Contractivity of Linear Fractional Transformations

Reinhold Heckmann *

Abstract

One possible approach to exact real arithmetic is to use linear fractional transformations (LFT's) to represent real numbers and computations on real numbers. Recursive expressions built from LFT's are only convergent (i.e., denote a well-defined real number) if the involved LFT's are sufficiently contractive. In this paper, we define a notion of contractivity for LFT's. It is used for convergence theorems and for the analysis and improvement of algorithms for elementary functions.

Keywords : Exact Real Arithmetic, Linear Fractional Transformations

1 Introduction

Linear Fractional Transformations (LFT's) provide an elegant approach to real number arithmetic [8, 17, 11, 14, 12, 6]. One-dimensional LFT's $x \mapsto \frac{ax+c}{bx+d}$ are used in the representation of real numbers and to implement basic unary functions, while two-dimensional LFT's $(x, y) \mapsto \frac{axy+cx+ey+g}{bxy+dx+fy+h}$ provide binary operations such as addition and multiplication, and can be combined to obtain infinite expression trees denoting transcendental functions.

LFT's can be modeled within linear algebra. If the four parameters of a one-dimensional LFT are written as a (2,2)-matrix (shortly called *matrix*), functional composition becomes matrix multiplication. Likewise, the eight parameters of a two-dimensional LFT can be written as a (2,4)-matrix (called *tensor*). Basic computational steps can be realized as variants of matrix multiplication.

In the LFT approach, real numbers are represented by infinite products of matrices. Two variants differing in the choice of these matrices have been considered by the group of Edalat and Potts at Imperial College [13, 6]. In the general approach, the first matrix is arbitrary, while the remaining matrices are *positive*, i.e., satisfying $a, b, c, d \geq 0$. The digit approach (also called *exact floating point*) is more restrictive: the first matrix in the product is one of four *sign matrices*, and the remaining matrices are members of a fixed finite set of positive *digit matrices*.

*FB 14 – Informatik, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany, e-mail: heckmann@cs.uni-sb.de

We present part of the existing framework of the LFT approach in Section 2. This provides the background for understanding the results in the remainder of this paper.

Digit matrices are made such that every infinite product of digit matrices converges, i.e., denotes a single well-defined real number (possibly ∞). Yet there are divergent general products, for instance the product of identity matrices. A general product only converges if its matrices denote LFT's that are sufficiently contractive. In Section 3, we derive a notion of contractivity for matrices, and prove a sufficient criterion for the convergence of infinite products of matrices.

In the LFT approach, non-rational elementary functions can be represented by infinite tensor expressions: $fx = t_0 x(t_1 x(t_2 x(\ldots)))$. In general, these expressions may or may not converge for a fixed real argument x. As in the case of matrices, convergence can be guaranteed if the tensors (two-dimensional LFT's) in the expression are sufficiently contractive. In Section 4, a notion of contractivity for tensors is derived from that for matrices and used to prove a sufficient criterion for convergence of tensor expressions. In Section 5, tensor contractivity is used in the analysis of some tensor expressions proposed by Edalat's group. In certain cases, it is possible to modify these tensor expressions in order to achieve better convergence.

2 Exact Real Arithmetic by Linear Fractional Transformations

In this section, we present the framework of exact real arithmetic by LFT's [8, 17, 11]. After a general introduction, we specialize to the version used by the group of Edalat and Potts at Imperial College [14, 12, 13, 16, 6].

2.1 From Digit Streams to Linear Fractional Transformations

There are many ways to represent real numbers as infinite objects [3, 2, 4, 5]. Here, we are only concerned with representations as infinite streams of "digits". These streams are evaluated incrementally; at any given time, only a finite prefix of the stream is known.

There are several different stream representations which can be grouped into two large families: variations of the familiar decimal representation [1, 3, 2, 5, 7, 11, 10], and continued fraction expansions [8, 17, 9].

For the first family, consider the usual decimal representation.[1] A number such as $0.142\cdots$ can be unraveled from left to right as follows:

$$0.142\cdots = \tfrac{1}{10}(1 + 0.42\cdots);\ 0.42\cdots = \tfrac{1}{10}(4 + 0.2\cdots);\ 0.2\cdots = \tfrac{1}{10}(2 + 0.\cdots)$$

[1] This representation is not suitable for practical purposes, as it lacks redundancy, and thus, most arithmetic functions are not computable. However, it provides a familiar example.

Thus, every digit d corresponds to an affine map α_d with $\alpha_d(x) = \frac{1}{10}(d+x) = \frac{x+d}{10}$. A number of the form $0.\cdots$ can be any element of the closed interval $[0,1]$, and so, a number of the form $0.142\cdots$ can be any element of the interval $(\alpha_1 \circ \alpha_4 \circ \alpha_2)[0,1] = [0.142, 0.143]$. In general, the infinite stream $0.d_1d_2d_3\cdots$ represents the unique real number in the intersection $\bigcap_{n=1}^{\infty}(\alpha_{d_1} \circ \cdots \circ \alpha_{d_n})[0,1]$.

In the classical continued fraction expansion [18], irrational numbers in the interval $[0,\infty]$ can be written as $a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}$ with natural numbers a_n. Every number a corresponds to the rational function ρ_a with $\rho_a(x) = a + \frac{1}{x} = \frac{ax+1}{x}$. Similar to the case above, an infinite continued fraction corresponds to the intersection $\bigcap_{n=1}^{\infty}(\rho_{a_1} \circ \cdots \circ \rho_{a_n})[0,\infty]$.

