Translation of Taylor series into LFT expansions

Reinhold Heckmann

1 Introduction

In Exact Real Arithmetic, real numbers are represented as potentially infinite streams of information units, called *digits*. In this paper, we work in the framework of *Linear Fractional Transformations* (LFT's, also known as Möbius transformations) that provide an elegant approach to real number arithmetic (Gosper 1972, Vuillemin 1990, Nielsen and Kornerup 1995, Potts and Edalat 1996, Edalat and Potts 1997, Potts 1998b). One-dimensional LFT's are used as digits and to implement basic unary functions, while two-dimensional LFT's provide binary operations such as addition and multiplication, and can be combined to obtain infinite expression trees denoting transcendental functions. Peter Potts (1998a, 1998b) derived these expression trees from *continued fraction expansions* of the transcendental functions. In contrast, we show how to derive LFT expression trees from *power series expansions*, which are available for a greater range of functions.

In Section 2, we present the LFT approach in some detail. Section 3 contains the main results of the paper. We first derive an LFT expansion from a power series using Horner's scheme (Section 3.1). The results are not very satisfactory. Thus, we show how LFT expansions may be modified using algebraic transformations (Section 3.2). A particular such transformation, presented in Section 3.3, yields satisfactory results for standard functions, as shown in the final examples section 4.

2 Exact real arithmetic by linear fractional transformations

In this section, we recall the framework of exact real arithmetic via LFT's (Gosper 1972, Vuillemin 1990, Nielsen and Kornerup 1995). We do not follow exactly the version used by the group of Edalat and Potts at Imperial College (Potts and Edalat 1996, Edalat and Potts 1997, Potts 1998a, Potts 1998b), but change the *base interval* from $[0,\infty]$ to $[-1,1]$. The reasons for this change and its pros and cons were discussed in (Heckmann 2000).

2.1 LFT's and matrices

Linear Fractional Transformations (LFT's) are functions $x \mapsto \frac{ax+c}{bx+d}$ from reals to reals, parameterised by numbers a, b, c, and d. In general, these numbers may be reals, but in this paper, we shall only consider LFT's with integer parameters, as it is usually done in practical implementations of exact real arithmetic.

More precisely, LFT's are functions from ${\mathbb{R}^\star}_\perp$ to ${\mathbb{R}^\star}_\perp$ where $\mathbb{R}^\star = \mathbb{R} \cup \{\infty\}$ is the one-point compactification of the real line, and ${\mathbb{R}^\star}_\perp = \mathbb{R}^\star \cup \{\perp\}$ is $\mathbb{R}^\star$ plus an additional 'undefined' value $\perp$. The value ∞ arises as $r/0$ with $r \neq 0$, and on the other hand, the value of the LFT at ∞ is defined as a/b. The undefined value $\perp$ arises as $0/0$, and the value of any LFT at $\perp$ is $\perp$.

The four parameters of an LFT may be presented as a 2-2-matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ of integers, hereafter *matrix*. Thus, any matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ denotes an LFT $\langle M \rangle$ with $\langle M \rangle(x) = \frac{ax+c}{bx+d}$. Clearly, common factors of the four entries of the matrix do not matter; the matrices M and kM denote the same LFT if k is a non-zero integer. Thus, we get an equivalence $M \cong kM$ for $k \neq 0$. In particular, $M \cong -M$ holds. As a slight normalisation, we usually present matrices in a way such that the lower right entry is non-negative ($d \geq 0$). In the sequel, we shall identify matrices and LFT's, writing $M(x)$ instead of $\langle M \rangle(x)$.

Multiplication of matrices preserves equivalence and corresponds to composition of LFT's. LFT's with non-zero determinant (non-singular LFT's) map $\mathbb{R}^\star$ to $\mathbb{R}^\star$ and are invertible. An integer representation of the inverse is given by $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)^* = \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$.

