mathematics

MDPI

Article

Analysis, Evaluation and Exact Tracking of the Finite Precision Error Generated in Arbitrary Number of Multiplications

Constantin Papaodysseus [1,*], Dimitris Arabadjis [2], Fotios Giannopoulos [1], Athanasios Rafail Mamatsis [1] and Constantinos Chalatsis [1]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechneiou 9, 15780 Athens, Greece; fgian@mail.ntua.gr (F.G.); trmamatsis@mail.ntua.gr (A.R.M.); khalatsi@mail.ntua.gr (C.C.)

[2] School of Engineering, University of West Attica, Petrou Ralli & Thivon 250 Egaleo, 12241 Athens, Greece; darampatzis@uniwa.gr

* Correspondence: cpapaod@cs.ntua.gr; Tel.: +30-21-0772-1476

Abstract: In the present paper, a novel approach is introduced for the study, estimation and exact tracking of the finite precision error generated and accumulated during any number of multiplications. It is shown that, as a rule, this operation is very "toxic", in the sense that it may force the finite precision error accumulation to grow arbitrarily large, under specific conditions that are fully described here. First, an ensemble of definitions of general applicability is given for the rigorous determination of the number of erroneous digits accumulated in any quantity of an arbitrary algorithm. Next, the exact number of erroneous digits produced in a single multiplication is given as a function of the involved operands, together with formulae offering the corresponding probabilities. In case the statistical properties of these operands are known, exact evaluation of the aforementioned probabilities takes place. Subsequently, the statistical properties of the accumulated finite precision error during any number of successive multiplications are explicitly analyzed. A method for exact tracking of this accumulated error is presented, together with associated theorems. Moreover, numerous dedicated experiments are developed and the corresponding results that fully support the theoretical analysis are given. Eventually, a number of important, probable and possible applications is proposed, where all of them are based on the methodology and the results introduced in the present work. The proposed methodology is expandable, so as to tackle the round-off error analysis in all arithmetic operations.

Keywords: finite precision error in a single multiplication; finite precision error in successive multiplications; exact tracking of round-off error; finite precision error; multiplication with finite word length; statistical properties of finite precision error; loss of significance during multiplication

Citation: Papaodysseus, C.; Arabadjis, D.; Giannopoulos, F.; Mamatsis, A.R.; Chalatsis, C. Analysis, Evaluation and Exact Tracking of the Finite Precision Error Generated in Arbitrary Number of Multiplications. *Mathematics* **2021**, *9*, 1199. https://doi.org/10.3390/math9111199

Academic Editors: Simeon Reich and Raimondas Ciegis

Received: 3 March 2021
Accepted: 19 May 2021
Published: 25 May 2021

1. Introduction

All contemporary computing machines store both integer and floating-point numbers with a finite number of digits. This piece of fixed-sized data that is handled as a unity by the instruction set or the processor's hardware is called finite word; the number of bits that form this piece of data, is frequently called "finite word length" or "employed precision". In addition, on the hardware level, a computer performs fundamental operations, using a finite word length. Nowadays, dedicated software programs have been developed, which perform operations with a finite number of digits, the value of which is chosen by the programmer and/or the user, the only limitation being the memory and time constraints. We shall also use for this number of digits the term "finite word length" or "employed precision".

Due to the fact that the precision with which all arithmetic operations are made is always limited, a numerical error is, as a rule, accumulated during the execution of most algorithms. In particular, in various algorithms and corresponding applications, the

obtained results may be totally erroneous and/or unreliable due to the aforementioned reasons, which are inherent to all computing machines. We stress that this problem exists even when an arbitrarily large finite word length is employed for the execution of the algorithm, as it will become evident from the analysis introduced in the present paper. For this reason, we will use the term "finite precision error" for this numerical error; various authors and researchers also use the term "quantization error", "round-off error" or other equivalent terms.