The formal similarity between the two approaches presented above leads to the following generalization [8, 17, 14, 12, 13, 16, 6]: real numbers in some *base interval* I are represented by infinite streams of digits. Digits are certain *Linear Fractional Transformations* (LFT's) $x \mapsto \frac{ax+c}{bx+d}$, parameterized by numbers a, b, c, d (in practical cases usually integers). The meaning of an infinite stream $\tau_1, \tau_2, \ldots$ of LFT's is the intersection $\bigcap_{n=1}^{\infty}(\tau_1 \circ \cdots \circ \tau_n)(I)$. This intersection is filtered (decreasing) if $\tau_n(I) \subseteq I$ holds for all digits τ_n. The stream is *convergent* if the intersection is a singleton set; the real number in this set is the number denoted by the stream. Not every stream is convergent; consider for instance the stream consisting of an infinite repetition of the identity LFT $x \mapsto x$.

2.2 LFT's and Matrices

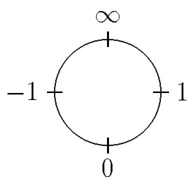
Every 2-2-matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ of real numbers denotes an LFT τ given by $\tau x = \frac{ax+c}{bx+d}$. In the sequel, we shall identify M and τ and write Mx for matrices M and reals x. All the matrices used in representations of real numbers and computations with real numbers will be integer matrices.

LFT's described by non-singular matrices, i.e., matrices M with determinant $\det M = ad - bc \neq 0$, are considered as functions from $\mathbf{R}^\star$ to itself, where $\mathbf{R}^\star = \mathbf{R} \cup \{\infty\}$ is the one-point compactification of the real line. The value ∞ arises as $r/0$ with $r \neq 0$, and on the other hand, $M(\infty)$ is defined to be a/b. For LFT's described by singular matrices, an additional 'number' $\perp$ (undefined) is needed which arises as $0/0$ and thus can be represented by the zero vector. Hence, the value of $M(\perp)$ is $\perp$. Therefore, non-singular LFT's denote functions from $\mathbf{R}^\star_\perp$ to itself, where $\mathbf{R}^\star_\perp = \mathbf{R}^\star \cup \{\perp\}$. With these definitions, composition of LFT's corresponds to matrix multiplication. Thus, the infinite streams of LFT's from the end of Section 2.1 can be considered as infinite products of matrices.

The mapping from matrices to LFT's is not one-to-one; for, M and rM denote the same LFT if $r \neq 0$. Here, rM means that all entries of M are multiplied by the number r, a process called *scaling*. Because of $\det(rM) = r^2 \det M$, the determinant of a matrix is not invariant under scaling, but its sign (1, 0, or -1) is, i.e., the sign of the determinant of M is a well-defined property of the LFT M. LFT's M with $\det M \neq 0$ (non-singular LFT's) are invertible. Thus, non-singular LFT's form a group under composition.

2.3 Intervals

The set $\mathbf{R}^\star$ can be visualized as a circle. Intervals $[u, v]$ are anti-clockwise arcs from u to v, e.g., $[0, 1] = \{x \in \mathbf{R} \mid 0 \leq x \leq 1\}$, and $[1, 0] = \{x \in \mathbf{R} \mid 1 \leq x \text{ or } x \leq 0\} \cup \{\infty\}$.



LFT's M with $\det M > 0$ preserve the orientation of their arguments around the circle, and thus, $M[u, v] = [Mu, Mv]$ holds for such LFT's. On the other hand, LFT's with $\det M < 0$ swap the orientation, whence $M[u, v] = [Mv, Mu]$ holds for such LFT's.

2.4 General Normal Products

In the representation of real numbers by LFT's (matrices), the base interval $[0, \infty]$ is used. The reason is that this base interval admits a simple check for the inclusion property: $\tau[0, \infty] \subseteq [0, \infty]$ holds if and only if τ can be represented by a matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ with $a, b, c, d \geq 0$ and $a{+}b,\ c{+}d > 0$; the latter two conditions are needed to avoid the value $\perp = 0/0$ for arguments in the interval $[0, \infty]$ (note that in particular, $M(0) = c/d$ and $M(\infty) = a/b$). Matrices satisfying these conditions are called *positive*.[2]

In the general approach to LFT representation, real numbers (elements of $\mathbf{R}^\star$) are represented as infinite products of integer matrices $\prod_{n=1}^{\infty} M_n$ where M_1 is arbitrary and all other matrices are positive, making the intersection $\bigcap_{n=1}^{\infty} M_1 M_2 \cdots M_n[0, \infty]$ filtered (decreasing). If M_1 is positive as well, then this intersection is a subset of $[0, \infty]$.

2.5 Sign and Digit Matrices

In the digit approach ('exact floating point'), the matrices in the representing infinite product are restricted to a finite set. The first matrix is one of four *sign matrices*, while the remaining ones come from a finite set of positive *digit matrices*.

There are four possible sign matrices, corresponding to rotations by 0°, 90°, 180°, and 270°. They can be explicitly described as follows:

$$
\begin{array}{llll}
S_+ & = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) & S_+\,[0, \infty] & = [0, \infty] \\
S_\infty & = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right) & S_\infty[0, \infty] & = [1, -1] \\
S_- & = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) & S_-\,[0, \infty] & = [\infty, 0] \\
S_0 & = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right) & S_0\,[0, \infty] & = [-1, 1]
\end{array}
$$

[2]In many papers, the additional conditions $a + b, c + d > 0$ do not occur since only non-singular matrices are considered.

S_0 and S_∞ are inverse to each other (up to scaling); $S_0 \cdot S_\infty = S_\infty \cdot S_0 = 2\mathrm{E}$ holds.

There are many possible sets of digit matrices, one for every base $r > 1$. Edalat and Potts [6] also discuss non-integer bases, but their implementation uses base $r = 2$. Here, we only consider integer bases $r > 1$.