An LFT $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ is called *affine* if $b = 0$ and $d \neq 0$. By normalisation, any affine LFT can be written as $\left(\begin{smallmatrix} a & c \\ 0 & 1 \end{smallmatrix}\right)$ which we'll sometimes abbreviate as $\langle c; a \rangle$; the corresponding function is $\lambda x.\, c + ax$. Composition (multiplication) and inversion of affine LFT's are obtained as follows:

$$\langle c;\, a \rangle \cdot \langle c';\, a' \rangle \;=\; \langle c + ac';\, aa' \rangle \tag{1}$$

$$\langle c;\, a \rangle^* \;=\; \langle -c/a;\, 1/a \rangle \quad \text{if } a \neq 0 \tag{2}$$

Note that $\langle c; a \rangle(B) = [c \pm a]$ where $B = [-1, 1] = [0 \pm 1]$ is the base interval. Here and in the sequel, we use a representation of intervals by center c and radius r, written as $[c \pm r]$, defined as

$$[c \pm r] \;=\; \{x \in \mathbb{R} \mid |x - c| \leq |r|\} \quad . \tag{3}$$

Note that we allow negative radii with the convention $[c \pm r] = [c \pm (-r)] = [c \pm |r|]$. In the formulae above, this was realised by using $|r|$ on the right hand side. Negative radii will prove convenient when dealing with power series, where the individual summands may be negative as well as positive.

An LFT M is *refining* if $M(B) \subseteq B$ where B is the base interval. (This notion was studied in (Edalat and Potts 1997) for arbitrary intervals.) Thus,

$$\begin{aligned} \langle c; a \rangle \text{ is refining} \quad &\iff \quad [c \pm a] \subseteq B \iff c - |a| \geq -1 \;\&\; c + |a| \leq 1 \\ &\iff \quad -c \leq 1 - |a| \;\&\; c \leq 1 - |a| \iff |c| + |a| \leq 1 \end{aligned} \tag{4}$$

Consequently, $\left(\begin{smallmatrix} a & c \\ 0 & d \end{smallmatrix}\right)$ is refining iff $|a| + |c| \leq |d|$.

The *contractivity* of a refining matrix M is $\sup_{x \neq y \in B} \frac{|M(x) - M(y)|}{|x - y|}$. The contractivity of an affine matrix $\left(\begin{smallmatrix} a & c \\ 0 & d \end{smallmatrix}\right)$ is $|a/d|$, and hence the contractivity of $\langle c; a \rangle$ is $|a|$.

2.2 Representing reals by LFT's

Using the base interval $B = [-1,1]$, a sequence $A_0, A_1, \ldots$ of LFT's induces a sequence $I_1, I_2, \ldots$ of intervals by

$$I_n \;=\; A_0 \cdots A_{n-1}(B). \tag{5}$$

The sequence of intervals is *nested* if $I_n \supseteq I_{n+1}$ for all $n \geq 1$. The inclusion $I_n \supseteq I_{n+1}$ means $A_0 \cdots A_{n-1}(B) \supseteq A_0 \cdots A_n(B)$. If the matrices $A_0, \ldots, A_{n-1}$ are non-singular, this is equivalent to $A_n(B) \subseteq B$. Therefore in the non-singular case, the sequence of intervals is nested iff all LFT's A_n with $n \geq 1$ are refining. Note that A_0 need not be refining.

A nested sequence of intervals $I_n = [c_n \pm r_n]$ converges to a single point x iff $r_n \to 0$ as $n \to \infty$. Even if the sequence of intervals $[c_n \pm r_n]$ is not nested, it may converge to a single point x in the sense that $c_n \to x$ and $r_n \to 0$. But this is not sufficient to say that the intervals approximate x (consider for instance the case $I_n = [2/n \pm 1/n] = [1/n, 3/n]$). Therefore, we require the interval sequence to be nested. In theory, it is sufficient that the interval sequence is *eventually nested*, i.e., $I_n \supseteq I_{n+1}$ for almost all n, because we may omit the first few intervals and obtain a nested sequence. Yet this procedure is not always practical as we shall see later.

In the sequel, we shall write $I_n \to y$ if the two end points converge to y, $I_n \downarrow y$ if $I_n \to y$ and the intervals are eventually nested, and $I_n \Downarrow y$ if $I_n \to y$ and the intervals are nested from the beginning.

