MATEMATIČKI VESNIK
МАТЕМАТИЧКИ ВЕСНИК
Corrected proof
Available online 09.04.2021

research paper
оригинални научни рад

DYADIC FLOATING POINT

Miloš Milovanović

Abstract. The paper is aimed to elaborate the floating point multiresolution, considering convergence that allows some more fractions than otherwise. It implies a calculation concerning infinite strings of digits, which is not implementable in the standard representation, but requires a dyadic one. Such a view is much more convenient for regarding convergence because of specific norm whose logarithm follows the multiresolution scale. Arithmetic operations are performed in almost the same manner as the standard floating point method. Conversions from one representation to another are discussed in details. The main advantage of the method concerns an opportunity of representing constructible angles in the Euclidean plane, which is significant inter alia for computational geometry. A basic application also concerns two's complement representation of negative numbers, which is accurate only if one implies convergence in regard to the norm. In that respect, it offers a consistent realization of methods the computer science already provides.

1. Introduction

Floating point is a method for representing rational numbers in the computer science. It appears in a few variants from the single precision to the double one, implying usually the 2-based representation. The exponential notation tends to reflect the *multiresolution* of real numbers, which is a term that has been intimately related to the wavelet analysis [6]. It was defined by Stéphane Mallat and Yves Meyer to typify decomposition of the functional space $L^2(\mathbb{R})$ in an ascending sequence of subspaces having trivial intersection and complete closed union. Concerning the real numbers, multiresolution applies to the underlying domain of the Lebesgue space.

However, mapping the continuum onto a discrete structure has not been done without any trouble and there are some famous bugs [7]. The fractions representable in such a manner are only the ones whose denominators are powers of two, whilst those like $\frac{1}{3}, \frac{1}{5}, \ldots$ are reduced to the approximate values having maximal denominator. The paper is aimed to elaborate the floating point multiresolution, considering

2020 Mathematics Subject Classification: 68U01, 68M07, 65D18.
Keywords and phrases: Dyadic numbers; multiresolution; computational geometry.

convergence in a norm. Although the method is still about rational numbers, the convergence supports representing more fractions than otherwise. It implies a calculation concerning infinite strings of digits, that is not implementable in the standard but requires a dyadic representation. The exponential notation in that regard should not reflect structure of the real numbers, but the dyadic ones that also include the subset of rationals.

The theory of dyadic numbers is presented in Appendix, elaborating their history and applications. Some concepts are also mentioned immediately along the text.

2. The floating point of reals

Under the terms floating point, one implies notation $\pm a \times 2^i$, whereby a is the mantissa and i the integer exponent ranging from e_{min} to e_{max}. The mantissa is presented in the form $a = .a_1 a_2 \dots a_d$ of d binary digits $a_k = \begin{cases} 0 \\ 1 \end{cases}$, each of them having a position value 2^{-k}. For the uniqueness of representation, the first bit is regarded to be the unit $a_1 = 1$. The author considers the length d to be a power of two, and it is quite convenient to assume $d = 2^5$. According to that, a is specified up to the resolution 2^{-32}.

The method reflects the multiresolution of real numbers $\mathbb{R}$, that is a sequence of the approximate subsets $\mathcal{A}_i$ satisfying axioms:

$$\begin{gathered} \mathcal{A}_i \supset \mathcal{A}_{i+1}; \qquad x \in \mathcal{A}_i \Leftrightarrow 2x \in \mathcal{A}_{i+1}; \qquad \cap_{-\infty}^{+\infty} \mathcal{A}_i = \{0\}; \\ \overline{\cup_{-\infty}^{+\infty} \mathcal{A}_i} = \mathbb{R}, \text{ implying closure in regard to the standard norm;} \qquad (1) \\ \mathcal{A}_0 = \mathbb{Z} \text{ is the basic subset consisted of integers.} \end{gathered}$$

Defining a sequence of the detail subsets to be $\mathcal{D}_i = \mathcal{A}_i \setminus \mathcal{A}_{i+1}$, one gets an alternative axiomatization of the same structure:

$$\begin{gathered} \mathcal{D}_i \text{ are mutually exclusive;} \qquad x \in \mathcal{D}_i \Leftrightarrow 2x \in \mathcal{D}_{i+1}; \\ \overline{\cup_{-\infty}^{+\infty} \mathcal{D}_i} = \mathbb{R} \text{ implying closure in regard to the standard norm;} \qquad (2) \\ \mathcal{D}_0 = \mathbb{Z} + \frac{1}{2} \text{ is the basic subset consisted of semi-integers.} \qquad (3) \end{gathered}$$

