Electronic Notes in Theoretical Computer Science 13 (1998)
URL: http://www.elsevier.nl/locate/entcs/volume13.html 20 pages

Big Integers and Complexity Issues in Exact Real Arithmetic ⋆

Reinhold Heckmann

FB 14 – Informatik, Universität des Saarlandes
Postfach 151150, D-66041 Saarbrücken, Germany

e-mail: heckmann@cs.uni-sb.de

Abstract

One possible approach to exact real arithmetic is to use linear fractional transformations to represent real numbers and computations on real numbers. We show how to determine the digits that can be emitted from a transformation, and present a criterion which ensures that it is possible to emit a digit. Using these results, we prove that the obvious algorithm to compute n digits from the application of a transformation to a real number has complexity $O(n^2)$, and present a method to reduce this complexity to that of multiplying two n bit integers.

1 Introduction

Linear Fractional Transformations (LFT's) provide an elegant approach to real number arithmetic [5,14,9,12,10,4]. One-dimensional LFT's $x \mapsto \frac{ax+c}{bx+d}$ are used as digits 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. In Section 2, we present the LFT approach in some detail. This provides the background for understanding the results in the remainder of this paper.

LFT's can be modelled 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*). We refer to matrices and tensors collectively as *transformers*. Basic

⋆ Most of the results in this paper were found during a visiting fellowship of the author at Imperial College, London. This visit was organised by Abbas Edalat and funded by EPSRC.

computational steps such as consuming one digit of the argument(s) (*absorption*) or producing one digit of the result (*emission*) can be realised as variants of matrix multiplication applied to a transformer and a digit matrix.

Usually, all the transformers employed in real number arithmetic have integer components. In Section 3, we reiterate the main result of [6]: if the difference of the column sums of a transformer is not zero, at least one entry of the transformer has bit size $\Omega(n)$ after n digits have been emitted (law of big numbers).

In Section 4, we first show how to check whether any digit can be emitted from a given transformer, and how to determine this digit. Then, we introduce attributes of a matrix—*shrink factor* and *contractivity*—which are useful for predicting when emission is possible. Using these results, we are able to show that in the cases not covered by the law of big numbers, the entries of a matrix are bounded by a constant (Section 5).

In Section 6, we discuss the impact of these results on the complexity of real number computation. In particular, we consider the time needed to compute n digits from the application of an LFT to a real number. The obvious evaluator that handles each digit individually needs time $O(n^2)$ if the law of big numbers applies, and time $O(n)$ otherwise. To reduce the quadratic complexity, we propose to combine many digits in a small basis to one digit in a large basis. By this method, the complexity is reduced to that of multiplying two n bit integers.

2 Exact Real Arithmetic by Linear Fractional Transformations

In this section, we present the framework of exact real arithmetic via LFT's [5,14,9], specialised to the version used by the group of Edalat and Potts at Imperial College [12,10,11,13,4].

2.1 LFT's and Matrices

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

It is useful to present the four parameters of an LFT as a 2-2-matrix $A = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ of integers, hereafter *matrix*. Every matrix denotes an LFT $\langle A \rangle$, given by $\langle A \rangle(x) = \frac{ax+c}{bx+d}$. LFT's described by non-singular matrices, i.e., matrices A with determinant $\det A = ad - bc \neq 0$, are considered as endofunctions of $\mathbb{R}^\star = \mathbb{R} \cup \{\infty\}$, the one-point compactification of the real line. The value ∞ arises as $r/0$ with $r \neq 0$, and on the other hand, $\langle A \rangle(\infty)$ is defined as a/b. For LFT's described by singular matrices, an additional 'number' $\perp$ (undefined) is needed which arises as $0/0$. The value of $\langle A \rangle(\perp)$ is defined as $\perp$.

The mapping $A \mapsto \langle A \rangle$ is not one-to-one; for, $\langle A \rangle = \langle kA \rangle$ holds for all integers $k \neq 0$. We shall write $A \cong B$ if $\langle A \rangle = \langle B \rangle$, or equivalently $B = kA$ for some $k \neq 0$. A matrix is called *k-reducible* if the integer k is a common factor of its four components. Division of a k-reducible matrix by k is called *reduction by k*. A matrix is *in lowest terms* if there is no common factor other than 1 and -1. All matrices different from $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ are equivalent to a matrix in lowest terms.

Composition of LFT's corresponds to matrix multiplication: $\langle A \rangle \circ \langle B \rangle = \langle A \cdot B \rangle$. The equivalence relation '$\cong$' is a congruence w.r.t. multiplication.

Because of the equation $\det(rA) = r^2 \det A$, the determinant of a matrix is not invariant under equivalence '$\cong$', but its sign (1, 0, or -1) is, i.e., the sign of the determinant of A is a well-defined property of the LFT $\langle A \rangle$. LFT's with non-zero determinant (non-singular LFT's) are invertible. To obtain an integer representation of $\langle A \rangle^{-1}$, the *pseudo-inverse* A^* can be used. It is defined by

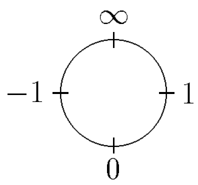
$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}^* = \begin{pmatrix} d & -c \\ -b & a \end{pmatrix} \tag{1}$$

Clearly, $\det(A^*) = \det A$ holds. The main property of the pseudo-inverse operation is

$$A \cdot A^* = A^* \cdot A = \det A \cdot \mathrm{E} \tag{2}$$

where $\mathrm{E} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ is the identity matrix, and so, $A \cdot A^* = A^* \cdot A \cong \mathrm{E}$ if $\det A \neq 0$, whence $\langle A \rangle^{-1} = \langle A^* \rangle$.

2.2 Representing Reals by LFT's



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

Non-singular LFT's map intervals to intervals: if $\det M > 0$, then $\langle M \rangle [u, v]$ is $[\langle M \rangle u, \langle M \rangle v]$, while for $\det M < 0$, we get $\langle M \rangle [u, v] = [\langle M \rangle v, \langle M \rangle u]$. For the interval $[0, \infty]$, these formulae simplify to $\langle M \rangle [0, \infty] = [\frac{c}{d}, \frac{a}{b}]$ for $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ with $\det M > 0$, and $[\frac{a}{b}, \frac{c}{d}]$ for $\det M < 0$.

Thus, an infinite stream of non-singular matrices M_0, M_1, ... defines the interval

$$\bigcap_{n=0}^{\infty} \langle M_0 \cdot M_1 \cdot \ldots \cdot M_n \rangle [0, \infty] \ . \tag{3}$$

The intersection is filtered (decreasing) if $\langle M_n \rangle [0, \infty] \subseteq [0, \infty]$ holds for all $n > 0$. This inclusion property is equivalent to the condition that all entries of M_n are ≥ 0 or all are ≤ 0. Matrices with all entries ≥ 0 are called *positive*.

If almost all LFT's $\langle M_0 \rangle$, $\langle M_1 \rangle$, ... are sufficiently contractive, then the intersection in (3) shrinks to a singleton set. In this case, the stream of matrices or LFT's denotes a unique real number (it *converges*). In [7], some sufficient criteria for convergence are presented.

