, so $\mathbf{o}(u) = -1$ if F_u is decreasing and $\mathbf{o}(u) = +1$ if F_u is increasing.

Definition 4.16 *For an interval number system (F,W), define the* **expansion graph** *with vertices $(x,s) \in \overline{\mathbb{R}} \times \{-1,0,+1\}$ and labelled edges*

$$\begin{aligned}
(x,s) &\xrightarrow{a} (F_a^{-1}(x), s\cdot\mathbf{o}(a)), \text{ if } x \in W_a,\\
(x,s) &\xrightarrow{a} (F_a^{-1}(x), -\mathbf{o}(a)), \text{ if } x = \mathfrak{l}(W_a),\ s \leq 0,\\
(x,s) &\xrightarrow{a} (F_a^{-1}(x), +\mathbf{o}(a)), \text{ if } x = \mathbf{r}(W_a),\ s \geq 0.
\end{aligned}$$

Proposition 4.17 *Let (F,W) be an interval number system, $x \in \overline{\mathbb{R}}$, $u \in A^\omega$. Then $u \in \mathcal{S}_{F,W}$ and $\Phi(u) = x$ iff u is the label of a path with source $(x,0)$.*

Proof: Let u be the label of a path $(x,0) \xrightarrow{u_0} (x_1,s_1) \xrightarrow{u_1} \cdots$, so $x_n = F^{-1}_{u_{[0,n)}} \in \overline{W_{u_n}}$. If $s_n = 0$ for all n, then $x_n \in W_{u_n}$ and $x \in W_{u_{[0,n)}} \neq \emptyset$. Thus $u \in \mathcal{S}_{F,W}$ and $\Phi(u) = x$. Let n be the first integer with $s_{n+1} \neq 0$, so $x_n \notin W_{u_n}$. Then $x \in W_{u_{[0,n)}} \cap F_{u_{[0,n}}(\overline{W_{u_n}})$, so $W_{u_{[0,n+1)}} \neq \emptyset$ If $x_n = \mathfrak{l}(W_{u_n})$ and $\mathbf{o}(u_n) = -1$ then $s_{n+1} = +1$ and $x_{n+1} \neq \mathfrak{l}(W_{u_{n+1}})$, since otherwise no edge

would lead out of (x_{n+1}, s_{n+1}). This implies $W_{u_{[0,n+2)}} \neq \emptyset$. By induction we show that for each $m > n$ we have

$$\begin{aligned} \mathbf{o}(u_{[n,m)}) = -1 \quad &\Rightarrow \quad s_m = +1,\ x_m \neq \mathfrak{l}(W_{u_m}), \\ \mathbf{o}(u_{[n,m)}) = +1 \quad &\Rightarrow \quad s_m = -1,\ x_m \neq \mathbf{r}(W_{u_m}). \end{aligned}$$

In both cases we get $W_{[0,m]} \neq 0$, and $x \in \overline{W_{[0,m]}}$. Thus $u \in \mathcal{S}_{F,W}$ and $\Phi(u) = x$. If $x_n = \mathbf{r}(W_{u_n})$, the proof is analogous. Conversely if $u \in \mathcal{S}_{F,W}$, $\Phi(u) = x$, then $x_n \in \overline{W_{u_n}}$. If $x_n \in W_{u_n}$ for each n, then we get a path with $s_n = 0$. Let n be the first index such that $x_n \notin W_{u_n}$, say $x_n = \mathfrak{l}(W_{u_n})$. Given $m > n$ then $W_{u_n} \cap F_{[n,m)}(W_{u_m}) \neq \emptyset$ so we get

$$\begin{aligned} \mathbf{o}(u_{[n,m)}) = -1 \quad &\Rightarrow \quad x_m \neq \mathfrak{l}(W_{u_m}), \\ \mathbf{o}(u_{[n,m)}) = +1 \quad &\Rightarrow \quad x_m \neq \mathbf{r}(W_{u_m}). \end{aligned}$$

In the former case we set $s_m = +1$, in the latter case we set $s_m = -1$. This defines an infinite path with label u. If $x_n = \mathbf{r}(W_{u_n})$, the proof is analogous. □

4.3 Partition number systems

If the intervals W_a do not overlap, then we get an order on $\mathcal{S}_{F,W}$ which corresponds to the order on $\mathbb{R}$. We say that an open almost-cover $W = \{W_a \subset \mathbb{R} :\ a \in A\}$ is an **open partition** if $W_a \cap W_b = \emptyset$ for $a \neq b$. An open partition is uniquely specified by its set of **cutpoints**

$$\mathcal{E}(W) = \{\mathfrak{l}(W_a) :\ a \in A\} = \{\mathbf{r}(W_a) :\ a \in A\}.$$

Definition 4.18 *We say that an interval number system (F, W) is a* **partition number system***, if W is an open partition and for each $a \in A$ we have $\infty \notin W_a$ and $\infty \notin F_a^{-1}(W_a)$.*

Examples of partition number systems are the system of simple continued fractions of Definition 1.13 or the system of symmetric continued fractions of Definition 1.14. The system of signed continued fractions of Example 4.5 does not comply with Definition 4.18 since $\infty \in F_0^{-1}(W_0)$. However it can be modified to a partition number system if we take the alphabet $A = \{\overline{1}, \overline{0}, 0, 1\}$ with transformations $F_0(x) = F_{\overline{0}}(x) = -1/x$, $W_{\overline{0}} = (-1, 0)$, $W_0 = (0, 1)$.

When we work with partition number systems it is convenient to distinguish two infinities $-\infty = \frac{-1}{0}$ and $+\infty = \frac{1}{0}$ with the order on $\mathbb{R}$ extended by $-\infty < x < +\infty$ for every $x \in \mathbb{R}$. Assume that the alphabet $A = \{0, 1, \ldots, s\}$ of a partition number system respects the order on $\mathbb{R}$. This means that for the endpoints $l_a = \mathfrak{l}(W_a)$, $r_a = \mathbf{r}(W_a)$ we have

$$-\infty = l_0 < r_0 = l_1 < r_1 = l_2 < \cdots < r_{s-1} = l_s < r_s = +\infty.$$

We define the order $\prec$ on $\mathcal{S}_{F,W}$ by

$$\begin{aligned} u \prec v \quad \Leftrightarrow \quad &\exists n, u_{[0,n)} = v_{[0,n)},\ u_n < v_n,\ \mathbf{o}(u_{[0,n)}) = +1, \text{ or} \\ &\exists n, u_{[0,n)} = v_{[0,n)},\ u_n > v_n,\ \mathbf{o}(u_{[0,n)}) = -1, \end{aligned}$$

where $\mathbf{o}(u) = +1$ if F_u is increasing and $\mathbf{o}(u) = -1$ if F_u is decreasing. For $u = \lambda$ we set $\mathbf{o}(\lambda) = +1$, so $u_0 < v_0$ implies $u \prec v$. We write $u \preceq v$ if $u \prec v$ or $u = v$. Both inequalities $\prec$ and $\preceq$ are defined analogously between finite words of the same length. If $u, v \in \mathcal{S}_{F,W}$ and $u \preceq v$

then $u_{[0,n)} \preceq v_{[0,n)}$ for each n. By Proposition 4.17, each $x \in \overline{\mathbb{R}}$ has at most two expansions. If $u \in \mathcal{S}_{F,W}$, $\Phi(u) = x$ and n is the first index such that $x_n = F^{-1}_{u_{[0,n)}}(x)$ is a cutpoint of the partition then we have two possibilities for u_n but all u_m with $m > n$ are determined uniquely. We denote the two expansions by $\mathcal{E}_-(x)$ and $\mathcal{E}_+(x)$ and distinguish them by the requirement

$$\mathcal{E}_+(\infty) \prec \mathcal{E}_-(\infty),\ \mathcal{E}_-(x) \preceq \mathcal{E}_+(x) \text{ for } x \in \mathbb{R}$$

If the orbit of x never visits any cutpoint then x has a unique expansion $\mathcal{E}(x) = \mathcal{E}_-(x) = \mathcal{E}_+(x)$. Thus $\mathcal{E}_-(r_a)_0 = a = \mathcal{E}_+(l_a)_0$, in particular $\mathcal{E}_-(\infty)_0 = s$, $\mathcal{E}_+(\infty)_0 = 0$. If $u = \mathcal{E}_-(x)$ and $x_i = F^{-1}_{u_{[0,i)}}(x)$, then either $x_i \in W_{u_i}$ or $x_i = r_{u_i}$ provided $\mathbf{o}(u_{[0,i)}) = +1$ or $x_i = l_{u_i}$ provided $\mathbf{o}(u_{[0,i)}) = -1$. For $v = \mathcal{E}_+(x)$ and $x_i = F^{-1}_{v_{[0,i)}}(x)$ we have either $x_i \in W_{v_i}$ or $x_i = \mathfrak{l}(W_{v_i})$ provided $\mathbf{o}(v_{[0,i)}) = +1$ or $x_i = \mathfrak{r}(W_{v_i})$ provided $\mathbf{o}(v_{[0,i)}) = -1$. It follows that if $u \in \mathcal{L}^n_{F,W}$, and $x \in W_u$ then

$$\mathcal{E}(x)_{[0,n)} = \mathcal{E}_-(\mathfrak{r}(W_u))_{[0,n)} = \mathcal{E}_+(\mathfrak{l}(W_u))_{[0,n)} = u.$$

Examples of expansions in partition number systems can be seen in Figures 1.8 or 1.9. If $x, y \in W_a$, $x < y$ and $\mathbf{o}(a) = +1$ then $F_a^{-1}(x) < F_a^{-1}(y)$. This follows from the assumption $\infty \notin W_a$, $\infty \notin F_a^{-1}(W_a)$. By induction we get for any $u \in \mathcal{L}_{F,W}$

$$\begin{aligned} x, y \in W_u, x < y, \mathbf{o}(u) = +1 \quad &\Rightarrow \quad F_u^{-1}(x) < F_u^{-1}(y) \\ x, y \in W_u, x < y, \mathbf{o}(u) = -1 \quad &\Rightarrow \quad F_u^{-1}(x) > F_u^{-1}(y) \end{aligned}$$

Proposition 4.19 *Let (F, W) be a partition number system and $x, y \in \mathbb{R}$.*

1. *If $x < y$ then $\mathcal{E}_+(x) \prec \mathcal{E}_-(y)$.*
2. *$\mathcal{E}_+(\infty) \prec \mathcal{E}_-(x) \preceq \mathcal{E}_+(x) \prec \mathcal{E}_-(\infty)$.*
3. *If $\mathcal{E}_-(x) \prec \mathcal{E}_-(y)$ or $\mathcal{E}_+(x) \prec \mathcal{E}_+(y)$ then $x < y$.*
4. *If $u \in \mathcal{L}^n_{F,W}$ and $\mathcal{E}_+(x)_{[0,n)} \preceq u$ then $x < \mathfrak{r}(W_u)$.*
5. *If $u \in \mathcal{L}^n_{F,W}$ and $u \preceq \mathcal{E}_-(x)_{[0,n)}$ then $\mathfrak{l}(W_u) < u$.*

Proof: 1. If $x < y$ then $u = \mathcal{E}_+(x) \neq \mathcal{E}_-(y) = v$. Let n be the first integer such that $u_n \neq v_n$. If $\mathbf{o}(u_{[0,n)}) = +1$ then $x_n = F^{-1}_{u_{[0,n)}}(x) < F^{-1}_{u_{[0,n)}}(y) = y_n$, so $u_n < v_n$ and $u \prec v$. If $\mathbf{o}(u_{[0,n)}) = -1$ then $x_n > y_n$, so $u_n > v_n$ and $u \prec v$.
2. With the convention $-\infty < x < +\infty$, the argument of the preceding proof works for $-\infty < x$ and $x < +\infty$.
3. If $\mathcal{E}_-(x) \prec \mathcal{E}_-(y)$, then $x \neq y$. From $y < x$ we would get $\mathcal{E}_+(y) \prec \mathcal{E}_-(x) \prec \mathcal{E}_-(y)$ which is a contradiction. Thus $x < y$. The proof is similar if $\mathcal{E}_+(x) \prec \mathcal{E}_+(y)$.
4. Assume by contradiction $\mathfrak{r}(W_u) \leq x$. Then $u = \mathcal{E}_-(\mathfrak{r}(W_u))_{[0,n)} \prec \mathcal{E}_+(\mathfrak{r}(W_u))_{[0,n)} \preceq \mathcal{E}_+(x)_{[0,n)}$ which is a contradiction.
5. If $x \leq \mathfrak{l}(W_u)$ then $\mathcal{E}_-(x)_{[0,n)} \preceq \mathcal{E}_-(\mathfrak{l}(W_u))_{[0,n)} \prec \mathcal{E}_+(\mathfrak{l}(W_u))_{[0,n)} = u$ which is a contradiction.
□

The language of the subshift $\mathcal{S}_{F,W}$ is determined by the expansions of the endpoints l_a, r_a of W_a. Before the proof of the next theorem note that open intervals $I, J \subseteq \mathbb{R}$ have nonempty intersection iff $\max\{\mathfrak{l}(I), \mathfrak{l}(J)\} < \min\{\mathfrak{r}(I), \mathfrak{r}(J)\}$.

Theorem 4.20 *Let (F, W) be a partition number system. Then*

1. *$u \in A^+$ belongs to $\mathcal{L}_{F,W}$ iff $\mathcal{E}_+(l_{u_n})_{[0,|u|-n)} \preceq \sigma^n(u) \preceq \mathcal{E}_-(r_{u_n})_{[0,|u|-n)}$ for each $n < |u|$.*
2. *$u \in A^\omega$ belongs to $\mathcal{S}_{F,W}$ iff $\mathcal{E}_+(l_{u_n}) \preceq \sigma^n(u) \preceq \mathcal{E}_-(r_{u_n})$ for each n.*

Here we set $\sigma^n(u) = u_{[n,|u|)}$ for $u \in A^+$.

Proof: 1. Assume that $u \in \mathcal{L}_{F,W}$ and choose some $x \in W_u$. For $n < m = |u|$ we have $x_n = F^{-1}_{[0,n)}(x) \in W_{\sigma^n(u)} \subseteq W_{u_n}$ and either $\sigma^n(u) = \mathcal{E}_-(x_n)_{[0,|u|-n)}$ or $\sigma^n(u) = \mathcal{E}_+(x_n)_{[0,|u|-n)}$. By Proposition 4.19 it follows $\mathcal{E}_+(l_{u_n})_{[0,|u|-n)} \preceq \sigma^n(u) \preceq \mathcal{E}_-(r_{u_n})_{[0,|u|-n)}$. Conversely assume that $u \in A^m$ satisfies the condition. If $m = 1$ then $u \in \mathcal{L}_{F,W}$ is trivial. Assume that the statement is true for all v with $|v| < m$. Since $|\sigma(u)| < m$, we get $\sigma(u) \in \mathcal{L}_{F,W}$. By the assumption with $n = 0$ there exist $v, w \in A^{m-1}$ such that $u_0 v = \mathcal{E}_+(l_{u_0})_{[0,m)} \preceq u \preceq \mathcal{E}_-(r_{u_0})_{[0,m)} = u_0 w$. We consider two cases. If $\mathbf{o}(u_0) = +1$ then by Proposition 4.19 we get

$$\mathcal{E}_+(F^{-1}_{u_0}(l_{u_0}))_{[0,m-1)} = v \preceq \sigma(u) \preceq w = \mathcal{E}_-(F^{-1}_{u_0}(r_{u_0}))_{[0,m-1)},$$

so $F^{-1}_{u_0}(l_{u_0}) < \mathbf{r}(W_{\sigma(u)})$, $\mathfrak{l}(W_{\sigma(u)}) < F^{-1}_{u_0}(r_{u_0})$. Since $F^{-1}_{u_0}(l_{u_0}) < F^{-1}_{u_0}(r_{u_0})$, we have

$$\max\{F^{-1}_{u_0}(l_{u_0}), \mathfrak{l}(W_{\sigma(u)})\} < \min\{F^{-1}_{u_0}(r_{u_0}), \mathbf{r}(W_{\sigma(u)})\}.$$

It follows $W_{\sigma(u)} \cap F^{-1}_{u_0}(W_{u_0}) \neq \emptyset$, so $W_u \neq \emptyset$ and $u \in \mathcal{L}_{F,W}$. If $\mathbf{o}(u_0) = -1$ then

$$\mathcal{E}_-(F^{-1}_{u_0}(l_{u_0}))_{[0,m-1)} = v \succeq \sigma(u) \succeq w = \mathcal{E}_+(F^{-1}_{u_0}(r_{u_0}))_{[0,m-1)},$$

so $F^{-1}_{u_0}(l_{u_0}) > \mathfrak{l}(W_{\sigma(u)})$, $\mathbf{r}(W_{\sigma(u)}) > F^{-1}_{u_0}(r_{u_0})$. Since $F^{-1}_{u_0}(l_{u_0}) > F^{-1}_{u_0}(r_{u_0})$, we get

$$\max\{F^{-1}_{u_0}(r_{u_0}), \mathfrak{l}(W_{\sigma(u)})\} < \min\{F^{-1}_{u_0}(l_{u_0}), \mathbf{r}(W_{\sigma(u)})\}.$$

It follows that $W_{\sigma(u)} \cap F^{-1}_{u_0}(W_{u_0}) \neq \emptyset$, so $W_u \neq \emptyset$ and $u \in \mathcal{L}_{F,W}$.

2. is an immediate consequence of 1. □

4.4 Sofic expansion subshifts

We characterize interval number systems whose expansion subshifts are of finite type or sofic.

Theorem 4.21 (Kůrka [39]) *Let (F, W) be an interval number system. Then $\mathcal{S}_{F,W}$ is an SFT of order $m+1$ iff*

$$\forall a \in A, \forall u \in \mathcal{L}^m_{F,W}, (W_u \cap F_a^{-1}(W_a) \neq \emptyset \;\Rightarrow\; W_u \subseteq F_a^{-1}(W_a))$$

In this case $W_u = F_{u_{[0,n-m]}}(W_{u_{(n-m,n]}})$ for each $u \in \mathcal{L}_{F,W}$ with $|u| = n+1 > m$.

Proof: The condition can be equivalently stated in the form that $F_a(W_u) \cap W_a \neq \emptyset$ implies $F_a(W_u) \subseteq W_a$. Let $u \in A^{n+1}$, and assume that $u_{[i,i+m]} \in \mathcal{L}_{F,W}$ for all $i < n - m$. Then $\emptyset \neq W_{u_{[0,m]}} = W_{u_0} \cap F_{u_0}(W_{u_{[1,m]}})$, so $F_{u_0}(W_{u_{[1,m]}}) \subseteq W_{u_0}$ and $W_{u_{[0,m]}} = F_{u_0}(W_{u_{[1,m]}})$. It follows

$$\begin{aligned} W_u &= W_{u_{[0,m]}} \cap F_{u_{[0,m]}}(W_{u_{[m+1,n]}}) \\ &= F_{u_0}(W_{u_{[1,m]}}) \cap F_{u_{[0,m]}}(W_{u_{[m+1,n]}}) \\ &= F_{u_0}(W_{u_{[1,n]}}) = F_{u_{[0,1]}}(W_{u_{(2,n]}}) = \cdots \\ &= F_{u_{[0,n-m]}}(W_{u_{(n-m,n]}}) \neq \emptyset. \end{aligned}$$

Thus $u \in \mathcal{L}_{F,W}$, so we have proved that $\mathcal{S}_{F,W}$ is an SFT of order $m+1$. Conversely, assume that the condition is not satisfied, so let $a \in A$, $u \in \mathcal{L}^m_{F,W}$ be such that $W_u \cap F_a^{-1}(W_a) \neq \emptyset$ but $W_u \not\subseteq F_a^{-1}(W_a)$, so $W_u \setminus F_a^{-1}(W_a) \neq \emptyset$. Since $\lim_{n\to\infty} \max\{|W_v| : v \in \mathcal{L}^m_{F,W}\} = 0$, there exists

$v \in \mathcal{L}_{\mathcal{W}}$ such that $W_v \subseteq F_u^{-1}(W_u) \setminus F_{au}^{-1}(W_a) = F_u^{-1}(W_u \setminus F_a^{-1}(W_a))$, so $F_u(W_v) \subseteq W_u$ and $F_{au}(W_v) \cap W_a = \emptyset$. It follows $W_{uv} = W_u \cap F_u(W_v) = F_u(W_v) \neq \emptyset$ but $W_{auv} = W_a \cap F_a(W_u) \cap F_{au}(W_v) = \emptyset$. Thus $au \in \mathcal{L}_{F,W}$, $uv \in \mathcal{L}_{F,W}$, and $auv \notin \mathcal{L}_{F,W}$, so $\mathcal{S}_{F,W}$ is not an SFT of order $m+1$. □

The condition of Theorem 4.21 means that each endpoint of $F_a^{-1}(W_a)$ is an endpoint of some W_u, where $u \in \mathcal{L}_{F,W}^m$. In particular $\mathcal{S}_{F,W}$ is an SFT of order 2 iff each endpoint of $F_a^{-1}(W_a)$ is an endpoint of some W_b. In this case we have $W_u = F_{u_{[0,n)}}(W_{u_n})$ for each $u \in A^{n+1}$.

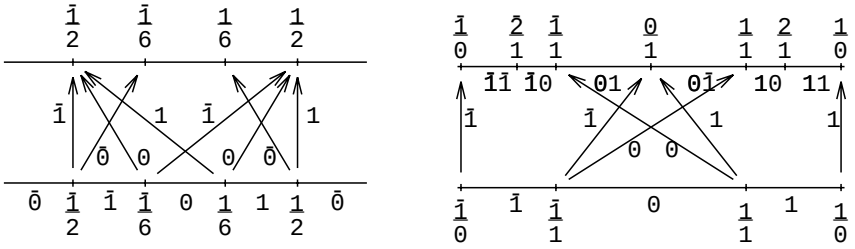


Figure 4.5: Expansion subshifts of finite type: the ternary system of Example 4.4 of order 2 (left) and the system of signed continued fractions of Example 4.5 of order 3 (right).

The ternary signed system (F, Σ_D) of Proposition 1.6 and Example 4.4 has the expansion subshift of order 2, since each endpoint of $F_a^{-1}(W_a)$ is an endpoint of some W_b (see Figure 4.5 left). Indeed $F_{\overline{1}}^{-1}(W_{\overline{1}}) = F_0^{-1}(W_0) = F_1^{-1}(W_1) = (\frac{-1}{2}, \frac{1}{2})$, $F_{\overline{0}}^{-1}(W_{\overline{0}}) = (\frac{1}{6}, \frac{1}{-6})$. The system of signed continued fractions from Example 4.5 has the expansion subshift of order 3: Each endpoint of W_u with $|u| = 2$ is an endpoint of some W_a (see Figure 4.5 right).

Theorem 4.22 *If (F, W) is an interval number system and $\{W_a : a \in A\}$ is a cover of $\overline{\mathbb{R}}$, then $\mathcal{S}_{F,W}$ is not an SFT.*

Proof: By the assumption, $\{W_u : u \in \mathcal{L}_{F,W}^n\}$ is a cover for each n. If x is an endpoint of some $F_a^{-1}(W_a)$, and $m > 0$ then there exists $u \in \mathcal{L}_{F,W}^m$ with $x \in W_u$, so $W_u \cap F_a^{-1}(W_a) \neq \emptyset$ but $W_u \not\subseteq F_a^{-1}(W_a)$. Thus $\mathcal{S}_{F,W}$ is not an SFT of order $m+1$. □

Thus interval number systems whose expansion subshifts are of finite type cannot be redundant.

Theorem 4.23 (Kůrka [39]) *Let (F, W) be an interval number system with alphabet A. Then the expansion subshift $\mathcal{S}_{F,W}$ is sofic iff there exists an open partition $V = \{V_p \subseteq \overline{\mathbb{R}} : p \in B\}$ such that if $F_a(V_q) \cap V_p \cap W_a \neq \emptyset$, then $F_a(V_q) \subseteq V_p \cap W_a$. In this case, $\mathcal{S}_{F,W}$ is the subshift of the labelled graph $G_{F,W,V}$ with vertices $p \in B$ and labelled edges $p \xrightarrow{a} q \Leftrightarrow F_a(V_q) \subseteq V_p \cap W_a$.*

Proof: Let $V = \{V_p : p \in B\}$ be an open partition with the assumed properties and let $p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_1} \cdots \xrightarrow{u_{n-1}} p_n \xrightarrow{u_n}$ be a path in the graph $G_{F,W,V}$. Then $V_{p_n} \cap W_{u_n} \neq \emptyset$ and for each $k < n$ we have $F_{u_k}(V_{p_{k+1}} \cap W_{u_{k+1}}) \subseteq F_{u_k}(V_{p_{k+1}}) \subseteq V_{p_k} \cap W_{u_k}$. We get

$$\begin{aligned} \emptyset &\neq F_{u_{[0,n)}}(V_{p_n} \cap W_{u_n}) \subseteq F_{u_{[0,n-1)}}(V_{p_{n-1}} \cap W_{u_{n-1}}) \subseteq \cdots \subseteq F_{u_0}(V_{p_1} \cap W_{u_1}) \subseteq V_{p_0} \cap W_{u_0}, \\ \emptyset &\neq F_{u_{[0,n)}}(V_{p_n} \cap W_{u_n}) \subseteq F_{u_{[0,n-1)}}(W_{u_{n-1}}) \cap \cdots \cap F_{u_0}(W_{u_1}) \cap W_{u_0} \subseteq W_{u_{[0,n)}}, \end{aligned}$$

Thus $u_{[0,n)} \in \mathcal{L}_{F,W}$. On the other hand, assume that $W_u \neq \emptyset$ and let us construct a path in the graph with the label u. There exists $p_0 \in B$ such that $\emptyset \neq V_{p_0} \cap W_u$ and there exists p_1 such that

$$\emptyset \neq V_{p_1} \cap F_{u_0}^{-1}(V_{p_0} \cap W_u) \subseteq V_{p_1} \cap F_{u_0}^{-1}(V_{p_0} \cap W_{u_0}).$$

By the assumption, $F_{u_0}(V_{p_1}) \subseteq V_{p_0} \cap W_{u_0}$. Thus we have an edge $p_0 \xrightarrow{u_0} p_1$ and we get $V_{p_1} \cap F_{u_0}^{-1}(W_u) \neq \emptyset$. There exists p_2 such that

$$\emptyset \neq V_{p_2} \cap F_{u_1}^{-1}(V_{p_1} \cap F_{u_0}^{-1}(W_u)) \subseteq V_{p_2} \cap F_{u_1}^{-1}(V_{p_1} \cap W_{u_1}).$$

Thus we have an edge $p_1 \xrightarrow{u_1} p_2$ and $V_{p_2} \cap F_{u_{[0,2)}}^{-1}(W_u) \neq \emptyset$. We continue by induction. Assume that we have constructed $p_k \in B$ with $V_{p_k} \cap F_{u_{[0,k)}}^{-1}(W_u) \neq \emptyset$. Then there exists p_{k+1} such that

$$\emptyset \;\neq\; V_{p_{k+1}} \cap F_{u_k}^{-1}(V_{p_k} \cap F_{u_{[0,k)}}^{-1}(W_u)) \subseteq V_{p_{k+1}} \cap F_{u_k}^{-1}(V_{p_k} \cap W_{u_k}),$$

so we have an edge $p_k \xrightarrow{u_k} p_{k+1}$ and $V_{p_{k+1}} \cap F_{u_{[0,k+1)}}^{-1}(W_u) \neq \emptyset$. We have constructed a path with label u, so we have established that $\mathcal{S}_{F,W}$ is the subshift of the graph $G_{F,W,V}$.
Conversely assume that $\mathcal{S}_{F,W}$ is sofic. Recall that the follower set of a word $u \in \mathcal{L}_{F,W}$ is $\mathcal{F}_u = \{v \in A^\omega : \; uv \in \mathcal{S}_{F,W}\}$. Since $\mathcal{S}_{F,W}$ is sofic, the set $\{\mathcal{F}_u : \; u \in \mathcal{L}_{F,W}\}$ of follower sets is finite. Given $u, v \in \mathcal{L}_{F,W}$, then $\mathcal{F}_u = \mathcal{F}_v$ iff $W_{uw} \neq \emptyset \Leftrightarrow W_{vw} \neq \emptyset$ for any $w \in \mathcal{L}_{F,W}$. This is equivalent to $F_u^{-1}(W_u) \cap W_w \neq \emptyset \Leftrightarrow F_v^{-1}(W_v) \cap W_w \neq \emptyset$ for any $w \in \mathcal{L}_{F,W}$. Since the length of W_w tends to zero as $|w| \to \infty$, we get $\mathcal{F}_u = \mathcal{F}_v$ iff $F_u^{-1}(W_u) = F_v^{-1}(W_v)$, so $\{F_u^{-1}(W_u) : \; u \in \mathcal{L}_{F,W}\}$ is a finite set. Each $F_u^{-1}(W_u)$ is either an open interval or a finite union of open intervals. Denote by $\mathcal{E}$ the finite set of all endpoints of all these intervals and let $\{V_p : \; p \in B\}$ be the open interval partition whose cutpoints are exactly $\mathcal{E}$. Assume that $V_p \cap W_a \neq 0$ and let x be its endpoint. Then x is either an endpoint of W_a or an endpoint of some (interval of) $F_u^{-1}(W_u) \cap W_a$. In the former case, $F_a^{-1}(x)$ is an endpoint of $F_a^{-1}(W_a)$, in the latter case, $F_a^{-1}(x)$ is an endpoint of some interval of

$$F_a^{-1}(W_a \cap F_u^{-1}(W_u)) = F_{ua}^{-1}(F_u(W_a) \cap W_u) = F_{ua}^{-1} W_{ua}$$

Thus in either case, $F_a^{-1}(x) \in \mathcal{E}$, so it is an endpoint of some V_q. This means that if $V_q \cap F_a^{-1}(V_p \cap W_a) \neq \emptyset$ then $V_q^\circ \subseteq F_a^{-1}(V_p \cap W_a)$. Thus we have proved that V satisfies the conditions of the theorem. □

If the conditions of Theorem 4.23 are satisfied, then we say that $V = \{V_p : \; p \in B\}$ is an **open SFT partition** for (F, W).

Theorem 4.24 *A partition number system (F, W) has a sofic subshift $\mathcal{S}_{F,W}$ iff $\mathcal{E}_-(\mathfrak{l}(W_a))$ and $\mathcal{E}_+(\mathfrak{l}(W_a))$ are periodic sequences for each $a \in A$.*

Proof: The condition implies that $\mathcal{E}_-(\mathbf{r}(W_a))$ and $\mathcal{E}_+(\mathbf{r}(W_a))$ are also periodic sequences. If all trajectories of all edpoints of W_a are periodic, then the points of these trajectories form a finite set $\mathcal{E}$ and we define $V = \{V_p : \; p \in B\}$ as the open partition whose endpoints are the points of $\mathcal{E}$. Assume by contradiction that $F_a(V_q) \cap V_p \cap W_a \neq \emptyset$ and $F_a(V_q) \not\subseteq V_p \cap W_a$. Then for one of the endpoints x of $V_p \cap W_a$ we have $F_a^{-1}(x) \in V_q$ and this is a contradiction since x belongs to a trajectory of an endpoint of some W_a. Conversely, let $V = \{V_p : \; p \in B\}$ be an open partition which satisfies the conditions of Theorem 4.23. Assume by contradiction that an endpoint x of some W_a does not have periodic expansion $u = \mathcal{E}_-(x)$ or $u = \mathcal{E}_+(x)$. Let n be the first integer such that $x_{n+1} = F_{u_{[0,n]}}^{-1}(x)$ is not an endpoint of any V_p, so there exists $b \in A$ such that x_n is an endpoint of some $V_p \cap W_b$ and $x_{n+1} \in V_q$ for some $q \in B$. Then $x_n \in F_b(\overline{V_q}) \cap \overline{V_p} \cap \overline{W_b}$, so $F_b(V_q) \cap V_p \cap W_b \neq \emptyset$, but $F_b(V_q) \not\subseteq V_p \cap W_b$ and this is a contradiction. □

Note that the binary signed system from Example 4.3 is not an interval number system. If we take the cover $W_{\overline{0}} = (\frac{1}{2}, \frac{1}{-2})$, $W_{\overline{1}} = (\frac{-1}{1}, \frac{-1}{4})$, $W_0 = (\frac{-1}{2}, \frac{1}{2})$, $W_1 = (\frac{1}{4}, \frac{1}{2})$, then $\mathcal{S}_{F,W} \not\subseteq \mathbb{X}_F$. We obtain an interval number system with another cover:

p	a	V_p	$F_a^{-1}(V_p)$	q
0	$\overline{0}$	$(-3,-2)$	$(-\frac{3}{2},-1)$	2
1	$\overline{0}$	$(-2,-\frac{3}{2})$	$(-1,-\frac{3}{4})$	3
1	$\overline{1}$	$(-2,-\frac{3}{2})$	$(-3,-2)$	0
2	$\overline{1}$	$(-\frac{3}{2},-1)$	$(-2,-1)$	1, 2
3	$\overline{1}$	$(-1,-\frac{3}{4})$	$(-1,-\frac{1}{2})$	3, 4
4	$\overline{1}$	$(-\frac{3}{4},-\frac{1}{2})$	$(-\frac{1}{2},0)$	5
5	$\overline{1}$	$(-\frac{1}{2},0)$	$(0,1)$	6, 7, 8
5	0	$(-\frac{1}{2},0)$	$(-1,0)$	3, 4, 5
6	0	$(0,\frac{1}{2})$	$(0,1)$	6, 7, 8
6	1	$(0,\frac{1}{2})$	$(-1,0)$	3, 4, 5
7	1	$(\frac{1}{2},\frac{3}{4})$	$(0,\frac{1}{2})$	6
8	1	$(\frac{3}{4},1)$	$(\frac{1}{2},1)$	7, 8
9	1	$(1,\frac{3}{2})$	$(1,2)$	9, A
A	1	$(\frac{3}{2},2)$	$(2,3)$	B
A	$\overline{0}$	$(\frac{3}{2},2)$	$(\frac{3}{4},1)$	8
B	$\overline{0}$	$(2,3)$	$(1,\frac{3}{2})$	9
C	$\overline{0}$	$(3,-3)$	$(\frac{3}{2},-\frac{3}{2})$	$A, B, C, 0, 1$

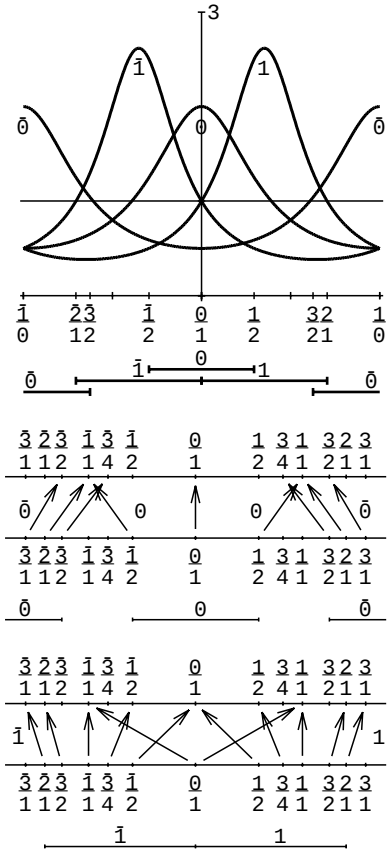


Figure 4.6: The open SFT partition and the labelled graph of the binary signed interval system from Example 4.25

Example 4.25 *The* **binary signed interval system** (F, W) *has alphabet* $A = \{\overline{1}, 0, 1, \overline{0}\}$, *transformations* $F_{\overline{1}}(x) = \frac{x-1}{2}$, $F_0(x) = \frac{x}{2}$, $F_1(x) = \frac{x+1}{2}$, $F_{\overline{0}}(x) = 2x$, *and intervals* $W_{\overline{1}} = (-2, 0)$, $W_0 = (\frac{-1}{2}, \frac{1}{2})$, $W_1 = (0, 2)$, $W_{\overline{0}} = (\frac{3}{2}, \frac{3}{-2})$.

Since $\mathbf{V}(F_{\overline{1}}) = (-2, 0)$, $\mathbf{V}(F_0) = (\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$, $\mathbf{V}(F_1) = (0, 2)$, $\mathbf{V}(F_{\overline{0}}) = (\sqrt{2}, -\sqrt{2})$, we get a number system by by Theorem 4.12. The expansion subshift $\mathcal{S}_{F,W}$ is sofic. Its SFT partition $V = \{V_p : p \in B\}$ has endpoints -3, -2, $\frac{-3}{2}$, -1, $\frac{-3}{4}$, $\frac{-1}{2}$, 0, $\frac{1}{2}$, $\frac{3}{4}$, 1, $\frac{3}{2}$, 2, 3. Since W is a cover, $\mathcal{S}_{F,W}$ is not a SFT by Theorem 4.22. Indeed, each cylinder interval W_{0^n} contains the endpoint 0 of V. The graph $G_{F,W,V}$ is given in Figure 4.6. Each row of the table gives all edges (p, a, q) with source $p \in B$ and label a.

Proposition 4.26 *Let* (F, W) *be an interval number system with sofic expansion subshift* $\mathcal{S}_{F,W}$ *and let* $V = \{V_p : p \in B\}$ *be its open SFT cover. Then*

1. *If* $p \xrightarrow{u} q$ *is a path in* $G_{F,W,V}$, *then* $F_u(V_q) \subseteq V_p \cap W_u$ *and* $F_u(\overline{V_q}) \subseteq \overline{V_p} \cap \overline{W_u}$.
2. $\Phi(\mathcal{F}_p) = \overline{V_p} = \bigcup\{F_a(\overline{V_q}) : p \xrightarrow{a} q\}$.
3. $\Phi([u]) = \overline{W_u} = \bigcup\{F_u(\overline{V_q}) : \xrightarrow{u} q\}$

Proof: 1. By the proof of Theorem 4.23 we have $F_u(V_q) \subseteq V_p \cap W_u$, so $F_u(\overline{V_q}) \subseteq \overline{V_p \cap W_u} \subseteq \overline{V_p} \cap \overline{W_u}$.
2. If $p \xrightarrow{a} q$, then $F_a(V_q) \subseteq V_p$. If $x \in \overline{V_p}$, then there exists $a \in A$ with $x \in \overline{W_a}$ and $V_p \cap W_a \neq \emptyset$. There exists $q \in B$ with $F_a^{-1}(x) \in \overline{V_q}$, so $x \in F_a(\overline{V_q})$ and $p \xrightarrow{a} q$. Thus we have proved $\overline{V_p} = \bigcup\{F_a(\overline{V_q}) : p \xrightarrow{a} q\}$. We show $\Phi(\mathcal{F}_p) = \overline{V_p}$. For $x \in \overline{V_p}$ there exists $p = p_0 \xrightarrow{u_0} p_1$ with $F_{u_0}^{-1}(x) \in \overline{V_{p_1}}$. We continue in this construction and obtain an infinite path $p = p_0 \xrightarrow{u_0} p_1 \cdots$ such that $F_{u_{[0,n)}}^{-1}(x) \in \overline{V_{p_n}}$, so $x \in F_{u_{[0,n)}}(\overline{V_{p_n}}) \subseteq \overline{W_{u_{[0,n)}}}$. Thus $x \in \Phi(\mathcal{F}_p)$, so $\overline{V_p} \subseteq \Phi(\mathcal{F}_p)$.

Conversely, if $x \in \Phi(\mathcal{F}_p)$, let $p = p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_0} \cdots$ be an infinite path with $\Phi(u) = x$. For each n we have $F_{u_{[0,n)}}(V_{p_n}) \subseteq V_{p_0} \cap W_{u_{[0,n)}} \neq \emptyset$. Choose an $x_n \in V_{p_0} \cap W_{u_{[0,n)}}$, then $\lim_{n\to\infty} x_n = x$, so $x \in \overline{V_p}$. Thus we have proved $\Phi(\mathcal{F}_p) = \overline{V_p}$.
3. If $x \in \overline{W_u}$, then we construct a path $p \xrightarrow{u} q$ with $x \in \overline{V_p} \subseteq \overline{W_{u0}}$ similarly as in the proof of Theorem 4.23. The only difference is that we choose at each step p_k with $F^{-1}_{u_{[0,k)}}(x) \in \overline{V_{p_k}} \cap F^{-1}_{u_{[0,k)}}(\overline{W_u})$. The opposite inclusion $\bigcup\{F_u(\overline{V_q}) : \xrightarrow{u} q\} \subseteq \overline{W_u}$ follows from 1. □

4.5 Sofic number systems

Proposition 4.26 has a partial converse. Let (F, Σ) be a number system with a sofic subshift Σ and let $G = (B, E)$ be a labelled graph with $\Sigma = \Sigma_G$. Then the sets $\{\Phi(\mathcal{F}_p) : p \in B\}$ satisfy the same conditions as the sets $\overline{V_p}$ in Proposition 4.26, but they need not be intervals. If they are intervals, then their endpoints can be obtained as Φ-values of periodic paths and these periodic paths are determined by selectors. A **selector** for a labelled graph $G = (B, E)$ is a mapping $K : B \to E$ which selects at each vertex $p \in B$ an outgoing edge $K(p) = (p, a, q)$ with source p. A selector K determines for each vertex $p \in B$ a path $p = p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_0} \cdots$ defined by $p_0 = p$, $K(p_i) = (p_i, u_i, p_{i+1})$. This path is periodic, since there exist $i < j$ with $p_i = p_j$ and then $p_{i+k} = p_{j+k}$, $u_{i+k} = u_{j+k}$ for all $k \geq 0$. We denote by $K^p = u_{[0,i)}(u_{[i,j)})^\omega$ the label of the path of K.

Theorem 4.27 *Let F be an iterative system over A and let $G = (B, E)$ be an A-labelled graph such that (F, Σ_G) is a number system. For $p \in B$ consider the closed sets $V_p = \Phi(\mathcal{F}_p)$. Then*
1. *$V_p = \bigcup\{F_a(V_q) : p \xrightarrow{a} q\}$ for each $p \in B$.*
2. *$\Phi([u]) = \bigcup\{F_u(V_q) : \xrightarrow{u} q\}$ for each $u \in \mathcal{L}_G$.*
3. *If all V_p are intervals then there exist selectors L, R, such that for each $p \in B$ either $V_p = \overline{\mathbb{R}}$ or V_p is a proper closed interval with $\mathfrak{l}(V_p) = \Phi(L^p)$ and $\mathbf{r}(V_p) = \Phi(R^p)$.*

Proof: 1. If $p \xrightarrow{a} q$ and $u \in \mathcal{F}_q$ then $au \in \mathcal{F}_p$ and $F_a(\Phi(u)) = \Phi(au) \in \Phi(\mathcal{F}_p) = V_p$, so $F_a(V_q) \subseteq V_p$. Conversely, if $au \in \mathcal{F}_p$, then there exists an edge $p \xrightarrow{a} q$ with $u \in \mathcal{F}_q$ and $\Phi(au) = F_a(\Phi(u)) \in F_a(V_q)$. Thus $V_P = \bigcup\{F_a(V_q) : p \xrightarrow{a} q\}$.
2. If $\xrightarrow{u} q$ and $v \in \mathcal{F}_q$ then $F_u(\Phi(v)) = \Phi(uv) \in \Phi([u])$, so $F_u(V_q) \subseteq \Phi([u])$. Conversely, if $uv \in [u]$, then there exists a path $p \xrightarrow{u} q \xrightarrow{v}$ and $\Phi(uv) = F_u(\Phi(v)) \in F_u(V_q)$, so $\Phi([u]) = \bigcup\{F_u(V_q) : \xrightarrow{u} q\}$ for each $u \in \mathcal{L}_G$.
3. Assume that each V_p is an interval and denote by $B_0 = \{p \in B : |V_p| < 1\}$ the set of vertices whose intervals are proper. For $p \in B_0$ denote by $l_p = \mathfrak{l}(V_p)$, $r_p = \mathbf{r}(V_p)$. Since $l_p \in V_p$, by item 1, there exists an edge $p \xrightarrow{a} q$ and $x \in V_q$ with $l_p = F_a(x)$. It follows that V_q is also a proper interval and either $x = l_q$ provided F_a is increasing or $x = r_q$ provided F_a is decreasing. We define the left selector L on p as $L(p) = (p, a, q)$. Analogously there exists an edge $p \xrightarrow{b} s$ such that $F_b(r_p) = r_s$ provided F_b is increasing and $F_b(r_p) = l_s$ provided F_b is decreasing, and we define $R(p) = (p, b, s)$. If $V_p = \overline{\mathbb{R}}$, we define $L(p)$ and $R(p)$ arbitrarily. Thus L, R are selectors for $G = (B, E)$. For $p \in B_0$, there exists $q \in B_0$ such that $p \xrightarrow{u} q \xrightarrow{v} q$ and $L^p = uv^\omega$. For every k we have $\Phi(L^p) = F_u(\mathbf{s}(F_v)) = F_{uv^k}(\mathbf{s}(F_v)) \in \Phi([uv^k])$. Depending on the orientations of F_u and F_v we have either $l_p = F_u(l_q) = F_{uv^k}(l_q) \in \Phi([uv^k])$ or $l_p = F_{uv^k}(r_q) \in \Phi([uv^k])$. Since $\lim_{k\to\infty} |\Phi([uv^k])| = 0$, we get $\Phi(L^p) = l_p$ and similarly $\Phi(R^p) = r_p$. □

If L, R are selectors from Theorem 4.27, then for each selector K and for each $p \in B$ we have $\Phi(K^p) \in \Phi(\mathcal{F}_p) \subseteq [\Phi(L^p), \Phi(R^p)]$. Since there is only a finite number of selectors for a given

labelled graph, the left and right selectors L, R from Theorem 4.27 can be found effectively. We define now a class of number systems with sofic subshifts whose sets $\Phi(\mathcal{F}_p)$ are intervals. We say that $V = \{V_p \subseteq \overline{\mathbb{R}} : p \in B\}$ is a closed interval cover, if each V_p is a closed interval and $\bigcup_{p\in B} = \overline{\mathbb{R}}$.

Definition 4.28 *A* **sofic number system** *of order $n \geq 1$ over an alphabet A is a triple (F, G, V), where*

1. *$F = \{F_a \in \mathbb{M}(\mathbb{R}) : a \in A\}$ is an iterative system.*
2. *$G = (B, E)$ is a finite A-labelled graph.*
3. *$V = \{V_p \subseteq \overline{\mathbb{R}} : p \in B\}$ is a closed interval cover of $\overline{\mathbb{R}}$ such that $V_p = \bigcup\{F_a(V_q) : p \xrightarrow{a} q\}$.*
4. *$V_q \subseteq \overline{\mathbf{U}(F_u)}$ whenever $p \xrightarrow{u} q$ and $|u| = n$.*
5. *If $G = (B, E, \mathbf{i})$ is an initialized graph, then $V_{\mathbf{i}} = \overline{\mathbb{R}}$.*
6. *If $\{\mathrm{int}_{V_p}(F_a(V_q)) : p \xrightarrow{a} q\}$ is a cover of V_p, then we say that (F, G, V) is a* **redundant sofic number system**.

Theorem 4.29 *Let (F, G, V) be a sofic number system. Then*

1. *(F, Σ_G) is a number system, i.e., $\Sigma_G \subseteq \mathbb{X}_F$ and $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ is continuous and surjective.*
2. *$\Phi(\mathcal{F}_p) = V_p$ for each $p \in B$*
3. *$x = \Phi(u)$ iff there exists an infinite path (p, u) such that $x \in \bigcap F_{u_{[0,i)}}(V_{p_i})$.*
4. *$\Phi([u]) = \bigcup\{F_u(V_q) : \xrightarrow{u} q\}$ for each $u \in \mathcal{L}_G$.*
5. *If $G = (B, E, \mathbf{i})$ is an initialized graph then $\Phi([u]) = \bigcup\{F_u(V_q) : \mathbf{i} \xrightarrow{u} q\}$.*
6. *If $G = (B, E, \mathbf{i})$ is a deterministic graph and $\mathbf{i} \xrightarrow{u} q$, then $\Phi([u]) = F_u(V_q)$.*
7. *If (F, G, V) is a redundant sofic system then $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ is a redundant mapping.*

Proof: We assume that the order is $n = 1$, since the proof in the case of a general order is similar. Thus we assume that $V_q \subseteq \overline{\mathbf{U}(F_a)}$ whenever $p \xrightarrow{a} q$. By Proposition 3.33 there exists a real increasing function $\psi : [0, 1] \to [0, 1]$ such that $\psi(0) = 0$, $\psi(t) < t$ for $t > 0$ and for each $a \in A$ and for each interval $I \subseteq \mathbf{U}(F_a)$ we have $|F_a(I)| \leq \psi(|I|)$. Let $u \in \Sigma_G$ and let $p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_1} p_2 \xrightarrow{u_2} \cdots$ be an infinite path with label u. For $0 < m < n$ we have

$$\begin{aligned} F_{u_{[m,n)}}(V_{p_n}) &\subseteq F_{u_{[m,n-1)}}(V_{p_{n-1}}) \subseteq \cdots \subseteq V_{p_m} \subseteq \overline{\mathbf{U}(F_{u_{m-1}})}, \\ |F_{u_{[0,n)}}(V_{p_n})| &= |F_{u_0} F_{u_{[1,n)}}(V_{p_n})| \leq \psi(|F_{u_{[1,n)}}(V_{p_n})|) \leq \psi^2(|F_{u_{[2,n)}}(V_{p_n})|) \leq \cdots \\ &\leq \psi^n(|V_{p_n}|). \end{aligned}$$

Thus $\lim_{n\to\infty} |F_{u_{[0,n)}}(V_{p_n})| = 0$. Since $F_{u_{[0,n+1)}}(V_{p_{n+1}}) \subseteq F_{u_{[0,n)}}(V_{p_n})$, there exists a unique point $x \in \bigcap_n F_{u_{[0,n)}}(V_{p_n})$. Since $F_{u_{[0,n)}}(V_{p_n}) \subseteq V_{p_0}$, by Proposition 3.8 there exist points $x_n \in F_{u_{[0,n)}}(V_{p_n})$ such that

$$(F^{-1}_{u_{[0,n)}})^\bullet(x_n) \geq |V_{p_0}|/|F_{u_{[0,n)}}(V_{p_n})| \geq |V_{p_n}|/\psi^n(|V_{p_n}|),$$

so $\lim_{n\to\infty} x_n = x$, $\lim_{n\to\infty}(F^{-1}_{u_{[0,n)}})^\bullet(x_n) = \infty$. By Theorem 3.41 we get $\Phi(u) = x$. Thus we have proved $\Sigma_G \subseteq \mathbb{X}_F$. Since $F_{u_{[0,n)}}(V_{p_n}) \subseteq V_{p_0}$, we get $\Phi(\mathcal{F}_{p_0}) \subseteq V_{p_0}$. Conversely, we construct for each $x = x_0 \in V_p$ a path $p = p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_1} p_2 \xrightarrow{u_2} \cdots$ such that $x = \Phi(u)$. If $p_0 \xrightarrow{u_{[0,n)}} p_n$ has been already constructed and $x_n = F^{-1}_{u_{[0,n)}}(x) \in V_{p_n}$, then we find an edge $p_n \xrightarrow{u_n} p_{n+1}$ with $x_n \in F_{u_n}(V_{p_{n+1}})$ and set $x_{n+1} = F^{-1}_{u_n}(x_n)$. Then $x = \Phi(u)$, so we have proved $V_p \subseteq \Phi(\mathcal{F}_p)$ for each $p \in B$. Since V is a cover, $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ is surjective. We show that $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ is continuous. For $u \in \mathcal{L}^n_G$ there exists a finite number of paths $p_{0,j} \xrightarrow{u_0} p_{1,j} \xrightarrow{u_1} \cdots \xrightarrow{u_{n-1}} p_{n,j}$ with label u. Set $W_i = \bigcup_j V_{i,j} \subseteq \overline{\mathbf{U}(F_{u_{i-1}})}$. Then $\Phi(u) \in F_u(W_n)$

and $|F_u(W_n)| \le \psi^n(|\mathbf{U}(F_{u_{m-1}})|) \le \psi^n(1)$. Thus $|\Phi(u)| \le \psi^{|u|}(1)$ and therefore Φ is continuous. Thus we have proved 1,2,3.
4. If $x \in \Phi([u])$ then there exists an infinite path with prefix $p \xrightarrow{u} q$ so $x \in F_u(V_q)$. Conversely, if $x \in F_u(V_q)$, then u can be extended to an infinite path and $x \in \Phi([u])$.
5. If $x \in \Phi([u])$ then there exists an infinite path with prefix $\mathbf{i} \xrightarrow{u} q$ so $x \in F_u(V_q)$. Conversely, if $x \in F_u(V_q)$, then u can be extended to an infinite path and $x \in \Phi([u])$.
6. is an immediate consequence of 5.
7. If $x \in \Phi([u])$, then there exists a path with label u and target p such that $x \in F_u(V_p)$, so $F_u^{-1}(x) \in V_p$. By the assumption there exists an edge $p \xrightarrow{a} q$ such that $F_u^{-1}(x) \in \mathrm{int}_{V_p}(F_a(V_q))$ so $x \in \mathrm{int}_{F_u(V_p)}(F_{ua}(V_q)) \subseteq \mathrm{int}_{\Phi([u])}(\Phi([ua]))$. By Theorem 2.27, Φ is redundant. □

Note that the mapping $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ may be redundant even if the system (F, G, V) is not redundant, This may happen in an interval number system (F, W) with a cover W, whose expansion subshift $\mathcal{S}_{F,W}$ is sofic. Then $\{V_p : p \in B\}$ need not be a cover. However, the subshift may have another graph G with another almost cover V and (F, G, V) may be redundant.

p	a	V_p	$F_a^{-1}(V_p)$	q
$\overline{1}$	$\overline{1}$	$(\infty, -1)$	$(\infty, 0)$	$\overline{1}, \overline{0}$
$\overline{0}$	0	$(-1, 0)$	$(1, \infty)$	1
0	0	$(0, 1)$	$(\infty, -1)$	$\overline{1}$
1	1	$(1, \infty)$	$(0, \infty)$	0, 1

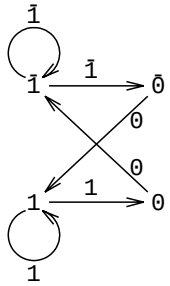


Figure 4.7: The labelled graph of $\mathcal{S}_{F,W}$ for the system of signed continued fractions of Example 4.5
.

Consider the number system of signed continued fractions $(F, \Sigma_D) = (F, \mathcal{S}_{F,W})$ of Example 4.5 with forbidden words $D = \{00, \overline{1}1, 1\overline{1}, \overline{1}0\overline{1}, 101\}$ and intervals $W_{\overline{1}} = (\infty, -1)$, $W_0 = (-1, 1)$, $W_1 = (1, \infty)$. Its open SFT partition has alphabet $B = \{\overline{1}, \overline{0}, 0, 1\}$ and intervals $V_{\overline{1}} = (\infty, -1)$, $V_{\overline{0}} = (-1, 0)$, $V_0 = (0, 1)$, $V_1 = (1, \infty)$. The graph $G_{F,W,V}$ is neither initialized nor right-resolving: all edges with the same source carry the same label (see Figure 4.7) and we have

$$\begin{aligned}
F_{\overline{1}}^{-1}(\overline{V_{\overline{1}}}) &= [\infty, 0] = [\infty - 1] \cup [-1, 0] = \overline{V_{\overline{1}}} \cup \overline{V_{\overline{0}}} \\
F_0^{-1}(\overline{V_{\overline{0}}}) &= [1, \infty] = \overline{V_1} \\
F_0^{-1}(\overline{V_0}) &= [\infty, -1] = \overline{V_{\overline{1}}} \\
F_1^{-1}(\overline{V_1}) &= [0, \infty] = [0, 1] \cup [1, \infty] = \overline{V_0} \cup \overline{V_1}.
\end{aligned}$$

The vertices of the deterministic labelled graph of Σ_D are proper prefixes of the forbidden words $B = \{\lambda, \overline{1}, 0, 1, \overline{1}0, 10\}$. In figure 4.8 we give the left and right selectors constructed according to Theorem 4.27, corresponding intervals V_q and their preimages by the labels of

p	a	q	L^q	R^q	V_q	$F_a(V_q)$
$\lambda, 0, \overline{1}, 10$	$\overline{1}$	$\overline{1}$	$\overline{1}^\omega$	01^ω	$[\infty, 0]$	$[\infty, -1]$
λ	0	0	$10\overline{1}^\omega$	$\overline{1}01^\omega$	$[1, -1]$	$[-1, 1]$
$\lambda, 0, 1, \overline{1}0$	1	1	$0\overline{1}^\omega$	1^ω	$[0, \infty]$	$[1, \infty]$
$\overline{1}$	0	$\overline{1}0$	$10\overline{1}^\omega$	1^ω	$[1, \infty]$	$[-1, 0]$
1	0	10	$\overline{1}^\omega$	$\overline{1}01^\omega$	$[\infty, -1]$	$[0, 1]$

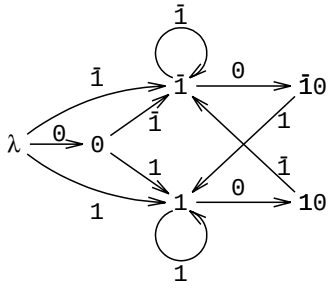


Figure 4.8: The deterministic graph of the number system of signed continued fractions.

ingoing edges. We have

$$\begin{aligned}
V_\lambda &= \overline{\mathbb{R}} = F_{\overline{1}}(V_{\overline{1}}) \cup F_0(V_0) \cup F_1(V_1) \\
V_{\overline{1}} &= [\infty, 0] = [\infty, -1] \cup [-1, 0] = F_{\overline{1}}(V_{\overline{1}}) \cup F_0(V_{\overline{1}0}) \\
V_0 &= [1, -1] = [1, \infty] \cup [\infty, 1] = F_1(V_1) \cup F_{\overline{1}}(V_{\overline{1}}) \\
V_1 &= [0, \infty] = [0, 1] \cup [1, \infty] = F_0(V_{10}) \cup F_1(V_1) \\
V_{\overline{1}0} &= [1, \infty] = F_1(V_1) \\
V_{10} &= [\infty, -1] = F_{\overline{1}}(V_{\overline{1}})
\end{aligned}$$

The interval cylinders are obtained from the unique paths with source $\mathbf{i} = \lambda$: $\Phi([\overline{1}]) = F_{\overline{1}}V_{\overline{1}} = [\infty, -1]$, $\Phi([0]) = F_0V_0 = [-1, 1]$, $\Phi([1]) = F_1V_1 = [1, \infty]$. Thus $\Sigma_G = \Sigma_D$ and (F, G, V) is a sofic number system.

Consider the binary signed system with alphabet $A = \{\overline{1}, 0, 1, \overline{0}\}$ and forbidden words $D = \{\overline{1}0, 0\overline{0}, 1\overline{0}, \overline{0}0, \overline{1}1, 1\overline{1}\}$. The vertices of the deterministic labelled graph are prefixes of the forbidden words $B = \{\lambda, \overline{1}, 0, 1, \overline{0}\}$. The graph together with the V-intervals is in Figure 4.9. We have

$$\begin{aligned}
V_\lambda &= \overline{\mathbb{R}} = F_{\overline{1}}(V_{\overline{1}}) \cup F_0(V_0) \cup F_1(V_1) \cup F_{\overline{0}}(V_{\overline{0}}) \\
V_{\overline{1}} &= [-1, \tfrac{1}{2}] = [-1, -\tfrac{1}{4}] \cup [-\tfrac{1}{2}, \tfrac{1}{2}] = F_{\overline{1}}(V_{\overline{1}}) \cup F_0(V_0) \\
V_0 &= [-1, 1] = [-1, -\tfrac{1}{4}] \cup [-\tfrac{1}{2}, \tfrac{1}{2}] \cup [\tfrac{1}{4}, 1] = F_{\overline{1}}(V_{\overline{1}}) \cup F_0(V_0) \cup F_1(V_1) \\
V_1 &= [-\tfrac{1}{2}, 1] = [-\tfrac{1}{2}, \tfrac{1}{2}] \cup [\tfrac{1}{4}, 1] = F_0(V_0) \cup F_1(V_1)
\end{aligned}$$

Thus $\Sigma_G = \Sigma_D$ and (F, G, V) is a sofic number system.

p	a	q	L^q	R^q	V_q	F_aV_q
$0, \overline{0}, \overline{1}$	$\overline{1}$	$\overline{1}$	$\overline{1}^\omega$	01^ω	$[-1, \frac{1}{2}]$	$[-1, -\frac{1}{4}]$
$0, 1, \overline{1}$	0	0	$\overline{1}^\omega$	1^ω	$[-1, 1]$	$[-\frac{1}{2}, \frac{1}{2}]$
$0, 1, \overline{0}$	1	1	$0\overline{1}^\omega$	1^ω	$[-\frac{1}{2}, 1]$	$[\frac{1}{4}, 1]$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$10\overline{1}^\omega$	$\overline{1}01^\omega$	$[\frac{1}{4}, -\frac{1}{4}]$	$[\frac{1}{2}, -\frac{1}{2}]$

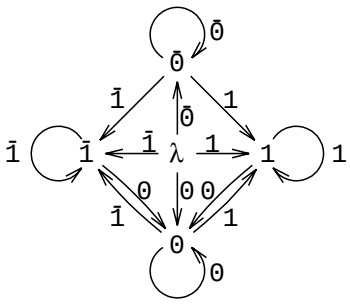


Figure 4.9: The deterministic graph of the binary signed system

4.6 The contraction and length quotients

The speed of convergence of a number system is expressed by its length quotients which measure the dependence of the cylinder interval length on the word length. The length quotients are related to the contraction quotients which measure the growth of the derivations of the composite transformations. To develop the theory of these quotients we need the subadditive Lemma 4.30.

