May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

Computational Aspects of Various Number Representations

Nikolai N. Nepejvoda[1], Ivan N. Grigorevsky[1], Andrei V. Klimov[1,2], Yuri A. Klimov[1,2], Sergei A. Romanenko[1,2]
[1]Ailamazyan Program Systems Institute of Russian Academy of Sciences, Pereslavl-Zalessky, Russia.
[2]Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia.

Article Info
Volume 83
Page Number: 28224 - 28231
Publication Issue:
May-June 2020

Article History
Article Received: *11 May 2020*
Revised: *19 May 2020*
Accepted: *29 May 2020*
Publication: *12 June 2020*

Abstract

Modern mathematical modeling problems, which exploit the power of supercomputers to the utmost extent, require increasing digit capacity of real numbers. However, the traditional representation of the reals is not efficient enough in the case of large bit sizes due to carry propagation. To make the arithmetic operations digit-wise parallel, several alternative numeration systems has been proposed and intensively studied during the last decades, starting with the seminal result by L. E. J. Brouwer a century ago. This paper continues this line of research and reveals some computational aspects of various number representation systems: exponential, recurrent (or linear), aliquot, overlaying exponential. Several classic results are generalized to broader classes of numeration systems. In particular, the representability theorem by A. Rényi for exponential systems is generalized to include convergent and nonconvergent numeration systems, uniform and non-uniform bases, and arbitrary sets of digits. As a contribution for practical use, a symmetric optimal positional overlay system is proposed, and an addition algorithm is described.

Keywords: *number representation, numeration systems, digit-wise parallel operations, effective computability.*

I. Introduction

While computers and supercomputers are becoming more powerful and mathematical modeling requirements are increasing, the usual precision of real numbers at 32, 64, and even 128 bits feels insufficient. Sometimes digit capacity of up to thousands of bits is needed. Moreover, some mathematical problems require varying precision, which changes during calculations depending on the accuracy achieved.

This raises the problem of efficiently implementing arithmetic operations with digit-wise parallelism. However, the conventional positional number systems have a serious drawback. Because of carry propagation, possibly along all digits, even a parallel digit-wise implementation of addition has the worst time proportional to the accuracy of the arguments.

This is not just a practical problem. It arose from deep theoretical questions. As L. E. J. Brouwer discovered a century ago [4], the carry prohibits abstract computability of the operations on the real numbers (which we discuss and generalize in Section 5 below). He invented an overlaying numeration system that is free from this disadvantage due to redundancy of representation. This result was not used in practice and remained unexplored and undeveloped for decades.

May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

Nowadays there is a stream (not too large, but somewhat long and deep) of works studying algorithmic and algebraic aspects of various numeration systems (see [2, 9, 11, 18] and many in-between). In this paper we focus on connections and interrelations of ‘almost standard’ commonly studied systems, as well as some more exotic ones, and generalize the existing results with practical goals in mind. In particular, we compare exponential systems with the standard set of digits with additive systems where the sequence of bases is less regular, and with systems where the digits represent overlapping intervals rather than single values.

The main contributions of this paper are as follows:

1. Generalization of the representability theorem for non-standard positional numeration systems to include arbitrary number bases and arbitrary sets of digits (Section III).
2. Generalization of the abstract computability theorem and its extension to aliquot number systems (Section V).
3. Development of a symmetric optimal overlaying positional system along with the additive operation for it (Section VI).
4. Further development of the notion of a recurrent numeration system that was introduced in our earlier paper [14] (Section VII).

II. Main definitions

Definition 2.1. By an *additive numeration system* we mean a sequence of bases b_i, $-\infty < i < +\infty$, $0 < |b_i| < |b_{i+1}|$, and a finite set of *digits* including 0. Each digits d has its *value* $V(d)$, which is usually represented by its denotation (say $2, -2, 3/7, \pi$). We require that the value of digit 0 is 0 and $b_0 = 1$. The integer i is called the *position* of d_i and b_i.

A representation of a real number is a sequence of digits where all digits with sufficiently large positions are 0. The maximal index of a non-zero digit is called the *order* of the representation. A representation is *finite* or *precise* if all i with sufficiently large negative i are also 0. The minimal index $-i$ of a non-zero digit is called the *precision* of a representation, and this representation is called the i-th approximation.