Because of the usage of matrix multiplication in (3), we consider a stream of matrices converging to a real number x as a (formal) infinite product, and write $x = \prod_{n=0}^{\infty} M_n$. Many real numbers can be elegantly represented by such infinite products, e.g., $\sqrt{2} = \prod_{n=0}^{\infty} \left(\begin{smallmatrix} 1 & 2 \\ 1 & 1 \end{smallmatrix}\right)$ or $\mathrm{e} = \prod_{n=0}^{\infty} \left(\begin{smallmatrix} 2n+2 & 2n+1 \\ 2n+1 & 2n \end{smallmatrix}\right)$. 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. The group of Edalat and Potts at Imperial College [11,4] proposed such a standard form, where the first matrix M_0 must be one of four *sign matrices*, while the remaining ones are taken from a finite set of *digit matrices*. Digit matrices are positive and contracting, so that the intersection in (3) is decreasing and converges to a real number.

The four possible sign matrices correspond to rotations of the unit circle by 0°, 90°, 180°, and 270°. They can be explicitly described as follows:

$$\begin{aligned} S_+ &= \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) & \langle S_+\rangle\,[0,\infty] &= [0,\infty] \\ S_\infty &= \left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right) & \langle S_\infty\rangle[0,\infty] &= [1,-1] \\ S_- &= \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) & \langle S_-\rangle\,[0,\infty] &= [\infty,0] \\ S_0 &= \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right) & \langle S_0\rangle\;[0,\infty] &= [-1,1] \end{aligned}$$

S_0 and S_∞ are pseudo-inverse to each other; $S_0 \cdot S_\infty = S_\infty \cdot S_0 = 2\mathrm{E}$ holds.

There are many possible sets of digit matrices, one for every base $r > 1$. The implementation of Edalat and Potts [4] uses base $r = 2$. In this paper, we consider integer bases $r \geq 2$.

Fix an integer $r \geq 2$. Every real number in the interval $[-1, 1]$ has a representation as $\sum_{n=1}^{\infty} k_n r^{-n}$ with integer digits k_n satisfying $|k_n| < r$. (Digits may be negative [1].) These digits correspond to affine maps $x \mapsto \frac{x+k}{r}$ that are LFT's $\langle A_k^r\rangle$ with $A_k^r = \left(\begin{smallmatrix} 1 & k \\ 0 & r \end{smallmatrix}\right)$, mapping the interval $[-1, 1]$ into $[\frac{k-1}{r}, \frac{k+1}{r}]$. These image intervals have length $2/r$ and cover $[-1, 1]$. The image intervals coming from successive values of k overlap in a common interval of length $1/r$. This provides the redundancy needed in exact real arithmetic.

Since the base interval in (3) is $[0, \infty]$ and not $[-1, 1]$, the maps $\langle A_k^r\rangle$ have to be transformed into that interval. This can be done by composition with the maps $\langle S_\infty\rangle$ and $\langle S_0\rangle$, which are mutually inverse bijections between $[-1, 1]$ and $[0, \infty]$. Thus, the actual digit matrices are

$$D_k^r \;=\; S_\infty \cdot A_k^r \cdot S_0 \;=\; \begin{pmatrix} r+k+1 & r+k-1 \\ r-k-1 & r-k+1 \end{pmatrix} \tag{4}$$

and their pseudo-inverses are given by

$$(D_k^r)^* \;=\; \begin{pmatrix} 1-k+r & 1-k-r \\ 1+k-r & 1+k+r \end{pmatrix}\;. \tag{5}$$

Since the two entries in the top row of D_k^r differ by 2, these matrices are either in lowest terms or 2-reducible. The latter case occurs iff the parities of r and

Table 1
Digit matrices for base 2

k	A_k^2	D_k^2	lowest terms	$\langle D_k^2\rangle([0,\infty])$
-1	$\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 \\ 2 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$	$[0,1]$
0	$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$	$[\frac{1}{3},3]$
1	$\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 4 & 2 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$	$[1,\infty]$

k are different. In base $r = 2$ for instance, the digit matrices with $k \neq 0$ are 2-reducible, while that with $k = 0$ is not (see Table 1).

2.3 Compressing Digits

It is a familiar property of number systems used to represent integers that n digits in base r can be combined to one digit in base r^n. A similar result holds for the digit matrices presented above. Multiplying two digit matrices yields:

$$D_k^r\, D_{k'}^{r'} \;=\; S_\infty\, A_k^r\, S_0\, S_\infty\, A_{k'}^{r'}\, S_0 \;=\; 2\, S_\infty\, A_k^r\, A_{k'}^{r'}\, S_0 \;=\; 2\, D_{kr'+k'}^{rr'} \tag{6}$$

Here, the second equality is due to $S_0\, S_\infty = 2\mathrm{E}$, and the third due to

$$A_k^r \cdot A_{k'}^{r'} = \begin{pmatrix} 1 & k \\ 0 & r \end{pmatrix} \cdot \begin{pmatrix} 1 & k' \\ 0 & r' \end{pmatrix} = \begin{pmatrix} 1 & k' + kr' \\ 0 & rr' \end{pmatrix} \tag{7}$$

together with the estimation $|kr' + k'| \leq (r-1)r' + (r'-1) = rr' - 1$.

Iterating (6) leads to

$$D_{k_1}^r \cdot \ldots \cdot D_{k_n}^r \;=\; 2^{n-1}\, D_k^{r^n} \quad \text{where} \quad k = \sum_{i=1}^{n} k_i r^{n-i} \;. \tag{8}$$

Hence, we obtain:

(i) The product of n digit matrices in base r is always 2^{n-1}-reducible.

(ii) After 2^{n-1}-reduction, the result is a digit matrix in base r^n.

2.4 Computation by LFT's

LFT's can be used not only to represent real numbers, but also to perform computations with real numbers. For the sake of simplicity, we only present computations within the interval $[0,\infty]$ where real numbers can be represented by a stream of digit matrices without a leading sign matrix.

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. Recall that an LFT maps $[0,\infty]$ into itself iff it can be represented by a positive matrix (all components ≥ 0).

Given a positive matrix M, the actual computation of $\langle M\rangle(x)$ is performed by a sequence of *absorptions* and *emissions*. Absorption means that M consumes the first digit D_1 of x, thereby becoming $M \cdot D_1$, which is positive again. It corresponds to the equality

$$M \cdot (D_1 \cdot D_2 \cdot \ldots) \;=\; (M \cdot D_1) \cdot (D_2 \cdot \ldots) \;. \tag{9}$$

Emission means that M produces one further digit D of the result, thereby becoming $D^* \cdot M$. It corresponds to the equivalence

$$(D_1 \cdot \ldots \cdot D_n) \cdot M \;\cong\; (D_1 \cdot \ldots \cdot D_n \cdot D) \cdot (D^* \cdot M) \;. \tag{10}$$

Emission of a digit D is allowed only if $D^* \cdot M$ is positive. For small bases such as $r = 2$, it is possible to check for each digit matrix D_k^r individually whether it may be emitted. Of course, this method is unsuitable for large bases. We will return to this issue in Section 4.1.

Because of the built-in redundancy, there are often two, sometimes even three different candidates for emission. In this case, it does not matter which one is chosen.

A possible strategy for the computation of $\langle M \rangle(x)$ is as follows: emit digits until no further emission is possible, then absorb one digit of x, again emit digits until no longer possible, etc. Later, we shall see that $O(n)$ absorptions are sufficient to obtain n emitted digits (Theorem 4.5).