Lemma 4.30 *Let $\{a_n : n \geq 1\}$ be a sequence of real numbers such that $a_{n+m} \leq a_n + a_m$. Then there exists a limit $a = \lim_{n\to\infty} \frac{a_n}{n}$ and $a \leq \frac{a_m}{m}$ for each m.*

Proof: For a fixed m, let $n = m \cdot q_n + r_n$, where $q_n = \lfloor n/m \rfloor$ is the integer part of n/m, and $0 \leq r_n < m$ is the remainder. Since $a_n \leq q_n \cdot a_m + a_{r_n}$, we get

$$\begin{aligned} \limsup_{n\to\infty} \frac{a_n}{n} &\leq a_m \cdot \lim_{n\to\infty} \frac{q_n}{n} + \lim_{n\to\infty} \frac{a_{r_n}}{n} = \frac{a_m}{m} \\ \limsup_{n\to\infty} \frac{a_n}{n} &\leq \liminf_{m\to\infty} \frac{a_m}{m} \end{aligned}$$

so the limit $a = \lim_{n\to\infty} \frac{a_n}{n}$ exists and $a \leq \frac{a_m}{m}$ for each m. □

Proposition 4.31 *Let (F, Σ) be a number system. For $n > 0$ set*

$$\begin{aligned} \mathbf{q}_n &= \inf\{|F_u^\bullet(\Phi(v))| : uv \in \Sigma, |u| = n\} \\ \mathbf{Q}_n &= \sup\{|F_u^\bullet(\Phi(v))| : uv \in \Sigma, |u| = n\} \end{aligned}$$

Then

1. $0 < \mathbf{q}_n \cdot \mathbf{q}_m \leq \mathbf{q}_{n+m}$.
2. $0 < \mathbf{Q}_{n+m} \leq \mathbf{Q}_n \cdot \mathbf{Q}_m$.
3. *There exists the limit $\mathbf{q} = \lim_{n\to\infty} \sqrt[n]{\mathbf{q}_n}$ called the* **lower contracting quotient** *of (F, Σ).*
4. *There exists the limit $\mathbf{Q} = \lim_{n\to\infty} \sqrt[n]{\mathbf{Q}_n}$ called the* **upper contracting quotient** *of (F, Σ).*
5. *For each n we have $\mathbf{q}_n \leq \mathbf{q} \leq \mathbf{Q} \leq \mathbf{Q}_n$*

Proof: The function $Q_n : \Sigma \to \mathbb{R}$ defined by $Q_n(u) = |(F_{u_{[0,n)}})^\bullet(\Phi(\sigma^n(u))|$ is continuous and positive. Since Σ is compact, the function has a positive minimum $\mathbf{q}_n$ and a positive maximum $\mathbf{Q}_n$. Let $uvw \in \Sigma$, $|u| = n$, $|v| = m$. Since $F_{uv}^\bullet(\Phi(w)) = F_u^\bullet(\Phi(vw)) \cdot F_v^\bullet(\Phi(w))$, we get $\mathbf{q}_n \cdot \mathbf{q}_m \leq F_{uv}^\bullet(\Phi(w)) \leq \mathbf{Q}_n \cdot \mathbf{Q}_m$, so $\mathbf{q}_n \cdot \mathbf{q}_m \leq \mathbf{q}_{n+m}$, $\mathbf{Q}_{n+m} \leq \mathbf{Q}_n \cdot \mathbf{Q}_m$. We apply Lemma 4.30 to $-\ln \mathbf{q}_n$ and $\ln \mathbf{Q}_n$ to get the existence of limits $\mathbf{q}$, $\mathbf{Q}$ with $\mathbf{q}_n \leq \mathbf{q} \leq \mathbf{Q} \leq \mathbf{Q}_n$. □

Proposition 4.32 *If (F, W) is an interval number system then*

$$\begin{aligned} \mathbf{q}_n &= \min\{F_u^\bullet(x) : u \in \mathcal{L}_{F,W}^n, x \in F_u^{-1}(\overline{W_u})\} \\ \mathbf{Q}_n &= \max\{F_u^\bullet(x) : u \in \mathcal{L}_{F,W}^n, x \in F_u^{-1}(\overline{W_u})\} \end{aligned}$$

Proof: If $uv \in \Sigma$ then $F_u(\Phi(v)) = \Phi(uv) \in \Phi([u]) = \overline{W_u}$, so $\Phi(v) \in F_u^{-1}(\overline{W_u})$. There exists $uv \in \Sigma$ such that

$$\mathbf{q}_n = F_u^\bullet(\Phi(v)) \geq \min\{F_u^\bullet(x) : u \in \mathcal{L}_{F,W}^n, x \in F_u^{-1}(\overline{W_u})\}.$$

Conversely, the minimum of all $F_u^\bullet(x)$ on $F_u^{-1}(\overline{W_u})$ is attained at some $x \in F_u^{-1}(\overline{W_u})$ with $u \in \mathcal{L}_{F,W}^n$. Since $\overline{W_u} = \Phi([u])$, there exists $uv \in [u]$ with $x = F_u^{-1}(\Phi(uv)) = \Phi(v)$, so

$$\min\{F_u^\bullet(x) :\ u \in \mathcal{L}_{F,W}^n, x \in F_u^{-1}(\overline{W_u})\} = F_u^\bullet(\Phi(v)) \geq \mathbf{q}_n$$

For $\mathbf{Q}_n$ the proof is analogous. □

Proposition 4.33 *If (F, G, V) is a sofic number system then*

$$\begin{aligned} \mathbf{q}_n &= \min\{F_u^\bullet(x) :\ x \in V_q, \xrightarrow{u} q, |u| = n\} \\ \mathbf{Q}_n &= \max\{F_u^\bullet(x) :\ x \in V_q, \xrightarrow{u} q, |u| = n\} \end{aligned}$$

Proof: There exists $uv \in \Sigma$ such that $|u| = n$, $\mathbf{q}_n = F_u^\bullet(\Phi(v))$. There exists a path $p \xrightarrow{u} q \xrightarrow{v}$ such that $v \in \mathcal{F}_q$, $\Phi(v) \in \Phi(\mathcal{F}_q) = V_q$, so $\mathbf{q}_n \geq \min\{F_u^\bullet(x) :\ x \in V_q, \xrightarrow{u} q, |u| = n\}$. Conversely, there exists a path $p \xrightarrow{u} q$ and $x \in V_q$, where $F_u^\bullet(x)$ attains its minimum. Since $V_q = \Phi(\mathcal{F}_q)$, there exists a $v \in \mathcal{F}_q$ with $x = \Phi(v)$, so $\min\{F_u^\bullet(x) :\ x \in V_q, \xrightarrow{u} q, |u| = n\} \geq \mathbf{q}_n$. For $\mathbf{Q}_n$, the proof is analogous. □

Definition 4.34 *Let (F, Σ) be a Möbius number system. The lower and upper* **length quotients** *are defined by*

$$\begin{aligned} \mathbf{l}_n &= \min\{|\Phi[u]| :\ u \in \mathcal{L}_G^n\}, \\ \mathbf{L}_n &= \max\{|\Phi[u]| :\ u \in \mathcal{L}_G^n\} \\ \mathbf{l} &= \liminf_{n\to\infty} \sqrt[n]{\mathbf{l}_n} \\ \mathbf{L} &= \limsup_{n\to\infty} \sqrt[n]{\mathbf{L}_n} \end{aligned}$$

Proposition 4.35 *For an interval number system (F, W) we have $\mathbf{L} \leq \sqrt[m]{\mathbf{Q}_m}$ for each $m > 0$.*

Proof: For each $u \in \mathcal{L}_{F,W}^{n+1}$ we have $|\Phi([u])| = |W_u| \leq |F_{u_{[0,n)}}(W_{u_n})|$. For a fixed m let $n = km + j$ with $0 \leq j < m$. We get

$$\begin{aligned} |W_u| &\leq |F_{u_{[0,m)}}(W_{u_{[m,n]}})| \leq \mathbf{Q}_m \cdot |W_{u_{[m,n]}}| \leq \cdots \leq \mathbf{Q}_m^k \cdot |W_{u_{[km,n]}}| \\ \sqrt[n]{|W_u|} &\leq \mathbf{Q}_m^{\frac{k}{km+j}} \cdot C^{\frac{1}{km+j}} \end{aligned}$$

where $C = \max\{|W_u| :\ |u| < m\}$. As $n \to \infty$, $k \to \infty$ and the right-hand side converges to $\sqrt[m]{\mathbf{Q}_m}$. □

Theorem 4.36 *If (F, G, V) is a sofic number system and $m > 0$ then*

$$\sqrt[m]{\mathbf{q}_m} \leq \mathbf{l} \leq \mathbf{L} \leq \sqrt[m]{\mathbf{Q}_m}.$$

Proof: For a fixed m denote by

$$\begin{aligned} C_0 &= \min\{|F_u(V_q)| : \xrightarrow{u} q, |u| < m\} \\ C_1 &= \max\{|F_u(V_q)| : \xrightarrow{u} q, |u| < m\} \end{aligned}$$

Assume that $p \xrightarrow{u} q$ is a path in G and $n = km + j$, where $0 \le j < m$. Then

$$|F_u(V_q)| \le \mathbf{Q}_m \cdot |F_{\sigma^m(u)}(V_q)| \le \cdots \le \mathbf{Q}_m^k \cdot |F_{\sigma^{km}(u)}(V_q)| \le C_1 \cdot \mathbf{Q}_m^k$$

and similarly $|F_u(V_q)| \ge C_0 \cdot \mathbf{q}_m^k$, so

$$C_0^{\frac{1}{n}} \cdot \mathbf{q}_m^{\frac{k}{km+j}} \le \sqrt[n]{|F_u(V_q)|} \le C_1^{\frac{1}{n}} \cdot \mathbf{Q}_m^{\frac{k}{km+j}}$$

As $n \to \infty$, the left-hand side converges to $\sqrt[m]{\mathbf{q}_m}$ and the right-hand side converges to $\sqrt[m]{\mathbf{Q}_m}$. □

Example 4.37 *For the binary signed number system of Example 4.3 we have* $\mathbf{l} = \mathbf{L} = \frac{1}{2}$.

Proof: $\mathrm{sz}(a,b) = \frac{a \cdot b}{\det(b,a)}$. For $m \ge 0$, $u \in \{\overline{1}, 0, 1\}^n$ we get $\Phi[\overline{0}^m u] = F_{\overline{0}^m u}(V_p)$ where $[-\frac{1}{2}, \frac{1}{2}] \subset V_p \subseteq [-1, 1]$, so

$$2^m[\varphi(u) - 2^{-n-1}, \varphi(u) + 2^{-n-1}] \subseteq \Phi[\overline{0}^m u] \subseteq 2^m[\varphi(u) - 2^{-n}, \varphi(u) + 2^{-n}]$$

where $\varphi(u) = \sum_{i<|u|} u_i 2^{-i-1}$, so $|\varphi(u)| \le 1$. On both sides we have an interval of the form $I_n = [\frac{2^n a_n - b_n}{2^n}, \frac{2^n a_n + b_n}{2^n}]$, where $|a_n| \le 1$, $\frac{1}{2} \le b_n \le 1$. We get $\mathrm{sz}(I_n) = \frac{2^{2n}(a_n^2+1) - b_n^2}{2^{n+1} b_n}$, $2^{n-1} - 2^{-n-1} \le \mathrm{sz}(I_n) \le 2^{n+1}$. From the estimate $\frac{1}{4 \cdot \mathrm{sz}(I)} \le |I| \le \frac{1}{\pi \cdot \mathrm{sz}(I)}$ we obtain $\lim_{n\to\infty} \sqrt[n]{|I_n|} = \frac{1}{2}$, so $\lim_{|u|\to\infty} \sqrt[m+|u|]{|\Phi[\overline{0}^m u]|} = \frac{1}{2}$. For $\overline{0}^m$ we have $\Phi[\overline{0}^m] = F_{\overline{0}}^m(V_{\overline{0}}) = [\frac{2^{m-2}}{1}, \frac{2^{m-2}}{-1}]$ with $\mathrm{sz}(\Phi[\overline{0}^m]) = \frac{2^{2m-4}-1}{2^{m-1}}$, so $\lim_{m\to\infty} \sqrt[m]{|\Phi[\overline{0}^m]|} = \frac{1}{2}$. Thus $\mathbf{l} = \mathbf{L} = \frac{1}{2}$. □

Proposition 4.38 *For the system of symmetric continued fractions from Example 4.6 we have* $\mathbf{l} \le \frac{3-\sqrt{5}}{2} \doteq 0.312$, $\mathbf{L} = 1$.

Proof: The deterministic graph of the system has vertices $B = \{\lambda, \overline{1}, 1\}$ and edges

$$\overline{1} \overset{\overline{1},\overline{0}}{\leftarrow} \overline{1} \overset{\overline{1},\overline{0}}{\leftarrow} \lambda \overset{0,1}{\to} 1 \overset{0,1}{\to} 1$$

with intervals $V_{\overline{1}} = [\infty, 0]$, $V_1 = [0, \infty]$. For $u = (10)^n$ we have $F_{10}^n = [\frac{f_{2n+1}}{f_{2n}}, \frac{f_{2n}}{f_{2n-1}}]$, where f_n are the Fibonacci numbers defined by $f_1 = f_2 = 1$, $f_{n+2} = f_n + f_{n-1}$. It follows

$$\begin{aligned} \Phi[(10)^n] &= F_{10}^n V_1 = [\tfrac{f_{2n}}{f_{2n-1}}, \tfrac{f_{2n+1}}{f_{2n}}] \\ \mathrm{sz}(\Phi[(10)^n]) &= f_{2n}(f_{2n+1} + f_{2n-1}) \approx \alpha^{-4n}(\alpha + \alpha^{-1})/5, \end{aligned}$$

where $\alpha = \frac{\sqrt{5}-1}{2} \doteq 0.618$, so $\lim_{n\to\infty} \sqrt[2n]{|\Phi[(10)^n]|} = \alpha^2$, and $\mathbf{l} \le \alpha^2$. For $u = 1^n$ we have $\Phi[1^n] = F_1^n(V_1) = [\frac{n}{1}, \frac{1}{0}]$, with $\mathrm{sz}(\Phi[1^n]) = n$ so $\frac{1}{4n} \le |\Phi[1^n]| \le \frac{1}{\pi n}$. It follows $\lim_{n\to\infty} \sqrt[n]{|\Phi[1^n]|} = 1$, so $\mathbf{l} = 1$. □

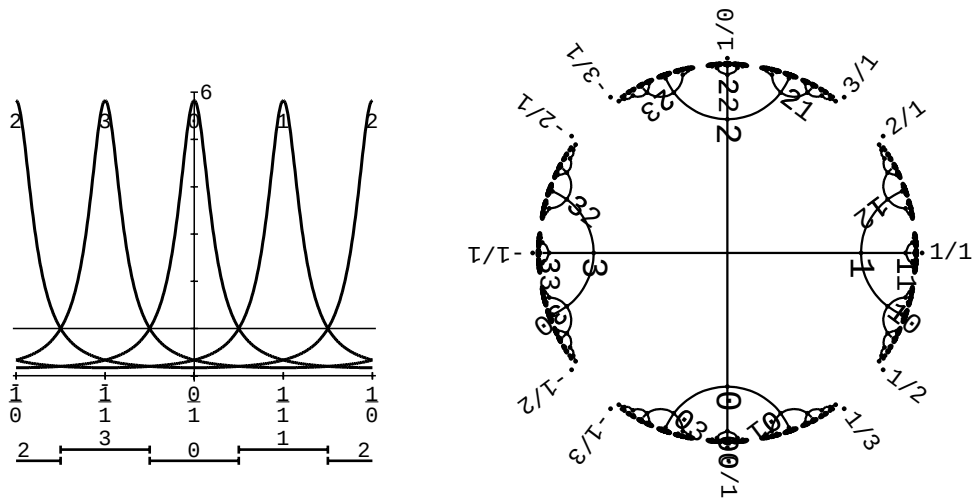


Figure 4.10: The square interval number system with $r = (\sqrt{2}-1)^2$ and $W_0 = \mathbf{V}(F_0) = (1-\sqrt{2}, \sqrt{2}-1)$. The expansion subshift is the circular subshift with speed 1.

4.7 Polygonal number systems

Polygonal number systems consist of hyperbolic transformations whose fixed points form vertices of a regular polygon. The parameters of these systems are the number $n \geq 3$ of vertices and the similarity quotient $0 < r < 1$ of the transformations. We denote by $Q_r(x) = rx$ the similarity with quotient r and by R_n the rotation by angle $\frac{2\pi}{n}$. For $a \in A = \{0, 1, \ldots, n-1\}$, we get

$$R_n^a = \begin{bmatrix} \cos\frac{\pi a}{n} & \sin\frac{\pi a}{n} \\ -\sin\frac{\pi a}{n} & \cos\frac{\pi a}{n} \end{bmatrix}, \; Q_r = \begin{bmatrix} r & 0 \\ 0 & 1 \end{bmatrix}$$

Definition 4.39 *The* **polygonal iterative system** *with $n \geq 3$ vertices and quotient $0 < r < 1$ has alphabet $A = \{0, 1, \ldots, n-1\}$ and transformations $F_a = R_n^a Q_r R_n^{-a}$.*

The transformations of the system are

$$\begin{aligned} F_a &= \begin{bmatrix} r\cos^2\frac{\pi a}{n} + \sin^2\frac{\pi a}{n} & (1-r)\sin\frac{\pi a}{n}\cos\frac{\pi a}{n} \\ (1-r)\sin\frac{\pi a}{n}\cos\frac{\pi a}{n} & r\sin^2\frac{\pi a}{n} + \cos^2\frac{\pi a}{n} \end{bmatrix} \\ &= \begin{bmatrix} (1+r) - (1-r)\cos\frac{2\pi a}{n} & (1-r)\sin\frac{2\pi a}{n} \\ (1-r)\sin\frac{2\pi a}{n} & (1+r) + (1-r)\cos\frac{2\pi a}{n} \end{bmatrix} \end{aligned}$$

The expansion interval of Q_r is $\mathbf{V}(Q_r) = (-\sqrt{r}, \sqrt{r})$ with the length

$$|\mathbf{V}(Q_r)| = \frac{1}{\pi}\operatorname{arccotg}\frac{1-r}{2\sqrt{r}} = \frac{1}{\pi}\arccos\frac{1-r}{1+r}.$$

To get an interval number system we take an interval $W_0 = (-s, s)$ with $s \leq \sqrt{r}$ and

$$W_a = R_n^a(W_0) = \left(\frac{-s\cos\frac{\pi a}{n} + \sin\frac{\pi a}{n}}{s\sin\frac{\pi a}{n} + \cos\frac{\pi a}{n}}, \frac{s\cos\frac{\pi a}{n} + \sin\frac{\pi a}{n}}{-s\sin\frac{\pi a}{n} + \cos\frac{\pi a}{n}}\right).$$

All W_a have the same length $|W_a| = \frac{1}{\pi}\arccos\frac{1-s^2}{1+s^2}$. These intervals should overlap, so their length should be larger than $\frac{1}{n}$. This condition gives

$$\sqrt{r} \geq s \geq \sqrt{\frac{1-\cos\frac{\pi}{n}}{1+\cos\frac{\pi}{n}}} = \tan\frac{\pi}{2n}$$

which implies $r \geq \tan^2\frac{\pi}{2n}$. For example for $n=3$ we get $r \geq \frac{1}{3}$, for $n=4$ we get $r \geq 3-2\sqrt{2} \approx 0.172$ (Figure 4.10), and for $n=6$ we get $r \geq 7-4\sqrt{3} \approx 0.072$. Thus we have

Proposition 4.40 *If $n \geq 3$, $A = \{0,1,\ldots,n-1\}$, $F_a = R_n^a Q_r R_n^{-a}$, $W_a = R_n^a(-s,s)$, $\tan\frac{\pi}{2n} \leq s \leq \sqrt{r} < 1$, then (F,W) is an interval number system.*

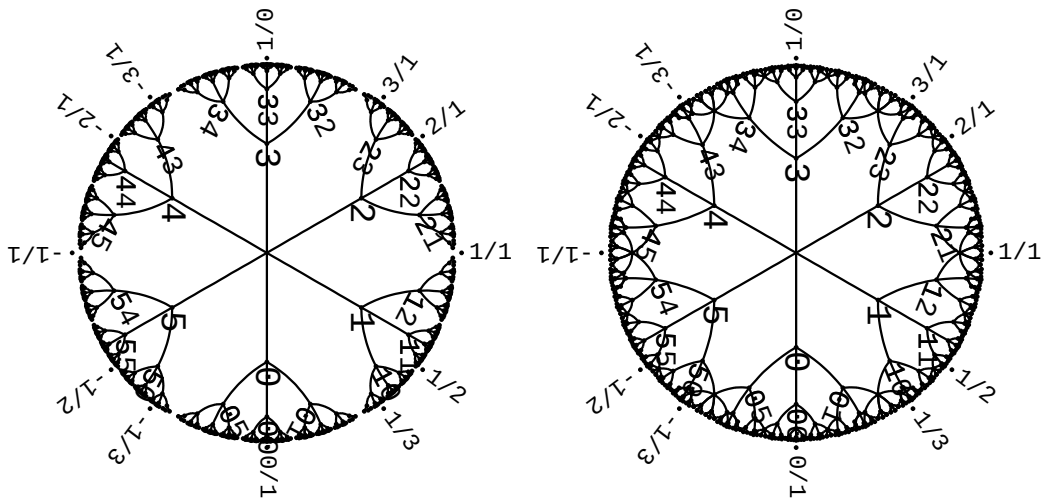


Figure 4.11: The sexagon number systems with circular SFT with speed 1 and parameter $r = 2-\sqrt{3}$ (left) and $r = \frac{1}{3}$ (right)

We consider now polygonal systems (F,Σ), with **circular SFT** Σ which are symmetric with respect to rotations and allow a limited speed around the circle. The circular subshift Σ_1 with speed 1 allows only transitions to neighboring letters, so the forbidden words are

$$D = \{ab \in A^2 :\ b \notin \{\mathrm{mod}_n(a-1), a, \mathrm{mod}_n(a+1)\}\}.$$

For example, with $n=4$, the forbidden words are $D = \{02,13,20,31\}$. With $n=5$, the forbidden words are $D = \{02,03,13,14,24,20,30,31,41,42\}$. The left and right selectors are $L(a) = \mathrm{mod}_n(a-1)$, $R(a) = \mathrm{mod}_n(a+1)$, so $L^0 = (0(n-1)\cdots 21)^\omega$, $R^0 = (012\cdots(n-1))^\omega$. For $v = 012\cdots(n-1)$ we get $F_v = Q_r(R_nQ_rR_n^{-1})(R_n^2Q_rR_n^{-2})\cdots(R_n^{n-1}Q_rR_n^{1-n}) = (Q_rR_n)^n$. We have

$$Q_rR_n = \begin{bmatrix} r\cos\frac{\pi}{n} & r\sin\frac{\pi}{n} \\ -\sin\frac{\pi}{n} & \cos\frac{\pi}{n} \end{bmatrix},\ \mathrm{trc}(Q_rR_n) = \frac{(r+1)^2\cos^2\frac{\pi}{n}}{r}$$

Thus $\mathrm{trc}(Q_rR_n) \geq 4$ iff $r^2\cos^2\frac{\pi}{n} + 2r(\cos^2\frac{\pi}{n} - 2) + \cos^2\frac{\pi}{n} \geq 0$. This quadratic inequality has discriminant $D = 4\sin^2\frac{\pi}{n}$ and solutions

$$\frac{2-\cos^2\frac{\pi}{n} \pm 2\sin^2\frac{\pi}{n}}{\cos^2\frac{\pi}{n}} = \frac{(1\pm\sin\frac{\pi}{n})^2}{1-\sin^2\frac{\pi}{n}} = \frac{1\pm\sin\frac{\pi}{n}}{1\mp\sin\frac{\pi}{n}}.$$

If $r > r_n = \frac{1-\sin\frac{\pi}{n}}{1+\sin\frac{\pi}{n}}$ then Q_rR_n is elliptic and $\Sigma_1 \not\subseteq \mathbb{X}_F$. If $r = r_n$ then Q_rR_n is parabolic and if $r < r_n$ then Q_rR_n is hyperbolic. The stable fixed point of Q_rR_n is then

$$s_{r,n} = \frac{(1-r)\cos\frac{\pi}{n} - \sqrt{(1+r)^2\cos^2\frac{\pi}{n} - 4r}}{2\sin\frac{\pi}{n}}.$$

The vertices of the deterministic automaton for Σ_1 are the prefixes of the forbidden words $B = \{\lambda, 0, 1, \ldots, n-1\}$. For the V-intervals we get $V_a = R_n^a V_0$, where $V_0 = (-s_{r,n}, s_{r,n})$. The value mapping $\Phi : \Sigma \to \overline{\mathbb{R}}$ is surjective provided the intervals V_a cover $\overline{\mathbb{R}}$, i.e., if the length of V_0 is at least $\frac{1}{n}$, i.e., if

$$s_{r,n} \geq \sqrt{\frac{1-\cos\frac{\pi}{n}}{1+\cos\frac{\pi}{n}}} = \frac{1-\cos\frac{\pi}{n}}{\sin\frac{\pi}{n}}.$$

or

$$\sqrt{(1+r)^2\cos^2\frac{\pi}{n} - 4r} \leq (3-r)\cos\frac{\pi}{n} - 2$$

The right-hand side of this inequality must be positive which gives the condition $r < \frac{3\cos\frac{\pi}{n}-2}{\cos\frac{\pi}{n}}$. For $n = 3$ we get $r < -1$ which is impossible so there exists no polygonal number system with $n = 3$ and Σ_1. If $n \geq 4$ then the right-hand side of the inequality is positive, and we get after a little of algebra the condition $r \geq \frac{2\cos\frac{\pi}{n}-1}{2\cos\frac{\pi}{n}+1}$. Since $\frac{1-\sin\frac{\pi}{n}}{1+\sin\frac{\pi}{n}} \leq \frac{3\cos\frac{\pi}{n}-2}{\cos\frac{\pi}{n}}$ for $n \geq 4$, we get

Proposition 4.41 *If $n \geq 4$, $A = \{0, 1, \ldots, n-1\}$, $F_a = R_n^a Q_r R_n^{-a}$, Σ_1 is the circular subshift with speed 1 and*

$$\frac{2\cos\frac{\pi}{n}-1}{2\cos\frac{\pi}{n}+1} \leq r \leq \frac{1-\sin\frac{\pi}{n}}{1+\sin\frac{\pi}{n}}.$$

then (F, Σ_1) is a number system.

Proof: The condition implies that the sets V_a obtained by the selectors cover $\overline{\mathbb{R}}$. To show that (F, Σ_1) is a sofic number system we have to prove the condition 4 of Definition 4.28 that $V_q \subseteq \mathbf{U}(F_a)$ provided $\xrightarrow{a} q$. This reads $V_{a-1} \cup V_a \cup V_{a+1} \subseteq \mathbf{U}(F_a)$. This is satisfied provided $V_1 = R_n(V_0) \subseteq \mathbf{U}(F_0)$ and this is equivalent with $V_0 \subseteq R_n^{-1}(\mathbf{U}(F_0))$. Since $\mathbf{U}(F_0) = (-\frac{1}{\sqrt{r}}, \frac{1}{\sqrt{r}})$, the condition reads

$$s_{r,n} \leq \frac{\cos\frac{\pi}{n} - \sqrt{r}\sin\frac{\pi}{n}}{\sin\frac{\pi}{n} + \sqrt{r}\cos\frac{\pi}{n}}$$

for all r with $\frac{2\cos\frac{\pi}{n}-1}{2\cos\frac{\pi}{n}+1} \leq r \leq \frac{1-\sin\frac{\pi}{n}}{1+\sin\frac{\pi}{n}}$. This can be proved by elementray methods. □

In particular for $n = 4$ we get unique $r = (\sqrt{2}-1)^2$ (Figure 4.10). For $n = 6$ we get $2 - \sqrt{3} \leq r \leq \frac{1}{3}$. The systems with these extreme values are in Figure 4.11. To obtain more convergent systems we take a smaller circular subshift $\Sigma_{1/2}$ with speed $\frac{1}{2}$. The subshift forbids the same words as Σ_1 and moreover the words 012, $0(n-1)(n-2)$, 123, $10(n-1)\ldots$. For the right selector we get $L^0 = v^\omega$ with $v = 01122\cdots(n-1(n-1)0$. For F_v we get $F_v = Q_r(R_nQ_r^2R_n^{-1})(R_n^2Q_r^2R_n^{-2})\cdots(R_n^{n-1}Q_r^2R_n^{1-n}) = (Q_rR_nQ_r)^n$. For each $n \geq 3$ there exist sofic polygonal number systems with the subshift $\Sigma_{1/2}$.

4.8 Discrete groups

Regular transformations with a projective metric form the metric space $\mathbb{MR}$) and the composition operation is continuous. Thus $\mathbb{M}(\mathbb{R})$ is a continuous group. An iterative system $F = \{F_a \in \mathbb{M}(\mathbb{R}) : \ a \in A\}$ determines a subgroup of $\mathbb{M}(\mathbb{R})$: the smallest subgroup of $\mathbb{M}(\mathbb{R})$ which contains all F_a. We say that this is a **discrete group**, if it is discrete subspace of $\mathbb{M}(\mathbb{R})$, i.e., if each its element is isolated (see Beardon [4], Katok [28]). An important example of a discrete group is the **modular group** of transformations with integer coefficients and unit determinant (see Section 6.3)

$$\mathbb{M}^1(\mathbb{Z}) = \{M(x) = \frac{ax+b}{cx+d} : \ a,b,c,d \in \mathbb{Z}, \det(M) = ad - bc = 1\}$$

For example, the systems of signed continued fractions or symmetric continued fractions generate the modular group. Some polygonal systems determine discrete groups as well. We consider discrete polygonal systems with $2n$ transformations which determine tesellation of the hyperbolic disc by regular m-gons. For $F_a = R_{2n} Q_r R_{2n}^{-1}$ we have $F_a F_{a+n} = \mathrm{Id}$. A discrete system occurs if the points $A_0 = 0$, $A_1 = \widehat{F}_{n-1}(0)$, $A_2 = \widehat{F}_{2(n-1)}(0), \ldots$ form vertices of a regular polygon.

Definition 4.42 *Let n, m be integers with $\frac{1}{n} + \frac{2}{m} < 1$. The $(2n, m)$-discrete polygonal system has alphabet $A = \{0, 1, \ldots, 2n-1\}$ and transformations $F_a = R_{2n}^a Q_r R_{2n}^{-a}$, where*

$$r = r_{2n,m} = \frac{1 - \sqrt{1 - \sin^2 \frac{\pi}{2n} / \cos^2 \frac{\pi}{m}}}{1 + \sqrt{1 - \sin^2 \frac{\pi}{2n} / \cos^2 \frac{\pi}{m}}}$$

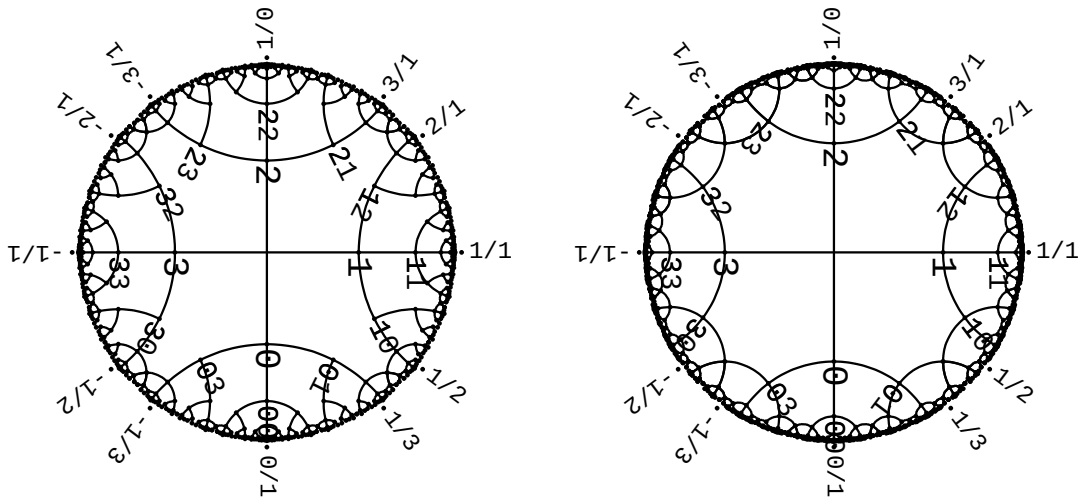


Figure 4.12: The discrete polygonal $(4,5)$-system with $r_{4,5} = \frac{1-\sqrt{\sqrt{5}-2}}{1+\sqrt{\sqrt{5}-2}} \approx 0.346$ (left) and the $(4,6)$-system with $r_{4,6} = 2 - \sqrt{3} \approx 0.268$ (right).

Proposition 4.43 *The $(2n, m)$-discrete polygonal system generates a discrete group which satisfies*

$$\begin{aligned} F_a F_{a+n} &= \mathrm{Id}, \\ F_0 F_{n+1} F_{2(n+1)} \cdots F_{(m-1)(n+1)} &= R_{2n}^{mn+m}, \\ F_0 F_{n-1} F_{2(n-1)} \cdots F_{(m-1)(n-1)} &= R_{2n}^{mn-m} \end{aligned}$$

(the additions are modulo $2n$).

Proof: Denote by $A_0 = 0$ and $A_i = \widehat{F}_0\widehat{F}_{n-1}\widehat{F}_{2(n-1)}\cdots\widehat{F}_{(i-1)(n-1)}(0)$. We derive a condition on r which implies that $A_m = 0$, so $A_0, \ldots, A_{m-1}$ form the vertices of a regular m-gon, whose inner angles at vertices A_i are π/n. Denote by $a = \varrho(0, \widehat{F}_0(0))$ the hyperbolic length of the sides of the polygon, by S its center and by B_0 the middle of the hyperbolic line A_0A_1. The hyperbolic triangle SA_0B_0 has angles $\frac{\pi}{2n}$ at A_0, $\frac{\pi}{2}$ at B_0 and $\frac{\pi}{m}$ at S. Its side has length $A_0B_0 = \frac{a}{2}$. By the second cosine rule and Proposition 3.28 we get

$$\frac{1}{\sqrt{1-|\widehat{F}(0)|^2}} = \cosh\frac{a}{2} = \frac{\cos\frac{\pi}{2n}\cos\frac{\pi}{2} + \cos\frac{\pi}{m}}{\sin\frac{\pi}{2n}\sin\frac{\pi}{2}} = \frac{\cos\frac{\pi}{m}}{\sin\frac{\pi}{2n}}$$

Since $\widehat{F}_0(0) = \frac{i(1-r)}{1+r}$ we get

$$r = \frac{1-|\widehat{F}(0)|}{1+|\widehat{F}(0)|} = \frac{1-\sqrt{1-1/\cosh^2(a/2)}}{1+\sqrt{1-1/\cosh^2(a/2)}}$$

and the formula for r follows. For the transformation $G_- = F_0F_{n-1}F_{2(n-1)}\cdots F_{(m-1)(n-1)}$ we have

$$G_- = Q_r(R_{2n}^{n-1}Q_rR_{2n}^{1-n})\cdots(R_{2n}^{(m-1)(n-1)}Q_rR_{2n}^{(m-1)(1-n)}) = (Q_rR_{2n}^{n-1})^mR^{mn+m}$$

(We use $R_{2n}^{2n} = \mathrm{Id}$). We compute the trace

$$\mathrm{trc}(Q_rR_{2n}^{n-1}) \;=\; \frac{(r+1)^2}{r}\cdot\cos^2\frac{\pi(n-1)}{2n} = 4\cosh^2\frac{a}{2}\cdot\sin^2\frac{\pi}{2n} = 4\cos^2\frac{\pi}{m}$$

Thus $Q_rR_{2n}^{n-1}$ is an elliptic transformation with rotation angle $\mathrm{rot}(Q_rR_{2n}^{n-1}) = \frac{2\pi}{m}$ and therefore $(Q_rR_{2n}^{n-1})^m = \mathrm{Id}$. Thus $G_- = R_{2n}^{mn+m}$. Similarly we get for

$$G_+ = Q_r(R_{2n}^{n+1}Q_rR_{2n}^{-n-1})\cdots(R_{2n}^{(m-1)(n+1)}Q_rR_{2n}^{-(m-1)(n+1)}) = (Q_rR_{2n}^{n+1})^mR^{mn-m} = R^{mn-m}.$$

□

Note that $R^{mn+m} = R^m$, $R^{mn-m} = R^{-m}$ for m even and $R^{mn+m} = R^{n+m}$, $R^{mn-m} = R^{n-m}$ for m odd. In Figure 4.12 left we see the $(4,5)$ discrete system with $r_{4,5} = \frac{1-\sqrt{\sqrt{5}-2}}{1+\sqrt{\sqrt{5}-2}} \approx 0.346$ and the circular subshift $\Sigma_{1/2}$ with speed $\frac{1}{2}$ and forbidden words

$$D = \{02, 13, 20, 31, 012, 123, 230, 301, 032, 103, 210, 321\}.$$

In Figure 4.12 right we see the $(4,6)$ discrete system with $r_{4,6} = 2-\sqrt{3} \approx 0.268$ and the subshift with forbidden words

$$D = \{02, 13, 20, 31, 0321, 0123, 1032, 1230, 2103, 2301, 3201, 3012\}.$$

Chapter 5

Arithmetical algorithms

If (F, Σ) is a number system with redundant value mapping $\Phi : \Sigma \to \overline{\mathbb{R}}$, then each continuous mapping $G : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ can be lifted to a continuous mapping $F : \Sigma \to \Sigma$ such that $\Phi \circ F = G \circ \Phi$ (Proposition 2.16). A mapping $F : \Sigma \to \Sigma$ is continuous iff there exists a sequence of mappings $\{f_k : \mathcal{L}^{n_k}(\Sigma) \to A : \ k \geq 0\}$ such that $F(u)_n = f_n(u_{[0,n_k)})$ for each $u \in \Sigma$ and $k \geq 0$. If there exists an algorithm which for each n computes f_n, then we say that F is an **algorithmic mapping**. In this case there exists an algorithm which computes $F(u)$ for each input word $u \in \Sigma$. The algorithm successively reads letters of the input word u and when it reads the prefix of u of length k_n, it writes the letter $F(u)_n$ to the output. Thus the algorithm works in infinite time but each finite prefix of the oputput is computed in a finite time from a finite prefix of the input. Each algorithmic mapping is continouous but there exist continous mappings which are not algorithmic (see Weihrauch [68]).

Not every continuous mapping $G : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ has an algorithmic lifting. Assume that we want to compute a unary arithmetical operation, or Möbius transformation $G(x) = \frac{ax+b}{cx+d}$. This is possible if a, b, c, d are **alhorithmic numbers** and if the entries of the projective matrices which define the number system (F, Σ) are algorithmic as well. This condition is satisfied if all these entries are rational numbers (see Chapter 6) or algebraic numbers (see Chapter 6). Moreover, the subshift Σ should be an algorithmic subset of A^ω. In the present chapter we present arithmetical algorithms for sofic number systems (F, G, V) such that the entries of the projective matrices F_a and V_p are either rational or algebraic numbers.

In this case there exist also algorithms which compute binary arithmetical operations like addition or multiplication. There is, however, one difference with the unary arithmetical operations. Binary arithmetical operations are not defined everywhere. For example $\infty + \infty$ or $0 \cdot \infty$ are undefined expressions. If the addition algorithm is run on inputs which represent ∞ then it never produces any output. With this exception, binary algorithms work similarly as unary algorithms: Each finite prefix of the output is computed in a finite time from finite prefixes of the inputs. The idea of such an online computation of arithmetic operations comes from an unpublished manuscript of Gosper [21] and has been elaborated by Kornerup and Matula [34], [33] and Vuillemin [66].

5.1 Intervals

In section 3.2 we determine a proper interval $I = (a, b)$ by the ordered pair of its endpoints $a, b \in \overline{\mathbb{R}}$ as $(a, b) = \{x \in \overline{\mathbb{R}} : \ \det(a, x) \cdot \det(x, b) \cdot \det(b, a) > 0\}$ (Definition 3.2). When we work with intervals in arithmetical algorithms, this notation is not convenient. For example, for a decreasing transformation $M \in \mathbb{M}^-(\mathbb{R})$ we get $M(I) = (M(b), M(a))$ so we have to distinguish

the sign of $\det(M)$ when we map an interval by a transformation. A better possibility, which leads to an efficient matrix calculus (see Kůrka [40]), is to define the open interval with endpoints a, b as the set $\{ay_0 + by_1 : \ y_0, y_1 > 0\}$ of **convex combinations** of a, b. The two disjoint intervals $I = (a, b)$ and $J = (b, a)$ are then represented by the matrices $P = \left[\begin{smallmatrix} a_0 & b_0 \\ a_1 & b_1 \end{smallmatrix}\right]$ and $Q = \left[\begin{smallmatrix} a_0 & -b_0 \\ a_1 & -b_1 \end{smallmatrix}\right]$ (see Figure 5.1). The order of columns is arbitrary. Matrices $\left[\begin{smallmatrix} a_0 & b_0 \\ a_1 & b_1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} b_0 & a_0 \\ b_1 & a_1 \end{smallmatrix}\right]$ represent the same interval. A nonzero multiple λP of P represents the same interval as P, so proper intervals are represented by **regular projective matrices**, i.e., by the elements of the projective space $\mathbb{M}(\mathbb{R})$. We get $x \in I$ iff $x = Py$ for some vector y with **positive sign**: the sign of $y \in \overline{\mathbb{R}}$ is the sign of the product $y_0 y_1$: $\mathrm{sgn}(y) = \mathrm{sgn}(y_0 y_1) \in \{-1, 0, 1\}$. Thus $x \in I$ iff $\mathrm{sgn}(P^{-1}x) > 0$. To get also improper intervals we apply this definition also to singular matrices and even to the zero matrix $\mathbf{0} = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ (the zero-dimensional subspace of the vector space $\mathbb{R}^{2\times 2}$). Denote by

$$\overline{\mathbb{M}}(\mathbb{R}) = \mathbb{P}(\mathbb{R}^{2\times 2}) \cup \{\mathbf{0}\}$$

the set of all subspaces of $\mathbb{R}^{2\times 2}$ of dimension at most 1. Recall that the (pseudo)inverse of a matrix is defined by $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right]$. If P is not regular then PP^{-1} is the zero matrix. The stable and unstable point of the zero matrix is by definition $\frac{0}{0}$.

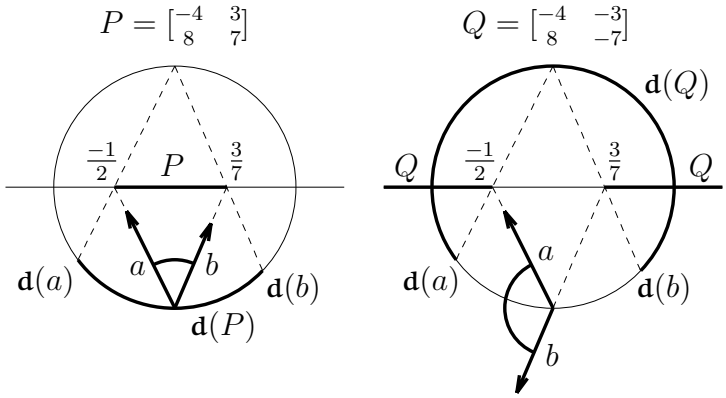


Figure 5.1: The stereographic projection of intervals.

Definition 5.1 *The open and closed intervals of a matrix $P \in \overline{\mathbb{M}}(\mathbb{R})$ are defined by*

$$\begin{aligned} P^o &= \{x \in \overline{\mathbb{R}} : \ \mathrm{sgn}(P^{-1}x) > 0\}, \\ P^c &= \{x \in \overline{\mathbb{R}} : \ \mathrm{sgn}(P^{-1}x) \geq 0\}. \end{aligned}$$

The left and right endpoints of $P = \left[\begin{smallmatrix} a_0 & b_0 \\ a_1 & b_1 \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$ are $\mathbf{l}(P) = \frac{a_0}{a_1}$, $\mathbf{r}(P) = \frac{b_0}{b_1}$ provided $\det(P) < 0$ and $\mathbf{l}(P) = \frac{b_0}{b_1}$, $\mathbf{r}(P) = \frac{a_0}{a_1}$ provided $\det(P) > 0$.

Proposition 5.2 *Let $P = \left[\begin{smallmatrix} a_0 & b_0 \\ a_1 & b_1 \end{smallmatrix}\right] \in \overline{\mathbb{M}}(\mathbb{R})$. Then*

$$P^o = \begin{cases} (\frac{a_0}{a_1}, \frac{b_0}{b_1}) & \text{if } \det(P) < 0, \\ (\frac{b_0}{b_1}, \frac{a_0}{a_1}) & \text{if } \det(P) > 0, \\ \emptyset & \text{if } \det(P) = 0, \mathrm{sgn}(\mathbf{u}(P)) \leq 0 \\ \overline{\mathbb{R}} \setminus \{\mathbf{s}(P)\} & \text{if } \det(P) = 0, \mathrm{sgn}(\mathbf{u}(P)) > 0 \end{cases}$$

$$P^c = \begin{cases} [\frac{a_0}{a_1}, \frac{b_0}{b_1}] & \text{if } \det(P) < 0, \\ [\frac{b_0}{b_1}, \frac{a_0}{a_1}] & \text{if } \det(P) > 0, \\ \{\mathbf{s}(P)\} & \text{if } \det(P) = 0, \mathrm{sgn}(\mathbf{u}(P)) < 0 \\ \overline{\mathbb{R}} & \text{if } \det(P) = 0, \mathrm{sgn}(\mathbf{u}(P)) \geq 0 \end{cases}$$

In particular for the zero matrix we have $\mathbf{0}^o = \emptyset$, $\mathbf{0}^c = \overline{\mathbb{R}}$.

Proof: We have $P^{-1}x = \frac{b_1x_0 - b_0x_1}{a_0x_1 - a_1x_0}$, so

$$\begin{aligned}\det(a,x)\cdot\det(x,b)\cdot\det(b,a) &= (a_0x_1 - a_1x_0)(b_1x_0 - b_0x_1)(b_0a_1 - b_1a_0)\\ &= -(P^{-1}x)_1\cdot(P^{-1}x)_0\cdot\det(P)\\ \det(b,x)\cdot\det(x,a)\cdot\det(a,b) &= (P^{-1}x)_1\cdot(P^{-1}x)_0\cdot\det(P)\end{aligned}$$

and we get the statement for P regular. If $P = \left[\begin{smallmatrix} s_0u_1 & -s_0u_0 \\ s_1u_1 & -s_1u_0\end{smallmatrix}\right]$ is singular with $u = \mathbf{u}(P)$ and $s = \mathbf{s}(P)$ then

$$P^{-1}x = \begin{bmatrix} -s_1u_0 & s_0u_0 \\ -s_1u_1 & s_0u_1\end{bmatrix}\cdot\frac{x_0}{x_1} = \frac{u_0(-s_1x_0 + s_0x_1)}{u_1(-s_1x_0 + s_0x_1)}$$

If $\operatorname{sgn}(u) < 0$ then $P^o = \emptyset$, $P^c = \{s\}$. If $\operatorname{sgn}(u) = 0$ then $P^o = \emptyset$, $P^c = \overline{\mathbb{R}}$. If $\operatorname{sgn}(u) > 0$ then $P^o = \overline{\mathbb{R}} \setminus \{s\}$, $P^c = \overline{\mathbb{R}}$. □

Denote by

$$\begin{aligned}\mathbb{R}^+ &= (0,\infty) = \{x \in \overline{\mathbb{R}} :\ \operatorname{sgn}(x) > 0\} = \mathrm{Id}^o\\ \overline{\mathbb{R}}^+ &= [0,\infty] = \{x \in \overline{\mathbb{R}} :\ \operatorname{sgn}(x) \geq 0\} = \mathrm{Id}^c\end{aligned}$$

Denote by $\neg = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1\end{smallmatrix}\right]$ the **negation matrix**. We have $\neg^{-1} = \neg$ and $\neg x = \frac{x_0}{-x_1}$ for $x \in \overline{\mathbb{R}}$, so $\operatorname{sgn}(x) \geq 0$ iff $\operatorname{sgn}(\neg x) \leq 0$. Moreover,

$$P = \begin{bmatrix} a & b \\ c & d\end{bmatrix} \Rightarrow \neg P = \begin{bmatrix} a & b \\ -c & -d\end{bmatrix},\ P\neg = \begin{bmatrix} a & -b \\ c & -d\end{bmatrix}$$

Thus multiplying by $\neg$ from the left changes the signs of the bottom row and multiplying by $\neg$ from the right changes the signs of the right column.

Proposition 5.3 *For* $P \in \overline{\mathbb{M}}(\mathbb{R})$ *we have* $(P\neg)^c = \overline{\mathbb{R}} \setminus P^o$, $(P\neg)^o = \overline{\mathbb{R}} \setminus P^c$, $P^o \cap (P\neg)^o = \emptyset$, $P^c \cup (P\neg)^c = \overline{\mathbb{R}}$.

Proof: We have $x \in (P\neg)^c$ iff $\operatorname{sgn}(\neg P^{-1}x) = \operatorname{sgn}((P\neg)^{-1}x) \geq 0$ iff $\operatorname{sgn}(P^{-1}x) \leq 0$ iff $x \notin P^o$ and similarly, $x \in (P\neg)^o$ iff $x \notin P^c$. For $P = \left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right]$ we get $P^{-1}x = \frac{dx_0 - bx_1}{-cx_0 + ax_1}$, $(P\neg)^{-1}x = \frac{dx_0 - bx_1}{cx_0 - ax_1}$, so $P^o \cap (P\neg)^o = \emptyset$, $P^c \cup (P\neg)^c = \overline{\mathbb{R}}$. □

Definition 5.4 *Define the sign of a matrix* $M \in \overline{\mathbb{M}}(\mathbb{R})$ *by*

$$\operatorname{sgn}(M) = \begin{cases} 1 & \text{if } \exists\lambda \neq 0, \forall i,j, \lambda M_{ij} > 0\\ 0 & \text{if } \exists\lambda \neq 0, \forall i,j, \lambda M_{ij} \geq 0 \text{ and } \exists i,j, M_{ij} = 0\\ -1 & \text{if } \exists i,j,k,l, M_{ij} < 0 < M_{kl}\end{cases}$$

Proposition 5.5 *If* $P, Q \in \mathbb{M}(\mathbb{R})$ *then* $\operatorname{sgn}(Q^{-1}P) \geq 0$ *iff* $P^o \subseteq Q^o$ *iff* $P^c \subseteq Q^c$.

Proof: If P, Q are regular then $P^o \subseteq Q^o$ iff $P^c \subseteq Q^c$ since $P^c = P^o \cup \{\mathbf{l}(P), \mathbf{r}(P)\}$. If $\operatorname{sgn}(Q^{-1}P) \geq 0$, $x \in P^c$ then $\operatorname{sgn}(Q^{-1}x) = \operatorname{sgn}((Q^{-1}P)\cdot(P^{-1}x)) \geq 0$, so $x \in Q^c$ and therefore $P^c \subseteq Q^c$. Conversely assume by contradiction that $P^c \subseteq Q^c$ and $\operatorname{sgn}(Q^{-1}P) < 0$. If $P = \left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right]$ then $\frac{a}{c} \in P^c \subseteq Q^c$, $\frac{b}{d} \in P^c \subseteq Q^c$, so $\operatorname{sgn}(Q^{-1}\frac{a}{c}) \geq 0$, $\operatorname{sgn}(Q^{-1}\frac{b}{d}) \geq 0$. This means that both columns of $Q^{-1}P$ have nonnegative sign, and since $\operatorname{sgn}(Q^{-1}P) < 0$, they have the opposite sign. It follows that $\operatorname{sgn}(Q^{-1}P\neg) > 0$ and therefore $(P\neg)^c \subseteq Q^c$. We get $\overline{\mathbb{R}} = P^c \cup (P\neg)^c \subseteq Q^c$ and this is a contradiction since Q is assumed to be regular. □

Definition 5.6 *For $P, Q \in \overline{\mathbb{M}}(\mathbb{R})$ we write $P \subseteq Q$ if $\operatorname{sgn}(Q^{-1}P) \geq 0$. The image of a set $I \subseteq \overline{\mathbb{R}}$ by a transformation $M \in \overline{\mathbb{M}}(\mathbb{R})$ is defined by*

$$M(I) = \{y \in \overline{\mathbb{R}} : \exists x \in I : y = Mx\} = \{M(x) : x \in I\} \cap \overline{\mathbb{R}}.$$

If M is a singular transformation and $I = \{\mathbf{u}(M)\}$, then $M(I) = \emptyset$. If I contains a point different from $\mathbf{u}(M)$ then $M(I) = \{\mathbf{s}(M)\}$. For the zero transformation we have $\mathbf{0}(I) = \emptyset$ for every set $I \subseteq \overline{\mathbb{R}}$.

Proposition 5.7
1. If $P, Q \in \overline{\mathbb{M}}(\mathbb{R})$ then $P(Q^c) \subseteq (PQ)^c$.
2. If $P, Q \in \mathbb{M}(\mathbb{R})$ then $P(Q^c) = (PQ)^c$.

Proof: We use Proposition 3.37.
1. If $P = \mathbf{0}$ then $P(Q^c) = \emptyset$.
2. If $Q = \mathbf{0}$ then $(PQ)^c = \overline{\mathbb{R}}$.
3. Let $P \in \mathbb{M}^0(\mathbb{R})$, $Q \in \mathbb{M}(\mathbb{R})$. Then $P(Q^c) = \{\mathbf{s}(P)\}$, $\mathbf{s}(PQ) = \mathbf{s}(P)$, so either $(PQ)^c = \{\mathbf{s}(P)\}$, or $(PQ)^c = \overline{\mathbb{R}}$.
4. Let $Q \in \mathbb{M}^0(R)$. Then either $PQ = \mathbf{0}$ and $(PQ)^c = \overline{\mathbb{R}}$ or $PQ \in \mathbb{M}^0(\mathbb{R})$ and then $\mathbf{u}(PQ) = \mathbf{u}(Q)$. If $\mathbf{u}(Q) \geq 0$ then $(PQ)^c = \overline{\mathbb{R}}$. If $\mathbf{u}(Q) < 0$ then $P(Q^c) = \{P(\mathbf{s}(Q))\} = (PQ)^c$.
5. Let $P, Q \in \mathbb{M}(\mathbb{R})$. We have $y \in (PQ)^c$ iff $\operatorname{sgn}(Q^{-1}P^{-1}y) \geq 0$ iff $P^{-1}y \in Q^c$. This is equivalent to $y = PP^{-1}y \in P(Q^c)$. □

Proposition 5.8 *Let $P, Q \in \mathbb{M}(\mathbb{R})$ be regular matrices.*
1. If $M \in \overline{\mathbb{M}}(\mathbb{R})$ and $\operatorname{sgn}(Q^{-1}MP) \geq 0$, then $M(P^c) \subseteq Q^c$.
2. If $M \in \mathbb{M}(\mathbb{R})$ and $M(P^c) \subseteq Q^c$, then $\operatorname{sgn}(Q^{-1}MP) \geq 0$.

Proof: 1. If $M = \mathbf{0}$ is the zero matrix then $M(P^c) = \emptyset \subseteq Q^c$. If M is singular then $\operatorname{sgn}(Q^{-1}MP) \geq 0$ implies $\mathbf{s}(M) = \mathbf{s}(MP) \in Q^c$ so $M(P^c) = \{\mathbf{s}(M)\} \subseteq Q^c$. If M is regular then $\operatorname{sgn}(Q^{-1}MP) \geq 0$ implies $M(P^c) = (MP)^c \subseteq Q^c$ by Propositions 5.5 and 5.7.
2. If M is regular, then $(MP)^c = M(P^c) \subseteq Q^c$ by Proposition 5.7. By Proposition 5.5 we get $\operatorname{sgn}(Q^{-1}MP) \geq 0$. □

Proposition 5.9 *Define the* **size** *of a regular projective matrix $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$ by $\operatorname{sz}(P) = \frac{ab+cd}{|ad-bc|}$. Then $\operatorname{sz}(P) = \operatorname{sz}(P^c)$ (see Definition 3.2) and*

$$\begin{aligned} |P| &= \frac{1}{\pi}\operatorname{arccotg} \operatorname{sz}(P) = \frac{1}{2} - \frac{1}{\pi}\arctan \operatorname{sz}(P) \\ &= \begin{cases} \frac{1}{\pi}\arctan\frac{1}{\operatorname{sz}(P)} & \text{if } \operatorname{sz}(P) > 0 \\ \frac{1}{2} & \text{if } \operatorname{sz}(P) = 0 \\ \frac{1}{\pi}\arctan\frac{1}{\operatorname{sz}(P)} + 1 & \text{if } \operatorname{sz}(P) < 0 \end{cases} \end{aligned}$$

For the length of small intervals we have an estimate

$$\operatorname{sz}(P) > 1 \Leftrightarrow |P| < \frac{1}{4} \Rightarrow \frac{1}{4 \cdot \operatorname{sz}(P)} \leq |P| \leq \frac{1}{\pi \cdot \operatorname{sz}(P)}$$

Proof: We use Definition 3.2 and Proposition 3.3. If $\det(P) < 0$ then $\det(\left[\begin{smallmatrix} b & a \\ d & c \end{smallmatrix}\right]) = -\det(P) > 0$. If $\det(P) > 0$ then $\det(\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]) = \det(P) > 0$. In both cases we get $|P^c| = \frac{1}{\pi} \arccos \frac{ab+cd}{\sqrt{(a^2+c^2)(b^2+d^2)}}$. The rest of the proof follows from well-known trigonometric formulas. □

We have seen that a projective matrix can be regarded as a transformation, i.e., as a selfmap of $\overline{\mathbb{R}}$ or as an interval, i.e., a subset of $\overline{\mathbb{R}}$. We turn now to its third interpretation as an operator on intervals. We say that $M \in \mathbb{M}(\mathbb{R})$ is a **nonnegative projective matrix** if its sign is nonnegative. If M is nonnegative and $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{R})$, then by definition, $PM \subseteq P$, i.e., $(PM)^o \subseteq P^o$ and $(PM)^c \subseteq P^c$. For example if $M_0 = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $M_1 = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ then

$$PM_0 = \begin{bmatrix} a & a+b \\ c & c+d \end{bmatrix},\ PM_1 = \begin{bmatrix} a+b & b \\ c+d & d \end{bmatrix}.$$

If $\det(P) < 0$ then PM_0 is a left part of P and PM_1 is a right part of P. Consider an interval number system (F, W) over A. The intervals W_a are assumed proper and open so we represent them by matrices: from now on we assume that $W_a \in \mathbb{M}(\mathbb{R})$. For $u \in A^{n+1}$ we have

$$W_u = W^o_{u_0} \cap F_{u_0}(W^o_{u_1}) \cap F_{u_{[0,2)}}(W^o_{u_2}) \cap \cdots \cap F_{u_{[0,n)}}(W^o_{u_n}).$$

If $\mathcal{S}_{F,W}$ is a SFT of order 2, then $W_u = F_{u_{[0,n)}}(W^o_{u_n}) = (F_{u_{[0,n)}} W_{u_n})^o$ is an open interval (see Theorem 4.21) which is represented by the matrix $F_{u_{[0,n)}} W_{u_n}$. If $u \in \mathcal{S}_{F,W}$ is an infinite word, the intervals $W_{u_{[0,n)}}$ give ever better approximation to $\Phi(u)$. We can compute $W_{u_{[0,n+1)}}$ from $W_{u_{[0,n)}}$ by the **cut matrices** $H_{ab} = W_a^{-1} F_a W_b$. If $ab \in \mathcal{L}^2_{F,W}$ then $F_a W_b \subseteq W_a$, so H_{ab} is a nonnegative matrices and for $u \in \mathcal{L}^{n+1}_{F,W}$ we get

$$W_u = W_{u_0} H_{u_0 u_1} H_{u_1 u_2} \cdots H_{u_{n-1} u_n}$$

Indeed $W_{uab} = F_u F_a W_b = F_u W_a W_a^{-1} F_a W_b = W_{ua} H_{ab}$ is obtained by "cutting" W_{ua} by H_{ab}. For example for the number system of symmetric continued fractions of Definition 1.14 we get $H_{a0} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $H_{a1} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ for $a \in \{0, 1\}$ and $H_{a\overline{1}} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $H_{a\overline{0}} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ for $a \in \{\overline{1}, \overline{0}\}$. Thus each interval $W_u = (\frac{a}{c}, \frac{b}{d})$ is divided into two intervals $W_{u0} = (\frac{a}{c}, \frac{a+b}{c+d})$ and $W_{u1} = (\frac{a+b}{c+d}, \frac{b}{d})$ (see Figure 5.2).

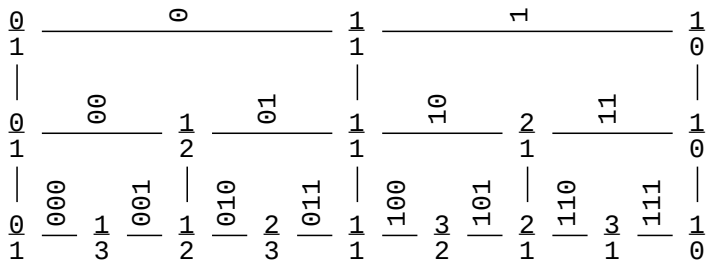


Figure 5.2: The cylinder intervals of the number system of symmetric continued fractions.