Fix an integer $r > 1$. Every real number in the interval $[-1, 1]$ can be represented as $\sum_{n=1}^{\infty} k_n r^{-n}$ with integer digits k_n satisfying $|k_n| < r$. (Digits may be negative [1].) As in Section 2.1, these digits correspond to affine maps $A_k^r = \left(\begin{smallmatrix} 1 & k \\ 0 & r \end{smallmatrix}\right)$.

Since the base interval is not $[-1, 1]$, but $[0, \infty]$, the maps A_k^r have to be transformed into maps of that interval. This can be done by composition with the maps S_∞ and S_0, which are mutually inverse bijections between $[-1, 1]$ and $[0, \infty]$. Thus, the actual digit matrices are

$$D_k^r = S_\infty \cdot A_k^r \cdot S_0 = \begin{pmatrix} r+k+1 & r+k-1 \\ r-k-1 & r-k+1 \end{pmatrix} . \qquad (1)$$

2.6 Tensors

LFT's can be used not only to represent real numbers, but also to perform computations with real numbers. Using suitable LFT's $x \mapsto \frac{ax+c}{bx+d}$, basic functions such as $x \mapsto x+1$, $x \mapsto 2x$, and $x \mapsto \frac{1}{x}$ can be easily expressed. To compute sums, products, etc., *two-dimensional LFT's* are employed. They are characterized by 8 parameters, and thus can be represented by 2-4-matrices, called *tensors*. A tensor $T = \left(\begin{smallmatrix} a & c & e & g \\ b & d & f & h \end{smallmatrix}\right)$ denotes the function $T : {\mathbf{R}_\perp^\star}^2 \to \mathbf{R}_\perp^\star$ given by $Txy = \frac{axy+cx+ey+g}{bxy+dx+fy+h}$. For tensors, the notion of positivity can be defined in analogy with the case of matrices: a two-dimensional LFT maps $[0, \infty]^2$ to $[0, \infty]$ if and only if it can be represented by a positive tensor, i.e., a tensor with non-negative components and positive column sums.

It is straightforward how to represent addition, subtraction, multiplication, and division by suitable integer tensors [8, 17, 14, 12, 13]. Infinite expressions built from integer tensors may also be used to represent transcendental functions [15], e.g., $\ln x = T_0 x(T_1 x(T_2 x(\ldots)))$ where $T_0 = \left(\begin{smallmatrix} 1 & 1 & -1 & -1 \\ 0 & 1 & 1 & 0 \end{smallmatrix}\right)$ and $T_n = \left(\begin{smallmatrix} n & 2n+1 & n+1 & 0 \\ 0 & n+1 & 2n+1 & n \end{smallmatrix}\right)$ for $n > 0$.

3 Contractivity of Matrices

An infinite product of matrices $\prod_{n=1}^{\infty} M_n$ denotes the set-theoretic intersection $\bigcap_{n=1}^{\infty} M_1 M_2 \cdots M_n [0, \infty]$. If this intersection is a singleton set, we say the product *converges*; otherwise, it *diverges*. For instance, the product of identity matrices diverges since $\mathrm{E}^n[0, \infty] = \mathrm{E}[0, \infty] = [0, \infty]$ for all n in $\mathbf{N}$.

Now, we try to derive a criterion sufficient to guarantee convergence of an infinite product of positive matrices. Intuitively, the LFT's denoted by the matrices should be contracting functions to ensure the shrinking of the corresponding sequence of intervals to a single point.

3.1 Calculation of the Contractivity

To measure the contractivity of an LFT, a metric on $[0,\infty]$ is needed. In [6, 15], the sign matrix $S_0 : [0,\infty] \to [-1,1]$ with $S_0 = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ is used to derive a suitable metric for $[0,\infty]$ from the usual metric on $[-1,1]$. This metric is also useful for the purposes of the paper at hand. It is given by

$$d(x,y) = |S_0 x - S_0 y| = \left| \frac{x-1}{x+1} - \frac{y-1}{y+1} \right| = \frac{2|x-y|}{(x+1)(y+1)}$$

To calculate the contractivity of a non-singular positive matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$, we compute

$$\begin{aligned} d(Mx, My) &= \frac{2\left| \frac{ax+c}{bx+d} - \frac{ay+c}{by+d} \right|}{\left(\frac{ax+c}{bx+d} + 1\right)\left(\frac{ay+c}{by+d} + 1\right)} \\ &= \frac{2\,|(ax+c)(by+d) - (ay+c)(bx+d)|}{(ax+c+bx+d)(ay+c+by+d)} \\ &= \frac{2\,|ad-bc| \cdot |x-y|}{((a+b)x+c+d)((a+b)y+c+d)} \end{aligned}$$

Combining these two equations, we obtain

$$\frac{d(Mx,My)}{d(x,y)} = |\det M| \cdot \frac{(x+1)(y+1)}{((a+b)x+c+d)((a+b)y+c+d)} \tag{2}$$

Since we want to derive a sufficient criterion for convergence, we are interested in worst case behavior, i.e., we look for the maximum of the above expression over x and y in $[0,\infty]$.

Note that $\frac{x+1}{(a+b)x+c+d} = M'x$ where $M' = \left(\begin{smallmatrix} 1 & 1 \\ a+b & c+d \end{smallmatrix}\right)$. If $c+d \geq a+b$, matrix M' has positive determinant, thus M' is increasing, and therefore attains its maximum at $x = \infty$, yielding $\frac{1}{a+b}$. Otherwise, the maximum is attained at 0, yielding $\frac{1}{c+d}$. In any case, the maximum is $1/\min(a+b, c+d)$. Since the same analysis is possible for $\frac{y+1}{(a+b)y+c+d}$, the final result is

$$\sup_{x\in[0,\infty]} \sup_{y\in[0,\infty]} \frac{d(Mx,My)}{d(x,y)} = \frac{|\det M|}{(\min(a+b,c+d))^2} \tag{3}$$

This result also holds for singular positive matrices, since such matrices denote constant functions, and so the left hand side is 0.