A sequence of matrices that generates a nested sequence of intervals converging to x may be considered as a representation of x. Because of the usage of matrix multiplication in (5), we write $x = \prod_{n=0}^{\infty} A_n$, using a (formal) infinite product notation. Many real numbers can be elegantly represented by such infinite products.

To control the information flow in computations with reals, it turned out to be useful to convert these representations into a kind of standard form, using a leading *exponent matrix* (Heckmann 2000) or *sign matrix* followed by a sequence of *digit matrices* (Potts and Edalat 1997, Edalat and Potts 1997). For the purposes of the present paper, it is not necessary to delve into the details of these special representations. Instead, we may assume that real numbers are represented by arbitrary infinite products where all matrices but the first one are refining (to obtain a nested sequence of intervals) with contractivities bounded by some $c < 1$ (to ensure the interval sequence shrinks to a single point).

2.3 Tensors

To represent sums, products, and quotients, *two-dimensional LFT's* are employed. They are characterised by 8 integer parameters, and thus can be represented by 2-4-matrices of integers, called *tensors*. A tensor $T = \begin{pmatrix} a & c & e & g \\ b & d & f & h \end{pmatrix}$ denotes the bistrict function $T : \mathbb{R}^\star{}_\perp \times \mathbb{R}^\star{}_\perp \to \mathbb{R}^\star{}_\perp$ given by $T(x,y) = \frac{axy+cx+ey+g}{bxy+dx+fy+h}$. Again, common factors do not matter. The tensor $\begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ realises addition, $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ multiplication, and $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ division. Given arguments x and y represented by infinite products $\prod_{n=0}^{\infty} A_n$ and $\prod_{n=0}^{\infty} B_n$, there are methods to absorb the matrices A_n and B_n gradually into a tensor T and to emit matrices C_n from the tensor, thereby producing a representation $\prod_{n=0}^{\infty} C_n$ for the result $T(x,y)$.

For fixed x, the function $T|_x := (y \mapsto T(x,y))$ is an LFT with $T|_x = \begin{pmatrix} ax+e & cx+g \\ bx+f & dx+h \end{pmatrix}$. Conversely, any matrix M_x whose entries linearly depend on a parameter x can be turned into a tensor T such that $M_x(y) = T(x,y)$. The representations of the values of transcendental functions will generally involve such parametrised matrices. For instance, we shall later derive the representation

$$e^x \;=\; \begin{pmatrix} x & x+1 \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} x & x \\ 0 & n+1 \end{pmatrix}$$

from the Taylor series $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Such a product of parametrised matrices is open to two interpretations: For rational arguments x, it is (equivalent to) an infinite product of integer matrices, e.g., $e = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} 1 & 1 \\ 0 & n+1 \end{pmatrix}$. For general (real) arguments however, it should be turned into the infinite tensor expression $e^x = T_0(x, T_1(x, T_2(x, \ldots)))$ with $T_0 = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $T_n = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & n+1 \end{pmatrix}$.

3 From power series to LFT expansions

We first recall the basic properties of power series. Let $a_0, a_1, \ldots$ be a sequence of real numbers which is used to form the power series $\sum_{n=0}^{\infty} a_n x^n$ for $x \in \mathbb{R}$. The number

$$R \;=\; \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}} \tag{6}$$

(with $\frac{1}{0} = \infty$ and $\frac{1}{\infty} = 0$) is called the *convergence radius* of the power series. The reason for this name are the following properties:

- If $|x| < R$, then $\sum_{n=0}^{\infty} a_n x^n$ converges (to some real number);
- if $|x| > R$, then $\sum_{n=0}^{\infty} a_n x^n$ does not converge;
- for $|x| = R$, anything can happen.

If almost all coefficients a_n are non-zero, then R can be bounded by

$$\liminf_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \;\le\; R \;\le\; \limsup_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \quad . \tag{7}$$

In particular, if $\left| \frac{a_n}{a_{n+1}} \right|$ converges, its limit is R. For instance, in $e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$, the quotient $\frac{a_n}{a_{n+1}} = \frac{(n+1)!}{n!} = n+1$ goes to infinity, whence the convergence radius of this series is ∞.