Similarly we can compute the intervals $\Phi([u])$ in a sofic number system (F, G, V). If $G = (B, E, \mathbf{i})$ is an initialized graph (i.e., $u \in \mathcal{L}_G$ iff $\mathbf{i} \xrightarrow{u}$), then $V_{\mathbf{i}} = \overline{\mathbb{R}}$. For each noninitial state $p \in B$, V_p is a proper closed interval which we represent by a matrix $V_p \in \mathbb{M}(\mathbb{R})$. For an edge $(p, a, q) \in E$ we define the **cut matrix** $H_{p,a,q}$ by

$$H_{p,a,q} = \begin{cases} F_a V_q & \text{if} \quad p = \mathbf{i} \\ V_p^{-1} F_a V_q & \text{if} \quad p \neq \mathbf{i} \end{cases}.$$

If $p \neq \mathbf{i}$ then $F_a V_q \subseteq V_p$, so $H_{p,a,q}$ is a nonnegative matrix. For a path $p_0 \xrightarrow{u_0} p_1 \xrightarrow{u_1} \cdots \xrightarrow{u_{n-1}} p_n$ we get

$$F_u V_{p_n} = \begin{cases} V_{p_0} H_{p_0,u_0,p_1} H_{p_1,u_1,p_2} \cdots H_{p_{n-1},u_{n-1},p_n} & \text{if} \quad p_0 \neq \mathbf{i}, \\ H_{p_0,u_0,p_1} H_{p_1,u_1,p_2} \cdots H_{p_{n-1},u_{n-1},p_n} & \text{if} \quad p_0 = \mathbf{i}. \end{cases}$$

5.2 The unary algorithm

Given a redundant sofic number system (F, G, V), we consider the **unary algorithm**, which computes a unary arithmetical operation $x \mapsto Mx$, where $M \in \mathbb{M}(\mathbb{R})$ is a Möbius transformation The input is a path $(p, u) \in \Sigma_{|G|}$ and the output is a path $(q, v) \in \Sigma_{|G|}$ such that $\Phi(v) = M\Phi(u)$. The computation takes infinite time but each finite prefix $(q_{[0,n)}, v_{[0,n)})$ of the output path is computed in a finite time from a finite prefix $(p_{[0,k_n)}, u_{[0,k_n)})$ of the input path.

The algorithm works by searching a path in the labelled **unary graph** whose edges are labelled by pairs $(a, b) \in (A \cup \{\lambda\})^2$ of input and output letters. An edge with label (a, λ) represents an **absorption** of a letter a from the input, an edge with label (λ, b) represents an **emission** of a letter b to the output. The label $(u, v) \in (A^*)^2$ of a finite path is the concatenation of the labels of its edges. Such a path represents the change of state upon reading the word u from the input and writing the word v to the output. We assume that the graph $G = (B, E, \mathbf{i})$ is initialized, i.e., $\mathbf{i} \in B$ and $u \in \Sigma_G$ iff there is a path with source $\mathbf{i}$ and label u.

Definition 5.10 *The unary graph of a sofic number system (F, G, V) with initialized graph $G = (B, E, \mathbf{i})$ is defined as follows: Its vertices are $(X, p, q) \in \mathbb{M}(\mathbb{R}) \times B^2$, its labelled edges are*

$$\begin{array}{rlll} \textit{absorption:} & (X,p,q) \xrightarrow{(a,\lambda)} (XH_{p,a,r}, r, q), & \textit{if} & p \xrightarrow{a} r \\ \textit{emission:} & (X,p,q) \xrightarrow{(\lambda,a)} (F_a^{-1}X, p, r) & \textit{if} & p \neq \mathbf{i}, q \xrightarrow{a} r, X \subseteq F_a V_r. \end{array}$$

The test $X \subseteq F_a V_r$ is evaluated by computing the sign of the matrix $(F_a V_r)^{-1} X$. Such a test is algorithmic provided the entries of F_a, V_r and X belong to a **computable ordered field** (see Section 7dfnordfield), for example to the field of rational numbers (see Chapter 6). Recall that the **cut matrix** of an edge $p \xrightarrow{a} q$ is $H_{p,a,q} = F_a V_q$ provided $p = \mathbf{i}$ and $H_{p,a,q} = V_p^{-1} F_a V_q$ otherwise. Define the **admissible set** of a vertex $(X, p, q) \in \mathbb{M}(\mathbb{R}) \times B^2$ by

$$\mathcal{A}(X,p,q) = \begin{cases} \emptyset & \text{if} \quad p = \mathbf{i} \\ \{(a,r) \in A \times B : \ q \xrightarrow{a} r, X \subseteq F_a V_r\} & \text{if} \quad p \neq \mathbf{i} \end{cases}$$

A redundant sofic system (F, G, V) has a **threshold** $\tau > 0$ such that $\mathcal{A}(X, p, q) \neq \emptyset$ whenever $p \neq \mathbf{i}$ and $|X| < \tau$. The threshold is the minimum of the Lebesgue numbers of the covers $\{\mathrm{int}_{V_p^c}(F_a(V_q^c)) : \ p \xrightarrow{a} q\}$ of V_p^c.

Proposition 5.11 *If $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{(u,v)} (Y, p, q)$ is a finite path, then $\mathbf{i} \xrightarrow{u} p$, $\mathbf{i} \xrightarrow{v} q$, $Y = F_v^{-1} X F_u V_p$. If $p \neq \mathbf{i} \neq q$ then $Y \subseteq V_q$. If $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{u,v}$ is an infinite path and $u, v \in A^\omega$, then $u, v \in \Sigma_G$ and $X(\Phi(u)) = \Phi(v)$.*

Proof: In the initial state $(X, \mathbf{i}, \mathbf{i})$ no emission is aplicable, so the first edge must be an absorption. If $\mathbf{i} \xrightarrow{u} a$ is a path in G and $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{(u,\lambda)} (XF_u V_p, p, \mathbf{i}) \xrightarrow{(\lambda,a)} (Y, p, q)$ is a path in the unary graph up to the first emission, then $XF_u V_p \subseteq F_a V_q$, so $Y = F_a^1 XF_u V_p \subseteq V_q$. Assume that the condition is satisfied for a path $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{(u,v)} (Y, p, q)$. If $(Y, p, q) \xrightarrow{(a,\lambda)} (Z, r, q)$ is an absorption, then $Z = YH_{p,a,r} = F_v^{-1} XF_u V_p H_{p,a,r} = F_v^{-1} XF_{ua} V_r$ and $Z \subseteq Y \subseteq V_q$. If $(Y, p, q) \xrightarrow{(\lambda,a)} (Z, p, r)$ is an emission, then $Z = F_a^{-1} Y = F_{va}^{-1} XF_u V_p$. Since $Y \subseteq F_a V_r$, we get $Z \subseteq V_r$. Let $(u, v) \in (A^\omega)^2$

be a label of an infinite path with source $(X, \mathbf{i}, \mathbf{i})$. Then for each n there exists k_n and a path $(X, \mathbf{i}, \mathbf{i}) \stackrel{(u_{[0,k_n)}, v_{[0,n)})}{\longrightarrow} (Y_n, p_n, q_n)$. Thus $F^{-1}_{v_{[0,n)}} X F_{u_{[0,k_n)}} V_{p_n} \subseteq V_{q_n}$ and therefore $X F_{u_{[0,k_n)}} V_{p_n} \subseteq F_{v_{[0,n)}} V_{q_n}$. We get $X(\Phi(u)) \in (X F_{u_{[0,k_n)}} V_{p_n})^c \subseteq (F_{v_{[0,n)}} V_{q_n})^c$, and $\Phi(v) \in (F_{v_{[0,n)}} V_{q_n})^c$. Since the length of these intervals converges to zero, we get $\Phi(v) = X(\Phi(u))$. □

A path $(X_0, \mathbf{i}, \mathbf{i}) \stackrel{u_0,v_0}{\longrightarrow} (X_1, p_1, q_1) \stackrel{u_1,v_1}{\longrightarrow} (X_2, p_2, q_2) \stackrel{u_2,v_2}{\longrightarrow} \cdots$ in the unary graph projects to an input path $\mathbf{i} \stackrel{u_0}{\longrightarrow} p_1 \stackrel{u_1}{\longrightarrow} p_2 \stackrel{u_2}{\longrightarrow} \cdots$ and to an output path $\mathbf{i} \stackrel{v_0}{\longrightarrow} q_1 \stackrel{v_1}{\longrightarrow} q_2 \stackrel{v_2}{\longrightarrow} \cdots$. Some edges in these projected paths are of the form $p \stackrel{\lambda}{\longrightarrow} p$. The unary graph represents a nondeterministic algorithm for computing the symbolic representation of M. From each state (X, p, q) there leads several absorption edges and none, one or several emission edges. To get a deterministic algorithm, we consider a **selector** s which at each state chooses an emission, i.e., an element of $\mathcal{A}(X, p, q)$ provided $\mathcal{A}(X, p, q) \neq \emptyset$. If $\mathcal{A}(X, p, q) = \emptyset$ then s chooses an absorption. This is indicated by $s(X, p, q) = \mathrm{x}$.

Definition 5.12

1. *A* **unary selector** *for a sofic number system (F, G, V) is a mapping $s : \mathbb{M}(\mathbb{R}) \times B^2 \to (A \times B) \cup \{x\}$ such that if $s(X, p, q) = (a, r) \in A \times B$ then $(a, r) \in \mathcal{A}(X, p, q)$.*
2. *If $s(X, p, r) = x$ then we say that (X, p, q) is an* **absorption state** *of s, otherwise (X, p, q) is an* **emission state***.*
3. *A selector s is* **greedy** *if $s(X, p, q) \in \mathcal{A}(X, p, q)$ whenever $\mathcal{A}(X, p, q) \neq \emptyset$..*

If all entries of matrices F_a, V_p are integers, the state matrices X can be stored with integer entries whose GCD (greatest common divisor) is 1 (see Chapter 6). After each step, the entries of the state matrix X are cancelled by their common GCD. If the admissible set contains more than one element, a reasonable selection is the choice of the edge $p \stackrel{a}{\longrightarrow} p'$ which gives the smallest norm of the result $F_a^{-1} X$. A selector s determines for each input transformation $M \in \mathbb{M}(\mathbb{R})$ and input path $(p, u) \in \Sigma_{|G|}$ a unique output path $\Theta_{M,s}(p, u) = (q, v) \in \Sigma_{|G|} \cup \mathcal{L}_{|G|}$ such that

$$(M, \mathbf{i}, \mathbf{i}) \stackrel{u_0,v_0}{\longrightarrow} (X_1, p_1, q_1) \stackrel{u_1,v_1}{\longrightarrow} (X_2, p_2, q_2) \stackrel{u_2,v_2}{\longrightarrow} \cdots$$

is an infinite path in the unary graph. Here u_i, v_i may be empty, so they are not necessarily the i-th letters of u or v. If $s(X_i, p_i, q_i) = \mathrm{x}$, then $u_i \neq \lambda$, $v_i = \lambda$. If $s(X_i, p_i, q_i) \neq \mathrm{x}$ then $u_i = \lambda$ and $(v_i, q_{i+1}) = s(X_i, p_i, q_i)$. The image $\Theta_{M,s}(p, u)$ of an infinite path may be a finite path. In redundant systems with a greedy selector, an infinite input yields an infinite output:

Theorem 5.13 *If (F, G, V) is a redundant sofic number system, then for any greedy selector s and an initial state matrix $M \in \mathbb{M}(\mathbb{R})$ the mapping $\Theta_{M,s} : \Sigma_{|G|} \to \Sigma_{|G|}$ is continuous. If $(q, v) = \Theta_{M,s}(p, u)$, then $M(\Phi(u)) = \Phi(v)$.*

Proof: Since (F, G, V) is redundant it has a threshold $\tau > 0$ which is the minimum of the Lebesgue numbers of $\{\mathrm{int}_{V_p^c}(F_a(V_q^c)) : p \stackrel{a}{\longrightarrow} q\}$. Thus if $I \subseteq V_p^c$ and $|I| \leq \tau$ then there exists $p \stackrel{a}{\longrightarrow} q$ such that $I \subseteq F_a(V_q^c)$. We show that each infinite path of the selector contains an infinite number of both absorptions and emissions. Assume by contradiction that (X_i, p_i, q_i) is an infinite path which consists only of absorptions, so its label is (u, λ) with $u \in \Sigma_G$. Since $\lim_{n\to\infty} |F_{u_{[0,n)}} V_{p_n}| = 0$, we get $\lim_{n\to\infty} |X_0 F_{u_{[0,n)}} V_{p_n}| = 0$ by the continuity of X_0, and therefore $|X_0 F_{u_{[0,n)}} V_{p_n}| \leq \tau$ for some n, which is a contradiction. Assume that there exists an infinite path consisting only of emissions. Then by Proposition 3.33 the length of the intervals X_i grows until it exceeds the length of any V_q, and this is a contradiction. The rest of the proof follows from Proposition 5.11. □

p	V_p	a	q	F_aV_q	$H_{p,a,q}$
λ	$\overline{\mathbb{R}}$	0	0	$[\frac{-1}{2}, \frac{1}{2}]$	$[\frac{-1}{2}, \frac{1}{2}]$
		1	1	$[\frac{1}{4}, \frac{2}{2}]$	$[\frac{1}{4}, \frac{2}{2}]$
		$\overline{0}$	$\overline{0}$	$[\frac{1}{2}, \frac{1}{-2}]$	$[\frac{1}{2}, \frac{1}{-2}]$
		$\overline{1}$	$\overline{1}$	$[\frac{-2}{2}, \frac{-1}{4}]$	$[\frac{-2}{2}, \frac{-1}{4}]$
0	$[\frac{-1}{1}, \frac{1}{1}]$	$\overline{1}$	$\overline{1}$	$[\frac{-2}{2}, \frac{-1}{4}]$	$[\frac{4}{0}, \frac{5}{3}]$
		0	0	$[\frac{-1}{2}, \frac{1}{2}]$	$[\frac{3}{1}, \frac{1}{3}]$
		1	1	$[\frac{1}{4}, \frac{2}{2}]$	$[\frac{3}{5}, \frac{0}{4}]$
1	$[\frac{-1}{2}, \frac{1}{1}]$	0	0	$[\frac{-1}{2}, \frac{1}{2}]$	$[\frac{3}{0}, \frac{1}{4}]$
		1	1	$[\frac{1}{4}, \frac{2}{2}]$	$[\frac{1}{2}, \frac{0}{2}]$
$\overline{0}$	$[\frac{1}{4}, \frac{1}{-4}]$	1	1	$[\frac{1}{4}, \frac{2}{2}]$	$[\frac{4}{0}, \frac{5}{3}]$
		$\overline{0}$	$\overline{0}$	$[\frac{1}{2}, \frac{1}{-2}]$	$[\frac{3}{1}, \frac{1}{3}]$
		$\overline{1}$	$\overline{1}$	$[\frac{-2}{2}, \frac{-1}{4}]$	$[\frac{3}{5}, \frac{0}{4}]$
$\overline{1}$	$[\frac{-1}{1}, \frac{1}{2}]$	$\overline{1}$	$\overline{1}$	$[\frac{-2}{2}, \frac{-1}{4}]$	$[\frac{2}{0}, \frac{2}{1}]$
		0	0	$[\frac{-1}{2}, \frac{1}{2}]$	$[\frac{4}{1}, \frac{0}{3}]$

X^c	$XH_{p,u,p'}\quad F_v^{-1}X$	$p \xrightarrow{u} p'$	$q \xrightarrow{v} q'$
$[3.00, 0.33]$	$\begin{bmatrix}3 & 1\\1 & 3\end{bmatrix}\begin{bmatrix}1 & 1\\2 & -2\end{bmatrix}$	$0 \xrightarrow{\overline{0}} \overline{0}$	
$[0.71, -0.20]$	$\begin{bmatrix}5 & 1\\7 & -5\end{bmatrix}\begin{bmatrix}4 & 5\\0 & 3\end{bmatrix}$	$\overline{0} \xrightarrow{1} 1$	
$[0.71, 1.40]$	$\begin{bmatrix}1 & 0\\0 & 2\end{bmatrix}\begin{bmatrix}5 & 7\\7 & 5\end{bmatrix}$		$\lambda \xrightarrow{\overline{0}} \overline{0}$
$[0.36, 0.70]$	$\begin{bmatrix}2 & -1\\0 & 1\end{bmatrix}\begin{bmatrix}5 & 7\\14 & 10\end{bmatrix}$		$\overline{0} \xrightarrow{1} 1$
$[-0.29, 0.40]$	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}-2 & 2\\7 & 5\end{bmatrix}$		$1 \xrightarrow{0} 0$
$[-0.57, 0.80]$	$\begin{bmatrix}-4 & 4\\7 & 5\end{bmatrix}\begin{bmatrix}1 & 0\\2 & 2\end{bmatrix}$	$1 \xrightarrow{1} 1$	
$[0.24, 0.80]$	$\begin{bmatrix}4 & 8\\17 & 10\end{bmatrix}\begin{bmatrix}1 & 0\\2 & 2\end{bmatrix}$	$1 \xrightarrow{1} 1$	
$[0.54, 0.80]$	$\begin{bmatrix}2 & -1\\0 & 1\end{bmatrix}\begin{bmatrix}20 & 16\\37 & 20\end{bmatrix}$		$0 \xrightarrow{1} 1$
$[0.08, 0.60]$	$\begin{bmatrix}3 & 12\\37 & 20\end{bmatrix}\begin{bmatrix}1 & 0\\2 & 2\end{bmatrix}$	$1 \xrightarrow{1} 1$	
$[0.35, 0.60]$	$\begin{bmatrix}2 & -1\\0 & 1\end{bmatrix}\begin{bmatrix}27 & 24\\77 & 40\end{bmatrix}$		$1 \xrightarrow{1} 1$
$[-0.30, 0.20]$	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}-23 & 8\\77 & 40\end{bmatrix}$		$1 \xrightarrow{0} 0$
$[-0.60, 0.40]$	$\begin{bmatrix}-46 & 16\\77 & 40\end{bmatrix}\begin{bmatrix}1 & 0\\2 & 2\end{bmatrix}$	$1 \xrightarrow{1} 1$	
$[-0.09, 0.40]$	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}-14 & 32\\157 & 80\end{bmatrix}$		$0 \xrightarrow{0} 0$
$[-0.18, 0.80]$	$\begin{bmatrix}-28 & 64\\157 & 80\end{bmatrix}\begin{bmatrix}1 & 0\\2 & 2\end{bmatrix}$	$1 \xrightarrow{1} 1$	
$[0.32, 0.80]$	$\begin{bmatrix}2 & -1\\0 & 1\end{bmatrix}\begin{bmatrix}100 & 128\\317 & 160\end{bmatrix}$		$0 \xrightarrow{1} 1$
$[-0.37, 0.60]$	$\begin{bmatrix}-117 & 96\\317 & 160\end{bmatrix}\begin{bmatrix}3 & 1\\0 & 4\end{bmatrix}$	$1 \xrightarrow{0} 0$	
$[-0.37, 0.28]$	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}-117 & 89\\317 & 319\end{bmatrix}$		$1 \xrightarrow{0} 0$
$[-0.74, 0.56]$	$\begin{bmatrix}-234 & 178\\317 & 319\end{bmatrix}\begin{bmatrix}3 & 0\\5 & 4\end{bmatrix}$	$0 \xrightarrow{1} 1$	
$[0.07, 0.56]$	$\begin{bmatrix}94 & 356\\1273 & 638\end{bmatrix}\begin{bmatrix}3 & 1\\0 & 4\end{bmatrix}$	$1 \xrightarrow{0} 0$	
$[0.07, 0.40]$	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}94 & 506\\1273 & 1275\end{bmatrix}$		$0 \xrightarrow{0} 0$

input matrix $M = \begin{bmatrix}3 & 1\\1 & 3\end{bmatrix}$

input: $u = \overline{0}111111010$, $p_n = 0$, $F_uV_{p_n} = \begin{bmatrix}505 & 507\\256 & 256\end{bmatrix} = [1.97266, 1.98047]$

result: $MF_uV_{p_n} = \begin{bmatrix}1771 & 1777\\1273 & 1275\end{bmatrix} = [1.39120, 1.39373]$

output: $v = \overline{0}101100100$, $q_m = 0$, $F_vV_{q_m} = \begin{bmatrix}355 & 357\\256 & 256\end{bmatrix} = [1.38672, 1.39453]$

Table 5.1: The computation of the unary algorithm (right) in the binary signed system from Figure 4.9 (left).

If G is a deterministic graph then each word $u \in \Sigma_G$ determines a unique path $\mathbf{i} \xrightarrow{u}$ with label u. Thus there exists a continuous mapping $\Theta_{M,s} : \Sigma_G \to \Sigma_G$ such that $\Phi\Theta_{M,s} = M\Phi$. In Table 5.1 we give the graph of the binary signed system from Figure 4.9 (left) and the computation of the unary algorithm in the system (right).

For a nonredundant system, the unary algorithm with a greedy selector need not work. It may happen that ever smaller intervals X contain a point which does not belong to the interior of any F_aV_r, so the condition $X \subseteq F_aV_r$ is never met and the output remains finite. In this case X is a subset of a union $F_aV_{q_0} \cup F_bV_{q_1}$ of two neighboring intervals. Thus we know that the output is either a or b and we may pursue both these possibilities in two parallel branches. These two branches may coexist indefinitely, giving two output words v, w such that $\Phi(M(u)) = \Phi(v) = \Phi(w)$. It may also happen that at some later step one of the branches ceases to represent an output with $\Phi(M(u)) = \Phi(v)$ and is therefore closed. In these parallel branches with states (X, p, q) we do not always have $X \subseteq V_q$ but only $\emptyset \neq X \cap V_q$. If $\emptyset = X \cap V_q$, then the branch is closed. On the other hand if $X \subseteq V_q$, then the branch represents the correct computation and the other branch is closed.

The nondeterministic algorithm based on these principles is given by the **branching unary graph** in Definition 5.14. Since the two branches have different output words, we incorporate the output word to the state. Thus a state (or a vertex of the graph) is (X, p, q, v), where $v \in \mathcal{L}_G$ is the output word. The edges are labelled only by the input letters. A vertex of the graph is either a single state (X, p, q, v) or a pair of states $((X_0, p, q_0, v), (X_1, p, q_1, w))$ with the same input vertex p. The initial state is $(X, \mathbf{i}, \mathbf{i}, \lambda)$. There are branching edges from a single state to a pair of states and closing edges which close one of the branches. If the vertex is a pair of states, an absorption is applied to both states simultaneously. On the other hand, an emissions is applied only to one of the states.

Definition 5.14 *The branching unary graph of a sofic number system (F, G, V) with deterministic graph $G = (V, E, \mathbf{i})$ is defined as follows: Its vertices are either $(X, p, q, v) \in \mathbb{M}(\mathbb{R}) \times B^2 \times \mathcal{L}_G$, or pairs $((X_0, p, q_0, v), (X_1, p, q_1, w))$ of vertices. The labelled edges are*

absorption:	(X, p, q, v)	$\xrightarrow{a}$	$(XH_{p,a,p'}, p', q, v)$,	*if* $p \xrightarrow{a} p'$
absorption:	(X_0, p, q_0, v) (X_1, p, q_1, w)	$\xrightarrow{a}$	$(X_0H_{p,a,p'}, p', q_0, v)$ $(X_1H_{p,a,p'}, p', q_1, w)$	*if* $p \xrightarrow{a} p'$
emission:	(X, p, q, v)	$\xrightarrow{\lambda}$	$(F_a^{-1}X, p, q', va)$	*if* $p \neq \mathbf{i}, q \xrightarrow{a} q'$ $\emptyset \neq X \cap V_q \subseteq F_aV_{q'}$,
branching:	(X, p, q, v)	$\xrightarrow{\lambda}$	$(F_a^{-1}X, p, q_0, va)$ $(F_b^{-1}X, p, q_1, vb)$	*if* $p \neq \mathbf{i}, q \xrightarrow{a} q_0, q \xrightarrow{b} q_1$ $X \subseteq V_q \cap (F_aV_{q_0} \cup F_bV_{q_1})$,
closing:	(X_0, p, q_0, v) (X_1, p, q_1, w)	$\xrightarrow{\lambda}$	(X_1, p, q_1, w)	*if* $\emptyset = X_0 \cap V_{q_0}$ *or* $X_1 \subseteq V_{q_1}$
closing:	(X_0, p, q_0, v) (X_1, p, q_1, w)	$\xrightarrow{\lambda}$	(X_0, p, q_0, v)	*if* $\emptyset = X_1 \cap V_{q_1}$ *or* $X_0 \subseteq V_{q_0}$

To obtain a deterministic algorithm, we should define a selector which selects one of the possible edges. The closing edges should be chosen whenever they are applicable: the branch to be closed does not represent any possible output. A branching edge should be chosen if the interval X becomes too small. One possibility is to define small open intervals V_{q_0,q_1} which contain the common endpoints of $F_aV_{q_0} \cap F_bV_{q_1}$ and opt for the branching when $X \subseteq V_{q_0,q_1}$. For appropriate selectors, the input word $u \in \Sigma_G$ yields an infinite path with label u. The words v, w of the states of the path give either a single output v with $\Phi(M(u)) = \Phi(v)$ or two output words with $\Phi(M(u)) = \Phi(v) = \Phi(w)$.

5.3 Bilinear tensors

Binary arithmetical operations like addition or multiplication are obtained from bilinear functions $T : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$. While a linear function $M : \mathbb{R}^2 \to \mathbb{R}^2$ is a 1-contravariant and 1-covariant tensor, a bilinear function is a 1-contravariant and 2-covariant tensor given by $T(x,y)_k = \sum_{i=0}^{1}\sum_{j=0}^{1} T_{kij}x_iy_j$ (see e.g., Bishop and Goldberg [6]). The tensor T determines a function $T : \overline{\mathbb{R}} \times \overline{\mathbb{R}} \to \overline{\mathbb{R}} \cup \{\frac{0}{0}\}$ defined by

$$\begin{aligned} T(x,y) &= \frac{(T_{000}x_0 + T_{010}x_1)y_0 + (T_{001}x_0 + T_{011}x_1)y_1}{(T_{100}x_0 + T_{110}x_1)y_0 + (T_{101}x_0 + T_{111}x_1)y_1} \\ &= \frac{(T_{000}y_0 + T_{001}y_1)x_0 + (T_{010}y_0 + T_{011}y_1)x_1}{(T_{100}y_0 + T_{101}y_1)x_0 + (T_{110}y_0 + T_{111}y_1)x_1} \end{aligned}$$

For example $T(x,y) = x + y = \frac{x_0y_1+x_1y_0}{x_1y_1}$. A nonzero multiple of a tensor defines the same function on $\overline{\mathbb{R}} \times \overline{\mathbb{R}}$, so tensors are conceived as points of the projective space $\mathbb{P}(\mathbb{R}^{2\times2\times2})$. Denote by $\overline{\mathbb{T}}(\mathbb{R}) = \mathbb{P}(\mathbb{R}^{2\times2\times2}) \cup \{\mathbf{0}\}$ the set of all projective tensors of dimension at most 1. We write tensors as (2×4)-matrices $T = \left[\begin{smallmatrix} T_{000} & T_{010} & T_{001} & T_{011} \\ T_{100} & T_{110} & T_{101} & T_{111} \end{smallmatrix}\right]$. For a tensor T and projective vectors $x, y, z \in \overline{\mathbb{R}}$ we have projective matrices T^*x, T_*y, zT obtained by different kinds of multiplication:

$$(T^*x)_{kj} = \sum_i T_{kij}x_i,\ (T_*y)_{ki} = \sum_j T_{kij}y_j,\ (zT)_{ij} = \sum_k z_kT_{kij}.$$

Then $(T^*x)y = (T_*y)x = T(x,y)$.

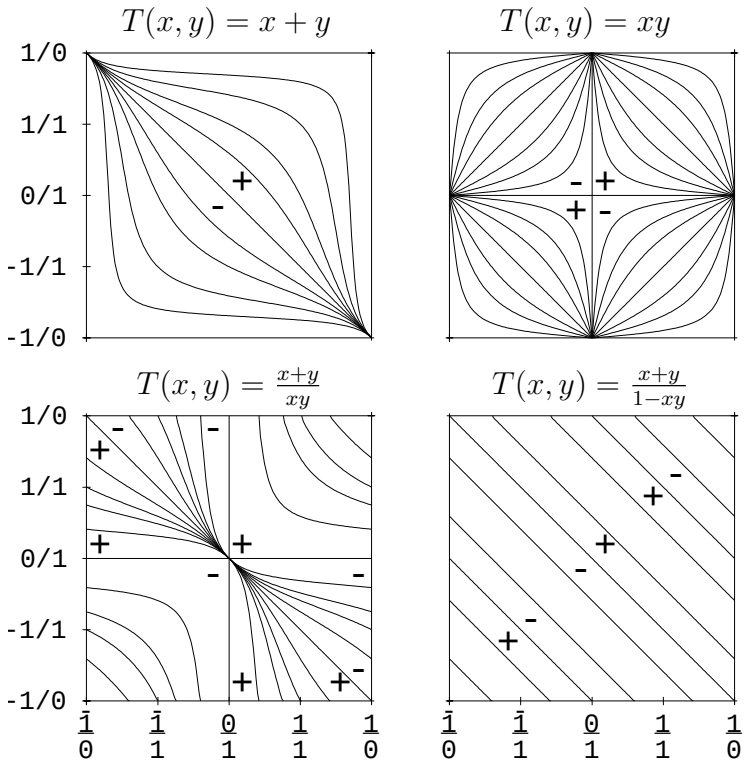


Figure 5.3: Level curves of bilinear tensors with marked positivity and negativity regions. The straight lines in the last case follow from the formula $\frac{\tan t+\tan s}{1-\tan t\cdot\tan s} = \tan(t+s)$.

$T(x,y)$	T^*x	$\det(T^*x)$	D_x	$S(T)$
$x+y$	$\left[\begin{smallmatrix} x_1 & x_0 \\ 0 & x_1 \end{smallmatrix}\right]$	x_1^2	0	$\{(\frac{1}{0},\frac{1}{0})\}$
xy	$\left[\begin{smallmatrix} x_0 & 0 \\ 0 & x_1 \end{smallmatrix}\right]$	x_0x_1	1	$\{(\frac{0}{1},\frac{1}{0}),(\frac{1}{0},\frac{0}{1})\}$
$\frac{x+y}{xy}$	$\left[\begin{smallmatrix} x_1 & x_0 \\ x_0 & 0 \end{smallmatrix}\right]$	$-x_0^2$	0	$\{(\frac{0}{1},\frac{0}{1})\}$
$\frac{x+y}{1-xy}$	$\left[\begin{smallmatrix} x_1 & x_0 \\ -x_0 & x_1 \end{smallmatrix}\right]$	$x_0^2+x_1^2$	-4	$\emptyset$

Table 5.2: Singular points of tensors

Bilinear tensors can be classified according to the number of their singular marginal matrices. For a given tensor T consider the quadratic form

$$\begin{aligned}\det(T^*x) &= \det\begin{bmatrix} T_{000}x_0+T_{010}x_1 & T_{001}x_0+T_{011}x_1 \\ T_{100}x_0+T_{110}x_1 & T_{101}x_0+T_{111}x_1 \end{bmatrix} \\ &= Ax_0^2+Bx_0x_1+Cx_1^2.\end{aligned}$$

Denote by $D_x(T)=B^2-4AC$ the discriminant of $\det(T^*x)$. If $A=B=C=0$ then T^*x is singular for every $x\in\overline{\mathbb{R}}$. Assume that at least one of the A,B,C is nonzero. If $D_x<0$ then T^*x is regular for every $x\in\overline{\mathbb{R}}$. If $D_x=0$ then there exists one point $x\in\overline{\mathbb{R}}$ with singular T^*x. If $D_x>0$ then there exist two points x with singular T^*x. If T^*x is singular and $y=\mathbf{u}(T^*x)$ then $T(x,y)=\frac{0}{0}$. We say that (x,y) is a **singular point** of T. Denote by $S(T)$ the set of singular points of a tensor (see Table 5.2). A tensor may be visualized by its **level curves**

$$T^{-1}(z)=\{(x,y)\in\overline{\mathbb{R}}^2:\ T(x,y)=z\}.$$

In singular points with $T(x,y)=\frac{0}{0}$ the level curves intersect (see Figure 5.3).

For a tensor T and a matrix P we define tensors T^*P, T_*P and PT by

$$(T^*P)_{kij}=\sum_p T_{kpj}P_{pi},\ (T_*P)_{kij}=\sum_q T_{kiq}P_{qj},\ (PT)_{kij}=\sum_r P_{kr}T_{rij}.$$

Then $(T^*P)^*x=T^*(Px)$, $(T_*P)_*y=T_*(Py)$. The operations with the first and second argument commute, so we adopt notations

$$\begin{aligned}T(x,y) &= (T^*x)y=(T_*y)x,\\ T(x,Q) &= (T^*x)Q=(T_*Q)^*x,\\ T(P,y) &= (T_*y)P=(T^*P)_*y,\\ T(P,Q) &= (T_*P)^*Q=(T^*Q)_*P.\end{aligned}$$

The multiplication from the left commutes with the multiplication from the right, so we write $PT^*Q=P(T^*Q)=(PT)^*Q$ for $P,Q\in\overline{\mathbb{M}}(\mathbb{R})$. For vectors $x,y\in\overline{\mathbb{R}}$ we have $(xT)y=x(T_*y)$, $x(yT)=y(T^*x)$. For a matrix $M=\left[\begin{smallmatrix} M_{00} & M_{01} \\ M_{10} & M_{11}\end{smallmatrix}\right]$ we denote its left and right columns by $M_{-0}=\frac{M_{00}}{M_{10}}$, $M_{-1}=\frac{M_{01}}{M_{11}}$, and the upper and lower row by $M_{0-}=\frac{M_{00}}{M_{01}}$, $M_{1-}=\frac{M_{10}}{M_{11}}$, so $(M_{-j})_i=M_{ij}$, $(M_{i-})_j=M_{ij}$. Similarly for a tensor T we denote by T_{k--}, T_{-i-}, T_{--j} the **marginal matrices** obtained from T by fixing a coordinate, and T_{-ij}, T_{k-j}, T_{ki-} **marginal vectors** obtained by fixing two coordinates. A simple algebra shows that the tensor $T(P,Q)$ consists of T-images of the endpoints of P and Q:

Proposition 5.15 *For a tensor T and matrices P, Q we have*

$$T(P,Q)_{-i-} = T(P_{-i}, Q),\ T(P,Q)_{--j} = T(P, Q_{-j}),\ T(P,Q)_{-ij} = T(P_{-i}, Q_{-j}).$$

Proof:

$$(T(P,Q)_{-i-})_{kj} \ = \ T(P,Q)_{kij} = \sum_{pq} T_{kpq} P_{pi} Q_{qj} = \sum_{pq} T_{kpq} (P_{-i})_p Q_{qj} = T(P_{-i}, Q)_{kj}$$

and similarly in other cases. □

Definition 5.16 *The image of sets $I, J \subseteq \overline{\mathbb{R}}$ by a tensor T is defined by*

$$\begin{aligned} T(I,J) &= \{T(x,y) :\ x \in I, y \in J\} \cap \overline{\mathbb{R}} \\ &= \{z \in \overline{\mathbb{R}} :\ \exists x \in I, \exists y \in J, z = T(x,y)\} \end{aligned}$$

In arithmetical algorithms we verify whether the image $T(I,J)$ of intervals I, J is included in a given interval K. We have an inclusion criterion which is formally similar to the criterion of the inclusion of intervals. The sign of a tensor is defined similarly as the sign of a matrix: it is nonnegative if there exists nonzero λ such that all λT_{kij} are nonnegative.

Proposition 5.17 (Algebraic inclusion criterion) *Let $T \in \overline{\mathbb{T}}(\mathbb{R})$ be a tensor and $P, Q, R \in \mathbb{M}(\mathbb{R})$ regular matrices. If $\operatorname{sgn}(R^{-1}T(P,Q)) \geq 0$ then $T(P^c, Q^c) \subseteq R^c$.*

Proof: Let $x \in P^c$, $y \in Q^c$ and $z = T(x,y) \in \overline{\mathbb{R}}$. Since P is regular, for $u = P^{-1}x$ we have $\operatorname{sgn}(u) \geq 0$ and $x = Pu$, so

$$(T^*x)Q = (T^*(Pu))Q = ((T^*P)^*u)Q = ((T^*P)_*Q)^*u = T(P,Q)^*u.$$

It follows $\operatorname{sgn}(R^{-1}(T^*x)Q) = \operatorname{sgn}(R^{-1}(T(P,Q)^*u) \geq 0$, so $(T^*x)(Q^c) \subseteq R^c$ by Proposition 5.8 and therefore $z \in R^c$. Thus we have proved $T(P^c, Q^c) \subseteq R^c$. □

Theorem 5.17 has a converse for regular tensors.

Definition 5.18 *We say that T is a* **regular tensor***, if for each $x, y, z \in \overline{\mathbb{R}}$, the matrices zT, T^*x, T_*y are nonzero. Denote by $\mathbb{T}(\mathbb{R})$ the space of regular tensors.*

A tensor is regular iff its pairs of marginal matrices are linearly independent, i.e., if $T_{0--} \neq T_{1--}$, $T_{-0-} \neq T_{-1-}$ and $T_{--0} \neq T_{--1}$ are different points of the projective space $\mathbb{P}(\mathbb{R}^{2\times 2})$. Examples of regular tensors are $\left[\begin{smallmatrix}1&0&0&0\\0&0&0&1\end{smallmatrix}\right]$ (multiplication), $\left[\begin{smallmatrix}0&1&1&0\\0&0&0&1\end{smallmatrix}\right]$ (addition), or $\left[\begin{smallmatrix}0&1&0&0\\0&0&1&0\end{smallmatrix}\right]$ (division).

Proposition 5.19 *If T is a regular tensor and M is a regular matrix, then MT, T^*M and T_*M are regular tensors.*

Proof: If $x \in \overline{\mathbb{R}}$, $Mx \in \overline{\mathbb{R}}$ and $(T^*M)^*x = T^*(Mx)$ is nonzero. Since T_*x is nonzero, $(T^*M)_*x = (T_*x)M$ is nonzero. Since zT is nonzero, $z(T^*M) = M^T(zT)$ is nonzero (here M^T is the transposed matrix of M. Thus we have proved that T^*M is regular. Similarly we show that T_*M is regular. Since $(MT)_*x = M(T_*x)$, $(MT)^*x = M(T^*x)$, $z(MT) = (zM)T$, we get that MT is regular. □

Proposition 5.20 *If T is a regular tensor then* $\det(T^*x)$ *is a nonzero quadratic form.*

Proof: Assume by contradiction that

$$\det(T^*x) = \det\begin{bmatrix} T_{000}x_0 + T_{010}x_1 & T_{001}x_0 + T_{011}x_1 \\ T_{100}x_0 + T_{110}x_1 & T_{101}x_0 + T_{111}x_1 \end{bmatrix}.$$

is a zero quadratic form, so we have zero coefficients at x_0^2 and x_1^2:

$$\det\begin{bmatrix} T_{000} & T_{001} \\ T_{100} & T_{101} \end{bmatrix} = 0,\ \det\begin{bmatrix} T_{010} & T_{011} \\ T_{110} & T_{111} \end{bmatrix} = 0$$

It follows that there exist a_i, b_i, T_{ij} such that

$$T^*x = \begin{bmatrix} a_0T_{00}x_0 + b_0T_{10}x_1 & a_0T_{01}x_0 + b_0T_{11}x_1 \\ a_1T_{00}x_0 + b_1T_{10}x_1 & a_1T_{01}x_0 + b_1T_{11}x_1 \end{bmatrix}$$

For the coefficient at x_0x_1 we get $(a_0b_1 - a_1b_0) \cdot (T_{00}T_{11} - T_{01}T_{10}) = 0$. If $a_0b_1 - a_1b_0 = 0$ then $a_1T_{0--} = a_0T_{1--}$ so T_{0--} and T_{1--} are linearly dependent. If $T_{00}T_{11} - T_{01}T_{10} = 0$ then $T_{11}T_{--0} = T_{10}T_{--1}$, so T_{--0} and T_{--1} are linearly dependent. In both cases, T is not regular and this is a contradiction. □

Proposition 5.21 *If T is a regular tensor, P, Q, R are regular matrices and $T(P^c, Q^c) \subseteq R^c$, then* $\operatorname{sgn}(R^{-1}T(P, Q)) \geq 0$.

Proof: We show that for $x \in [0, \infty]$ we have $(T^*(Px))(Q^c) \subseteq R^c$. Indeed if $z \in (T^*(Px))(Q^c)$ then there exists $y \in Q^c$ such that $z = (T^*(Px))(y) = T^*(Px, y)$. Since $Px \in P^c$, we get $z \in T(P^c, Q^c) \subseteq R^c$. Since T is a regular tensor, $T^*(Px)$ is a nonzero matrix and therefore $M(x) = R^{-1}(T^*(Px))Q$ is a nonzero matrix too. We can therefore norm it and assume that $||M(x)||^2 = \sum_{ij} M(x)_{ij}^2 = 1$. Since $(T^*(Px))(Q^c) \subseteq R^c$, by Theorem 5.8 we get $\operatorname{sgn}(R^{-1}(T^*(Px))Q) \geq 0$ whenever $T^*(Px)$ is a regular matrix. By Proposition 5.20, $\det(T^*(Px))$ is a nonzero quadratic form, so there exist at most two $x \in [0, \infty]$ such that $T^*(Px)$ is a singular matrix. Since each $M(x)_{ij}$ is continuous function of $x \in [0, \infty]$, there exists λ such that $\lambda M(x)_{ij} \geq 0$ for all i, j and $x \in [0, \infty]$, so $\operatorname{sgn}(R^{-1}T(P, Q)) \geq 0$. □

The intervals P^c, Q^c form a rectangle in $\overline{\mathbb{R}}^2$ whose vertices are (P_{-0}, Q_{-0}), (P_{-0}, Q_{-1}), (P_{-1}, Q_{-0}), (P_{-1}, Q_{-1}). Since all MT are monotone, $T(P^c, Q^c)$ is the image of the sides of this rectangle. We have $T(P_{-1}, Q_{-0}) \in T(P^c, Q_{-0}) \cap T(P_{-1}, Q^c)$, $T(P_{-1}, Q_{-1}) \in T(P^{-1}, Q^c) \cap T(P^c, Q_{-1})$, $T(P_{-0}, Q_{-1}) \in T(P^c, Q_{-1}) \cap T(P_{-0}, Q^c)$, $T(P_{-0}, Q_{-0}) \in T(P_{-0}, Q^c) \cap T(P^c, Q_{-0})$, so $T(P^c, Q_{-0})$, $T(P_{-1}, Q^c)$, $T(P^c, Q_{-1})$, $T(P_{-0}, Q^c)$ are contiguous intervals.

Theorem 5.22 (Geometric inclusion criterion) *If T is a regular tensor and P, Q are regular matrices, then*

$$T(P^c, Q^c) = T(P^c, Q_{-0}) \cup T(P_{-1}, Q^c) \cup T(P^c, Q_{-1}) \cup T(P_{-0}, Q^c)$$

Proof: The right-hand side $Y = T(P^c, Q_{-0}) \cup T(P_{-1}, Q^c) \cup T(P^c, Q_{-1}) \cup T(P_{-0}, Q^c)$ is a union of contiguous intervals, so it is a (possibly full) interval which is included in $T(P^c, Q^c)$. Conversely let $z \in T(P^c, Q^c)$, so there exist $x \in P^c$, $y \in Q^c$ such that $z = T(x, y)$. Assume that T^*x is regular. Then $T(x, y)$ is a linear combination of $T(x, Q_{-0})$ and $T(x, Q_{-1})$ which both belong to Y. It follows that z belongs to Y as well. Assume that T^*x is singular. Since it has at most one unstable point, either $z = T(x, Q_{-0}) \in Y$ or $z = T(x, Q_{-1}) \in Y$. □

Definition 5.23 *For a tensor $T \in \mathbb{T}(\mathbb{R})$ and a matrix $M \in \mathbb{M}(\mathbb{R})$ we write $T \subseteq M$ if* $\operatorname{sgn}(M^{-1}T) \geq 0$.

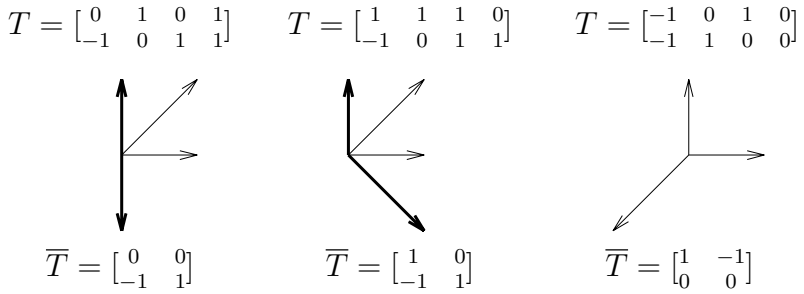


Figure 5.4: The matrix convex hull of a tensor

We are going to construct for a tensor T its **matrix convex hull** $\overline{T}$ which is a matrix such that $\operatorname{sgn}(Q^{-1}T) \geq 0$ iff $\operatorname{sgn}(Q^{-1}\overline{T}) \geq 0$ for each regular matrix Q. Let $u, v \in \mathbb{R}^2$ be vectors with $\det(u, v) > 0$. This means that the counterclockwise oriented angle from u to v is less than $\pi = 180°$. We say that a vector $w \in \mathbb{R}^2$ is a **convex combination** of u and v, if $w = ux_0 + vx_1$ for some $x_0, x_1 \geq 0$. This can be written as $w = [u, v]x$, where $[u, v]$ is the matrix with columns u, v and w, x are column vectors. Then we get

$$\begin{aligned} x &= [u,v]^{-1}w = \frac{1}{\det(u,v)} \begin{bmatrix} v_1 & -v_0 \\ -u_1 & u_0 \end{bmatrix} \cdot \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \\ &= \frac{1}{\det(u,v)} \begin{bmatrix} w_0v_1 - w_1v_0 \\ u_0w_1 - u_1w_0 \end{bmatrix} = \frac{1}{\det(u,v)} \begin{bmatrix} \det(w,v) \\ \det(u,w) \end{bmatrix} \end{aligned}$$

so w is a convex combination of u, v iff $\det(u, w) \geq 0$ and $\det(w, v) \geq 0$. For a regular matrix Q we have $Q^{-1}w = Q^{-1}[u, v]x$. It follows that $\operatorname{sgn}(Q^{-1}[u, v]) \geq 0$ iff $\operatorname{sgn}(Q^{-1}[u, v, w]) \geq 0$.

Proposition 5.24 *Let T be a $(2 \times n)$-matrix with $n \geq 3$. There exists a (2×2)-matrix $\overline{T}$ such that* $\operatorname{sgn}(Q^{-1}T) \geq 0$ *iff* $\operatorname{sgn}(Q^{-1}\overline{T}) \geq 0$ *for each regular (2×2)-matrix Q. We say that $\overline{T}$ is a* **matrix convex hull** *of T.*

Proof: If T is the zero matrix then $\overline{T}$ is also the zero matrix. Assume that T is nonzero. If a nonzero column u of T is a negative multiple of another column v of T then $\operatorname{sgn}(Q^{-1}T) \geq 0$ for no regular matrix Q, so we can take $\overline{T} = [u, v]$ (see Figure 5.4 left). If a column u of T is a nonnegative multiple of another column of T, or if it is a convex combination of two other columns of T, then we can omit it and obtain a $(2 \times (n-1))$-matrix T' such that $\operatorname{sgn}(Q^{-1}T) \geq 0$ iff $\operatorname{sgn}(Q^{-1}T') \geq 0$ for each regular matrix Q. We show that if $n \geq 4$ and no column of T is a nonzero multiple of another column of T, then a column of T is a convex combination of two other columns of T. Indeed for a column u of T there exist two different columns v, w of T such that $\operatorname{sgn}(\det(u, v)) = \operatorname{sgn}(\det(u, w))$. By a permutation of u, v, w we can attain $\det(u, w) > 0$, $\det(w, v) > 0$, $\det(u, v) > 0$, so w is a convex combination of u and v. Thus we successively omit columns which are convex combinations of other columns till we get a matrix which cannot be further reduced in this way. If this matrix has two columns we are done (see Figure 5.4 center). If it has three columns u, v, w then they can be permuted so that $\det(u, v) > 0$, $\det(v, w) > 0$, $\det(w, v) > 0$ and $\operatorname{sgn}(Q^{-1}[u, v, w]) \geq 0$ for no regular matrix Q. Thus we can take for $\overline{T}$ any singular matrix with $\operatorname{sgn}(\mathbf{u}(\overline{T})) > 0$, for example $\overline{T} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}$ (see Figure 5.4 right). □

Note that the matrix convex hull is not determined by T uniquely.

```
def s(X,p,q,r):
  if p == i or q == i: return xy
  for r -c-> r':
    if sgn(V_{r'}^{-1} F_c^{-1} X) ≥ 0: return c
  x, y=False,False
  for r -c-> r':
    Y = V_{r'}^{-1} F_c^{-1} X
    s_0, s_1 = sgn(Y_{-0-}), sgn(Y_{-1-})
    s_2, s_3 = sgn(Y_{--0}), sgn(Y_{--1})
    if (s_0 ≥ 0 ∨ s_1 ≥ 0) & (s_2 < 0 ∨ s_3 < 0) : x = True
    if (s_2 ≥ 0 ∨ s_3 ≥ 0) & (s_0 < 0 ∨ s_1 < 0) : y = True
  if (x & y) ∨ (¬x & ¬y): return xy
  if x: return x
  if y: return y
```

Table 5.3: The **balanced greedy selector** for the binary algorithm

5.4 The binary algorithm

The binary arithmetical algorithm for the addition, subtraction, multiplication, division and other bilinear functions works similarly as the unary algorithm by searching a path in the **binary graph**. The states (vertices) of the binary graph consist of binary tensors and states of the input and output paths.

Definition 5.25 *The binary graph for a sofic number system (F, G, V) is defined as follows: Its vertices are $(X, p, q, r) \in \mathbb{T}(\mathbb{R}) \times B^3$. The labelled edges are*

$$\begin{array}{rlll}
x-\textit{absorption:} & (X,p,q,r) \overset{(a,\lambda,\lambda)}{\longrightarrow} & (X^* H_{p,a,p'}, p', q, r), & \textit{if} \quad p \xrightarrow{a} p' \\
y-\textit{absorption:} & (X,p,q,r) \overset{(\lambda,a,\lambda)}{\longrightarrow} & (X_* H_{q,a,q'}, p, q', r), & \textit{if} \quad q \xrightarrow{a} q' \\
\textit{emission:} & (X,p,q,r) \overset{(\lambda,\lambda,a)}{\longrightarrow} & (F_a^{-1} X, p, q, r') & \textit{if} \quad p \neq \mathbf{i} \neq q,\ r \xrightarrow{a} r', \\
 & & & X \subseteq F_a V_{r'},
\end{array}$$

The first rule is an **x-absorption** of a letter of the first argument, the second rule is an **y-absorption** of a letter of the second argument, and the third rule is an **emission** of a letter of the output. The label of an edge is a triple consisting of x-input, y-input and output. The label of a path is the concatenation of the labels of its edges.

Proposition 5.26 *If $(X, \mathbf{i}, \mathbf{i}, \mathbf{i}) \overset{(u,v,w)}{\longrightarrow} (Y, p, q, r)$ is a finite path, then $\mathbf{i} \xrightarrow{u} p$, $\mathbf{i} \xrightarrow{v} q$, $\mathbf{i} \xrightarrow{w} r$, $Y = F_w^{-1} X(F_u V_p, F_v V_q)$. If $p \neq \mathbf{i}$, $q \neq \mathbf{i}$ and $r \neq \mathbf{i}$, then $Y \subseteq V_r$. If $(X, \mathbf{i}, \mathbf{i}, \mathbf{i}) \overset{(u,v,w)}{\longrightarrow}$ is an infinite path with $u, v, w \in A^\omega$, then $u, v, w \in \Sigma_G$ and $X(\Phi(u), \Phi(v)) = \Phi(w)$.*

Proof: The first emission must be preceeded by an x-absorption and an y-absorption. If $(X, \mathbf{i}, \mathbf{i}, \mathbf{i}) \overset{(u,v,\lambda)}{\longrightarrow} (Y, p, q, \mathbf{i}) \overset{(\lambda,\lambda,a)}{\longrightarrow} (Z, p, r)$ is the shortest path with an emission, then $Y = X(F_u V_p, F_v V_q) \subseteq F_a V_r$, so $Z = F_a^{-1} X(F_u V_p, F_v V_q) \subseteq V_r$. Assume that the condition is satisfied for $(X, \mathbf{i}, \mathbf{i}, \mathbf{i}) \overset{(u,v,w)}{\longrightarrow} (Y, p, q, r)$. If $(Y, p, q, r) \overset{(a,\lambda,\lambda)}{\longrightarrow} (Z, p', q, r)$ is an x-absorption, then $Z = Y^* H_{p,a,p'} = F_w^{-1} X(F_u V_p H_{p,a,p'}, F_v V_q) = F_w^{-1} X(F_{ua} V_{p'}, F_v V_q)$. If $(Y, p, q, r) \overset{(\lambda,a,\lambda)}{\longrightarrow} (Z, p, q', r)$ is an y-absorption, then $Z = Y_* H_{q,a,q'} = F_w^{-1} X(F_u V_p, F_v V_q H_{q,a,q'}) = F_w^{-1} X(F_u V_p, F_{va} V_{q'})$. If $(Y, p, q, r) \overset{(\lambda,\lambda,a)}{\longrightarrow} (F_a^{-1} Y, p, q, r')$ is an emission, then $Y \subseteq F_a V_s$, so $F_a^{-1} Y \subseteq V_s$. If $u, v, w \in A^\omega$ and (u, v, w) is a label of an infinite path with source $(X, \mathbf{i}, \mathbf{i}, \mathbf{i})$, then for each n there exist j_n, k_n

such that $(X, \mathbf{i}, \mathbf{i}, \mathbf{i}) \xrightarrow{(u_{[0,j_n)}, v_{[0,k_n)}, w_{[0,n)})} (Y_n, p_n, q_n, r_n)$ is a path, so $Y_n = F^{-1}_{w_{[0,n)}} X(F_{u_{[0,j_n)}} V_n, F_{v_{[0,k_n)}} V_n)$, $Y_n \subseteq V_{r_n}$, so $X(F_{u_{[0,j_n)}} V_{p_n}, F_{v_{[0,k_n)}} V_{q_n}) \subseteq F_{w_{[0,n)}} V_{r_n}$. We get $\Phi(w), X(\Phi(u), \Phi(v)) \in F_{w_{[0,n)}}(V_{r_n})$, so $\Phi(w) = X(\Phi(u), \Phi(v))$. □

The binary graph represents a nondeterministic algorithm for arithmetic operations. To get a deterministic algorithm, we use a selector $s : \mathbb{T}(\mathbb{R}) \times B^3 \to (A \times B) \cup \{\mathrm{x}, \mathrm{y}, \mathrm{xy}\}$ which chooses an admissible emission or an absorption. If $s(X, r) = (c, r') \in A \times B$, then the algorithm performs an emission with edge $r \xrightarrow{c} r'$. Otherwise the algorithm performs either an x-absorption or an y-absorption or both. The simplest greedy selector chooses an emission whenever possible and both the x-absorption and y-absorption if no emission is possible. But then it may happen that the length of the x-intervals X_{--j} becomes disproportionate with the length of the y-intervals X_{-i-}.

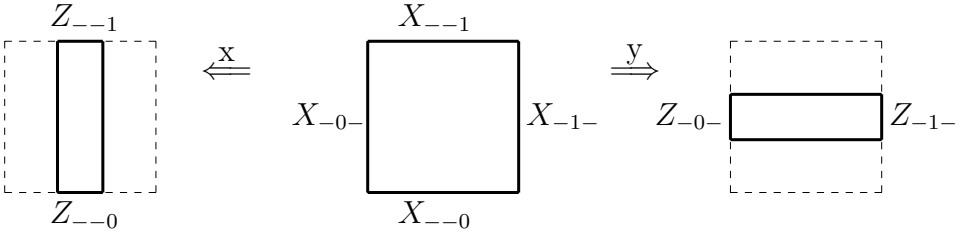


Figure 5.5: The x-absorption (left) and y-absorption (right)

In Table 5.3 we give in a Python-like syntax the **balanced greedy selector** which keeps the length of x-intervals and y-intervals balanced. The selector choses an emission whenever possible. If not it chooses either an x-absorption or an y-absorption or both. To choose a convenient kind of absorption we consider all edges $r \xrightarrow{c} r'$ and evaluate the tensor $Y = V^{-1}_{r'} F^{-1}_c X$. If for some c, i, j, $\mathrm{sgn}(Y_{-i-}) \geq 0$, and $\mathrm{sgn}(Y_{--j}) < 0$, then $X_{-i-} \subseteq F_c V_{r'}$ but $X_{--j} \not\subseteq F_c V_{r'}$ and we select an x-absorption to get a smaller interval $Z_{--j} = X_{--j} H_{p,a,p'}$ in the next step (Figure 5.5 left). If $\mathrm{sgn}(Y_{--j}) \geq 0$, and $\mathrm{sgn}(Y_{-i-}) < 0$, then $X_{--j} \subseteq F_c V_{r'}$ but $X_{-i-} \not\subseteq F_c V_{r'}$ and we select an y-absorption to get a smaller interval $Z_{-i-} = X_{-i-} H_{q,b,q'}$ in the next step (Figure 5.5 right). We select both x-absorption and y-absorption if both or none of these two conditions is satisfied.

A sample run of the algorithm is in Table 5.4. It shows the convex closure $\overline{X}$ of the state tensor, the state tensor X itself together with the matrices which act upon it and the input and output paths. In the first step we start with the multiplication tensor $X = \left[\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right]$, whose marginal matrices are included in no $F_a V_q$, so both x-absorption and y-absorption are used. The same situation occurs in the second and third steps. In the fourth step we get a tensor with interval $[0.56, 3.00]$ which is included in $F_{\overline{0}} V_3 = \left[\begin{smallmatrix} 1 & 1 \\ 2 & -2 \end{smallmatrix}\right]$ so the emission of $\overline{0}$ is chosen. In the next step with tensor $X = \left[\begin{smallmatrix} 9 & 6 & 18 & 12 \\ 32 & 16 & 16 & 8 \end{smallmatrix}\right]$ we have $X_{--1} = \left[\begin{smallmatrix} 18 & 12 \\ 16 & 8 \end{smallmatrix}\right] \subseteq F_{\overline{0}} V_3 = \left[\begin{smallmatrix} 1 & 1 \\ 2 & -2 \end{smallmatrix}\right]$ but neither $X_{-0-} = \left[\begin{smallmatrix} 9 & 18 \\ 32 & 16 \end{smallmatrix}\right]$ nor $X_{-1-} = \left[\begin{smallmatrix} 6 & 12 \\ 16 & 8 \end{smallmatrix}\right]$ is included in $F_{\overline{0}} V_3$, so the y-absorption is closed to get smaller X_{-i-} intervals.

In contrast to the unary algorithm, the binary algorithm in redundant systems is not guaranteed to produce an infinite output. This happens if we try to compute indefinite expressions like $\frac{0}{0}$, $0 \cdot \infty$ or $\infty + \infty$. In this case the algorithm reads ever longer prefixes of the input words without producing any output. Nevertheless, in redundant systems the algorithm gives the correct result whenever the computed result belongs to $\overline{\mathbb{R}}$.

Proposition 5.27 *Let (F, G, V) be a sofic number system with initialized graph G and let*

$\overline{X}$	$(X^*H_u)_*H_v$ $F_w^{-1}X$	u	v	w
$[0.00,\infty]$	$\left[\begin{smallmatrix}1&0&0&0\\0&0&0&1\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&2\\4&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&1\\2&-2\end{smallmatrix}\right]$	$\lambda \xrightarrow{1} 1$	$\lambda \xrightarrow{\overline{0}} \overline{0}$	
$[0.12,-0.12]$	$\left[\begin{smallmatrix}1&2&1&2\\8&4&-8&-4\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\1&3\end{smallmatrix}\right]$	$1 \xrightarrow{0} 0$	$\overline{0} \xrightarrow{\overline{0}} \overline{0}$	
$[0.25,-0.25]$	$\left[\begin{smallmatrix}1&3&1&3\\4&4&-4&-4\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&0\\5&4\end{smallmatrix}\right]*\left[\begin{smallmatrix}4&5\\0&3\end{smallmatrix}\right]$	$0 \xrightarrow{1} 1$	$\overline{0} \xrightarrow{1} 1$	
$[0.56,3.00]$	$\left[\begin{smallmatrix}1&0\\0&2\end{smallmatrix}\right].\left[\begin{smallmatrix}9&6&18&12\\16&8&8&4\end{smallmatrix}\right]$			$\lambda \xrightarrow{\overline{0}} \overline{0}$
$[0.28,1.50]$	$\left[\begin{smallmatrix}9&6&18&12\\32&16&16&8\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]$		$1 \xrightarrow{0} 0$	
$[0.28,1.12]$	$\left[\begin{smallmatrix}9&6&27&18\\32&16&32&16\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\1&3\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$0 \xrightarrow{0} 0$	
$[0.49,0.94]$	$\left[\begin{smallmatrix}2&-1\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}63&36&105&60\\128&64&128&64\end{smallmatrix}\right]$			$\overline{0} \xrightarrow{1} 1$
$[-0.02,0.88]$	$\left[\begin{smallmatrix}-1&4&41&28\\64&32&64&32\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\1&3\end{smallmatrix}\right]$		$0 \xrightarrow{0} 0$	
$[0.15,0.69]$	$\left[\begin{smallmatrix}19&20&61&44\\128&64&128&64\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&0\\5&4\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$0 \xrightarrow{1} 1$	
$[0.45,0.69]$	$\left[\begin{smallmatrix}2&-1\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}461&280&298&176\\1024&512&512&256\end{smallmatrix}\right]$			$1 \xrightarrow{1} 1$
$[-0.10,0.38]$	$\left[\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}-51&24&42&48\\512&256&256&128\end{smallmatrix}\right]$			$1 \xrightarrow{0} 0$
$[-0.20,0.75]$	$\left[\begin{smallmatrix}-51&24&42&48\\256&128&128&64\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$1 \xrightarrow{0} 0$	
$[-0.01,0.56]$	$\left[\begin{smallmatrix}-3&48&183&144\\512&256&512&256\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\1&3\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$0 \xrightarrow{0} 0$	
$[0.18,0.47]$	$\left[\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}375&288&753&480\\2048&1024&2048&1024\end{smallmatrix}\right]$			$0 \xrightarrow{0} 0$
$[0.37,0.94]$	$\left[\begin{smallmatrix}2&-1\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}375&288&753&480\\1024&512&1024&512\end{smallmatrix}\right]$			$0 \xrightarrow{1} 1$
$[-0.27,0.88]$	$\left[\begin{smallmatrix}-137&32&241&224\\512&256&512&256\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}4&5\\0&3\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$0 \xrightarrow{\overline{1}} \overline{1}$	
$[-0.07,0.41]$	$\left[\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}-146&128&851&832\\2048&1024&4096&2048\end{smallmatrix}\right]$			$1 \xrightarrow{0} 0$
$[-0.14,0.81]$	$\left[\begin{smallmatrix}-146&128&851&832\\1024&512&2048&1024\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]*\left[\begin{smallmatrix}2&2\\0&1\end{smallmatrix}\right]$	$1 \xrightarrow{0} 0$	$\overline{1} \xrightarrow{\overline{1}} \overline{1}$	
$[-0.14,0.40]$	$\left[\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}-292&244&559&1637\\2048&2048&4096&4096\end{smallmatrix}\right]$			$0 \xrightarrow{0} 0$
$[-0.29,0.80]$	$\left[\begin{smallmatrix}-292&244&559&1637\\1024&1024&2048&2048\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&0\\5&4\end{smallmatrix}\right]*\left[\begin{smallmatrix}4&0\\1&3\end{smallmatrix}\right]$	$0 \xrightarrow{1} 1$	$\overline{1} \xrightarrow{0} 0$	
$[0.23,0.80]$	$\left[\begin{smallmatrix}1873&1742&4931&3274\\8192&4096&8192&4096\end{smallmatrix}\right]*\left[\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&0\\5&4\end{smallmatrix}\right]$	$1 \xrightarrow{1} 1$	$0 \xrightarrow{1} 1$	
$[0.56,0.80]$	$\left[\begin{smallmatrix}2&-1\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}36733&21596&22958&13096\\65536&32768&32768&16384\end{smallmatrix}\right]$			$0 \xrightarrow{1} 1$
$[0.12,0.60]$	$\left[\begin{smallmatrix}3965&5212&6574&4904\\32768&16384&16384&8192\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&1\\0&4\end{smallmatrix}\right]$	$1 \xrightarrow{0} 0$	$1 \xrightarrow{0} 0$	
$[0.12,0.44]$	$\left[\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right].\left[\begin{smallmatrix}3965&8271&10087&14397\\32768&32768&32768&32768\end{smallmatrix}\right]$			$1 \xrightarrow{0} 0$
$[0.24,0.88]$	$\left[\begin{smallmatrix}3965&8271&10087&14397\\16384&16384&16384&16384\end{smallmatrix}\right]*\left[\begin{smallmatrix}4&5\\0&3\end{smallmatrix}\right]*\left[\begin{smallmatrix}3&0\\5&4\end{smallmatrix}\right]$	$0 \xrightarrow{\overline{1}} \overline{1}$	$0 \xrightarrow{1} 1$	

input tensor $M = \left[\begin{smallmatrix}1&0&0&0\\0&0&0&1\end{smallmatrix}\right]$ (multiplication)

input: $u = 1011111101101\overline{1}$, $p = \overline{1}$, $F_uV_p = \left[\begin{smallmatrix}6122&12247\\8192&16384\end{smallmatrix}\right] = [0.747, 0.747]$

input: $v = \overline{00}10001001\overline{11}0101$, $q = 1$, $F_vV_q = \left[\begin{smallmatrix}8617&4310\\4096&2048\end{smallmatrix}\right] = [2.105, 2.113]$

result: $M(F_uV_p, F_vV_q) = \left[\begin{smallmatrix}52753274&105532399&26385820&52784570\\33554432&67108864&16777216&33554432\end{smallmatrix}\right] = [1.573, 1.572, 1.575, 1.573]$

output: $w = \overline{0}110010010$, $r = 0$, $F_wV_r = \left[\begin{smallmatrix}401&403\\256&256\end{smallmatrix}\right] = [1.566, 1.574]$

Table 5.4: The computation of the binary algorithm in the binary signed system

s be the balanced greedy selector. If T is a tensor and (p,u), (q,v) are paths of G such that $T(\Phi(u),\Phi(v)) \in \overline{\mathbb{R}}$, then the binary algorithm computes an infinite path (r,w) such that $T(\Phi(u),\Phi(v)) = \Phi(w)$.