In the floating point, however, (3) should be replaced by the set containing 32-bit strings from the binary point on. Also, in the axiom (2) there is no need for closure since the union is finite and therefore the closed set. According to that, the floating point $\mathbb{F}$ satisfies axioms:

$$\begin{gathered} \mathcal{D}_i \text{ are mutually exclusive;} \\ x \in \mathcal{D}_i \Leftrightarrow 2x \in \mathcal{D}_{i+1}; \qquad \cup_{e_{min}}^{e_{max}} \mathcal{D}_i = \mathbb{F}; \\ \mathcal{D}_0 = \{\pm .1 a_2 \dots a_d\} \text{ is the basic subset} \qquad (4) \end{gathered}$$

consisted of numbers up to the resolution 2^{-32} between $\pm\frac{1}{2}$ and ± 1.

Although resembling to the real numbers multiresolution, the main distinction

concerns absence of closure in regard to the norm. In that respect $\mathbb{F}$ is reduced to a discrete lattice consisting of the power-of-two denominators only, which should be avoided through an implementation of the limit process in the axiom (4):

$$\mathcal{D}_0 = \left\{ \pm.\overrightarrow{1a_2 \dots a_d} \right\} \text{ is the basic subset} \tag{5}$$

consisted of the 32-bit strings repeating periodically rightwards from the binary point on.

Implying $m = 1a_2 \dots a_d$, it follows $.\overrightarrow{m} = m \times 2^{-d} + \dots + m \times 2^{-2d} + \dots = \frac{m \times 2^{-d}}{1-2^{-d}} = \frac{m}{2^d-1}$. The deviation from $.m = \frac{m}{2^d}$ is therefore $.\overrightarrow{m} - .m = \frac{m}{2^d(2^d-1)} \leq 2^{-d}$, whereby the equality is reached for $.\overrightarrow{m} = .\overrightarrow{11 \dots 1} = 1$. Hence, one states $.\overrightarrow{m} \approx .m$ assuming an adequacy up to the resolution 2^{-d}.

However, a usual calculation in the floating point is disabled because of the limit process (5). It is troubled by the infinite string of binary digits tending rightwards, which makes the notation virtually unusable. Fortunately, there is a manner to avoid the problem considering the leftward periodization $.\overleftarrow{m} = m \times 1 + m \times 2^d + \dots = \frac{m}{1-2^d}$ that implies convergence in regard to an alternative norm [12]. Thereby the property $\overleftarrow{m}. = -.\overrightarrow{m}$ holds, since $\frac{m}{1-2^d} = -\frac{m}{2^d-1}$.

3. Conversion to the dyadic representation

In terms of the leftward periodicity, the axiom (5) is replaced by

$$\mathcal{D}_0 = \left\{ \pm\overleftarrow{b_d \dots b_2 1}. \right\} \text{ is the basic subset} \tag{6}$$

consisted of the 32-bit strings repeating periodically leftwards from the binary point on. The first bit is considered to be a unit for the uniqueness of representation $\pm b \times 2^j$. The convergence implies an alternative norm on $\mathbb{F}$, termed the *dyadic* one, whose $\frac{1}{2}$-based logarithm is defined by the valuation $\|x\|^{(2)} = j \Leftrightarrow x \in \mathcal{D}_j$ (see Appendix 6). Respecting convergence in the norm, the floating point reflects multiresolution of the dyadic numbers $\mathbb{D}$, whose approximate subsets $\mathcal{A}_j$ satisfy axioms:

$$\begin{gathered} \mathcal{A}_j \supset \mathcal{A}_{j+1}; \qquad x \in \mathcal{A}_j \Leftrightarrow 2x \in \mathcal{A}_{j+1}; \\ \cap_{-\infty}^{+\infty} \mathcal{A}_j = \{0\}; \qquad \cup_{-\infty}^{+\infty} \mathcal{A}_j = \mathbb{D}; \\ \mathcal{A}_0 = \{\dots b_2 b_1.\} \text{ is the basic subset} \end{gathered} \tag{7}$$

consisted of the dyadic integers that are bit sequences extended leftwards from the binary point on, implying convergence in regard to the dyadic norm.