3.1 The Horner evaluation of power series

Horner's rule to evaluate polynomials can easily be adapted to obtain an LFT expansion from a power series $f(x) = \sum_{n=0}^{\infty} a_n x^n$ with convergence radius R:

$$\sum_{n=0}^{\infty} a_n x^n \;=\; a_0 + x \cdot \sum_{n=1}^{\infty} a_n x^{n-1} \;=\; \langle a_0; x\rangle \left(\sum_{n=0}^{\infty} a_{n+1} x^n \right) \tag{8}$$

with the affine LFT $\langle a_0;x\rangle = \lambda y.\, a_0 + xy$. This suggests that $f(x)$ is given by

$$f(x) \;=\; \prod_{n=0}^{\infty} \langle a_n; x\rangle \;=\; \prod_{n=0}^{\infty} \begin{pmatrix} x & a_n \\ 0 & 1 \end{pmatrix} \tag{9}$$

To verify this guess, we first determine $A_0 \cdots A_{n-1}$ for $A_i = \langle a_i; x\rangle$:

$$\langle a_0;x\rangle \cdots \langle a_{n-1};x\rangle \;=\; \left\langle \sum_{i=0}^{n-1} a_i x^i;\, x^n \right\rangle \tag{10}$$

This is obvious for $n = 1$. The inductive step follows from $\left\langle \sum_{i=0}^{n-1} a_i x^i;\, x^n \right\rangle \langle a_n;x\rangle = \left\langle \sum_{i=1}^{n-1} a_i x^i + x^n \cdot a_n;\, x^n \cdot x \right\rangle$ by (1).

From (10), we immediately obtain $I_n = A_0 \cdots A_{n-1}(B) = [s_n \pm x^n]$ where $s_n = \sum_{i=0}^{n-1} a_i x^i$. For $|x| < R$, we know $s_n \to f(x)$ as $n \to \infty$, and $x^n \to 0$ holds if and only if $|x| < 1$. Thus, $I_n \to f(x)$ holds for $|x| < \min(R, 1)$. For $|x| \geq 1$, the radii x^n do not tend to 0. Thus, we have already lost part of the convergence area if $R > 1$.

The LFT's $\langle a_n;x\rangle$ are non-singular iff $x \neq 0$. For $x = 0$, we have $f(x) = a_0$ and $I_n = I_1 = [a_0 \pm 0]$. Thus, we always have a nested sequence of intervals converging to the right value in this singular situation.

The LFT's $\langle a_n;x\rangle$ are refining iff $|a_n| + |x| \leq 1$. Thus in case $x \neq 0$, the generated sequence of intervals is nested iff $|x| \leq 1 - |a_n|$ for all $n \geq 1$, iff $|x| \leq 1 - \sup_{n\geq 1} |a_n|$, and it is eventually nested if $|x| < 1 - \limsup_{n\to\infty} |a_n|$.

Summary of the properties of the Horner evaluation: For $|x| < R$, we have

- $I_n \to f(x)$ if $|x| < 1$;
- $I_n \downarrow f(x)$ if $|x| < 1 - \limsup_{n\to\infty} |a_n|$;
- $I_n \Downarrow f(x)$ if $|x| \leq 1 - \sup_{n\geq 1} |a_n|$.

Let us consider some examples:

- For $f(x) = \frac{1}{1-x} = \sum_{n=0}^{\infty} 1 \cdot x^n$ with convergence radius $R = 1$, the LFT expansion is $\prod_{n=0}^{\infty} \langle 1;x\rangle = \prod_{n=0}^{\infty} \begin{pmatrix} x & 1 \\ 0 & 1 \end{pmatrix}$. The generated intervals are $I_n = [\sum_{i=0}^{n-1} x^i \pm x^n] = [\frac{1-x^n}{1-x} \pm x^n]$. For $|x| < 1$, the center points converge to $\frac{1}{1-x}$ and the radii to 0; thus both end points converge to the correct value $\frac{1}{1-x}$. Yet A_n is refining only if $x = 0$.