Proof: It suffices to show that in the computed path there must be an infinite number of emissions. If not then the output of the path is a finite word w. When all emissions are done we have a state tensor $X = F_w^{-1} T(F_{u_{[0,n)}}, F_{v_{[0,m)}})$. The length n of the x-input grows with x-absorptions and the length m of the y-input grows with y-absorptions. We show that both m and n grow to infinity. If not, from some step onwards, only one kind of absorptions is chosen, say x-absorptions. This means that the length of X_{--j} intervals converges to zero and ultimately, since the system is redundant, these intervals are contained in some $F_c V_{r'}$. If the intervals X_{-i-} are included in $F_c V_{r'}$, then the selector choses the emission of c. If not then the selector choose an y-absorption. Thus there is an infinite number of both x-absorptions and y-absorptions. But then the length of the interval X^c converges to zero and X has to be included in some $F_c V_{r'}$ so that an emission is available. □

5.5 Polynomials

With the binary algorithm, we can compute polynomial and rational functions. However, they can be also computed directly. A polynomial is a complex function $p(x) = p_0 + p_1 x + \cdots + p_n x^n$, where $p_i \in \mathbb{C}$. If we define $p(\infty) = \infty$, then $p : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a continuous function. Often we write a polynomial as an infinite sum $p(x) = \sum_{i \geq 0} p_i x^i$, where only a finite number of coefficients p_i are nonzero. The **degree** $\deg(p)$ of p is the largest n such that $p_n \neq 0$ and its **leading coefficient** is $\ell(p) = p_{\deg(p)}$. For the constant **zero polynomial** $p(x) = 0$ we set $\deg(p) = -1$. We say that p is a **monic** polynomial if its leading coefficient is $\ell(p) = 1$. Denote by $\mathbb{C}[x]$ the set of all polynomials and by $\mathbb{R}[x]$ the set of polynomials with real coefficients. As algebraic structures, both $\mathbb{C}[x]$ and $\mathbb{R}[x]$ are **commutative rings** with a unit. The addition, subtraction and multiplication are defined pointwise by $(p+q)(x) = p(x) + q(x)$, $(p-q)(x) = p(x) - q(x)$, $(pq)(x) = p(x) \cdot q(x)$. For the coefficients we get $(p+q)_n = p_n + q_n$, $(p-q)_n = p_n - q_n$, $(pq)_n = \sum_{i=0}^{n} p_i q_{n-i}$. The multiplication of $p \in \mathbb{C}[x]$ by $a \in \mathbb{C}$ is $(ap)(x) = a \cdot p(x)$, or $(ap)_n = a \cdot p_n$, so $\mathbb{C}[x]$ is also a vector space over $\mathbb{C}$ and $\mathbb{R}[x]$ is a vector space over $\mathbb{R}$.

If $r = pq$, we say that p divides r and write $p|r$. A linear polynomial $p(x) = x - a$ divides r iff a is a root of r, i.e., if $r(a) = 0$. By the fundamental theorem of algebra, every polynomial of positive degree has a real or complex root. It follows that each polynomial can be written as $p(x) = a(x-c_1)^{r_1} \cdots (x-c_m)^{r_m}$, where $a, c_i \in \mathbb{C}$ are complex numbers and $r_1 + \cdots + r_m = \deg(p)$. Polynomials can be divided with remainder: For every nonzero polynomials t, s there exist unique polynomials q (quotient) and r (remainder) such that $t = sq + r$ and $\deg(r) < \deg(s)$. Nonzero polynomials s, t have the **greatest common divisor** (GCD) $p = \gcd(s,t)$ which is the monic polynomial of highest degree which divides both s and t. If a polynomial divides both p and q, then it divides also $\gcd(p,q)$. The GCD of two polynomials can be found by the Euclidean algorithm. If p_0, p_1 are given nonzero polynomials, there exists a unique sequence of polynomials $p_2, \ldots, p_n, p_{n+1}$ such that $n \geq 1$, $p_{i-1} = p_i q_i + p_{i+1}$ for some $q_i \in \mathbb{C}[x]$, $\deg(p_{i+1}) < \deg(p_i)$, and $p_{n+1}(x) = 0$, so $p_{n-1} = p_n q_n$. Then p_n is a constant multiple of $\gcd(p_0, p_1)$.

Proposition 5.28 *If $p, q \in \mathbb{C}[x]$ are nonzero polynomials then there exist polynomials s, t such that $ps + qt = \gcd(p,q)$.*

Proof: Set $M = \{ps + qt : \ s,t \in \mathbb{C}[x]\}$ and let r be a nonzero monic polynomial of M with the lowest degree. For each nonzero $ps + qt \in M$ there exist u, v with $ps + qt = ru + v$ and

$\deg(v) < \deg(r)$. Since $v \in M$ we get $v = 0$, so r divides all elements of M in particular p and q. On the other hand if r divides p and q it divides also r, so $r = \gcd(p, q)$. □

The derivation of a polynomial $p(x) = p_0 + p_1 x + \cdots + p_n x^n$ is

$$p'(x) = p_1 + 2p_2 x + \cdots + np_n x^{n-1}.$$

For $p(x) = (x-c)^r q(x)$ we get $p'(x) = r(x-c)^{r-1}q(x) + (x-c)^r q'(x)$. Thus for $p(x) = a(x-c_1)^{r_1}\cdots(x-c_m)^{r_m}$ we get

$$\gcd(p, p') = a(x-c_1)^{r_1-1}\cdots(x-c_m)^{r_m-1}.$$

The number of real roots of a real a polynomial can be determined by the Sturm Theorem (see Waerden [65]). Define the **variance** $w(a_0, \ldots, a_n)$ of a finite sequence of real numbers as its number of sign changes. To get the variance, delete first all zeros, so

$$w(a_0, \ldots, a_{i-1}, 0, a_{i+1}, \ldots, a_n) = w(a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n).$$

For a sequence which does not contain zeros we have

$$w(a_0, \ldots, a_n) = |\{i < n : \ a_i a_{i+1} < 0\}|.$$

Given a polynomial $p(x)$ define its **Sturm chain** as a finite sequence p_i of polynomials defined by $p_0 = p$, $p_1 = p'$, $p_{i-1} = p_i q_i - p_{i+1}$, where $\deg(p_{i+1}) < \deg(p_i)$. Thus the Sturm chain is just the Euclidean sequence of p, p' except that the remainders are taken negative. The last element of the chain satisfies $p_{m-1} = p_m q_m$, so p_m is a constant multiple of $\gcd(p, p')$.

Theorem 5.29 *Let $p \in \mathbb{R}[x]$ be a real polynomial with the Sturm chain $p = p_0, \ldots, p_m$, let $a < b$ be real numbers which are not the roots of p. Then the number of roots of p (counted without multiplicities) in the interval $I = (a, b)$ is*

$$|\{x \in I : \ p(x) = 0\}| = w(p_0(a), \ldots, p_m(a)) - w(p_0(b), \ldots, p_m(b))$$

Proof: Since p_m is a constant multiple of $\gcd(p, p')$, there exist polynomials r_i with $p_i = r_i p_m$. Since $p(a) \neq 0 \neq p(b)$, we have $p_m(a) \neq 0 \neq p_m(b)$, $r_0(a) \neq 0 \neq r_0(b)$, $r_m(a) = r_m(b) = 1$. By passing from p_i to r_i the variations do not change: $w(p_0(a), \ldots, p_m(a)) = w(r_0(a), \ldots, r_m(a))$ and similarly $w(p_0(b), \ldots, p_m(b)) = w(r_0(b), \ldots, r_m(b))$. If $J \subseteq I$ is an interval in which no r_i has a root, then $w(r_0(c), \ldots, r_m(c))$ is constant on J. We evaluate how w changes at $c \in I$ in which one of the r_i is zero. If $r_i(c) = 0$, with $0 < i < m$, then $r_{i+1}(c) \neq 0$, since otherwise we would get $r_{i+2}(c) = 0$ and by induction $r_m(c) = 0$ which is a contradiction. Thus both $r_{i+1}(c)$, $r_{i-1}(c)$ are nonzero and therefore they are nonzero also in some interval which contains c. This implies that $w(r_{i-1}(x), r_i(x), r_{i+1}(x))$ is constant in such an interval. Assume now that $r_0(c) = 0$. Then $p_0(c) = 0$ and $p(x) = (x-c)^k s(x)$ for some $k \geq 1$ and a polynomial $s(x)$ with $s(c) \neq 0$. We get $p'(x) = k(x-c)^{k-1}s(x) + (x-c)^k s'(x)$. If we divide p and p' by $(x-c)^{k-1}$, we get

$$s_0(x) = (x-c)s(x), \ s_1(x) = ks(x) + (x-c)s'(x).$$

For $x < c$ we have $\operatorname{sgn}(s_0(x)) = -\operatorname{sgn}(s(c))$, $\operatorname{sgn}(s_1(x)) = \operatorname{sgn}(s(c))$, while for $x > c$ we get $\operatorname{sgn}(s_0(x)) = \operatorname{sgn}(s_1(x)) = \operatorname{sgn}(s(c))$. Thus as x passes through c, $w(s_0(x), s_1(x))$ diminishes by one. Since $r_i(c)$ are nonzero multiples of $s_i(c)$, the same happens for r_i. Thus for each root c of $p(x)$ in I, the variance $w(r_0(x), \ldots, r_m(x))$ diminishes by one when x passes through c. □

5.6 Rational functions

A **rational function** $R : \overline{\mathbb{R}} \to \overline{\mathbb{R}} \cup \{\frac{0}{0}\}$ of degree at most $q \geq 0$ is a ratio of two polynomials of degree at most q, or a function of the form

$$R(x) = \frac{R_{00}x_0^q + R_{01}x_0^{q-1}x_1 + \cdots + R_{0q}x_1^q}{R_{10}x_0^q + R_{11}x_0^{q-1}x_1 + \cdots + R_{1q}x_1^q}.$$

A rational function is **regular** if the numerator and denominator polynomials are relatively prime. In this case $R(x) \neq \frac{0}{0}$ for every $x \in \overline{\mathbb{R}}$ and R is a mapping $R : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$. We do not adopt the assumption of regularity in general, since its verification would unnecessarily complicate the transcendent algorithm of Section 8.3. For each rational function R there exists an equivalent regular rational function r, which is obtained from R by cancelling the common factors of the numerator and denominator of R. A rational function R of degree at most q is given by a $2 \times (q+1)$-matrix $R = (R_{kj})_{k=0,1,j=0,\ldots,q}$, so $R(x)_k = \sum_{i=0}^{q} R_{ki}x_0^{q-i}x_1^i$. If M is a transformation, then both compositions $RM = R \circ M$ and $MR = M \circ R$ are rational functions of degree at most q, which is regular provided both R and M are regular. The composition MR is obtained by the product of the matrices $(MR)_{ki} = \sum_{j=0}^{1} M_{kj}R_{ji}$. To obtain the composition RM, let $y_i = \sum_{j=0}^{1} M_{ij}x_j$, so

$$\begin{aligned}
R(y)_k &= \sum_{p=0}^{q} R_{kp}(M_{00}x_0 + M_{01}x_1)^{q-p}(M_{10}x_0 + M_{11}x_1)^p \\
&= \sum_{p=0}^{q} R_{kp} \cdot \sum_{i=0}^{q-p} \binom{q-p}{i} M_{00}^{q-p-i}M_{01}^i x_0^{q-p-i}x_1^i \cdot \sum_{j=0}^{p} \binom{p}{j} M_{10}^{p-j}M_{11}^j x_0^{p-j}x_1^j \\
&= \sum_{r=0}^{q}\sum_{i=0}^{r}\sum_{p=r-i}^{q-i} R_{kp}\binom{q-p}{i}\binom{p}{r-i} M_{00}^{q-p-i}M_{01}^i M_{10}^{p-r+i}M_{11}^{r-i}x_0^{q-r}x_1^r
\end{aligned}$$

where $r = i + j$. Since $0 \leq i \leq q - p$, $0 \leq j \leq p$, we get $j = r - i \leq p \leq q - i$. Thus the composition RM is defined by

$$(RM)_{kr} = \sum_{i=0}^{r}\sum_{p=r-i}^{q-i} R_{kp}\binom{q-p}{i}\binom{p}{r-i} M_{00}^{q-p-i}M_{01}^i M_{10}^{p-r+i}M_{11}^{r-i}$$

If S is a rational function of degree p, then $R \circ S$ and $S \circ R$ are rational functions of degree $q \cdot p$.

Rational functions are obtained from tensors. A bilinear tensor T is **symmetric** if $T_{ijk} = T_{ikj}$ for each i, j, k. For a rational function R of degree 2 there exists a symmetric tensor $T = \left[\begin{smallmatrix} R_{00} & R_{01}/2 & R_{01}/2 & R_{02} \\ R_{10} & R_{11}/2 & R_{11}/2 & R_{12} \end{smallmatrix}\right]$, such that $R(x) = T(x, x)$. For each interval $I \subseteq \overline{\mathbb{R}}$ we have

$$R(I) = \{R(x) : \ x \in I\} \subseteq \{T(x, y) : \ x, y \in I\} = T(I, I).$$

If P, Q are regular matrices and $\operatorname{sgn}(Q^{-1}RP) \geq 0$ then $\operatorname{sgn}(Q^{-1}T(P, P)) \geq 0$ and by Theorem 5.17 $R(P^c) \subseteq T(P^c, P^c) \subseteq Q^c$.

To get the inclusion criterion for rational functions of degree 2 or more, we have to generalize Theorem 5.17 to tensors of higher degrees. For example, **trilinear** tensors $T(x, y, z)_k = \sum_{i,j,l} T_{kijl}x_i y_j z_l$ determine functions $T : \overline{\mathbb{R}}^3 \to \overline{\mathbb{R}} \cup \{\frac{0}{0}\}$. For $x \in \overline{\mathbb{R}}$ we get a bilinear tensor T^*x

and for matrices P, Q, R we get a trilinear tensors T^*P, $T(P,Q,R)$ defined by

$$\begin{aligned}(T^*x)_{kjl} &= \sum_i T_{kijl}x_i \\ (T^*P)_{kijl} &= \sum_r T_{krjl}P_{ri} \\ T(P,Q,R)_{kijl} &= \sum_{r,s,t} T_{krst}P_{ri}Q_{sj}R_{tl}\end{aligned}$$

The image of intervals $I, J, K \subseteq \overline{\mathbb{R}}$ by a trilinear tensor T is

$$T(I,J,K) = \{T(x,y,z) : \ x \in I, y \in J, z \in K\} \cap \overline{\mathbb{R}}.$$

Proposition 5.30 *Let T be a trilinear tensor, P, Q, R, S regular matrices. If $\operatorname{sgn}(S^{-1}T(P,Q,R)) \geq 0$, then $T(P^c, Q^c, R^c) \subseteq S^c$.*

Proof: Let $\operatorname{sgn}(S^{-1}T(P,Q,R)) \geq 0$, $x \in P^c$, $u = P^{-1}x$, so $x = Pu$ and $\operatorname{sgn}(u) \geq 0$. We have

$$(T^*x)(Q,R) = (T^*(Pu))(Q,R) = ((T^*P)^*u)(Q,R) = T(P,Q,R)^*u$$

Since $\operatorname{sgn}(S^{-1}T(P,Q,R)^*u) \geq 0$, we get by Theorem 5.17 $(T^*x)(Q^c, R^c) \subseteq S^c$. If $y \in Q^c$, $z \in R^c$ then $T(x,y,z) = (T^*x)(y,z) \in S^c$, so $T(P^c, Q^c, R^c) \subseteq S^c$. □

For a rational function R of degree 3 there exists a symmetric trilinear tensor T given by $T_{kijl} = R_{k,(i+j+l)}/\binom{3}{i+j+l}$ such that $R(x) = T(x,x,x)$ for any $x \in \overline{\mathbb{R}}$. Thus if $\operatorname{sgn}(Q^{-1}RP) \geq 0$ then $R(P^c) \subseteq Q^c$. More generally, a q-linear tensor $T_{k,i_1,\ldots,i_q}$ of q variables $x^{(1)}, \ldots, x^{(q)} \in \overline{\mathbb{R}}$ is given by

$$T(x^{(1)}, \ldots, x^{(q)})_k = \sum_{i_1,\ldots,i_q} T_{k,i_1,\ldots,i_q} x^{(1)}_{i_1}, \ldots, x^{(q)}_{i_q}.$$

If $\operatorname{sgn}(Q^{-1}T(P_1, \ldots, P_q)) \geq 0$ then $T(P_1^c, \ldots, P_q^c) \subseteq Q^c$. For a rational function R of order q there exists a symmetric q-linear tensor T of q variables such that $R(x) = T(x, \ldots, x)$. We obtain a simple criterion for the inclusion:

Theorem 5.31 *Let $R : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ be a rational function and P, Q regular matrices. If $\operatorname{sgn}(Q^{-1}RP) \geq 0$, then $R(P^c) \subseteq Q^c$.*

Chapter 6

Integer vectors and matrices

When we compute arithmetical algorithms in a sofic number system, we perform arithmetical operations with the entries of its transformations, intervals and vectors. These operations are algorithmic, provided the entries of the matrices are rational numbers. Since we work with projective matrices and vectors, we can assume that their entries are integers whose greatest common divisor is 1. Then each projective tensor, matrix or vector with rational entries has exactly two representations with coprime integers.

6.1 Determinant, norm and length

Denote by $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ the set of integers and by

$$\overline{\mathbb{Q}} = \{x \in \mathbb{Z}^2 \setminus \{\tfrac{0}{0}\} : \ \gcd(x) = 1\}$$

the set of (homogeneous coordinates of) rational numbers which we understand as a subset of $\overline{\mathbb{R}}$. Here $\gcd(x) > 0$ is the greatest common divisor of x_0 and x_1. Each rational number has exactly two representations in $\overline{\mathbb{Q}}$, $x = \frac{x_0}{x_1} = \frac{-x_0}{-x_1}$. In contrast to the norm of vectors $x \in \overline{\mathbb{R}}$, the **norm** $||x|| = \sqrt{x_0^2 + x_1^2}$ of $x \in \overline{\mathbb{Q}}$ does not depend on the representation of x. We have the cancellation map $\mathbf{d} : \mathbb{Z}^2 \setminus \{\frac{0}{0}\} \to \overline{\mathbb{Q}}$ given by $\mathbf{d}(x) = \frac{x_0/\gcd(x)}{x_1/\gcd(x)}$. Denote by $\mathbb{Z}^{2\times 2}$ the set of 2×2 matrices with integer entries and by

$$\mathbb{M}(\mathbb{Z}) = \{M \in \mathbb{Z}^{2\times 2} : \ \gcd(M) = 1, \ \det(M) \neq 0\}.$$

Each matrix of $\mathbb{M}(\mathbb{R})$ with rational entries has exactly two representations $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] = \left[\begin{smallmatrix} -a & -b \\ -c & -d \end{smallmatrix}\right]$ in $\mathbb{M}(\mathbb{Z})$. For $x \in \overline{\mathbb{Q}}$ we distinguish $M \cdot x \in \mathbb{Z}^2$ given by $(M \cdot x)_i = \sum_j M_{ij} x_j$ from $Mx = \mathbf{d}(M \cdot x) \in \overline{\mathbb{Q}}$. For $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{Z}^{2\times 2} \setminus \{\mathbf{0}\}$ denote by $\mathbf{d}(M) = \left[\begin{smallmatrix} a/g & b/g \\ c/g & d/g \end{smallmatrix}\right]$, where $g = \gcd(M) > 0$ is the greatest common divisor of the entries of M. Thus we have the cancellation map $\mathbf{d} : \mathbb{Z}^{2\times 2} \setminus \{\mathbf{0}\} \to \mathbb{M}(\mathbb{Z})$. We distinguish the matrix multiplication $M \cdot N$ from the multiplication $MN = \mathbf{d}(M \cdot N)$ in $\mathbb{M}(\mathbb{Z})$. The **determinant** and **norm** of $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{Z}^{2\times 2}$ are defined by

$$\det(M) = ad - bc, \ ||M|| = \sqrt{a^2 + b^2 + c^2 + d^2}$$

and do not depend on the representation of M in $\mathbb{M}(\mathbb{Z})$. We have

$$\det(M \cdot N) = \det(M) \cdot \det(N), \ ||M \cdot N|| \leq ||M|| \cdot ||N||,$$

so $|\det(MN)| \leq |\det(M)| \cdot |\det(N)|$, $||MN|| \leq ||M|| \cdot ||N||$. The pseudo-inverse of $M = [\frac{a}{c}, \frac{b}{d}]$ is $M^{-1} = [\frac{d}{-c}, \frac{-b}{a}] = [\frac{-d}{c}, \frac{b}{-a}]$. We have $M \cdot M^{-1} = \det(M) \cdot \mathrm{Id}$, $MM^{-1} = \mathrm{Id} = [\frac{1}{0}, \frac{0}{1}]$.

Proposition 6.1 *If $M, N \in \mathbb{M}(\mathbb{Z})$, then $g = \gcd(M \cdot N)$ divides both $\det(M)$ and $\det(N)$.*

Proof: Clearly g divides $M^{-1} \cdot M \cdot N = \det(M) \cdot N$. Since $\gcd(N) = 1$, g divides $\det(M)$. For a similar reason, g divides $\det(N)$. □

Recall that by Proposition 5.9 the size and length of a matrix $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{Z})$ are defined by $\mathrm{sz}(P) = \frac{ab+cd}{|ad-bc|}$, $|P| = \frac{1}{2} - \frac{1}{\pi} \arctan\, \mathrm{sz}(P)$ and we have an estimate

$$\mathrm{sz}(P) \geq 1 \Leftrightarrow |P| \leq \frac{1}{4} \;\Rightarrow\; \frac{1}{4 \cdot \mathrm{sz}(P)} \leq |P| \leq \frac{1}{\pi \cdot \mathrm{sz}(P)}$$

Lemma 6.2 *If $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{M}(\mathbb{Z})$ is an integer matrix, then*

$$\begin{aligned} \max\{|a|,|b|,|c|,|d|\} \quad &\leq \quad \max\{|ab+cd|, |ad-bc|\}, \\ \sqrt{2 \cdot |\det(P) \cdot \mathrm{sz}(P)|} \leq \; ||P|| \quad &\leq 2 \cdot |\det(P)| \cdot \max\{|\mathrm{sz}(P)|, 1\}, \\ ||P|| \quad &\leq |\det(P)| \cdot \max\left\{\tfrac{1}{|P|}, \tfrac{1}{1-|P|}\right\}. \end{aligned}$$

Proof: If $a = 0$, then $|bc| = |\det(P)| \neq 0$, and

$$0 < |b|, |c| \leq |bc| = |ad - bc|,\ |d| \leq |cd| = |ab + cd|.$$

If $b = 0$, then $ad = \det(P) \neq 0$, and $0 < |a|, |d| \leq |ad| = |ad - bc|$, $|c| \leq |cd| = |ab + cd|$. Similarly we prove the inequality if $c = 0$ or $d = 0$. Assume now that all a, b, c, d are nonzero. If $\mathrm{sgn}(ab) \cdot \mathrm{sgn}(cd) > 0$ then $|a| \cdot |b| + |c| \cdot |d| = |ab + cd|$, so $\max\{|a|, |b|, |c|, |d|\} \leq |ab + cd|$. If $\mathrm{sgn}(ab) \cdot \mathrm{sgn}(cd) < 0$ then

$$\mathrm{sgn}(ad) \cdot \mathrm{sgn}(bc) = \mathrm{sgn}(abcd) = \mathrm{sgn}(ab) \cdot \mathrm{sgn}(cd) < 0,$$

so $|a| \cdot |d| + |b| \cdot |c| = |ad - bc|$ and $\max\{|a|, |b|, |c|, |d|\} \leq |ad - bc|$. Thus we have proved the first inequality in all cases. From $(a \pm b)^2 + (c \pm d)^2 \geq 0$ we get

$$2 \cdot |\det(P) \cdot \mathrm{sz}(P)| = 2|ab + cd| \leq ||P||^2,$$

so $\sqrt{2 \cdot |\det(P) \cdot \mathrm{sz}(P)|} \leq ||P||$. If $\max\{|a|, |b|, |c|, |d|\} \leq K$ then $||P|| \leq 2K$. Thus

$$\begin{aligned} |ab + cd| \leq |ad - bc| \quad &\Rightarrow \quad ||P|| \leq 2|\det(P)|, \\ |ab + cd| \geq |ad - bc| \quad &\Rightarrow \quad ||P|| \leq 2|ab + cd| = 2|\det(P) \cdot \mathrm{sz}(P)|. \end{aligned}$$

Thus $||P|| \leq 2 \cdot |\det(P)| \cdot \max\{|\mathrm{sz}(P)|, 1\}$. To prove the last inequality we distinguish three cases. If $|P| \leq \frac{1}{4}$ then $\frac{ab+cd}{|ad-bc|} \geq 1$ so $ab + cd \geq |ad - bc|$. From $\max\{|a|, |b|, |c|, |d|\} \leq ab + cd$ we get by Proposition 5.9

$$||P|| \leq 2(ab + cd) = 2 \cdot \mathrm{sz}(P) \cdot |\det(P)| \leq \frac{|2\det(P)|}{\pi|P|} \leq \frac{|\det(P)|}{|P|}.$$

If $\frac{1}{4} \leq |P| \leq \frac{1}{2}$, then $|\arctan \frac{ab+cd}{ad-bc}| \leq \frac{\pi}{4}$, so $|ab+cd| \leq |ad-bc|$. It follows $\max\{|a|, |b|, |c|, |d|\} \leq |ad - bc|$, so $||P|| \leq 2|\det(P)| \leq \frac{|\det(P)|}{|P|}$. If $|P| \geq \frac{1}{2}$ then for the matrix $Q = \left[\begin{smallmatrix} a & -b \\ c & -d \end{smallmatrix}\right]$ we have $|Q| = 1 - |P|$ so $|Q| \leq \frac{1}{2}$ and $||P|| = ||Q|| \leq \frac{|\det(Q)|}{|Q|} = \frac{|\det(P)|}{1-|P|}$. □

Lemma 6.3 *If $P \in \mathbb{M}(\mathbb{Z})$ and $x \in P^o \cap \overline{\mathbb{Q}}$, then*
1. $6 \cdot ||x||^2 \cdot |P| \cdot |\det(P)| \geq 1$.
2. If $|P| < \frac{1}{2}$ *then* $||P|| \leq \sqrt{3} \cdot ||x|| \cdot |\det(P)|$, *and* $|\mathrm{sz}(P)| \leq \frac{3}{2}||x||^2 \cdot |\det(P)|$.

Proof: If $|P| \geq \frac{1}{2}$, then the first inequality is satisfied trivially. Let $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and $|P| < \frac{1}{2}$, so $ab + cd > 0$. For $x = \frac{p}{q}$ we have $P^{-1}x = \left[\begin{smallmatrix} -d & b \\ c & -a \end{smallmatrix}\right] \cdot \frac{p}{q} = \frac{\beta}{\alpha}$, where $\alpha = pc - aq$, $\beta = qb - pd$. Since $x \in P^\circ$, $\mathrm{sgn}(\frac{\beta}{\alpha}) > 0$. Replacing x by $\frac{-p}{-q}$ if necessary, we can assume that $\alpha > 0$ and $\beta > 0$. Since $|P| < \frac{1}{2}$, either $\frac{0}{1} \notin P^c$ or $\frac{1}{0} \notin P^c$. Assume first $\frac{1}{0} \notin P^o$. Then $\mathrm{sgn}(P^{-1} \cdot \frac{1}{0}) = \mathrm{sgn}(\frac{-d}{c}) < 0$, so $cd > 0$. Since $x \neq \frac{1}{0}$, we have $q \neq 0$ and

$$q \det(P) = qad - qbc = (pc - \alpha)d - (pd + \beta)c = -(\alpha d + \beta c)$$

so $\alpha|d| + \beta|c| = |\alpha d + \beta c| = |q \det(P)|$. Since $\alpha, \beta, |c|, |d|$ are positive integers, we get

$$\alpha + \beta \leq |q \det(P)|, \ |c + d| = |c| + |d| \leq |q \det(P)|.$$

From $a + b = \frac{pc - \alpha + pd + \beta}{q}$ we get

$$|a + b| \leq \frac{|pc| + \alpha + |pd| + \beta}{|q|} \leq (|p| + 1) \cdot |\det(P)|.$$

Since $ab + cd > 0$ we get

$$||P||^2 < (a + b)^2 + (c + d)^2 \leq ((|p| + 1)^2 + q^2) \cdot \det(P)^2 \leq 3 \cdot ||x||^2 \cdot \det(P)^2$$

so we have proved $||P|| \leq \sqrt{3} \cdot ||x|| \cdot |\det(P)|$. It follows $4(ab + cd) \leq 2||P||^2 \leq 6||x||^2 \cdot \det(P)^2$, so $|\mathrm{sz}(P)| = \frac{ab + cd}{|\det(P)|} \leq \frac{3}{2} \cdot ||x||^2 \cdot |\det(P)|$. Similarly if $0 \notin P$ then $p \neq 0$, $\mathrm{sgn}(P^{-1} \cdot \frac{0}{1}) = \mathrm{sgn}(\frac{b}{-a}) < 0$, so $ab > 0$. We get $p \det(P) = -(\alpha b + \beta a)$, so $(\alpha|b| + \beta|a|) = |p \det(P)|$. It follows $\alpha + \beta \leq |p \det(P)|$, $|a| + |b| \leq |p \det(P)|$,

$$c + d \leq \frac{|aq| + |bq| + \alpha + \beta}{|p|} \leq (|q| + 1) \cdot |\det(P)|.$$

We get again $||P||^2 \leq 3 \cdot ||x||^2 \cdot \det(P)^2$ and $|\mathrm{sz}(P)| \leq \frac{3}{2} \cdot ||x||^2 \cdot |\det(P)|$.
The inequality $6 \cdot ||x||^2 \cdot |P| \cdot |\det(P)| \geq 1$ is satisfied whenever $|P| \geq \frac{1}{4}$. If $|P| \leq \frac{1}{4}$ then by Proposition 5.9 we get $|P| \geq \frac{1}{4|\mathrm{sz}(P)|} \geq \frac{1}{6||x||^2 \cdot |\det(P)|}$. □

6.2 Rational number systems

We consider number systems whose transformations have rational entries. By multiplying by the common denominator, we can assume that $F_a \in \mathbb{M}(\mathbb{Z})$. In interval number systems (F, W) or sofic number systems we assume also $W_a \in \mathbb{M}(\mathbb{Z})$ or $V_p \in \mathbb{M}(\mathbb{Z})$. For interval number systems with integer entries we use the concept of **rational expansion interval**

Definition 6.4 *The rational expansion interval of $M \in \mathbb{M}(\mathbb{Z})$ is defined by*

$$\mathbf{R}(M) = \{x \in \overline{\mathbb{Q}} : \ (M^{-1})^\bullet(x) > |\det(M)|\}.$$

Proposition 6.5 *Let $M \in \mathbb{M}(\mathbb{Z})$.*

1. $\mathbf{R}(M) \subseteq \mathbf{V}(M)$ *is a (possibly empty) open interval.*
2. $0, \infty \notin \mathbf{R}(M)$*, so either* $\mathbf{R}(M) \subseteq (\infty, 0)$ *or* $\mathbf{R}(M) \subseteq (0, \infty)$*.*
3. If $M^\bullet(0) = \det(M)$ *then* $M(0) \in \{0, \infty\}$
4. If $M^\bullet(\infty) = \det(M)$ *then* $M(\infty) \in \{0, \infty\}$
5. If $x \in \mathbf{R}(M) \cap \overline{\mathbb{Q}}$*, then* $||M^{-1}(x)|| < ||x||$*.*

Proof: Let $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$.
1. $\mathbf{R}(M)$ is an interval by the proof of Proposition 3.31.
2. We have $(M^{-1})^\bullet(0) = \frac{\det(M)}{b^2+a^2} \leq \det(M)$, so $0 \notin \mathbf{R}(M)$. We have $(M^{-1})^\bullet(\infty) = \frac{\det(M)}{c^2+d^2} \leq \det(M)$, so $\infty \notin \mathbf{R}(M)$.
3. If $M^\bullet(0) = \frac{\det(M)}{b^2+d^2} = \det(M)$ then $M(0) = \frac{b}{d} \in \{0, \infty\}$.
4. If $M^\bullet(\infty) = \frac{\det(M)}{a^2+c^2} = \det(M)$ then $M(\infty) = \frac{a}{c} \in \{0, \infty\}$.
5. If $x \in \mathbf{R}(M) \cap \overline{\mathbb{Q}}$ then $(M^{-1})^\bullet(x) = \frac{\det(M)\cdot||x||^2}{||M^{-1}(x)||^2} > \det(M)$, so $||M^{-1}(x)|| < ||x||$. □

Definition 6.6 *We say that* (F, W) *is a* **rational interval number system** *of order* $n \geq 1$*, if* $F_a, W_a \in \mathbb{M}(\mathbb{Z})$ *for each* $a \in A$ *and* $W_u \subseteq \mathbf{R}(F_u)$ *for each* $u \in \mathcal{L}^n_{F,W}$*.*

The system of **symmetric continued fractions** of Definition 1.14 is a a rational number system of order 1. Since all its transformations have unit determinant, we have $\mathbf{R}(F_a) = \mathbf{V}(F_a)$. For the same reason, the system of signed continued fractions from Example 4.5 is a rational number system of order 2.

Theorem 6.7 *A rational interval number system is neither redundant nor expansive.*

Proof: Since $0, \infty \notin W_a$ for any $a \in A$, the system is not redundant. We show that $\mathbf{Q}_n = \max\{F_u^\bullet(x) : \ u \in \mathcal{L}^n_{F,W}, x \in F_u^{-1}(\overline{W_u})\} = 1$. Let u be any expansion of 0, and $x_n = F^{-1}_{u_{[0,n)}}(0)$. Then $u_i \in \{0, \infty\}$ and $(F^{-1}_{u_n})^\bullet(x_n) = 1$, so $\mathbf{Q}_n = 1$. □

Theorem 6.8 (Delacourt and Kůrka [12]) *If* (F, W) *is a rational interval number system, then each rational number* $x \in \overline{\mathbb{Q}}$ *has a periodic expansion and* $\mathcal{S}_{F,W}$ *is a SFT.*

Proof: We prove the theorem for the order $n = 1$ since the case of a general order is similar. Thus we assume that $W_a \subseteq \mathbf{R}(F_a)$. If $u \in \mathcal{S}_{F,W}$ is an expansion of $x \in \overline{\mathbb{Q}}$, then for $x_n = F^{-1}_{u_{[0,n)}}(x) \in \overline{W_{u_n}} \subseteq \overline{\mathbf{R}(F_{u_n})}$ we have by Proposition 6.5 $||x_{n+1}|| \leq ||x_n||$ so there exists $m \geq 0$ and $n > 0$ such that $x_n = x_m$. Then $u_{[0,m)}(u_{[m,n)})^\omega$ is a periodic expansion of x. Thus each rational number has a periodic expansion.
We show that $\mathcal{S}_{F,W}$ is a sofic subshift. Define by induction

$$\begin{aligned} \mathcal{E}_0 &= \{\mathfrak{l}(W_a), \mathfrak{r}(W_a), \ a \in A\}, \\ \mathcal{E}_{n+1} &= \{F_a^{-1}(x) : \ a \in A, x \in \overline{W_a} \cap \mathcal{E}_n\}. \end{aligned}$$

If $x \in \overline{W_a} \cap \mathcal{E}_n$, then $||F_a^{-1}(x)|| \leq ||x||$ by Proposition 6.5, so there exists n such that $\mathcal{E}_{n+1} = \mathcal{E}_n$. Let $V = \{V_p \subseteq \overline{\mathbb{R}} : \ p \in B\}$ be the open interval partition with endpoints $\mathcal{E}(V) = \mathcal{E}_n$. If $V_p \cap W_a \neq \emptyset$, then both endpoints of $F_a^{-1}(V_p \cap W_a)$ belong to $\mathcal{E}_n$, so if $V_q \cap F_a^{-1}(V_p^\circ \cap W_a) \neq \emptyset$, then $V_q \subseteq F_a^{-1}(V_p \cap W_a)$. By Theorem 4.23, $\mathcal{S}_{F,W}$ is a sofic subshift and its labelled graph is $G = (B, E)$ with vertices B and edges $p \xrightarrow{a} q$ iff $V_q \subseteq F_a^{-1}(V_p \cap W_a)$. This graph determines

the SFT $\Sigma_{|G|} \subseteq E^\omega$ of order two such that $(p,a,q)(r,b,s) \in \mathcal{L}^2_{|G|}$ iff $q = r$. A path in this graph is a pair $(p,u) \in B^\omega \times A^\omega$ such that $p_i \xrightarrow{u_i} p_{i+1}$ for each i. This implies $V_{p_i} \cap W_{u_i} \neq \emptyset$. We have a factor map $\pi : \Sigma_{|G|} \to \Sigma_G = \mathcal{S}_{F,W}$ which is the projection $\pi(p,u) = u$. We show that π is bijective, i.e., that for each $u \in \mathcal{S}_{F,W}$ there exists a unique $p \in B^\omega$ such that $(p,u) \in \Sigma_{|G|}$. For a given $u \in \mathcal{S}_{F,W}$ denote by $x = \Phi(u)$ and $x_n = F^{-1}_{u_{[0,n)}}(x) \in \overline{W_{u_n}}$. If x is irrational, then all x_n are irrational and for each n there exists a unique p_n such that $x_n \in V_{p_n}$, so $(p,u) \in \mathcal{S}_{F,W,V}$. If x is rational then all x_n are rational. Since $x_n \in \overline{\mathbf{R}(F_{u_i})}$, we have $||x_{n+1}|| \leq ||x_n||$. If $x_n \in \mathbf{R}(F_{u_i})$ then $||x_{n+1}|| < ||x_n||$, so there exists only a finite number of indices n with $x_n \in \mathbf{R}(F_{u_n})$. Thus there exists n_0 such that for all $n \geq n_0$, x_n is an endpoint of $\mathbf{R}(F_{u_n})$ and therefore also an endpoint of W_{u_n}. It follows that there exists a unique p_n such that $x \in V_{p_n}$ and $V_{p_n} \cap W_{u_n} \neq \emptyset$. For each $m \leq n$ there exists unique p_m such that $x_m \in \overline{V_{p_m}} \cap \overline{W_{u_m}}$ and $F^{-1}_{u_{[m,n)}}(V_{p_m} \cap W_{u_m}) \cap V_{p_n} \neq 0$: either x_m is an inner point of V_{p_m} or x_m is an endpoint of V_{p_m} but for the other p'_m with $x_m \in \overline{V_{p'_m}} \cap \overline{W_{u_m}}$ we get $F^{-1}_{u_{[m,n)}}(V_{p'_m} \cap W_{u_m}) \cap V^\circ_{p_n} = 0$. Thus the projection $\pi : \Sigma_{|G|} \to \mathcal{S}_{F,W}$ is bijective. Since a homomorphic image of a SFT is a SFT, $\mathcal{S}_{F,W}$ is a subshift of finite type. □

6.3 Modular systems

Definition 6.9 *A transformation $M \in \mathbb{M}(\mathbb{Z})$ is modular, if* $\det(M) = 1$. *We say that (F,Σ) is a* **modular number system**, *if each F_a is a modular transformation.*

The number system of signed continued fractions and the number system of symmetric continued fractions are modular systems. For a modular transformation we have $\mathbf{R}(M) = \mathbf{V}(M)$, so a modular interval number system is rational. Thus if (F,W) is a modular interval number system, then each rational number has a periodic expansion and $\mathcal{S}_{F,W}$ is a SFT. On the other hand a modular system is neither redundant nor expansive. Despite this fact, we show that the unary algorithm works in modular systems for the Möbius transformations with integer entries. For each input the algorithm gives an infinite output and the size of the state matrix of the algorithm remains bounded during the computation. This implies that the algorithm has linear time complexity and can be computed by a finite state transducer. We first prove an auxiliary Lemma.

Lemma 6.10 *Consider the unary graph in a modular sofic number system (F,G,V) of order* 1.

1. *If $(X,p,q) \xrightarrow{a,\lambda} (Y,r,q)$ is an absorption and $p \neq \mathbf{i}$ then $|Y| < |X|$.*
2. *If $(X,p,q) \xrightarrow{\lambda,a} (Y,p,r)$ is an emission then $|Y| > |X|$ and $||Y|| \leq ||X||$.*

Proof: 1. Since $H_{p,a,r}$ is a nonnegative matrix, we have $Y = XH_{p,a,r} \subset X$, so $|Y| < |X|$.
2. Since $V_r \subseteq \mathbf{U}(F_a)$, we have $X \subseteq F_a V_r \subseteq \mathbf{V}(F_a)$ and $|Y| > |X|$. For each $x \in X^c$ we have $(F_a^{-1})^\bullet(x) \geq 1$ and therefore $||F_a^{-1}(x)|| \leq ||x||$. In particular this holds for both endpoints of X which implies $||F_a^{-1}X|| \leq ||X||$. □

In Theorem 5.13 we have proved that a redundant sofic system (F,G,V) has a threshold and the unary algorithm with greedy selector computes a mapping $\Theta_{M,s} : \Sigma_{|G|} \to \Sigma_{|G|}$ such that if $\Theta_{M,s}(p,u) = (q,v)$ then $M\Phi(u) = \Phi(v)$. Modular systems have a weaker property: they have local thresholds.

Proposition 6.11 *A greedy selector in a modular sofic number system has a* **local threshold**. *This means that for each $X \in \mathbb{M}(\mathbb{Z})$ there exists a threshold $\tau(X) > 0$ such that in the computation (X_i, p_i, q_i) with initial state $(X, \mathbf{i}, \mathbf{i})$ each absorption state (X_i, p_i, q_i) satisfies $|X_i| \geq \tau(X)$*

Proof: Set

$$\begin{aligned} C &= 6 \cdot \max\{||\mathbf{l}(F_a V_r)||^2, ||\mathbf{r}(F_a V_r)||^2 : \xrightarrow{a} r\}, \\ D &= \max\{|\det(V_p)| : p \in B\} \end{aligned}$$

Let $(X_0, p_0, q_0) \overset{(u_0,v_0)}{\longrightarrow} (X_1, p_1, q_1) \overset{(u_1,v_1)}{\longrightarrow} \cdots$ be a path in the unary graph computed by a greedy selector s (here $u_i, v_i \in A \cup \{\lambda\}$). Since each F_a is modular, and $X_i = F^{-1}_{v_{[0,i)}} X_0 F_{u_{[0,i)}} V_{p_i}$, we get $|\det(X_i)| \leq D \cdot |\det(X_0)|$. If (X_i, p_i, q_i) is an absorption state, then X_i contains an endpoint x of some $F_a V_r$, so by Lemma 6.3

$$CD \cdot |\det(X_0)| \cdot |X_i| \geq 6 \cdot ||x||^2 \cdot |\det(X_i)| \cdot |X_i| \geq 1.$$

Thus $|X_i| \geq \tau(X_0) = \frac{1}{CD|\det(X_0)|}$. □

Corollary 6.12 *In modular sofic systems, the unary algorithm with a greedy selector computes for each $M \in \mathbb{M}(\mathbb{Z})$ a continuous mapping $\Theta_{M,s} : \Sigma_G \to \Sigma_G$ such that $\Phi\Theta_{M,s} = M\Phi$.*

Proof: By Proposition 6.11 each computation of the unary algorithm contains an infinite number of both absorptions and emissions, so analogously as in Theorem 5.13 we prove that the unary algorithm with a greedy selector computes for each $M \in \mathbb{M}(\mathbb{Z})$ a continuous mapping $\Theta_{M,s} : \Sigma_{|G|} \to \Sigma_{|G|}$ such that if $\Theta_{M,s}(p, u) = (q, v)$ then $\Phi(v) = M(\Phi(u)$. By Theorem 6.8, $\Sigma_{|G|}$ is conjugated to Σ_G, and we get the result. □

We show now that in each computation of a greedy selector, the norm of the state matrix remains bounded.

Theorem 6.13 (Delacourt and Kůrka [12]) *Let s be a greedy selector in a sofic modular number system. Then for each $X \in \mathbb{M}(\mathbb{Z})$ there exists a bound $\nu(X) > 0$ such that for each computation $(X_0, \mathbf{i}, \mathbf{i}) \overset{u_0,v_0}{\longrightarrow} (X_1, p_1, q_1) \overset{u_1,v_1}{\longrightarrow} (X_2, p_2, q_2) \cdots$ we have $||X_i|| \leq \nu(X_0)$ for each i.*

Proof: Denote by $\tau(X_0)$ the local threshold from Proposition 6.11. Let C, D be the constants from its proof and set

$$\begin{aligned} L &= \max\{|\mathbf{U}(F_a)| : a \in A\}, \\ H &= \max\{||H_{p,a,q}|| : p \xrightarrow{a} q\} \end{aligned}$$

Each path $(X_0, \mathbf{i}, \mathbf{i}) \overset{u_0,v_0}{\longrightarrow} (X_1, p_1, q_1) \overset{u_1,v_1}{\longrightarrow} (X_2, p_2, q_2) \cdots$ of a greedy selector contains an infinite number of both absorptions and emissions and $|\det(X_i)| \leq D \cdot |\det(X_0)|$ for each i. If (X_n, p_n, q_n) is an emission state and (X_i, p_i, q_i) are absorption states for $n < i < m$, then $|X_n| \leq |F_{u_n} V_{q_{n+1}}| \leq |\mathbf{V}(F_{u_n})| < \frac{1}{2}$ and by Lemma 6.10,

$$1 > L > |X_{n+1}| > |X_{n+2}| > \cdots > |X_m|.$$

If n_0 is the time of the first emission, then $|X_n| \le L$ for each $n \ge n_0$. If $n \ge n_0$, (X_n, p_n, q_n) is an absorption state and (X_i, p_i, q_i) are emission states for $n < i < m$, then by Proposition 6.11, $|X_n| > \tau(X_0)$, so

$$\begin{aligned} ||X_n|| &\le M = D \cdot |\det(X_0)| \cdot \max\{\tfrac{1}{\tau(X_0)}, \tfrac{1}{1-L}\} \\ M \cdot H &\ge ||X_{n+1}|| > ||X_{n+1}|| > \cdots > ||X_m||. \end{aligned}$$

Denote by $\nu_n(X_0) = \max\{||X_i|| : \ 0 \le i \le n\}$ and let n_1 be the time of the first absorption with $n_1 > n_0$. Then $||X_i|| \le \nu(X_0) = \max\{M \cdot H, \nu_{n_1}(X_0)\}$ for every i. □

A special case of Theorem 6.13 for simple continued fractions has been proved by Raney [57]. Since there is only a finite number of matrices $X \in \mathbb{M}(\mathbb{Z})$ whose norm $||X||$ does not exceed a given bound, there is only a finite number of vertices (X_i, p_i, q_i) which appear in the computation of the unary algorithm. This means that the computation of the unary algorithm can be done by a **finite state transducer** which is a finite automaton with an output function.

6.4 Finite state transducers

Definition 6.14 *A finite state transducer over an alphabet A is a quadruple $\mathcal{T} = (Q, \delta, \tau, \mathbf{i})$, where $(Q, \delta, \mathbf{i})$ is an accepting automaton (Definition 2.28) and $\tau : A \times Q \to A^*$ is a partial output function with the same domain as δ.*

For each $u \in A^*$ we have a partial mapping $\tau_u : Q \to A^*$ defined by induction: $\tau_\lambda(p) = \lambda$, $\tau_{ua}(p) = \tau_u(p)\tau(a, \delta_u(p))$ (concatenation). The output mapping works also for infinite words. If u is a prefix of v, then $\tau_u(p)$ is a prefix of $\tau_v(p)$, so for each $p \in Q$ and $u \in A^\omega$ there exists a unique $\tau_u(p) \in A^* \cup A^\omega$ such that each $\tau_{u_{[0,n)}}(p)$ is its prefix. A finite transducer determines a labelled oriented graph, whose vertices are elements of Q. There is an oriented edge $p \stackrel{(a,v)}{\longrightarrow} q$ iff $\delta_a(p) = q$ and $\tau_a(p) = v$. The label of a path is the concatenation of the labels of its edges, so there is a path $p \stackrel{(u,v)}{\longrightarrow} q$ iff $\delta_u(p) = q$ and $\tau_u(p) = v$.

It follows from Theorem 6.13 that for a given sofic modular system (F, G, V), a greedy selector s and an initial transformation $M \in \mathbb{M}(\mathbb{Z})$ there exists a finite state transducer $\mathcal{T} = (Q, \delta, \tau, \imath)$ which computes M. The state set Q consists of the absorption states $(X, p, q) \in \mathbb{M}(\mathbb{Z}) \times B^2$ such that $||X|| \le \nu(M)$, where $\nu(M)$ is the bound from Theorem 6.13. For a given $(X, p, q) \in Q$ and $a \in A$ with $p \stackrel{a}{\longrightarrow} p_0$ take the path $(X, p, q) \stackrel{a,\lambda}{\longrightarrow} (X_0, p_0, v_0) \stackrel{\lambda, v_0}{\longrightarrow} \cdots \stackrel{\lambda, v_{n-1}}{\longrightarrow} (X_n, p_n, q_n)$ such that (X_i, p_i, q_i) are emission states for $i < n$ and (X_n, p_n, q_n) is an absorption state. Then we define partial mappings $\delta : Q \times A \to Q$, $\tau : Q \times A \to A^*$ by $\delta((X, p, q), a) = (X_n, p_n, q_n)$ and $\tau((X, p, q), a) = v_0 \cdots v_{n-1}$. If $n = 0$ then $\tau((X, p, q), a) = \lambda$. The initial state of the transducer is $\imath = (M, \mathbf{i}, \mathbf{i})$.

Definition 6.15 *Let (F, G, V) be a sofic number system with an alphabet A and an initialized graph $G = (B, E, \mathbf{i})$. We say that a finite state transducer $\mathcal{T} = (Q, \delta, \tau, \imath)$ extends G if there is a projection $\pi : Q \to B$ such that $\pi(\imath) = \mathbf{i}$, and if $\delta(p, a) = q$ then $\pi(p) \stackrel{a}{\longrightarrow} \pi(q)$ in G. We say that $\mathcal{T}$ computes a real function $g : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$, if $\mathcal{T}$ extends G and for any $u \in \Sigma_G$ we have $v = \tau_u(\imath) \in \Sigma_G$ and $\Phi(v) = g(\Phi(u))$.*

If we define $\Theta_g(u) = \tau_u(\imath)$, then $\Theta : \Sigma_G \to \Sigma_G$ is a continuous function which satisfies $\Phi\Theta_g = g\Phi$.

Corollary 6.16 *For a modular sofic number system (F, G, V) with a deterministic graph G and a transformation $M \in \mathbb{M}(\mathbb{Z})$ there exists a finite state transducer which computes M.*

On the other hand, we show that Möbius transformations are the only functions which are computable by finite state transducers in sofic number systems. This has been proved in Konečný [31], who assumes that the function in question is differentiable and has nonzero derivative at the fixed point of the transformations. A similar result has been obtained by Kůrka and Vávra [45] for the case of analytic functions. Recall that $\mathcal{F}_p = \{u \in \Sigma_G^* : p \xrightarrow{u}\}$ is the follower set of $p \in B$.

Proposition 6.17 *Assume that a finite state transducer $\mathcal{T} = (Q, \delta, \tau, \imath)$ computes a real function g in a sofic number system (F, G, V) with an initialized graph G. Then for every state $p \in Q$ there exists a real function $g_p : V_p \to \overline{\mathbb{R}}$ such that if $w \in \mathcal{F}_p$ and $\tau_w(p) = z$, then $\Phi(z) = g_p(\Phi(w))$. We say that $\mathcal{T}$ computes g_p at state p. If $u, v \in \mathcal{L}(\Sigma)$, and $p \xrightarrow{(u,v)} q$, then $g_q = F_v^{-1} g_p F_u$.*

Proof: Assume that $\imath \xrightarrow{(u,v)} p \xrightarrow{(w,z)}$ with $w, z \in \Sigma_G$ and set $g_p = F_v^{-1} g F_u$. Then

$$g_p \Phi(w) = F_v^{-1} g F_u \Phi(w) = F_v^{-1} g \Phi(uw) = F_v^{-1} \Phi(vz) = \Phi(z),$$

so $\mathcal{T}$ computes g_p at p. If $p \xrightarrow{(u,v)} q \xrightarrow{(w,z)}$, then $F_v^{-1} g_p F_u \Phi(w) = F_v^{-1} g_p \Phi(uw) = F_v^{-1} \Phi(vz) = \Phi(z)$, so $\mathcal{T}$ computes $F_v^{-1} g_p F_u$ at q and is equal to g_q. □

Lemma 6.18 *Assume that a finite transducer $\mathcal{T} = (Q, \delta, \tau, \imath)$ computes a nonconstant rational function g in a sofic number system (F, G, V) and let $p \xrightarrow{u,v} p$ be a path in $\mathcal{T}$. Then F_u and F_v are either hyperbolic or decreasing transformations.*

Proof: By Proposition 6.17 $\mathcal{T}$ computes at the vertex p a function $g_p : \Phi(\mathcal{F}_p) \to \overline{\mathbb{R}}$ with $g_p F_u = F_v g_p$. If g is rational, then g_p is a rational function defined on the interval $\Phi(\mathcal{F}_p)$ which extends to a unique rational function defined on whole $\overline{\mathbb{R}}$. Since $\pi(p) \xrightarrow{u} \pi(p)$ is a path in G, we get $u^\omega \in \Sigma_G$, so F_u is not elliptic. Since $\Theta(u^\omega) = v^\omega$ we get $v^\omega \in \Sigma_G$, so F_v is not elliptic. We show that neither F_u nor F_v is parabolic. We distinguish three cases. 1. If both F_u and F_v are parabolic, they are conjugated to the translation $T^1(x) = x + 1$, so there exist transformations f_0, f_1 such that $F_u = f_0 T^1 f_0^{-1}$, $F_v = f_1 T^1 f_1^{-1}$. For the rational function $h = f_1^{-1} g_p f_0$ we get

$$T^1 h \;=\; T^1 f_1^{-1} g_p f_0 = f_1^{-1} F_v g_p f_0 = f_1^{-1} g_p F_u f_0 = f_1^{-1} g_p f_0 T^1 = h T^1,$$

so $h(x + 1) = h(x) + 1$. It follows that the rational function $h_0(x) = h(x) - x$ is periodic: $h_0(x + 1) = h(x + 1) - x - 1 = h(x) - x = h_0(x)$. However, no rational function is periodic.
2. If F_u is parabolic and F_v is hyperbolic or decreasing, then there exist transformations f_0, f_1 such that $F_u = f_0 T^1 f_0^{-1}$, $F_v = f_1 Q_r f_1^{-1}$, where $Q_r(x) = rx$ and $0 \neq r \neq 1$. For the rational function $h = f_1^{-1} g_p f_0$ we get $Q_r h = h T^1$, which implies $h(x + n) = h(x) \cdot r^n$ for each integer n. If $r = -1$ then h is a periodic function with period 2, which is impossible. If $r \neq -1$ and $h(x) \neq 0$, then we get $\lim_{n\to\infty} h(x + n) \neq \lim_{n\to-\infty} h(x + n)$: one of these limits is zero and the other is infinity. This means that h is not continuous at ∞ which is a contradiction. Thus $h(x) = 0$ for all x and g_p is a constant function.
3. If F_u is hyperbolic or decreasing and F_v is parabolic, then there exist transformations f_0, f_1 such that $F_u = f_0 Q_r f_0^{-1}$, $F_v = f_1 T^1 f_1^{-1}$, where $Q_r(x) = rx$ and $0 \neq r \neq 1$. For the rational function $h = f_1^{-1} g_p f_0$ we get $T^1 h = h Q_r$, i.e., $h(x)+1 = h(rx)$. For $x = 0$ we get $h(0)+1 = h(0)$ which is a contradiction. □

Lemma 6.19 *Let g be a real rational function of degree $n \geq 2$, and let $F_0, F_1, F_2, F_3 \in \mathbb{M}(\mathbb{R})$ be hyperbolic or decreasing transformations such that $F_0 g = g F_1$, $F_2 g = g F_3$. Then F_2 has the same fixed points as F_0 and F_3 has the same fixed points as F_1.*

Proof: There exist transformations f_0, f_1 and r_0, r_1 different from 0 and 1 such that $F_0 = f_0 Q_{r_0} f_0^{-1}$, $F_1 = f_1 Q_{r_1} f_1^{-1}$. For the rational function $h = f_0^{-1} g f_1$ we get

$$Q_{r_0} h = Q_{r_0} f_0^{-1} g f_1 = f_0^{-1} F_0 g f_1 = f_0^{-1} g F_1 f_1 = f_0^{-1} g f_1 Q_{r_1} = h Q_{r_1},$$

so $h(r_1^m x) = h(x) r_0^m$. The only rational functions which satisfy this equation are of the form $h(x) = p x^n$. From $\deg(g) \geq 2$ we get $n \geq 2$ and

$$f_0^{-1} F_2 f_0 h = f_0^{-1} F_2 g f_1 = f_0^{-1} g F_3 f_1 = h f_1^{-1} F_3 f_1.$$

Setting $f_0^{-1} F_2 f_0 = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, $f_1^{-1} F_3 f_1 = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ we get

$$(apx^n + b)(Cx + D)^n = p(cpx^n + d)(Ax + B)^n$$

Comparing the coeficients at x^{2n} and x^{2n-1} we get $aC^n = pcA^n$, $aC^{n-1}D = pcA^{n-1}B$, so $pcA^n D = aC^n D = pcA^{n-1}BC$, and $pcA^{n-1}(AD - BC) = 0$. Thus $cA = 0$ and it follows $aC = 0$. Comparing the coeficients at x and x^0, we get $bCD^{n-1} = pdAB^{n-1}$, $bD^n = pdB^n$, so $pdAB^{n-1}D = bcD^n = pdCB^n$ and $pdB^{n-1}(AD - BC) = 0$. Thus $dB = 0$ and it follows $bD = 0$. We have therefore proved $cA = aC = dB = bD = 0$. Since both matrices are regular, either $A = D = a = d = 0$ or $B = C = b = c = 0$. In the former case, F_2 and F_3 would be elliptic which is excluded by the assumption. Thus $B = C = b = c = 0$, so both $f_0^{-1} F_2 f_0$ and $f_1^{-1} F_3 f_1$ have the fixed points 0 and ∞. It follows that F_2 has the same fixed points as F_0 and F_3 has the same fixed points as F_1. □

Theorem 6.20 (Kůrka and Vávra [45]) *A rational function of degree 2 or more cannot be computed by a finite state transducer in a sofic number system.*

Proof: Assume that a finite state transducer $\mathcal{T} = (Q, \delta, \tau, \imath)$ computes a rational function h of degree $\deg(h) \geq 2$ in (F, G, V). Then each vertex p computes a rational function of the same degree. Take any infinite path $\imath \xrightarrow{u,v}$ in $\mathcal{T}$. There exists a state $p \in Q$ which occurs infinitely many times in this path, so we have an infinite sequence of finite words $u^{(i)}, v^{(i)}$ such that

$$i \xrightarrow{(u^{(0)},v^{(0)})} p \xrightarrow{(u^{(1)},v^{(1)})} p \xrightarrow{(u^{(2)},v^{(2)})} p \cdots$$

By Lemma 6.19, all $F_{u^{(i)}}$ with $i > 0$ are either hyperbolic or decreasing and have the same fixed points. It follows that $\Phi(u) = F_{u^{(0)}}(s)$, where s is one of the fixed points of $F_{u^{(1)}}$. However, the set of such points $\Phi(u)$ is countable, while the mapping $\Phi : \Sigma_G \to \overline{\mathbb{R}}$ is assumed to be surjective. This is a contradiction. □

6.5 Bimodular systems

As an examples of a rational number system which is not modular, consider the **bimodular number system** which extends the binary signed system and consists of all transformations $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ with $\det(M) = ad - bc = 2$, $\mathrm{tr}(M) = a + d = 3$ and $||M||^2 = a^2 + b^2 + c^2 + d^2 = 6$ (see Kůrka [39]).