2.5 Tensors

To compute sums, products, etc., *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 $T = \left(\begin{smallmatrix} a & c & e & g \\ b & d & f & h \end{smallmatrix}\right)$ denotes the function $\langle T \rangle : \mathbb{R}^\star{}_\perp \times \mathbb{R}^\star{}_\perp \to \mathbb{R}^\star{}_\perp$ given by $\langle T \rangle(x, y) = \frac{axy+cx+ey+g}{bxy+dx+fy+h}$. For tensors, the notions of reducible, reduction, and lowest terms can be defined analogous to the case of matrices. Likewise for positivity: a two-dimensional LFT maps $[0, \infty]^2$ to $[0, \infty]_\perp$ iff it can be represented by a positive tensor, i.e., a tensor with components ≥ 0. Because of these analogies, we refer to matrices and tensors collectively as *transformers*.

It is easy to represent addition, subtraction, multiplication, and division by suitable tensors [5,14,12,10,11]. Tensors may also be used to represent transcendental functions, e.g., $\arctan x = \langle T_0 \rangle(x, \langle T_1 \rangle(x, \langle T_2 \rangle(x, \ldots)))$ where $T_n = \left(\begin{smallmatrix} 0 & 1 & 0 & 0 \\ (n+1)^2 & 0 & 0 & 2n+1 \end{smallmatrix}\right)$. It remains to show how to actually compute $\langle T \rangle(x, y)$ for a given positive integer tensor T [10,11].

Emissions can be done as in the one-dimensional case: in emitting a digit D, tensor T is replaced by $D^* \cdot T$, which is a tensor again. Emission of D is allowed only if $D^* \cdot T$ is positive.

Since digits can be absorbed from both arguments, there are two kinds of *absorptions*: absorption of a digit D from the left argument transforms T into $T \circledL D$, while absorption from the right argument yields $T \circledR D$. Right absorption can be defined by writing a tensor T as a row $(T^{\mathrm{L}}, T^{\mathrm{R}})$ of two matrices, and specifying

$$(T^{\mathrm{L}}, T^{\mathrm{R}}) \circledR M \;=\; (T^{\mathrm{L}}M,\; T^{\mathrm{R}}M) \;. \tag{11}$$

To define left absorption, let $T^\times$ be T with the two middle columns exchanged. Then let $T \circledL M = (T^\times \circledR M)^\times$, and so

$$(T \circledL M)^\times = T^\times \circledR M \qquad \text{and} \qquad (T \circledR M)^\times = T^\times \circledL M \;. \tag{12}$$

Later, we shall see that D-emissions and D-absorptions have many properties in common. Thus, we introduce a common name: a *D-transaction* at a transformer is either a D-emission or a D-absorption.

3 The Appearance of Big Integers

Naively, one may think that the entries of a transformer become bigger by absorptions, and become smaller again by emissions if common factors are cancelled out (*reduction*). However, practical experiments have shown that the size of the biggest entry usually increases with the number of transactions. This impression was confirmed by a formal analysis in [6]. For the sake of completeness, we repeat a shortened version of the proof of this important result in this section.

3.1 Big Numbers in Matrices

In this subsection, we derive lower bounds for the entries of a matrix after n transactions and all possible reductions. This is done by observing how the determinant and the so-called column difference are changed by transactions and reductions, and by deriving a reduction invariant from this.

Determinants are easy because of $\det(A \cdot B) = \det A \cdot \det B$, which implies $\det(M \cdot D_k^r) = \det((D_k^r)^* \cdot M) = \det D_k^r \cdot \det M$. How big is $\det D_k^r$? Since $\det S_0 = \det S_\infty = 2$ and $\det A_k^r = \det \left(\begin{smallmatrix} 1 & k \\ 0 & r \end{smallmatrix}\right) = r$, we have $\det D_k^r = \det(S_\infty A_k^r S_0) = 4r$.

In the following list, let M be a matrix, and let M' be the result of applying a transaction or reduction to M.

- Transaction with D_k^r: $\det M' = 4r \det M$,
- Reduction by k: $\det M' = \frac{1}{k^2} \det M$.

The *column difference* of a matrix is $\mathrm{cd} \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right) = (a+b) - (c+d)$. Digit matrices and their inverses have column difference 0. Consider the product of $(D_k^r)^*$ (5) with a vector $\left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$:

$$\begin{pmatrix} 1-k+r & 1-k-r \\ 1+k-r & 1+k+r \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (1-k)(u+v) + r(u-v) \\ (1+k)(u+v) - r(u-v) \end{pmatrix} \tag{13}$$

which implies

$$(D_k^r)^* \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u' \\ v' \end{pmatrix} \quad \Longrightarrow \quad u' + v' = 2(u+v) \;. \tag{14}$$

From this, $\mathrm{cd}((D_k^r)^* \cdot M) = 2\,\mathrm{cd}\,M$ follows. It is straightforward to verify that $\mathrm{cd}(M \cdot D_k^r) = 2\,\mathrm{cd}\,M$ holds as well. Thus, we obtain:

- Transaction with D_k^r: $\mathrm{cd}\,M' = 2\,\mathrm{cd}\,M$,
- Reduction by k: $\mathrm{cd}\,M' = \frac{1}{k}\,\mathrm{cd}\,M$.

Hence, the properties of having zero or non-zero column difference are transaction invariants.

For a matrix M with $\mathrm{cd}\,M \neq 0$, the quotient $\mathrm{qcd}\,M = \frac{\det M}{(\mathrm{cd}\,M)^2}$ is a well-defined rational number. By a transaction with D_k^r, this quotient is multiplied by $\frac{4r}{2^2} = r$; and a k-reduction yields a factor of $\frac{1/k^2}{(1/k)^2} = 1$. Thus, the quotient qcd is invariant under reductions, and is multiplied by r in every transaction. Therefore, if M_0 is some initial matrix with $\mathrm{cd}\,M_0 \neq 0$, and M_n the result of applying n transactions in base r to M_0, and all possible reductions, then $\mathrm{qcd}\,M_n = r^n\,\mathrm{qcd}\,M_0$. This equation can be turned into an integer equation by multiplying by the denominators:

$$\det M_n \cdot (\mathrm{cd}\,M_0)^2 = r^n \cdot \det M_0 \cdot (\mathrm{cd}\,M_n)^2 \tag{15}$$

If $\mathrm{cd}\,M_0 \neq 0$, then $\mathrm{cd}\,M_n \neq 0$, too. As an integer, $(\mathrm{cd}\,M_n)^2$ is at least 1. This gives a lower bound for the determinant:

$$|\det M_n| \;\geq\; \frac{|\det M_0|}{(\mathrm{cd}\,M_0)^2} \cdot r^n \quad . \tag{16}$$

The determinant does not directly give information about the sizes of the entries of a matrix. A better measure is the maximum of their absolute values: $\left\|\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)\right\| = \max(|a|, |b|, |c|, |d|)$. A lower bound for the determinant of a matrix M can be turned into a lower bound for the norm $\|M\|$ using the inequality $\|M\| \geq \sqrt{\frac{1}{2}|\det M|}$, which follows from the definition of the determinant as $\det\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right) = ad - bc$. Thus, we obtain from (16):

$$\|M_n\| \;\geq\; \sqrt{\frac{|\det M_0|}{2(\mathrm{cd}\,M_0)^2}} \cdot \left(\sqrt{r}\right)^n \quad . \tag{17}$$

Thus, if in addition $\det M_0 \neq 0$, *even if all possible reductions are performed, the entries of the matrix are bound to grow exponentially in the number of transactions.*

It is more useful to consider the bit sizes of the entries instead of the entries themselves. The bit size of a number m is $\log m$.