A number x is represented in an additive system if there exist an integer N and a sequence of digits d_i such that $d_i = 0$ for all $i > N$ and

$$\sum_{-\infty}^{+\infty} b_i \cdot d_i = x.$$

We say that an additive system is *convergent* if for each $x > 0$ there is a minimal index i such that for each real y such that $|y| > x$, y requires at least the order i to be represented.

To the best of our knowledge, all the systems studied in recent publications are convergent.

Comments and examples on Definition 2.1. All traditional exponential systems are additive. Their bases are α^i where $|\alpha| > 1$; their digits are usually subsequent integers from $-l$ to u. They are studied in dozens of papers, e.g., [7, 9, 16].

Simple examples of ‘slightly non-standard’ exponential systems are systems of Frougny et al. [1, 7]. They proved that a rational base is in almost all aspects worse than many algebraic irrational bases. For example, natural numbers other than 0 and 1 are not representable in such systems. But if we have a system with a rational base $p/q, p > q$, and digits

$$0, \frac{1}{q}, \frac{2}{q}, \cdots, \frac{p-1}{q}$$

then we can represent integer numbers in finite form and operations become a bit easier.

A more general class of additive systems are *recurrent* systems where

$$b_n = a_1 \cdot b_{n-1} + \cdots + a_k \cdot b_{n-k}.$$

In particular, the famous Fibonacci and n-bonacci systems [17] are recurrent ones. Such systems are called *linear* in works [8, 19]. (This name seems somewhat misleading as *exponential* systems are a simple special case of *linear* ones: $b_n = \alpha \cdot b_{n-1}$.)

An example of a more exotic but interesting in some aspects additive system that is not linear, is the *aliquot* (or *Egyptian*) system (for Egyptian fractions see [5]) where $b_{-i} = 1/i$ and digits are 0,1. It is easily extended to a system for all (not only positive) real

May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

numbers by putting, e.g., $b_{-i} = (-1)^i/i$. Here we see that the sequence of numbers is not always correct, e.g., the sequence $\cdots,1,\cdots,1,1,1,0,0,\cdots,0,\cdots$ represents nothing. Thus this system is not convergent.

III. Representability theorem

The following theorem generalizes to arbitrary additive systems the fundamental result by A. Rényi [18] for exponential systems. Its restriction to exponential systems also generalizes the Rényi's theorem.

The n-*th span* is the segment

$$\left[\min_j d_j \cdot b_n, \max_j d_j \cdot b_n\right].$$

Its length is denoted by l_n. An (n,j)-*th gap* $g_{n,j} = \min_i |d_j - d_i| \cdot b_n$.

Theorem 1 (Representability). If

1. $l_n \geqslant b_{n+1} - b_n$, for all n, and
2. $g_{n,j} \leqslant b_n$, for all n, j,

then all $x \geqslant 0$ are representable in our system. If additionally $\min_j d_j \leqslant -1$, then all real x are representable.

Proof. Step 1 ('Integer' part): For positive numbers, find the maximal $\sum_{i=0}^n d_i \cdot b_i \leqslant x$. For all real numbers, find the sum such that $|\sum_{i=0}^n d_i \cdot b_i - x|$ is minimal. In both cases, the search is finite.

Step 2 ('Fractional' part): Let we have the i-th approximation $\sum_{-i}^n d_i \cdot b_i$. Recursively find the next digit d_{-i-1} by minimizing the distance of the partial representation from x. For positive numbers, find the miminal non-negative $x - \sum_{-i-1}^n d_i \cdot b_i$. For all real numbers, minimize $|x - \sum_{-i-1}^n d_i \cdot b_i|$.

Thus, we have an algorithm for sequentially getting the digits of the representation. □

IV. Overlaying

The real numbers are inherently a completely different data type rather than integers or algebraic numbers. They are imprecise. They represent, from the physical and other natural sciences point of view, an infinite sequence of more and more precise measurements. So in practice there is no such real number as 2 (even the power in the Newton's law of gravitation is $2 \pm \varepsilon$).

Algorithmically and constructively the real numbers are infinite effective sequences of embedded intervals

$$[a_{n+1}, b_{n+1}] \subset [a_n, b_n], \lim_{n\to\infty}(b_n - a_n) = 0.$$

Interval representations of real numbers have recently been studied in [13, 20]. Here we consider a certain class of such representations and compare operations on them with operations on the conventional representations.