In the last axiom of (7), the sign $\pm$ is omitted since a negative number is represented by the use of two's complement, i.e., $-\dots b_2 b_1. = \dots \widetilde{b_2}\widetilde{b_1}. + 1$ implying $\widetilde{b_k} = 1 - b_k$ is one's complement value of a binary digit. Due to the infinite extension leftwards, it is easy to prove $-x + x = 0$. The basic subset of the floating point (6) is hence writable in the form

$$\mathcal{D}_0 = \left\{ \overleftarrow{b_d \dots b_1}. + \widetilde{b_1} \right\}, \tag{8}$$

wherein b_1 is the bit $\pm = \begin{cases} 1 \\ 0 \end{cases}$ defining a sign of the number.

The floating point of a real number is the notation $\pm a \times 2^i$, wherein the exponent i is an integer and the mantissa a corresponds to a periodized value $.\overrightarrow{a_1 \dots a_d}$ up to the resolution 2^{-d}. Its conversion to the dyadic representation $b \times 2^j$, whereat $b = \overleftarrow{b_d \dots b_2 b_1}. + \widetilde{b_1}$, takes place in the following steps.

- $\pm.\overrightarrow{a_1 \dots a_d} \times 2^i = \mp\overleftarrow{a_1 \dots a_d}. \times 2^i = \mp\overleftarrow{a_{d-l+1} \dots a_d a_1 \dots a_{d-l}}. \times 2^{i+l}$, whereby $a_{d-l} = 1$ and $a_{d-l+1} = \dots = a_d = 0$, i.e., there are l zeros at the right end of the sequence.
- $-\overleftarrow{a_{d-l+1} \dots a_d a_1 \dots a_{d-l}}. = \overleftarrow{\widetilde{a_{d-l+1}} \dots \widetilde{a_d} \widetilde{a_1} \dots \widetilde{a_{d-l}}}. + 1.$

In accordance to that, the exponent is $j = i + l$ and the mantissa string is

$$b_d \dots b_1 = \begin{cases} a_{d-l+1} \dots a_d a_1 \dots a_{d-l}, & \text{for the } - \text{ sign} \\ \widetilde{a_{d-l+1}} \dots \widetilde{a_d} \widetilde{a_1} \dots \widetilde{a_{d-l}}, & \text{for the } + \text{ sign} \end{cases}.$$

Thereby the first bit b_1 of the string stores the information about the sign of a number.

The convenience of the dyadic representation concerns the sign incorporated, wherewith the addition and subtraction are easily implemented. Multiplication is done in the same manner as of the integers, since $\overleftarrow{m}. \times \overleftarrow{n}. = \frac{-m}{2^d-1} \times \frac{-n}{2^d-1} \approx \frac{m \times n/2^d}{2^d-1} = -\overleftarrow{p}.$, wherein p is the integer rounding of $m \times n/2^d$. The division is also likewise in the standard representation, since $\overleftarrow{m}. \div \overleftarrow{n}. = \frac{-m}{2^d-1} \div \frac{-n}{2^d-1} = m \div n \approx a \times 2^i = b \times 2^j$ implying conversion to the dyadic one.

4. The inverse conversion to reals

The inverse conversion to the floating point of reals is done straightforwardly. The dyadic representation $b \times 2^j$, whereat $b = \overleftarrow{b_d \dots b_2 b_1}. + \widetilde{b_1}$, is converted to $\pm a \times 2^i$ in the following manner.

- The sign is determined by the first b_1, considering $\left.\begin{matrix} 0 \\ 1 \end{matrix}\right\} = \mp$.
- The mantissa $a = .\overrightarrow{1 a_2 \dots a_d}$ is consisted of the string

  $$a_1 a_2 \dots a_d = \begin{cases} b_{d-l} \dots b_1 b_d \dots b_{d-l+1}, & b_1 = 0 \\ \widetilde{b_{d-l}} \dots \widetilde{b_1} \widetilde{b_d} \dots \widetilde{b_{d-l+1}}, & b_1 = 1 \end{cases}$$

  whereby $b_{d-l} = \widetilde{b_1}$ and $b_{d-l+1} = \dots = b_d = b_1$, i.e., there are l bits equal to b_1 at the left end of the sequence.
- The exponent $i = j - l$.

The standard representation $\pm.1a_2 \dots a_d \times 2^i$ is approximated up to the resolution 2^{-d}.

It is more appropriate, however, to regard the mantissa in terms of the periodized bit sequence $.\overrightarrow{1a_2 \dots a_d} = \frac{m}{2^d-1}$. The floating point multiresolution $\mathbb{F}$, in that manner considering the basic subset (5), is the sequence of detail subsets

$$\mathcal{D}_i = \left\{ \pm \frac{m}{2^{-i}(2^d-1)} \right\} \tag{9}$$

whereby $m = 1a_2 \dots a_d$ is a 32-bit integer starting by the unit. Since d is assumed to be 2^5, it holds $2^d - 1 = 2^{2^5} - 1 = (2^{2^4}+1) \times (2^{2^4}-1) = \dots = F_4 \times F_3 \times F_2 \times F_1 \times F_0$ implying $F_k = 2^{2^k} + 1$ are the Fermat numbers.