Thus, for $x \neq 0$, the sequence of intervals is not nested—not even eventually nested. In fact, for $0 < x < 1$,

$$\left|\frac{1}{1-x} - \frac{1-x^n}{1-x}\right| = \frac{|x^n|}{|1-x|} > |x^n|$$

which means that $f(x) = \frac{1}{1-x}$ is not even contained in any of the intervals I_n!

- For $f(x) = e^x = \sum_{n=0}^{\infty} \frac{1}{n!} \cdot x^n$ with convergence radius $R = \infty$, the LFT expansion is $\prod_{n=0}^{\infty} \left\langle \frac{1}{n!}; x \right\rangle = \prod_{n=0}^{\infty} \begin{pmatrix} n!x & 1 \\ 0 & n! \end{pmatrix}$ with uncomfortably big entries. The general theory tells us that the end points of I_n tend to e^x iff $|x| < 1$ (since for $|x| \geq 1$, the radii do not converge to 0). Matrix A_n is refining iff $|x| \leq 1 - \frac{1}{n!}$. On the negative side, this means that the intervals are only nested if $x = 0$, since A_1 is only refining if $x = 0$. On the positive side, it means that for $|x| < 1$, the interval sequence is eventually nested since $|x| \leq 1 - \frac{1}{n!}$ will hold for sufficiently large n. Hence, in contrast to the first example, the function value e^x is contained in almost all intervals I_n if $|x| < 1$.

3.2 Transformation of the LFT expansion

The LFT expansions derived from Horner's rule in the previous section are not very good: they are rarely nested, loose out part of the convergence circle, and contain uncomfortably large entries for important functions like e^x. There is a general way to transform LFT expansions which may lead to improvements.

Given a sequence $A_0, A_1, A_2, \ldots$ of matrices, the corresponding sequence of intervals $I_1, I_2, \ldots$ is given by $I_n = P_n(B)$ $(n \geq 1)$ where $P_n = A_0 \cdots A_{n-1}$ is the product of the first n matrices. Now, we may choose any sequence $U_1, U_2, \ldots$ of *non-singular* matrices and throw in products $U_i U_i^*$ of U_i and its inverse U_i^* into the original product:

$$P_n = A_0 \cdots A_{n-1} = A_0 U_1 U_1^* A_1 U_2 U_2^* \cdots U_{n-1} U_{n-1}^* A_{n-1}$$

With the new names $Q_n = P_n U_n$ for $n \geq 1$, $B_0 = A_0 U_1$, and $B_i = U_i^* A_i U_{i+1}$ for $i \geq 1$, we obtain

$$Q_n = P_n U_n = B_0 B_1 \cdots B_{n-1}.$$

This means we have defined a new sequence $B_0, B_1, \ldots$ of matrices which denotes the sequence of intervals $J_n = Q_n(B)$.

We want to apply this technique to the affine matrices $A_n = \langle a_n; x \rangle$ coming from Horner's rule. Since we want to preserve affinity here, we choose affine matrices $U_n = \langle c_n; r_n \rangle$ (with $r_n \neq 0$ for non-singularity). Then

$$A_n U_{n+1} = \langle a_n; x \rangle \langle c_{n+1}; r_{n+1} \rangle = \langle a_n + c_{n+1} x;\ r_{n+1} x \rangle$$

Therefore

$$B_0 = A_0 U_1 = \langle a_0 + c_1 x;\ r_1 x \rangle$$

while multiplication with U_n^* from the left yields

$$B_n = \left\langle -\frac{c_n}{r_n}; \frac{1}{r_n} \right\rangle \langle a_n + c_{n+1} x;\ r_{n+1} x \rangle = \left\langle \frac{a_n - c_n + c_{n+1} x}{r_n};\ \frac{r_{n+1}}{r_n} x \right\rangle$$

The matrices Q_n are given by

$$Q_n = P_n U_n = \langle s_n; x^n \rangle \langle c_n; r_n \rangle = \langle s_n + c_n x^n;\ r_n x^n \rangle$$

so that the denoted intervals are $J_n = [(s_n + c_n x^n) \pm r_n x^n]$. Note that for $x = 0$, we have $J_n = [a_0 \pm 0] \Downarrow a_0 = f(0)$.