Definition 2.2 (Exponential interval representations of real numbers with overlaying). Consider natural numbers with a *base* $\nu > 1$, and an *overlay* ε, $0 \leqslant \varepsilon < 1$. A *system with overlays* (ν, ε) on a segment $[low, high]$ is a system with ν digits defined as intervals produced by the equations below. Let ξ be the length of our desired interval. Denote by δ the shift of a digit relatively to the previous one. Then we have

$$\delta = \xi - \epsilon;\ \epsilon = \xi \cdot \varepsilon;\ \nu \cdot \delta + \xi = 1. \qquad (1)$$

$$\xi = \frac{1}{(\nu-1)\cdot(1-\varepsilon)+1};\ l = \text{high} - \text{low}. \qquad (2)$$

The digits of this representation are as follows (Figure 1):

$$[0, \xi] = [0 \cdot \delta, 0 \cdot \delta + \xi],$$
$$[1 \cdot \delta, 1 \cdot \delta + \xi],$$
$$\cdots$$
$$[(\nu - 1) \cdot \delta, (\nu - 1) \cdot \delta + \xi] = [1 - \xi, 1].$$

The digits of the next position are descried by the similar fragmentation of each of the segments:

$$[\text{low} + i \cdot \delta \cdot l, \text{low} + (i \cdot \delta + \xi) \cdot l].$$

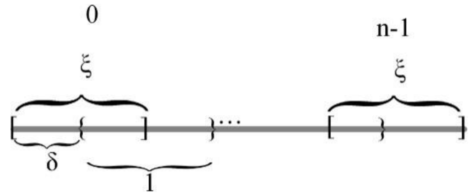

Figure 1: Breaking into digits

May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

Such a system was first proposed by L. E. J. Brouwer [4]. In his system ⟨0,1,2,0.5⟩, the segment [0,1] is divided into the overlaying segments [0,2/3], [1/3,1], and so on. For example, [0,2/3] is divided into [0,4/9], [2/9,2/3]. The conventional systems are a special class of overlaying with the zero overlay. For example, if we add extra digits to these systems (e.g., digit -1 to the binary system obtaining the system with 3 digits $0,1,-1$), they become similar to overlaying ones but with non-uniform partial overlays.

V. Abstract computability

The first question for a numeration system is which operations are computable in the system. Further questions concern the complexity of computation.

To consider abstract computation issues uniformly and independently from a specific notion of an algorithm, we use the term *effective operations* to denote any class of operations conforming with *Brouwer's principle of finite information* (in the special case of sequences, it is called *Brouwer's continuity principle*; we remove unnecessary binding a specific data structure and rectify the idea of Brouwer):

> *Finite information about the result is computed requiring only finite information about the arguments.*

and extending it with the following requirement of modern computer programming and science:

> *External procedures are used as closed modules whose internal structure is inaccessible to the caller.*

There is a set of results that gradually became more and more general from 1920 to 2015 (at least).

Here we consider the weakest requirement that a representation allows for effective computability of some operations.

We say that a system is *non-redundant* if for all n, j we have

$$l_n + g_{n,j} < b_{n+1} + g_{n,0}.$$

Theorem 2 (Brouwer, 1921 [4], Banach, Mazur, 1937 [3], Uspensky, 1960 [22], 2014 [15], revised and generalized here).

1. There is an effective transformation (Figure 2) of every convergent positional system with an effective sequence of bases and rational digits into the general constructive representation of the real numbers (a sequence of embedded segments).
2. There is no effective transformation of the general representation into any convergent positional system.
3. There is no algorithm of multiplication and division in any convergent positional system and no algorithm of addition for any non-redundant system.
4. For any effective function of real numbers there exists an algorithm to transform an arbitrary aliquot representation of arguments into an aliquot representation of the result.

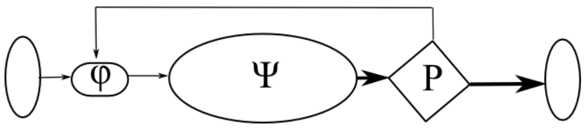

Figure 2: Effective transformation

Sketch of proof.