Theorem 3.1 (Law of big numbers) *Let M be a matrix with non-zero determinant and non-zero column difference. After n transactions at M, at least one entry of the result has bit size $\Omega(n)$, even if all possible reductions are performed.*

The law of big numbers means that the usage of big integers is unavoidable in exact real arithmetic, in the signed digit approach of Edalat's group. It applies even in the simplest cases. For instance, doubling of an unsigned real is effected by the matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ that has determinant 2 and column difference 1, halving by $\left(\begin{smallmatrix}1&0\\0&2\end{smallmatrix}\right)$ with determinant 2 and column difference -1, and addition of 1 by the matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ with determinant 1 and column difference -1.

The law of big numbers does not apply to matrices with zero column difference. The simplest example is the identity matrix $\mathrm{E} = \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$. According to (2), after a D-absorption, a subsequent D-emission, and a reduction by $\det D$, the identity matrix is recovered. Repeating this cycle, we see that there are arbitrarily long sequences of transactions at the identity matrix which do not lead to entries bigger than $4r$. In [6], it was an open problem whether such a fixed bound can be found for any matrix with column difference 0. Meanwhile, this question was settled positively; we present a proof in Section 5.

3.2 Big Numbers in Tensors

In this subsection, we derive analogues of the results of the previous section for tensors. The proceeding is similar, but a major obstacle is that tensors do not have determinants. Fortunately there is a suitable substitute.

We start by introducing an analogue to the column difference of a matrix. Writing a tensor T as a row $(T^{\mathrm{L}}, T^{\mathrm{R}})$ of two matrices, its *column difference* $\mathrm{cd}\, T$ is defined from the column differences of the two matrices: $\mathrm{cd}(T^{\mathrm{L}}, T^{\mathrm{R}}) = \mathrm{cd}\, T^{\mathrm{L}} - \mathrm{cd}\, T^{\mathrm{R}}$, i.e.,

$$\mathrm{cd}\begin{pmatrix} a & c & e & g \\ b & d & f & h \end{pmatrix} = (a+b) - (c+d) - (e+f) + (g+h) \ .$$

From (11) and the properties of cd, we obtain for all digit matrices D

$$\mathrm{cd}((T^{\mathrm{L}}, T^{\mathrm{R}}) \circledR D) = \mathrm{cd}(T^{\mathrm{L}} \cdot D) - \mathrm{cd}(T^{\mathrm{R}} \cdot D) = 2\, \mathrm{cd}(T^{\mathrm{L}}, T^{\mathrm{R}}) \ .$$

By $(T \circledR D)^{\times} = T^{\times} \copyright D$ (12) and $\mathrm{cd}(T^{\times}) = \mathrm{cd}\, T$, we obtain the corresponding formula $\mathrm{cd}(T \copyright D) = 2\, \mathrm{cd}\, T$. From (14), $\mathrm{cd}(D^{*} \cdot T) = 2\, \mathrm{cd}\, T$ follows for all digit matrices D. Therefore, 'cd' for tensors behaves exactly as 'cd' for matrices:

- Transaction with D_k^r: $\quad \mathrm{cd}\, T' = 2\, \mathrm{cd}\, T$,
- Reduction by k: $\quad \mathrm{cd}\, T' = \frac{1}{k}\, \mathrm{cd}\, T$.

Again, the properties of having zero or non-zero column difference are transaction invariants.

A suitable substitute for the determinant of a matrix is the *column determinant* $\mathrm{cdet}\, T$ of a tensor T, defined by

$$\mathrm{cdet}\begin{pmatrix} a & c & e & g \\ b & d & f & h \end{pmatrix} = (a+b)(g+h) - (c+d)(e+f) \ . \tag{18}$$

Because of (14), all four column sums are doubled by an emission, and so, $\mathrm{cdet}(D^{*} \cdot T) = 4\, \mathrm{cdet}\, T$ holds for all tensors T and digit matrices D. Note that in contrast to the determinant of matrices, the factor is not $\det D^{*} = 4r$, but only 4. On the other side, the column determinant is multiplicative w.r.t. absorptions; for any tensor T and matrix M,

$$\mathrm{cdet}(T \copyright M) = \mathrm{cdet}(T \circledR M) = \mathrm{cdet}\, T \cdot \det M \tag{19}$$

holds. Here, the first equality follows from (12) and $\operatorname{cdet}(T^\times) = \operatorname{cdet} T$, while the proof of the second equality is a straightforward, but tedious exercise in algebraic manipulations.

Summarising and specialising to the case of digit matrices, we obtain:

- Emission of D^r_k: $\quad \operatorname{cdet} T' = 4 \operatorname{cdet} T$,
- Absorption of D^r_k: $\quad \operatorname{cdet} T' = 4r \operatorname{cdet} T$,
- Reduction by k: $\quad \operatorname{cdet} T' = \frac{1}{k^2} \operatorname{cdet} T$.

In contrast to matrices, emissions and absorptions behave differently.

For a tensor T with $\operatorname{cd} T \neq 0$, we consider the quotient $\operatorname{qcd} T = \frac{\operatorname{cdet} T}{(\operatorname{cd} T)^2}$. This quotient is invariant under reductions and emissions. Every absorption yields a factor of r. Therefore, if T_0 is some initial tensor with $\operatorname{cd} T_0 \neq 0$, and T_n the result of applying n absorptions, any number of emissions, and all possible reductions to T_0, then $\operatorname{qcd} T_n = r^n \operatorname{qcd} T_0$. As in the case of matrices, a lower bound for the column determinant follows:

$$|\operatorname{cdet} T_n| \;\geq\; \frac{|\operatorname{cdet} T_0|}{(\operatorname{cd} T_0)^2} \cdot r^n \quad . \tag{20}$$

For tensors T, we define a norm $\|T\|$ as the maximum of the absolute values of the eight entries. Because of $\|T\| \geq \sqrt{\frac{1}{8}|\operatorname{cdet} T|}$, we obtain

$$\|T_n\| \;\geq\; \sqrt{\frac{|\operatorname{cdet} T_0|}{8(\operatorname{cd} T_0)^2}} \cdot \left(\sqrt{r}\right)^n \quad . \tag{21}$$

Formulating this in terms of bit sizes yields:

Theorem 3.2 (Law of big numbers for tensors) *Let T be a tensor with non-zero column determinant and non-zero column difference. After n absorptions and any number of emissions at T, at least one entry of the result has bit size $\Omega(n)$, even if all possible reductions are performed.*

The tensors that realise the four basic arithmetic operations satisfy the hypotheses of the law of big numbers:

Addition:	$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\operatorname{cdet} = -1$	$\operatorname{cd} = -1$
Subtraction:	$\begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\operatorname{cdet} = 1$	$\operatorname{cd} = 1$
Multiplication:	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\operatorname{cdet} = 1$	$\operatorname{cd} = 2$
Division:	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	$\operatorname{cdet} = -1$	$\operatorname{cd} = -2$

Yet the tensor for the mean value operation is different:

Mean value:	$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$	$\operatorname{cdet} = -1$	$\operatorname{cd} = 0$

Does this mean that $\frac{1}{2}x$, which leads to big numbers when computed with $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, can be computed as $\frac{0+x}{2}$ avoiding big numbers? The answer is no, at least in the case $r = 2$. Let T^{R} be the matrix on the right hand side of the tensor T. The equations $(D^* \cdot T)^{\mathrm{R}} = D^* \cdot T^{\mathrm{R}}$ and $(T \circledR D)^{\mathrm{R}} = T^{\mathrm{R}} \cdot D$ hold for all tensors T and digit matrices D. This means that the right half of $\left(\begin{smallmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{smallmatrix}\right)$ behaves exactly as the halving matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ during emissions and absorptions from the right. Since the number 0 is represented by the infinite product $(D^2_{-1})^\omega$ or $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 2 \end{smallmatrix}\right)^\omega$, and $\left(T \circledL \left(\begin{smallmatrix} 1 & 0 \\ 1 & 2 \end{smallmatrix}\right)\right)^{\mathrm{R}} = 2T^{\mathrm{R}}$, the correspondence is only changed by a common factor during absorptions from the left. Hence, after any number of transactions, the right half of the resulting tensor is a multiple of the matrix resulting from $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ by the corresponding sequence of transactions. Thus, it has entries which are at least as big as the entries of the matrix, which are big by Theorem 3.1.

