Electronic Notes in Theoretical Computer Science 6 (1998)
URL: http://www.elsevier.nl/locate/entcs/volume6.html 15 pages

A New Representation for Exact Real Numbers [1]

Abbas Edalat Peter John Potts

Department of Computing
Imperial College of Science, Technology, and Medicine
London SW7 2BZ, England

E. N. T. C. S.

Elsevier Science B. V.

Abstract

We develop the theoretical foundation of a new representation of real numbers based on the infinite composition of linear fractional transformations (lft), equivalently the infinite product of matrices, with non-negative coefficients. Any rational interval in the one point compactification of the real line, represented by the unit circle S^1, is expressed as the image of the base interval $[0, \infty]$ under an lft. A sequence of shrinking nested intervals is then represented by an infinite product of matrices with integer coefficients such that the first so-called sign matrix determines an interval on which the real number lies. The subsequent so-called digit matrices have non-negative integer coefficients and successively refine that interval. Based on the classification of lft's according to their conjugacy classes and their geometric dynamics, we show that there is a canonical choice of four sign matrices which are generated by rotation of S^1 by $\pi/4$. Furthermore, the ordinary signed digit representation of real numbers in a given base induces a canonical choice of digit matrices.

1 Introduction

It is well-known that in floating-point computation the accumulation of round-off errors can lead to highly inaccurate results. Interval analysis [11] has been used to partially circumvent this problem by maintaining a pair of bounding floating-point numbers that is guaranteed to contain the real number or interval in question. However, this interval can get unjustifiably large and thereby convey very little information.

In principle, exact real arithmetic provides an alternative technique for real number computation and the verification of numerical algorithms: Here, in contrast to the fixed point or the floating point format, a real number is

[1] This work has been supported by EPSRC at Imperial College.

represented by an *infinite sequence* of, say, digits, such that each digit gives a better approximation to the real number in question. Any algorithm for the computation of a function takes one such infinite sequence as input and outputs another infinite sequence. Any element of the output sequence must be obtained by reading some finite segment of the input sequence. This basic criterion rules out the ordinary decimal representation of real numbers for exact computation. Indeed, it is easy to see that for example multiplication by 3 is not computable in decimal representation: Given the input $.3333\cdots$ one cannot produce the first element of the output sequence. This first digit should be 0 if after a finite sequence of 3's a digit less than or equal to 2 occurs and should be 1 if eventually a digit greater or equal to 4 occurs. It is well-known that the same problem exists in any other base. This is due to the fact that the ordinary digit representation of real numbers in a given base has no non-trivial redundancy: Apart from some exceptional numbers, the representation of real numbers is unique.

One therefore has to look for a redundant representation of real numbers. The most well-known example is the signed digit system in a given base. The signed binary system, for example, is generated by the digits $\{-1, 0, 1\}$. Any number r, say in $[-1, 1]$, can be represented as an infinite sequence $.d_1d_2d_3\cdots$ with

$$r = \frac{d_1}{2} + \frac{d_2}{2^2} + \frac{d_3}{2^3}\cdots$$

The representation is redundant and can be expressed by the three affine maps

$$\begin{aligned} f_k : [-1,1] &\to [-1,1] \\ x \quad &\mapsto \frac{x+k}{2}, \end{aligned}$$

for $k = -1, 0, 1$. In fact, for the real number r above we have:

$$\{r\} = \bigcap_{n\geq 1} f_{d_1} f_{d_2} \cdots f_{d_n}[-1, 1].$$

We can therefore identify the expansion $.d_1d_2d_3\cdots$ of r with the infinite composition $f_{d_1} f_{d_2}\cdots$ generated by the iterated function system (IFS) f_{-1}, f_0, f_1 on $[-1, 1]$. Avizienis [1] and Wiedmer [17] have developed efficient algorithms for basic arithmetic operations in the signed digit system. The signed binary system is also the basis of the first extension of PCF with a real number data type in the work of Di Gianantonio [4,5].

In the late 1980's two other frameworks for exact real number computation were proposed. In the Boehm and Cartwright's approach [3,2], developed and implemented recently by Valerie Menissier [10], a computable real number is approximated by B-adic numbers of the form k/B^n where B is the base, n is a natural number and k is an integer. For any basic function in analysis a feasible algorithm has been presented in order to produce an approximation to the value of the function at a given computable real number up to any threshold of accuracy. However, the computation is not incremental in the sense that to obtain any more accurate approximation one has to compute from scratch.

Kornerup and Matula [9] and Vuillemin [16], proposed a representation of computable real numbers by redundant continued fractions and presented various incremental algorithms for basic arithmetic operations using the earlier work of Gosper [7] and for some transcendental functions.

Any continued fraction expansion of a real number can be expressed as an infinite composition of linear fractional transformations (lft) of the form

$$f : x \mapsto \frac{ax+c}{bx+d} : \mathbb{R}^* \to \mathbb{R}^*, \tag{1}$$

where $\mathbb{R}^*$ is the real line extended with the point at infinity and $a, b, c, d \in \mathbb{Z}$. In fact, any continued fraction expansion

$$r = a_0 + \cfrac{b_0}{a_1 + \cfrac{b_1}{a_2 + \cfrac{b_2}{a_3 + \ddots}}}$$

of a real number r can be expressed as $r = \phi_0(r_0)$ with

$$r_0 = a_1 + \cfrac{b_1}{a_2 + \cfrac{b_2}{a_3 + \ddots}}$$

and $\phi_0(x) = a_0 + \frac{b_0}{x}$. Iterating the above scheme, we obtain $r = \phi_0\phi_1 \cdots \phi_n(r_n)$ with

$$r_n = a_{n+1} + \cfrac{b_{n+1}}{a_{n+2} + \cfrac{b_{n+2}}{a_{n+3} + \ddots}}$$

and $\phi_i(x) = a_i + \frac{b_i}{x}$ for $0 \le i \le n$. One can therefore identify the original continued fraction for r with the infinite composition $\phi_0\phi_1\phi_2\cdots$. Such a representation of real numbers was already present in [16]. Nielson and Kornerup [12] later developed a general framework for exact arithmetic by representing real numbers by redundant infinite composition of linear fractional transformations (lft). Escardo's extension of PCF [6] is based on the redundant representation of a real number in $[0, 1]$ as an infinite composition of contracting affine maps $f : [0, 1] \to [0, 1]$ with rational coefficients; these maps represent a particular class of lft's.