3.3 The standard choice for c_n and r_n

There are two possibilities to reduce the term $a_n - c_n + c_{n-1}x$ in B_n to just one summand, namely $c_n = 0$ and $c_n = a_n$. In the first case, J_n becomes $[s_n \pm r_n x^n]$, while in the second, we have $J_n = [(s_n + a_n x^n) \pm r_n x^n] = [s_{n+1} \pm r_n x^n]$. This shows that the two possibilities are actually not that different. We prefer the second possibility because it appears to be one step ahead of the first.

The choice $c_n = a_n$ yields

$$B_0 = \begin{pmatrix} r_1 x & a_0 + a_1 x \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B_n = \begin{pmatrix} r_{n+1} x & a_{n+1} x \\ 0 & r_n \end{pmatrix} \quad \text{for } n \geq 1.$$

The resulting intervals were already determined to be $J_n = [s_{n+1} \pm r_n x^n]$. Thus, $J_n \to f(x)$ if $|x| < R$ and $r_n x^n \to 0$. The matrices B_n are refining for $|x| \leq \frac{|r_n|}{|r_{n+1}| + |a_{n+1}|}$.

We now come to a choice for r_n. If all a_n for $n \geq 1$ are non-zero (a_0 may be 0), we may choose $r_n = k a_n$ with some constant $k > 0$. This leads to

$$B_0 = \begin{pmatrix} k a_1 x & a_0 + a_1 x \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B_n = \begin{pmatrix} k a_{n+1} x & a_{n+1} x \\ 0 & k a_n \end{pmatrix} = \begin{pmatrix} kx & x \\ 0 & k q_n \end{pmatrix} \quad \text{for } n \geq 1$$

where $q_n = \frac{a_n}{a_{n+1}}$. The resulting intervals are $J_n = [s_{n+1} \pm k a_n x^n]$. If $|x| < R$, then $\sum_{n=0}^{\infty} a_n x^n$ converges, which implies $a_n x^n \to 0$. Thus $J_n \to f(x)$ holds for all $|x| < R$.

The matrices B_n are refining for $|x| \leq \frac{|k a_n|}{|k a_{n+1}| + |a_{n+1}|} = \frac{k}{k+1} |q_n|$. Thus the interval sequence is eventually nested if $|x| < \frac{k}{k+1} \liminf_{n \to \infty} |q_n|$. Now if $|q_n|$ converges, its limit is R, so that the condition for eventually nesting becomes $|x| < \frac{k}{k+1} R$. The condition for a nested sequence is similar, with $\liminf$ replaced by $\inf$.

- $J_n \to f(x)$ for $|x| < R$;
- $J_n \downarrow f(x)$ for $|x| < \frac{k}{k+1} \liminf_{n \to \infty} |q_n|$ (or $|x| < \frac{k}{k+1} R$ if $|q_n|$ converges);
- $J_n \Downarrow f(x)$ for $|x| \leq \frac{k}{k+1} \inf_{n \geq 1} |q_n|$.

The factor $\frac{k}{k+1}$ increases with k, yet the simplest possibility is $k = 1$ with $\frac{k}{k+1} = \frac{1}{2}$. Values of $|x|$ near R should be avoided anyway since the contractivity of B_n is $|x|/|q_n|$ which converges to $|x|/R$ if $|q_n|$ is convergent; in this case, we have poor contractivity for $|x| \approx R$, and the restriction to $|x| \leq \frac{1}{2} R$ is reasonable.

4 Case studies

Throughout this section, we use the name q for $\inf_{n\geq 1}|q_n|$ which is important for determining the region where a nested sequence of intervals is obtained.

4.1 Exponential function

Recall $e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$. Thus $q_n = \frac{a_n}{a_{n+1}} = \frac{1/n!}{1/(n+1)!} = n+1$, whence $R = \infty$, while $q = \inf_{n\geq 1}|q_n| = 2$. In the standard approach with $k = 1$, we obtain

$$e^x \;=\; \begin{pmatrix} x & x+1 \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} x & x \\ 0 & n+1 \end{pmatrix} \quad \text{nested for } |x| \leq 1. \tag{11}$$

All B_n with $n \geq 1$ have contractivity $\frac{|x|}{n+1} \leq \frac{1}{n+1}$, which seems reasonable. The function can be extended from $[-1,1]$ to the whole real line using the equation $e^{2x} = (e^x)^2$.