$F_0, \dots, F_4$, which appear in the product, are the only Fermat numbers known to be prime. A denominator of the reduced fractions from (9) is therefore obtained multiplying a power of two by the Fermat primes. According to the Gauss theorem [5], such a fraction exactly fits to constructible angle of the Euclidean geometry in regard to the full angle labelled by the unit. Although the multiresolution of constructible angles is far from being exhausted like that, (9) corresponds to the sub-multiresolution consisted by rationals of the structure. In that respect, the method should take an appropriate significance concerning the computational geometry (Figure 1).

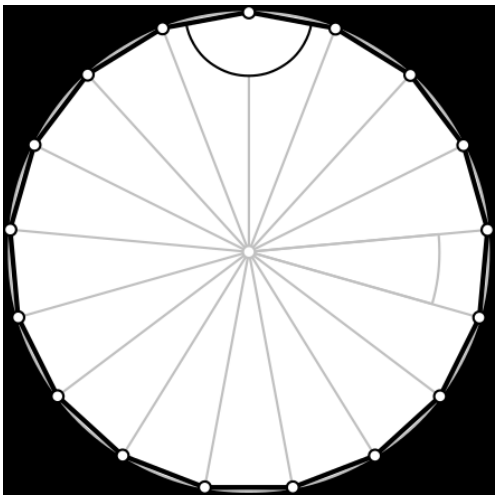

Figure 1: Regular 17-gon, which is constructible due to the Fermat number $F_2 = 17$. The labelled angles are exactly represented by the floating point multiresolution (9).

5. Conclusion

The standard floating point is a method for representing fractions, reduced only to the ones whose denominators are powers of two. However, it is aimed to reflect the multiresolution of real numbers defined by axioms (1). In that regard, it is convenient to consider convergence of periodized bit sequence in terms of the binary expansion

(5). The standard convergence is inapplicable, since it troubles calculation by the infinite sequence of binary digits tending rightwards. Fortunately, it corresponds to an alternative convergence of the same sequence in a leftward manner, implying multiresolution of the dyadic numbers that follows axioms (7).

The conversion and operations in the dyadic representation are easily implemented. The inverse conversion provides an opportunity to represent fractions whose denominators are not only the powers of two. Such a structure fits to constructible angles in regard to the full one, which makes the method significant for computational geometry.

6. Appendix – p-adic numbers (history and applications)

The emergence of p-adic numbers is almost concurrent to the real ones. Both of them appeared in the XIX century, having a long prehistory that goes back to the XVII at least. To be the beginner of real numbers is considered René Descartes, who postulated the number line. The concept is definitely established by Richard Dedekind, whose cuts of rationals constitute real numbers [2].

On the other hand, rudiments of p-adic numbers are cognizable in *Arithemtica infinitorum* by John Wallis [14] and also in papers by Leonard Euler, who were dealing with regularization of divergent series and calculation of their sums. Some papers by Ernst Kummer contain an implicit use as well. However, their formulation is responded to Kurt Hensel who considered expansion of rational functions in terms of the irreducible element powers [4]. He defined a p-adic number to be the series $x = \sum_{k \geq j} b_k p^k$ that converges in the norm $|x|^{(p)} = p^{-\|x\|^{(p)}}$ wherein $|x|^{(p)} = j$ is valuation signifying the less index k for which $b_k \neq 0$ [10].

According to the Ostrowski theorem [9], any norm defined on the rational number x is either the standard $|x|$, or the p-adic one $|x|^{(p)}$ whereat p is a prime. The closure implying topology induced by the first norm gives rise to reals, and by the second one to p-adic numbers [13]. Recalling that a norm $|\cdot|$ is the function satisfying axioms:

$$|x| = 0 \Leftrightarrow x = 0; \quad |x \times y| = |x| \times |y|; \quad |x + y| \leq |x| + |y|; \tag{10}$$

one observes that p-adic norm also satisfies $|x + y|^{(p)} \leq \max(|x|^{(p)}, |y|^{(p)})$ termed the *ultra-norm* relation, which is a stringent inequality than the last axiom of (10). Having significant applications in physics and biology [11], the ultra-metricity seems to be generating property of hierarchical structures [3]. Multiresolution of the floating point method is certainly one of them.