1. According to the definition of a convergent system, for each $\varepsilon > 0$ we can get a segment of length $\leqslant \varepsilon$ included into the previous segments.
2. If the segments are located near the place where one digit replaces another, we cannot determine this digit in general case. To show this, we can use the so called *method of provocation* (Figure 3): wait until the procedure to compute a representation produces a digit and change the next approximation so that this digit is wrong. (A special case of this method is known as *diagonalization*. We prefer the more general term *provocation* from [12, 13], which hides the details of specific encoding and representation and emphasizes the main idea.)

TEST Engineering & Management
May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

3. Analogously by the method of provocation.
4. According to the results by W. Sierpinski [21], there is an algorithm for primitive recursive computation of the sum of various $1/n$, $n > k$, to obtain any rational number p/q (Egyptian fraction). So after getting a new segment, we represent its lower bound extending the previous representation. □

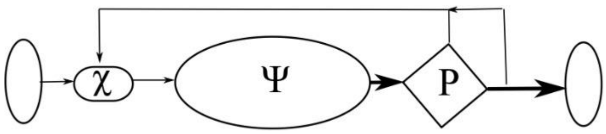

Fig 3. The method of provocation

Theorem 3 (revised and improved form of the theorem from [15]). In any overlaying system with overlay > 0, each effective function of real numbers is effectively computable.

Sketch of proof. At each step, take the digit whose segment is greater than the segment of the standard result. To provide this, it suffices to wait until the length or partial result of the segment will be less than the current overlay. □

Thus, overlaying systems have the advantage of aliquot systems without their disadvantages: the astronomically large length of representations for numbers x with $|x| > 100$ and the existence of meaningless representations.

VI. Optimal overlaying system for parallel addition

Theorem 4. In the overlaying system with 3 digits and overlay $\varepsilon = 0.5$ (which we refer to as the *three-halves system*), we can compute $\frac{x+y}{2}$ completely in parallel with memory 0 and anticipation 2 (the terminology from, e.g., [9]).

Proof. For brevity, let the segment $[0,1/2]$ be denoted by 0, $[1/2,1]$ by 1, $[1/4,3/4]$ by $*$. Table 1 makes our statement clear. The current digit is the first, the lower digit is the second. Note that $01 = {*}0$, ${*}1 = 10$. The notation $\begin{smallmatrix}0\\ *\end{smallmatrix}, \begin{smallmatrix}*\\ 1\end{smallmatrix}, \begin{smallmatrix}\bar{1}\\ 0\end{smallmatrix}$ represents that both results are correct. We can optimize our algorithm taking the more convenient one. □

+	00	0*	01	**	*1	1*	11
00	0	0	0	0	0 *	*	*
0*	0	0	0	0 *	*	*	*
01	0	0	0 *	*	*	*	* 1
**	0	0 *	*	*	*	* 1	1
*1	0 *	*	*	*	* 1	1	1
1*	*	*	*	* 1	1	1	1
11	*	*	* 1	1	1	1	1

Table 1: The addition table for the three-halves system

The three-halves system has some shortcomings. First, it represents only positive numbers. If we try to represent negative numbers by the common way: sign $-$ before a denotation, then our number space becomes $(-\infty, 0] \cup [0, +\infty)$ which is not isomorphic constructively to real numbers $\mathbb{R}$. Consequently, the operation $+$ is not computable for this representation: we cannot compute the sign of $-0.0000\ldots$ and $+0.0000\ldots$ based on finite information about arguments because some of the next digits can ruin the partial result. Secondly (A. Shvorin, private communication), although $x - y$ is algorithmically computable when $x > y$, there is no automaton for this computation as it cannot be local.

These shortcomings are overcome in a symmetric three-halves system (proposed by A. Shvorin). Here we have 3 digits:

$[-1,0]$ denoted by $\bar{1}$,

$[-0.5, 0.5]$ by 0,

$[0,1]$ by 1.

May-June 2020
ISSN: 0193-4120 Page No. 28224 - 28231

Theorem 5. In the *symmetric* overlaying system with 3 digits and overlay $\varepsilon = 0.5$, we can compute $\frac{x+y}{2}$ and $\frac{x-y}{2}$ completely in parallel with memory 0 and anticipation 2.

Table 2 gives the method of addition. The method of subtraction is easily derived by inverting the second argument. Here $\bar{1}1 = 0\bar{1}$, $1\bar{1} = 01$.