The quantity on the right hand side of (3) is called the *contractivity* of the matrix M, abbreviated as con M:[3]

$$M = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \Longrightarrow \operatorname{con} M \;=\; \frac{|\det M|}{(\min(a+b, c+d))^2} \tag{4}$$

Note that the contractivity of a matrix is invariant under scaling (multiplication of all four entries by a number $r > 0$). This is as expected since it was derived as a property of the LFT denoted by the matrix.

The relationship between distance and contractivity is expressed by the following proposition:

Proposition 3.1 *For every positive matrix M and every x, y in $[0, \infty]$*

$$d(Mx, My) \leq \operatorname{con} M \cdot d(x, y)$$

Equality holds if M has equal column sums, i.e., $a + b = c + d$.

Proof: The estimation follows from (3) and (4). If $a + b = c + d$, then $x + 1$ and $y + 1$ can be canceled in (2) which leads to the claimed equality. □

3.2 Contractivity and Composition

From the derivation of the contractivity, it is obvious that the contractivity of a composition of two LFT's is at most as large as the product of the contractivities of the two LFT's. The following proposition adds the details.

Proposition 3.2 *Let M_1 and M_2 be positive matrices. Then* $\operatorname{con}(M_1 \cdot M_2) \leq \operatorname{con} M_1 \cdot \operatorname{con} M_2$ *holds.*

Proof: We provide an explicit proof to see when equality holds. Since $\det(M_1 \cdot M_2) = \det M_1 \cdot \det M_2$, it suffices to consider the denominators. Let $M_1 = \left(\begin{smallmatrix} a_1 & c_1 \\ b_1 & d_1 \end{smallmatrix}\right)$ and $M_2 = \left(\begin{smallmatrix} a_2 & c_2 \\ b_2 & d_2 \end{smallmatrix}\right)$. Then $M_1 \cdot M_2 = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ with $a = a_1 a_2 + c_1 b_2$, $b = b_1 a_2 + d_1 b_2$, etc. Thus

$$a + b = (a_1 + b_1)a_2 + (c_1 + d_1)b_2 \geq m_1(a_2 + b_2) \geq m_1 m_2 \tag{5}$$

where $m_i = \min(a_i + b_i, c_i + d_i)$. Here, the first '$\geq$' relation relies on the positivity of M_2 ($a_2, b_2 \geq 0$). Analogously, $c + d \geq m_1 m_2$ holds, and so $\min(a + b, c + d) \geq m_1 m_2$ as required. □

In the special case of $a_1 + b_1 = c_1 + d_1$, the first '$\geq$' relation in (5) can be replaced by equality. Using this,

$$\min(a + b, c + d) = \min(m_1(a_2 + b_2), m_1(c_2 + d_2)) = m_1 m_2$$

follows, giving the following result:

[3] In earlier papers and talks of the author, the *reciprocal* of this quantity was called contractivity. The present definition has some advantages over this, including the fact that con M cannot get infinite any longer, and it is closer to established usage.

Proposition 3.3 *Let M_1 and M_2 be positive matrices where M_1 has equal column sums ($a_1 + b_1 = c_1 + d_1$). Then* $\operatorname{con}(M_1 \cdot M_2) = \operatorname{con} M_1 \cdot \operatorname{con} M_2$ *holds.*

Matrices with equal column sums not only admit equality in Prop. 3.1 and Prop. 3.3, but also have a particularly simple formula for their contractivity.

First, the minimum in the denominator of $\operatorname{con}\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ may be removed if $a + b = c + d$, yielding $\frac{|\det|}{(a+b)^2}$. Next, from $a + b = c + d$, we obtain $d = a + b - c$, and so the determinant is

$$ad - bc = a^2 + ab - ac - bc = (a + b)(a - c) \ .$$

Inserting this into the formula for con yields

$$\operatorname{con}\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{|a - c|}{a + b} \quad \text{if } \ a + b = c + d \ . \tag{6}$$

3.3 A Convergence Criterion

Let $\delta(I)$ be the diameter of an interval $I \subseteq I_0$ in the metric d. Note that $\delta(I_0) = 2$ since $d(0, \infty) = |(-1) - 1| = 2$. With this notion, we obtain by iterating Prop. 3.1:

Proposition 3.4 *For a sequence $(M_i)_{i \in \mathbf{N}}$ of positive matrices, let $I_n = (M_1 \circ \cdots \circ M_n)(I_0)$ where $I_0 = [0, \infty]$. Then $\delta(I_n) \leq 2 \prod_{i=1}^{n} \operatorname{con} M_n$; equality holds if all the matrices M_i have equal column sums.*

Hence $\lim_{n \to \infty} \delta(I_n) = 0$ holds if $\prod_{i=1}^{\infty} \operatorname{con} M_n = 0$. Therefore, we obtain the following convergence criterion:

Theorem 3.5 *An infinite product $\prod_{n=1}^{\infty} M_n$ of positive matrices converges if $\prod_{n=1}^{\infty} \operatorname{con} M_n = 0$ holds.*

Usually, we shall not directly apply this criterion, but one of the following two corollaries:

Corollary 3.6 *Let $\prod_{n=1}^{\infty} M_n$ be an infinite product of positive matrices. If there is a constant $c < 1$ such that $\operatorname{con} M_n \leq c$ for all but a finite number of indices n, then the product converges to a real number.*