Example 6.21 *The bimodular number system has alphabet* $A = \{0, 1, 2, 3, 4, 5, 6, 7\}$ *and transformations with matrices*

$$F_0 = [\tfrac{1}{1}, \tfrac{0}{2}],\ F_1 = [\tfrac{1}{0}, \tfrac{1}{2}],\ F_2 = [\tfrac{2}{1}, \tfrac{0}{1}],\ F_3 = [\tfrac{2}{0}, \tfrac{1}{1}),$$
$$F_4 = [\tfrac{2}{0}, \tfrac{-1}{1}],\ F_5 = [\tfrac{2}{-1}, \tfrac{0}{1}],\ F_6 = [\tfrac{1}{0}, \tfrac{-1}{2}],\ F_7 = [\tfrac{1}{-1}, \tfrac{0}{2}]$$

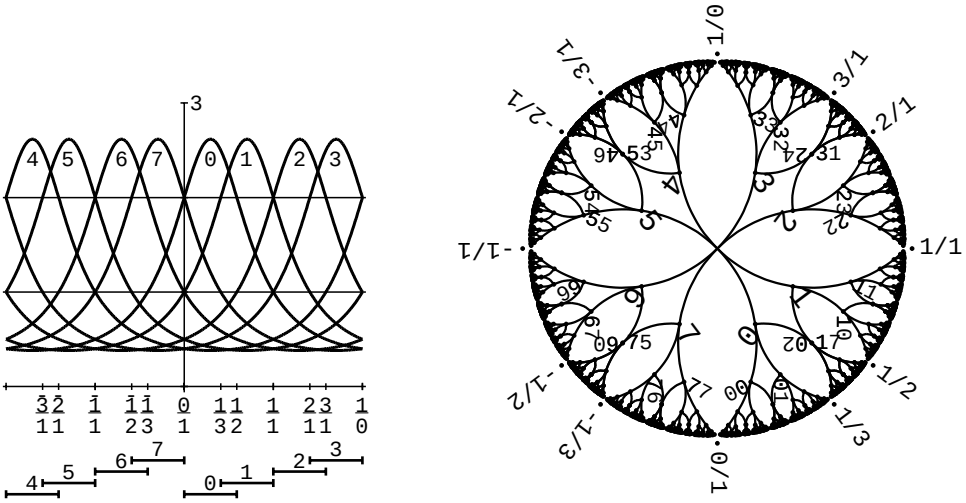


Figure 6.1: The small bimodular system

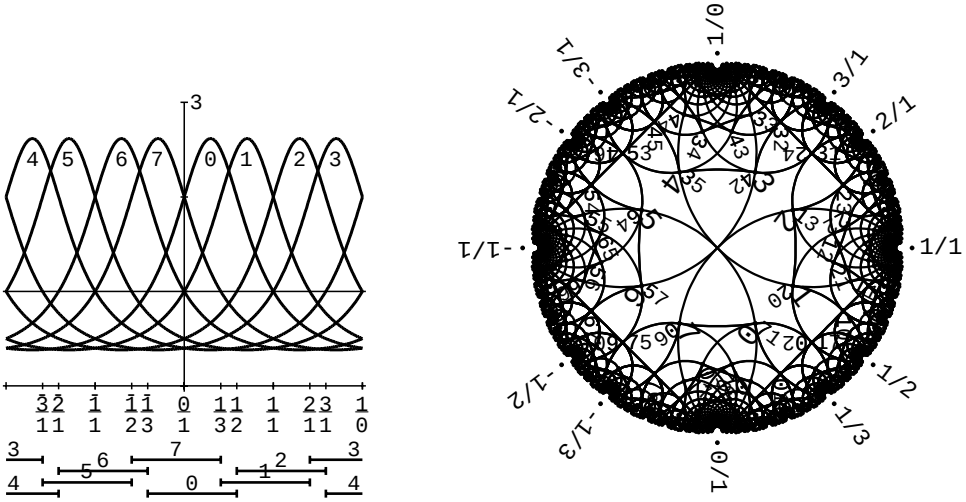


Figure 6.2: The large bimodular system

There exists several number systems with transformations of Example 6.21. The **small bimodular system** $(F, \mathcal{R})$ is the interval number system with intervals $W_a = \mathbf{R}(F_a)$. They form an almost-cover so $(F, \mathcal{R})$ is a rational system with an SFT expansion subshift $\mathcal{S}_{F,W}$ (see Figure 6.1) of order 3 with forbidden words

$$\begin{aligned} D \;=\; & \{03, 04, 05, 06, 07, 12, 13, 14, 15, 16, 20, 21, 25, 26, 27, 30, 34, 35, 36, 37, \\ & 40, 41, 42, 43, 47, 50, 51, 52, 56, 57, 61, 62, 63, 64, 65, 70, 71, 72, 73, 74, \\ & 024, 175, 246, 317, 460, 531, 602, 753\} \end{aligned}$$

a	F_a	$\mathbf{R}(F_a)$	$F_a^{-1}(\mathbf{R}(F_a))$	$\mathbf{V}(F_a)$	$F_a^{-1}(\mathbf{V}(F_a))$
0	$\left[\begin{smallmatrix}1 & 0\\ 1 & 2\end{smallmatrix}\right]$	$(0, \frac{1}{2})$	$(0, 2)$	$(\frac{-1}{3}, 1)$	$(\frac{-1}{2}, \infty)$
1	$\left[\begin{smallmatrix}1 & 1\\ 0 & 2\end{smallmatrix}\right]$	$(\frac{1}{3}, 1)$	$(\frac{-1}{3}, 1)$	$(0, 2)$	$(-1, 3)$
2	$\left[\begin{smallmatrix}2 & 0\\ 1 & 1\end{smallmatrix}\right]$	$(1, 3)$	$(1, -3)$	$(\frac{1}{2}, \infty)$	$(\frac{1}{3}, -1)$
3	$\left[\begin{smallmatrix}2 & 1\\ 0 & 1\end{smallmatrix}\right]$	$(2, \infty)$	$(\frac{1}{2}, \infty)$	$(1, -3)$	$(0, -2)$
4	$\left[\begin{smallmatrix}2 & -1\\ 0 & 1\end{smallmatrix}\right]$	$(\infty, -2)$	$(\infty, \frac{-1}{2}$	$(3, -1)$	$(2, 0)$
5	$\left[\begin{smallmatrix}2 & 0\\ -1 & 1\end{smallmatrix}\right]$	$(-3, -1)$	$(3, -1)$	$(\infty, \frac{-1}{2})$	$(-1, -\frac{1}{3})$
6	$\left[\begin{smallmatrix}1 & -1\\ 0 & 2\end{smallmatrix}\right]$	$(-1, \frac{-1}{3})$	$(-1, \frac{1}{3})$	$(-2, 0)$	$(-3, 1)$
7	$\left[\begin{smallmatrix}1 & 0\\ -1 & 2\end{smallmatrix}\right]$	$(-\frac{1}{2}, 0)$	$(-2, 0)$	$(-1, \frac{1}{3})$	$(\infty, \frac{1}{2})$

Table 6.1: The transformations and intervals of the small and large bimodular interval systems

p	a	q	L^p	R^p	V_q	F_aV_q
$7, 0, 1$	0	0	$(76543210)^\omega$	$(12345670)^\omega$	$[\frac{1}{-2}, \frac{-2}{-1}]$	$[\frac{1}{-3}, \frac{-2}{-4}]$
$0, 1, 2$	1	1	$(07654321)^\omega$	$(23456701)^\omega$	$[\frac{1}{-3}, \frac{-3}{-1}]$	$[\frac{-2}{-6}, \frac{-4}{-2}]$
$1, 2, 3$	2	2	$(10765432)^\omega$	$(34567012)^\omega$	$[\frac{-1}{-3}, \frac{-3}{1}]$	$[\frac{-2}{-4}, \frac{-6}{-2}]$
$2, 3, 4$	3	3	$(21076543)^\omega$	$(45670123)^\omega$	$[\frac{-1}{-2}, \frac{-2}{1}]$	$[\frac{-4}{-2}, \frac{-3}{1}]$
$3, 4, 5$	4	4	$(32107654)^\omega$	$(56701234)^\omega$	$[\frac{-2}{-1}, \frac{-1}{2}]$	$[\frac{-3}{-1}, \frac{-4}{2}]$
$4, 5, 6$	5	5	$(43210765)^\omega$	$(67012345)^\omega$	$[\frac{-3}{-1}, \frac{-1}{3}]$	$[\frac{-6}{2}, \frac{-2}{4}]$
$5, 6, 7$	6	6	$(54321076)^\omega$	$(70123456)^\omega$	$[\frac{3}{-1}, \frac{-1}{-3}]$	$[\frac{4}{-2}, \frac{2}{-6}]$
$6, 7, 0$	7	7	$(65432107)^\omega$	$(01234567)^\omega$	$[\frac{2}{-1}, \frac{-1}{-2}]$	$[\frac{2}{-4}, \frac{-1}{-3}]$

Table 6.2: The circular bimodular system with SFT subshift of speed 1 and order 2

Its expansion quotient is $\mathbf{Q} = 2$ and it is not redundant. The **large bimodular system** $(F, \mathcal{V})$ is an interval number system with intervals $W_a = \mathbf{V}(F_a)$ (see Figure 6.2). Its expansion quotient is $\mathbf{Q} = 1$ and it is redundant. Its expansion subshift is not SFT but it is sofic with the same SFT partition as $(F, \mathcal{R})$, with endpoints 0, $\frac{1}{3}$, $\frac{1}{2}$, 1, 2, 3, ∞, -3, -2, -1, $-\frac{1}{2}$, and $-\frac{1}{3}$. See Table 6.1 for the intervals of both systems. There is also an interval partition system with endpoints

$$0, \sqrt{2}-1, 1, \sqrt{2}+1, \infty, -\sqrt{2}-1, -1, -\sqrt{2}+1.$$

Then there is a redundant bimodular sofic system with the circular subshift Σ_1 with speed 1 and order 2 (see Section 4.7) with forbidden words

$$\begin{aligned} D &= \{ab \in A^2 : \bmod_8(a-b) \in \{2,3,4,5,6\} \\ &= \{02, 03, 04, 05, 06, 13, 14, 15, 16, 17, 20, 24, 25, 26, 27, 30, 31, 35, 36, 37, \\ &\quad 40, 41, 42, 46, 47, 50, 51, 52, 53, 57, 60, 61, 62, 63, 64, 71, 72, 73, 74, 75\} \end{aligned}$$

The deterministic automaton has states $B = \{\mathbf{i}, 0, 1, 2, 3, 4, 5, 6, 7\}$. The intervals V_a together with intervals F_aV_q are given in Table 6.2 and Figure 6.3. The system is not an interval number system. If we take intervals $W_a = \Phi([a])$, we get a different interval number system whose expansion subshift is sofic but not of finite type.

If the unary algorithm is computed in a nonmodular system whose transfomattions have integer entries then the determinant of the state matrix need not remain bounded. If $X, F \in$

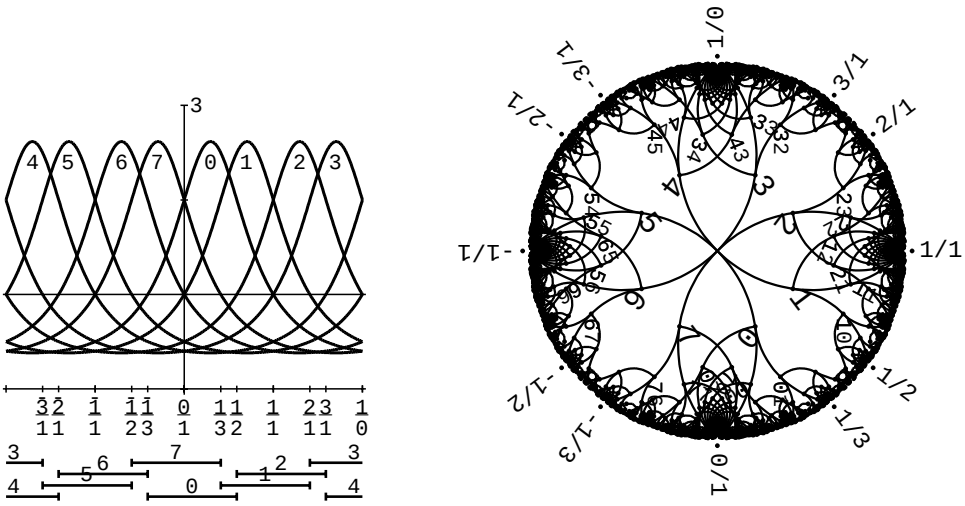


Figure 6.3: The circular bimodular system (F, Σ_1) with the circular subshift of speed 1.

$\mathbb{M}(\mathbb{Z})$ and $Y = FX = \mathbf{d}(F \cdot X)$ then either $\det(Y) = \det(F) \cdot \det(X)$ provided $\gcd(F \cdot X) = 1$ or $\det(Y) = \det(X)/\det(F)$ provided $\gcd(F \cdot X) = \det(F)$ and $\det(F)$ is a prime (see Proposition 6.1). When the unary algorithm is computed in the large bimodular system, the determinant and norm of the state matrix remain small most of the time so the unary algorithm has asymptotically linear time complexity. This is due to the fact that some compositions of the transformations are modular - see Proposition 6.22, whose proof is a simple verification. A selector which takes advantage of this scheme in the large bimodular system takes a small threshold τ and applies an absorption whenever $|XV_p| > \tau$. If $|XV_p| < \tau$ then the selector chooses the an emission of the letter a with the smallest norm of the matrix $F_a^{-1}X$. If τ is sufficiently small then there are usually several possible emissions letters a and the smallest norm of $F_a^{-1}X$ is achieved by cancellation of $F_a^{-1} \cdot X$ by 2 (see Kůrka [42], Kůrka and Delacourt [43]).

Proposition 6.22 *For the bimodular number system of Example 6.21, set* $A_0 = \{1, 2, 5, 6\}$, $A_1 = \{0, 3, 4, 7\}$.
1. If $a_0 \in A_0$, $a_1 \in A_1$, *then* $\det(F_{a_0 a_1}) = 1$.
2. $F_{14} = F_{27} = F_{50} = F_{63} = \mathrm{Id}$.
3. Both $\{\mathbf{V}(F_a) : a \in A_0\}$ *and* $\{\mathbf{V}(F_a) : a \in A_1\}$ *are almost-covers.*

While statistically, cancellations occur frequently in the large bimodular system, there are exceptional cases in which they do not occur at all, so that the determinant and norm of the state matrices steadily grows. We prove this results for general expansive number systems whose transformations have determinants 1 or 2.

Lemma 6.23 *Assume* $F \in \mathbb{M}(\mathbb{Z})$ *and* $|\det(F)| \leq 2$.
1. If $|F^\bullet(0)| > 1$, *then either* $F = \left[\begin{smallmatrix} 2 & 0 \\ c & \pm 1 \end{smallmatrix}\right]$, $F(0) = 0$, *or* $F = \left[\begin{smallmatrix} a & \pm 1 \\ 2 & 0 \end{smallmatrix}\right]$, $F(0) = \infty$.
2. If $|F^\bullet(\infty)| > 1$, *then either* $F = \left[\begin{smallmatrix} 0 & 2 \\ -1 & d \end{smallmatrix}\right]$, $F(\infty) = 0$, *or* $F = \left[\begin{smallmatrix} 1 & b \\ 0 & 2 \end{smallmatrix}\right]$, $F(\infty) = \infty$.

Proof: Let $F = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. If $|F^\bullet(0)| = \frac{|\det(F)|}{b^2+d^2} > 1$, then $|\det(F)| = 2$ since b, d cannot be both zero. Thus $b^2 + d^2 < 2$, so $b, d \in \{-1, 0, 1\}$ and either $b = 0$ or $d = 0$. It follows that either $F = \left[\begin{smallmatrix} 2 & 0 \\ c & \pm 1 \end{smallmatrix}\right]$ or $F = \left[\begin{smallmatrix} a & \pm 1 \\ 2 & 0 \end{smallmatrix}\right]$. If $|F^\bullet(\infty)| = \frac{|\det(F)|}{a^2+c^2} > 1$, then $|\det(F)| = 2$, $a, c \in \{-1, 0, 1\}$ and

either $F = \left[\begin{smallmatrix} \pm 1 & b \\ 0 & 2 \end{smallmatrix}\right]$, or $F = \left[\begin{smallmatrix} 0 & 2 \\ \pm 1 & d \end{smallmatrix}\right]$. □

Theorem 6.24 (Kůrka and Vávra [45]) *Let (F, G, V) be an expansive sofic number system such that $F_a \in \mathbb{M}(\mathbb{Z})$ and $|\det(F_a)| \leq 2$ for each $a \in A$. Then there exists a transformation $M \in \mathbb{M}(\mathbb{Z})$ and an input word $u \in \Sigma_G$ such that in the computation of the unary algorithm on input matrix M and input word u, no cancellation ever occurs.*

Proof: Denote by mod_2 the modulo function whose value is 0 on even numbers and 1 on odd numbers. The modulo function works on integer matrices as well. Choose any transformation M such that $M(0) = 0$ and $\mathrm{mod}_2(M) = (\frac{0}{1}, \frac{0}{0})$, e.g., $M(x) = \frac{2x}{x+2}$. Pick a word $u \in \Sigma_G$ with $\Phi(u) = 0$ and assume that we have a finite automaton which computes M on u with the result v, so $\Phi(v) = 0$. The computation of the automaton determines a path with vertices $(G_{n,m} V_{p_n}, p_{u_n}, q_{v_m})$, where $G_{n,m} = F^{-1}_{v_{[0,m)}} M F_{u_{[0,n)}}$ and in each transition we have either $G_{n,m} \overset{(u_n,\lambda)}{\longrightarrow} G_{n+1,m}$ or $G_{n,m} \overset{(\lambda,v_n)}{\longrightarrow} G_{n,m+1}$. We show by induction that during the process no cancellation ever occurs: either $\det(G_{n+1,m}) = 2\det(G_{n,m})$ or $\det(G_{n,m+1}) = 2\det(G_{n,m})$. Denote by $x_n = \Phi(\sigma^n(u)) = F^{-1}_{u_{[0,n)}} \Phi(u) = F^{-1}_{u_{[0,n)}}(0)$, so $x_0 = 0$ and $y_m = F^{-1}_{v_{[0,m)}} M \Phi(u) = F^{-1}_{v_{[0,m)}}(0)$, so $y_0 = 0$. Denote by $H_{n,m} = \mathrm{mod}_2(G_{n,m})$. We show by induction that $x_n, y_m \in \{0, \infty\}$, and $H_{n,m}$ is determined by x_n, y_m by the table

x_n, y_m	$0,0$	$0,\infty$	$\infty,0$	∞,∞
$H_{n,m}$	$(\frac{0}{1}, \frac{0}{0})$	$(\frac{1}{0}, \frac{0}{0})$	$(\frac{0}{0}, \frac{0}{1})$	$(\frac{0}{0}, \frac{1}{0})$

If $x_n = y_m = 0$, then $(F^{-1}_{u_n})^\bullet(0) > 0$ since the system is expansive, so by Lemma 6.23 either $x_{n+1} = 0$ and then $H_{n+1,m} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{0}{c}, \frac{0}{1})^{-1} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{1}{c}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0})$, or $x_{n+1} = \infty$ and then $H_{n+1,m} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{a}{0}, \frac{1}{0})^{-1} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{0}{0}, \frac{1}{a}) = (\frac{0}{0}, \frac{0}{1})$. Similarly $(F^{-1}_{v_m})^\bullet(0) > 0$ so by Lemma 6.23 either $y_{m+1} = 0$ and then $H_{n,m+1} = (\frac{0}{c}, \frac{0}{1}) \cdot (\frac{0}{1}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0})$, or $y_{m+1} = \infty$ and then $H_{n,m+1} = (\frac{a}{0}, \frac{1}{0}) \cdot (\frac{0}{1}, \frac{0}{0}) = (\frac{1}{0}, \frac{0}{0})$. Similarly in other cases:

$$
\begin{array}{lcll}
(x_n, y_m) = (0,0) & \Rightarrow & (x_{n+1}, y_m) = (0,0), & H_{n+1,m} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{1}{c}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0}) \\
 & \Rightarrow & (x_{n+1}, y_m) = (\infty,0), & H_{n+1,m} = (\frac{0}{1}, \frac{0}{0}) \cdot (\frac{0}{0}, \frac{1}{a}) = (\frac{0}{0}, \frac{0}{1}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (0,0), & H_{n,m+1} = (\frac{0}{c}, \frac{0}{1}) \cdot (\frac{0}{1}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (0,\infty), & H_{n,m+1} = (\frac{a}{0}, \frac{1}{0}) \cdot (\frac{0}{1}, \frac{0}{0}) = (\frac{1}{0}, \frac{0}{0}) \\
(x_n, y_m) = (0,\infty) & \Rightarrow & (x_{n+1}, y_m) = (0,\infty), & H_{n+1,m} = (\frac{1}{0}, \frac{0}{0}) \cdot (\frac{1}{c}, \frac{0}{0}) = (\frac{1}{0}, \frac{0}{0}) \\
 & \Rightarrow & (x_{n+1}, y_m) = (\infty,\infty), & H_{n+1,m} = (\frac{1}{0}, \frac{0}{0}) \cdot (\frac{0}{0}, \frac{1}{a}) = (\frac{0}{0}, \frac{1}{0}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (0,0), & H_{n,m+1} = (\frac{0}{1}, \frac{0}{d}) \cdot (\frac{1}{0}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (0,\infty), & H_{n,m+1} = (\frac{1}{0}, \frac{b}{0}) \cdot (\frac{1}{0}, \frac{0}{0}) = (\frac{1}{0}, \frac{0}{0}) \\
(x_n, y_m) = (\infty,0) & \Rightarrow & (x_{n+1}, y_m) = (0,0), & H_{n+1,m} = (\frac{0}{0}, \frac{0}{1}) \cdot (\frac{d}{1}, \frac{0}{0}) = (\frac{0}{1}, \frac{0}{0}) \\
 & \Rightarrow & (x_{n+1}, y_m) = (\infty,0), & H_{n+1,m} = (\frac{0}{0}, \frac{0}{1}) \cdot (\frac{0}{0}, \frac{b}{1}) = (\frac{0}{0}, \frac{0}{1}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (\infty,0), & H_{n,m+1} = (\frac{0}{c}, \frac{0}{1}) \cdot (\frac{0}{0}, \frac{0}{1}) = (\frac{0}{0}, \frac{0}{1}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (\infty,\infty), & H_{n,m+1} = (\frac{a}{0}, \frac{1}{0}) \cdot (\frac{0}{0}, \frac{0}{1}) = (\frac{0}{0}, \frac{1}{0}) \\
(x_n, y_n) = (\infty,\infty) & \Rightarrow & (x_{n+1}, y_m) = (0,\infty), & H_{n+1,m} = (\frac{0}{0}, \frac{1}{0}) \cdot (\frac{d}{1}, \frac{0}{0}) = (\frac{1}{0}, \frac{0}{0}) \\
 & \Rightarrow & (x_{n+1}, y_m) = (\infty,\infty), & H_{n+1,m} = (\frac{0}{0}, \frac{1}{0}) \cdot (\frac{0}{0}, \frac{b}{1}) = (\frac{0}{0}, \frac{1}{0}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (\infty,0), & H_{n,m+1} = (\frac{0}{1}, \frac{0}{d}) \cdot (\frac{0}{0}, \frac{1}{0}) = (\frac{0}{0}, \frac{0}{1}) \\
 & \Rightarrow & (x_n, y_{m+1}) = (\infty,\infty), & H_{n,m+1} = (\frac{1}{0}, \frac{b}{0}) \cdot (\frac{0}{0}, \frac{1}{0}) = (\frac{0}{0}, \frac{1}{0}).
\end{array}
$$

It follows that in all cases $\det(G_{n,m}) = 2^{n+m}\det(G)$. If $n + m \neq n' + m'$, then $G_{n,m} \neq G_{n',m'}$ and the corresponding states of the automaton must be different. Thus the number of states cannot be finite. □

n	$\mathbf{b}(n)$	$\mathbf{p}(n)$
0	0	–
1	1	0
2	10	100
3	11	101
4	100	11000
5	101	11001
6	110	11010
7	111	11011
8	1000	1110000
9	1001	1110001
10	1010	1110010
11	1011	1110011
12	1100	1110100
∞	–	1^ω

u	$\mathbf{c}(u)$	$\mathbf{dc}(u)$	v	$\mathbf{d}(v)$	$\mathbf{cd}(v)$
0	00	0	0	λ	λ
1	01	1	1	λ	λ
2	10	2	00	0	00
3	11	3	01	1	01
00	001	00	10	2	10
01	000	01	11	3	11
10	010	10	000	01	000
11	011	11	001	00	001
22	101	22	010	10	010
23	100	23	011	11	011
32	110	32	100	23	100
33	111	33	101	22	101
000	001	00	110	32	110
001	00100	001	111	33	111

Table 6.3: The binary and prefix codes (left) and the compression and decompression codes (right)

6.6 Binary continued fractions

In modular systems both the unary and binary arithmetical algorithms are computed faster than in nonmodular systems. But modular systems have the disadvantege of slow convergence: their upper contraction quotient is $\mathbf{Q} = 1$ We therefore modify the symmetric system of continued fractions by coding words $0^{a_0}1^{a_1}\cdots$ as sequences of binary representations of the integers a_n. We represent the digits $\overline{0}, \overline{1}$ by numbers $2, 3$, so we work with the alphabet $A = \{0, 1, 2, 3\}$ and the subshift $\Sigma_D = \{0, 1\}^\omega \cup \{2, 3\}^\omega$. For $a \in \{0, 1\}$ we denote by $\overline{a} = 1 - a \in \{0, 1\}$ its complement. For integers $n \in \mathbb{Z}$ and $k > 0$ we denote by $|n|_k = n \bmod k$. Denote by $\mathbf{b} : \mathbb{N} \to \{0, 1\}^+$ the binary code defined by $\mathbf{b}(0) = 0$ and $\mathbf{b}(n) = u \in \{0, 1\}^{k+1}$, where $2^k \le n < 2^{k+1}$ and $n = 2^k u_0 + \cdots + u_k$. Define the prefix code $\mathbf{p} : \{1, 2, \ldots, \infty\} \to \{0, 1\}^+ \cup \{1^\omega\}$ by $\mathbf{p}(\infty) = 1^\omega$ and $\mathbf{p}(n) = 1^k 0u$, where $2^k \le n < 2^{k+1}$, $|u| = k$, and $n = 2^k + 2^{k-1}u_0 + \cdots u_{k-1}$, so $\mathbf{b}(n) = 1u$ (see Table 6.3 left). Define the compression code $\mathbf{c} : \Sigma_D \to \{0, 1\}^\omega$ by

$$\begin{aligned}
\mathbf{c}(0^{a_0}1^{a_1}0^{a_2}\cdots) &= 00\mathbf{p}(a_0)\mathbf{p}(a_1)\mathbf{p}(a_2)\cdots \\
\mathbf{c}(1^{a_0}0^{a_1}1^{a_2}\cdots) &= 01\mathbf{p}(a_0)\mathbf{p}(a_1)\mathbf{p}(a_2)\cdots \\
\mathbf{c}(2^{a_0}3^{a_1}2^{a_2}\cdots) &= 10\mathbf{p}(a_0)\mathbf{p}(a_1)\mathbf{p}(a_2)\cdots \\
\mathbf{c}(3^{a_0}2^{a_1}3^{a_2}\cdots) &= 11\mathbf{p}(a_0)\mathbf{p}(a_1)\mathbf{p}(a_2)\cdots
\end{aligned}$$

Here all a_i are positive. The sequence $\{a_n : n \ge 0\}$ may be finite if its last element is ∞. The compression code is bijective and has an inverse **decompression code** $\mathbf{d} : \{0, 1\}^\omega \to \Sigma_D$. Both codes are continuous in the Cantor topology, so they act also on finite words: For $u \in \Sigma_D$, $\mathbf{c}(u) \in \{0, 1\}^*$ is the longest common prefix of all $\mathbf{c}(v)$ with $v \in [u]$ and the length of $\mathbf{c}(u)$ goes to infinity with $|u| \to \infty$. Similarly, for $u \in \{0, 1\}^*$, $\mathbf{d}(u)$ is the longest common prefix of all $\mathbf{d}(v)$ with $[u]$. Thus $\mathbf{dc}(u)$ is a prefix of u and $\mathbf{cd}(u)$ is a prefix of u (see Table 6.3).

Both codes can be computed by transducers, which are infinite graphs whose labels are pairs u/v of input and output words. The states of the compression transducer are $(s, a, n) \in \{0, 1\}^2 \times \mathbb{N}$ where $s \in \{0, 1\}$ is the sign, $a \in \{0, 1\}$ is the digit and $n \in \mathbb{N}$ counts the number of

digits. The initial state is $\mathbf{i} = (0, 0, 0)$. The transducer accepts on input either single letters, or more generally words of the form a^k, where $a \in A$ The transitions are

$$\begin{aligned}
(0,0,0) &\xrightarrow{a^k/\lfloor \frac{a}{2} \rfloor |a|_2 1^{\lfloor \log_2(k) \rfloor}} (\lfloor \tfrac{a}{2} \rfloor, |a|_2, k),\\
(\lfloor \tfrac{a}{2} \rfloor, |a|_2, n) &\xrightarrow{a^k/1^{\lfloor \log_2(n+k) \rfloor - \lfloor \log_2(n) \rfloor}} (\lfloor \tfrac{a}{2} \rfloor, |a|_2, n+k) \text{ if } n > 0\\
(\lfloor \tfrac{a}{2} \rfloor, \overline{|a|_2}, n) &\xrightarrow{a^k/0\sigma(\mathbf{b}(n))1^{\lfloor \log_2(k) \rfloor}} (\lfloor \tfrac{a}{2} \rfloor, |a|_2, k) \text{ if } n > 0
\end{aligned}$$

For example we have a path

$$(0,0,0) \xrightarrow{2^2/101} (1,0,2) \xrightarrow{2/\lambda} (1,0,3) \xrightarrow{3/01} (1,1,1) \xrightarrow{3^5/11} (1,1,6) \xrightarrow{3/\lambda} (1,1,7)$$

which yields $\mathbf{c}(2^3 3^7) = 1010111$. The inverse decompression code $\mathbf{d} = \mathbf{c}^{-1} : \{0,1\}^\omega \to \Sigma_D$ is computed by a transducer with states $(s, a, b, n) \in \{0.1\}^3 \times (\{-1\} \cup \mathbb{N})$, where s is the sign, a is the digit, n is the count of the letters, $b = 0$ if the count increases and $b = 1$ if the count decreases. The initial state is $\mathbf{j} = (0, 0, 0, -1)$ and the transitions are

$$\begin{aligned}
(0,0,0,-1) &\xrightarrow{s/\lambda} (s,0,0,0)\\
(s,0,0,0) &\xrightarrow{a/2s+a} (s,a,0,1)\\
(s,a,0,n) &\xrightarrow{1/(2s+a)^n} (s,a,0,2n) \text{ if } n > 0\\
(s,a,0,n) &\xrightarrow{0/\lambda} (s,a,1,\tfrac{n}{2}) \text{ if } n > 1\\
(s,a,0,1) &\xrightarrow{0/2s+\overline{a}} (s,\overline{a},0,1)\\
(s,a,1,n) &\xrightarrow{1/(2s+a)^n} (s,a,1,\tfrac{n}{2}) \text{ if } n > 1\\
(s,a,1,n) &\xrightarrow{0/\lambda} (s,a,1,\tfrac{n}{2}) \text{ if } n > 1\\
(s,a,1,1) &\xrightarrow{1/2s+a,2s+\overline{a}} (s,\overline{a},0,1)\\
(s,a,1,1) &\xrightarrow{0/2s+\overline{a}} (s,\overline{a},0,1)
\end{aligned}$$

If we feed the decompression transducer with the word $\mathbf{c}(2^3 3^7) = 1010111$, we get a path

$$\begin{aligned}
&(0,0,0,-1) \xrightarrow{1/\lambda} (1,0,0,0) \xrightarrow{0/2} (1,0,0,1) \xrightarrow{1/2} (1,0,0,2) \xrightarrow{0/\lambda}\\
&(1,0,1,1) \xrightarrow{1/23} (1,1,0,1) \xrightarrow{1/3} (1,1,0,2) \xrightarrow{1/3^2} (1,1,0,4)
\end{aligned}$$

giving $\mathbf{d}(1010111) = 2^3 3^4$ which is a prefix of $2^3 3^7$.

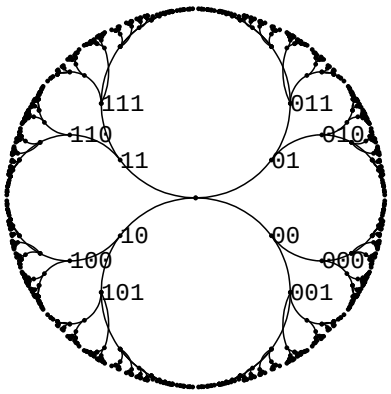
The **binary continued fraction** system (BCF) is defined by the value mapping $\Psi = \Phi \circ \mathbf{c} : \{0,1\}^\omega \to \overline{\mathbb{R}}$, where $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is the value mapping of the CF system. We define the length quotients similarly as in Möbius number systems with Φ replaced by Ψ. Moreover we define the mean length quotient by

$$L_n = 2^{-n} \sum_{u \in \{0,1\}^n} |\Psi[u]|.$$

Some values of these quotients are in Figure 6.4.

We now modify the general binary algorithm for the BCF system. Recall that we have transformations with matrices $F_0 = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$, $F_1 = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $F_2 = \left[\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right]$, $F_3 = \left[\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right]$. The graph has vertices $B = \{\lambda, 0, 1\}$ and the intervals have matrices $V_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $V_1 = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. For the matrices $H_a = V^{-1}_{\lfloor \frac{a}{2} \rfloor} F_a V_{\lfloor \frac{a}{2} \rfloor}$ we get

$$H_0 = H_3 = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right], \; H_1 = H_2 = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right].$$



n	$\sqrt[n]{\underline{L}_n}$	$\sqrt[n]{L_n}$	$\sqrt[n]{\overline{L}_n}$
1	0.500	0.500	0.500
2	0.500	0.500	0.500
3	0.468	0.500	0.528
4	0.446	0.535	0.620
5	0.433	0.527	0.600
6	0.424	0.540	0.654
7	0.417	0.533	0.630
8	0.413	0.540	0.668
9	0.409	0.535	0.647

Figure 6.4: The binary continued fractions (left) and the lengths of its cylinders (right)

For $u \in \{0,1\}^n$ we set $H_u = H_{u_0} \circ \cdots \circ H_{u_{n-1}}$. The states of the modified binary graph are (X, p, q, r), where $X \in \mathbb{T}(\mathbb{R})$, $p, q \in \{0,1\}^3 \times \{-1, 0, 1, 2, \ldots\}$ are states of the decompression transducer, and $r \in \{0,1\}^2 \times \mathbb{N}$ is a state of the compression transducer. The initial state is $(T, \mathbf{j}, \mathbf{j}, \mathbf{i})$ where $\mathbf{j} = (0, 0, 0, -1)$ is the initial state of the decompression transducer and $\mathbf{i} = (0, 0, 0)$ is the initial state of the compression transducer. The transitions are

$$
\begin{aligned}
(X,p,q,r) &\xrightarrow{a,\lambda,\lambda} (X^*F_b, p', q, r) \text{ if } p_3 < 0, p \xrightarrow{a/b} p' \\
(X,p,q,r) &\xrightarrow{\lambda,a,\lambda} (X_*F_b, p, q', r) \text{ if } q_3 < 0, q \xrightarrow{a/b} q' \\
(X,p,q,r) &\xrightarrow{a,\lambda,\lambda} (X^*H_b, p', q, r) \text{ if } p_3 \geq 0, p \xrightarrow{a/b} p' \\
(X,p,q,r) &\xrightarrow{\lambda,a,\lambda} (X_*H_b, p, q', r) \text{ if } q_3 \geq 0, q \xrightarrow{a/b} q' \\
(X,p,q,r) &\xrightarrow{\lambda,\lambda,b} (F_a^{-k}X, p, q, r') \text{ if } p_3, q_3 \geq 0, \\
&\qquad\qquad X \subseteq F_a^k V_{\lfloor \frac{a}{2} \rfloor},\ r \xrightarrow{a^k/b} r'
\end{aligned}
$$

If $(T, \mathbf{j}, \mathbf{j}, \mathbf{i}) \xrightarrow{u,v,w}$ is an infinite path with infinite $u, v, w \in \{0,1\}^\omega$, then $\Psi(w) = T(\Psi(u), \Psi(v))$. We consider a selector $s : \mathbb{T}(\mathbb{R}) \to A^+ \cup \{\text{'x'}, \text{'y'}, \text{'xy'}\}$ which depends only on the state $X \in \mathbb{T}(\mathbb{R})$. It searches a letter $a \in A$ such that $X \subseteq F_a V_{\lfloor \frac{a}{2} \rfloor}$. If it finds such a letter, it finds the maximum k such that $X \subseteq F_a^k V_{\lfloor \frac{a}{2} \rfloor}$ and outputs a^k. This can be done in $\lfloor \log_2 k \rfloor$ steps. If no such letter exists then it finds one of the absorptions similarly as the selector of the general binary algorithm (see Kůrka [41]).

Chapter 7

Algebraic number fields

Arithmetical algorithms considered in Chapter 5 are based on the arithmetical operations with the entries of the matrices of the number systemin question. If these entries are not integers or rationals, we need arithmetical algorithms which work with them. Such algorithms exist for algebraic numbers. Algebraic numbers can be represented by vectors of rational numbers, and arithmetical operations with them are based on matrix calculus.

7.1 Polynomials with rational coefficients

Recall that a polynomial is a complex function $p(x) = \sum_{i\geq 0} p_i x^i$, where p_i are complex numbers and $p_i = 0$ for $i > \deg(p)$. Polynomials can be added, subtracted and multiplied and they form the ring $\mathbb{C}[x]$. The subring of polynomials with real coefficients is denoted by $\mathbb{R}[x]$. Similarly we denote by $\mathbb{Q}[x]$ the ring of polynomials with rational coefficients and by $\mathbb{Z}[x]$ the ring of polynomials with integer coefficients. The **content** of a polynomial $p \in \mathbb{Z}[x]$ is the greatest common divisor of its coefficients. A polynomial of $\mathbb{Z}[x]$ is **primitive**, if its content is 1. It is **irreducible** if it cannot be written as a product of two polynomials of $\mathbb{Z}[x]$ of positive degree. Similarly, a polynomial of $\mathbb{Q}[x]$ is **irreducible** if it cannot be written as a product of two polynomials of $\mathbb{Q}[x]$ of positive degree.

Proposition 7.1 (Gauss) *The product of two primitive polynomials of $\mathbb{Z}[x]$ is primitive.*

Proof: Let $p(x) = \sum_i p_i x^i$, $q(x) = \sum_i q_i x^i$, $r(x) = p(x)q(x) = \sum_i r_i x^i$, and assume that $k > 1$ is a prime number which divides the content of r. Let p_n be the first coefficient of p not divisible by k and let q_m be the first coefficient of q not divisible by k. In the sum

$$r_{n+m} = p_n q_m + p_{n-1} q_{m+1} + p_{n+1} q_{m-1} + \cdots$$

every term except the first is divisible by k. Since $k|r_{n+m}$ we get $k|p_n q_m$, so k divides either p_n or q_m which is a contradiction. □

Proposition 7.2 *A polynomial $r \in \mathbb{Z}[x]$ is irreducible in $\mathbb{Z}[x]$ iff it is irreducible in $\mathbb{Q}[x]$.*

Proof: If r is irreducible in $\mathbb{Q}[x]$ then it is clearly irreducible in $\mathbb{Z}[x]$. Conversely assume that r is reducible in $\mathbb{Q}[x]$, so $r = pq$ with $p, q \in \mathbb{Q}[x]$ of positive degree. Then we can write $p = \frac{1}{a}s$, $q = \frac{1}{b}t$, where a, b are positive integers and $s, t \in \mathbb{Z}[x]$ are primitive polynomials. It follows that $abr = st$ is a primitive polynomial, so $ab = 1$ and r is reducible in $\mathbb{Z}[x]$. □

Proposition 7.3 *An irreducible polynomial $p \in \mathbb{Z}[x]$ does not have multiple roots.*

Proof: If $p(x)$ is in $\mathbb{C}[x]$ divisible by $(x-a)^2$, where $a \in \mathbb{C}$, then $(x-a)$ divides both p and $p' \in \mathbb{Z}[x]$, so $r = \gcd(p, p')$ has positive degree. The Euclidean algorithm which computes gcd uses only field operations, so $r \in \mathbb{Q}[x]$. Thus p is reducible in $\mathbb{Q}[x]$ and this is a contradiction. □

The irreducibility of a polynomial $p \in \mathbb{Z}[x]$ can be tested algorithmically. We use the fact that a polynomial of degree n is uniquely determined by its value at $n+1$ distinct points. If $c_0, c_1, \ldots, c_n$ are distinct complex numbers and $a_0, \ldots, a_n$ are arbitrary complex numbers, the unique polynomial of degree n which satisfies $p(c_i) = a_i$ is given by the formula

$$p(x) = \sum_{i=0}^{n} \frac{a_i(x-c_0)\cdots(x-c_{i-1})(x-c_{i+1})\cdots(x-c_n)}{(c_i-c_0)\cdots(c_i-c_{i-1})(c_i-c_{i+1})\cdots(c_i-c_n)}.$$

Let $p \in \mathbb{Z}[x]$ and $\deg(p) = n$. If p is reducible, then it has a factor r of degree at most $\lfloor \frac{n}{2} \rfloor$. To test whether p has a factor of degree $m \leq \lfloor \frac{n}{2} \rfloor$, take $m+1$ distinct integers $c_0, \ldots, c_m$. We have $r(c_i) | p(c_i)$ and there is only a finite number of integers which divide $p(c_i)$. Thus for each sequence of integers b_i which divide $p(c_i)$ we take the polynomial r which satisfies $r(c_i) = b_i$ and test whether r divides p.

Definition 7.4 *We say that $\alpha \in \mathbb{C}$ is an algebraic number if there exists a polynomial $p \in \mathbb{Q}[x]$ such that $p(\alpha) = 0$. The* **degree** *of α is the smallest integer d such that there exists a polynomial $p \in \mathbb{Q}[x]$ with $p(\alpha) = 0$ and $\deg(p) = d$.*

Proposition 7.5 *For an algebraic number α there exists a unique monic irreducible polynomial $p \in \mathbb{Q}[x]$ with $p(\alpha) = 0$. We say that p is the* **minimal polynomial** *of α. If $q \in \mathbb{Q}[x]$ and $q(\alpha) = 0$ then p divides q.*

Proof: Let $p \in \mathbb{Q}[x]$ be a monic polynomial of smallest degree which satisfies $p(\alpha) = 0$. Then p is irreducible, since otherwise α would be a root of one of its factors. If $q(\alpha) = 0$, then $x - \alpha$ divides in $\mathbb{C}[x]$ both p and q, so $r = \gcd(p, q)$ has a positive degree and $r(\alpha) = 0$. Thus $\deg(r) = \deg(p)$ and therefore p divides q. □

7.2 Extension fields

Assume that $K \subseteq \mathbb{C}$ is a subfield of the field of the complex numbers. This means that if $x, y \in K$ then the sum $x + y$, difference $x - y$, and product xy belong to K, and if $y \neq 0$ then also $x/y \in K$. The smallest subfield of $\mathbb{C}$ is the field $\mathbb{Q}$ of rational numbers. Each subfield K of C contains $\mathbb{Q}$ as a subfield. If $\mathbb{Q} \subseteq K \subset L \subseteq \mathbb{C}$ are two subfields, then L is a vector space over K: If $u_i \in L$ and $a_i \in K$ then $\sum_i a_i u_i \in L$. If L as a K-vector space has a finite dimension n, we say that L is a **finite field extension** of K and write $n = [L : K]$. In particular we say that $K \subseteq \mathbb{C}$ is an **algebraic number field** if it has finite dimension over $\mathbb{Q}$. If K is an algebraic number field of dimension n, then each $\alpha \in K$ is an algebraic number of degree at most n. Indeed, the numbers $1, \alpha, \ldots, \alpha^n$ are linearly $\mathbb{Q}$-dependent, so there exist $p_i \in \mathbb{Q}$ with $\sum_{i \leq n} p_i \alpha^i = 0$. Given a field $K \subseteq \mathbb{C}$ and a set $M \subseteq \mathbb{C}$ then the intersection of all subfields of $\mathbb{C}$ which contain $M \cup K$ as a subset is the extension field $K(M)$ of K **generated** by M. In particular a **simple field extension** of K is by definition an extension $K(\alpha)$ by a single $\alpha \in \mathbb{C} \setminus K$. See e.g., Ireland and Rosen [26] for an introduction to the theory of extension fields.

Proposition 7.6 *If $\mathbb{Q} \subseteq K \subset L \subset M \subseteq \mathbb{C}$ are subfields of $\mathbb{C}$ and if L is finite field extension of K and M is a finite field extension of L, then M is a finite field extension of K and $[M : K] = [M : L] \cdot [L : K]$.*

Proof: Let $\{u_0, \ldots, u_{n-1}\}$ be a basis of L over K and let $\{v_0, \ldots v_{m-1}\}$ be a basis of M over L. Then $\{u_i v_j : \ i < n, j < m\}$ is a basis of M over K. □

For example $\mathbb{Q}(\sqrt{2}) = \{x_0 + x_1\sqrt{2} : \ x_i \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt{3}) = \{x_0 + x_1\sqrt{3} : \ x_i \in \mathbb{Q}\}$ are algebraic fields of dimension 2 and $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \{x_0 + x_1\sqrt{2} + x_2\sqrt{3} + x_3\sqrt{6} : \ x_i \in \mathbb{Q}\}$ is an algebraic field of dimension 4. We show that $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is a simple field extension. For $\alpha = \sqrt{2} + \sqrt{3}$ we get $\alpha^2 = 5 + 2\sqrt{6}$ so it is a root of an irreducible monic polynomial $p(x) = x^4 - 10x^2 + 1$. Since $\alpha^3 = 11\sqrt{2} + 9\sqrt{3}$, we get $\sqrt{2} = \frac{\alpha^3 - 9\alpha}{2}$, $\sqrt{3} = \frac{11\alpha - \alpha^3}{2}$, so $\mathbb{Q}(\sqrt{2} + \sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$.

Proposition 7.7 *If $L \subseteq \mathbb{C}$ is a finite field extension of $K \subset L$, then it is a simple field extension of K. In particular, every algebraic number field is a simple field extension of $\mathbb{Q}$.*

Proof: we show first that for each $\alpha, \beta \in \mathbb{C} \setminus K$ there exists $\gamma \in \mathbb{C}$ such that $K(\gamma) = K(\alpha, \beta)$. Let p, q be the minimal polynomials of α and β of degree n and m. Denote by $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ the distinct roots of p and by $\beta = \beta_0, \ldots, \beta_{m-1}$ the distinct roots of q in $\mathbb{C}$. There exists $c \in \mathbb{Q}$ such that $\alpha + c\beta \neq \alpha_i + c\beta_j$ for all $0 < i < n$, $0 < j < m$. Set $\gamma = \alpha + c\beta$, $r(x) = p(\gamma - cx)$. Then β is the only common root of r and q. Indeed if β_j is a root of r then $\gamma - c\beta_j \neq \alpha_i$, so $p(\gamma - c\beta_j) \neq 0$. It follows that $\gcd(r, q) = x - \beta$. The coefficients of both r and q are in the field $K(\gamma)$ and the GCD is computed using only the field operations of $K(\gamma)$. It follows $\beta \in K(\gamma)$ and therefore $\alpha = \gamma - c\beta \in K(\gamma)$ as well. Thus $K(\gamma) = K(\alpha, \beta)$. To prove that L is a simple field extension of K, take any $\alpha_0 \in L \setminus K$. Then $K(\alpha_0) \subseteq L$. If $K(\alpha_0) \neq L$, take any $\alpha_1 \in L \setminus K(\alpha_0)$, and so on. Since the dimensions of these fields increase, after a finite number of steps we get $L = K(\alpha_0, \ldots, \alpha_p)$, so there exists $\alpha \in L \setminus K$ with $L = K(\alpha)$. □

Example 7.8 *Each algebraic number field of dimension 2 is of the form*

$$\mathbb{Q}(\sqrt{d}) = \{x_0 + x_1\sqrt{d} : \ x_0, x_1 \in \mathbb{Q}\},$$

where d is a **squarefree** *integer. This means that d is not divisible by any r^2 with $r > 1$. The arithmetical operation in $\mathbb{Q}(\sqrt{d})$ are given by*

$$\begin{aligned}
(x_0 + x_1\sqrt{d}) \pm (y_0 + y_1\sqrt{d}) &= (x_0 \pm y_0 + (x_1 \pm y_1)\sqrt{d}) \\
(x_0 + x_1\sqrt{d})(y_0 + y_1\sqrt{d}) &= (x_0y_0 + x_1y_1d + (x_0y_1 + x_1y_0)\sqrt{d}) \\
\frac{1}{x_0 + x_1\sqrt{d}} &= \frac{x_0 - x_1\sqrt{d}}{x_0^2 - x_1^2 d}
\end{aligned}$$

Proof: If $[K : \mathbb{Q}] = 2$ and $\alpha \in K \setminus \mathbb{Q}$ then there exist rational a, b with $\alpha^2 + a\alpha + b = 0$, so $\alpha = \frac{1}{2}(-a \pm \sqrt{a^2 - 4b})$, and $K = \mathbb{Q}(\sqrt{d})$ with $d = a^2 - 4b$. If $d = r^2c$, then $K = \mathbb{Q}(\sqrt{c})$. □

We have seen that each element of $\mathbb{Q}(\sqrt{d})$ is given by a pair $x = (x_0, x_1)$ of rational numbers, so $\mathbb{Q}(\sqrt{d})$ is isomorphic to $\mathbb{Q}^2$.

Proposition 7.9 *Let α be an algebraic number of degree $n > 1$. Then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = n$, $\mathbb{Q}(\alpha)$ over $\mathbb{Q}$ has the* **power basis** *$\{1, \alpha, \ldots, \alpha^{n-1}\}$ and $\mathbb{Q}(\alpha) = \{q(\alpha) : \ q \in \mathbb{Q}_n[x]\}$, where $\mathbb{Q}_n[x] = \{q \in \mathbb{Q}[x] : \ \deg(q) < n\}$.*

Proof: Clearly $\alpha \in \{q(\alpha) : \ q \in \mathbb{Q}_n[x]\} \subseteq \mathbb{Q}(\alpha)$. We show that $\{q(\alpha) : \ q \in \mathbb{Q}_n[x]\}$ is a field. Let $p(x) = -p_0 - p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$ be the minimal polynomial of α. If $q, r \in \mathbb{Q}_n[x]$ then $q+r, q-r \in \mathbb{Q}_n[x]$, so $\{q(\alpha) : \ q \in \mathbb{Q}_n[x]\}$ is closed with respect to addition and subtraction. For the product we have $qr \in \mathbb{Q}_{2n-1}[x]$. Using successively the identity $\alpha^n = p_0 + p_1\alpha + \cdots + p_{n-1}\alpha^{n-1}$ we reduce $(qr)(\alpha)$ to an expression which does not contain any power of α higher than $n-1$. Thus there exists $s \in \mathbb{Q}_n[x]$ such that $(qr)(\alpha) = s(\alpha)$. If $q \in \mathbb{Q}_n[x]$ is a nonzero polynomial, then $\gcd(p, q) = 1$. By Proposition 5.28 there exist $s, t \in \mathbb{Q}[x]$ such that $ps + qt = 1$. Thus $1 = p(\alpha)s(\alpha) + q(\alpha)t(\alpha) = q(\alpha)t(\alpha)$. Using the identity $\alpha^n = p_0 + p_1\alpha + \cdots + p_{n-1}\alpha^{n-1}$ we find a polynomial $r \in \mathbb{Q}_n[x]$ such that $r(\alpha) = t(\alpha)$, so $r(\alpha) = 1/q(\alpha)$. Thus $\{q(\alpha) : \ q \in \mathbb{Q}_n[x]\}$ is the smallest field which contains α and coincides with $\mathbb{Q}(\alpha)$. As a vector space over $\mathbb{Q}$, $\mathbb{Q}(\alpha)$ is generated by 1, α, α^2, ..., α^{n-1}. Since these numbers are $\mathbb{Q}$-independent, they form a basis, so $[\mathbb{Q}(\alpha) : \mathbb{Q}] = n$. □

Let α be an algebraic number with minimal polynomial $p(x) = -p_0 - p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$ of degree n. A polynomial $q \in \mathbb{Q}[x]$ with degree $\deg(q) < n$ is determined by the vector of its coefficients so $\mathbb{Q}(\alpha)$ is isomorphic with $\mathbb{Q}^n$. A row vector or a $(1 \times n)$-matrix $x = [x_0, \ldots, x_{n-1}] \in \mathbb{Q}^n$ represents the number $\beta = \sum_{i<n} x_i\alpha^i = x \cdot w \in \mathbb{Q}(\alpha)$, where $w = [1, \alpha, \ldots, \alpha^{n-1}]^T$ is a column vector or a $(n \times 1)$-matrix. Thus the isomorphism from $\mathbb{Q}^n$ to $\mathbb{Q}(\alpha)$ is given by $x \mapsto x \cdot w$. The addition and subtraction in $\mathbb{Q}(\alpha)$ corresponds to the addition and subtraction in $\mathbb{Q}^n$ and multiplication by $a \in \mathbb{Q}$ in $\mathbb{Q}(\alpha)$ corresponds to multiplication by a in $\mathbb{Q}^n$: For $a \in \mathbb{Q}$, $x, y \in \mathbb{Q}^n$ we have $(x \pm y) \cdot w = x \cdot w \pm y \cdot w$, $(ax) \cdot w = a(x \cdot w)$. The product $xy \in \mathbb{Q}^n$ which corresponds to the product in $\mathbb{Q}(\alpha)$ is defined by $(xy) \cdot w = (x \cdot w)(y \cdot w)$. To obtain the product xy we first multiply x and y as polynomials so we get a vector in $\mathbb{Q}^{2n-1}$. This polynomial product can be obtained by matrix multiplication $x \cdot B(y)$ where $B(y)$ is an $(n \times (2n-1))$-matrix given by

$$B(y)_{ij} = \begin{cases} y_{j-i} & \text{if} \quad 0 \le j - i < n \\ 0 & \text{otherwise} \end{cases}$$

For $n = 3$ we have

$$B(y) = \begin{bmatrix} y_0 & y_1 & y_2 & 0 & 0 \\ 0 & y_0 & y_1 & y_2 & 0 \\ 0 & 0 & y_0 & y_1 & y_2 \end{bmatrix}$$

so $x \cdot B(y) = [x_0y_0, x_0y_1 + x_1y_0, x_0y_2 + x_1y_1 + x_2y_1, x_1y_2 + x_2y_1, x_2y_2]$. Then we reduce the polynomial $x \cdot B(y)$ to degree at most $n-1$ using repeatedly the identity $\alpha^n = p_0 + p_1\alpha + \cdots + p_{n-1}\alpha^{n-1}$. This reduction is also represented by matrix multiplication. If $m \ge n$, then the reduction of $z \in \mathbb{Q}^m$ to $w \in \mathbb{Q}^{m-1}$ is given by

$$w = z_0 + \cdots + z_{m-1}\alpha^{m-1} + z_m\alpha^{m-n}(p_0 + p_1\alpha + \cdots + p_{n-1}\alpha^{n-1}),$$

so $w = z \cdot P(m)$ where $P(m)$ is an $((m+1) \times m)$-matrix given by

$$P(m)_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ p_{j-m+n} & \text{if} \quad i = m, 0 \le j - m + n < n \\ 0 & \text{otherwise} \end{cases}$$

The reduction of a vector $z \in \mathbb{Q}^{2n-1}$ to a vector of $\mathbb{Q}^n$ is then represented by the $((2n-1) \times n)$-matrix $P = P(2n-2) \cdots P(n+1) \cdot P(n)$. For example for $n = 3$ we get

$$P = P(4) \cdot P(3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & p_0 & p_1 & p_2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ p_0 & p_1 & p_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ p_0 & p_1 & p_2 \\ p_0 p_2 & p_0 + p_1 p_2 & p_1 + p_2^2 \end{bmatrix}$$

Thus the multiplication in $\mathbb{Q}^n$ is given by $xy = x \cdot B(y) \cdot P = y \cdot B(x) \cdot P$. The division is given by $\frac{x}{y} = x \cdot (B(y) \cdot P)^{-1}$. Here we write the matrix multiplication with dot $\cdot$. For an algebraic number field $\mathbb{Q}(\sqrt{d})$ of dimension 2 we have

$$\begin{aligned} B(y) \cdot P &= \begin{bmatrix} y_0 & y_1 & 0 \\ 0 & y_0 & y_1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ d & 0 \end{bmatrix} = \begin{bmatrix} y_0 & y_1 \\ d y_1 & y_0 \end{bmatrix} \\ \begin{bmatrix} x_0 & x_1 \end{bmatrix} \cdot B(y) \cdot P &= \begin{bmatrix} x_0 y_0 + d x_1 y_1 & x_0 y_1 + x_1 y_0 \end{bmatrix} \end{aligned}$$

Besides this vector representation of algebraic numbers there is also a matrix representation. It uses an $(n \times n)$-rational matrix to represent an algebraic number of degree n and the multiplication of algebraic numbers corresponds to the multiplication of matrices. Let $K \subseteq \mathbb{C}$ be an algebraic number field of dimension n. Given $\beta \in K$, the multiplication $x \mapsto \beta x$ is a linear mapping on the vector space K over $\mathbb{Q}$ so it is represented by an $(n \times n)$-matrix. Let $w = [w_0, \ldots, w_{n-1}]^T$ be a basis of K over $\mathbb{Q}$ conceived as a column vector or a $(n \times 1)$-matrix. Then for each $\beta \in K$ there exists an $(n \times n)$-matrix $M_w(\beta)$ such that $\beta w_i = \sum_{j<k} M_w(\beta)_{ij} w_j$, or $\beta w = M_w(\beta) \cdot w$. This means that β is an eigenvalue of $M_w(\beta)$ and w is the corresponding right eigenvector.

Proposition 7.10 *Let $w \in K^n$ be a basis of an algebraic number field K, let $M_w(\beta)$ be the matrix representation of $\beta \in K$ such that $\beta w = M_w(\beta) \cdot w$. Then for each $\beta, \gamma \in K$, $a \in \mathbb{Q}$ we have*

1. *$M_w(1)$ is the identity matrix.*
2. *$M_w(\beta \pm \gamma) = M_w(\beta) \pm M_w(\gamma)$,*
3. *$M_w(a\beta) = a M_w(\beta)$,*
4. *$M_w(\beta\gamma) = M_w(\beta) \cdot M_w(\gamma)$,*
5. *$M_w(1/\beta) = M_w(\beta)^{-1}$ provided $\beta \neq 0$.*

Proof: Multiplication by 1 is the identity mapping. Since $(\beta \pm \gamma)w = (M_w(\beta) \pm M_w(\gamma)) \cdot w$, we get $M_w(\beta \pm \gamma) = M_w(\beta) \pm M_w(\gamma)$. Since

$$(\beta\gamma)w = \gamma(\beta w) = \gamma M_w(\beta) \cdot w = M_w(\beta) \cdot (\gamma w) = M_w(\beta) \cdot M_w(\gamma) \cdot w,$$

we get $M_w(\beta\gamma) = M_w(\beta) \cdot M_w(\gamma)$. Since $M_w(\beta) \cdot M_w(1/\beta)$ is the identity matrix, $M_w(1/\beta) = M_w(\beta)^{-1}$. □

If $K = \mathbb{Q}(\alpha)$ we denote by $M_\alpha(\beta)$ the matrix representation of $\beta \in K$ by the power basis $w = [1, \alpha, \ldots, \alpha^{n-1}]^T$. By Proposition 7.10 we get

$$M_\alpha \left(\sum_{i<n} y_i \alpha^i \right) = \sum_{i<n} y_i M_\alpha(\alpha)^i$$

so each $M_\alpha(\beta)$ can be obtained from $M_\alpha(\alpha)$. Let $p(x) = -p_0 = p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$ be the minimal polynomial of α. Since $\alpha\alpha^i = \alpha^{i+1}$ and $\alpha\alpha^{n-1} = \sum_{i<n} p_i\alpha^i$ we get

$$M_\alpha(\alpha)_{ij} = \begin{cases} 1 & \text{if } i < n-1, j = i+1 \\ p_j & \text{if } i = n-1 \\ 0 & \text{otherwise} \end{cases}$$

Indeed for $w = [1, \alpha, \ldots, \alpha^{n-1}]^T$ we have

$$\alpha w = \begin{bmatrix} \alpha \\ \alpha^2 \\ \vdots \\ \alpha^{n-1} \\ \alpha^n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ p_0 & p_1 & p_2 & \cdots & p_{n-1} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \alpha \\ \vdots \\ \alpha^{n-2} \\ \alpha^{n-1} \end{bmatrix} = M_\alpha(\alpha) \cdot w$$

For an algebraic number field $\mathbb{Q}(\sqrt{d})$ of dimension 2 we have $p_0 = d$, $p_1 = 0$, so

$$M_{\sqrt{d}}(\sqrt{d}) = \begin{bmatrix} 0 & 1 \\ d & 0 \end{bmatrix}$$

$$M_{\sqrt{d}}(x_0 + x_1\sqrt{d}) = \begin{bmatrix} x_0 & x_1 \\ dx_1 & x_0 \end{bmatrix}$$

Indeed

$$\begin{bmatrix} x_0 & x_1 \\ dx_1 & x_0 \end{bmatrix} \cdot \begin{bmatrix} y_0 & y_1 \\ dy_1 & y_0 \end{bmatrix} = x_0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + x_1 \begin{bmatrix} 0 & 1 \\ d & 0 \end{bmatrix} = \begin{bmatrix} x_0y_0 + dx_1y_1 & x_0y_1 + x_1y_0 \\ d(x_0y_1 + x_1y_0) & x_0y_0 + dx_1y_1 \end{bmatrix}$$

While the matrices $M_w(\beta)$ depend on the choice of the basis w, their trace and determinant depend only on β.