+	$\overline{11}$	$\bar{1}0$	$\bar{1}1$	00	01	10	11
$\overline{11}$	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$ 0	0	0
$\bar{1}0$	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$ 0	0	0	0
$\bar{1}1$	$\bar{1}$	$\bar{1}$	$\bar{1}$ 0	0	0	0	0 1
0 0	$\bar{1}$	$\bar{1}$ 0	0	0	0	0 1	1
0 1	$\bar{1}$ 0	0	0	0	0 1	1	1
1 0	0	0	0	0 1	1	1	1
1 1	0	0	0 1	1	1	1	1

Table 2: The addition table for the symmetric three-halves system

The symmetric system has one more pleasant property. One of the representations of a real number can be easily obtained from its usual binary representation:

1. If a number does not start with $[-]0$, add a leading zero.
2. Replace the leading zero by 01 for the positive number or by $0\bar{1}$ for the negative one.
3. For the positive number replace each other 0 with $\bar{1}$.
4. For the negative number replace each 1 with $\bar{1}$, each other 0 with 1.

Example. The number 0.1011 is transformed into $01.1\bar{1}11\overline{111}\ldots$ The number -101 is transformed to $0\bar{1}1\bar{1}.1111\ldots$ All representations are infinite because we have no precise numbers.

VII. First steps to recurrent systems

Consider recurrent systems for the integer numbers with

$$b_n = a_1 \cdot b_{n-1} + \cdots + a_k \cdot b_{n-k}$$

and standard digits $0,1,\cdots,a_1+1$ where for all i $a_i \geqslant a_{i+1}$, $b_0 = 1$, $b_{-n} = 0$ for all n. For such a system, there exists an algorithm of addition with memory 1 and anticipation k, and this result cannot be improved by any number of extra digits (generalization of the result from [15]).

But expanding a recurrent system to the real numbers is a hard problem. Recall [6] that each linear recurrent sequence is representable in the form

$$\sum_{i=1}^{k} q_i \cdot r_i^k$$

where r_i are the roots of the characteristic equation

$$x^k - x^{k-1} \cdot a_1 - \cdots - a_k = 0.$$

Thus, a recurrent representation can be extended to negative indices if all non-zero q_i are positive and correspond to positive roots greater than 1. For such systems, the representability and abstract computability theorems hold if

$$\sum_{q_i \neq 0} r_i \leqslant \max_j c_j - \min_j c_j.$$

These systems are convergent.

For example, recurrent sequence $b_n = 5b_{n-1} - 6b_{n-2}$ generates a recurrent representation system

$$b_n = \frac{1}{2} \cdot (2^n + 3^n)$$

and digits $0,1,-1,2,-2$.

An open problem. Are there any good arithmetic algorithms for recurrent representation systems for the reals other than exponential ones?

Nevertheless, recurrent systems can be used, and are used, for fixed point numbers (e.g., the Fibonacci system).

VIII. Conclusion

Future post-silicon computers require new methods of programming and new representation of data. In this regard, it is necessary to study not only specific cases of number representation but their entire zoo. Recent investigations gave a lot of isolated results, and now it is time to combine them together, since the new computing will require different structures for processing by different processors (say, optic or biological).

In this paper we presented the following main results:

1. We generalized the representability theorem by A. Rényi [18] for exponential systems to include convergent and nonconvergent, uniform and non-uniform bases, and arbitrary sets of digits. Even the restriction of our theorem to exponential systems gives a more general result than the Rényi's theorem.
2. We also generalized and extended the abstract computability theorem to aliquot numeration systems.
3. We described a novel algorithm of additive operations for symmetric optimal overlaying system.
4. We outlined first steps to efficient algorithms for the recurrent numeration systems that we introduced in an earlier publication [14]. We also pointed to the open problem of construction of good recurrent representation systems for the real numbers.

Acknowledgments. We are particularly grateful to Artem Shvorin for his valuable ideas, suggestions, and comments on this paper. Our special thanks are due to Mark Hatkevich for developing computer models of overlaying and redundant systems (previous versions of them are published in [10]).

This research is financially supported by the Russian Federation represented by Ministry of Education and Science (grant id: RFMEFI61319X0092).