Corollary 3.7 *If $\prod_{n=1}^{\infty} M_n$ is an infinite product of positive matrices such that $\lim_{n \to \infty} \operatorname{con} M_n < 1$, then the product converges to a real number.*

The estimation $\delta(I_n) \leq 2 \prod_{i=1}^{n} \operatorname{con} M_n$ from Prop. 3.4 shows that the smaller the contractivities $\operatorname{con} M_n$ are, the smaller the intervals I_n will be, i.e., the quicker the convergence of the infinite product of matrices is.

n	$\prod_{i=1}^{n} M_i$	I_n	$\delta(I_n)$	$2\prod_{i=1}^{n} \operatorname{con} M_i$
1	$\left(\begin{smallmatrix}1 & 2\\ 1 & 1\end{smallmatrix}\right)$	[1.0000, 2.0000]	$1.333 \cdot 2^{-2}$	2^{-1}
2	$\left(\begin{smallmatrix}3 & 4\\ 2 & 3\end{smallmatrix}\right)$	[1.3333, 1.5000]	$1.829 \cdot 2^{-5}$	2^{-3}
3	$\left(\begin{smallmatrix}7 & 10\\ 5 & 7\end{smallmatrix}\right)$	[1.4000, 1.4286]	$1.255 \cdot 2^{-7}$	2^{-5}
4	$\left(\begin{smallmatrix}17 & 24\\ 12 & 17\end{smallmatrix}\right)$	[1.4117, 1.4167]	$1.722 \cdot 2^{-10}$	2^{-7}
5	$\left(\begin{smallmatrix}41 & 58\\ 29 & 41\end{smallmatrix}\right)$	[1.4137, 1.4147]	$1.182 \cdot 2^{-12}$	2^{-9}

Table 1: Approximation of $\sqrt{2}$

3.4 Some Examples

Recall from Section 2.5 that digit matrices are given by $D_k^r = \left(\begin{smallmatrix}r+k+1 & r+k-1\\ r-k-1 & r-k+1\end{smallmatrix}\right)$. These matrices have equal column sums, and so (6) applies giving $\operatorname{con} D_k^r = \frac{2}{2r} = \frac{1}{r}$. Hence, infinite products of digits converge as they should. Moreover, we know from the equality statement in Prop. 3.4 that for a sequence $(D_{k_n}^r)_{n\in\mathbf{N}}$ of digits, the size of the nth approximating interval is exactly $2r^{-n}$ (in the metric d).

In [12], $\sqrt{2}$ is described as $\prod_{n=1}^{\infty} \left(\begin{smallmatrix}1 & 2\\ 1 & 1\end{smallmatrix}\right)$. Since $\operatorname{con} \left(\begin{smallmatrix}1 & 2\\ 1 & 1\end{smallmatrix}\right) = \frac{|1-2|}{2^2} = \frac{1}{4}$, this product is convergent. Prop. 3.4 gives the estimation $\delta(I_n) \le 2/4^n$. In Table 1, this upper bound is compared with the actual size of the approximation intervals. In the intervals, the lower end is rounded down and the upper end is rounded up to obtain a proper inclusion. Recall that the size $\delta(I_n)$ of I_n does not refer to the usual distance, but to the distance d introduced in Section 3.1. Thus, $\delta([1,2])$ is not 1, but $\frac{1}{3}$. The power of 2 occurring in the value of $\delta(I_n)$ indicates how many digit matrices in base $r = 2$ have been calculated so far.

In [15], e is described as $\prod_{n=1}^{\infty} \left(\begin{smallmatrix}2n & 2n-1\\ 2n-1 & 2n-2\end{smallmatrix}\right)$. The determinant of these matrices is -1, whence their contractivity is $1/(4n-3)^2$, ensuring convergence. In Table 2, the actual size of the approximation intervals is compared with the upper bound from Prop. 3.4. The convergence is much quicker than in Table 1, at the expense of bigger numbers in the matrices.

From a formula for arctan given in [15], $\frac{\pi}{4} = \arctan 1 = \prod_{n=1}^{\infty} \left(\begin{smallmatrix}0 & 1\\ n^2 & 2n-1\end{smallmatrix}\right)$ follows. For $n \ge 2$, these matrices have contractivity $\frac{n^2}{(2n)^2} = \frac{1}{4}$, and thus, the product converges. The same paper contains another infinite product for π involving big integers of size about $10939058860032000\, n^4$, where for large n, the contractivity of each matrix is about $1/151931373056000$.

n	$\prod_{i=1}^{n} M_i$	I_n	$\delta(I_n)$	$2\prod_{i=1}^{n} \operatorname{con} M_i$
1	$\binom{2\ 1}{1\ 0}$	$[2, \infty]$	$1.333 \cdot 2^{-1}$	$1.000 \cdot 2^{+1}$
2	$\binom{11\ 8}{4\ 3}$	$[2.666666666, 2.750000000]$	$1.422 \cdot 2^{-8}$	$1.280 \cdot 2^{-4}$
3	$\binom{106\ 87}{39\ 32}$	$[2.717948717, 2.718750000]$	$1.899 \cdot 2^{-14}$	$1.011 \cdot 2^{-10}$
4	$\binom{1457\ 1264}{536\ 465}$	$[2.718279569, 2.718283583]$	$1.217 \cdot 2^{-21}$	$1.532 \cdot 2^{-18}$
5	$\binom{25946\ 23225}{9545\ 8544}$	$[2.718281822, 2.718281836]$	$1.905 \cdot 2^{-30}$	$1.357 \cdot 2^{-26}$

Table 2: Approximation of e

4 Contractivity of Tensors

A tensor $T = \begin{pmatrix} a & c & e & g \\ b & d & f & h \end{pmatrix}$ denotes an LFT T in two variables: $Txy = \frac{axy+cx+ey+g}{bxy+dx+fy+h}$. For fixed x, this becomes an LFT $T|_x$ in the variable y, where the matrix $T|_x$ is given by $T|_x = \begin{pmatrix} ax+e & cx+g \\ bx+f & dx+h \end{pmatrix}$. In the special case $x = \infty$, this matrix should be replaced by $T|_\infty = \binom{a\ c}{b\ d}$.