The search for an incremental and efficient framework for real number computation is a challenging exercise. A new, feasible and incremental representation of real numbers based on the composition of linear fractional transformations with non-negative integer coefficients has been introduced in [14,15]; it has provided efficient algorithms for exact computation of all elementary functions. This has also led to an extension of PCF with a real number data

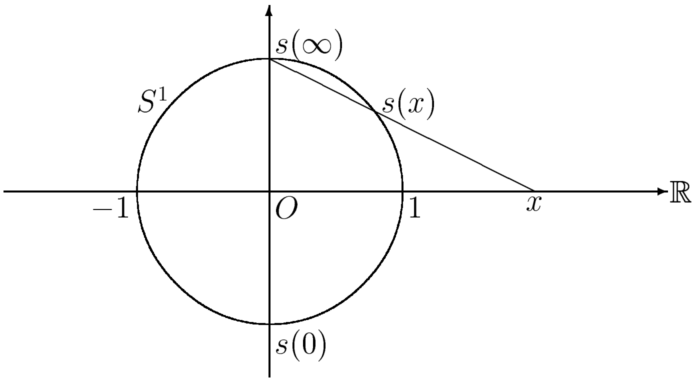

Fig. 1. Stereographic projection.

type interpreted as the domain of intervals of the one point compactification of non-negative real numbers [13].

In this paper, we examine the theoretical basis of the new representation of real numbers and show that with reasonable assumptions the framework in [14,15] is a most suitable framework for exact real arithmetic.

2 The new representation of real numbers

There are a number of equivalent definitions of a computable real number. The most convenient one for us is to consider a real number as the intersection of a shrinking nested sequence of rational intervals; we then say that the real number is computable if there is a master program which generates all these rational intervals. It follows from the definition that the usual predicates such as $=$, $\leq$ and $<$ on computable real numbers are not decidable. One consequence is that, since there is no test for zero, we cannot avoid dividing say 1 by 0. Therefore, any suitable framework for exact real arithmetic must regard ∞ as an ordinary real number. We therefore work with the extended real line $\mathbb{R}^*$ regarded as the one-point compactification of the real line. A simple model for $\mathbb{R}^*$ is the unit circle S^1 in the plane with its centre at the origin equipped with the subspace Euclidean topology of the plane. Given any point $x \in \mathbb{R}$ lying on the horizontal axis, the line joining the top point of S^1 and x intersects S^1 at a unique point $s(x)$ as in Figure 1. We define $s(\infty)$ to be the top point of S^1. Then the map $s : \mathbb{R}^* \to S^1$ is a homeomorphism and is called the stereographic projection.

The usual ordering of the real numbers induces the anti-clockwise orientation on S^1. The interval $[a, b] \subset S^1$ is defined to be the closed arc going anti-clockwise from a to b. With this convention, the interval $[a, a]$ denotes the singleton $\{a\}$ rather than $\mathbb{R}^*$. An obvious metric on $\mathbb{R}^*$, compatible with its topology, is given by taking the distance between $x, y \in \mathbb{R}^*$ to be the length of the shorter arc connecting $s(x)$ and $s(y)$ on S^1. However, this distance involves transcendental functions. A suitable metric ρ on $\mathbb{R}^*$ which is easy to compute is defined as follows. For extended reals x and y which are both non-negative or both non-positive, we put

$$\rho(x,y)=\left|\frac{|x|-1}{|x|+1}-\frac{|y|-1}{|y|+1}\right|.$$

Otherwise, if x and y have different signs, then

$$\rho(x,y)=\min(\rho(x,0)+\rho(0,y),\rho(x,\infty)+\rho(\infty,y)).$$

Similar to terms like $1/0$, we also cannot avoid expressions such as $\infty-\infty$, $0/0$ and 0^0 which must all be denoted by $\perp=\mathbb{R}^*$. This leads us naturally to the domain $\mathbf{I}\mathbb{R}^*=\{[a,b]\subset\mathbb{R}^*\}\cup\{\mathbb{R}^*\}$ of the intervals of $\mathbb{R}^*$ ordered by reverse inclusion. Any continuous function $f:\mathbb{R}^*\to\mathbb{R}^*$ has a canonical extension $\hat{f}:\mathbf{I}\mathbb{R}^*\to\mathbf{I}\mathbb{R}^*$, given by $\hat{f}(A)=f(A)=\{f(x)|x\in A\}$. For convenience, we always write $\hat{f}$ simply as f and often denote $f(A)$ simply by fA.

We will use the class of lft's or Möbius transformations with real coefficients to encode any sequence of shrinking nested intervals and, hence, any real number. The choice of lft's for this purpose is crucial to develop efficient and elegant algorithms for all elementary functions in this framework [14,15]. The set of all real lft's, denoted by $\mathbb{M}$, consists of maps f given in Equation 1 with $a,b,c,d\in\mathbb{R}$ and $ad-bc\neq 0$. An lft is a homeomorphism of $\mathbb{R}^*$; it is orientation preserving if $ad-bc>$ and orientation reversing if $ad-bc<0$. Sometimes we need to consider the extension $\tilde{f}:z\mapsto\dfrac{az+c}{bz+d}:\mathbb{C}^*\to\mathbb{C}^*$ of f to the extended complex plane $\mathbb{C}^*$.