Proposition 7.11 *Let K be an algebraic number field. Then the* **trace** $T_K(\beta) = \mathrm{tr}(M_w(\beta))$, *and* **norm** $N_K(\beta) = \det(M_w(\beta))$ *of $\beta \in K$ do not depend on the basis w of K.*

Proof: Let $w = [w_0, \ldots, w_{n-1}]^T$, $v = [v_0, \ldots, v_{n-1}]^T$ be two bases of K. There exists a regular matrix $A \in \mathbb{Q}^{n\times n}$ such that $v_i = \sum_{j<n} A_{ij}w_j$ or $v = A \cdot w$, $w = A^{-1} \cdot v$, so

$$\beta v = A \cdot \beta w = A \cdot M_w(\beta) \cdot w = A \cdot M_w(\beta) \cdot A^{-1} \cdot v.$$

Thus $M_v(\beta) = A \cdot M_w(\beta) \cdot A^{-1}$ and $\mathrm{tr}(M_v(\beta)) = \mathrm{tr}(M_w(\beta))$, $\det(M_v(\beta)) = \det(M_w(\beta))$. □

For an algebraic α we write $N_\alpha(\beta) = N_{\mathbb{Q}(\alpha)}(\beta)$, $T_\alpha(\beta) = T_{\mathbb{Q}(\alpha)}(\beta)$. For example for a positive squarefree number $d \in \mathbb{N}$ we get

$$T_{\sqrt{d}}(x_0 + x_1\sqrt{d}) = 2x_0,$$

$$N_{\sqrt{d}}(x_0 + x_1\sqrt{d}) = x_0^2 - dx_1^2.$$

If $p(x) = -p_0 = p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$ is the minimal polynomial of α, then from $M_\alpha(\alpha)$ we get $T_\alpha(\alpha) = p_{n-1}$, $N_\alpha(\alpha) = (-1)^{n+1}p_0$.

Proposition 7.12 *Let K be an algebraic number field of dimension $n > 1$. For $x, y \in K$, $a \in \mathbb{Q}$ we have $T_K(ax) = aT_K(x)$, $T_K(x+y) = T_K(x) + T_K(y)$, $N_K(ax) = a^n N_K(x)$, $N_K(xy) = N_K(x) \cdot N_K(y)$.*

Proof: For each basis w of K we have $M_w(ax) = aM_w(x)$, $M_w(x+y) = M_w(x) + M_w(y)$, $M_w(xy) = M_w(x) \cdot M_w(y)$. □

A **field embedding** $\sigma : K \to \mathbb{C}$ of a field $K \subseteq \mathbb{C}$ is an injective mapping which preserves the field operations, so $\sigma(x \pm y) = \sigma(x) \pm \sigma(y)$, $\sigma(xy) = \sigma(x)\sigma(y)$, and $\sigma(x/y) = \sigma(x)/\sigma(y)$ provided $y \neq 0$. Any field embedding σ fixes $\mathbb{Q}$: for $a \in \mathbb{Q}$ we have $\sigma(a) = a$. It follows that for any polynomial $q \in \mathbb{Q}[x]$ and any $\beta \in K$ we have $\sigma(q(\beta)) = q(\sigma(\beta))$. Let α be an algebraic number with minimal polynomial $p(x) = -p_0 - p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$. If $\sigma : \mathbb{Q}(\alpha) \to \mathbb{C}$ is a field embedding, then $p(\sigma(\alpha)) = \sigma(p(\alpha)) = 0$. Thus $\sigma(\alpha)$ is a root of p and σ is uniquely determined by $\sigma(\alpha)$. Since there are n distinct roots $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in \mathbb{C}$ of p, there are n field embeddings $\sigma_0, \ldots, \sigma_{n-1} : \mathbb{Q}(\alpha) \to \mathbb{C}$ defined by $\sigma_i(q(\alpha)) = q(\alpha_i)$, and σ_0 is the identity. We say that α_i are **conjugated** to $\alpha = \sigma_0(\alpha)$. Since $p(x) = \prod_{i=0}^{n-1}(x - \alpha_i)$, we get

$$\begin{aligned}
\sum_i \alpha_i &= p_{n-1} \\
\sum_{i<j} \alpha_i\alpha_j &= -p_{n-2} \\
\sum_{i<j<k} \alpha_i\alpha_j\alpha_k &= p_{n-3} \\
&\vdots \\
\alpha_0 \cdots \alpha_{n-1}(\tfrac{1}{\alpha_0} + \cdots \tfrac{1}{\alpha_{n-1}}) &= (-1)^n p_1 \\
\alpha_0 \cdots \alpha_{n-1} &= (-1)^{n+1} p_0
\end{aligned}$$

From the formula for $M_\alpha(\alpha)$ we get

Proposition 7.13 *If α is an algebraic number of degree n with minimal polynomial $p(x) = -p_0 - p_1 x - \cdots - p_{n-1}x^{n-1} + x^n$ and conjugated roots $\alpha = \alpha_0, \ldots, \alpha_{n-1}$ then*

$$T_\alpha(\alpha) = p_{n-1} = \sum_{i<n} \alpha_i$$

$$N_\alpha(\alpha) = (-1)^{n+1} p_0 = \prod_{i<n} \alpha_i$$

For example for $\alpha = \alpha_0 = \sqrt{d}$ we have $\alpha_1 = -\sqrt{d}$ and $\sigma_1(x_0 + x_1\sqrt{d}) = x_0 - x_1\sqrt{d} \in \mathbb{Q}(\sqrt{d})$, so both $\sigma_0, \sigma_1 : \mathbb{Q}(\sqrt{d}) \to \mathbb{Q}(\sqrt{d})$ are automorphisms. For $\alpha = \sqrt[3]{2}$ we have $\alpha_1 = \frac{\sqrt[3]{2}(-1+i\sqrt{3})}{2}$, $\alpha_2 = \frac{\sqrt[3]{2}(-1-i\sqrt{3})}{2}$, so $\sigma_1(\mathbb{Q}(\alpha)) \neq \mathbb{Q}(\alpha)$. If α is algebraic and $\beta \in \mathbb{Q}(\alpha)$, then β is algebraic too, since $1, \beta, \ldots, \beta^n$ are linearly dependent over $\mathbb{Q}$. The degree m of the minimal polynomial Q of β divides n, since $\mathbb{Q}(\beta) \subseteq \mathbb{Q}(\alpha)$.

Proposition 7.14 *Let $\mathbb{Q} \subseteq K \subset L \subset \mathbb{C}$ be algebraic number fields of dimensions $[K : \mathbb{Q}] = m$, $[L : K] = k$, $[L : \mathbb{Q}] = n = mk$. Then for each embedding $\sigma : K \to \mathbb{C}$ there exist exactly k embeddings $\sigma_0, \ldots, \sigma_{k-1} : L \to \mathbb{C}$ which extend σ, i.e., $\sigma_i(x) = \sigma(x)$ for $x \in K$.*

Proof: Let $\sigma : K \to \mathbb{C}$ be a field embedding. There exists $\alpha \in L \setminus K$ such that $L = K(\alpha)$. The minimal polynomial $p(x) = \sum_{i<k} p_i x^i$ of α over K has degree k and coefficients $p_i \in K$. If $c \in \mathbb{C}$ is a root of p, then

$$\sum_{i<k} \sigma(p_i)\sigma(c)^i = \sigma\left(\sum_{i<k} p_i c^i\right) = \sigma(0) = 0,$$

so $\sigma(c)$ is a root of $q(x) = \sum_{i<k} \sigma(p_i)x^i$. It follows that $q(x)$ has k distinct roots. If $\tau : L \to \mathbb{C}$ is an embedding which extends σ, then

$$0 = \tau(p(\alpha) = \sum_{i<k} \tau(p_i)\tau(\alpha)^i = \sum_{i<k} \sigma(p_i)\tau(\alpha)^i = q(\tau(\alpha))$$

so $\tau(\alpha)$ is a root of $q(x)$ and τ is uniquely determined by $\tau(\alpha)$. Since q has k roots, there exists exactly k embeddings τ of σ. □

Proposition 7.15 *Let K be an algebraic number field of dimension $n > 1$ and let $\sigma_i : K \to \mathbb{C}$, $i = 0, 1, \ldots, n-1$ be its distinct embeddings. If $\beta \in K$ is an algebraic number of degree $m \leq n$ with minimal polynomial $q(x) = -q_0 - \cdots - q_{m-1}x^{m-1} + x^m$, then m diviede n and*

$$\begin{aligned} T_K(\beta) &= \sum_{i<n} \sigma_i(\beta) = \frac{n}{m} \cdot q_{m-1}, \\ N_K(\beta) &= \prod_{i<n} \sigma_i(\beta) = (-1)^n \cdot (-q_0)^{\frac{n}{m}}. \end{aligned}$$

In particular $T_K(a) = na$, $N_K(a) = a^n$ for $a \in \mathbb{Q}$.

Proof: Let $K = \mathbb{Q}(\alpha)$, and assume that $\sigma_0, \ldots, \sigma_{n-1}$ are distinct embeddings of $\mathbb{Q}(\alpha)$ to $\mathbb{C}$ such that the restrictions of $\sigma_0, \ldots, \sigma_{m-1}$ to $\mathbb{Q}(\beta)$ are distinct embeddings of $\mathbb{Q}(\beta)$ to $\mathbb{C}$. We have $\beta\alpha^i = \sum_{j<n} M_\alpha(\beta)_{ij}\alpha^j$ and

$$\sum_{j<n} I_{jk}\sigma_k(\beta)\sigma_j(\alpha^i) = \sigma_k(\beta)\sigma_k(\alpha^i) = \sigma_k(\beta\alpha^i) = \sum_{j<n} M_\alpha(\beta)_{ij}\sigma_k(\alpha^j)$$

where I is the identity matrix. Define $(n \times n)$-matrices $S(\alpha)$, $N(\beta)$ by $S(\alpha)_{ij} = \sigma_j(\alpha^i)$, $N(\beta)_{jk} = I_{jk}\sigma_k(\beta)$. Then $S(\alpha) \cdot N(\beta) = M_\alpha(\beta) \cdot S(\alpha)$ and $N(\beta)$ is a diagonal matrix with diagonal formed by $\sigma_j(\beta)$. By Propositions 7.13, 7.14 we get

$$\begin{aligned} T_K(\beta) &= \operatorname{tr}(M_\alpha(\beta)) = \operatorname{tr}(N(\beta)) = \sum_{j<n} \sigma_j(\beta) = \frac{n}{m} \sum_{j<m} \sigma_j(\beta) \\ &= \frac{n}{m} q_{m-1}, \\ N_K(\beta) &= \det(M_\alpha(\beta)) = \det(N(\beta)) = \prod_{j<n} \sigma_j(\beta) = (\prod_{j<m} \sigma_j(\beta))^{\frac{n}{m}} \\ &= ((-1)^{m+1}q_0)^{\frac{n}{m}} = (-1)^n \cdot (-q_0)^{\frac{n}{m}}. \end{aligned}$$

For $a \in \mathbb{Q}$ we have $m = 1$, $q_0 = a$, so $T_K(a) = na$, $N_K(a) = (-1)^n \cdot (-a)^n = a^n$. □

Proposition 7.16 *Let K be an algebraic number field of dimension n and $\sigma_0, \ldots, \sigma_{n-1}$ the distinct field embeddings of K into $\mathbb{C}$. For a vector $w \in K^n$ define $(n \times n)$-matrices $S(w)$, $\tau(w)$ by $S(w)_{ij} = \sigma_j(w_i)$, $\tau(w)_{ij} = T_K(w_i w_j)$. Then*

$$\det(\tau(w)) = \det(S(w))^2 = \Delta(w)$$

is called the **discriminant** *of w.*

Proof:

$$\tau(w)_{ij} = \sum_{k<n} \sigma_k(w_i)\sigma_k(w_j) = \sum_{k<n} S(w)_{ik} S(w)^T_{kj} = (S(w) \cdot S(w)^T)_{ij},$$

so $\tau(w) = S(w) \cdot S(w)^T$ and therefore $\det(\tau(w)) = \det(S(w))^2$. □

Proposition 7.17 *A vector $w \in K^n$ is a basis of K over $\mathbb{Q}$ iff $\Delta(w) \neq 0$.*

Proof: If $w_0, \ldots, w_{n-1}$ are linearly dependent then $\sum_{i<n} a_i w_i = 0$ for some nonzero $a \in \mathbb{Q}^n$. For each j we get $\sum_{i<n} a_i T_K(w_i w_j) = T_K(\sum_{i<n} a_i w_i w_j) = T_K(0) = 0$, so $\det(\tau(w)) = 0$. Conversely assume that $w \in K^n$ is a basis and $\det(\tau(w)) = 0$. Then there exist nonzero a_i such that $\sum_{i<n} a_i T_K(w_i w_j) = 0$. If $\alpha = \sum_{i<n} a_i w_i$, then $T(\alpha w_j) = 0$ for all w_j. Since w is a basis of K, we get $T_K(\alpha\beta) = 0$ for all $\beta \in K$ which is a contradiction since $T_K(1) \neq 0$. □

Proposition 7.18 *Assume that $w, v \in K^n$ are bases of K over $\mathbb{Q}$ and A is the transformation matrix with $v_i = \sum_{j<n} A_{ij} w_j$. Then $\Delta(v) = \det(A)^2 \cdot \Delta(w)$.*

Proof: By Proposition 7.12 we have

$$T_K(v_i v_j) = \sum_{k<n} \sum_{l<n} A_{ik} A_{jl} T_K(w_k w_l) = (A\tau(w)A^T)_{ij}$$

so $\tau(v) = A \cdot \tau(w) \cdot A^T$ and $\det(\tau(v)) = \det(\tau(w)) \cdot \det(A)^2$ □

7.3 Computable ordered fields

Definition 7.19 *An* **ordered field** *is a pair (K, P), where K is a field and $P \subset K$ is its subset (of positive elements) which satisfies the following conditins:*
1. For every $x \in K$ either $x \in P$ or $-x \in P$ or $x = 0$.
2. If $x, y \in P$ then $x + y \in P$ and $xy \in P$.
3. We say that (K, P) is a **computable ordered field**, *if the operations of addition, subtraction, multiplication and division are algorithmic and if the set P of positive elements is computable.*

While $\mathbb{R}$ or $\mathbb{Q}$ together with its sets of positive elements are ordered fields, the field $\mathbb{C}$ of complex numbers is not orderable. There exists no set $P \subset \mathbb{C}$ such that $(\mathbb{C}, P)$ is an ordered field. On the other hand, each subfield of $\mathbb{R}$ is an ordered field with the order inherited from $\mathbb{R}$, If (K, P) is an ordered field, then the inequality is defined by $x < y$ iff $y - x \in P$ and $x \leq y$ if $x < y$ or $x = y$.

Theorem 7.20 *Each real algebraic number field $K \subset \mathbb{R}$ is a computable ordered field.*

Proof: The operations of addition, subtraction, multiplication and division are performed on $\mathbb{Q}^n$ and they are clearly algorithmic. It rests the inequality or the set of positive elements P. Let $K = \mathbb{Q}(\alpha)$ and let p be the minimal polynomial for α. Since all the fields $\sigma_i(K)$ are isomorphic, we must distinguish α from its conjugate roots. We can do it by specifying an interval I with rational endpoints such that α is the only root of P in I. This can be verified

by the Sturm theorem 5.29. Given an element $\beta = q(\alpha) \in \mathbb{Q}(\alpha)$ where $q \in \mathbb{Q}_n[x]$, we have to decide whether $q(\alpha) > 0$ or $q(\alpha) < 0$. If q has no root in I, then $q(\alpha) > 0$ iff q is positive on both endpoints of I. If q has a root in I, then we take a smaller interval $I_1 \subset I$ with rational endpoints which contains α and repeat the test with I_1. Since α is irrational we find finally an interval which contains α but q has no root in I_k. Then the sign of $q(\alpha)$ is the sign of q at any endpoint of I_k. □

All arithmetical algorithms use only the field operations with the entries of the matrices F_a and V_p and the comparisions.

Corollary 7.21 *Let α be a real algebraic number and let (F, G, V) be a sofic number system such that all entries of matrices F_a, V_p are in $\mathbb{Q}(\alpha)$. Then all arithmetic algorithms work properly.*

7.4 Algebraic integers

Definition 7.22 *An algebraic number α is an* **algebraic integer** *if its minimal polynomial belongs to $\mathbb{Z}[x]$, i.e., if there exist $p_i \in \mathbb{Z}$ such that $-p_0 - p_1\alpha - \cdots - p_{n-1}\alpha^{n-1} + \alpha^n = 0$.*

Denote by $\overline{\mathbb{Q}}$ the set of algebraic integers. For an algebraic number field K denote by $\mathbb{Z}_K$ the set of algebraic integers of K. For an algebraic number α denote by $\mathbb{Z}_\alpha = \mathbb{Z}_{\mathbb{Q}(\alpha)}$ the set of algebraic integers of $\mathbb{Q}(\alpha)$. For $\alpha \in \mathbb{C}$ denote by $\mathbb{Z}(\alpha) = \{q(\alpha) : \ q \in \mathbb{Z}[x]\}$ the smallest ring which contains α.

Proposition 7.23 *If α is an algebraic number, then there exists a positive integer $k \in \mathbb{Z}$ such that $k\alpha$ is an algebraic integer.*

Proof: Let $p \in \mathbb{Q}[x]$ be the minimal polynomial of α. Denote by $k > 0$ the GCD of the denominators of the coefficients of p, so $p(x) = (p_0 + \cdots + p_n x^n)/k$ for some $p_i \in \mathbb{Z}$. Then $p_n\alpha$ is a root of $q(x) = p_0 p_n^{n-1} + p_1 p_n^{n-2} x + \cdots + p_{n-2} p_n x^{n-2} + p_{n-1} x^{n-1} + x^n$. □

Proposition 7.24 *Let d be a positive squarefree integer greater than 1.*

1. If $\mathrm{mod}_4(d) = 2$ or $\mathrm{mod}_4(d) = 3$ then $\mathbb{Z}_{\sqrt{d}} = \{x_0 + x_1\sqrt{d} : \ x_0, x_1 \in \mathbb{Z}\}$.

1. If $\mathrm{mod}_4(d) = 1$ then $\mathbb{Z}_{\sqrt{d}} = \{x_0 + x_1\frac{\sqrt{d}-1}{2} : \ x_0, x_1 \in \mathbb{Z}\}$.

Proof: A number $x = x_0 + x_1\sqrt{d}$ with $x_0, x_1 \in \mathbb{Q}$ is an algebraic integer iff its minimal polynomial $x^2 - 2x_0 x + x_0^2 - dx_1^2$ has integer coefficients, iff $T_{\sqrt{d}}(x) = 2x_0 \in \mathbb{Z}$ and $N_{\sqrt{d}}(x) = x_0^2 - dx_1^2 \in \mathbb{Z}$. It follows $4dx_0^2 \in \mathbb{Z}$, so $4dx_1^2 \in \mathbb{Z}$ and since d is squarefree, $2x_1 \in \mathbb{Z}$. For the integers $y_0 = 2x_0$, $y_1 = 2x_1$ we have $4|(y_0^2 - dy_1^2)$. If $d = 4k + 2$ then $4|(y_0^2 - 2y_1^2)$. If $d = 4k + 3$ then $4|(y_0^2 - 3y_1^2)$. In both cases this is possible only if both y_0, y_1 are even, so $x_0, x_1 \in \mathbb{Z}$. If $d = 4k + 1$ then $4|(y_0^2 - y_1^2)$ and y_0, y_1 must have the same parity. We get

$$x_0 + x_1\sqrt{d} = \frac{y_0 + y_1}{2} + \frac{\sqrt{d} - 1}{2} y_1,$$

where $\frac{y_0+y_1}{2} \in \mathbb{Z}$ and $y_1 \in \mathbb{Z}$. □

Definition 7.25 *A set $M \subseteq \mathbb{C}$ is a* **free $\mathbb{Z}$-module** *if it is a group with respect to the addition, i.e., if $x+y \in M$ and $x-y \in M$ whenever $x, y \in M$. A free $\mathbb{Z}$-module M is* **finitely generated** *if there exists m and $w_0, \ldots, w_{m-1} \in M$ such that each $x \in M$ can be written as $x = \sum_{i<m} x_i w_i$ with $x_i \in \mathbb{Z}$.*

Proposition 7.26 *$\alpha \in \mathbb{C}$ is an algebraic integer iff there exists a finitely generated free $\mathbb{Z}$-module $M \subset \mathbb{C}$ such that $\alpha M = \{\alpha x : \ x \in M\} \subseteq M$.*

Proof: If $\alpha \in \overline{\mathbb{Q}}$ is an algebraic integer of degree m then $\mathbb{Z}(\alpha) = \{q(\alpha) : \ q \in \mathbb{Z}_m[x]\}$. Indeed, using the equality $\alpha = p_0 + p_1\alpha + \cdots + p_{n-1}\alpha^{n-1}$, we can find for any $q \in \mathbb{Z}(\alpha)$ some $r \in \mathbb{Z}_m(\alpha)$ with $q(\alpha) = r(\alpha)$. Clearly $\alpha\mathbb{Z}(\alpha) \subseteq \mathbb{Z}(\alpha)$ and $1, \alpha, \ldots, \alpha^{m-1}$ are generators of $\mathbb{Z}(\alpha)$. Conversely assume that M is a finitely generated module with $\alpha M \subseteq M$. Let $w_0, \ldots, w_{m-1} \in \mathbb{C}$ be generators of M. Then there exist $C_{ij} \in \mathbb{Z}$ such that $\alpha w_i = \sum_j C_{ij} w_j$. If $w = [w_0, \ldots, w_{m-1}]^T$ is the column vector and C is the matrix with entries C_{ij} then we have $C \cdot w = \alpha w$, so w is the right eigenvector of C with the eigenvalue α. It follows that $\det(I\alpha - C) = 0$, where I is the identity matrix. Thus α is a root of a monic polynomial $p(x) = \det(Ix - C)$ which belongs to $\mathbb{Z}[x]$. It follows that the minimal polynomial of α belongs to $\mathbb{Z}[x]$. □

Proposition 7.27 *The set $\overline{\mathbb{Q}}$ of algebraic integers is a subring of $\mathbb{C}$.*

Proof: Let α be an arithmetic integer of degree n and let β be an arithmetic integer of degree m. Then $\mathbb{Z}(\alpha, \beta) = \{\sum_{ij} x_{ij}\alpha^i\beta^j : \ x_{ij} \in \mathbb{Z}\}$ is a finitely generated module, $(\alpha+\beta)\mathbb{Z}(\alpha, \beta) \subseteq \mathbb{Z}(\alpha, \beta)$ and $(\alpha\beta)\mathbb{Z}(\alpha, \beta) \subseteq \mathbb{Z}(\alpha, \beta)$. Thus $\alpha + \beta \in \overline{\mathbb{Q}}$ and $\alpha\beta \in \overline{\mathbb{Q}}$. □

Corollary 7.28 *For each algebraic number field K the set $\mathbb{Z}_K = K \cap \overline{\mathbb{Q}}$ of its algebraic integers is a ring.*

Definition 7.29 *Let K be an algebraic number field of dimension n. We say that $w \in (\mathbb{Z}_K)^n$ is an integral basis of $\mathbb{Z}_K$ if $\mathbb{Z}_K = \{x_i w_i : \ x_i \in \mathbb{Z}\}$.*

Proposition 7.30 *If K is an algebraic number field then $\mathbb{Z}_K$ has an integral basis.*

Proof: By Proposition 7.23 there exists an algebraic integer $\alpha \in K$ such that $K = \mathbb{Q}(\alpha)$. Thus there exist bases of K over $\mathbb{Q}$ which consist of algebraic integers. If w is such a basis then $T_K(w_i w_j) \in \mathbb{Z}$ are integers by Proposition 7.15, so $\Delta(w) \in \mathbb{Z}$. Take a basis $w \in (\mathbb{Z}_K)^n$ with minimal absolute value of the discriminant $\Delta(w)$. We show that w is an integral basis. Since w is a $\mathbb{Q}$-basis for K, for each $x \in \mathbb{Z}_K$ there exist unique $x_i \in \mathbb{Q}$ such that $x = \sum_{i<n} x_i w_i$. We show that $x_i \in \mathbb{Z}$. If not we can reorder w_i so that $x_0 \notin \mathbb{Z}$. There exists $m \in \mathbb{Z}$ such that $0 < y_0 = x_0 - m < 1$. Take $v_0 = x - mw_0 \in \mathbb{Z}_K$ and $v_i = w_i$ for $i > 0$. Then $v \in (\mathbb{Z}_K)^n$ is a basis for K. Since $v_0 = y_0 w_0 + x_1 w_1 + \cdots + x_{n-1} w_{n-1}$, we get $v = A \cdot w$ where

$$A = \begin{bmatrix} y_0 & x_1 & x_2 & \cdots & x_{n-1} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Since $\det(A) = y_0 < 1$, we get $|\Delta(v)| < |\Delta(w)|$ by Proposition 7.18 and this is a contradiction. □

Corollary 7.31 *All integral bases of an algebraic number field K have the same discriminant which is called the* **discriminant** $\Delta(K)$ *of K.*

The quadratic field $\mathbb{Q}(\sqrt{d})$ with squarefree $d = 4k+2$ or $d = 4k+3$ has integral basis $w = [1, \sqrt{d}]^T$. For the discriminant we get

$$\tau(w) = \begin{bmatrix} 2 & 0 \\ 0 & 2d \end{bmatrix},\ \Delta(\mathbb{Q}(\sqrt{d})) = \det(\tau(w)) = 4d$$

If $d = 4k+1$, then $\mathbb{Q}(\sqrt{d})$ has integral basis $w = [1, \frac{\sqrt{d}-1}{2}]^T$ and discriminant

$$\tau(w) = \begin{bmatrix} 2 & -1 \\ -1 & (d+1)/2 \end{bmatrix},\ \Delta(\mathbb{Q}(\sqrt{d})) = \det(\tau(w)) = d.$$

7.5 Pisot and Salem numbers

Definition 7.32 *We say that a real algebraic integer $\alpha > 1$ is a* **Pisot number** *if for all its conjugates we have $|\sigma_i(\alpha)| < 1$. A real algebraic integer $\alpha > 1$ is a* **Salem number** *if for all its conjugates we have $|\sigma_i(\alpha)| \leq 1$ and there exists a conjugate with $|\sigma_i(\alpha)| = 1$.*

Each ordinary integer $n \geq 2$ is a Pisot number. The golden mean $\alpha = \frac{\sqrt{5}+1}{2} \doteq 1.618$ is a Pisot number. Its minimal polynomial is $p(x) = x^2 - x - 1$ and its conjugate is $\frac{1-\sqrt{5}}{2} \doteq -0.618$. The smallest known Salem number is the largest real root of the polynomial

$$p(x) = x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1$$

which is approximatgely 1.176. We are going to show that every algebraic number field contains a Pisot number which generates it.

A **lattice** is a finitely generated subgroup $L \subseteq \mathbb{R}^n$ with pointwise addition. Thus a lattice is a set $L = \{\sum_i x_i v_i : \ x_i \in \mathbb{Z}\}$ where $v_i \in \mathbb{R}^n$ are linearly $\mathbb{R}$-independent vectors. The determinant of the lattice is the determinant of its matrix $V_{ij} = (v_j)_i$ whose columns are the vectors v_j. The simplest lattice is $\mathbb{Z}^n$ which is generated by the identity matrix I. We say that a set $X \subseteq \mathbb{R}^n$ is **convex** if for every $x, y \in X$ and $0 < t < 1$ we have $tx + (1-t)y \in X$. We say that X is **symmetric** if $-x \in X$ whenever $x \in X$ (see Micciancio and Goldwasser [51]).

Proposition 7.33 *Let $X \subseteq \mathbb{R}^n$ be a convex symmetric set with volume* $\mathrm{vol}(X) > 2^n$. *Then X contains a nonzero point of $\mathbb{Z}^n$.*

Proof: Consider a mapping $f : X \to \mathbb{R}^n$ given by $f(x)_i = |x_i|_2 \in [0, 2]$. This mapping preserves the volume and its image is included in a cube $f(X) \subseteq [0, 2]^n$ with volume 2^n. Thus there exist different $x, y \in X$ with $f(x) = f(y)$, i.e., $y = x + 2u$ for some $u \in \mathbb{Z}^n$ with $u \neq 0$. Since X is symmetric, we get $-x \in X$ and since X is convex we have $u = \frac{1}{2}(x + 2u - x) = \frac{1}{2}(y - x) \in X$. □

Theorem 7.34 (Minkowski) *Let $L \subseteq \mathbb{R}^n$ be a lattice and let $X \subseteq \mathbb{R}^n$ be a symmetric convex set with volume* $\mathrm{vol}(X) > 2^n \cdot |\det(L)|$. *Then X contains a nonzero point of L.*

Proof: Let $L = \{\sum_i x_i v_i : \ x_i \in \mathbb{Z}\}$ be a lattice generated by linearly independent vectors $v_i \in \mathbb{R}^n$. Define the square matrix V by $V_{ij} = (v_j)_i$. Then $L = \{V \cdot x : \ x \in \mathbb{Z}^n\}$ and $V^{-1}(L) = \mathbb{Z}^n$. If $X \subseteq \mathbb{R}^n$ is convex and symmetric then $V^{-1}(X)$ is convex and symmetric and its volume is $\mathrm{vol}(X)/|\det(L)|$. Thus $V^{-1}(X)$ contains a nonzero point of $\mathbb{Z}^n$ by Proposition 7.33, and therefore X contains a nonzero point of L. □

Theorem 7.35 (Salem) *In each algebraic number field K there exists a Pisot number $\alpha \in K$ such that $K = \mathbb{Q}(\alpha)$.*

Proof: Let $w \in K^n$ be an integral basis of $\mathbb{Z}_K$ and let $\sigma_0, \ldots, \sigma_{n-1}$ be the distinct embeddings of K into $\mathbb{C}$. We have a matrix $S(w)$ defined by $S(w)_{ij} = \sigma_j(w_i)$. Take a lattice L whose generators are the rows of $S(w)$, so $L = \{\sum_i y_i(\sigma_0(w_i), \sigma_1(w_i), \ldots, \sigma_{n-1}(w_i)) : \ y_i \in \mathbb{Z}\}$. Then $\Delta(K) = \det(S(w))^2 = \det(L)^2$. For $0 < \delta < 1$, $B > \sqrt{\Delta(K)}/\delta^{n-1}$ consider the set

$$X = \{x \in \mathbb{R}^n : \ |x_0| < B, \forall i > 0, |x_i| < \delta\}.$$

Then the volume of X is $2^n B\delta^{n-1} > 2^n|\det(L)|$ so by Theorem 7.34 there exists a nonzero $x \in L \cap X$ and there exist $y_i \in \mathbb{Z}$ such that $x_j = \sum_i y_i\sigma_j(w_i)$. Since X is symmetric, we can assume $x_0 \geq 0$. Thus $\alpha = x_0 = y_0w_0 + \cdots + y_{n-1}w_{n-1} \in \mathbb{Z}_K$, $0 < \alpha < B$, $|\sigma_i(\alpha)| = |x_j| < \delta$ for $0 < j < n$. By Proposition 7.15, $\prod_j \sigma_j(\alpha) \in \mathbb{Z} \setminus \{0\}$, and therefore $\alpha > 1$. Thus $\alpha \in \mathbb{Z}_K$ is a Pisot number. We show that $K = \mathbb{Q}(\alpha)$. If not then $m = [K : \mathbb{Q}(\alpha)] < n$ and α would appear n/m times among the conjugates $\sigma_i(\alpha)$. However, this is not the case since $|\sigma_i(\alpha)| < 1$ for all $i > 0$. □

7.6 Positional systems

A positional number system for a bounded interval (see Section 1.4) is defined by a real base $\beta > 1$ and a finite contiguous set of digits $A = [r, s] = \{r, r + 1, \ldots, s - 1, s\} \subset \mathbb{Z}$ with $s - r \geq \beta - 1$. If $s - r > \beta$, then the system is redundant. The base β need not be an integer. The study of positional system with noninteger bases has been initialized by Rényi [58]. The surjective and continuous value mapping $\Phi : A^\omega \to [\frac{r}{\beta-1}, \frac{s}{\beta-1}]$ is given by $\Phi(u) = \sum_{i\geq 0} u_i\beta^{-i-1}$. Thus we have a sofic number system whose graph has a single vertex λ and edges $\lambda \xrightarrow{a} \lambda$ for all $a \in A$. For V_λ we get

$$\begin{aligned} V_\lambda &= [b_0, b_1] = [\tfrac{r}{\beta-1}, \tfrac{s}{\beta-1}] \\ F_a(V_\lambda) &= [\tfrac{b_0+a}{\beta}, \tfrac{b_1+a}{\beta}] = [\tfrac{a}{\beta} + \tfrac{r}{\beta(\beta-1)}, \tfrac{a}{\beta} + \tfrac{s}{\beta(\beta-1)},] \end{aligned}$$

Among the expansions of $x \in [b_0, b_1]$ we consider the smallest (in lexicographic order) which we call the **lazy expansion** and the largest which we call the **greedy expansion**. The **lazy function** $L_\beta : [b_0, b_1] \to [b_0, b_1]$ and the **lazy expansion map** $\mathcal{E}_l : [b_0, b_1] \to A^\omega$ are defined by

$$\begin{aligned} L_\beta(x) &= \beta x - \mathbf{a}_l(x) \\ \mathcal{E}_l(x)_i &= \mathbf{a}_l(L_\beta^i(x)), \text{ where} \\ \mathbf{a}_l(x) &= \min\{a \in A : \ x \in F_a(V_\lambda)\} \end{aligned}$$

Proposition 7.36 *If $\beta > 1$, $A = [r, s] \subseteq \mathbb{Z}$, $s - r \geq \beta - 1$ then*
1. $\mathbf{a}_l(x) = \max\{r, \lceil \beta x - b_1 \rceil\}$.
2. $\Phi(\mathcal{E}_l(x)) = x$ for every $x \in V_\lambda$
3. $\mathcal{E}_l(\Phi(u)) \preceq u$ for every $u \in A^\omega$.

Proof: 1. We have $\mathbf{a}_l(x) = r$ iff $x \leq \frac{b_1+r}{\beta}$ iff $\beta x - b_1 \leq r$ iff $\lceil \beta x - b_1 \rceil \leq r$. For $a > r$ we have $\mathbf{a}_l(x) = a$ iff $\frac{b_1+a-1}{\beta} < x \leq \frac{b_1+a}{\beta}$ iff $a - 1 < \beta x - b_1 \leq a$ iff $\lceil \beta x - b_1 \rceil = a$, so we get $\mathbf{a}_l(x) = \max\{r, \lceil \beta x - b_1 \rceil\}$.

2,3. We use Theorem 4.29. We have $\Phi(u) = x$ iff there exists a sequence $x_i \in V_\lambda$ with $x_{i+1} = F_{u_i}^{-1}(x_i)$. Since $L_\beta(x) = F_a^{-1}(x)$ for $a = \mathbf{a}_l(x)$, the result follows. □

The lazy expansion map is in fact the left expansion in a partition number system (see Definition 4.18 and Figure 7.1 left).

Definition 7.37 *The* **lazy partition number system** *with base $\beta > 1$ and alphabet $A = [r, s] \subset \mathbb{Z}$ is given by transformations $F_a(x) = \frac{x+a}{\beta}$ and intervals*

$$W_a = \begin{cases} (b_0, \frac{b_1+r}{\beta}) & \text{for} \quad a = r \\ (\frac{b_1+a-1}{\beta}, \frac{b_1+a}{\beta}) & \text{for} \quad r < a \le s \end{cases}$$

Proposition 7.38 *For the lazy partition number system (F, W) with base $\beta > 1$ and alphabet $A = [r, s]$ we have*
1. $\mathcal{E}_-(x) = \mathcal{E}_l(x)$, so $\mathcal{E}_-(x)_i = \mathbf{a}_l(L_\beta^i(x))$
2. $\mathcal{E}_-(\Phi(u)) \preceq u$ for any $u \in A^\omega$.
3. $u \in A^\omega$ belongs to $\mathcal{S}_{F,W}$ iff $\sigma^{n+1}(u) \succeq \mathcal{E}_+(b_1 - 1)$ whenever $u_n > p$.
4. The subshift $\mathcal{S}_{F,W}$ is sofic iff $\mathcal{E}_+(b_1 - 1)$ is periodic.

Proof: Items 1,2 follow from $\mathbf{a}_l(x) = \min\{a \in A : \ x \in \overline{W_a}\}$.
3. We use Theorem 4.20. For $a \in A$ we have $\mathcal{E}_-(r_a) = as^\omega$. If $a > r$ then $\mathcal{E}_+(l_a) = a\mathcal{E}_+(b_1 - 1)$, for $a = r$ we have $\mathcal{E}_+(l_r) = r^\omega$. If $u_n = r$, then the condition $\mathcal{E}_+(l_r) = r^\omega \preceq \sigma^n(u) \preceq rs^\omega = \mathcal{E}_-(r_r)$ is always satisfied. If $u_n > r$ then the condition $\mathcal{E}_+(l_a) = a\mathcal{E}_+(b_1 - 1) \preceq \sigma^n(u) \preceq as^\omega = \mathcal{E}_-(r_r)$ is satisfied iff $\sigma^{n+1}(u) \succeq \mathcal{E}_+(b_1 - 1)$.
Item 4 follows from Theorem 4.24. □

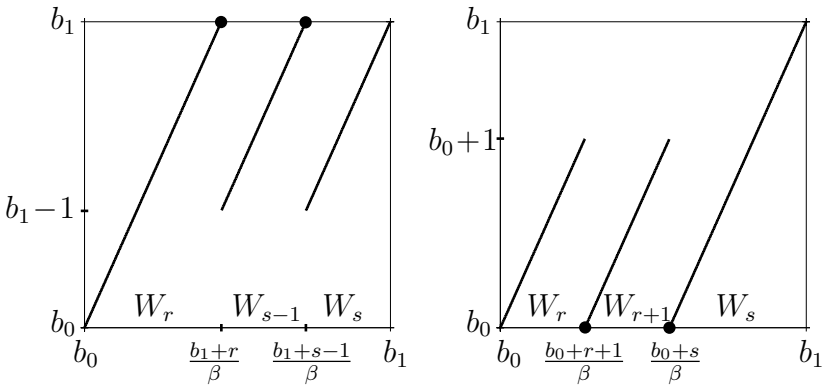


Figure 7.1: The lazy function L_β and the lazy partition $\{W_a : \ a \in A\}$ (left), the greedy function G_β and the greedy partition $\{W_a : \ a \in A\}$ (right) in a positional system with noninteger base

The **greedy function** $G_\beta : [b_0, b_1] \to [b_0, b_1]$ and the **greedy expansion map** $\mathcal{E}_g : [b_0, b_1] \to A^\omega$ are given by

$$\begin{aligned} G_\beta(x) &= \beta x - \mathbf{a}_g(x) \\ \mathcal{E}_g(x)_i &= \mathbf{a}_g(G_\beta^i(x)), \text{ where} \\ \mathbf{a}_g(x) &= \max\{a \in A : \ x \in F_a(V_\lambda)\} \end{aligned}$$

Proposition 7.39 *If $\beta > 1$, $A = [r, s] \subseteq \mathbb{Z}$, $s - r \ge \beta - 1$ then*

1. $\mathbf{a}_g(x) = \min\{s, \lfloor \beta x - b_0 \rfloor\}$
2. $\Phi(\mathcal{E}_g(x)) = x$ *for every* $x \in V_\lambda$
3. $u \preceq \mathcal{E}_g(\Phi(u))$ *for every* $u \in A^\omega$.

Proof: 1. We have $\mathbf{a}_g(x) = s$ iff $\frac{b_0+s}{\beta} \le x$ iff $s \le \beta x - b_0$ iff $s \le \lfloor \beta x - b_0 \rfloor$. For $a < s$ we have $\mathbf{a}_g(x) = a$ iff $\frac{b_0+a}{\beta} \le x < \frac{b_0+a+1}{\beta}$ iff $a \le \beta x - b_0 < a + 1$ iff $\lfloor \beta x - b_0 \rfloor = a$, so we get $\mathbf{a}_g(x) = \min\{s, \lfloor \beta x - b_0 \rfloor\}$.
2,3 follow from Theorem 4.29. □

The greedy expansion map is the right expansion in a partition number system (see Figure 7.1 right)

Definition 7.40 *The* **greedy partition number system** *with base* $\beta > 1$ *and alphabet* $A = [r, s] \subset \mathbb{Z}$ *is given by transformations* $F_a(x) = \frac{x+a}{\beta}$ *and intervals*

$$W_a = \begin{cases} (\frac{b_0+a}{\beta}, \frac{b_0+a+1}{\beta}) & \text{for} \quad r \le a < s \\ (\frac{b_0+s}{\beta}, b_1) & \text{for} \quad a = s \end{cases}$$

Proposition 7.41 (Parry [53]) *For the greedy partition number system* (F, W) *with base* $\beta > 1$ *and alphabet* $A = [r, s]$ *we have*
1. $\mathcal{E}_+(x) = \mathcal{E}_g(x)$, *so* $\mathcal{E}_+(x)_i = \mathbf{a}_g(G^i_\beta(x))$
2. $u \preceq \mathcal{E}_+(\Phi(u))$ *for any* $u \in A^\omega$.
3. $u \in A^\omega$ *belongs to* $\mathcal{S}_{F,W}$ *iff* $\sigma^{n+1}(u) \preceq \mathcal{E}_-(b_0 + 1)$ *whenever* $u_n < q$.
4. The subshift $\mathcal{S}_{F,W}$ *is sofic iff* $\mathcal{E}_-(b_0 + 1)$ *is periodic.*

Proof: The proof is similar to the proof of Theorem 7.38. For $a \in A$ we have $\mathcal{E}_+(l_a) = ar^\omega$. If $a < s$ then $\mathcal{E}_-(r_a) = a\mathcal{E}_-(b_0 + 1)$, otherwise $\mathcal{E}_-(r_s) = s^\omega$. If $u_n = s$, then the condition $\mathcal{E}_+(l_q) = sr^\omega \preceq \sigma^n(u) \preceq s^\omega = \mathcal{E}_-(r_s)$ is always satisfied. If $a = u_n < s$ then the condition $\mathcal{E}_+(l_a) = ar^\omega \preceq \sigma^n(u) \preceq a\mathcal{E}_-(b_0 + 1)$ is satisfied iff $\sigma^{n+1}(u) \preceq \mathcal{E}_-(b_0 + 1)$. □

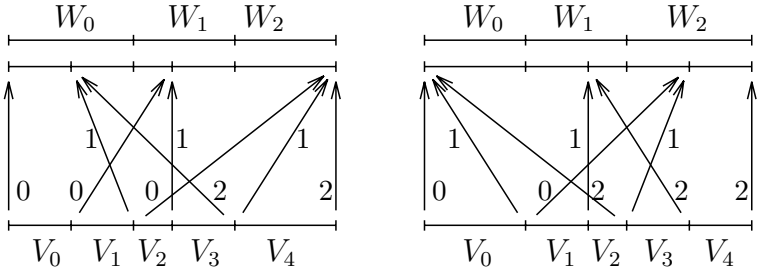


Figure 7.2: The β-system with $\beta = \frac{3+\sqrt{5}}{2}$.

As an example consider the positional system with base $\beta = \frac{3+\sqrt{5}}{2} \doteq 2.618$ and alphabet $A = [0, 1, 2]$. The intervals W_a of the lazy partition have endpoints 0, $\frac{b_1}{\beta} \doteq 0.472$, $\frac{b_1+1}{\beta} \doteq 0.854$, $b_1 = \frac{2}{\beta} = \sqrt{5} - 1 \doteq 1.236$. Since $F_0^{-1}(b_1 - 1) = \beta - 2 \in W_1$ and $F_1^{-1}(\beta - 2) = \beta - 2$, we get $\mathcal{E}_+(b_1 - 1) = 01^\omega$, so the lazy subshift $\mathcal{S}_{F,W}$ is sofic. Its SFT partition $V = \{V_a : a \in [0, 5]\}$ has

endpoints 0, $b_1 - 1 \doteq 0.236$, $\frac{b_1}{\beta} \doteq 0.472$, $\beta - 2 \doteq 0.618$, $\frac{b_1+1}{\beta} \doteq 0.854$, $b_1 \doteq 1.236$ and graph (see Figure 7.2 left)

$$\begin{aligned}
[0, b_1 - 1] = \overline{V_0} &= F_0(\overline{V_0} \cup \overline{V_1} \cup \overline{V_2}) \\
[b_1 - 1, \tfrac{b_1}{\beta}] = \overline{V_1} &= F_0(\overline{V_3} \cup \overline{V_4}) \\
[\tfrac{b_1}{\beta}, \beta - 2] = \overline{V_2} &= F_1(\overline{V_1} \cup \overline{V_2}) \\
[\beta - 2, \tfrac{b_1+1}{\beta}] = \overline{V_3} &= F_1(\overline{V_3} \cup \overline{V_4}) \\
[\tfrac{b_1+1}{\beta}, b_1] = \overline{V_4} &= F_2(\overline{V_1} \cup \overline{V_2} \cup \overline{V_3} \cup \overline{V_4})
\end{aligned}$$

The intervals W_a of the greedy partition have endpoints 0, $\frac{1}{\beta} \doteq 0.382$, $\frac{2}{\beta} \doteq 0.764$, and $b_1 = \frac{2}{\beta} = \sqrt{5} - 1 \doteq 1.236$. Since $F_2^{-1}(1) = \beta - 2 \in W_1$ and $F_1^{-1}(\beta - 2) = \beta - 2$, we get $\mathcal{E}_-(1) = 21^\omega$, so the greedy subshift $\mathcal{S}_{F,W}$ is sofic. Its SFT partition has cutpoints 0, $\frac{1}{\beta} \doteq 0.382$, $\beta - 2 = 0.618$, $\frac{2}{\beta} \doteq 0.764$, 1 and b_1 and graph (see Figure 7.2 right)

$$\begin{aligned}
[0, \tfrac{1}{\beta}] = \overline{V_0} &= F_0(\overline{V_0} \cup \overline{V_1} \cup \overline{V_2} \cup \overline{V_3}) \\
[\tfrac{1}{\beta}, \beta - 2] = \overline{V_1} &= F_1(\overline{V_0} \cup \overline{V_1}) \\
[\beta - 2, \tfrac{2}{\beta}] = \overline{V_2} &= F_1(\overline{V_2} \cup \overline{V_3}) \\
[\tfrac{2}{\beta}, 1] = \overline{V_3} &= F_2(\overline{V_0} \cup \overline{V_1}) \\
[1, b_1] = \overline{V_4} &= F_2(\overline{V_2} \cup \overline{V_3} \cup \overline{V_4})
\end{aligned}$$

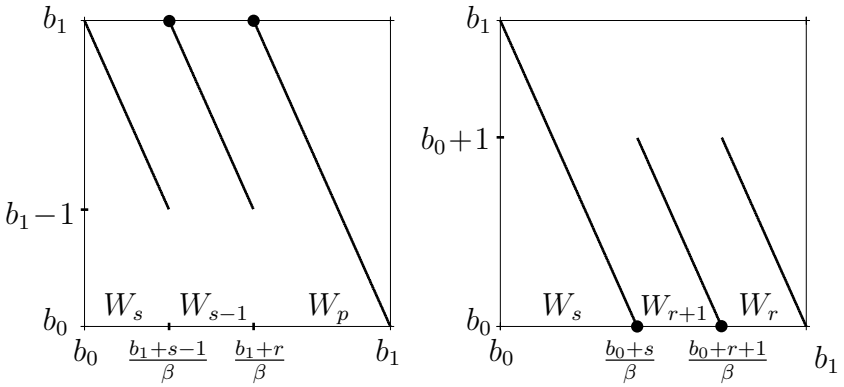


Figure 7.3: The lazy expansion function L_β (left) and the greedy expansion function G_β (right) in a positional system with negative noninteger base

The positional system with negative base $\beta < -1$ and the alphabet $A = [r, s]$ determines a mapping $\Phi : A^\omega \to [b_0, b_1]$ where $b_0 = \frac{s\beta+r}{\beta^2-1}$ and $b_1 = \frac{r\beta+s}{\beta^2-1}$, so

$$\begin{aligned}
V_\lambda &= [b_0, b_1] = [\tfrac{s\beta+r}{\beta^2-1}, \tfrac{r\beta+s}{\beta^2-1}] \\
F_a(V_\lambda) &= [\tfrac{b_1+a}{\beta}, \tfrac{b_0+a}{\beta}]
\end{aligned}$$

For the lazy and greedy expansions we get the same formulas as for $\beta > 1$.

$$\begin{aligned}
\mathbf{a}_l(x) &= \min\{a \in A :\ x \in F_a(V_\lambda)\} = \max\{r, \lceil \beta x - b_1 \rceil\} \\
L_\beta(x) &= \beta x - \mathbf{a}_l(x), \\
\mathcal{E}_l(x)_i &= \mathbf{a}_l(L_\beta^i), \\
\mathbf{a}_g(x) &= \max\{a \in A :\ x \in F_a(V_\lambda)\} = \min\{s, \lfloor \beta x - b_0 \rfloor\} \\
G_\beta(x) &= \beta x - \mathbf{a}_g(x), \\
\mathcal{E}_g(x)_i &= \mathbf{a}_g(G_\beta^i).
\end{aligned}$$

Indeed we have $\mathbf{a}_l(x) = r$ iff $\frac{b_1+r}{\beta} \leq x$ iff $\beta x - b_1 \leq r$ iff $\lceil \beta x - b_1 \rceil \leq r$. For $a > r$ we have $\mathbf{a}_l(x) = a$ iff $\frac{b_1+a}{\beta} \leq x < \frac{b_1+a-1}{\beta}$ iff $a - 1 < \beta x - b_1 \leq a$ iff $\lceil \beta x - b_1 \rceil = a$, so we get $\mathbf{a}_l(x) = \max\{r, \lceil \beta x - b_1 \rceil\}$ and similarly for $\mathbf{a}_g(x)$. The lazy partition system has intervals

$$W_a = \begin{cases} (\frac{b_1+r}{\beta}, b_1) & \text{for} \quad a = r \\ (\frac{b_1+a}{\beta}, \frac{b_1+a-1}{\beta}) & \text{for} \quad r < a \leq s \end{cases}$$

The greedy partition system has intervals

$$W_a = \begin{cases} (b_0, \frac{b_0+s}{\beta}) & \text{for} \quad a = s \\ (\frac{b_0+a+1}{\beta}, \frac{b_0+a}{\beta}) & \text{for} \quad r \leq a < s \end{cases}$$

When we iterate the greedy function (with $\beta > 1$) then for all $x < b_1$ there exists n_0 such that $G_\beta^n(x) \in [b_0, b_0 + 1]$ for all $n \geq n_0$. This is why the dynamics of the function G_β has been studied in Rényi [58] on this restricted interval.

Definition 7.42 *The* **restricted greedy partition number system** *with noninteger base $\beta > 1$ is given by the alphabet $A = [0, \lfloor \beta \rfloor]$, transformations $F_a(x) = \frac{x+a}{\beta}$ and intervals*

$$W_a = \begin{cases} (\frac{a}{\beta}, \frac{a+1}{\beta}) & \text{if} \quad a < \lfloor \beta \rfloor \\ (\frac{q}{\beta}, 1) & \text{if} \quad a = \lfloor \beta \rfloor \end{cases}$$

The value function Φ is defined on the subshift $\mathcal{S}_{F,W}$ and its range is the unit interval $[0, 1]$.

Proposition 7.43 *In the restricted greedy system with $\beta > 1$ we have $u \in \mathcal{S}_{F,W}$ iff for each $k \geq 0$ we have $\sigma^k(u) \preceq \mathcal{E}_-(1)$. The expansion subshift is sofic iff $\mathcal{E}_-(1)$ is periodic.*

Proof: We use Theorem 4.20. We have $l_a = a/\beta$ for $0 < a < n$ and $r_a = (a + 1)/\beta$ for $0 \leq a < q = \lfloor \beta \rfloor$, $r_q = 1$. We have $\mathcal{E}_+(l_a) = a0^\omega$ for each $a \in A$, $\mathcal{E}_-(r_a) = a\mathcal{E}_-(1)$ for $a < q$ and $\mathcal{E}_-(r_q) = \mathcal{E}_-(1)$. The condition $\mathcal{E}_+(l_{u_k}) \preceq \sigma^k(u) \preceq \mathcal{E}_-(r_{u_k})$ yields

$$\begin{aligned}
u_k < b \quad &\Rightarrow \quad u_k 0^\omega \preceq u_k \sigma^{k+1}(u) \preceq u_k \mathcal{E}_-(1) \Leftrightarrow \sigma^{k+1}(u) \preceq \mathcal{E}_-(1) \\
u_k = b \quad &\Rightarrow \quad u_k 0^\omega \preceq u_k \sigma^{k+1}(u) \preceq \mathcal{E}_-(1) \Leftrightarrow \sigma^k(u) \preceq \mathcal{E}_-(1)
\end{aligned}$$

Thus the condition is equivalent to $\sigma^k(u) \preceq \mathcal{E}_-(1)$ for all $k \geq 0$. □

Theorem 7.44 (Schmidt [61]) *If $\beta > 1$ is a Pisot number then every $x \in \mathbb{Q}(\beta) \cap [0, 1]$ has a periodic expansion in the restricted greedy system with base β.*

Proof: Denote by $\sigma_0, \ldots, \sigma_{n-1}$ the n distinct embeddings of $\mathbb{Q}(\beta)$ into $\mathbb{C}$ with σ_0 the identity. Let $x \in [0,1)$, $x_m = G_\beta^m(x)$, $u_m = \lfloor \beta G_\beta^m(x) \rfloor$, where $G_\beta(x) = \beta x - \lfloor \beta x \rfloor$. Since $x_m = (x_{m+1} + u_m)/\beta$ we get by induction for each m and $j < n$

$$\begin{aligned} x &= \sum_{i<m} u_i \beta^{-i-1} + x_m \beta^{-m} \\ x_m &= x\beta^m - \sum_{i<m} u_i \beta^{m-i-1} \\ \sigma_j(x_m) &= \sigma_j(x)\sigma_j(\beta)^m - \sum_{i<m} u_i \sigma_j(\beta)^{m-i-1} \end{aligned}$$

For $j = 0$ we have $0 \le x_m < 1$. Since $|u_i| < \beta$, for $j > 0$ we have

$$|\sigma_j(x_m)| \le |\sigma_j(x)| + \beta \sum_{i<m} |\sigma_j(\beta)|^{m-i-1} \le |\sigma_j(x)| + \frac{\beta}{1-\eta},$$

where $\eta = \max\{|\sigma_j(\beta)| : j > 0\} < 1$. There exists an integer $q > 0$ such that qx_0 is an algebraic integer. Since $x_{m+1} = \beta x_m - u_m$, it follows that each qx_m is an algebraic integer. Let $w = [w_0, \ldots, w_{n-1}]^T$ be an integral basis for $\mathbb{Z}_\beta$. Thus for each m there exist integers $x_{m,0}, \ldots x_{m,n-1}$ such that $x_m = \frac{1}{q}\sum_{j<n} x_{m,j} w_j$ and therefore $\sigma_k(x_m) = \frac{1}{q}\sum_{j<n} x_{m,j}\sigma_k(w_j)$ for each $k < n$. Denote by $S(w)$ the regular matrix $S(w)_{jk} = \sigma_k(w_j)$. Then

$$\begin{aligned} [\sigma_0(x_m), \ldots, \sigma_{n-1}(x_m)] &= \frac{1}{q}[x_{m,0}, \ldots, x_{m,n-1}] \cdot S(w) \\ [x_{m,0}, \ldots, x_{m,n-1}] &= q[\sigma_0(x_m), \ldots, \sigma_{n-1}(x_m)] \cdot S(w)^{-1} \end{aligned}$$

The right-hand side is bounded, i.e., there exists $M > 0$ such that $|\sigma_j(x_m)| < M$ for all $j < n$ and all $m \ge 0$. It follows that the left-hand side is bounded too and there exists m and $k > 0$ such that $x_{m+k} = x_m$. Thus x_i is a periodic sequence and u_i is a periodic sequence as well. □

Corollary 7.45 *If $\beta > 1$ is a Pisot number, then both the lazy and greedy partition number systems with base β have sofic expansion subshift.*

Theorem 7.46 (Schmidt [61]) *If $\beta > 1$ and each $x \in \mathbb{Q} \cap [0,1)$ has a periodic expansion in the restricted greedy system with base β, then β is a Pisot number or a Salem number.*

Proof: First we show that β is algebraic. If $0 < q < 1$ is a rational number whose expansion $u \in A^\omega$ has preperiod m and period $n > 0$, then

$$q = \sum_{i<m} u_i \beta^{-i-1} + \frac{\beta^{-m}}{1-\beta^{-n}} \sum_{i<n} u_{m+i} \beta^{-i-1}$$

which is an algebraic equation for β. Assume by contradiction that β has a conjugate $\gamma = \sigma(\beta)$ with $|\gamma| > 1$. If $x \in \mathbb{Q} \cap [0,1)$ then for $x_m = G_\beta^m(x)$ we get

$$\begin{aligned} x - \sum_{i<m} u_i \beta^{-i-1} &= x_m \cdot \beta^{-m} \\ x - \sum_{i<m} u_i \gamma^{-i-1} &= \sigma(x_m) \cdot \gamma^{-m} \\ \left| x - \sum_{i<m} u_i \gamma^{-i-1} \right| &\le \max\{|\sigma(x_m)| : m \ge 0\} \cdot |\gamma|^{-m} \end{aligned}$$

Since $\{x_m : \ m \geq 0\}$ is a periodic sequence, the right hand side converges to zero, so $x = \sum_{i=0}^{\infty} u_i \gamma^{-i-1}$. For each k there exists a rational $x \in \mathbb{Q}$ such that $\frac{1}{\beta} < x < \frac{1}{\beta} + \frac{1}{\beta^{k+1}}$. Its greedy expansion u satisfies $u_0 = 1$, $u_1 = \cdots = u_{k-1} = 0$, and, as we have proved, $x = \sum_{i=0}^{\infty} u_i \gamma^{-i-1}$. It follows

$$|\beta^{-1} - \gamma^{-1}| \ = \ \left| \sum_{i=k}^{\infty} u_i (\beta^{-i-1} - \gamma^{-i-1}) \right| \leq \frac{\beta}{\beta^k(\beta-1)} + \frac{\beta}{|\gamma|^k(|\gamma|-1)}$$

As $k \to \infty$, the right-hand side converges to zero, so we get $\beta = \gamma$ which is a contradiction. □

To get a number system for whole $\overline{\mathbb{R}}$ we add digit $\overline{0}$ and sometimes the sign digit $-$. The transformations are

$$F_a(x) = \begin{cases} \frac{x+a}{\beta} & \text{if} \quad a \in [r,s] \\ \beta x & \text{if} \quad a = \overline{0} \\ -x & \text{if} \quad a = - \end{cases}$$

Suitable subshifts depend on β and A. We identify some simple SFT for symmetric alphabets $A = [-s, s] \cup \{\overline{0}\}$.

Proposition 7.47 *Let $1 < \beta \leq 2$, $A = \{\overline{1}, 0, 1, \overline{0}\}$ and set*

$$\begin{aligned} D \ &= \ \{a\overline{0} : \ a \in \{\overline{1}, 0, 1\}\} \cup \{10^n\overline{1}, \overline{1}0^n 1 : \ n \leq \mathbf{n}_\beta\}, \text{ where} \\ \mathbf{n}_\beta \ &= \ \left\lfloor \frac{-\ln(\beta-1)}{\ln \beta} \right\rfloor . \end{aligned}$$

Then (F, Σ_D) is a number system.

Proof: The value mapping $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is clearly surjective and continuous at every $x \neq \infty$. We show that it is continuous at ∞. The smallest number in $\Phi([1])$ is $x = \Phi(10^n\overline{1})$ where $n = \mathbf{n}_\beta + 1$. Since $n > \frac{-\ln(\beta-1)}{\ln \beta}$ and therefore $\beta^n > \frac{1}{\beta-1}$, we get

$$x = \frac{1}{\beta} - \frac{1}{\beta^{n+2}} - \frac{1}{\beta^{n+3}} - \cdots = \frac{1}{\beta} - \frac{1}{\beta^{n+1}(\beta-1)} > 0$$

Similarly the largest number in $\Phi([\overline{1}])$ is $-x$, so $\Phi([\overline{0}^m]) = [x\beta^m, -x\beta^m]$. The angle length of this interval converges to 0 as $m \to \infty$, so Φ is continuous at ∞. □

For $\frac{\sqrt{5}+1}{2} < \beta \leq 2$ we get $\mathbf{n}_\beta = 0$, so the forbidden set is $D = \{\overline{1}\overline{0}, 0\overline{0}, 1\overline{0}, \overline{0}0, 1\overline{1}, \overline{1}1\}$. This case includes the binary signed system of Example 4.3. All these systems are redundant.