Similarly, for fixed y, an LFT $T|^y$ in the variable x results where $T|^y = \begin{pmatrix} ay+c & ey+g \\ by+d & fy+h \end{pmatrix}$. For $y = \infty$, this matrix should be replaced by $T|^\infty = \begin{pmatrix} a & e \\ b & f \end{pmatrix}$.

If T is positive, then $T|_x$ and $T|^y$ are positive for all x, y in $[0, \infty]$ – even for 0 and ∞ – and thus, their contractivities $\operatorname{con} T|_x$ and $\operatorname{con} T|^y$ are well-defined.

In [12, 15], many transcendental functions are defined by infinite tensor expressions of the kind $fx = T_0 x(T_1 x(T_2 x \ldots))$. For x given by a (convergent) infinite product of matrices $M_1 M_2 \cdots$, the meaning of fx is

$$\bigcap_{m=1}^{\infty} \bigcap_{n=1}^{\infty} T_0 I_m (T_1 I_m (\ldots (T_n I_m [0, \infty]) \ldots)) \tag{7}$$

where I_m is the interval $M_1 \cdots M_m [0, \infty]$. However, the cited papers do not contain an analysis for which values of x the expression fx converges, i.e., the intersection (7) is a singleton set.

The question of convergence of the tensor expression fx for a fixed value x can be reduced to the question of convergence of the corresponding product of matrices $T_0|_x \cdot T_1|_x \cdot \ldots$. The key for proving this is the following observation about tensors:

Proposition 4.1 *Let T be a positive tensor and $(I_n)_{n\in\mathbf{N}}$ and $(J_n)_{n\in\mathbf{N}}$ be shrinking sequences of subintervals of $[0, \infty]$. Then $\bigcap_{n=1}^{\infty} T I_n J_n = T(\bigcap_{n=1}^{\infty} I_n)(\bigcap_{n=1}^{\infty} J_n)$ holds.*

Proof: This is a consequence of the continuity of $T : [0, \infty] \times [0, \infty] \to [0, \infty]$ and the compactness of the intervals I_n and J_n. □

By repeated applications of this proposition, intersection (7) can be simplified as follows:

$$\begin{aligned}
&\bigcap_{m=1}^{\infty}\bigcap_{n=1}^{\infty} T_0 I_m(T_1 I_m(\ldots(T_n I_m[0,\infty])\ldots))\\
&= \bigcap_{n=1}^{\infty} T_0(\bigcap_{m=1}^{\infty} I_m)(T_1(\bigcap_{m=1}^{\infty} I_m)(\ldots(T_n(\bigcap_{m=1}^{\infty} I_m)[0,\infty])\ldots))\\
&= \bigcap_{n=1}^{\infty} T_0\{x\}(T_1\{x\}(\ldots(T_n\{x\}[0,\infty])\ldots))\\
&= \bigcap_{n=1}^{\infty} T_0|_x(T_1|_x(\ldots(T_n|_x[0,\infty])\ldots))
\end{aligned}$$

From this, we immediately obtain:

Theorem 4.2 *Let $(T_n)_{n\in\mathbf{N}}$ be a sequence of positive tensors and x be an element of $[0,\infty]$, given by a convergent product $\prod_{n=0}^{\infty} M_n$ of positive matrices. Then the tensor expression $T_0(\prod_{n=0}^{\infty} M_n)(T_1(\prod_{n=0}^{\infty} M_n)(\ldots))$ converges if and only if the product $\prod_{n=0}^{\infty}(T_n|_x)$ of positive matrices converges.*

To prove convergence of $\prod_{n=0}^{\infty}(T_n|_x)$, Theorem 3.5 and its two Corollaries can be used.

5 Case Studies: Analysis of Tensor Expressions

Here, we analyze some of the tensor expressions in [15]. Most of them have good contractivities. In the case of square root, the contractivity can be considerably increased by modifying the original tensor expression.

5.1 Square Root

In [15, Section 11.1], $\sqrt{x}$ is given by the tensor expression $Tx(Tx(\ldots))$ with $T = \left(\begin{smallmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{smallmatrix}\right)$. Using Theorem 4.2 and Corollary 3.6, we shall show that this tensor expression converges for $x \in (0,\infty)$. However, there are slightly different tensors that show better convergence, i.e., smaller contractivity values than T.

Why does the above tensor expression represent $\sqrt{x}$? Assuming that it converges to some number y, this number must be a solution of the equation $y = Txy$, yielding $y^2 = x$ or $y = \sqrt{x}$.

A little computation shows that other tensors can do this job as well. Let $T = \left(\begin{smallmatrix} a & c & e & g \\ b & d & f & h \end{smallmatrix}\right)$ be an arbitrary tensor. Then $y = Txy$ is equivalent to

$$y(bxy + dx + fy + h) = axy + cx + ey + g$$

or after some reordering

$$bxy^2 + fy^2 + (d-a)xy + (h-e)y = cx + g \quad .$$

If $b = 0$, $d = a$, $h = e$, $g = 0$, and $f = c \neq 0$, this equation reduces to $y^2 = x$ as desired. Therefore, the most general form of a suitable tensor is $T = \left(\begin{smallmatrix} a & c & e & 0 \\ 0 & a & c & e \end{smallmatrix}\right)$. Positivity is ensured by $a, c, e > 0$. By Theorem 4.2, we need to calculate $\operatorname{con} T|_x$ to obtain information about the convergence behavior.

$$\operatorname{con} T|_x = \operatorname{con} \begin{pmatrix} ax+e & cx \\ c & ax+e \end{pmatrix} = \frac{|(ax+e)^2 - c^2x|}{(ax+e+\min(c, cx))^2} \tag{8}$$

For $x \leq 1$, the minimum in this formula evaluates to cx, giving

$$\frac{|(ax+e)^2 - c^2x|}{((a+c)x+e)^2} \quad . \tag{9}$$

For $x \geq 1$, the minimum is c, which gives

$$\frac{|(ax+e)^2 - c^2x|}{(ax+e+c)^2} = \frac{|(a+eu)^2 - c^2u|}{(a+(e+c)u)^2} \tag{10}$$

where $u = 1/x$.