The dyadic number means a specific case considering the prime $p = 2$, corresponded to the binary code $\ldots b_1 b_0 . \ldots b_j = \sum_{k \geq j} b_k 2^k$. A basic application in the computer science concerns two's complement representation of negative numbers, which implies one's complement of the register increased by unit. The method was suggested by John von Neumann in the proposal for an electronic stored-program digital computer [8]. Its implementation was the *Electronic Delay Storage Automatic Calculator (EDSAC)* realized in 1949 by Maurice Wilkes and his team at the Univer-

sity of Cambridge Mathematical Laboratory. However, the method of complementing a d-bit register in order to calculate the opposite value is accurate only if one implies leftward periodization that converges in regard to the dyadic norm. In that respect, a value of the d-bit register corresponds to $\overleftarrow{m}. = \frac{m}{1-2^d}$ and its opposite is $\overleftarrow{\widetilde{m}}. + 1 = \frac{(2^d-1)-m}{1-2^d} + 1 = -\frac{m}{1-2^d} = -\overleftarrow{m}.$ whereat $\widetilde{m} = (2^d - 1) - m$ signifies the one's complement value. Considering normalization in regard to the unit, one obtains a basic set of the floating point multiresolution (8). Dyadic numbers are therefore at the very core of computation, offering a consistent realization of methods it already provides [1].

Acknowledgement. Supported by the Ministry of Education, Sceince and Technological Development of the Republic of Serbia through Mathematical Institute of the Serbian Academy of Sciences and Arts.

References

[1] V. Arashin, A. Khrennikov, *Applied Algebraic Dynamics*, de Gruyter Expositions in Mathematics 49, Walter de Gruyter, Berlin – New York, 2009.

[2] R. Dedekind, *Essays in the Theory of Numbers. I Continuity and Irrational Numbers; II The Nature and Meaning of Numbers*, The Open Court Publishing Company, Chicago, 1901.

[3] B. Dragovich, A. Khrennikov, A. Yu. S. V. Kozyrev, I. V. Volovich, *On p-adic Mathematical Physics*, p-adic Numbers, Ultrametric Analysis and Applications, **1(1)** (2009), 1–17.

[4] K. Hensel, *Über eine neue Begründung der Theorie der algebraischen Zahlen*, Jahresbericht der Deutschen Mathematiker-Vereinigung, **6(3)** (1897), 83–88.

[5] M. Křížek, F. Luca, L. Somer, *17 Lectures on Fermat Numbers: From Number Theory to Geometry*, Springer, New York, 2001.

[6] S. Mallat, *A Wavelet Tour of Signal Processing. The Sparse Way*, Elsevier, Amsterdam, 2009.

[7] J.-M. Muller, N. Bisebarre, F. de Dinechin, C.-P. Jeannerod, V. Lefèvre, G. Melquiond, N. Revol, D. Stehlé, S. Torres, *Handbook of Floating Point Arithmetic*, Springer, New York, 2010.

[8] J. von Neumann, *First Draft of a Report on the EDVAC*, Moore School of Electrical Engineering, University of Pennsylvania, June 30th 1945.

[9] A. Ostrowski, *Über einige Lösungen der Funktionalgleichung* $\varphi(x) \cdot \varphi(y) = \varphi(xy)$, Acta Mathematica, **41(1)** (1916), 271–284.

[10] Y. Perrin, *A journey through the history of p-adic numbers*, in: A. Escassut, C. Perez-Garcia, K. Shamseddine (eds.), *Advances in Ultrametric Analysis*, 14th International Conferences on p-adic Functional Analysis, June 30 - Jully 4, 2016. Université d'Auvergne, Aurillac, France, Contemporary Mathematics 704, American Mathematical Society, New York, 2018, 261–272.

[11] R. Rammal, G. Toulouse, M. A. Virasoro, *Ultra-metricity for Physicists*, Reviews of Modern Physics, **58(3)** (1986), 765–788.

[12] A. Rich, *Leftist Numbers*, The College Mathematics Journal, **39(5)** (2008), 330–336.

[13] A. M. Robert, *A Course in p-adic Analysis*, Springer-Verlag, New York, 2000.

[14] J. Wallis, *Arithmatica infinitorum*, Typis Leon, Lichfield Academiae typographi, London, 1656.

(received 19.10.2019; in revised form 08.12.2020; available online 09.04.2021)

Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, Belgrade, Serbia

E-mail: milosm@mi.sanu.ac.rs