We recall some elementary properties of $\mathbb{M}$ which are similar to those of complex lft's given for example in [8]. Under composition of maps, $\mathbb{M}$ is a group of homeomorphisms of $\mathbb{R}^*$. If $GL(2,\mathbb{R})$ denotes the general linear group of 2×2 non-singular matrices with real coefficients, then the mapping $\Theta:GL(2,\mathbb{R})\to\mathbb{M}$ which maps the matrix $\begin{pmatrix}a&c\\b&d\end{pmatrix}$ to the lft ϕ with $\phi(x)=\dfrac{ax+c}{bx+d}$ is a group-homomorphism. The kernel K of Θ consists of all matrices of the form λI where $\lambda\neq 0$ and I is the identity matrix. Therefore, $\mathbb{M}\cong GL(2,\mathbb{R})/K$. All this means that we can identify any lft up to scaling with a 2×2 matrix. Furthermore, $\mathbb{R}^*$ can be identified with the projective real line, i.e. the set of one dimensional subspaces of $\mathbb{R}^2$. In fact, any such subspace V is spanned by a vector $v=\begin{pmatrix}k\\l\end{pmatrix}\in V$ with $k,l\in\mathbb{R}$ not both zero. The ratio $k/l\in\mathbb{R}^*$ is independent of the choice of $v\in V$. Hence, one can identify V with k/l. The vector $\begin{pmatrix}k\\l\end{pmatrix}$ is said to represent $x=k/l\in\mathbb{R}^*$ in *homogeneous coordinates*. We can normalise this vector by dividing it by $\sqrt{k^2+l^2}$ to obtain $\begin{pmatrix}\sin\alpha\\\cos\alpha\end{pmatrix}$, where $0\leq\alpha<\pi$ with $\tan\alpha=k/l$. In Figure 1, α is represented as the angle $s(0)\widehat{s(\infty)}s(x)$, between the line segments $s(\infty)s(0)$ and $s(\infty)s(x)$. Note that the angle $s\widehat{(0)O}s(x)$, between the line segments $Os(0)$ and $Os(x)$ is, 2α. Therefore, as α increases from 0 to π, x increases from 0 to infinity and back through negative numbers to 0, while $s(x)$ goes from $s(0)$ anticlockwise once around S^1. The action of an lft in these coordinates is reduced to matrix multiplication. Indeed, for the lft ϕ above, we have $\phi(\frac{k}{l}) = \frac{ak+cl}{bk+dl}$, which in homogeneous coordinates can be simply written as multiplication by a representative matrix:

$$\begin{pmatrix} k \\ l \end{pmatrix} \mapsto \begin{pmatrix} ak+cl \\ bk+dl \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} k \\ l \end{pmatrix}.$$

Therefore, we can freely move, on the one hand, between $k/l \in \mathbb{R}^*$ and its homogeneous representation $\begin{pmatrix} k \\ l \end{pmatrix}$ and on the other, between the lft $x \mapsto \dfrac{ax+c}{bx+d}$ and its matrix representation $\begin{pmatrix} a & c \\ b & d \end{pmatrix}$. In both cases the representation is unique up to scaling.

A basic property of the group $\mathbb{M}$ is that it is 3-transitive. This means that for any pair of distinct triples (x_1, x_2, x_3) and (y_1, y_2, y_3) with $x_i, y_i \in \mathbb{R}^*$ $(i = 1, 2, 3)$ there exists a unique lft $\phi \in \mathbb{M}$ with $y_i = \phi(x_i)$ for $i = 1, 2, 3$. An immediate consequence is the following property.

Proposition 2.1 *Given two non-trivial intervals $[p, q]$ and $[r, s]$ with $p \neq q$ and $r \neq s$, there exists an lft $\phi \in \mathbb{M}$ with $\phi([p, q]) = [r, s]$.*

It follows that if we fix a base interval, then we can express, or encode, all other non-trivial intervals as the image of this base interval under an lft. The most efficient base interval is $[0, \infty]$ as no computation is needed to determine the lft in the proposition. Indeed $x \mapsto \dfrac{rx+s}{x+1}$ and $\mapsto \dfrac{sx+r}{x+1}$ both map $[0, \infty]$ to $[r, s]$, the first reverses the orientation whilst the second preserves it. Furthermore, if $[r, s]$ is a rational interval $[\dfrac{a}{b}, \dfrac{c}{d}]$, then the maps $x \mapsto \dfrac{ax+c}{bx+d}$ and $x \mapsto \dfrac{cx+a}{dx+b}$ have integer coefficients and map $[0, \infty]$ onto $[\dfrac{a}{b}, \dfrac{c}{d}]$ respectively reversing and preserving the orientation. Note that, if k and l are positive integers, the maps $x \mapsto \dfrac{kax+lc}{kbx+ld}$ and $x \mapsto \dfrac{kcx+la}{kdx+lb}$ will satisfy the same property. However, the lft will be unique up to change of orientation if we require that the sum of the absolute values of its coefficients be minimal.

3 Refining semi-groups of lft's

In this section, we study refinement of intervals by lft's. An lft $\phi \in \mathbb{M}$ *refines* an interval $[p, q] \subset \mathbb{R}^*$ if $\phi[p, q] \subseteq [p, q]$. Clearly, the identity transformation refines any interval. We say a subset $S \subseteq \mathbb{M}$ refines $[p, q]$ if each element of S refines $[p, q]$. Clearly, if S refines $[p, q]$, then the semi-group generated by S also refines $[p, q]$. It follows that there is a largest sub-semigroup of $\mathbb{M}$ which refines a given interval.