For $x = 0$, (9) simplifies to $e^2/e^2 = 1$, and for $x = \infty$ or $u = 0$, (10) becomes $a^2/a^2 = 1$. Thus, our criteria give no direct information about convergence for these special two values, but convergence can be shown by a closer analysis.

For $x = 0$, we get $M := T|_0 = \left(\begin{smallmatrix} e & 0 \\ c & e \end{smallmatrix}\right)$ with $e, c > 0$. Note that $M(0) = 0$ and $M(\frac{e}{kc}) = \frac{e^2/kc}{c(e/kc)+e} = \frac{e^2}{ce+kce} = \frac{e}{(k+1)c}$. Thus, starting from $[0,\infty] = [0, \frac{e}{0c}]$, we obtain $M^k[0,\infty] = [0, \frac{e}{kc}]$, from which we see that the product $\prod_{n=1}^{\infty} M$ converges to the real number 0. For $x = \infty$, convergence to ∞ can be shown analogously.

In contrast to the cases $x = 0$ or $x = \infty$, we can achieve $\operatorname{con} T|_x < 1$ for $0 < x < \infty$ by suitable choices of the parameters a, c, and e. In the following, we look for parameter values which minimize $\operatorname{con} T|_x$.

Comparing (9) and (10), one sees that for $a = e$, $\operatorname{con} T|_x = \operatorname{con} T|_{1/x}$ holds. In the sequel, we restrict ourselves to this symmetric case and $x \leq 1$. In this case, (9) becomes

$$\frac{|a^2x^2 + (2a^2 - c^2)x + a^2|}{((a+c)x+a)^2} \quad . \tag{11}$$

The tensor proposed in [15] corresponds to $a = c = 1$ which gives

$$\operatorname{con} T|_x = \frac{x^2+x+1}{(2x+1)^2} = \frac{x^2+x+1}{4x^2+4x+1} \quad .$$

This fraction is < 1 for all x in $[0, 1]$ except $x = 0$.

Note that for fixed a, the denominator of (11) increases with increasing c, while the numerator decreases, provided that the quadratic expression in it is not negative. Thus, c should be chosen as large as possible under the side condition that $a^2x^2 + (2a^2 - c^2)x + a^2$ is still positive for x in $[0, 1]$. The largest such value is $c = 2a$, which yields a numerator of $a^2(x^2 - 2x + 1) = a^2(x-1)^2$, whence $\operatorname{con} T|_x = \frac{(x-1)^2}{(3x+1)^2}$. Thus, we propose to replace $\left(\begin{smallmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{smallmatrix}\right)$ by $\left(\begin{smallmatrix} 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \end{smallmatrix}\right)$ in the computation of square roots. The resulting gain in contractivity for certain values of x in $[0, 1]$ is shown in Table 3.

x	0	1/6	1/3	1/2	2/3	5/6	1
$\frac{x^2+x+1}{(2x+1)^2}$	1	$\frac{43}{64} \approx 0.67$	$\frac{13}{25} = 0.52$	$\frac{7}{16} \approx 0.44$	≈ 0.39	≈ 0.36	$\frac{1}{3} \approx 0.33$
$\frac{(x-1)^2}{(3x+1)^2}$	1	$\frac{25}{81} \approx 0.31$	$\frac{1}{9} \approx 0.11$	$\frac{1}{25} = 0.04$	1/81	1/441	0

Table 3: Contractivity for certain values of x

5.2 Logarithm

According to [15], a tensor expression for the natural logarithm $\ln x$ is given by $T_0 x(T_1 x(\ldots))$ with $T_0 = \left(\begin{smallmatrix} 1 & 1 & -1 & -1 \\ 0 & 1 & 1 & 0 \end{smallmatrix}\right)$ and $T_n = \left(\begin{smallmatrix} n & 2n+1 & n+1 & 0 \\ 0 & n+1 & 2n+1 & n \end{smallmatrix}\right)$ for $n > 0$. Tensor T_0 is not positive reflecting the fact that $\ln x$ is not in $[0, \infty]$ for x in $[0, 1)$. To apply Theorem 4.2, we only consider the positive tensors T_n with $n > 0$.

$$\begin{aligned} \operatorname{con} T_n|_x &= \operatorname{con} \begin{pmatrix} nx+n+1 & (2n+1)x \\ 2n+1 & (n+1)x+n \end{pmatrix} \\ &= \frac{|(nx+n+1)((n+1)x+n) - (2n+1)^2 x|}{(\min(nx+3n+2, (3n+2)x+n))^2} \end{aligned}$$