3.3 Discussion

The laws of big numbers as derived above apply to unsigned reals only. For instance, halving in the zero interval $[-1, 1]$ with base $r = 2$ means putting D^2_0 in front of the unsigned part of the argument, an operation possible without employing big integers.

Of course, our results crucially depend on the choice of the digit matrices. All digit matrices for all bases have zero column difference, and this fact is implicitly used in the derivations of the formulae for the cd values after transactions. A completely different choice of digit matrices, with non-zero column difference, may change everything. Also, the results may look different if irrational bases are used such as the golden ratio. However, we believe that big numbers cannot be avoided even in these cases, although we do not have a proof.

4 Emission from Matrices and Tensors

Let r be a fixed basis, i.e., an integer greater than 1. In order to perform emissions, we need to know for a given positive transformer A whether there is an integer k with $|k| < r$ such that $(D^r_k)^* \cdot A \geq 0$. As already mentioned, this question can be answered by examining all possible values of k. Of course, this method is only efficient if r is small.

In this section, we present a direct method to find a suitable value of k if it exists. After this, we introduce two attributes of a matrix M, the *shrink factor* $\operatorname{shr} M$ and the *contractivity* $\operatorname{con} M$ that allow the prediction of the existence of suitable values of k. Thus, we obtain an algorithm suitable for large bases r, and moreover, valuable theoretical insights for a closer analysis of the basic computational processes in the LFT framework.

4.1 Computing Digits that can be Emitted

To study $(D^r_k)^* \cdot A$ for matrices or tensors A, it suffices to consider the simpler case where A is replaced by a column vector of the original transformer. From

(13), we know that

$$(22) \qquad \begin{pmatrix} u' \\ v' \end{pmatrix} = (D_k^r)^* \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (1-k)(u+v) + r(u-v) \\ (1+k)(u+v) - r(u-v) \end{pmatrix} .$$

Assuming that $\binom{u}{v}$ is positive, i.e., $u, v \geq 0$, and different from $\binom{0}{0}$, we want to ensure that $u', v' \geq 0$ or $u', v' \leq 0$. From (22), $u' + v' = 2(u+v)$ follows, and so, the case $u', v' \leq 0$ need not be considered. Condition $u' \geq 0$ or $(1-k)(u+v) + r(u-v) \geq 0$ is equivalent to $k \leq r\frac{u-v}{u+v} + 1$, while $v' \geq 0$ iff $k \geq r\frac{u-v}{u+v} - 1$. Together, this means $\left|k - r\frac{u-v}{u+v}\right| \leq 1$.

Let $q = r\frac{u-v}{u+v}$. We want to compute the *emission set* of q, i.e., the set of all integers k satisfying $|k-q| \leq 1$ and also the general digit condition $|k| \leq r-1$. If q happens to be an integer, there are three integers k satisfying $|k-q| \leq 1$, namely q itself, $q-1$, and $q+1$. If q is not an integer, there are merely two such integers, namely $\lfloor q \rfloor$ and $\lceil q \rceil$. Determining these candidates is possible by integer operations (including the division $(r(u-v))/(u+v)$). Since $u, v \geq 0$, $|q| = r\frac{|u-v|}{u+v} \leq r$ holds. Thus, the digit condition $|k| \leq r-1$ can never rule out all integer solutions of $|k-q| \leq 1$, and so, every q arising in the situation considered here has a non-empty emission set.

Applying these results to the case of a positive non-singular matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$, we see that $(D_k^r)^* \cdot M \geq 0$ iff

$$(23) \qquad \left|k - r\frac{a-b}{a+b}\right| \leq 1 \qquad \text{and} \qquad \left|k - r\frac{c-d}{c+d}\right| \leq 1 .$$

Hence, we look for the intersection of the emission sets of $q_1 = r\frac{a-b}{a+b}$ and $q_2 = r\frac{c-d}{c+d}$. If q_1 and q_2 are too far apart, this intersection is empty, and no digit can be emitted. If q_1 and q_2 are close together, the intersection may have more than one element. In this case, it does not matter which one is chosen for the emission.

For a tensor $T = \left(\begin{smallmatrix} a & c & e & g \\ b & d & f & h \end{smallmatrix}\right)$, the set of digits that may be emitted is the intersection of *four* emission sets, belonging to numbers $q_1, \ldots, q_4$ derived from the four columns.

4.2 Guaranteed Emission

If a digit k can be emitted from a matrix, then $|q_1 - q_2| \leq |q_1 - k| + |k - q_2| \leq 2$. Thus, emission is impossible if $|q_1 - q_2| > 2$. On the other hand, emission is surely possible if $|q_1 - q_2| \leq 1$. For a proof of this claim, several cases have to be distinguished. Recall $|q_1|, |q_2| \leq r$. Hence, if $q_1 > r-1$, then $q_1 \in [r-1, r]$, whence $q_2 \in [r-2, r]$, and thus, $k = r-1$ is a solution. The cases $q_2 > r-1$, $q_1 < -r+1$ and $q_2 < -r+1$ can be handled similarly. The remaining case is $|q_1|, |q_2| \leq r-1$. Then also $|q| \leq r-1$ holds for $q = \frac{1}{2}(q_1 + q_2)$. Let k be an integer with $|k| \leq r-1$ and $|k - q| \leq \frac{1}{2}$; such an integer always exists. Then we have $|k - q_1| \leq |k - q| + |q - q_1| \leq \frac{1}{2} + \frac{1}{2}|q_2 - q_1| \leq 1$, and analogously $|k - q_2| \leq 1$.

Lemma 4.1 *For a positive non-singular matrix* $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$*, let* $q_1 = r\frac{a-b}{a+b}$ *and* $q_2 = r\frac{c-d}{c+d}$*. If* $|q_1 - q_2| \le 1$*, then a digit in base* r *can be emitted. Conversely, if an* r*-digit can be emitted, then* $|q_1 - q_2| \le 2$ *must hold.*

4.3 The Shrink Factor of a Matrix

Because of its importance, we shall analyse the value of $|q_1 - q_2|$ in more detail.

$$\begin{aligned} |q_1 - q_2| &= r\left|\frac{a-b}{a+b} - \frac{c-d}{c+d}\right| \\ &= r\left|\frac{(ac+ad-bc-bd)-(ac-ad+bc-bd)}{(a+b)(c+d)}\right| \\ &= 2r\frac{|ad-bc|}{(a+b)(c+d)} \end{aligned} \tag{24}$$

Let us introduce a name for the fraction in (24).