Consider the interval $[0,\infty]$. Let $\mathbb{M}^+ \subseteq \mathbb{M}$ be the set of lft's whose coefficients are all non-negative or, equivalently, all non-positive. It is easy to see that $\mathbb{M}^+$ is the refining semi-group of $[0,\infty]$. Furthermore, we have:

Proposition 3.1 *$[0,\infty]$ is the unique interval which has $\mathbb{M}^+$ as its refining semi-group.*

Proof. Suppose $\mathbb{M}^+$ is the refining semi-group of an interval $A \subset \mathbb{R}^*$ and let $A = \phi[0,\infty]$ with $\phi : x \mapsto \dfrac{ax+c}{bx+d}$. We have, by assumption, $\psi A \subseteq A$ for all $\psi \in \mathbb{M}^+$. Therefore, $\phi^{-1}\psi\phi[0,\infty] \subseteq [0,\infty]$ for all $\psi \in \mathbb{M}^+$, or equivalently, $\gamma = \phi^{-1}\psi\phi \in \mathbb{M}^+$ for all $\psi \in \mathbb{M}^+$. Taking ψ to be the map $x \mapsto \dfrac{1}{x}$ we get

$$\gamma : x \mapsto \frac{(-ac+bd)x + d^2 - c^2}{(a^2-b^2)x + ac - bd},$$

which implies $ac - bd = 0$. Taking ϕ to be the map $x \mapsto x+1$, we get

$$\gamma : x \mapsto \frac{(ad+bd-bc)x + d^2}{-b^2x - bc - bd + ad},$$

which means that either $b=0$ or $d=0$. We cannot have $b=d=0$ since this together with $ac=bd$ implies $ad-bc=0$. If $b=0$ and $d\neq 0$, then $c=0$, $d \neq 0$ and $ad > 0$. Thus $\phi(x) = \dfrac{ax}{d}$ and therefore $A = \phi[0,\infty] = [0,\infty]$. If, on the other hand, $d=0$ and $b \neq 0$, then $a=0$, $c \neq 0$ and $bc > 0$. Hence, $\phi(x) = \dfrac{c}{bx}$ and we have again $A = \phi[0,\infty] = [0,\infty]$. □

What happens if we change the base interval? If B is any non-trivial interval, then $B = \phi[0,\infty]$ for some lft $\phi \in \mathbb{M}$ and we have:

Corollary 3.2 *The refining semi-group of $B = \phi[0,\infty]$ is given by*

$$\phi\mathbb{M}^+\phi^{-1} = \{\phi\psi\phi^{-1} | \psi \in \mathbb{M}^+\},$$

in other words, $\gamma \in \mathbb{M}$ refines B iff $\phi^{-1}\gamma\phi \in \mathbb{M}^+$.

The above corollary implies that the refining semi-group of the interval $[0,\infty]$, namely $\mathbb{M}^+$, is distinguished by having the simplest characterization.

We now consider $[0,\infty]$ as the base interval and characterize the refinement of intervals when they are expressed as images of $[0,\infty]$ under lft's.

Proposition 3.3 *For lft's ϕ and ψ we have $\phi[0,\infty] \supseteq \psi[0,\infty]$ iff $\psi = \phi\gamma$ for some $\gamma \in \mathbb{M}^+$.*

It follows that for any shrinking sequence of nested intervals $[p_0,q_0] \supseteq [p_1,q_1] \supseteq [p_2,q_2] \supseteq \cdots$ we have $[p_n,q_n] = \phi_0\phi_1\cdots\phi_n[0,\infty]$ where $\phi_0 \in \mathbb{M}$ and $\phi_i \in \mathbb{M}^+$ for $1 \leq i \leq n$. Therefore, the sequence can be expressed as an infinite composition of lft's, or equivalently infinite product of matrices, $\phi_0\phi_1\phi_2\cdots$.

A rational number can be represented by a vector with integer coefficients. A finite product $\phi_0\phi_1\phi_2\cdots\phi_n$, with $\phi_0 \in \mathbb{M}$, $\phi_i \in \mathbb{M}^+$ for $1 \leq i \leq n-1$ and ϕ_n a vector with non-negative coefficients, represents a finite sequence of shrinking nested intervals collapsing into a rational number.

We have therefore shown that any real number can be represented as the intersection $\bigcap_{n\geq 0} \phi_0\phi_1\phi_2\cdots\phi_n[0,\infty]$ with $\phi_0 \in \mathbb{M}$ and $\phi_i \in \mathbb{M}^+$ $(i \geq 1)$ such that ϕ_n has integer coefficients for all $n \geq 0$. If $\phi_n : x \mapsto \dfrac{ax+c}{bx+d}$, then in matrix notation, the real number can be expressed as the infinite product

$$\begin{pmatrix} a_0 & c_0 \\ b_0 & d_0 \end{pmatrix} \begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & c_2 \\ b_2 & d_2 \end{pmatrix} \begin{pmatrix} a_3 & c_3 \\ b_3 & d_3 \end{pmatrix} \cdots.$$

We call this a *normal product*. It gives a simple representation and a convenient operational semantics for the lazy representation of the computable reals: finite segments of the above matrix product give incremental interval approximations to the real number represented by the matrix product. More specifically the first matrix tells us that the result is contained in the interval $[\frac{a_0}{b_0}, \frac{c_0}{d_0}]$ or $[\frac{c_0}{d_0}, \frac{a_0}{b_0}]$ according to the sign of the determinant of the matrix. The other matrices will successively refine this interval to give better and better approximations to the real number. The first matrix is called a *sign* matrix whereas the other matrices are *digit* matrices. The *information* contained in an lft $\phi : x \mapsto \dfrac{ax+c}{bx+d} : \mathbb{R}^* \to \mathbb{R}^*$ is defined by $\mathsf{info}(\phi) = \phi[0,\infty]$.

4 Geometric dynamics of lft's

Lft's can be classified according to their conjugacy classes and their geometric dynamics on $\mathbb{R}^*$. The following results for real lft's are obtained from the corresponding results for lft's with complex coefficients [8]. Consider the lft $\phi : x \mapsto \dfrac{ax+c}{bx+d} : \mathbb{R}^* \to \mathbb{R}^*$. We consider the fixed points of ϕ which are solutions of the quadratic $bx^2 + x(d-a) - c = 0$. If $b = 0$, then there is a fixed point at $x = \infty$ and another one at $x = \dfrac{c}{d-a}$ if $a \neq d$. If $b \neq 0$, then the roots of the quadratic are given by $\frac{a-d\pm\sqrt{(a+d)^2-4(ad-bc)}}{2b}$. We can distinguish the following three cases:

- If $(a+d)^2 = 4(ad-bc)$, then there is a unique fixed point at $\dfrac{a-d}{2b}$. Note that, in this case, we necessarily have $ad - bc > 0$.
- If $(a+d)^2 > 4(ad-bc)$, then there are two real roots. Here we can have $ad - bc > 0$ or $ad - bc < 0$.
- If $(a+d)^2 < 4(ad-bc)$, then there are two complex conjugate roots. In this case, we always have $ad - bc > 0$.