With a higher value of k, the area where the intervals are nested is slightly increased, but not beyond $q = 2$. With $k = 3$ for instance, we obtain

$$e^x \;=\; \begin{pmatrix} 3x & x+1 \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} 3x & x \\ 0 & 3(n+1) \end{pmatrix} \quad \text{nested for } |x| \leq \tfrac{3}{2}.$$

The contractivity is not affected by varying k. Even with $k = 1$, representation (11) yields an eventually nested convergent sequence of intervals for all x; the condition $|x| \leq 1$ is only needed to ensure that the sequence is nested from the beginning.

As already mentioned, a representation such as (11) is open to two different interpretations. For rational arguments x, it is (equivalent to) an infinite product of integer matrices, e.g., $e = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} 1 & 1 \\ 0 & n+1 \end{pmatrix}$. In the rational case, it is not absolutely necessary that all matrices but the first one are refining. Consider for instance $x = 2$, where the first three matrices $\begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$, and $\begin{pmatrix} 2 & 2 \\ 0 & 3 \end{pmatrix}$ are not refining, while the next matrix $\begin{pmatrix} 2 & 2 \\ 0 & 4 \end{pmatrix}$ and all subsequent ones are refining. We may multiply the first three matrices into one which gives $e^2 = \begin{pmatrix} 4 & 19 \\ 0 & 3 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} 2 & 2 \\ 0 & n+3 \end{pmatrix}$ which generates a nested sequence of intervals.

For general (real) arguments however, representation (11) should be turned into the infinite tensor expression $e^x = T_0(x, T_1(x, T_2(x, \ldots)))$ with

$$T_0 \;=\; \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad T_n \;=\; \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & n+1 \end{pmatrix}$$

In this case, the above trick to obtain a nested sequence of intervals even for $|x| > 1$ cannot be applied because there is no way to multiply the first few tensors into one (at least in the standard LFT approach). If you try to multiply the corresponding matrices, whose entries depend linearly on x, the entries of the product matrix will be non-linear polynomials in x.

4.2 Cosine and hyperbolic cosine

The power series of the cosine function is $\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}(x^2)^n$. By writing this in terms of x^2 instead of x, zero coefficients are avoided. We get $a_0 = 1$, $a_1 = -\frac{1}{2}$, and $q_n = -(2n+1)(2n+2)$, whence $q = 12$, which for the standard case gives a bound of 6 for a nested sequence. Hence with a bit of scaling

$$\cos x \;=\; \begin{pmatrix} -u & 2-u \\ 0 & 2 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} -u & -u \\ 0 & 2(n+1)(2n+1) \end{pmatrix} \qquad \text{nested for } u = x^2 \leq 6.$$

The function can be extended to the whole real line by $\cos(2x) = 2\cos^2 x - 1$.
The power series for cosh is similar to that for cos: $\cosh x = \sum_{n=0}^{\infty} \frac{1}{(2n)!}(x^2)^n$. Hence, the results look similar; only replace $-u$ by u throughout. The extension formula is again $\cosh(2x) = 2\cosh^2 x - 1$.

4.3 Sine and hyperbolic sine

The power series for sine is $\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$. Here, the exponents of x are odd. They can be expressed in terms of x^2 if one factor of x is taken out: $\frac{\sin x}{x} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(x^2)^n$. We get $a_0 = 1$, $a_1 = -\frac{1}{3}$, and $q_n = -(2n+2)(2n+3)$, whence $q = 20$, and thus

$$\frac{\sin x}{x} \;=\; \begin{pmatrix} -u & 3-u \\ 0 & 3 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} -u & -u \\ 0 & 2(n+1)(2n+3) \end{pmatrix} \qquad \text{nested for } u = x^2 \leq 10.$$

For extension to the whole real line, the formula $\sin(2x) = 2\sin x \cos x$ may be used which unfortunately involves cosine as well. Again, sinh is analogous; just replace $-u$ by u.