Proposition 7.48 *Consider a number system with $\beta > 2$ and symmetric alphabet $A = [-s, s]$ with $s \geq 1$. Assume that an integer $s_0 \leq s$ satisfies $\frac{s}{\beta-1} < s_0 \leq \frac{2s}{\beta-1}$ and define the forbidden set D by $D = \{a\overline{0} : \ a \in [-s, s]\} \cup \{\overline{0}a : \ |a| < s_0\}$. Then (F, Σ_D) is a number system.*

Proof: Set $s_0 = \lfloor \frac{s}{\beta-1} \rfloor + 1$. The intervals $[-s, -s_0]$, $[s_0, s]$ have at least one element and $\overline{0}a$ is forbidden iff $|a| < s_0$. The intervals

$$\begin{aligned} \Phi(\{u \in [-s,s]^\omega : \ -s \leq u_0 \leq -s_0\}) \ &= \ [\tfrac{-s}{\beta-1}, \tfrac{-s_0}{\beta} + \tfrac{s}{\beta(\beta-1)}] = [-b_1, -b_0] \\ \Phi(\{u \in [-s,s]^\omega : \ s_0 \leq u_0 \leq s\}) \ &= \ [\tfrac{s_0}{\beta} - \tfrac{s}{\beta(\beta-1)}, \tfrac{s}{\beta-1}] = [b_0, b_1] \end{aligned}$$

do not contain zero since $b_0 > 0$. This implies that Φ is continuous at ∞. Since $\beta b_0 \leq b_1$, the intervals $[b_0, b_1]$, $[\beta b_0, \beta b_1]$, $[\beta^2 b_0, \beta^2 b_1], \ldots$ overlap and $\Phi : \Sigma_D \to \overline{\mathbb{R}}$ is surjective. □

For $s_0 = 1$ we get the condition $2 \leq s + 1 < \beta \leq 2s + 1$. A special case is the ternary signed system from Example 4.4 with $s = 1$, $\beta = 3$.

7.7 Positional arithmetic

In positional number systems the arithmetical algorithms are simplified since these systems consist of **linear transformations** of the form $M(x) = ax + b$. A Möbius transformation $M = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is linear iff $M(\infty) = \infty$ iff $c = 0$. Linear transformations form a subgroup of $\mathbb{M}(\mathbb{R})$. If the unary algorithm computes a linear transformation $M(x) = ax + b$ in a positional number system, all $F_v^{-1} M F_u$ are linear transformations. Similarly if a binary algorithm computes a bilinear tensor of the form $T(x, y) = T_0xy + T_1x + T_2y + T_3$, then all compositions T^*F_u, T_*F_v, $F_w^{-1}T$ are of this form. We show that in this case the arithmetical algorithms work with a bounded delay. This means that there exists $\delta > 0$ such that the prefix $w_{[0,n)}$ of the result depends on the prefixes $u_{[0,n+\delta)}$, $v_{[0,n+\delta)}$ of the operands. Suppose we add two numbers in a positional system from Proposition 7.47 or Proposition 7.48. If the operands are $u = v = \overline{0}^\omega$ then the addition algorithm does not give any output since $\infty + \infty$ is an indeterminate expression. If $u = \overline{0}^\omega$ and v has a prefix $\overline{0}^n$, then after reading the first letter of v different from $\overline{0}$, the algorithm starts writing $\overline{0}$ to the output and in infinite time outputs $\overline{0}^\omega$. If the operands are $\overline{0}^n u$, $\overline{0}^m v$, where $u, v \in [-s, s]^+$ do not contain $\overline{0}$, then the output is in the form $\overline{0}^p w$, where $w \in [-s, s]^+$ and $|w| + \delta \geq \min\{|u|, |v|\}$. With multiplication, the situation is similar. Since $\infty \cdot \infty = \infty$, the inputs $u = v = \overline{0}^\omega$ yield $\overline{0}^\omega$ but the input $u = 0^\omega$, $v = \overline{0}^\omega$ does not give any output since $0 \cdot \infty$ is an indeterminate expression. In all other cases, the algorithm works eventually with a bounded delay. To get this result we work with the Euclidean length of an interval $I = [a, b] \subset \mathbb{R}$ which we denote by $|I|_e = b - a$. If $F_a = \frac{x+a}{\beta}$ is a transformation of a positional number system, then $|F_a(I)|_e = |I|_e/\beta$ and $|F_a^{-1}(I)|_e = \beta|I|_e$.

Lemma 7.49 *Let $T(x, y) = T_0xy + T_1x + T_2y + T_3$ be a bilinear tensor, $K > 0$ and let $I, J \subset [-K, K]$ be bounded intervals. Then*

$$|T(I, J)|_e \leq (|T_0|K + |T_1|) \cdot |I|_e + (|T_0|K + |T_2|) \cdot |J|_e$$

Proof: If $x_0, x_1 \in I$, $y_0, y_1 \in J$, then

$$\begin{aligned} |T(x_1, y_1) - T(x_0, y_0)| &\leq |T_0x_1(y_1 - y_0) + T_0(x_1 - x_0)y_0 + T_1(x_1 - x_0) + T_2(y_1 - y_0)| \\ &\leq (|T_0|K + |T_1|) \cdot |x_1 - x_0| + (|T_0|K + |T_2|) \cdot |y_1 - y_0| \end{aligned}$$

and the result follows. □

Proposition 7.50 *Assume that the binary algorithm works with a greedy selector in a redundant positional number system from Proposition 7.47 or Proposition 7.48 with symmetric alphabet $A = [-s, s] \cup \{\overline{0}\}$. Assume that the initial tensor is of the form $T(x, y) = T_0xy + T_1x + T_2y + T_3$. Then there exists a delay $\delta > 0$ such that if $(T, \mathbf{i}, \mathbf{i}, \mathbf{i}) \overset{u,v,w}{\longrightarrow} (X, p, q, r)$ and $u, v \in [-s, s]^*$ (no prefix of $\overline{0}$) then $|w| + \delta \geq \min\{|u|, |v|\}$.*

Proof: Since the inputs u, v do not contain $\overline{0}$, there exists $K > 0$ such that $V_p, V_q \subseteq [-K, K]$. Since $|F_u V_p|_e = |V_p|_e \cdot \beta^{-|u|}$, $|F_v V_q|_e = |V_q|_e \cdot \beta^{-|v|}$, there exists $C > 0$ such that

$$\begin{aligned} |T(F_u V_p, F_v V_q)|_e &\leq (|T_0|K + |T_1|) \cdot |V_p|_e \cdot \beta^{-|u|} + (|T_0|K + |T_2|) \cdot |V_q|_e \beta^{-|v|} \\ &\leq C \cdot \max\{\beta^{-|u|}, \beta^{-|v|}\} \\ |X|_e &= |F_w^{-1} T(F_u V_p, F_v V_q)|_e \leq C \cdot \beta^{|w| - \min\{|u|,|v|\}} \end{aligned}$$

If $|w|+d \leq \min\{|u|,|v|\}$, then $|X|_e \leq C\cdot\beta^{-d}$. There exists δ such that $C\cdot\beta^{-\delta+1}$ is less than the Lebesgue number (overlap) of the number system. It follows that if $(T,\mathbf{i},\mathbf{i},\mathbf{i}) \stackrel{u,v,w}{\longrightarrow} (X,p,q,r)$ and $|w|+\delta-1 \leq \min\{|u|,|v|\}$, then (X,p,q,r) is an emission state and the result follows. □

By Proposition 7.50, the addition in positional number systems works with a finite delay. For certain algebraic and integer bases β and sufficiently redundant alphabets we have a stronger result - the existence of a parallel addition algorithm with a delay $\delta > 0$. For given inputs $u,v \in \Sigma_G$ the algorithm computes $w \in \Sigma_G$ such that $\Phi(w) = \Phi(u)+\Phi(v)$ and w_i depends only on $u_{[i+\delta)}$ and $v_{[i+\delta)}$ (and not on the prefixes $u_{[0,i)}$ and $v_{[0,i)}$). Assume that $\beta > 1$ is a base, $A = [r,s] \subset \mathbb{Z}$ is an interval of integers with $r \leq 0 \leq s$ and $\Phi(x) = \sum_{i\geq 0} x_i\beta^{-i-1}$. For $x,y \in A^\omega$ we have $\Phi(x)+\Phi(y) = \Phi(z)$, where

$$z_i = x_i + y_i \in A+A = \{a+b:\ a,b\in A\} = [2r,2s].$$

To obtain an addition algorithm for the alphabet A we have to reduce $z \in (A+A)^*$ to w in the alphabet A with the same value $\Phi(w) = \Phi(z)$. However, because of the carry overs, the expansion of w may start already at position -1. We consider therefore larger spaces of symbolic sequences which start at arbitrary integer. Denote by

$$A^{*\omega} \ = \ \{x \in A^{\mathbb{Z}}:\ \exists k\in\mathbb{Z}, \forall i<k, x_i = 0\}.$$

For each finite alphabet $A \subset \mathbb{Z}$ and $\beta > 1$ the value map $\Phi_\beta : A^{*\omega} \to \mathbb{R}$ is given by $\Phi_\beta(x) = \sum_{i\in\mathbb{Z}} x_i\beta^{-i}$. We consider a reduction from a finite alphabet $B \subset \mathbb{Z}$ to A given by a a **sliding block code**. This is a mapping $F : B^{*\omega} \to A^{*\omega}$ given by $F(z)_i = f(z_{[i+l,i+r]})$ where $f : A^{r-l+1} \to A$ is a **local rule** which fixes zero, i.e., $f(0,\ldots,0) = 0$.

Definition 7.51 *Let $\beta > 1$ and $0 \in A \subset B \subset \mathbb{Z}$ be finite interval alphabets. We say that a sliding block code $F : B^{*\omega} \to A^{*\omega}$ performs a* **parallel reduction** *if $\Phi_\beta(F(x)) = \Phi_\beta(x)$ for every $x \in B^{*\omega}$.*

Proposition 7.52 (Avizienis [1]) *The Avizienis addition algorithm works for an integer base $\beta \geq 3$ and alphabet $A = [-a,a]$, where $\frac{\beta}{2} < a \leq \beta-1$. Given inputs $x,y \in A^{*\omega}$, the algorithm computes in parallel $w_i = x_i+y_i$ and then determines the quotients q_i and remainders r_i by the rule*

$$\begin{array}{rcccl} & w_i & \leq & -a & \Rightarrow \quad q_i=-1,\ \ r_i = w_i+\beta \\ -a & < & w_i & < & a \quad \Rightarrow \quad q_i=0,\ \ r_i = w_i \\ a & \leq & w_i & & \Rightarrow \quad q_i=1,\ \ r_i = w_i-\beta \end{array}$$

The block code $F : A^{*\omega}\times A^{*\omega} \to A^{*\omega}$ defined by $F(x,y)_i = z_i = r_i+q_{i+1}$ satisfies $\Phi_\beta(x)+\Phi_\beta(y) = \Phi_\beta(F(x,y))$.

Proof: We have $x_i+y_i = w_i = \beta q_i + r_i$. If $n = \min\{i\in\mathbb{Z}:\ x_i \neq 0 \text{ or } y_i \neq 0\}$, then

$$\Phi_\beta(x)+\Phi_\beta(y) \ = \ \sum_{i\geq n}(\beta q_i+r_i)\beta^{-i} = \sum_{i\geq n-1}(q_{i+1}+r_i)\beta^{-i} = \Phi_\beta(F(x,y))$$

We show that $F(x,y) \in A^{*\omega}$. If $-2a \leq w_i \leq -a$ then

$$a < -2a+\beta \leq r_i \leq -a+\beta < a$$

Since $|q_{i+1}| \leq 1$ we get $|z_i| \leq a$. If $-a < w_i < a$, then $|z_i| = |w_i| \leq a$. If $a \leq w_i \leq 2a$ then $-a < a-\beta \leq r_i \leq 2a-\beta < a$, so $|z_i| \leq |r_i|+|q_{i+1}| \leq a$. Thus $F(x,y) \in A^{*\omega}$. □

Proposition 7.53 (Chow and Robertson [8]) *The Chow and Robertson algorithm works for an even integer base $\beta = 2a$ and the alphabet $A = [-a, a]$. Given inputs $x, y \in A^{*\omega}$, the algorithm computes in parallel $w_i = x_i + y_i$ and then determines the quotients q_i and the remainders r_i by the rule*

$$\begin{array}{rcccccl} -\beta & \le & w_i & < & -a & \Rightarrow & q_i = -1, \quad r_i = w_i + \beta \\ -a & < & w_i & < & a & \Rightarrow & q_i = 0, \quad r_i = w_i \\ a & < & w_i & \le & \beta & \Rightarrow & q_i = 1, \quad r_i = w_i - \beta \\ & & w_i & = & -a, \quad w_{i+1} < 0 & \Rightarrow & q_i = -1, \quad r_i = a \\ & & w_i & = & -a, \quad w_{i+1} \ge 0 & \Rightarrow & q_i = 0, \quad r_i = -a \\ & & w_i & = & a, \quad w_{i+1} \le 0 & \Rightarrow & q_i = 0, \quad r_i = a \\ & & w_i & = & a, \quad w_{i+1} > 0 & \Rightarrow & q_i = 1, \quad r_i = -a \end{array}$$

The block code $F : A^{*\omega} \times A^{*\omega} \to A^{*\omega}$ defined by $F(x, y)_i = z_i = r_i + q_{i+1}$ satisfies $\Phi_\beta(x) + \Phi_\beta(y) = \Phi_\beta(F(x, y))$.

Proof: We have $x_i + y_i = w_i = \beta q_i + r_i$, so $\Phi_\beta(F(x, y)) = \Phi_\beta(x) + \Phi_\beta(y)$. We show that $F(x, y) \in A^{*\omega}$. In each case we have $|q_i| \le 1$. If $|w_i| \le a - 1$ or $|w_i| \ge a + 1$ then $|r_i| < a$, so $|z_i| \le a$. If $w_i = -a$, $w_{i+1} < 0$, then $q_{i+1} \le 0$ and $z_i = a + q_{i+1} \le a$. If $w_i = -a$, $w_{i+1} \ge 0$, then $q_{i+1} \ge 0$ and $z_i = -a + q_{i+1} \ge -a$. The proof is similar for $w_i = a$. □

For noninteger bases $\beta > 1$ we show that a sliding block code exists iff β is algebraic and no conjugate of β lies on the unit circle in the complex plane.

Proposition 7.54 *Let $0 \in A \subset B \subset \mathbb{Z}$. If there exists a reduction from B to A then β is an algebraic number.*

Proof: Let $F(z)_i = f(z_{[i+l,i+r]})$, where $f : A^{r-l+1} \to A$ is a local rule. Choose $b \in B \setminus A$ and set $x = b^\omega \in B^\omega \subset B^{*\omega}$, so $x_i = 0$ for $i < 0$. Denote by $a = f(b, \ldots, b) \in A$. For $y = F(x) \in A^{*\omega}$ we have $y_i = 0$ for $i < -r$ and $y_i = a$ for $i \ge -l$. We get

$$\frac{b\beta}{\beta - 1} = \Phi_\beta(x) = \Phi_\beta(y) = \sum_{i=-r}^{-l-1} y_i \beta^{-i} + \beta^{l+1} \frac{a}{\beta - 1}$$

or $b\beta = (\beta - 1)(y_{-l-1}\beta^{l+1} + \cdots + y_{-r}\beta^r) = a\beta^{l+1}$. If $n \ge l + 1$ is the largest integer with nonzero y_n, we get an algebraic equation for β of degree $n + 1$. If all y_n are zero, we get $a = b\beta^l$. Since $a \ne b$, we have $l > 0$ and we get an algebraic equation of degree l. □

Assume now that $\beta > 1$ is an algebraic number which is a root of a polynomial $p \in \mathbb{Z}[x]$. This need not be the minimal polynomial of β. The reduction algorithm presupposes special properties of p which are not always satisfied by the minimal polynomial. If $p(\beta) = 0$, then β is also a root of $x^k p(x)$ for any integer k, and we assume p in the form

$$p(x) = \sum_{i=l}^{r} p_i x^{-i} = \sum_{i \in \mathbb{Z}} p_i x^{-i},$$

where $p_i \in \mathbb{Z}$ for $l \le i \le r$ and $p_i = 0$ for $i \in \mathbb{Z} \setminus [l, r]$. If $\{q_i \in Z : \ i \in \mathbb{Z}\}$ is a bounded sequence of integers and $y_i = \sum_{k \in \mathbb{Z}} q_k p_{i+k}$ then for $\gamma = \sum_i y_i \beta^{-i}$ we get

$$\Phi(\gamma) \ = \ \sum_{k \in \mathbb{Z}} \sum_{i \in \mathbb{Z}} q_k \cdot p_{i+k} \beta^{-i} = \sum_{k \in \mathbb{Z}} \sum_{j \in \mathbb{Z}} q_k \cdot p_j \beta^{k-j} = \sum_{k \in \mathbb{Z}} q_k \beta^k \sum_{j \in \mathbb{Z}} p_j \beta^{-j} = 0$$

The reduction algorithm from B to $A \subset B$ chooses an appropriate sequence of q_i and for $x \in B^{*\omega}$ computes

$$F(x)_i = x_i - \sum_{k\in\mathbb{Z}} q_k p_{i+k}$$

For suitable sequences of q_i we get $F(x) \in A^{*\omega}$. For this purpose we need polynomials whose zeroth coefficient is sufficiently large. We say that a polynomial $p \in \mathbb{Z}[x]$ has a **dominant coefficient** p_n if $p_n \geq \sum_{i\neq n} |p_i|$. Then $p(x)x^{-n}$ is a polynomial with dominant zero coefficient. Assume that $\beta > 1$ is a root of a polynomial $p(x) = \sum_{i=l}^{r} p_i x^{-i}$, where $l \leq 0 \leq r$, with a dominant zero coefficient, so $p_0 > \sum_{i\neq 0} |p_i| = s$. We construct an addition algorithm in the alphabet $A = [-a-s, a+s]$ where $a = \lceil \frac{p_0-1}{2} \rceil$. Given inputs $x, y \in A^{*\omega}$, the algorithm first computes $z_i = x_i + y_i$ which is a word in the alphabet $A_0 = [-2a, 2a]$. Then it performs a series of reduction steps to a smaller alphabet until the reduced word in the alphabet A is attained.

Theorem 7.55 (Frougny et al. [18]) *Let p be a polynomial with dominant zero coefficient $p_0 > \sum_{i\neq 0} |p_i| = s$. Denote by $a = \lceil \frac{p_0-1}{2} \rceil$. If $B = [-b, b]$, where $b > a + s$, then the reduction algorithm $w \mapsto z$ given by*

$$z_i = w_i - \sum_{j\in\mathbb{Z}} q_{i+j}p_j, \text{ where } q_i = \begin{cases} 0 & \text{if } |w_i| \leq a \\ \operatorname{sgn}(w_i) & \text{otherwise} \end{cases}$$

reduces a word from $B^{*\omega}$ to a word in alphabet $C^{*\omega}$, where $C = [-c, c]$ and $c < b$.

Proof: Since the subtraction of $\sum_j q_{i+j}p_j$ does not change the value of the expansion, we have only to show that the result z belongs to a smaller alphabet. We have $|q_j| \leq 1$. If $|w_i| \leq a$ then $q_i = 0$ so $|z_i| \leq |w_i| + \sum_{j\neq 0} |p_j| \leq a + s < b$. Assume $a < w_i \leq b$. Since $p_0 \leq 2a + 1$ we get

$$\begin{aligned} z_i &= w_i - p_0 - \sum_{j\neq 0} q_{i+j}p_j \\ &\geq w_i - p_0 - s \geq (a+1) - (2a+1) - s = -a - s > -b \\ z_i &\leq w_i - p_0 + s \leq b - p_0 + s < b \end{aligned}$$

so $|z_i| < b$. If $w_i < -a$, the proof is analogous. □

Thus the repeated application of the reduction algorithm from Theorem 7.55 to $w = x + y$, where $x, y \in A^{*\omega}$ gives finally a word of $A^{*\omega}$.

Theorem 7.56 (Frougny et al. [18]) *An algebraic number $\alpha > 1$ is a root of a polynomial $p \in \mathbb{Z}[x]$ with a dominant coefficient iff $|\gamma| \neq 1$ for each conjugate γ of α.*

Proof: If $p(\alpha) = 0$ and γ is a conjugate of α with $|\gamma| = 1$ then $p(\gamma) = 0$ and for each $k \leq \deg(p)$ we have

$$|p_k| = |p_k\gamma^k| = \left| \sum_{j\neq k} p_j\gamma^j \right| \leq \sum_{j\neq k} |p_j|$$

so no coefficient is dominant. Conversely let

$$p(x) = -p_0 - p_1 x - \cdots - p_{n-1}x^{n-1} + x^n = \prod_{i<n} (x - \alpha_i)$$

be the minimal polynomial of α, so $p_i \in \mathbb{Q}$ and $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ are the conjugates of α. Since p is irreducible, we have $\alpha_i \neq \alpha_j$ for $i \neq j$. Assume that $|\alpha_i| \neq 1$ for each i. For each $m > 0$ consider the polynomial

$$q(m)(x) = \prod_{i<n}(x - \alpha_i^m) = -q(m)_0 - \cdots - q(m)_{n-1}x^{n-1} + x^n.$$

Then

$$\begin{aligned}
\sum_i \alpha_i^m &= q(m)_{n-1} \\
\sum_{i<j} \alpha_i^m \alpha_j^m &= -q(m)_{n-2} \\
\sum_{i<j<k} \alpha_i^m \alpha_j^m \alpha_k^m &= q(m)_{n-3} \\
&\vdots \\
\alpha_0^m \cdots \alpha_{n-1}^m &= (-1)^{n+1} q(m)_0
\end{aligned}$$

For

$$M(\alpha) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ p_0 & p_1 & p_2 & \cdots & p_{n-1} \end{bmatrix}$$

we have $\det(xI - M(\alpha)) = p(x)$, so the eigenvalues of $M(\alpha)$ are α_i, and the eigenvalues of $M(\alpha)^m$ are α_i^m. It follows $q(m)(x) = \det(xI - M(\alpha)^m)$, so $q(m) \in \mathbb{Q}[x]$. Reorder now the α_i so that $|\alpha_0| \geq |\alpha_1| \geq \cdots \geq |\alpha_{n-1}|$ and let k be the first index with $|\alpha_k| < 1$. If $|\alpha_i| > 1$ for all i, then we set $k = n$. For each subset $\{i_0, i_1, \ldots, i_j\} \subseteq \{0, 1, \ldots, n-1\}$ different from $\{0, 1, \ldots, k-1\}$ we have

$$\left|\frac{\alpha_{i_0} \cdots \alpha_{i_j}}{\alpha_0 \cdots \alpha_{k-1}}\right| < 1, \Rightarrow \lim_{m\to\infty} \left|\frac{\alpha_{i_0}^m \cdots \alpha_{i_j}^m}{\alpha_0^m \cdots \alpha_{k-1}^m}\right| = 0 \Rightarrow \lim_{m\to\infty} \frac{\sum_{j\neq k} |q(m)_j|}{|q(m)_k|} = 0$$

so $q(m)_j$ is a dominant coefficient of $q(m)$ for sufficiently large m. Thus the polynomial $q(m)(x^m)$ has a dominant coefficent and a root α. □

Theorem 7.57 (Frougny et al. [17]) *Let $\beta > 1$be an algebraic number which has a conjugate γ with $|\gamma| = 1$. If $0 \in A \subset \mathbb{Z}$ is an alphabet and $B = A + A$ then there exists no parallel reduction from $B^{*\omega}$ to $A^{*\omega}$.*

Proof: Assume by contradiction that the reduction is performed by a sliding block code $F(x)_i = f(x_{[i+l,i+r]})$, where $f : B^{r-l+1} \to A$ is a local rule with $l < 0 < r$. Denote by

$$S = \max\left\{ \left| \sum_{j=0}^{\max(-l,r)} a_j \gamma^j \right| : \ a_j \in A \right\}$$

If $\gamma^k = 1$, then the minimal polynomial of β divides $x^k - 1$ which is impossible since $\beta > 1$. Thus γ is not a root of unity. It follows that there exists an infinite number of indices i with $\Re(\gamma^i) > \frac{1}{2}$, and there exists $m > 0$ and $\{\varepsilon_j \in \{0,1\} : \ j \le m\}$ such that $\Re(\sum_{j=0}^m \varepsilon_j \gamma^j) > 3S$. Set

$$T = \max\left\{\left|\Re\left(\sum_{j=0}^m a_j\gamma^j\right)\right| : \ a_j \in A\right\}$$

so $T \ge 3S$. Take $x_j \in A$ such that $T = |\Re(\sum_{j=0}^m x_j\gamma^j)|$ and set $x = \sum_{j=0}^m x_j\beta^j$. Thus $T = |\Re(\varphi(x)|$, where $\varphi : \mathbb{Q}(\beta) \to \mathbb{Q}(\gamma)$ is the field homomorphism with $\varphi(\beta) = \gamma$. The sliding block code yields $z_j = F(x+x)_j \in A$ such that

$$\begin{aligned} x + x &= \sum_{j=-r}^{-1} z_j\beta^j + \sum_{j=0}^{m} z_j\beta^j + \sum_{j=m+1}^{m-l} z_j\beta^j \\ \varphi(x) + \varphi(x) &= \sum_{j=-r}^{-1} z_j\gamma^j + \sum_{j=0}^{m} z_j\gamma^j + \sum_{j=m+1}^{m-l} z_j\gamma^j \end{aligned}$$

We get

$$\begin{aligned} |\Re(\varphi(x))| + 3S &\le |\Re(\varphi(x))| + |\Re(\varphi(x))| = |\Re(\varphi(x) + \varphi(x))| \\ &\le |\Re(\textstyle\sum_{j=1}^{r} z_{j-r-1}\gamma^j)| + |\Re(\sum_{j=0}^{m} z_j\gamma^j)| + |\Re(\sum_{j=1}^{-l} z_{j+m}\gamma^j)| \\ &\le S + |\Re(\varphi(x))| + S \end{aligned}$$

and this is a contradiction. □

Corollary 7.58 *If $\beta > 1$ is an algebraic number, then there exists an alphabet $A = [r, s]$ and a parallel addition algorithm $F : A^{*\omega} \times A^{*\omega} \to A^{*\omega}$ iff β has no conjugate γ with $|\gamma| = 1$.*

The problem of finding the smallest alphabet A for which there exists an addition algorithm is treated in Frougny et al [19]. For the golden mean $\beta = \frac{\sqrt{5}-1}{2}$ we get $q(x) = x^4 - 3x^2 + 1$ with dominant coefficient q_2 or $q(x) = -x^2 + 3 - x^{-2}$ with dominant coefficient q_0. Thus $p_0 = 3$, $s = 2$, $a = 1$, so we get an addition algorithm in the alphabet $A = [-3, 3]$. The algorithm subtracts the word $\overline{1}030\overline{1}$ at any position i with $w_i > 1$ and adds it at any position i with $w_i < -1$. In this way the algorithm successively reduces a word in alphabet $[-6, 6]$ to alphabets $[-5, 5]$, $[-4, 4]$ and $[-3, 3]$. There exists also a more sophisticated addition algorithm in the alphabet $[-1, 1]$ (see Frougny et al [19]).

Chapter 8

Transcendent and iterative algorithms

Transcendent functions such as e^x or $\sin x$ can be expressed by power series, so they may be approximated by polynomials. However, better approximations can be obtained from a sequence of rational functions called Padé approximants (see Wall [67], Baker and Graves-Morris [2] or Jones and Thron [27]). Exact real algorithms for transcendent functions are based on these approximations.

8.1 Padé approximants

Padé approximants are rational functions derived from a power series $f(x) = c_0+c_1x+c_2x^2+\cdots$ which is treated as a **formal power series**: the questions of convergence are postponed. Formal power series can be added, subtracted and multiplied and they form a ring. The order $\lambda(f)$ of a formal power series $f(x) = \sum_{n\geq 0} c_n x^n$ is the least n such that $c_n \neq 0$. Clearly

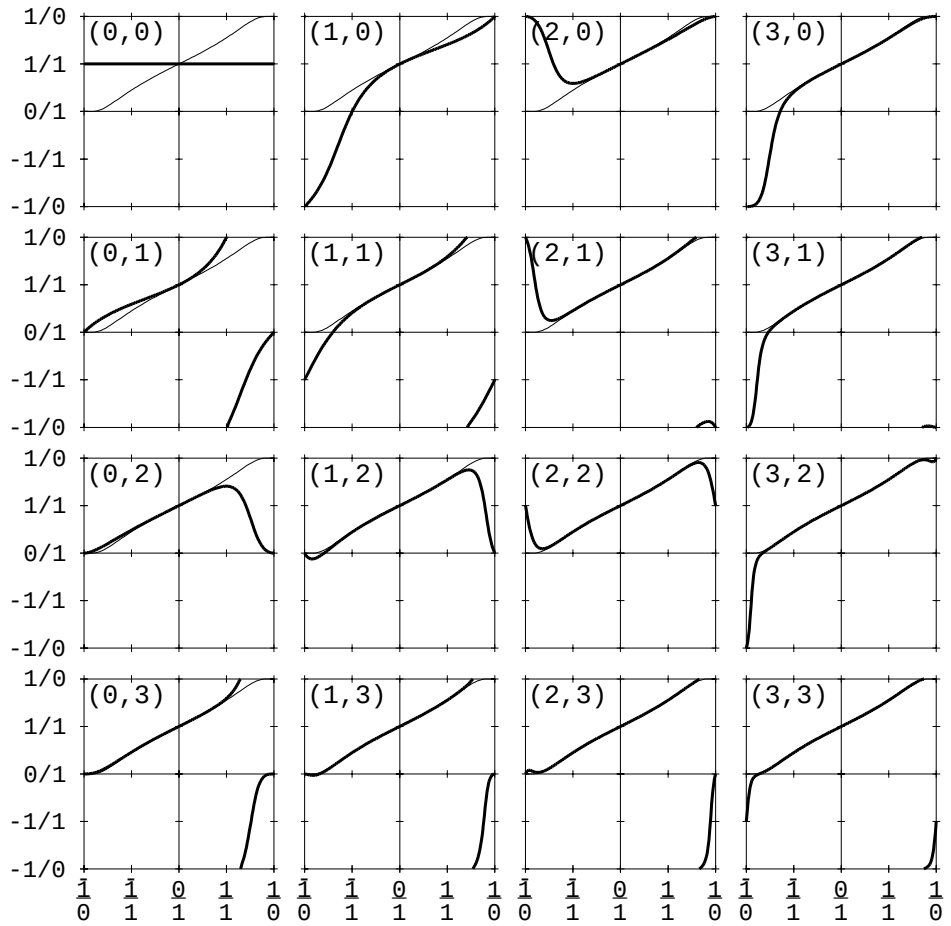
$$\begin{aligned} \lambda(f+g) &\geq \min\{\lambda(f), \lambda(g)\}, \\ \lambda(fg) &= \lambda(f) + \lambda(g). \end{aligned}$$

A **rational expression** is a pair (p, q) of polynomials. Rational expressions are equivalent ($(p_0, q_0) \sim (p_1, q_1)$) if $p_0q_1 = p_1q_1$. For each rational expression $r = (p, q)$ there exists a unique rational function $R(x) = \frac{P(x)}{Q(x)}$ such that $(P, Q) \sim (p, q)$. R is obtained by cancelling the common factors of p and q.

Definition 8.1 *Let f be a formal power series and $m, n \geq 0$ integers. We say that a rational expression $r_{m,n}(x) = (p_{m,n}(x), q_{m,n}(x))$ is the* **Padé approximant expression** *of f of order (m, n) if $\deg(p_{m,n}) \leq m$, $\deg(q_{m,n}) \leq n$ and $\lambda(fq_{m,n} - p_{m,n}) \geq m + n + 1$. A regular rational function $R_{m,n}(x) = \frac{P_{m,n}(x)}{Q_{m,n}(x)}$ is the* **Padé approximant** *of f of order (m, n) if it is equivalent to a Padé approximant expression of f of order (m, n).*

Proposition 8.2 *Each formal power series has Padé approximants of all orders. Two Padé approximant expressions of the same order are equivalent.*

Proof: Let $f(x) = c_0 + c_1x + c_2x^2 + \cdots$. We search for polynomials $p(x) = a_0 + \cdots + a_mx^m$, $q(x) = b_0 + \cdots + b_nx^n$, such that $\lambda(fq - p) \geq m + n + 1$. This condition gives a system of



	0	1	2	3
0	$\frac{1}{1}$	$\frac{x+1}{1}$	$\frac{x^2+2x+2}{2}$	$\frac{x^3+3x^2+6x+6}{6}$
1	$\frac{1}{-x+1}$	$\frac{x+2}{-x+2}$	$\frac{x^2+4x+6}{-2x+6}$	$\frac{x^3+6x^2+18x+24}{-6x+24}$
2	$\frac{2}{x^2-2x+2}$	$\frac{2x+6}{x^2-4x+6}$	$\frac{x^2+6x+12}{x^2-6x+12}$	$\frac{x^3+9x^2+36x+60}{3x^2-24x+60}$
3	$\frac{6}{-x^3+3x^2-6x+6}$	$\frac{6x+24}{-x^3+6x^2-18x+24}$	$\frac{3x^2+24x+60}{-x^3+9x^2-36x+60}$	$\frac{x^3+12x^2+60x+120}{-x^3+12x^2-60x+120}$

Figure 8.1: Padé approximants of orders (m, n) with $0 \le m, n \le 3$ (thick) of the exponential function (thin) $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.

equations for the unknowns a_i and b_i:

$$\begin{aligned} c_0 b_0 &= a_0 \\ c_1 b_0 + c_0 b_1 &= a_1 \\ &\vdots \\ c_m b_0 + c_{m-1} b_1 + \cdots + c_0 b_m &= a_m \end{aligned}$$

$$\begin{aligned} c_{m+1} b_0 + c_m b_1 + \cdots + c_{m-n+1} b_n &= 0 \\ c_{m+2} b_0 + c_{m+1} b_1 + \cdots + c_{m-n+2} b_n &= 0 \\ &\vdots \\ c_{m+n} b_0 + c_{m+n-1} b_1 + \cdots + c_m b_n &= 0. \end{aligned}$$

	0	1	2	3
0	$\frac{0}{1}$	$\frac{x}{1}$	$\frac{x}{1}$	$\frac{3x-x^3}{3}$
1	$\frac{0}{x}$	$\frac{x}{1}$	$\frac{x^2}{x}$	$\frac{3x-x^3}{x}$
2	$\frac{0}{x^2}$	$\frac{3x}{3+x^2}$	$\frac{3x}{3+x^2}$	$\frac{15x+4x^3}{15+9x}$
3	$\frac{0}{x^3}$	$\frac{3x}{3+x^2}$	$\frac{3x^2}{3x+x^3}$	$\frac{15x+4x^3}{15+9x}$

	0	1	2	3
0	$\frac{0}{1}$	$\frac{x}{1}$	$\frac{x}{1}$	$\frac{3x-x^3}{3}$
1	$\frac{0}{1}$	$\frac{x}{1}$	$\frac{x}{1}$	$\frac{3x-x^3}{3}$
2	$\frac{0}{1}$	$\frac{3x}{3+x^2}$	$\frac{3x}{3+x^2}$	$\frac{15x+4x^3}{15+9x}$
3	$\frac{0}{1}$	$\frac{3x}{3+x^2}$	$\frac{3x}{3+x^2}$	$\frac{15x+4x^3}{15+9x}$

Table 8.1: The Padé approximant expressions (left) and Padé approximants (right) of $f(x) = \arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$.

where $c_k = 0$ for $k < 0$ and $b_k = 0$ for $k > n$. The homogeneous system of the last n equations in $n+1$ unknowns $b_0, \ldots b_n$ has a nonzero solution and the first $m+1$ equations then determine the a_i, so we obtain p, q with $\lambda(fq - p) \geq m + n + 1$. Assume that p_1, q_1 are other polynomials of degree at most m, n such that $\lambda(fq_1 - p_1) \geq m + n + 1$. Then $\lambda(fqq_1 - pq_1) \geq m + n + 1$ and $\lambda(fqq_1 - p_1q) \geq m + n + 1$. Since $fqq_1 = fq_1q$, we get $\lambda(pq_1 - p_1q) \geq m + n + 1$. Since this is a polynomial of degree $m + n$, we get $pq_1 = p_1q$. □

If $c_0 \neq 0 \neq c_1$ then

$$\begin{aligned} R_{0,0}(x) &= \tfrac{c_0}{1}, & R_{1,0}(x) &= \tfrac{c_0+c_1x}{1} \\ R_{1,0}(x) &= \tfrac{c_0^2}{c_0-c_1x}, & R_{1,1}(x) &= \tfrac{c_0c_1+(c_1^2-c_0c_2)x}{c_1-c_2x} \end{aligned}$$

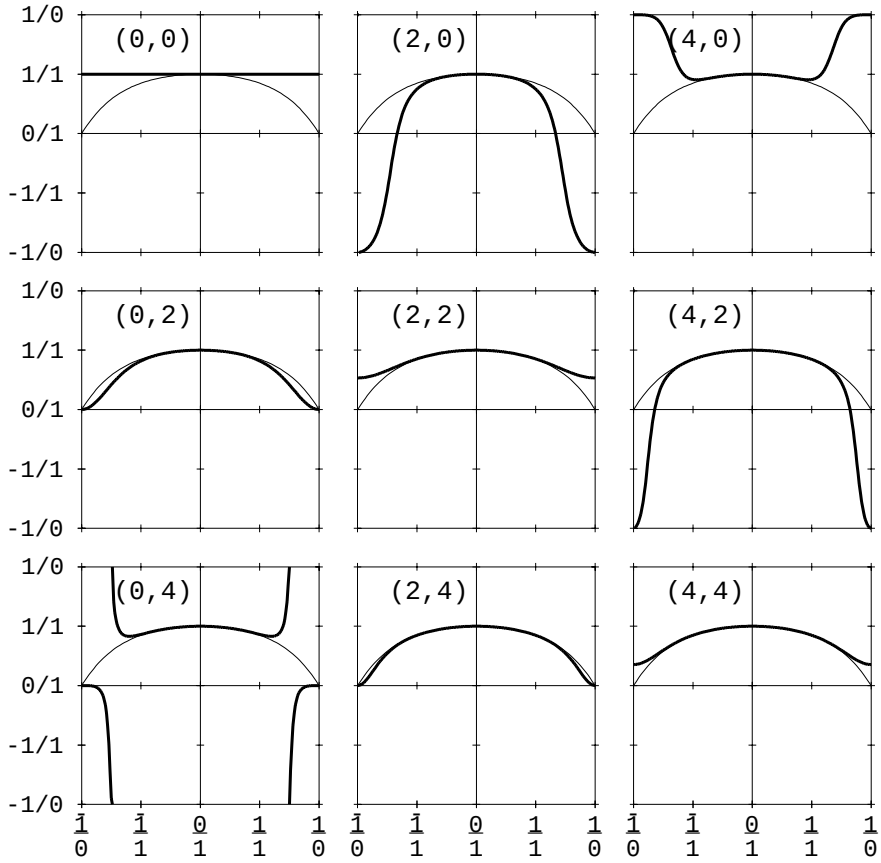
Padé approximants of the exponential functions are all different, so $R_{m,n} = \frac{P_{m,n}}{Q_{m,n}} = \frac{p_{m,n}}{q_{m,n}}$ (Figure 8.1). In the power series of $\arctan x$ there are only odd powers of x and the relation between $R_{m,n}$ and $r_{m,n}$ is more complicated (see Table 8.1). Note that the Padé approximants $R_{m,n} = \frac{P_{m,n}}{Q_{m,n}}$ do not necessarily satisfy the condition $\lambda(fQ_{m,n} - P_{m,n}) \geq m + n + 1$. For example for the Padé approximant $R_{0,1}(x) = \frac{0}{1}$ we have $\lambda(fQ - P) = \lambda(f) = 1$. If some powers are missing in the power series f, then the Padé table contains square blocks of identical rational functions:

Theorem 8.3 (Block Theorem) *Let f be a formal power series and let $R = \frac{P}{Q}$ be a regular rational function with $\deg(P) = m \geq 0$, $\deg(Q) = n \geq 0$, $\lambda(Qf - P) = m + n + r + 1$, where $0 \leq r$. Then $R_{i,j} = R$ iff $m \leq i \leq m + r$ and $n \leq j \leq n + r$. Moreover, $r_{m,j} = (P, Q) = r_{i,n}$ in this case.*

Proof: If $i < m$ or $j < n$ then $\deg(P_{i,j}) < m$ or $\deg(Q_{i,j}) < n$, so $R_{i,j} \neq R$. Assume that $R_{m+i,n+j} = R$ for some $i, j \geq 0$ such that either $i > r$ or $j > r$. Then $r_{m+i,n+j} = (PS, QS)$, where S is a polynomial which satisfies $\deg(PS) \leq m+i$, $\deg(QS) \leq n+j$ so $\deg(S) \leq \min\{i, j\}$. On the other hand we have $\lambda(fQS - PS) \geq m+n+i+j+1$. Since $\lambda(Qf - P) = m+n+r+1$, we get $\deg(S) \geq i + j - r$ which is either greater than j if $i > r$ or greater than i if $j > r$. This is a contradiction, so we have proved that if $R_{i,j} = R$ then $m \leq i \leq m + r$ and $n \leq j \leq n + r$. Since

$$\begin{aligned} \deg(p_{m+i,n+j}) &\leq m + i, \\ \deg(q_{m+i,n+j}) &\leq n + j, \\ \lambda(fq_{m+i,n+j} - p_{m+i,n+j}) &= m + n + r + 1 + \min\{i, j\} \geq m + n + i + j + 1 \end{aligned}$$

we have $r_{m+i,n+j} = (p_{m+i,n+j}, q_{m+i,n+j}) = (P(x)x^{\min\{i,j\}}, Q(x)x^{\min\{i,j\}})$ for $0 \leq i, j \leq r$. □



	0	1	2	3	4
0	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{3-x^2}{3}$	$\frac{3-x^2}{3}$	$\frac{15-5x^2+3x^4}{15}$
1	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{3-x^2}{3}$	$\frac{3-x^2}{3}$	$\frac{15-5x^2+3x^4}{15}$
2	$\frac{3}{3+x^2}$	$\frac{3}{3+x^2}$	$\frac{15+4x^2}{15+9x^2}$	$\frac{15+4x^2}{15+9x^2}$	$\frac{105+40x^2-4x^4}{105+75x^2}$
3	$\frac{3}{3+x^2}$	$\frac{3}{3+x^2}$	$\frac{15+4x^2}{15+9x^2}$	$\frac{15+4x^2}{15+9x^2}$	$\frac{105+40x^2-4x^4}{105+75x^2}$
4	$\frac{45}{45+15x^2-4x^4}$	$\frac{45}{45+15x^2-4x^4}$	$\frac{105+55x^2}{105+90x^2+9x^4}$	$\frac{105+55x^2}{105+90x^2+9x^4}$	$\frac{945+735x^2+64x^4}{945+1050x^2+224x^4}$

Figure 8.2: Padé approximants of orders (m, n) with $0 \le n, m \le 4$ of the function $f(x) = \frac{\arctan(x)}{x} = 1 - \frac{x^2}{3} + \frac{x^4}{5} - \frac{x^6}{7} + \cdots$ (thin).

The non-square rectangular first column in Table 8.1 right does not contradict Theorem 8.3 since for $R_{0,0} = \frac{0}{1}$ we have $\deg(0) = -1$. To avoid such a case we usually assume that $c_0 \neq 0$, so $R_{0,0} = \frac{c_0}{1}$ (see Figure 8.2 for the Padé approximants of $\arctan x/x$). The size r of the block in Theorem 8.3, may be infinite and then $Qf - P = 0$, so $f = \frac{P}{Q}$ is a rational function. Conversely, if $f = \frac{P}{Q}$, then r is infinite. If f is not a rational function, then for each m_0, n_0 there exist $m_1 > m_0$, $n_1 > n_0$ such that $R_{m_1,n_0} \neq R_{m_0,n_0} \neq R_{m_0,n_1}$. Thus we get infinite sequences of different Padé approximants with increasing indices m_i, n_i, which form staircases in the Padé table.

$$R_{m_0,n_0}, R_{m_1,n_0}, R_{m_1,n_1}, R_{m_2,n_1}, R_{m_2,n_2}, R_{m_3,n_2}, \ldots$$

$$R_{m_0,n_0}, R_{m_0,n_1}, R_{m_1,n_1}, R_{m_1,n_2}, R_{m_2,n_2}, R_{m_2,n_3}, \ldots$$

These sequences of Padé approximants can be expressed by continued fractions

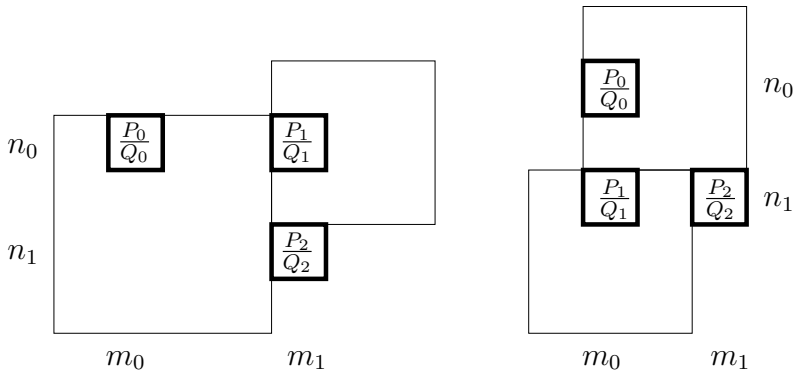


Figure 8.3: Square blocks in the Padé table (thin) and the Padé approximents (thick) from Theorem 8.4 (left) and Theorem 8.5 (right)

Theorem 8.4 *Let f be a formal power series which is not equal to any rational function and let $R_{m_0,n_0} = \frac{P_0}{Q_0}$ be its Padé approximant of order m_0, n_0. Assume that either $n_0 = 0$ or $R_{m_0,n_0-1} \neq R_{m_0,n_0}$. Let $m_1 > m_0$ be the first integer with $R_{m_0,n_0} \neq R_{m_1,n_0} = \frac{P_1}{Q_1}$ and let $n_1 > n_0$ be the first integer with $R_{m_1,n_0} \neq R_{m_1,n_1} = \frac{P_2}{Q_2}$ (see Figure 8.3 left). Then*

$$\begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} = \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & ax^k \\ 1 & \beta \end{bmatrix}$$

where $a \neq 0$, $k \leq n_1 - n_0$ and β is a polynomial of degree at most $\max\{0, (n_1 - n_0) - (m_1 - m_0)\}$. In particular if $(m_1, n_1) = (m_0 + 1, n_0 + 1)$ then $k = 1$ and $\beta = b$ is a constant.

Proof: Since $P_0Q_1 - P_1Q_0 \neq 0$, we use the pseudoinverse of $\left[\begin{smallmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{smallmatrix}\right]$ to compute

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} Q_1 & -P_1 \\ -Q_0 & P_0 \end{bmatrix} \cdot \begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} = \begin{bmatrix} 0 & P_2Q_1 - P_1Q_2 \\ P_0Q_1 - P_1Q_0 & P_0Q_2 - P_2Q_0 \end{bmatrix}$$

Since either $n_0 = 0$ or $R_{m_0,n_0-1} \neq R_{m_0,n_0}$, (n_0, m_0) belongs to the upper row of a square block of equal elements. By Theorem 8.3, all elements of the first row have the same Padé approximant expression, in particular $r_{m_0,n_0} = r_{m_1-1,n_0}$, which implies $\lambda(P_0 - fQ_0) \geq m_1 + n_0$. Similarly, R_{m_1,n_0} belongs to the first column of a square block of equal Padé approximants, so $r_{m_1,n_0} = r_{m_1,n_1-1}$ and $\lambda(P_1 - fQ_1) \geq m_1 + n_1$. Since $\lambda(P_2 - fQ_2) \geq m_1 + n_1 + 1$,

$$\begin{aligned} \lambda(C) &\geq \min\{\lambda(P_0Q_1 - fQ_0Q_1), \lambda(fQ_0Q_1 - P_1Q_0)\} \geq m_1 + n_0, \\ \lambda(B) &\geq \min\{\lambda(P_2Q_1 - fQ_1Q_2), \lambda(fQ_1Q_2 - P_1Q_2\} \geq m_1 + n_1, \\ \lambda(D) &\geq \min\{\lambda(P_0Q_2 - fQ_0Q_2), \lambda(fQ_0Q_2 - P_2Q_0\} \geq m_1 + n_0. \end{aligned}$$

Thus the orders of all B, C, D are at least $m_1 + n_0$. For the degrees we get $\deg(C) \leq m_1 + n_0$, $\deg(B) \leq m_1 + n_1$, $\deg(D) \leq \max\{m_1 + n_0, m_0 + n_1\}$. Thus $C = cx^{m_1+n_0}$, $B = bx^k$ with $k \leq m_1 + n_1$. Since $C = \det\left[\begin{smallmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{smallmatrix}\right]$, we have $\left[\begin{smallmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} Q_1/C & -P_1/C \\ -Q_0/C & P_0/C \end{smallmatrix}\right]$ and

$$\begin{aligned} \begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} &= \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix} \cdot \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} \\ &= \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & B/C \\ 1 & D/C \end{bmatrix} \end{aligned}$$

It follows $B/C = ax^k$ with $a \neq 0$, $k \leq n_1 - n_0$ and $\beta = D/C$ is a polynomial with $\deg(\beta) \leq \max\{0, (n_1 - n_0) - (m_1 - m_0)\}$. □

Theorem 8.5 *Let f be a formal power series which is not equal to any rational function and let $R_{m_0,n_0} = \frac{P_0}{Q_0}$ be its Padé approximant of order m_0, n_0. Assume that either $m_0 = 0$ or $R_{m_0-1,n_0} \neq R_{m_0,n_0}$. Let $n_1 > n_0$ be the first integer with $R_{m_0,n_0} \neq R_{m_0,n_1} = \frac{P_1}{Q_1}$ and let $m_1 > m_0$ be the first integer with $R_{m_0,n_1} \neq R_{m_1,n_1} = \frac{P_2}{Q_2}$ (see Figure 8.3 right). Then*

$$\begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} = \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & ax^k \\ 1 & \beta \end{bmatrix}$$

where $a \neq 0$, $k \leq m_1 - m_0$ and β is a polynomial of degree at most $\max\{0, (m_1-m_0)-(n_1-n_0)\}$. In particular if $(m_1, n_1) = (m_0 + 1, n_0 + 1)$ then $k = 1$ and $\beta = b$ is a constant.

The proof is analogous to the proof of Theorem 8.4. If the Padé approximants are all distinct, we can proceed along the diagonal $R_{00} = \frac{c_0}{1}$, $R_{10} = \frac{c_0+c_1x}{1}$, R_{11}, R_{21}, R_{22}, $R_{32}, \ldots$

$$\begin{aligned} f(x) &= \begin{bmatrix} c_0 & c_0 + c_1x \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x \\ 1 & b_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x \\ 1 & b_3 \end{bmatrix} \cdots \\ &= \begin{bmatrix} 1 & c_0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & c_1x \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x \\ 1 & b_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x \\ 1 & b_3 \end{bmatrix} \cdots \\ &= c_0 + \cfrac{c_1x}{1} + \cfrac{a_2x}{b_2} + \cfrac{a_3x}{b_3} + \cdots \end{aligned}$$

Alternatively we can express f as a continued fraction whose partial convergents are $R_{00} = \frac{c_0}{1}$, $R_{01} = \frac{c_0}{1-(c_1/c_0)x}$, R_{11}, $R_{12} \ldots$:

$$\begin{aligned} f(x) &= \begin{bmatrix} c_0 & c_0^2 \\ 1 & c_0 - c_1x \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x \\ 1 & b_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x \\ 1 & b_3 \end{bmatrix} \cdots \\ &= \begin{bmatrix} 0 & c_0 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & -c_1x \\ 1 & c_0 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x \\ 1 & b_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x \\ 1 & b_3 \end{bmatrix} \cdots \\ &= \cfrac{c_0}{1} - \cfrac{c_1x}{c_0} + \cfrac{a_2x}{b_2} + \cfrac{a_3x}{b_3} + \cdots \end{aligned}$$

Theorem 8.6 *Let f be a formal power series which is not equal to any rational function. Let $R_{m_0-1,n_0-1} \neq R_{m_0,n_0} = \frac{P_0}{Q_0}$, $R_{m_0,n_0} \neq R_{m_0+1,n_0+1} = \frac{P_1}{Q_1}$, $R_{m_0+1,n_0+1} \neq R_{m_0+2,n_0+2} = \frac{P_2}{Q_2}$. Then*

$$\begin{bmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{bmatrix} = \begin{bmatrix} P_0 & P_1 \\ Q_0 & Q_1 \end{bmatrix} \cdot \begin{bmatrix} 0 & ax^k \\ 1 & \beta \end{bmatrix}$$

where $a \neq 0$, $k \leq 2$ and β is a polynomial of degree at most 1.

Proof: We have $\lambda(P_0 - fQ_0) \geq m_0+n_0+1$, $\lambda(P_1 - fQ_1) \geq m_0+n_0+3$, $\lambda(P_2 - fQ_2) \geq m_0+n_0+5$. For the matrices $C = P_0Q_1 - P_1Q_0$, $B = P_2Q_1 - P_1Q_2$, $D = P_0Q_2 - P_2Q_0$ from the proof of Theorem 8.4 we get

$$\begin{aligned} \lambda(C) &\geq \min\{\lambda(P_0Q_1 - fQ_0Q_1), \lambda(fQ_0Q_1 - P_1Q_0)\} \geq m_0 + n_0 + 1, \\ \lambda(B) &\geq \min\{\lambda(P_2Q_1 - fQ_1Q_2), \lambda(fQ_1Q_2 - P_1Q_2\} \geq m_0 + n_0 + 3, \\ \lambda(D) &\geq \min\{\lambda(P_0Q_2 - fQ_0Q_2), \lambda(fQ_0Q_2 - P_2Q_0\} \geq m_1 + n_0 + 1. \end{aligned}$$

Since $\deg(C) \leq m_0+n_1+1$, $\deg(B) \leq m_0+n_0+3$, $\deg(D) \leq m_0+n_0+2$, the result follows. □

If in the Padé table all approximants are distinct, then we can express the formal power series f as a continued fraction whose partial convergents are $R_{00} = \frac{c_0}{1}$, $R_{11}(x) = \frac{c_0+(c_1-c_0c_2/c_1)x}{1-(c_2/c_1)x}$, $R_{22} \dots$

$$\begin{aligned} f(x) &= \begin{bmatrix} c_0 & c_0c_1+(c_1^2-c_0c_2)x \\ 1 & c_1-c_2x \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x^{k_2} \\ 1 & \beta_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x^{k_3} \\ 1 & \beta_3 \end{bmatrix} \cdots \\ &= \begin{bmatrix} 0 & c_0 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & -(c_1^2/c_0)x \\ 1 & c_1+((c_1^2/c_0)-c_2)x \end{bmatrix} \cdot \begin{bmatrix} 0 & a_2x^{k^2} \\ 1 & \beta_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & a_3x^{k_3} \\ 1 & \beta_3 \end{bmatrix} \cdots \\ &= \frac{c_0}{1} + \frac{-(c_1^2/c_0)x}{c_1+((c_1^2/c_0)-c_2)x} + \frac{a_2x^{k_2}}{\beta_2} + \frac{a_3x^{k_3}}{\beta_3} + \cdots \end{aligned}$$

where $k_i \leq 2$ and $\deg(\beta_i) \leq 1$. These expressions do not say anything about the convergence of these continued fractions nor about the convergence of the original formal power series f. Nevertheless, if convergent, the convergence is usually faster and has wider definition domain than the formal power series. For example for the exponential function we get a continued fraction which converges for every $x \in \mathbb{R}$ and its partial convergents form a staircase R_{00}, R_{10}, R_{11}, $R_{21}, \cdots$ in the Padé table.

$$\begin{aligned} e^x &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 2 \end{bmatrix} \cdots \\ &= 1 + \frac{x}{1} - \frac{x}{2} + \frac{x}{3} - \frac{x}{2} + \frac{x}{5} - \cdots - \frac{x}{2} + \frac{x}{(2n+1)} - \cdots \end{aligned}$$

Alternatively we get a continued fraction which converges for every $x \in \mathbb{R}$ and its partial convergents form a staircase R_{00}, R_{01}, R_{11}, $R_{12}, \cdots$ in the Padé table. The two expressions for e^x are related by the formula $e^x = 1/e^{-x}$.

$$\begin{aligned} e^x &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & -x \\ 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} 0 & x \\ 1 & 2 \end{bmatrix} \cdots \\ &= \frac{1}{1} - \frac{x}{1} + \frac{x}{2} - \frac{x}{3} + \cdots + \frac{x}{2} - \frac{x}{(2n+1)} + \cdots \end{aligned}$$

Using Theorem 8.6 we get a continued fraction which converges for every $x \in \mathbb{R}$ and its partial convergents form the main diagonal R_{00}, R_{11}, R_{22}, $R_{33}, \cdots$ in the Padé table.

$$\begin{aligned} e^{2x} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & 2x \\ 1 & 1-x \end{bmatrix} \cdot \begin{bmatrix} 0 & x^2 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 0 & x^2 \\ 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} 0 & x^2 \\ 1 & 7 \end{bmatrix} \cdots \\ &= 1 + \frac{2x}{1-x} + \frac{x^2}{3} + \frac{x^2}{5} + \frac{x^2}{7} + \cdots + \frac{x^2}{2n+1} + \cdots \end{aligned}$$

Many other convergence results have been obtained - see Wall [67] or Jones and Thron [27].

$$\begin{aligned}
(1+x)^a &= \frac{1}{1}-\frac{ax}{1}+\frac{(1+a)x}{2}+\frac{(1-a)x}{3}+\cdots+\frac{n(n+a)x}{2n}+\frac{n(n-a)x}{2n+1}+\cdots && x>-1\\
\ln(1+x) &= \frac{x}{1}+\frac{x}{2}+\frac{x}{3}+\frac{4x}{4}+\frac{4x}{5}+\cdots+\frac{n^2x}{2n}+\frac{n^2x}{2n+1}+\cdots && x>-1\\
\tan x &= \frac{x}{1}-\frac{x^2}{3}-\frac{x^2}{5}-\cdots-\frac{x^2}{2n+1}-\cdots && -\frac{\pi}{2}<x<\frac{\pi}{2}\\
\tanh x &= \frac{x}{1}+\frac{x^2}{3}+\frac{x^2}{5}+\cdots+\frac{x^2}{2n+1}+\cdots && x\in\mathbb{R}\\
\arctan x &= \frac{x}{1}+\frac{x^2}{3}+\frac{4x^2}{5}+\cdots+\frac{n^2x^2}{(2n+1)}+\cdots && x\in\mathbb{R}\\
\arg\tanh x &= \frac{1}{2}\ln\frac{1+x}{1-x}=\frac{x}{1}-\frac{x^2}{3}-\frac{4x^2}{5}-\cdots-\frac{n^2x^2}{(2n+1)}-\cdots && -1<x<1
\end{aligned}$$

8.2 Algebraic tensors

The approximation of transcendent functions by Padé approximants and continued fractions leads to the concept of an **algebraic tensor** which is a function $T:\overline{\mathbb{R}}\times\overline{\mathbb{R}}\to\overline{\mathbb{R}}\cup\{\frac{0}{0}\}$ of two real variables which is a rational function in the first variable and a Möbius transformation in the second variable. An algebraic tensor of degree $q\geq 0$ is given by

$$T(x,y)=\frac{(T_{000}x_0^q+T_{010}x_0^{q-1}x_1+\cdots+T_{0q0}x_1^q)y_0+(T_{001}x_0^q+T_{011}x_0^{q-1}x_1+\cdots+T_{0q1}x_1^q)y_1}{(T_{100}x_0^q+T_{110}x_0^{q-1}x_1+\cdots+T_{1q0}x_1^q)y_0+(T_{101}x_0^q+T_{111}x_0^{q-1}x_1+\cdots+T_{1q1}x_1^q)y_1}$$

so

$$T(x,y)_k=\sum_{i=0}^{q}\sum_{j=0}^{1}T_{kij}x_0^{q-i}x_1^iy_j$$

In particular, an algebraic tensor of degree 0 does not depend on x and is a transformation given by the matrix $T=\left[\begin{smallmatrix}T_{000} & T_{001}\\ T_{100} & T_{101}\end{smallmatrix}\right]$. For example, the unit matrix represents the projection $\mathrm{Id}(x,y)=y$. Given an algebraic tensor T of degree q, for each $x\in\overline{\mathbb{R}}$ we get a transformation T^*x given by $(T^*x)(y)=T(x,y)$. For each $y\in\overline{\mathbb{R}}$ we get a rational function T_*y of degree q given by $(T_*y)(x)=T(x,y)$. Thus

$$(T^*x)_{kj}=\sum_{i=0}^{q}T_{kij}x_0^{q-i}x_1^i,\ (T_*y)_{ki}=\sum_{j=0}^{1}T_{kij}y_j$$

For a transformation P we get algebraic tensors (T^*P), (T_*P), PT of degree q given by

$$(T^*P)(x,y)=T(Px,y),\ (T_*P)(x,y)=T(x,Py),\ (PT)(x,y)=P(T(x,y))$$

algebraic tensors satisfy similar identities as bilinear tensors, e.g.,

$$\begin{aligned}
T(x,y) &= (T^*x)y=(T_*y)x,\\
T(x,P) &= (T^*x)P=(T_*P)^*x,\\
T(P,y) &= (T_*y)P=(T^*P)_*y,\\
(T^*P)^*Q &= T^*(PQ),\\
(T_*P)_*Q &= T_*(PQ).
\end{aligned}$$

For a function expressed by a continued fraction we have

$$f(x) \;=\; \frac{a_0x}{b_0+}\,\frac{a_1x}{b_1+}\,\frac{a_2x}{b_2+}\cdots \;=\; \lim_{n\to\infty}(T_0^*x)\cdot(T_1^*x)\cdots(T_n^*x)(i)$$

where $T_n(x,y)=\frac{a_nx}{y+b_n}$. The composition of matrices T_n^*x leads to a new kind of tensor product. For algebraic tensors T,S we have $(T^*x)(S^*x)(y)=(T^*x)(S(x,y))=T(x,S(x,y))$. Thus for tensors T,S of degrees q,p we define the tensor $T*S$ of degree at most $q+p$ by

$$(T*S)(x,y)=T(x,S(x,y)).$$

Then

$$\begin{aligned}(T*S)(x,y)_n &= \sum_{i=0}^{q}\sum_{m=0}^{1} T_{nim}x_0^{q-i}x_1^iS(x,y)_m\\ &= \sum_{i=0}^{q}\sum_{m=0}^{1}\sum_{j=0}^{p}\sum_{k=0}^{1} T_{nim}S_{mjk}x_0^{q+p-i-j}x_1^{i+j}y_k\\ &= \sum_{r=0}^{p+q}\sum_{k=0}^{1}\sum_{i=\max(0,r-p)}^{\min(q,r)}\sum_{m=0}^{1} T_{nim}\cdot S_{m,r-i,k}x_0^{q+p-r}x_1^ry_k,\end{aligned}$$

where $r=i+j$. Thus

$$(T*S)_{nrk}=\sum_{i=\max(0,r-p)}^{\min(q,r)}\sum_{m=0}^{1} T_{nim}\cdot S_{m,r-i,k},\ 0\le r\le p+q$$

Proposition 8.7 *Let T,S,R be algebraic tensors, P,Q matrices and x a vector. Then*
1. $T*\mathrm{Id}=\mathrm{Id}*T=T$, $\mathrm{Id}^*P=\mathrm{Id}$, $\mathrm{Id}_*P=P$
2. $((T*S)*R)=(T*(S*R))$
3. $(T*S)^*x=(T^*x)\circ(S^*x)$
4. $(T*S)^*P=(T^*P)*(S^*P)$
5. $(T*S)_*Q=T*(S_*Q)$
6. $(PT)*S=P(T*S)$
7. $(T_*P)*S=T*(PS)$
8. $T*(PS)=(T_*P)*S$

Proof: $(\mathrm{Id} * T)(x,y) = \mathrm{Id}(x, T(x,y)) = T(x,y)$, $(T * \mathrm{Id})(x,y) = T(x, \mathrm{Id}(x,y)) = T(x,y)$,

$$\begin{aligned}
((T*S)*R)(x,y) &= (T*S)(x,R(x,y)) = T(x,S(x,R(x,y))) \\
&= T(x,(S*R)(x,y)) = (T*(S*R))(x,y) \\
((T*S)^*x)(y) &= (T*S)(x,y) = T(x,S(x,y)) = (T^*x)(S(x,y)) \\
&= (T^*x)((S^*x)(y)) = (T^*x)\cdot(S^*x)(y) \\
(T*S)^*P(x,y) &= (T*S)(Px,y) = T(Px,S(Px,y)) = (T^*P)(x,S(Px,y)) \\
&= (T^*P)(x,(S^*P)(x,y)) = ((T^*P)*(S^*P))(x,y) \\
(T*S)_*Q)(x,y) &= (T*S)(x,Qy) = T(x,S(x,Qy)) = T(x,(S_*Q)(x,y)) \\
&= (T*(S_*Q))(x,y) \\
((PT)*S)(x,y) &= (PT)(x,S(x,y)) = P(T(x,S(x,y))) = P((T*S)(x,y)) \\
((T_*P)*S)(x,y) &= (T_*P)(x,S(x,y)) = T(x,P(S(x,y)) = (T*(PS))(x,y) \\
T*(PS)(x,y) &= T(x,PS(x,y)) = (T_*P)(x,S(x,y)) = (T_*P)*S(x,y) \quad \square
\end{aligned}$$

The image of intervals I, J by an algebraic tensor is defined by

$$T(I,J) = \{T(x,y) :\ x \in I, y \in J\} \cap \overline{\mathbb{R}}.$$

Theorem 8.8 (Inclusion criterion) *If T is a algebraic tensor, P, Q, R are regular matrices and $\mathrm{sgn}(R^{-1}T(P,Q)) \geq 0$ then $T(P^c, Q^c) \subseteq R^c$.*

The proof is analogous to the proof of Theorem 5.31. We take the $(q+1)$-linear tensor

$$S(x^{(1)}, \ldots, x^{(q)}, y)_k = \sum_{i_1,\ldots,i_q,j} T_{k,i_1,\ldots,i_q,j} x^{(1)}_{i_1}, \ldots, x^{(q)}_{i_q} y_j$$

which is symmetric in the first q variables and $S(x,\ldots,x,y) = T(x,y)$. If $\mathrm{sgn}(R^{-1}T(P,Q)) \geq 0$ then $T(P^c,Q^c) \subseteq S(P^c,\ldots,P^c,Q^c) \subseteq R^c$.