Next we examine the conjugacy classes in $\mathbb{M}$. Recall that two elements u and v in a group are said to be conjugate written $u \sim v$ if the there is an element w with $u = wvw^{-1}$. This is an equivalence relation which partitions the group into conjugacy classes. In $GL(2,\mathbb{R})$, the trace and the determinant of a matrix are invariant under conjugacy. We denote the trace and the determinant of a matrix $M = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ by $\mathrm{tr}M = a + d$ and $\det M = ad - bc$ respectively. Since an lft is represented by a matrix up to a scaling factor, it follows that $\Theta(M) \sim \Theta(M')$ implies $\frac{(\mathrm{tr}M)^2}{\det M} = \frac{(\mathrm{tr}M')^2}{\det M'}$.

Given an lft $\phi \in \mathbb{M}$ and a matrix $M \in GL(2, \mathbb{R})$ with $\Theta(M) = \phi$, we define the *conjugacy invariance* of ϕ by $\mathrm{inv}\phi = \frac{(\mathrm{tr}M)^2}{\det M}$. Clearly $\mathrm{inv}\phi$ is independent of the representative matrix M; in fact if $\phi : x \mapsto \dfrac{ax + c}{bx + d}$ then $\mathrm{inv}\phi = \dfrac{(a + d)^2}{ad - bc}$. Therefore:

Proposition 4.1 *For $\phi, \phi' \in \mathbb{M}$, we have* $\mathrm{inv}\phi = \mathrm{inv}\phi'$ *if $\phi \sim \phi'$.*

A basic property of conjugacy in $\mathbb{M}$ is that it preserves fixed points and limits. If the lft's ϕ and ψ are conjugate with $\phi = \gamma\psi\gamma^{-1}$, then $x \in \mathbb{R}^*$ is a fixed point of ψ iff $\gamma(x)$ is a fixed point of ϕ. Furthermore $\psi^n(x) \to y$ as $n \to \infty$ iff $\phi^n(\gamma x) \to \gamma y$ as $n \to \infty$.

The identity element in a group forms a conjugacy class and the conjugacy classes in $\mathbb{M}$ are fully described by selecting a representative from each other class. We define the family $\phi_\lambda : x \mapsto \lambda x : \mathbb{R}^* \to \mathbb{R}^*$ in $\mathbb{M}$, for $\lambda \in L = \{\exp i\theta | 0 < \theta < 2\pi\} \cup \mathbb{R} \setminus \{0\}$.

- $\phi_1 : x \mapsto x + 1$ for $\lambda = 1$.
- $\phi_\lambda : x \mapsto \lambda x : \mathbb{R}^* \to \mathbb{R}^*$ for $\lambda \in \mathbb{R} \setminus \{0, 1, -1\}$.
- $\phi_\lambda : x \mapsto \dfrac{x \cos\frac{\theta}{2} + \sin\frac{\theta}{2}}{-x \sin\frac{\theta}{2} + \cos\frac{\theta}{2}}$ for $\lambda = \exp i\theta$ $(0 < \theta < 2\pi)$.

The lft $\phi_{\exp i\theta}$ represents the rotation of S^1 by θ. In fact, assume $x \in \mathbb{R}^*$ is represented in homogeneous coordinates by $\begin{pmatrix} \sin\frac{\alpha}{2} \\ \cos\frac{\alpha}{2} \end{pmatrix}$ where $\widehat{s(0)Os(x)} = \alpha$ as in Figure 1 of Section 2. Then its image $y = \phi_{\exp i\theta}(x)$ is represented by the vector $\begin{pmatrix} \sin\frac{\alpha+\theta}{2} \\ \cos\frac{\alpha+\theta}{2} \end{pmatrix}$. Therefore, we have $\widehat{s(0)Os(y)} = \alpha + \theta$ and hence $\widehat{s(x)Os(y)} = \theta$.

We note that $\phi_\lambda \sim \phi_{\lambda^{-1}}$ for all $\lambda \in L$. This is trivial when $\lambda = 1$ and for $\lambda \neq 1$ we have $\phi_\lambda = \psi\phi_{\lambda^{-1}}\psi$ where $\psi : x \mapsto 1/x$ satisfies $\psi = \psi^{-1}$. Also we have $\mathrm{inv}\phi_\lambda = \lambda + \dfrac{1}{\lambda} + 2$ for all $\lambda \in L$.

Proposition 4.2 *Any non-identity element in $\mathbb{M}$ is conjugate to ϕ_λ for some $\lambda \in L$.*

Proof. Let $\phi \in \mathbb{M}$ be a non-identity element. If ϕ has a unique fixed point, α say, then the map $\psi : x \mapsto \dfrac{1}{x - \alpha}$ maps this fixed point to ∞. We have $\psi\phi\psi^{-1} : x \mapsto x + a$ for some $a \in \mathbb{R} \setminus \{0\}$ and this is conjugate to ϕ_1 via the map $x \mapsto x/a$.

If ϕ has two distinct real fixed points α_1, α_2, then the map $\psi : x \mapsto \dfrac{x-\alpha_1}{x-\alpha_2}$ maps α_1, α_2 to $0, \infty$ respectively. Hence, $\psi\phi\psi^{-1} : x \mapsto \lambda x$ for some real $\lambda \neq 1$.