Definition 4.2 The *shrink factor* $\operatorname{shr} M$ of a positive non-singular matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ is

$$\operatorname{shr} M = \frac{|\det M|}{(a+b)(c+d)} \quad .$$

The motivation for choosing the name shrink factor is as follows: A positive non-singular matrix $M = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ induces an LFT $\langle M \rangle : [0, \infty] \to [0, \infty]$. By composition with the bijections $\langle S_0 \rangle$ and $\langle S_\infty \rangle$ between $[0, \infty]$ and $[-1, 1]$, we obtain a function $f_M = \langle S_0 M S_\infty \rangle : [-1, 1] \to [-1, 1]$. Note that $f_M(-1) = \langle S_0 M \rangle(0) = \langle S_0 \rangle(c/d) = \frac{(c/d)-1}{(c/d)+1} = \frac{c-d}{c+d}$, and similarly, $f_M(1) = \frac{a-b}{a+b}$. Thus, the length of the interval $f_M[-1, 1]$ is $\left|\frac{a-b}{a+b} - \frac{c-d}{c+d}\right| = 2 \operatorname{shr} M$ by a computation as in the beginning of this subsection. Therefore, f_M lets the interval $[-1, 1]$ shrink by a factor of $\operatorname{shr} M$, namely from length 2 to length $2 \operatorname{shr} M$.

From Lemma 4.1 and (24), we immediately obtain:

Proposition 4.3 *Let* M *be a positive non-singular matrix. If* $\operatorname{shr} M \le \frac{1}{2r}$*, then a digit in base* r *can be emitted. Conversely, if an* r*-digit can be emitted, then* $\operatorname{shr} M \le \frac{1}{r}$ *must hold.*

4.4 Properties of the Shrink Factor

For any positive non-singular matrix M,

$$\operatorname{shr} M \le 1 \tag{25}$$

holds. For, $(a+b)(c+d) \ge ad + bc \ge |ad - bc|$ holds for $a, b, c, d \ge 0$.

Digit matrices D_k^r (4) have a shrink factor of $\frac{4r}{2r \cdot 2r} = \frac{1}{r}$. Though they are not positive, inverse digit matrices $(D_k^r)^*$ (5) can be assigned a formal shrink factor of $\frac{4r}{2 \cdot 2} = r$.

Now, let us consider emission. By (14), the column sums of $(D_k^r)^* \cdot M$ are twice the column sums of M. Furthermore, $\det((D_k^r)^* \cdot M)$ is $\det M$ times

$\det(D_k^r)^* = 4r$. Thus, $\mathrm{shr}\,((D_k^r)^* \cdot M)$ is $\mathrm{shr}\,M$ times $\frac{4r}{2\cdot 2} = r$, and we obtain

$$\mathrm{shr}\,((D_k^r)^* \cdot M) \;=\; r \cdot \mathrm{shr}\,M \;. \tag{26}$$

From this result, one may conjecture that $\mathrm{shr}\,(A \cdot B) = \mathrm{shr}\,A \cdot \mathrm{shr}\,B$ holds, "proved" by the argument that if f_B shrinks $[-1,1]$ by $\mathrm{shr}\,B$ and f_A shrinks $[-1,1]$ by $\mathrm{shr}\,A$, then $f_{A\cdot B} = f_A \circ f_B$ shrinks $[-1,1]$ by $\mathrm{shr}\,A \cdot \mathrm{shr}\,B$. Yet this argument fails because in $f_{A\cdot B}[-1,1] = f_A(f_B[-1,1])$, function f_A is not applied to $[-1,1]$, but to some subinterval, and $\mathrm{shr}\,A$ does not provide any information about the shrink factor of that subinterval.

In fact, there is no relationship between $\mathrm{shr}(A \cdot B)$ and $\mathrm{shr}\,A \cdot \mathrm{shr}\,B$ in general; the former value may be smaller or larger than the latter. The only general property for positive non-singular matrices is

$$\mathrm{shr}\,(A \cdot B) \leq \mathrm{shr}\,A \tag{27}$$

because $f_B[-1,1] \subseteq [-1,1]$ and therefore $f_{A\cdot B}[-1,1] = f_A(f_B[-1,1]) \subseteq f_A[-1,1]$.

Even if B is a digit matrix, there is no better information about $\mathrm{shr}\,(A\cdot B)$. Consider for instance $B = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, which is the digit matrix D_1^2 in lowest terms. For $A = \left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$, we have $\mathrm{shr}\,A = \frac{ad}{ad} = 1$ and $\mathrm{shr}\,(A \cdot B) = \mathrm{shr}\left(\begin{smallmatrix} 2a & a \\ 0 & d \end{smallmatrix}\right) = \frac{2ad}{2a(a+d)} = \frac{d}{a+d}$, which by suitable choices of integers $a, d > 0$ may yield any rational number between 0 and 1.

4.5 The Contractivity of a Matrix

By (27), we know that the shrink factor after an absorption is not larger than before, but as the example above shows, we cannot claim any substantial decrease of the shrink factor, and so cannot conclude that after many absorptions, an emission will eventually be possible. To obtain such a result, the shrink factor has to be replaced by another property of a matrix, the *contractivity*.

Let M be a positive non-singular matrix. Recall $\mathrm{shr}\,M = \frac{|f_M(1) - f_M(-1)|}{|1-(-1)|}$, where the denominator 2 is written in a particular complex way. The disadvantage of this definition is that it does not provide any information about the shrinking of subintervals of $[-1,1]$. To obtain such information, we consider the supremum of the shrink factors of all such subintervals and call it the contractivity $\mathrm{con}\,M$ of M:

$$\mathrm{con}\,M \;=\; \sup\left\{ \frac{|f_M(x) - f_M(y)|}{|x-y|} \;\middle|\; x, y \in [-1,1],\ x \neq y \right\} \;. \tag{28}$$

This notion (in fact, its reciprocal) was introduced in [7] to obtain convergence criteria for infinite matrix and tensor expressions. In that paper, a direct formula for the contractivity is derived:

$$M = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \;\Longrightarrow\; \mathrm{con}\,M = \frac{|\det M|}{(\min(a+b, c+d))^2} \;. \tag{29}$$

From both formulae, it is obvious that $\mathrm{shr}\,M \leq \mathrm{con}\,M$ holds. By combining this with Prop. 4.3, we see that an r-digit can be emitted if $\mathrm{con}\,M \leq \frac{1}{2r}$.

However, a necessary condition for emission in the sense of the last part of Prop. 4.3 is not possible using contractivities, and the sufficient condition with con is much weaker than that with shr. Consider for instance the matrices $M_{uv} = \left(\begin{smallmatrix} 1 & uv \\ 0 & u \end{smallmatrix}\right)$ with $u, v \geq 1$. We have $\operatorname{shr} M_{uv} = \frac{u}{1 \cdot (uv+u)} = \frac{1}{v+1}$, and so, r-emission is guaranteed for any r if v is sufficiently big. On the other hand, $\operatorname{con} M_{uv} = \frac{u}{1^2} = u$ is arbitrarily large and never satisfies $\operatorname{con} M_{uv} \leq \frac{1}{2r}$ for any $r \geq 2$.