Consequently, a number of articles address the associated issues and the problems they generate. Thus, for example, authors in [1] study the finite precision error in the least mean square (LMS) adaptive algorithm and they show that the error's mean squared value is inversely proportional to the adaptation step size μ. Reference [2] introduces a fast algorithm for exponentially weighted least squares Kalman filtering, which suffers less from finite precision error drawbacks, intrinsic to this class of algorithms. Reference [3] presents algorithms for accurately converting floating-point numbers to decimal representation. Article [4] studies the finite precision effects on the execution of the Lanczos algorithm for solving the standard non-symmetric eigenvalue problem. The authors of [5] study round-off error propagation in an algorithm which computes the orthonormal basis of a Krylov subspace with Householder orthonormal matrices. Authors in [6] study the propagation of round-off error near the periodic orbits of a discretized linear area-preserving map. The round-off error probability distribution, considered as a function of time, is shown to be a calculable algebraic number. In [7] it is shown that there are theoretically convergent schemes that solve non-linear partial differential equations, which can produce numerical steady state solutions that do not correspond to steady state solutions of the boundary value problem. In [8], it is pointed out that the convergence of Gegenbauer polynomials at the endpoints is affected by round-off error; the article proposes both parameter optimization and reduction of the round-off error for the Gegenbauer reconstruction method. In [9], a set of specific semantics is introduced which describes the propagation of round-off error during a calculation. The authors of [10] give an estimation of the round-off error generated in long-time integration in a number of standard, nonlinear systems. Authors in [11] introduce an algorithm for the computation of the orthogonal Fourier–Mellin moments which is of linear complexity and is resistant to finite precision error effects. Reference [12] proposes a method for dealing with the instability of the digital frequency synthesis (DFS), caused by the round-off error. Article [13] presents bounds for round-off error, generated in various algorithms. Moreover, another approach is presented in [14], according to which, the evolution of round-off error in chaotic maps is treated as an additive noise to the expected exact solutions; the introduced method spots a threshold below which global errors may be ignored. Article [15] studies the round-off error generated during computation of Hardy's multiquadric and its related interpolators and proposes the use of arbitrary precision arithmetics to circumvent the associated finite precision error problems. Authors of [16] propose a fast, resistant to finite precision error method for evaluation of high order Zernike moments. Article [17] proposes a method for an improved scaling of finite precision error analysis. Finally, in [18,19], a preliminary form of the approach introduced here is presented.

In the present paper, we introduce a novel approach for studying and evaluating the finite precision error generated during the operation of multiplication. It is shown that, as a rule, this operation is very "toxic", in the sense that it may force the finite precision error accumulation to grow arbitrarily large. The exact amount of the generated number of erroneous digits added or subtracted to the result (product) of this operation is given. Consequently, the probabilities the number of erroneous digits of the product differ by k from the maximum number of erroneous digits of the operands are explicitly computed. In the process of doing so, a set of general definitions is given, applicable to all operations performed with finite word length by a computing machine. Then, the accumulation of erroneous digits after an arbitrary number of successive multiplications is extensively analyzed. Statistical properties of this accumulated error are stated that allow for exact

error prediction when the distribution of the associated operands are given. In addition, a number of theoretical results are introduced, which allow for the exact tracking of the generated and accumulated finite precision error during any number of multiplications in general. Numerous experimental results are presented, which fully support the presented theoretical analysis. We stress that the introduced methodology is expandable, so as to tackle all arithmetic operations.

2. A Set of Basic Definitions, Notations and Abbreviations

The entire analysis will be mainly made in the decimal arithmetic system, only because this system is far more familiar to most users. However, all results referred to in the present work hold perfectly well for the binary system too, or any other radix; the corresponding analysis and deductions may be obtained by means of a quite straightforward and slight modification of the approach introduced here.

In any arithmetic system, we assume that all numbers are expressed in scientific/canonical form. Thus, any number x is written as $mantissa \cdot 10^{\tau}$, in the decimal arithmetic system, where $|mantissa| \in [1, 10)$, $\tau \in \mathbb{Z}$; in the binary system, the same number is expressed as $mantissa \cdot 2^{\tau}$, where $|mantissa| \in [1, 2)$, $\tau \in \mathbb{Z}$. Independently of the employed radix, we will use the symbols $man(x)$ and $E(x)$ for the mantissa and the exponent of any quantity x respectively. We shall demonstrate in the following that the analysis introduced here based on the decimal radix offers accurate results and prediction for the multiplication(s) performed by computing machines.

Abbreviations 1. *We will use the acronym e. d. d. in place of "erroneous decimal digits" and c. d. d. for "correct decimal digits". In general, abbreviation "d. d." stands for "decimal digits". The abbreviation f. p. e. stands for "finite precision error". The symbols $\#edd(x)$ and $\#cdd(x)$ stand for the number of e. d. d. accumulated in quantity x, due to f. p. e., and its number of c. d. d. respectively.*

Notation 1. *The expressions the algorithm "has failed" or it "has been destroyed due to f. p. e." mean that the algorithm in hand offers completely unreliable results, at a certain iteration.*

Suppose that any two numbers α and β are given, both written in canonical form. In order to unambiguously verify if these two numbers share a common number of initial digits (i.e., stem of digits), starting from the most significant one, we shall employ the following:

Definition 1. *Consider two numbers, α, β of the same sign, written in scientific form, with the same number, n, of decimal digits in the mantissa:*

$$\alpha = man(\alpha) \cdot 10^{\tau},\ \beta = man(\beta) \cdot 10^{\rho}$$

where $\tau = E(\alpha)$ and $\rho = E(\beta)$. Let us assume, without any loss of generality, that $\tau \geq \rho$ holds. Then, these two numbers share the first μ, $\mu \in \mathbb{N}$ digits (they have the first μ digits in common) if and only if:

$$|\alpha - \beta| = w \cdot 10^{\tau - \mu}\ ,\ w \in [1, 10).$$

Consequently, α and β differ in the last λ, $\lambda \in \mathbb{N}$ digits if and only if:

$$|\alpha - \beta| = z \cdot 10^{\tau - (n - \lambda)}\ ,\ z \in [1, 10). \tag{2.1}$$

Evidently, in the binary system, the two numbers share the first μ, $\mu \in \mathbb{N}$ bits if and only if

$$|\alpha - \beta| = w \cdot 2^{E(\alpha) - \mu}\ ,\ w \in [1, 2), \tag{2.2}$$

where n is the finite word length, while they differ in the last $\lambda, \lambda \in \mathbb{N}$ digits if and only if

$$|\alpha - \beta| = z \cdot 2^{E(a) - (n - \lambda)}\ ,\ z \in [1, 2). \tag{2.3}$$

If the aforementioned relation offers a negative λ, then, by definition, $\lambda = 0$, namely α and β are identical as far as all their n digits are concerned.

Now, we shall give a couple of examples in order to clarify the content of Definition 1.

Example 1.

$$n_1 = 4.269587962\mathbf{400597}\cdot 10^4$$

$$n_2 = 4.269587962\mathbf{393951}\cdot 10^4$$

A simple inspection might lead someone to deduce that these two numbers differ by six (6) decimal digits. Actually and according to Definition 1, the following holds:

$$\tau = \rho = 4$$

$$n = 16$$

where the absolute difference is $|n_1 - n_2| = 6.6459\cdot 10^{-8}$

Hence,

$$\tau - (n - k) = -8 \Leftrightarrow k = 4$$

Therefore, the two aforementioned numbers differ in four (4) decimal digits, contrary to a probable initial expectation.

Example 2. *According to Definition 1, the two numbers with 32 decimal digits in the mantissa,*

$$n_1 = 6.98765432123456789\mathbf{77777777777777}\cdot 10^1,$$

$$n_2 = 6.98765432123456789\mathbf{01523451234533}\cdot 10^1\ ,$$

differ by 14 decimal digits, shown in bold, since $\tau = \rho = 1,\ n = 32$, while their absolute difference is $|n_1 - n_2| = 7.6254326543244\cdot 10^{-17}$; hence, $\tau - (n - k) = -5 \Leftrightarrow k = 14$.

Additionally, the two numbers $n_1 = 1.112324567422342\cdot 10^4$ and $n_2 = 1.112324567421112\cdot 10^4$ differ by 4 decimal digits in the mantissa, since $\tau = \rho = 4,\ n = 16$ and $|n_1 - n_2| = 1.230000634677708\cdot 10^{-8}$; therefore, $\tau - (n - k) = -8 \Leftrightarrow k = 4$.

Similarly, the two numbers $n_1 = 1.00000000\cdot 10^{\tau}$ and $n_2 = 9.99999999\cdot 10^{\tau-1}$ do not differ at all, since $|n_1 - n_2| = 9.999999717180685\cdot 10^{-10}$; hence $\tau - (n - k) = -10 \Leftrightarrow k = -1 - \tau < 0$.

Suppose, moreover, that all operations were made with infinite precision; then, let an arbitrary quantity α have the value $\widetilde{a}^c$, where superscript c indicates the ideally correct value of α. Next, suppose that the very same quantity α is calculated in a computing machine which performs the same operations as in the infinite case, using n digits in the mantissa; suppose that this machine generates the representation α_n for the specific quantity a. Then, a rigorous relation between α_n and $\widetilde{a}^c$ is obtained via the following:

Definition 2. *Let us assume that we restrict the infinite precision quantity $\widetilde{a}^c$ to its first n digits, obtaining quantity a^c. Let us also assume that comparison of α_n and a^c by means of Definition 1, manifests that these two quantities differ by λ_α digits. Then, we deduce that quantity α_n has the first $\lambda_c = (n - \lambda_\alpha)$ digits correct and all its other digits erroneous. The aforementioned statement holds for both the binary system, which is the base of contemporary computing machines, as well as for the decimal radix.*

A number of practical examples associated with the above Definition, will be given in Section 6.

It is known that a mantissa represented by a number of bits, say $\nu,\ \nu \in \mathbb{N}$, in a computing machine is approximated in the decimal radix by a number n of d. d., pretty close to the nearest integer of $\nu\cdot log_{10}2$. Since $\nu\cdot log_{10}2$ is, as a rule, not an integer, then the

number of correct digits of a quantity's decimal representation may fluctuate by one digit at most.