Finally, suppose the complex extension of ϕ has two distinct complex conjugate fixed points $a+ib$ and $a-ib$ with $b \neq 0$. Then the complex extension of $\psi : x \mapsto \dfrac{b}{x-a} : \mathbb{R}^* \to \mathbb{R}^*$ maps $a+ib$ and $a-ib$ to $-i$ and i respectively. We have $\psi\phi\psi^{-1} = \phi_\lambda$ where $\lambda = \exp i\theta$ with $\cos\frac{\theta}{2} = a/(a^2+b^2)$ and $\sin\frac{\theta}{2} = b/(a^2+b^2)$.

□

Proposition 4.3 *ϕ_λ is conjugate to ϕ_δ iff $\lambda = \delta$ or $\lambda = \delta^{-1}$.*

Corollary 4.4 *Two non-identity elements ϕ and ψ of $\mathbb{M}$ are conjugate iff $\operatorname{inv}\phi = \operatorname{inv}\psi$.*

We now classify the lft's in $\mathbb{M}$ according to their geometric and dynamic behaviour. Any non-identity element $\phi \in \mathbb{M}$, with $\phi : x \mapsto \dfrac{ax+c}{bx+d}$ say, is, as we have seen, conjugate to ϕ_λ for some $\lambda \in L$. Note that ϕ does not determine λ completely, since ϕ is also conjugate to $\phi_{\lambda^{-1}}$. The pair $\{\lambda, \lambda^{-1}\}$ is called the *multiplier* of ϕ. Two non-identity elements in $\mathbb{M}$ will be conjugate iff they have the same multipliers. The multiplier for ϕ is obtained by solving the quadratic $\text{inv}\phi = \text{inv}\phi_\lambda = \lambda + \lambda^{-1} + 2$, i.e. λ and λ^{-1} are solutions of $\lambda^2 + (2 - \text{inv}\phi)\lambda + 1 = 0$ which has discriminant $\Delta = \text{inv}\phi(\text{inv}\phi - 4)$.

As seen already, ϕ will have a unique fixed point x_0 iff $\text{inv}\phi = 4$, i.e. $\lambda = 1$. In this case, $\phi \sim \phi_1$ and ϕ is called *parabolic*. Since $\lim_{n\to\infty} \phi_1^n x = \infty$ for all $x \in \mathbb{R}^*$ it follows that $\lim_{n\to\infty} \phi^n x = x_0$ for all $x \in \mathbb{R}^*$.

If ϕ has two distinct fixed points then we distinguish between the following cases:

- $\phi \sim \phi_\lambda$ for some real λ with $|\lambda| \neq 1$. This is equivalent to $\text{inv}\phi > 4$ or $\text{inv}\phi < 0$. In this case, ϕ has two real fixed points. The map ϕ_λ has two fixed points at 0 and ∞. If $|\lambda| > 1$, the orbit of any point in $\mathbb{R}^*$ other than 0 tends to ∞ under ϕ_λ. The opposite happens if $|\lambda| < 1$. This means that one fixed point is an attractor and the other a repeller. The same therefore is true for ϕ. The map ϕ is called *hyperbolic* if $\text{inv}\phi > 4$, equivalently $ad - bc > 0$, or $\lambda > 0$. On the other hand, ϕ is called *loxodromic* if $\text{inv}\phi < 0$, equivalently $ad - bc < 0$, or $\lambda < 0$.
- $\phi \sim \phi_{\exp i\theta}$ for some $0 < \theta < 2\pi$. This is equivalent to $0 \leq \text{inv}\phi < 4$. The map ϕ is called *elliptic*. In this case, ϕ^n is the identity lft for some positive integer n iff θ is a rational multiple of 2π. Since $\lim_{m\to\infty} \phi^m_{\exp i\theta}(x) = \lim_{m\to\infty} \phi_{\exp im\theta}(x)$ does not exist for any $x \in \mathbb{R}^*$ the same is true for $\lim_{m\to\infty} \phi^m(x)$.

It follows immediately from the definitions that the four types of parabolic, hyperbolic, loxodromic and elliptic lft's are each invariant under conjugacy.

We note that an affine map $x \mapsto \lambda x + c$ with $\lambda \neq 0$ is parabolic if $\lambda = 1$ and $c \neq 0$, hyperbolic if $0 < \lambda < 1$ or $1 < \lambda$, loxodromic if $-1 < \lambda < 0$ or $\lambda < -1$, and elliptic if $\lambda = -1$. Furthermore, any hyperbolic or loxodromic lft is conjugate to a contracting linear map. An elliptic lft is, by definition, conjugate to a rotation $\phi_{\exp i\theta}$ of S^1. As in the complex case [8], a non-identity lft $\phi \in \mathbb{M}$ will have a finite order (i.e., $\phi^n = \mathrm{Id}$ for some $n \geq 2$) iff ϕ is elliptic; therefore, any finite subgroup of $\mathbb{M}$ consists of the identity map and elliptic lft's. For real lft's, there is a simple characterisation of all finite groups. Recall that the dihedral group of order $2k$ is the group generated by two elements p and q with $p^k = q^2 = e$ and $p^m h = hp^{2k-m}$, where e is the identity element.

Proposition 4.5 *If G is a finite subgroup of $\mathbb{M}$ of order n, then there exists an lft ψ such that the conjugate subgroup $\psi G \psi^{-1} = \{\psi g \psi^{-1} | g \in G\}$ satisfies the following:*

- *either $\psi G \psi^{-1}$ is the cyclic group of rotations of S^1 generated by $\phi_{\exp \frac{2\pi i}{n}}$, or*
- *n is even, say $n = 2k$, and $\psi G \psi^{-1}$ is the dihedral group generated by the rotation $\phi_{\exp \frac{2\pi i}{k}}$ and the reflection $x \mapsto -x$.*

5 Exact Floating Point

So far our representation allows arbitrary normal products of integer matrices $M_0 M_1 M_2 \cdots$ with $M_0 \in \mathbb{M}$ and $M_i \in \mathcal{M}^+$ for $i \geq 1$. This, in practice, results in some major problems. Firstly, intervals will be refined at an arbitrary rate, making any analysis of complexity of algorithms practically impossible. Secondly, matrix multiplication can quickly produce huge integers in a matrix quite disproportionate to the information contained in it.