This expression is symmetric in x, i.e., $\operatorname{con} T_n|_x = \operatorname{con} T_n|_{1/x}$ holds for all n in $\mathbf{N}$ and x in $[0, \infty]$. Therefore, we may restrict ourselves to the case $x \leq 1$. In this case, the minimum in the denominator reduces to its right argument. The numerator can be simplified as follows:

$$\begin{aligned} & |(nx+n+1)((n+1)x+n) - (2n+1)^2 x| \\ = \;& |n(n+1)x^2 + (n^2 + (n+1)^2 - (2n+1)^2)x + n(n+1)| \\ = \;& |n(n+1)x^2 - (2n^2+2n)x + n(n+1)| \\ = \;& n(n+1)(x-1)^2 \end{aligned}$$

Thus, we obtain for $x \leq 1$ a value of $\operatorname{con} T_n|_x = \frac{n(n+1)(x-1)^2}{((3n+2)x+n)^2}$. For $x = 0$, this simplifies to $\frac{n+1}{n} > 1$; therefore, convergence is not guaranteed in this case. (We conjecture that convergence holds, but cannot prove it.) For all x in $[0, 1]$, $\lim_{n\to\infty} \operatorname{con} T_n|_x = \frac{(x-1)^2}{(3x+1)^2}$ holds, the same expression as for the modified tensor for square root. The last row in Table 3 shows the values of this fraction for some values of x. For $x > 0$, the fraction is < 1, ensuring convergence of the tensor expression by Theorem 4.2 and Cor. 3.7. The best contractivity is achieved at $x = 1$, where $\operatorname{con} T_n|_1 = 0$ for all $n > 0$.

5.3 Exponential

In [15], a tensor expression for e^u with u in $[-1, 1]$ is given. Since $[-1, 1] = S_0[0, \infty]$ where $S_0 = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ is one of the four sign matrices, this tensor expression is in fact for $e^{S_0 x}$ with x in $[0, \infty]$: $e^{S_0 x} = T_0 x(T_1 x(\ldots))$ with $T_n = \begin{pmatrix} 2n+2 & 2n+1 & 2n & 2n+1 \\ 2n+1 & 2n & 2n+1 & 2n+2 \end{pmatrix}$. For $\operatorname{con} T_n|_x$, we obtain:

$$\begin{aligned} \operatorname{con} T_n|_x &= \operatorname{con} \begin{pmatrix} (2n+2)x + 2n & (2n+1)x + 2n + 1 \\ (2n+1)x + 2n + 1 & 2nx + 2n + 2 \end{pmatrix} \\ &= \frac{|((2n+2)x + 2n)(2nx + 2n + 2) - ((2n+1)x + 2n + 1)^2|}{(\min((4n+3)x + 4n + 1, (4n+1)x + 4n + 3))^2} \end{aligned}$$

Again, this is symmetric in x, the values for x and $1/x$ are identical. Thus, we restrict ourselves to the case $x \leq 1$ where the minimum in the denominator reduces to its left argument. The numerator can be simplified to $(x-1)^2$, so that $\operatorname{con} T_n|_x = \frac{(x-1)^2}{((4n+3)x+4n+1)^2}$. For $x = 0$, this is $1/(4n+1)^2$, for $x = 1/2$, it is $1/(12n+5)^2$, and for $x = 1$, it is 0 independent of n. For all x in $[0, 1]$, the expression is inverse quadratic in n and goes to 0 as n goes to ∞. Thus, we have very good contractivities in this example.

6 Conclusion

The introduction of a *contractivity* for LFT's leads to a sufficient criterion for the convergence of infinite matrix and tensor expressions. In addition to qualitative statements such as the expression converges or the expression may not converge, we obtain quantitative information from Prop. 3.4: the smaller the contractivity, the quicker the convergence. Surely, this will affect the complexity of evaluating an expression to a specified accuracy, yet the details of the relationship between contractivity and complexity are still to be found.

Acknowledgments

I was introduced to Exact Real Arithmetic and in particular to the LFT approach during a visiting fellowship at Imperial College, London, organized by Abbas Edalat and funded by EPSRC. I am most grateful to Abbas Edalat, Martín Escardó, Peter Potts, and Philipp Sünderhauf for their support during this visit. I also like to thank Fritz Müller in Saarbrücken who commented on a draft version of this paper.

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, R. Cartwright, M. Riggle, and M.J. O'Donell. Exact real arithmetic: A case study in higher order programming. In *ACM Symposium on Lisp and Functional Programming*, pages 162–173, 1986.

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

[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] A. Edalat and P. Potts. A new representation for exact real numbers. In S. Brookes and M. Mislove, editors, *MFPS '97*, volume 6 of *Electronic Notes in Theoretical Computer Science*, 1997. URL: `http://www.elsevier.nl/locate/entcs/volume6.html`.

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

[8] W. Gosper. Continued fraction arithmetic. Technical Report HAKMEM Item 101B, MIT Artificial Intelligence Memo 239, MIT, 1972.

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

[12] P. J. Potts and A. Edalat. Exact real arithmetic based on linear fractional transformations. Draft, Imperial College, available from `http://www-tfm.doc.ic.ac.uk/~pjp`, December 1996.

[13] P. J. Potts and A. Edalat. Exact real computer arithmetic. Draft, Imperial College, available from `http://www-tfm.doc.ic.ac.uk/~pjp`, March 1997.

[14] P. J. Potts. Computable real arithmetic using linear fractional transformations. Draft PhD Thesis, Imperial College, available from `http://www-tfm.doc.ic.ac.uk/~pjp`, June 1996.

[15] P. J. Potts. Efficient on-line computation of real functions using exact floating point. Available from `http://www-tfm.doc.ic.ac.uk/~pjp`, October 1997.

[16] P. Potts, A. Edalat, and M. Escardó. Semantics of exact real arithmetic. In *Proc. Twelfth Annual IEEE Symposium on Logic in Computer Science*, pages 248–257. IEEE, 1997.

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

[18] H.S. Wall. *Analytic Theory of Continued Fractions.* Van Nostrand Company, 1948.