Nevertheless, the contractivity is useful since it satisfies

$$\operatorname{con}(A \cdot B) \leq \operatorname{con} A \cdot \operatorname{con} B \quad . \tag{30}$$

For a proof, use (28), and note that for all x, y in $[-1, 1]$,

$$\frac{|f_{AB}(x) - f_{AB}(y)|}{|x-y|} = \frac{|f_A(f_B(x)) - f_A(f_B(y))|}{|f_B(x) - f_B(y)|} \cdot \frac{|f_B(x) - f_B(y)|}{|x-y|}$$

Thus, we obtain $\operatorname{con}(M \cdot D_k^r) \leq \operatorname{con} D_k^r \cdot \operatorname{con} M = \frac{1}{r} \operatorname{con} M$. For emission, $\operatorname{con}((D_k^r)^* \cdot M) = r \cdot \operatorname{con} M$ holds; the proof is analogous to that of the corresponding property of shr (26).

Let us summarise the results about the contractivity:

Proposition 4.4 *For every positive non-singular matrix M holds:*

(i) $0 \leq \operatorname{con} M < \infty$,

(ii) $\operatorname{con}(M \cdot D_k^r) \leq \frac{1}{r} \operatorname{con} M$,

(iii) $\operatorname{con}((D_k^r)^* \cdot M) = r \cdot \operatorname{con} M$,

(iv) *If* $\operatorname{con} M \leq \frac{1}{2r}$, *then an r-digit can be emitted.*

Given such a matrix M, a fixed number of absorptions suffices to obtain a matrix M' with $\operatorname{con} M' \leq \frac{1}{2}$. After the next absorption, we have a matrix M'' with $\operatorname{con} M'' \leq \frac{1}{2r}$. Then, an r-emission is possible, giving a matrix M''' with $\operatorname{con} M''' \leq \frac{1}{2}$. Repeating this, we see that from now on, at least one emission is possible after each absorption.

Theorem 4.5 *For every positive non-singular matrix M, there is a constant k such that at most $n + k$ digits must be absorbed to allow the emission of n digits.*

5 No Big Numbers in case of Zero Column Difference

The law of big numbers for matrices (Theorem 3.1) states that for non-singular matrices with non-zero column difference, the result after n transactions has at least one entry of bit size $\Omega(n)$, even if all possible reductions are performed. In this section, we prove a complementary result: for every positive non-singular matrix M with column difference zero and every argument x in $[0, \infty]$, the result of applying $\langle M \rangle$ to x can be computed with a bounded subset of the integers.

Consider a positive non-singular matrix $M_0 = \begin{pmatrix} a_0 & c_0 \\ b_0 & d_0 \end{pmatrix}$ with $\operatorname{cd} M_0 = 0$, i.e., $a_0+b_0 = c_0+d_0$. Comparing (29) and Definition 4.2, we see that $\mathrm{con}\, M_0 = \mathrm{shr}\, M_0$ holds. By (25), $\mathrm{con}\, M_0 \leq 1$ follows. For the following, recall Prop. 4.4. After one absorption, we get a matrix M_1 with $\mathrm{con}\, M_1 \leq \frac{1}{r} \leq \frac{1}{2}$. The next absorption yields a matrix M_2 with $\mathrm{con}\, M_2 \leq \frac{1}{2r}$. Then, an r-digit can be emitted, yielding M_3 with $\mathrm{con}\, M_3 \leq \frac{1}{2}$. After the next absorption, we have M_4 with $\mathrm{con}\, M_4 \leq \frac{1}{2r}$ etc. Thus, the computation sequence $A, A, E, A, E, A, E, \ldots$ is possible, where A means absorption and E emission. From the results in Section 3.1, we know that all the matrices M_i satisfy $\mathrm{cd}\, M_i = 0$. In the sequel, we assume that all matrices M_i are reduced to lowest terms. These reductions do not affect the arguments above since contractivity and zero column difference are invariant under reductions.

Consider the matrix M_{2n+1} after n emissions, i.e., $n+1$ digits have been absorbed and n digits have been emitted. According to (8) in Section 2.3, we may summarise all but the first absorbed r-digits into one r^n-digit K and all emitted r-digits into one r^n-digit K'. With $R = r^n$, the resulting matrix is $(D^R_{K'})^* \cdot M_1 \cdot D^R_K$ after the factor 2^{n-1} from (8) has been cancelled. Let us compute $M' = M_1 \cdot D^R_K$ first:

$$\begin{aligned}
&\begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix} \begin{pmatrix} R+K+1 & R+K-1 \\ R-K-1 & R-K+1 \end{pmatrix} \\
= &\begin{pmatrix} R(a_1+c_1)+(K+1)(a_1-c_1) & R(a_1+c_1)+(K-1)(a_1-c_1) \\ R(b_1+d_1)+(K+1)(b_1-d_1) & R(b_1+d_1)+(K-1)(b_1-d_1) \end{pmatrix} \\
=: &\begin{pmatrix} a' & c' \\ b' & d' \end{pmatrix} .
\end{aligned}$$

Now, we compute $M'' = (D^R_{K'})^* \cdot M'$:

$$\begin{aligned}
&\begin{pmatrix} 1-K'+R & 1-K'-R \\ 1+K'-R & 1+K'+R \end{pmatrix} \begin{pmatrix} a' & c' \\ b' & d' \end{pmatrix} \\
= &\begin{pmatrix} (1-K')(a'+b')+R(a'-b') & (1-K')(c'+d')+R(c'-d') \\ (1+K')(a'+b')-R(a'-b') & (1+K')(c'+d')-R(c'-d') \end{pmatrix} .
\end{aligned}$$

Note that $a_1+b_1-c_1-d_1 = 0$ implies $a'+b' = R(a_1+b_1+c_1+d_1) = 2R(a_1+b_1)$, and same with $c'+d'$. Modulo 2, the expressions $a_1-b_1+c_1-d_1$ and $a_1-b_1-c_1+d_1$ are equal to $a_1+b_1+c_1+d_1 = 2(a_1+b_1)$, and therefore, $a'-b'$ and $c'-d'$ are even. Thus, M'' is $2R$-reducible. After reduction, we obtain

$$\begin{pmatrix} (1-K')(a_1+b_1)+(a'-b')/2 & (1-K')(a_1+b_1)+(c'-d')/2 \\ (1+K')(a_1+b_1)-(a'-b')/2 & (1+K')(a_1+b_1)-(c'-d')/2 \end{pmatrix} .$$

Note that the sum of the four entries of this matrix is $4(a_1+b_1)$. Since the matrix is positive, this provides an upper bound for the entries. From this bound for the entries of M_{2n+1}, we easily get a bound for the entries of M_{2n+2}, and hence for the entries of all matrices.

6 The Complexity of LFT Application

The appearance of big integers affects the complexity of real number arithmetic. In this section, we study the time needed to compute n digits from the application of a matrix to a real number. Because of the law of big numbers, this time is $O(n^2)$ for matrices with non-zero column difference if the individual digits are handled one by one. By combining many digits in a small basis to one digit in a large basis, this quadratic complexity can be reduced to that of big integer multiplication. This kind of digit compression was already proposed by Peter Potts for absorptions, but not for emissions. Potts did not provide a complexity analysis.