In analogy with floating point formats, where number representations in a given base are generated by two sign symbols and a finite number of digits, we restrict the sign and digit matrices to a finite set of specific matrices. We will see that sign matrices should be rotations of S^1 which are elliptic lft's whereas digit matrices should be contracting maps which are hyperbolic lft's.

5.1 Sign Matrices

We start with sign matrices. Recall the definition of the information contained in an lft ϕ, i.e. $\mathsf{info}(\phi) = \phi[0, \infty]$. The information in sign matrices must overlap and cover S^1. If we further assume that they have the same length with respect to ρ and are evenly placed on S^1, then they will be generated by rotations of S^1. The lft $\phi_{\exp i\theta} : x \mapsto \dfrac{x \cos \frac{\theta}{2} + \sin \frac{\theta}{2}}{-x \sin \frac{\theta}{2} + \cos \frac{\theta}{2}}$ rotates S^1 by θ. Furthermore, $\phi_{\exp i\theta}$ generates a finite cyclic group iff θ is a rational multiple of 2π. Furthermore, our choice will be restricted if the lft is required to have integer coefficients.

Proposition 5.1 *Suppose θ is a non-integral rational multiple of 2π. Then the lft $\phi_{\exp i\theta}$ will have integer coefficients iff $\theta = \frac{\pi}{2}$ or $\theta = \pi$.*

For $\theta = \pi$, we get the cyclic group of order 2 consisting of $\phi_{\exp i\pi} : x \mapsto -\frac{1}{x}$ and the identity lft $\mathrm{Id} : x \mapsto x$. This gives the two intervals $\mathsf{info}(\phi_{\exp i\pi}) = [\infty, 0]$ and $\mathsf{info}(\mathrm{Id}) = [0, \infty]$ which are not overlapping. For $\theta = \pi/2$ we get the cyclic group of order 4 with elements $\phi_{\exp\frac{i\pi}{2}} : x \mapsto \frac{x+1}{-x+1}$, $\phi_{\exp i\pi} : x \mapsto -\frac{1}{x}$, $\phi_{\exp\frac{3\pi i}{2}} : x \mapsto \frac{x-1}{x+1}$ and $\mathrm{Id} : x \mapsto x$, with information $[1,-1], [\infty, 0], [-1, 1]$ and $[0, \infty]$ respectively. The simplest matrices representing these lft's are:

$$S_\infty = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad S_- = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad S_0 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad S_+ = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We therefore take these as our sign matrices. The group G generated by S_∞ and the reflection $x \mapsto -x$ is a dihedral group of order 8 and we have:

Proposition 5.2 *The group G is the unique finite group of lft's with integer coefficients which contains the cyclic group of sign matrices as a proper subgroup.*

5.2 Digit Matrices

We now would like to select an appropriate set of digit matrices from $\mathbb{M}^+$. Since composition of digit matrices are required to represent shrinking sequences of intervals, we will look for matrices which contract distances in $[0, \infty]$ with respect to the metric ρ. Digit matrices must overlap and cover $[0, \infty]$.

Note that for $x, y \in [0, \infty]$, we have $\rho(x, y) = |\frac{x-1}{x+1} - \frac{y-1}{y+1}| = |S_0(x) - S_0(y)|$ and S_0 is a homeomorphism from $[0, \infty]$ to its image $S_0[0, \infty] = [-1, 1]$. Let $\phi \in \mathbb{M}^+$ and consider the restriction $\phi : [0, \infty] \to [0, \infty]$. Then $S_0 \phi S_0^{-1}$ is a homeomorphism from $[-1, 1]$ onto itself.

Proposition 5.3 *The map $\phi : [0, \infty] \to [0, \infty]$ is contracting with respect to the ρ-metric iff $S_0 \phi S_0^{-1} : [-1, 1] \to [-1, 1]$ is contracting with respect to the Euclidean metric.*

It follows that for any base $b > 1$, the signed digit representation on $[-1, 1]$ in base b induces via the homeomorphism S_0 a suitable set of digit matrices in $\mathbb{M}^+$.

The signed digit system in base $b > 1$ in $[-1, 1]$ is generated by the contracting maps

$$\begin{array}{rcl} f_k : [-1,1] & \mapsto & [-1,1] \\ x & \mapsto & \dfrac{x+k}{b} \end{array}$$

with $k \in \mathrm{Dig}(b) = \{-b+n, b-n | n \in \mathbb{N}, 1 \le n \le \lfloor b \rfloor\}$, where $\lfloor b \rfloor$ is the integral part of b. Here, b can be allowed to be a rational or an irrational number. The case $b = 3/2$ was considered by Brower and the case $b = \frac{-1+\sqrt{5}}{2}$, the golden ratio, has been studied by di Gianantonio [4]. We now define the digit matrices in base b:

$$D_k = S_0^{-1} f_k S_0 = \begin{pmatrix} 1+b-k & 1-b+k \\ 1-b-k & 1+b+k \end{pmatrix}.$$

For example, for base 2, we have the four sign matrices S_+, S_∞, S_{-1} and S_0 together with the three digit matrices D_{-1}, D_0 and D_1.

Note that any finite composition of the affine contractions f_k, where $k \in \mathrm{Dig}(b)$, is an orientation preserving affine contraction. Therefore, any finite composition of digit matrices is conjugate to an orientation preserving affine contraction and is therefore hyperbolic.

Exact floating point in base b is defined as the representation of real numbers by infinite composition of lft's, or, equivalently, infinite product of matrices, such that the first matrix is one of the sign matrices above and the subsequent matrices are digit matrices.