4.4 Inverse of tangent

The power series for $\arctan x$ is $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1}$. As with $\sin x$, we consider $\frac{\arctan x}{x} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}(x^2)^n$ instead. Here, $a_0 = 1$, $a_1 = -\frac{1}{3}$, $q_n = -\frac{2n+3}{2n+1}$, whence $q = 1$, and thus

$$\frac{\arctan x}{x} \;=\; \begin{pmatrix} -u & 3-u \\ 0 & 3 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} -(2n+1)u & -(2n+1)u \\ 0 & 2n+3 \end{pmatrix} \qquad \text{for } u = x^2 \leq \tfrac{1}{2}.$$

Again, the hyperbolic variant artanh can be obtained by changing $-u$ into u throughout.

4.5 Natural logarithm

Recall $\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$. Thus $q_n = -\frac{n+1}{n}$ with $q = 1$, and we obtain after some scaling

$$\ln(1+x) \;=\; \begin{pmatrix} x & x \\ 0 & 1 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} -nx & -nx \\ 0 & n+1 \end{pmatrix} \qquad \text{nested for } |x| \leq 1/2.$$

An arbitrary positive argument of ln can be moved towards 1 by the equations $\ln(2x) = \ln x + \ln 2$ and $\ln(\frac{1}{2}x) = \ln x - \ln 2$. While $\ln 2 = \ln(1+1)$ is outside the scope of the above formula, it can be obtained as $\ln 2 = -\ln\frac{1}{2} = -\ln(1-\frac{1}{2})$ which gives

$$\ln 2 \;=\; \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} n & n \\ 0 & 2(n+1) \end{pmatrix}$$

where the matrices have contractivities of about $\frac{1}{2}$.

A better way to obtain $\ln 2$ is at follows: Recall $\operatorname{artanh} x = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$, and therefore, $\ln 2 = \ln\left(\frac{1+1/3}{1-1/3}\right) = 2\operatorname{artanh}(1/3) = \frac{2}{3}\frac{\operatorname{artanh}(1/3)}{1/3}$. Using the results of the previous subsection (with $u = 1/9$) yields

$$\ln 2 \;=\; \begin{pmatrix} 2 & 56 \\ 0 & 81 \end{pmatrix} \prod_{n=1}^{\infty} \begin{pmatrix} 2n+1 & 2n+1 \\ 0 & 9(2n+3) \end{pmatrix}$$

where the matrices have contractivities of about $\frac{1}{9}$.

References

Edalat, A., Potts, P. (1997): A new representation for exact real numbers. In: Thirteenth Annual Conference on Mathematical Foundations of Programming Semantics, MFPS XIII, Pittsburgh, Pennsylvania. Elsevier Science B.V., Amsterdam, URL: http://www.elsevier.nl/locate/entcs/volume6.html (Electronic Notes in Theoretical Computer Science, vol. 6)

Gosper, W. (1972): Continued fraction arithmetic. In: Technical Report HAKMEM Item 101B, MIT Artificial Intelligence Memo 239, MIT.

Heckmann, R. (2000): How many argument digits are needed to produce *n* result digits? In: Real Number Computation, RealComp '98, Indianapolis, Indiana. Elsevier Science B.V., Amsterdam, URL: http://www.elsevier.nl/locate/entcs/volume24.html (Electronic Notes in Theoretical Computer Science, vol. 24)

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

Potts, P.J. (1998a): Efficient on-line computation of real functions using exact floating point. Draft report, Imperial College, London.
URL: http://www.purplefinder.com/~potts/pub/phd/efficient.ps.gz

Potts, P.J. (1998b): Exact Real Arithmetic using Möbius Transformations. PhD thesis, Imperial College, London. URL: http://www.purplefinder.com/~potts/thesis.ps.gz

Potts, P.J., Edalat, A. (1996): Exact real arithmetic based on linear fractional transformations. Draft report, Imperial College, London.
URL: http://www.purplefinder.com/~potts/pub/phd/stanford.ps.gz

Potts, P.J., Edalat, A. (1997): Exact real computer arithmetic. Draft report, Imperial College, London. URL: http://www.purplefinder.com/~potts/pub/phd/normal.ps.gz

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