6.1 Basic Assumptions

In computing a real number y, we are interested in the time $T(n)$ needed to compute the first n digits of y. By digits, we mean digit matrices, plus possibly a sign matrix in front. If y is not a constant, but depends on some input value x, i.e., $y = f(x)$, we consider a fixed argument x and assume that *all digits of x are already computed and freely available.* This assumption means that we do not directly take into account the difference between two algorithms for f, one of which computes n digits of y from n digits of x, while the other one needs n^2 digits of x. Yet this difference has an indirect impact on the complexity; for, the second algorithm presumably needs additional time to digest the additional digits.

By the law of big numbers, big integers cannot be avoided except in some exceptional cases. Hence, we need to consider the complexity of big integer operations.

- Addition, subtraction, and comparison of two integers of bit size n take time $O(n)$.
- Multiplication of an integer of bit size n with a 'small' integer such as 2 or 3 takes time $O(n)$, too.
- Multiplication of two integers of bit size n requires more than $O(n)$ basic arithmetical operations. Any straightforward algorithm takes time $O(n^2)$. However, there are several faster algorithms in [8], including one which needs only $O(n \log n \log\log n)$ basic arithmetical operations, and one that simulates the multiplication in $O(n)$ operations on a pointer machine.

  In the sequel, let us assume a fixed algorithm for multiplication with complexity $C(n)$ better than $O(n^2)$.
- Integer division of a $2n$ bit integer by an n bit integer, yielding an n bit integer, is as complex as n bit multiplication times a constant [8]. Thus, we assume a complexity of $C(n)$ for integer division, too.

6.2 Digit by Digit Evaluation

For a matrix M, we want to estimate the complexity $T(n)$ of computing the first n digits of $\langle M\rangle(x)$. The straightforward method to compute this result is *digit by digit evaluation*: emit digits as long as possible, then absorb one digit of x, again emit digits as long as possible etc.

Assume that after some of these transactions, we have obtained the matrix M' with bit size s in its entries. To compute the next transaction, one has to check whether emission is possible which involves the computation of $(D_k^r)^* \cdot M'$ for all possible k, and if no emission is possible, an absorption has to be done by computing $M' \cdot D_k^r$ for some digit matrix D_k^r. Since we assumed a small basis r, all the entries of D_k^r are small, and so, all calculations can be done in time $O(s)$.

Assume that the start matrix M has non-zero column difference. From Theorem 4.5, we know that $O(n)$ digits have to be absorbed to emit n digits. By Theorem 3.1, the matrix resulting after $O(n)$ absorptions and n emissions has an entry of bit size $O(n)$. Hence, the next transaction needs time $O(n)$, and so, the overall time for the absorption and emission of n digits is $O(n^2)$.

The situation is different for matrices with column difference zero. In Section 5, we have seen that the entries of all matrices occurring during the computation can be bounded by a fixed upper bound, i.e., have bit size $O(1)$. Hence, every transaction needs time $O(1)$ in this case, and so, the overall time for the computation of n digits is merely $O(n)$.

6.3 Mass Absorption

The quadratic complexity of digit by digit evaluation can be reduced by handling many digits at once. For the following, assume the basis $r = 2$.

Let M be a positive non-singular matrix with non-zero column difference and small entries, and consider the task of computing n digits of $\langle M\rangle(u)$ for an unsigned real u (a stream of digit matrices). By Theorem 4.5, we know that $m = n + k$ digits of u have to be absorbed into M in order to compute the desired number of digits of the result. Let these m digits be $D_{k_1}^2 \cdots D_{k_m}^2$ with $|k_i| \leq 1$.

We have already pointed out that the computation of the result takes time $O(n^2)$ if the digits are absorbed and emitted one by one. Now assume that we first perform all m absorptions and start emitting the n digits of the result only afterwards. Thus, the first subtask is to compute $M \cdot D_{k_1}^2 \cdots D_{k_m}^2$ efficiently. To be more precise, let $M_0 = M$ and for $0 < i \leq m$, let M_i be either $M_{i-1} \cdot D_{k_i}^2$ directly, or the result which is obtained after all possible reductions have been performed on this matrix. In any case, the entries of M_{i-1} have bit size $O(i)$, and so it takes time $O(i)$ to compute $M_{i-1} \cdot D_{k_i}^2$. Therefore, the overall time to compute $M_1, M_2, \ldots, M_m$ is quadratic.

There is another way to obtain M_m. By (8),

$$D^2_{k_1} \cdots D^2_{k_m} = 2^{m-1} D^{2^m}_K = 2^{m-1} \begin{pmatrix} 2^m + K + 1 & 2^m + K - 1 \\ 2^m - K - 1 & 2^m - K + 1 \end{pmatrix}$$

where $K = \sum_{i=1}^m k_i r^{n-i}$. The factor 2^{m-1} can be cancelled immediately. The number K can be computed in time $O(m)$ as follows: collect the positive digits k_i into a number K^+, the negative digits into K^-, and compute $K = K^+ - K^-$. Since K has bit size $O(m)$, the computation of $D^{2^m}_K$ from K takes time $O(m)$, and so does the multiplication $M \cdot D^{2^m}_K$ since the entries of M are small.

- m digits can be absorbed into a matrix in time $O(m)$.

So far, we have not made any assumption about the concrete representation of big integers. A particularly fascinating approach is a representation in a redundant binary way, i.e., as $\sum_{i=1}^m k_i 2^i$, with $k_i \in \{-1, 0, 1\}$. Then the digit sequence $D^2_{k_1} \cdots D^2_{k_m}$ is in some sense identical with K (up to a possible reversal). Or put the other way round: instead of storing $D^2_{k_1}, \ldots, D^2_{k_m}$ as list of matrices or as list $[k_1, \ldots, k_m]$ of digits, store it as the number K together with the length m. This brings the digit matrix approach in close relationship with Boehm and Cartwright's functional approach [2,3].

6.4 Mass Emission

Mass absorption may bring down the cost of real number computation, but for a real gain, also mass emission is needed. For, no matter how quick $M' = M \cdot D^2_{k_1} \cdots D^2_{k_m}$ can be computed, the result has entries of bit size $\Omega(m) = \Omega(n)$, and so it takes time $\Omega(n^2)$ to emit n digits from M' if the digits are emitted one by one.

The solution is of course to emit one digit $D^{2^n}_K$ in base 2^n, and then split K into 2-digits $k_1, \ldots, k_n$. As we have seen in Section 4.1, the emission of a 2^n-digit from a matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ involves the approximation of $\frac{2^n(a-b)}{a+b}$ and $\frac{2^n(c-d)}{c+d}$ by integer divisions. Since these are divisions of a $n + O(n)$ bit integer by an $O(n)$ bit integer, they can be performed in time $C(n)$. If required at all, the number K can be split into $k_1, \ldots\ k_n$ by simply reading off the bits of K.

- By mass absorption and mass emission, the time needed to compute n digits of $\langle M \rangle(u)$ can be reduced from $O(n^2)$ to $C(n)$, the time needed for the multiplication of two integers of bit size n.

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] 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`.

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

[6] R. Heckmann. The appearance of big integers in exact real arithmetic based on linear fractional transformations. In *Proc. Foundations of Software Science and Computation Structures (FoSSaCS '98)*, volume 1378 of *LNCS*, pages 172–188. Springer-Verlag, 1998.

[7] R. Heckmann. Contractivity of linear fractional transformations. In J.-M. Chesneaux, F. Jézéquel, J.-L. Lamotte, and J. Vignes, editors, *Third Real Numbers and Computers Conference (RNC3)*, pages 45–59, April 1998.

[8] D. E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, 2nd edition, 1981.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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