Theorem 8.9 *Let $\{T_n :\ n \geq 0\}$ be a sequence of algebraic tensors and I, J intervals such that $T_n(I,J) \subseteq J$ for all n.*
1. If for each $x \in I$, T^*x is a contraction on J then there exists a limit

$$f(x) = \lim_{n\to\infty} (T_0^*x) \cdots (T_n^*x)(i).$$

2. If there exists a limit $f(x) = \lim_{n\to\infty}(T_0^*x)\cdots(T_n^*x)(i)$, then $f(x) \in \overline{J}$.

Proof: For each $x \in I$ and for each n we have $(T_n^*x)(J) \subseteq J$, so both statements follow by Proposition 3.43. $\square$

Definition 8.10 *For an algebraic tensor T and a matrix P we write $T \subseteq P$ if $\mathrm{sgn}(P^{-1}T) \geq 0$.*

Lemma 8.11 *Let T be an algebraic tensor and P, Q, R regular matrices. If $P \subseteq Q$ and $T^*Q \subseteq R$ then $T^*P \subseteq R$.*

Proof: $R^{-1}T^*P = R^{-1}T^*(QQ^{-1}P) = (R^{-1}T^*Q)^*(Q^{-1}P)$, so $\mathrm{sgn}(R^{-1}T^*P) \geq 0$. □

Lemma 8.12 *Let S, T be tensors and P, Q, R matrices. If $T(P,Q) \subseteq Q$ and $S(P,Q) \subseteq R$ then $(S*T)(P,Q) \subseteq R$.*

Proof: We have

$$\begin{aligned}(S*T)(P,Q) &= ((S*T)^*P)_*Q = ((S^*P)*(T^*P))_*Q = (S^*P)*T(P,Q)\\ &= (S^*P)*(QQ^{-1}T(P,Q)) = (S^*P_*Q)*(Q^{-1}T(P,Q))\\ &= S(P,Q)*(Q^{-1}T(P,Q))\end{aligned}$$

so we get $\mathrm{sgn}(R^{-1}(S*T)(P,Q)) = \mathrm{sgn}(R^{-1}S(P,Q)) \cdot \mathrm{sgn}(Q^{-1}T(P,Q)) \geq 0$ □

8.3 The transcendent algorithm

To compute a transcendent function in an interval $I = P^c$ we expres it as a limit $f(x) = \lim_{n\to\infty}(T_0^*x)\cdots(T_n^*x)(i)$, for some algebraic tensors T_n. This is possible if there exists an interval $J = Q^c$ such that $T_n(P,Q) \subseteq Q$ for all sufficiently large n. The algorithm uses states (vertices) $(X, Y, n, p, q) \in \mathbb{T}(\mathbb{R}) \times \mathbb{M}(\mathbb{R}) \times \mathbb{N} \times B^2$, where $X = T_0 * \cdots * T_n$, and $Y = F_u$ for some $\mathbf{i} \xrightarrow{u} p$.

Definition 8.13 *Let (F, G, V) be a sofic number system, $\{T_n : n \geq 0\}$ a sequence of algebraic tensors and P, Q regular matrices such that $T_n(P,Q) \subseteq Q$ for all $n \geq n_0$. The transcendent graph has vertices $(X, Y, n, p, q) \in \mathbb{T}(\mathbb{R}) \times \mathbb{M}(\mathbb{R}) \times \mathbb{N} \times B^2$. The labelled edges are*

$$\begin{array}{lll}(X,Y,n,p,q) \xrightarrow{(a,\lambda)} (X^*H_{p,a,p'}, YH_{p,a,p'}, n, p', q) & \text{if} & p \xrightarrow{a} p'\\ (X,Y,n,p,q) \xrightarrow{(\lambda,\lambda)} (X*(T_n^*Y), Y, n+1, p, q), & & \\ (X,Y,n,p,q) \xrightarrow{(\lambda,a)} (F_a^{-1}X, Y, n, p, q') & \text{if} & p \neq \mathbf{i},\ n \geq n_0,\ q \xrightarrow{a} q',\\ & & Y \subseteq P,\ X_*Q \subseteq F_aV_{q'}\end{array}$$

The first rule is the **digit absorption** of a letter from the input. The second rule is the **tensor absorption** of the n-th tensor and the third rule is an emission of an output letter.

Proposition 8.14 *Let (F, G, V) be a sofic number system, $\{T_n : n \geq 0\}$ a sequence of algebraic tensors and P, Q regular matrices such that $T_n(P,Q) \subseteq Q$ for all $n \geq n_0$. Set $S_n = T_0 * \cdots * T_{n-1}$, $S_0 = \mathrm{Id}$. If $(\mathrm{Id}, \mathrm{Id}, 0, \mathbf{i}, \mathbf{i}) \xrightarrow{(u,v)} (X, Y, n, p, q)$ is a path in the transcendent graph, then $\mathbf{i} \xrightarrow{u} p$, $\mathbf{i} \xrightarrow{v} q$, $Y = F_uV_p$, and $X = F_v^{-1}S_n^*Y$. If $p \neq \mathbf{i}$ then $Y \subseteq P$. If moreover $q \neq \mathbf{i}$ then $X_*Q \subseteq V_q$, or $S_n(Y,Q) \subseteq F_vV_q$.*

Proof: The first digit absorption and the first tensor absorption yield

$$\begin{aligned}(\mathrm{Id},\mathrm{Id},0,\mathbf{i},\mathbf{i}) &\xrightarrow{(a,\lambda)} (\mathrm{Id}, F_aV_p, 0, p', q) \xrightarrow{(\lambda,\lambda)} (T_0^*(F_aV_p), F_aV_p, 1, p', q)\\ (\mathrm{Id},\mathrm{Id},0,\mathbf{i},\mathbf{i}) &\xrightarrow{(\lambda,\lambda)} (T_0, \mathrm{Id}, 1, p, q) \xrightarrow{(\lambda,\lambda)} (T_0^*(F_aV_p), F_aV_p, 1, p', q)\end{aligned}$$

Assume by induction that the proposition holds for $(\mathrm{Id}, \mathrm{Id}, 0, \mathbf{i}, \mathbf{i}) \xrightarrow{(u,v)} (X, Y, n, p, q)$.
1. If $(X, Y, n, p, q) \xrightarrow{(a,\lambda)} (X', Y', n, p', q)$ is a digit absorption, then $Y' = YH_{p,a,p'} = F_{ua}V_{p'}$,

$$\begin{aligned}X' &= X^*H_{p,a,p'} = (F_v^{-1}S_n^*Y)^*H_{p,a,p'} = F_v^{-1}((S_n^*Y)^*H_{p,a,p'})\\ &= F_v^{-1}(S_n^*(YH_{p,a,p'})) = F_v^{-1}S_n^*Y'\end{aligned}$$

If $p \neq \mathbf{i}$ then $Y' \subseteq Y \subseteq V_q$. If moreover $q \neq \mathbf{i}$ then $S_n(Y,Q) \subseteq F_v V_q$ and therefore $S_n(Y',Q) \subseteq F_v V_q$ by Lemma 8.11.
2. If $(X,Y,n,p,q) \overset{(\lambda,\lambda)}{\longrightarrow} (X',Y,n+1,p,q)$ is a tensor absorption, then

$$X' \;=\; X * (T_n^* Y) = (F_v^{-1} S_n^* Y) * (T_n^* Y) = F_v^{-1}((S_n^* Y) * (T_n^* Y)) = F_v^{-1}(S_{n+1}^* Y)$$

If $q \neq \mathbf{i}$ then we have $S_n(Y,Q) \subseteq F_v V_q$, $T_n(Y,Q) \subseteq Q$, so $(S_n * T_n)(Y,Q) \subseteq F_v V_q$ by Lemma 8.12.
3. If $(X,Y,n,p,q) \overset{(\lambda,a)}{\longrightarrow} (X',Y,n,p,q')$ is an emission, then $X' = F_a^{-1} X = F_{va}^{-1} S_n^* Y$, $X'_* Q \subseteq V_{q'}$, so $S_n(Y,Q) \subseteq F_{va} V_{q'}$. □

Theorem 8.15 *Let $\{T_n : n \geq 0\}$ be a sequence of tensors and P, Q regular matrices such that $T_n(P,Q) \subseteq Q$ for all $n \geq n_0$. Assume that for each $x \in P^c$ there exists a limit $f(x) = \lim_{n\to\infty}(T_0^* x)\cdots(T_n^* x)(i)$. If $(\mathrm{Id},\mathrm{Id},0,\mathbf{i},\mathbf{i}) \overset{(u,v)}{\longrightarrow}$ is a path with infinite words u, v, then $u, v \in \Sigma_G$ and $f(\Phi(u)) = \Phi(v)$.*

Proof: For every k there exists n_k and m_k such that

$$(\mathrm{Id},\mathrm{Id},0,\mathbf{i},\mathbf{i}) \overset{(u_{[0,m_k)},v_{[0,k)})}{\longrightarrow} (X,Y,n_k,p_k,q_k),$$

so $S_{n_k}(F_{u_{[0,m_k)}} V_{p_k}, Q) \subseteq F_{v_{[0,k)}} V_{q_k}$. Let $x = \Phi(u) \in P^c$ and denote by

$$f_n(x) = \lim_{k\to\infty} (T_n^* x)\cdots(T_{n+k}^* x)(i).$$

Since $T_n(P,Q) \subseteq Q$ for $n \geq n_0$, we get $f_n(x) \in Q^c$ by Proposition 3.43 and Theorem 8.9. Since $x = \Phi(u) \in F_{u_{[0,m_k)}} V_{p_k}$ we get $f(x) = S_n(x, f_n(x)) \in F_{v_{[0,k)}} V_{q_k}$. Since $\Phi(v) \in F_{v_{[0,k)}} V_{q_k}$, we get $f(x) = \Phi(v)$. □

To get a deterministic algorithm we need a selector which chooses at each step one of the possible actions. A greedy selector chooses an emission whenever possible. If there is no emission possible, the selector chooses either a digit absorption or a tensor absorption. To carry out this decision, we use the concept of matrix convex hull of Proposition 5.24. Recall that the (2×2)-matrix $\overline{T}$ is the matrix convex hull of an $(n\times 2)$-matrix T, if for each regular matrix Q we have $\mathrm{sgn}(Q^{-1}T) \geq 0$ iff $\mathrm{sgn}(Q^{-1}\overline{T}) \geq 0$.

Definition 8.16 *The* **tensor convex hull** $\widetilde{T}$ *of an algebraic tensor T is the bilinear tensor $\widetilde{T}$ given by $\widetilde{T}_{ijk} = (\overline{T_{--k}})_{ij}$. If $\deg(T) \leq 1$ then $\widetilde{T} = T$.*

Thus $\widetilde{T}_{--0}$ is the matrix convex hull of the first q entries of T and $\widetilde{T}_{--1}$ is the matrix convex hull of the last q entries of T. Successive digit absorptions diminish both intervals $\widetilde{T}_{--0}$ and $\widetilde{T}_{--1}$, but these intervals may remain apart since the contraction of the Q-interval do not increase. Successive tensor absorptions on the other hand lead to greater contraction of the Q-interval which result in smaller distance between the intervals $\widetilde{T}_{--0}$ and $\widetilde{T}_{--1}$. Thus a reasonable selector for the transcendent algorithm is the balanced greedy selector of Table 5.3 applied to $\widetilde{X}$ of the state (X,Y,p,q,n), where y-absorption is replaced by tensor absorption.

Corollary 8.17 *Let T_n be a sequence of algebraic tensors and P, Q regular matrices and n_0 an integer such that for each $x \in P^c$, $n \geq n_0$, $T_n^* x$ is a contraction on Q^c. Then for each path $(p,u) \in \Sigma_{|G|}$ such that $x = \Phi(u) \in P^c$, the balanced greedy selector computes an infinite path with output (q,v) such that*

$$y = \Phi(v) = \lim_{n\to\infty} (T_0^* x)(T_1^* x)(T_2^* x)\cdots(T_n^* x)(i)$$

d	$\widetilde{X_*Q}$	Y	u_n	v_m
	$[\frac{0}{1}, \frac{1}{1}]$	$[\frac{1}{0}, \frac{0}{1}]$	$\lambda \overset{\overline{0}}{\to} \overline{0}$	
0				
	$[\frac{3}{1}, \frac{-1}{-3}, \frac{5}{3}, \frac{-3}{-5}]$	$[\frac{1}{2}, \frac{1}{-2}]$	$\overline{0} \overset{\overline{0}}{\to} \overline{0}$	
	$[\frac{2}{0}, \frac{0}{-2}, \frac{3}{1}, \frac{-1}{-3}]$	$[\frac{1}{1}, \frac{1}{-1}]$	$\overline{0} \overset{\overline{1}}{\to} \overline{1}$	
	$[\frac{3}{-5}, \frac{0}{-4}, \frac{2}{-6}, \frac{-2}{-6}]$	$[\frac{4}{-1}, \frac{2}{-2}]$	$\overline{1} \overset{0}{\to} 0$	
1				
	$[\frac{-1}{-31}, \frac{-3}{-21}, \frac{1}{-39}, \frac{-1}{-9}]$			$\lambda \overset{0}{\to} 0$
	$[\frac{-2}{-31}, \frac{-6}{-21}, \frac{2}{-39}, \frac{-2}{-9}]$			$0 \overset{0}{\to} 0$
	$[\frac{-4}{-31}, \frac{-12}{-21}, \frac{4}{-39}, \frac{-4}{-9}]$	$[\frac{3}{-1}, \frac{1}{-1}]$	$0 \overset{1}{\to} 1$	
2				
	$[\frac{-1808}{-16636}, \frac{-160}{-296}, \frac{-2640}{-19532}, \frac{-192}{-352}]$	$[\frac{7}{-4}, \frac{2}{-2}]$	$1 \overset{0}{\to} 0$	
	$[\frac{-452}{-4159}, \frac{-940}{-2885}, \frac{-660}{-4883}, \frac{-1148}{-3417}]$			$0 \overset{0}{\to} 0$
	$[\frac{-904}{-4159}, \frac{-1880}{-2885}, \frac{-1320}{-4883}, \frac{-2296}{-3417}]$	$[\frac{7}{-4}, \frac{5}{-4}]$	$0 \overset{1}{\to} 1$	
3				
	$[\frac{-90665024}{-200750088}, \frac{-883840}{-1345680}, \frac{-102683712}{-227977096}, \frac{-1004160}{-1530320}]$			$0 \overset{1}{\to} 1$
	$[\frac{19420040}{-200750088}, \frac{-422000}{-1345680}, \frac{22609672}{-227977096}, \frac{-478000}{-1530320}]$			$1 \overset{0}{\to} 0$
	$[\frac{38840080}{-200750088}, \frac{-844000}{-1345680}, \frac{45219344}{-227977096}, \frac{-956000}{-1530320}]$	$[\frac{23}{-16}, \frac{10}{-8}]$	$1 \overset{1}{\to} 1$	
	$[\frac{-531073520}{-2976887688}, \frac{-13504000}{-21530880}, \frac{-591561712}{-3383035528}, \frac{-15296000}{-24485120}]$	$[\frac{43}{-32}, \frac{20}{-16}]$	$1 \overset{0}{\to} 0$	
	$[\frac{-66384190}{-372110961}, \frac{-165488990}{-353507681}, \frac{-73945214}{-422879441}, \frac{-187025182}{-401922177}]$			$0 \overset{0}{\to} 0$
	$[\frac{-132768380}{-372110961}, \frac{-330977980}{-353507681}, \frac{-147890428}{-422879441}, \frac{-374050364}{-401922177}]$			$0 \overset{1}{\to} 1$
	$[\frac{106574201}{-372110961}, \frac{-308448279}{-353507681}, \frac{127098585}{-422879441}, \frac{-346178551}{-401922177}]$	$[\frac{43}{-32}, \frac{41}{-32}]$	$0 \overset{1}{\to} 1$	

input: $u = \overline{001}0101101$, $p = 1$, $F_uV_p = \left[\begin{smallmatrix}-167 & -82\\ 128 & 64\end{smallmatrix}\right] = [-1.305, -1.281]$
$\exp(2 \cdot F_uV_p) = [0.0735, 0.0771]$
output: $v = 0001001$, $q = 1$, $F_vV_q = \left[\begin{smallmatrix}17 & 10\\ 256 & 128\end{smallmatrix}\right] = [0.0664, 0.0781]$
invariant matrix: $Q = \left[\begin{smallmatrix}0 & 1\\ 1 & 1\end{smallmatrix}\right]$

Table 8.2: The transcendent algorithm with the exponential function in the binary signed system of Example 4.3.

Consider the computation of the exponential function according to

$$\begin{aligned} e^{2x} &= \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\cdot\begin{bmatrix}0 & 2x\\ 1 & 1-x\end{bmatrix}\cdot\begin{bmatrix}0 & x^2\\ 1 & 3\end{bmatrix}\cdot\begin{bmatrix}0 & x^2\\ 1 & 5\end{bmatrix}\cdot\begin{bmatrix}0 & x^2\\ 1 & 7\end{bmatrix}\cdots \\ &= \lim_{n\to\infty}(T_0^*x)(T_1^*x)(T_2^*x)\cdots(T_n^*x)(i) \end{aligned}$$

where

$$T_0^*x = \begin{bmatrix}1 & 1+x\\ 1 & 1-x\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\cdot\begin{bmatrix}0 & 2x\\ 1 & 1-x\end{bmatrix},\ T_n^*x = \begin{bmatrix}0 & x^2\\ 1 & 2n+1\end{bmatrix}$$

We have $\det(T_n^*x) = -x^2$, so T_n^*x is decreasing (or singular for $x=0$). As the invariant matrix we take $Q = \left[\begin{smallmatrix}0 & 1\\ 1 & 1\end{smallmatrix}\right]$ with $Q^c = [0,1]$. The P matrix should not contain ∞ but can be arbitrarily large. For $a>0$ set $P = \left[\begin{smallmatrix}-a & a\\ 1 & 1\end{smallmatrix}\right]$, so $P^c = (-a,a)$. Take n_0 with $2n_0+1 > a^2$. Then for each $n \geq n_0$, $T_n(P,Q) \subseteq Q$, and T_n^*x is contractive on Q for every $x \in P^c$. Indeed

$$|(T_n^*x)^\bullet(y)| = \frac{x^2(y^2+1)}{x^4+(y+2n+1)^2} \leq \frac{2a^2}{(2n+1)^2} < 1$$

Since and $T_n(x,0) = \frac{x^2}{2n+1} \in Q^c$, $T_n(x,1) = \frac{x^2}{2n+2} \in Q^c$, we get $T_n(P,Q) \subseteq Q$. A sample computation of the algorithm is in Table 8.2. The first column gives the degree of the state tensor, the second column gives its tensor convex hull.

For the function $\arctan x$ we have

$$\begin{aligned} \arctan x &= \frac{x}{1}+\frac{x^2}{3}+\frac{4x^2}{5}+\cdots+\frac{n^2x^2}{(2n+1)}+\cdots \qquad x\in\mathbb{R} \\ &= \begin{bmatrix}0 & x\\ 1 & 1\end{bmatrix}\cdot\begin{bmatrix}0 & x^2\\ 1 & 3\end{bmatrix}\cdot\begin{bmatrix}0 & 4x^2\\ 1 & 5\end{bmatrix}\cdot\begin{bmatrix}0 & 9x^2\\ 1 & 7\end{bmatrix}\cdots \\ &= \lim_{n\to\infty}(T_0^*x)(T_1^*x)(T_2^*x)\cdots(T_n^*x)(i) \end{aligned}$$

where

$$T_0^*x = \begin{bmatrix}0 & x\\ 1 & 1\end{bmatrix}\ T_n^*x = \begin{bmatrix}0 & n^2x^2\\ 1 & 2n+1\end{bmatrix}$$

The partial fractions are the Padé approximants R_{00}, R_{02}, R_{22}, $R_{24}\ldots$ of the function $\arctan(x)/x$ multiplied by x (see Figure 8.2). We have $\det(T^*x) = -n^2x^2$, so T_n^*x is decreasing (or singular for $x=0$). For a given interval $P = \left[\begin{smallmatrix}-a & a\\ 1 & 1\end{smallmatrix}\right]$ take $Q = \left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$ and n_0 with $2n_0+1 > a$. Then $T_n(P,Q) \subseteq Q$, but T_n^*x is not contractive on Q. However the composition

$$(T_{n-1}^*x)(T_n^*x) = \begin{bmatrix}(n-1)^2x^2 & (2n+1)(n-1)^2x^2\\ 2n-1 & n^2x^2+4n^2-1\end{bmatrix}$$

is contractive on Q, so the algorithm works (see Table 8.3).

For the function $\ln(1+x)$ we have

$$\begin{aligned} \ln(1+x) &= \frac{x}{1}+\frac{x}{2}+\frac{x}{3}+\frac{4x}{4}+\frac{4x}{5}+\cdots+\frac{n^2x}{2n}+\frac{n^2x}{2n+1}+\cdots \qquad x>-1 \\ &= \begin{bmatrix}0 & x\\ 1 & 1\end{bmatrix}\cdot\begin{bmatrix}0 & x\\ 1 & 2\end{bmatrix}\cdot\begin{bmatrix}0 & x\\ 1 & 3\end{bmatrix}\cdot\begin{bmatrix}0 & 4x\\ 1 & 4\end{bmatrix}\cdot\begin{bmatrix}0 & 4x\\ 1 & 5\end{bmatrix}\cdots \\ &= \lim_{n\to\infty}(T_0^*x)(T_1^*x)(T_2^*x)\cdots(T_n^*x)(i) \end{aligned}$$

d	$\widetilde{X_*Q}$	Y	u_n	v_m
	$[\frac{0}{1}, \frac{1}{0}]$	$[\frac{1}{0}, \frac{0}{1}]$	$\lambda \xrightarrow{\overline{0}} \overline{0}$	
0				
	$[\frac{1}{2}, \frac{1}{-2}, \frac{0}{2}]$	$[\frac{1}{2}, \frac{1}{-2}]$	$\overline{0} \xrightarrow{\overline{0}} \overline{0}$	
1				
	$[\frac{0}{0}, \frac{0}{0}, \frac{0}{0}, \frac{0}{0}]$	$[\frac{1}{1}, \frac{1}{-1}]$	$\overline{0} \xrightarrow{\overline{1}} \overline{1}$	
	$[\frac{72}{-72}, \frac{12}{-19}, \frac{4}{-1}, \frac{8}{-8}]$			$\lambda \xrightarrow{\overline{0}} \overline{0}$
	$[\frac{36}{-72}, \frac{6}{-19}, \frac{2}{-1}, \frac{4}{-8}]$	$[\frac{4}{-1}, \frac{2}{-2}]$	$\overline{1} \xrightarrow{0} 0$	
2				
	$[\frac{153}{-192}, \frac{19}{-48}, \frac{54}{-120}, \frac{9}{-24}]$			$\overline{0} \xrightarrow{\overline{1}} \overline{1}$
	$[\frac{57}{-96}, \frac{-5}{-24}, \frac{-6}{-60}, \frac{-3}{-12}]$	$[\frac{3}{-1}, \frac{1}{-1}]$	$0 \xrightarrow{1} 1$	
3				
	$[\frac{27608}{-952392}, \frac{-704}{-3264}, \frac{10496}{-87168}, \frac{-80}{-384}]$			$\overline{1} \xrightarrow{0} 0$
	$[\frac{55216}{-952392}, \frac{-1408}{-3264}, \frac{20992}{-87168}, \frac{-160}{-384}]$			$0 \xrightarrow{0} 0$
	$[\frac{110432}{-952392}, \frac{-2816}{-3264}, \frac{41984}{-87168}, \frac{-320}{-384}]$	$[\frac{7}{-4}, \frac{2}{-2}]$	$1 \xrightarrow{0} 0$	
	$[\frac{13804}{-119049}, \frac{-29620}{-68505}, \frac{5248}{-10896}, \frac{-2560}{-7440}]$			$0 \xrightarrow{0} 0$
	$[\frac{27608}{-119049}, \frac{-59240}{-68505}, \frac{10496}{-10896}, \frac{-5120}{-7440}]$	$[\frac{7}{-4}, \frac{5}{-4}]$	$0 \xrightarrow{1} 1$	
	$[\frac{-63309376}{-172704072}, \frac{-947840}{-1096080}, \frac{-643072}{-17614848}, \frac{-81920}{-119040}]$	$[\frac{23}{-16}, \frac{10}{-8}]$	$1 \xrightarrow{1} 1$	
4				
	$[\frac{-64673315584}{-119941739040}, \frac{-169256960}{-205451520}, \frac{-5988784384}{-9960752928}, \frac{-15165440}{-17537280}]$			$0 \xrightarrow{1} 1$
	$[\frac{-9404892128}{-119941739040}, \frac{-133062400}{-205451520}, \frac{-2016815840}{-9960752928}, \frac{-12793600}{-17537280}]$	$[\frac{43}{-32}, \frac{20}{-16}]$	$1 \xrightarrow{0} 0$	

input: $u = \overline{001}010110$, $p = 0$, $F_uV_p = \begin{bmatrix} -43 & -41 \\ 32 & 32 \end{bmatrix} = [-1.344, -1.281]$
$\arctan(F_uV_p) = [-0.931, -0.908]$
output: $v = \overline{01}0001$, $q = 1$, $F_vV_q = \begin{bmatrix} -31 & -14 \\ 32 & 16 \end{bmatrix} = [-0.969, -0.875]$
invariant matrix: $Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Table 8.3: The transcendent algorithm with the function $\arctan x$ in the binary signed system.

where

$$\begin{aligned} T_0^* x &= \begin{bmatrix} 0 & x \\ 1 & 1 \end{bmatrix} \\ T_n^* x &= \begin{bmatrix} 0 & n^2 x \\ 1 & 2n \end{bmatrix} \cdot \begin{bmatrix} 0 & n^2 x \\ 1 & 2n+1 \end{bmatrix} = \begin{bmatrix} n^2 x & n^2(2n+1)x \\ 2n & n^2 x + 2n(2n+1) \end{bmatrix} \\ &= \begin{bmatrix} nx & n(2n+1)x \\ 2 & nx + 4n + 2 \end{bmatrix} \end{aligned}$$

We have $\det(T_n^* x) = n^2 x^2$, so $T_n^* x$ is inreasing (or singular for $x = 0$). For a given interval $P = \left[\begin{smallmatrix} -a & a \\ 1 & 1 \end{smallmatrix}\right]$ take $Q = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ and n_0 with $2n_0 + 1 > a$. Then $T_n(P, Q) \subseteq Q$ and $T_n^* x$ is not contractive on Q.

8.4 Arithmetical expressions

Arithmetical algorithms considered so far can be combined to algorithms which compute arithmetical expressions like $x^2 + \arctan x$ or $\exp(x+y)/z$. An arithmetical expression can be parsed into a **circuit** which is an oriented graph whose vertices represent variables and either unary or binary arithmetical operations. Unary operations are either Möbius transformations or rational functions or transcendent functions. Binary operations are algebraic tensors.

Definition 8.18 *A circuit is a finite oriented graph (C, E), where C is the set of vertices and $E \subseteq C \times C$ is a set of edges. We assume the following properties*

1. *There are no loops (no paths from c to c for any $c \in C$).*
2. *The outdegree of each vertex is either 0, 1 or 2.*
3. *There exists exactly one vertex called root with indegree 0.*
4. *Each vertex (except the root) lies on a path which starts at the root.*
5. *The vertices with outdegree 1 are labelled either as unary or transcendent with a particular transcendent function.*
6. *The edges $c \to c'$ and $c \to c''$ of a vertex $c \in C$ with outdegree 2 are ordered.*

The leaves (vertices with outdegree 0) represent the variables, the vertices with outdegree 1 represent unary operations and the vertices with outdegree 2 represent the binary operations. We say that a vertex c' is an input to a vertex c if $c \to c'$ is an edge of the circuit. For example the expression $x^2 + \arctan x$ is represented by a circuit with states $C = \{0, 1, 2, 3\}$ and edges $1 \to 0$, $2 \to 0$, $3 \to 1$, $3 \to 2$. Here 0 is the leaf which represents the variable x, 1 represents the rational function x^2, 2 represents the transcendent function $\arctan x$, and 3 is the root and represents the addition $x + y$. A circuit represents an algorithm which computes at the root a function of n variables, where $n \geq 1$ is the number of leaves.

A circuit computes its function for a given a sofic number system (F, G, V) with an initialized graph $G = (B, E, \mathbf{i})$. During the computation it updates its compound state which consists of local states at vertices. A local state depends on the type of the vertex and has one of the following forms:
$(p \xrightarrow{u} q)$ (a finite path of the graph G) at leaves
$(Y, p \xrightarrow{u} q)$, where Y is a matrix, at vertices of type unary
$(X, p \xrightarrow{u} q)$, where X is an algebraic tensor, at vertices with outdegree 2
$(X, Y, n, p \xrightarrow{u} q)$, where X is an algebraic tensor, Y is a matrix and $n \in \mathbb{N}$, at vertices of type transcendent.

In all cases $p \xrightarrow{u} q$ is a finite path of G which represents the result computed by the algorithm at a given vertex c and not yet absorbed by vertices with input c. The path at a vertex may be of length 0, consisting from a single state $p \in B$. This is the case in the **initial compound state** whose all local states have the path **i**, the initial state of the graph G. A nondeterministic computation of a circuit is represented by a **circuit graph** whose vertices are compound states and whose edges are labeled by $(n+1)$-tuples $(u_0, \ldots, u_{n-1}, u_n) \in (A^*)^{n+1}$ of n inputs u_i (where n is the number of leaves) and one output u_n. The change of state of the i-th leaf from $p \xrightarrow{u} q$ to $p \xrightarrow{u} q \xrightarrow{v} r$ (i.e., to $p \xrightarrow{uv} r$) is an absorption edge with label $(\lambda, \ldots, v, \ldots, \lambda)$. The change of state of the root from $p \xrightarrow{uv} r = p \xrightarrow{u} q \xrightarrow{v} r$ to $q \xrightarrow{v} r$ is an emission edge with the label $(\lambda, \ldots, \lambda, u)$. Thus in the circuit graph we have labelled edges

$$\begin{aligned} s \xrightarrow{\lambda,\ldots,v,\ldots,\lambda} s' &\Leftrightarrow (p \xrightarrow{u} q) \to (p \xrightarrow{u} q \xrightarrow{v} r) \text{ at a leaf} \\ s \xrightarrow{\lambda,\ldots,\lambda,u} s' &\Leftrightarrow (X, \ldots, p \xrightarrow{u} q \xrightarrow{v} r) \to (X, \ldots, q \xrightarrow{v} r) \text{ at the root} \end{aligned}$$

There are also edges with label $(\lambda, \ldots, \lambda)$ which do not absorb any input nor emit any output but update the compound state by performing some steps of algorithms at particular vertices. These edges are of two types: absorptions and emissions. If at some vertex we have path $p \xrightarrow{u} q$ and the algorithm at this vertex performs an emission $q \xrightarrow{v} r$, then the state is changed to $p \xrightarrow{uv} r$. For example, consider a unary vertex $c \in C$ with the input vertex c'. Assume that state at c is $(X, p \xrightarrow{u} q)$ and the state at c' contains a path $p' \xrightarrow{u'} q'$ with source p'. If the unary algorithm has an emission edge $(X, p', q) \xrightarrow{\lambda,v} (F_v^{-1}X, p', r)$, then the state of c is changed to $(F_v^{-1}X, p \xrightarrow{uv} r)$. Similarly, assume that c is a vertex with outdegree 2 and input vertices c', c''. Assume that state at c is $(X, p \xrightarrow{u} q)$, the state at c' contains a path with source p' and the state at c'' contains a path with source p''. If the binary algorithm has an emission edge $(X, p', p'', q) \xrightarrow{\lambda,\lambda,v} (F_v^{-1}X, p', p'', r)$, then the state of c is changed to $(F_v^{-1}X, p \xrightarrow{uv} r)$. Analogously, an emission step of the transcendent algorithm changes the path at a vertex of type transcendent.

Finally there are compound state changes produced by absorption edges at the individual vertices. Let c be a vertex and let $c_1, \ldots, c_k$ be all vertices with input c. If the state of c contains a path $p \xrightarrow{u} q$, then all vertices c_i realize the absorption step with label $p \xrightarrow{u} q$ and the path of the state c is updated to q Thus if c_i is a vertex of type unary with state $(X, r \xrightarrow{v} s)$, then this state is updated to $(XH_{p,u,q}, r \xrightarrow{v} s)$. If c_i has outdegree 2 with state $(X, r \xrightarrow{v} s)$, then this state is updated either to $(X^*H_{p,u,q}, r \xrightarrow{v} s)$ or to $(X_*H_{p,u,q}, r \xrightarrow{v} s)$. We summarize that for the vertices of type unary we have state changes

$$\begin{aligned} (X, p \xrightarrow{u} q) \to (F_v^{-1}X, p \xrightarrow{uv} r) &\Leftrightarrow (X, p', q) \xrightarrow{\lambda,v} (F_v^{-1}X, p', r), \\ (X, p \xrightarrow{u} q) \to (XH_{p',u',q'}, p \xrightarrow{u} q) &\Leftrightarrow (X, p', q) \xrightarrow{u',\lambda} (XH_{p',u',q'}, q', q) \end{aligned}$$

provided $p' \xrightarrow{u'} q'$ in the input. For the vertex with outdegree 2 we have state changes

$$\begin{aligned} (X, p \xrightarrow{u} q) \to (F_v^{-1}X, p \xrightarrow{uv} r) &\Leftrightarrow (X, p', p'', q) \xrightarrow{\lambda,\lambda,v} (F_v^{-1}X, p', p'', r), \\ (X, p \xrightarrow{u} q) \to (X^*H_{p',u',q'}, p \xrightarrow{u} q) &\Leftrightarrow (X, p', p'', q) \xrightarrow{u',\lambda,\lambda} (X^*H_{p',u',q'}, q', p'', q) \\ (X, p \xrightarrow{u} q) \to (X_*H_{p'',u'',q''}, p \xrightarrow{u} q) &\Leftrightarrow (X, p', p'', q) \xrightarrow{\lambda,u'',\lambda} (X_*H_{p'',u'',q''}, p', q'', q) \end{aligned}$$

provided $p' \xrightarrow{u'} q'$ in the first input and $p'' \xrightarrow{u''} q''$ in the second input. For the vertex of type transcendent we have state changes

$$(X, Y, n, p \xrightarrow{u} q) \to (F_v^{-1}X, Y, n, p \xrightarrow{uv} r) \Leftrightarrow$$
$$(X, Y, n, p', q) \xrightarrow{\lambda, v} (F_v^{-1}X, Y, n, p', r),$$
$$(X, Y, n, p \xrightarrow{u} q) \to (X^* H_{p',u',q'}, Y H_{p',u',q'}, n, p \xrightarrow{u} q) \Leftrightarrow$$
$$(X, Y, n, p', q) \xrightarrow{u', \lambda} (X^* H_{p',u',q'}, Y^* H_{p',u',q'}, n, q', q)$$
$$(X, Y, n, p \xrightarrow{u} q) \to (X * (T_n^* Y), Y, n+1, p \xrightarrow{u} q) \Leftrightarrow$$
$$(X, Y, n, p', q) \xrightarrow{\lambda, \lambda} (X * (T_n^* Y), Y, n+1, p', q)$$

provided $p' \xrightarrow{u'} q'$ in the input While the emission edges (and t-absorption edges in the case of transcendent vertices) can be performed in any order, the absorptions must be done at each vertex with a given input c (sequentially or concurrently) and then the path of c is updated from $p' \xrightarrow{u'} q'$ to its target q'

The graph of a circuit with n leaves represents a nondeterministic algorithm for a (partial) function of n variables $f(x_0, \ldots, x_{n-1})$. If there is an infinite path with label $(u_0, \ldots, u_{n-1}, u_n) \in (A^\omega)^{n+1}$ from an initial state which represents f, then $\Phi(u_n) = f(\Phi(u_0), \ldots, \Phi(u_{n-1}))$. To obtain a deterministic algorithm, we have to use a selector which decides the sequence of particular actions on the base of the compound state.

8.5 Iterative algorithms

Many numerical algorithms are based on the iterative method. From an approximate solution is obtained a better approximation and this process is repeated. We say that x is a **fixed point** of a real function f if $f(x) = x$. The n-th iteration f^n of f is the composition of f with itself n times: we have a recurrent formula $f^{n+1}(x) = f(f^n(x))$. The convergence of a sequence $x_n = f^n(x)$ to a fixed point of f is pictured in Figure 8.4 left. For an initial x_0 we draw the vertical from the point $(x_0, 0)$ at the x-axis to the point (x_0, x_1) on the graph of f. Then we draw a horizontal line to the point (x_1, x_1) at the diagonal $y = x$, the vertical line to (x_1, x_2), etc.

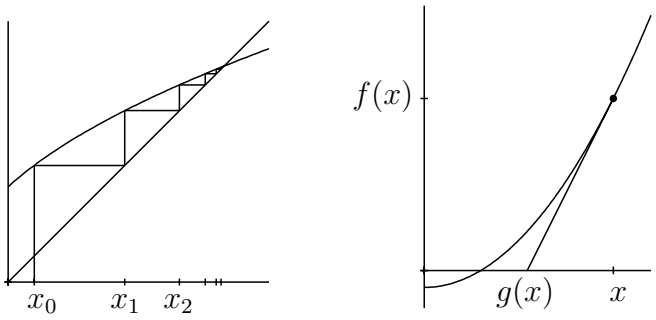


Figure 8.4: A stable fixed point (left) and the Newton iterative method (right).

Proposition 8.19 *Let f be a real function with a fixed point a and assume that f has a derivation at a with $|f'(a)| < 1$. Then there exists $\delta > 0$ such that for each x with $|x - a| < \delta$ we have $\lim_{n\to\infty} f^n(x) = a$.*

Proof: Take some q with $|f'(a)| < q < 1$. By the definition of derivative as a limit there exists $\delta > 0$ such that for any x with $0 \neq |x-a| < \delta$ we have $|\frac{f(x)-a}{x-a}| < q$, so $|f(x)-a| < q|x-a| < q\delta < \delta$. By induction we get $|f^n(x)-a| < q^n\delta$, so $\lim_{n\to\infty} f^n(x) = a$. □

We see from the proof that in a neighbourhood of the fixed point a, $|f^n(x)-a|$ is approximately a geometrical sequence with quotient $|f'(a)|$.

Proposition 8.20 *Let $f : [a,b] \to \mathbb{R}$ be a real differentiable function and $0 < q < 1$ be such that $|f'(x)| < q$ for each $x \in (a,b)$. If $(f(a)-a)(f(b)-b) < 0$ then f has a unique fixed point $c \in (a,b)$ and for each $x \in [a,b]$, we have $\lim_{n\to\infty} f^n(x) = c$.*

Proof: By the internmediate value theorem, f has a fixed point in (a,b). If it has two fixed points z, w in (a,b) then by the mean value theorem there exists x with $f'(x) = \frac{f(z)-f(w)}{z-w} = 1$ which is a contradiction. If $c \in (a,b)$ is the fixed point, and $x \in (a,b)$, then (by the mean value theorem) $|f(x)-a| < q|x-a|$, so $\lim_{n\to\infty} f^n(x) = a$. □

Propositions 8.19 and 8.20 hold as well with the standard (Euclidean) derivation $f'(x)$ replaced by the circle derivation $f^\bullet(x)$. A classical iterative algorithm is the **Newton iteration algorithm** for the solution of an equation $f(x) = 0$. If f is a differentiable real function and a is an approximate solution, then we approximate the function f by its tangent at a, i.e., by the linear function $h_a(x) = f(a) + f'(a)(x-a)$. We solve the equation $h_a(x) = 0$ to get $x = g(a) = a - f(a)/f'(a)$ (Figure 8.4 right). Thus for a given function f we iterate the function $g(x) = x - f(x)/f'(x)$. If the iteration process converges to a fixed point x of g with $g(x) = x$, then we have a solution of $f(x) = 0$. If f has a second derivation, then we get

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$$

Thus if a is a solution of $f(x) = 0$ then $g(a) = a$ and $g'(a) = 0$. This means that in a neighbourhood of the fixed point, $g^n(x)$ converge to a very fast: faster than any geometrical sequence.

Consider the task of finding a fixed point of a real function to an arbitrary precision. We assume that we have circuit (see Section 8.4) which computes the function f in a given sofic number system (F, G, V). Suppose that the assumptions of Proposition 8.20 are met, so there exists an interval I such that $f(x) - x$ have opposite signs at the endpoints of I and $|f^\bullet(x)| < q < 1$ for each $x \in I$. There exists a path $\mathbf{i} \xrightarrow{u} p$ such that $F_u V_p \subseteq I$ and $F_u V_p$ contains the fixed point of f. Since $|f^\bullet(x)| < q < 1$ in I, $F_u V_p$ is f-invariant, i.e., $f(F_u V_p) \subseteq F_u V_p$. This property can be verified by the circuit algorithm of f. It is satisfied provided there exists a path of the form $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{u,u} (X_u, p, p)$ in the circuit graph of f (here X is the compound state of the circuit which represents f). Once we have such an f-invariant path $\mathbf{i} \xrightarrow{u} p$, we can continue by induction. There exists an edge $p \xrightarrow{a} q$ such that $F_{ua} V_q \subseteq F_u V_p$ is f-invariant, i.e., there exists a path $(X_u, p, p) \xrightarrow{a,a} (X_{ua}, q, q)$ in the circuit graph. In this way we construct (in infinite time) an infinite path $(X, \mathbf{i}, \mathbf{i}) \xrightarrow{u,u}$ with $f(\Phi(u)) = \Phi(u)$. In Table 8.4 we see the computation of the stable fixed point of a hyperbolic Möbius transformation in the binary signed system.

input matrix: $\begin{bmatrix} 7 & 3 \\ 2 & 5 \end{bmatrix}$
fixed points: $1.822875655532, -0.822875655532$

u, p	$F_u V_p$	$F_u V_p$	$MF_u V_p$
$\overline{00}, 3$	$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$	$[1.0000, -1.0000]$	$[1.4286, -1.3333]$
$\overline{00}1, 2$	$\begin{bmatrix} 2 & 4 \\ 2 & 1 \end{bmatrix}$	$[1.0000, 4.0000]$	$[1.4286, 2.3846]$
$\overline{00}10, 1$	$\begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix}$	$[1.0000, 3.0000]$	$[1.4286, 2.1818]$
$\overline{00}100, 1$	$\begin{bmatrix} 3 & 5 \\ 2 & 2 \end{bmatrix}$	$[1.5000, 2.5000]$	$[1.6875, 2.0500]$
$\overline{00}100\overline{1}, 4$	$\begin{bmatrix} 6 & 15 \\ 4 & 8 \end{bmatrix}$	$[1.5000, 1.8750]$	$[1.6875, 1.8429]$
$\overline{00}100\overline{1}0, 1$	$\begin{bmatrix} 13 & 15 \\ 8 & 8 \end{bmatrix}$	$[1.6250, 1.8750]$	$[1.7424, 1.8429]$
$\overline{00}100\overline{1}01, 2$	$\begin{bmatrix} 57 & 30 \\ 32 & 16 \end{bmatrix}$	$[1.7812, 1.8750]$	$[1.8066, 1.8429]$
$\overline{00}100\overline{1}010, 1$	$\begin{bmatrix} 57 & 59 \\ 32 & 32 \end{bmatrix}$	$[1.7812, 1.8438]$	$[1.8066, 1.8309]$
$\overline{00}100\overline{1}0101, 2$	$\begin{bmatrix} 233 & 118 \\ 128 & 64 \end{bmatrix}$	$[1.8203, 1.8438]$	$[1.8219, 1.8309]$
$\overline{00}100\overline{1}01010, 1$	$\begin{bmatrix} 233 & 235 \\ 128 & 128 \end{bmatrix}$	$[1.8203, 1.8359]$	$[1.8219, 1.8279]$
$\overline{00}100\overline{1}01010\overline{1}, 4$	$\begin{bmatrix} 466 & 935 \\ 256 & 512 \end{bmatrix}$	$[1.8203, 1.8262]$	$[1.8219, 1.8242]$
$\overline{00}100\overline{1}01010\overline{11}, 4$	$\begin{bmatrix} 932 & 1867 \\ 512 & 1024 \end{bmatrix}$	$[1.8203, 1.8232]$	$[1.8219, 1.8230]$
$\overline{00}100\overline{1}01010\overline{11}0, 1$	$\begin{bmatrix} 1865 & 1867 \\ 1024 & 1024 \end{bmatrix}$	$[1.8213, 1.8232]$	$[1.8223, 1.8230]$
$\overline{00}100\overline{1}01010\overline{11}01, 2$	$\begin{bmatrix} 7465 & 3734 \\ 4096 & 2048 \end{bmatrix}$	$[1.8225, 1.8232]$	$[1.8227, 1.8230]$
$\overline{00}100\overline{1}01010\overline{11}010, 1$	$\begin{bmatrix} 7465 & 7467 \\ 4096 & 4096 \end{bmatrix}$	$[1.8225, 1.8230]$	$[1.8227, 1.8229]$
$\overline{00}100\overline{1}01010\overline{11}0101, 2$	$\begin{bmatrix} 29865 & 14934 \\ 16384 & 8192 \end{bmatrix}$	$[1.8228, 1.8230]$	$[1.8229, 1.8229]$
$\overline{00}100\overline{1}01010\overline{11}01010, 1$	$\begin{bmatrix} 29865 & 29867 \\ 16384 & 16384 \end{bmatrix}$	$[1.8228, 1.8229]$	$[1.8229, 1.8229]$
$\overline{00}100\overline{1}01010\overline{11}010100, 1$	$\begin{bmatrix} 59731 & 59733 \\ 32768 & 32768 \end{bmatrix}$	$[1.8228, 1.8229]$	$[1.8229, 1.8229]$

Table 8.4: The computation of the stable fixed point in the binary signed system.

Bibliography

[1] A. Avizienis. Signed-digit number representations for fast parallel arithmetic. *IRE Trans. Electron. Comput.*, EC-10:389–400, 1961.

[2] G. A. Baker, Jr. and P. Graves-Morris. *Padé Approximants*, volume 59 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, New York, 1996.

[3] M. P. Béal. Symbolic dynamics and finite automata. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 2, pages 483–503. Springer-Verlag, Berlin, 1997.

[4] A. F. Beardon. *The geometry of discrete groups*. Springer-Verlag, Berlin, 1995.

[5] V. Berthé and M. Rigo, editors. *Combinatorics, Automata and Number Theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010.

[6] R. L. Bishop and S. I. Goldberg. *Tensor analysis on manifolds*. Dover Publications, New York, 1980.

[7] A. Cauchy. Sur les moyens d'éviter les erreurs dans les calculs numériques. *Comptes Rendus d l'Académie des Sciences*, pages 789–798, 1840.

[8] C. Y. Chow and J. E. Robertson. Logical design of a redundant binary adder. In *IEEE 4th Symposium on Computer Arithmetic*, pages 109–115. IEEE Computer Society, 1978.

[9] H. S. M. Coxeter. *Projective geometry*. Springer-Verlag, Berlin, 2003.

[10] M. d. Ercegovac and T. Lang. *Digital arithmetic*. Morgan Kaufmann, 2003.

[11] K. Dajani and C. Kraaikamp. From greedy to lazy expansions and their driving dynamics. *Expositiones Mathematicae*, 20(4):315–327, 2002.

[12] M. Delacourt and P. Kůrka. Finite state transducers for modular Möbius number systems. In B. Rovan, V. Sassone, and P. Widmayer, editors, *MFCS 2012*, volume 7464 of *LNCS*, pages 323–334. Springer-Verlag, 2012.

[13] D. Dombek. *Non-standard representations of numbers*. PhD thesis, Czech Technical University in Prague, Prague, 2014.

[14] G. A. Edgar. *Measure, Topology, and Fractal Geometry*. Undergraduate Texts in Mathematics. Springer-Verlag, Berlin, 1990.

[15] Ch. Frougny. Parallel and on-line addition in negative base and some complex number systems. In *Proc. Euro-Par 96 Lyon*, volume 1124 of *LNCS*, pages 175–182, 1996.

[16] Ch. Frougny. On-the-fly algorithms and sequential machines. In *13th IEEE Symposium on Computer Arithmetic*, pages 260–265. IEEE Computer Society, 1997.

[17] Ch. Frougny, P. Heller, E. Pelantová, and M. Svobodová. k-block parallel addition versus 1-block parallel addition in non-standard numeration systems. *Theor. Comput. Sci.*, 543:52–67, 2014.

[18] Ch. Frougny, E. Pelantová, and M. Svobodová. Parallel addition in non-standard numeration systems. *Theoretical Computer Science*, 412:5714–5727, 2011.

[19] Ch. Frougny, E. Pelantová, and M. Svobodová. Minimal digit sets for parallel addition in non-standard numeration systems. *Journal of Integer Sequences*, 16(2):1–36, 2013. Article 13.2.17.

[20] Ch. Frougny and B. Solomyak. Finite beta-expansions. *Ergodic Theory and Dynamical Systems*, 12(4):713–723, 1992.

[21] R. W. Gosper. Continued fractions arithmetic. *Unpublished manuscript*, 1977. http://www.tweedledum.com/rwg/cfup.htm.

[22] R. Heckmann. Big integers and complexity issues in exact real arithmetic. *Electr. Notes Theor. Comput. Sci.*, 13, 1998.

[23] T. Hejda, Z. Masáková, and E. Pelantová. Greedy and lazy representations in negative base systems. *Kybernetika (Prague)*, 49(2):258–279, 2013.

[24] J. G. Hocking and G. S. Young. *Topology*. Dover Publications, New York, 1961.

[25] M. Iosifescu and C. Kraaikamp. *Metrical theory of continued fractions*, volume 547 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.

[26] K. Ireland and M. Rosen. *A classical introduction to modern number theory*. Graduate Texts in Mathematics. Springer-Verlag, Berlin, 1990.

[27] W. B. Jones and W. J. Thron. *Continued fractions*, volume 11 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, New York, 1984.

[28] S. Katok. *Fuchsian Groups*. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1992.

[29] A. Kazda. Convergence in Möbius number systems. *Integers*, 2:261–279, 2009.

[30] D. E. Knuth. *The art of computer programming. Seminumerical algorithms*, volume 2. Addison-Wesley, Reading, MA, 1981.

[31] M. Konečný. Real functions incrementally computable by finite automata. *Theoretical Computer Science*, 315(1):109–133, 2004.

[32] I. Koren. *Computer arithmetic algorithms*. Prentice Hall, 1993.

[33] P. Kornerup and D. W. Matula. *Finite precision number systems and arithmetic*. Cambridge University Press, Cambridge, 2010.

[34] P. Kornerup and D. W. Matula. An algorithm for redundant binary bit-pipelined rational arithmetic. *IEEE Transactions on Computers*, 39(8):1106–1115, August 1990.

[35] P. Kůrka. *Topological and symbolic dynamics*, volume 11 of *Cours spécialisés*. Société Mathématique de France, Paris, 2003.

[36] P. Kůrka. A symbolic representation of the real Möbius group. *Nonlinearity*, 21:613–623, 2008.

[37] P. Kůrka. Möbius number systems with sofic subshifts. *Nonlinearity*, 22:437–456, 2009.

[38] P. Kůrka. Expansion of rational numbers in Möbius number systems. In S. Kolyada, Y. Manin, and M. Moller, editors, *Dynamical Numbers: Interplay between Dynamical Systems and Number Theory*, volume 532 of *Contemporary Mathematics*, pages 67–82. American Mathematical Society, 2010.

[39] P. Kůrka. Stern-Brocot graph in Möbius number systems. *Nonlinearity*, 25:57–72, 2012.

[40] P. Kůrka. Exact real arithmetic for interval number systems. *Theoretical Computer Science*, 542:32–43, 2014.

[41] P. Kůrka. The exact real arithmetical algorithm in binary continued fractions. In *2015 IEEE 22nd Symposium on Computer Arithmetic ARITH-22*, pages 168–175. IEEE Computer Society, 2015.

[42] P. Kůrka. Fast arithmetical algorithms in Möbius number systems. *IEEE Transactions on computers*, 61(8):1097–1109, August 2012.

[43] P. Kůrka and M. Delacourt. The unary arithmetical algorithm in bimodular number systems. In *2013 IEEE 21st Symposium on Computer Arithmetic ARITH-21*, pages 127–134. IEEE Computer Society, 2013.

[44] P. Kůrka and A. Kazda. Möbius number systems based on interval covers. *Nonlinearity*, 23:1031–1046, 2010.

[45] P. Kůrka and T. Vávra. Analytic functions computable by finite state transducers. In M. Holzer and M. Kutrib, editors, *Implementation and Application of Automata*, volume 8587 of *LNCS*, pages 252–263. Springer-Verlag, 2014.

[46] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, Cambridge, 1995.

[47] M. Lothaire. *Combinatorics on words*, volume 17 of *Encyclopedia of mathematics and its applications*. Addison-Wesley, 1983.

[48] M. Lothaire. *Algebraic Combinatorics on words*, volume 90 of *Encyclopedia of mathematics and its applications*. Cambridge University Press, 2002.

[49] Z. Masáková and E. Pelantová. Purely periodic expansions in systems with negative base. *Acta Mathematica Hungarica*, 139(3):208–227, 2013.

[50] Z. Masáková, E. Pelantová, and T. Vávra. Arithmetics in number systems with negative base. *Theoretical Computer Science*, 412(8–10):835–845, 2011.

[51] D. Micciancio and S. Goldwasser. *Complexity of lattice problems: A cryptographic perspective*. Kluwer Academic Publishers, 2002.

[52] M. Niqui. Exact real arithmetic on the Stern-Brocot tree. *J. Discrete Algorithms*, 5(2):356–379, 2007.

[53] W. Parry. On the β-expansions of real numbers. *Acta Mathematica Academiae Scientiarum Hungaricae*, 11:401–416, 1960.

[54] O. Perron. *Die Lehre von den Kettenbrüchen*. B.G.Teubner, 1913.

[55] P. J. Potts. *Exact real arithmetic using Möbius transformations*. PhD thesis, University of London, Imperial College, London, 1998.

[56] P. J. Potts, A. Edalat, and M. H. Escardó. Semantics of exact real computation. In *Proceedings of the twelfth annual IEEE symposium in computer science*, pages 248–257, Warsaw, 1997.

[57] G. N. Raney. On contiuned fractions and finite automata. *Mathematische Annalen*, 206:265–283, 1973.

[58] A. Rényi. Representations for real numbers and their ergodic properties. *Acta Mathematica Academiae Scientiarum Hungaricae*, 8:477–493, 1957.

[59] M. Rigo. *Formal languages, automata and numeration systems 1 Introduction to combinatorics on words*. Wiley, 2014.

[60] M. Rigo. *Formal languages, automata and numeration systems 2 Applications to recognizability and decidability*. Wiley, 2014.

[61] K. Schmidt. On periodic expansions of Pisot numbers and Salem numbers. *Bulletin of the London Mathematical Society*, 12(4):269–278, 1980.

[62] R. A. Silverman. *Introductory complex analysis*. Dover Publications, New York, 1972.

[63] W. Steiner. Digital expansions with negative real bases. *Acta Mathematica Hungarica*, 139(1–2):106–119, 2013.

[64] K. S. Trivedi and M. D. Ercegovac. On-line algorithms for division and multiplication. *IEEE Transactions on Computers*, C-26(7):681–687, July 1977.

[65] B. L. van der Waerden. *Algebra*, volume I. Springer-Verlag, Berlin, 2003.

[66] J. E. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Transactions on Computers*, 39(8):1087–1105, August 1990.

[67] H. S. Wall. *Analytic theory of continued fractions*. AMS Chelsea Publishing, Providence, 2000.

[68] K. Weihrauch. *Computable analysis. An introduction*. EATCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin, 2000.

[69] B. Weiss. Subshifts of finite type and sofic systems. *Monatshefte für Mathematik*, 77:462–474, August 1990.

Notation

A^ω	the set of infinite words	10
$\mathbb{C}$	the complex plane	10
$\mathbf{d}$	stereographic projection	11
$\overline{\mathbb{R}}$	the extended real line	25
$H_{p,a,q}$	cut matrix	99
$\mathbf{L}$	upper length quotient	88
$\mathcal{L}_D$	the language of forbidden set	10
$\mathcal{L}_{F,W}$	expansion language	72
$\mathbf{l}$	lower length quotient	88
$M^\bullet(x)$	circle derivation	49
$\mathrm{nrm}(M)$	norm of a projective matrix	64
$\overline{Y}$	the closure of a set	25
P^c	closed interval	96
P^o	open interval	96
Φ	value mapping	69
$\mathbb{R}^n$	n-dimensional Euclidean space	25
$\mathbf{R}(M)$	rational expansion interval	119
$\mathbb{S}$	the unit circle	10
Σ_D	the subshift of forbidden set	10
$\mathcal{S}_{F,W}$	expansion subshift	72
$\mathbf{s}(M)$	stable point	65
$\mathrm{sz}(P)$	size of a projective matrix	98
$\mathrm{trc}(M)$	trace of a projective matrix	50
$\mathbf{U}(M)$	expanding interval	50
$\mathbf{u}(M)$	unstable point	65
$\mathbf{V}(M)$	contracting interval	50
$\mathbb{X}_F$	convergence space	69
Y°	the interior of a set	25

Index