References

[1] Akiyama S, Frougny C, Sakarovitch J (2008) Powers of Rationals Modulo 1 and Rational Base Number Systems. Isr. J. Math. 168.1:53–91. DOI: 10.1007/s11856-008-1056-4

[2] Avizienis A (1961) Signed-Digit Number Representations for Fast Parallel Arithmetic. IRE Trans. Electronic Comp. EC-10.3:381–400. DOI: 10.1109/TEC.1961.5219227

[3] Banach S, Mazur S (1937) Sur les fonctions calculables. Ann. Soc. Pol. de Math. 16.3–4:223–252 (in French)

[4] Brouwer LEJ (1921) Besitzt jede reele Zahl eine dezimalbruchentwicklung? Mathematische Annalen 83.3–4:201–210

[5] Creyaufmüller W (2016) Aliquot Sequences. URL: http://www.aliquot.de/aliquote.htm

[6] Cull P, Flahive M, Robson R (2005) Difference Equations: From Rabbits to Chaos. Springer. DOI: 10.1007/0-387-27645-9

[7] Frougny C, Pelantová E, Svobodová M (2013) Minimal Digit Sets for Parallel Addition in Non-Standard Numeration Systems. Journal of Integer Sequences 16.Article 13.2.17. URL: https://cs.uwaterloo.ca/journals/JIS/VOL16/Frougny/frougny3.html

[8] Frougny C, Solomyak B (1996) On Representation of Integers in Linear Numeration Systems. London Mathematical Society Lecture Note Series 228:345–368. URL: https://www.cambridge.org/gb/academic/subjects/mathematics/differential-and-integral-equations-dynamical-systems-and-co/ergodic-theory-and-zd-actions

[9] Frougny C et al (2019) On-line Algorithms for Multiplication and Division in Real and Complex Numeration Systems. Discret. Math. Theor. Comput. Sci. 21.3. URL: http://dmtcs.episciences.org/5569

[10] Khatkevich M (2016) Direct-Flow Adder Program Simulation. Program Systems: Theory and Applications 7.4(31):359–367. DOI: 10.25209/2079-3316-2016-7-4-359-367 (in Russian)

[11] Knuth DE (1960) An Imaginary Number System. Commun. ACM 3.4:245–247. DOI: 10.1145/367177.367233

[12] Nepeivoda NN (2011) Constructive Mathematics: Review of Progress, Lacks and Lessons. Part I. Logical Investigations 17:192–239 (in Russian)

[13] Nepejvoda NN (2019) Applied Logic. 3rd ed. Direct-Media, Moscow–Berlin. ISBN: 978-5-4499-0126-2. DOI: 10.23681/561272

[14] Nepejvoda NN (2019) Some Issues on Linear Numeration Systems. IOSR Journal of Mathematics (IOSR-JM) 15.6:47–50. URL: http://www.iosrjournals.org/iosr-jm/papers/Vol15-issue6/Series-4/E1506044750.pdf

[15] Nepejvoda NN, Grigorevsky IN, Lilitko EP (2014) On Representation of Real Numbers. Program Systems: Theory and Applications 5.4(22):105–121. URL: http://psta.psiras.ru/read/psta2014_4_105-121.pdf (in Russian)

[16] Parhami B (1993) On the Implementation of Arithmetic Support Functions for Generalized Signed-Digit Number Systems. IEEE Trans. Computers 42.3:379–384. DOI: 10.1109/12.210182

[17] Pisano L (2002) Fibonacci's Liber Abaci: A Translation into Modern English of the Book of Calculation. Ed. by Laurence Sigler. Springer. DOI: 10.1007/978-1-4613-0079-3

[18] Rényi A (1957) Representations for Real Numbers and Their Ergodic Properties. Acta Math. Acad. Sci. Hungary 8:477–493. DOI: 10.1007/BF02020331

[19] Shallit JO (1994) Numeration Systems, Linear Recurrences, and Regular Sets. Inf. Comput. 113.2:331–347. DOI: 10.1006/inco.1994.1076

[20] Shvorin AB (2015) Parallel Addition of Real Numbers in Overlaying Numeration Systems. Program Systems: Theory and Applications 6.2(25):101–117. DOI: 10.25209/2079-3316-2015-6-2-101-117 (in Russian)

[21] Sierpiński W (1956) Sur les décompositiones de nombres rationells en fractions primaires. Mathesis 65:16–32 (in French)

[22] Uspensky VA (1960) Lectures on Computable Functions. Nauka, Moscow (in Russian)