For each finite composition $D_{k_1}D_{k_2}\cdots D_{k_n}$ of digit matrices we have:

$$S_0 D_{k_1}D_{k_2}\cdots D_{k_n}[0,\infty] = f_{k_1}f_{k_2}\cdots f_{k_n}[-1,1].$$

Therefore for every infinite composition of digit matrices we obtain

$$\bigcap_{n\geq 0} S_0 D_{k_1}D_{k_2}\cdots D_{k_n}[0,\infty] = \bigcap_{n\geq 0} f_{k_1}f_{k_2}\cdots f_{k_n}[-1,1].$$

This gives us:

Proposition 5.4 *A real number with signed digit expansion $.k_1k_2k_3\cdots$ (with $k_j \in \mathrm{Dig}(b)$ for $j \geq 1$) is represented in exact floating point by the infinite product*

$$S_0 D_{k_1}D_{k_2}D_{k_3}\cdots.$$

We have already noted that rational numbers can be represented by integer vectors. Addition by a positive rational number $p/q \in \mathbb{Q}^+$ is represented by the parabolic lft $x \mapsto x + p/q$. Similarly, multiplication by $p/q \in \mathbb{Q}^+$ is represented by the lft $x \mapsto px/q$ which is hyperbolic for $p \neq q$. Recall that parabolic and hyperbolic matrices preserve orientation. Moreover, orientation preserving lft's in $\mathbb{M}^+$ have the following property under composition.

Proposition 5.5 *The composition of any two orientation preserving lft's in $\mathbb{M}^+$ is a hyperbolic or a parabolic or the identity lft.*

Proof. Suppose $\phi, \psi \in \mathbb{M}^+$ are orientation preserving and not inverses of each other with $\phi(x) = \dfrac{ax+c}{bx+d}$, where $a,b,c,d \geq 0$, and $\psi(x) = \dfrac{a'x+c'}{b'x+d'}$, where $a',b',c',d' \geq 0$. We can assume, by dividing the coefficients by $\sqrt{ad-bc}$ and $\sqrt{a'd'-b'c'}$ respectively, that $ad-bc = a'd'-b'c' = 1$. Note that for two positive real numbers r and s we have $rs \geq 1$ implies $r+s \geq 2$ with $r+s=2$ iff $r=s=1$. From $ad \geq 1$ and $a'd' \geq 1$ we get $aa'dd' \geq 1$. Hence, $\mathrm{inv}(\phi\psi) = (aa'+cb'+bc'+dd')^2 \geq (aa'+dd')^2 \geq 4$. We have $\mathrm{inv}(\phi\psi) = 4$ iff $cb'+bc' = 0$ and $aa' = dd' = 1$. Hence, $aa'dd' = 1$ which implies $ad = a'd' = 1$. Therefore, $\phi\psi$ is hyperbolic unless $a = 1/d = 1/a' = d'$ and $cb' = bc' = bc = b'c' = 0$, in which case it is parabolic. It is easy to check that for $\phi : x \mapsto a^2x + ca$ and $\psi : x \mapsto \dfrac{x}{a^2} + \dfrac{c}{a}$ with $a, c > 0$, the composition is the map $\phi\psi : x \mapsto x + 2ac$ which is parabolic. $\square$

Therefore any semi-group generated by hyperbolic and parabolic lft's, for example the semi-group generated by the digit matrices in a given base and the lft's for addition and multiplication by positive rational numbers, will consist of the identity map and parabolic and hyperbolic lft's.

Algorithms for computing elementary functions are developed in [14,15] using lft's with two arguments as proposed initially by Gosper [7].

References

[1] A. Avizienis. Signed-digit number representations for fast parallel arithmetic. *IRE Transactions on Electronic Computers*, 10:389–400, 1961.

[2] H.J. Boehm and R. Cartwright. Exact real arithmetic: Formulating real numbers as functions. In Turner. D., editor, *Research Topics in Functional Programming*, pages 43–64. Addison-Wesley, 1990.

[3] H.J. Boehm, R. Cartwright, M. Riggle, and O'Donnell M.J. Exact real arithmetic: A case study in higher order programming. In *ACM Symposium on Lisp and Functional Programming*, 1986.

[4] P. Di Gianantonio. *A functional approach to real number computation*. PhD thesis, University of Pisa, 1993.

[5] P. Di Gianantonio. Real number computability and domain theory. *Information and Computation*, 127(1):11–25, May 1996.

[6] M. H. Escardó. PCF extended with real numbers. *Theoretical Computer Science*, 162(1):79–115, August 1996.

[7] W. Gosper. *Continued Fraction Arithmetic*. HAKMEM Item 101B, MIT Artificial Intelligence Memo 239. MIT, 1972.

[8] G. Jones and D. Singerman. *Complex Functions*. Cambridge University Press, 1987.

[9] P. Kornerup and D. W. Matula. Finite precision lexicographic continued fraction number systems. In *Proc. 7th IEEE Symposium on Computer Arithmetic*, pages 207–214. IEEE Computer Society Press, 1985.

[10] V. Menissier-Morain. Arbitrary precision real arithmetic: design and algorithms. submitted to *J. Symbolic Computation*, 1996.

[11] R.E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, 1966.

[12] A. Nielsen and P. Kornerup. Msb-first digit serial arithmetic. *J. of Univ. Comp. Scien.*, 1(7):523–543, 1995.

[13] P. Potts, A. Edalat, and M. Escardó. Semantics of exact real arithmetic. In *Twelfth Annual IEEE Symposium on Logic in Computer Science*. IEEE, 1997.

[14] P. J. Potts and A. Edalat. Exact Real Arithmetic based on Linear Fractional Transformations, December 1996. Draft, Imperial College, available from `http://www-tfm.doc.ic.ac.uk/~pjp`.

[15] P. J. Potts and A. Edalat. Exact Real Computer Arithmetic, March 1997. Department of Computing Technical Report DOC 97/9, Imperial College, available from `http://theory.doc.ic.ac.uk/~ae`.

[16] J. E. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Transactions on Computers*, 39(8):1087–1105, 1990.

[17] E. Wiedmer. Computing with infinite objects. *Theoretical Computer Science*, 10:133–155, 1980.