3. Generation of Finite Precision Error in a Single Multiplication and Corresponding Probabilities

This Section presents a solution to the following problem: consider two arbitrary numbers, say α_n, β_n found in a computing machine that uses a finite word length of n decimal digits in the mantissa. Moreover, suppose due to an ensemble of previous calculations α_n has been computed with λ_α erroneous decimal digits (e. d. d.) in its mantissa, while β_n with λ_β e. d. d. in the mantissa. In addition, consider that multiplication $\gamma_n = \alpha_n\beta_n$ is executed in this computing machine. Then, so far, it has been an open question to determine the exact number of e. d. d. with which γ_n is evaluated; moreover, the corresponding probabilities that γ_n is computed with a specific number of e. d. d. must be evaluated.

3.1. Bounds and Evaluation of the Finite Precision Error Produced in a Single Multiplication

Consider any two quantities α, β having $\widetilde{a}^c$ and $\widetilde{\beta}^c$ ideally correct digits, should all operations and representations be made with infinite precision. Next, suppose that quantities α and β have been evaluated in a computing machine using n d. d. in the mantissa; we let the representations of these two numbers in this computing machine be a_n and β_n, respectively. In addition, following Definition 2, we let the restriction of $\widetilde{a}^c$ and $\widetilde{\beta}^c$ in this machine be α^c, β^c respectively. We would like to emphasize that the difference between a_n and a^c is the following: quantity α_n may have been evaluated with finite precision error, due to previous calculations. On the contrary, a^c is free of finite precision error since it is always considered to be a restriction of the ideally correct value of $\widetilde{a}^c$ in n decimal digits.

Consider, moreover, the product $\gamma = \alpha \cdot \beta$, executed both with infinite precision yielding product $\gamma^c = \alpha^c \cdot \beta^c$, as well as in a computing machine using n digits in the mantissa, generating $\gamma_n = \alpha_n \cdot \beta_n$. In addition, suppose that, due to previous calculations, α_n has been computed with λ_α erroneous decimal digits (e. d. d.), ($\Leftrightarrow \lambda_a^c = n - \lambda_\alpha$ correct decimal digits), while β_n with λ_β e. d. d. ($\Leftrightarrow \lambda_\beta^c = n - \lambda_\beta$ c. d. d.) due to the fact that all operations have been made with a finite word length. We note, as it will become evident in the following analysis, that the finite precision error generated in the multiplication process is located only in the mantissae of the involved terms. Hence, we may assume that α_n, β_n, α^c and β^c are plain mantissae, namely that $E(\alpha_n) = E(\beta_n) = \tau = 0$. In order to study the finite precision error generated in the computation of the product γ, we distinguish a number of cases, which are analytically presented below; in addition, a concise presentation of all these cases takes place in Tables 1 and 2, positioned in the end of the present sub-section. Thus:

Table 1. This refers to Case 1, with $\lambda_\alpha^c = \lambda_\beta^c = \lambda^c$. The first column under the title "sub-cases", the eventual values of $M_1 = |\alpha_n x + \beta_n y|$ are shown. In Line 3, in the right of the same title, the possible values of the product $|man(\alpha_n)man(\beta_n)|$ are presented. The obtained number of correct decimal digits (c. d. d.) of product $\gamma_n = a_n\beta_n$ is shown in bold in each corresponding square.

Case 1: Number of c. d. d. of Product γ_n When λ_α^c=λ_β^c=λ^c.		
Let M_1=$\lvert\alpha_n x$+$\beta_n y\rvert$.		
Sub-Cases	**(1.i)$\lvert man(\alpha_n)man(\beta_n)\rvert$<10**	**(1.ii)$\lvert man(\alpha_n)man(\beta_n)\rvert\geq$10**
(a) $100 \leq M_1 <$ UB	$\lambda^c - 2$	$\lambda^c - 1$
(b) $10 \leq M_1 < 100$	$\lambda^c - 1$	λ^c
(c) $1 \leq M_1 < 10$	λ^c	$\lambda^c + 1$
(d) $10^{-1} \leq M_1 < 1$	$\lambda^c + 1$	$\lambda^c + 2$
(e) $10^{-k} \leq M_1 < 10^{-(k-1)}$, $k = 2, 3, 4$	$\lambda^c + k$	$\lambda^c + k + 1$

Table 2. It refers to Case 2, where $\lambda_{\alpha}^{c} \neq \lambda_{\beta}^{c}$ and in particular $\lambda_{\alpha}^{c} > \lambda_{\beta}^{c}$, without any loss of generality. The first column under the title "sub-cases", the eventual values of $M_2 = \left|\alpha_n x + \beta_n y \cdot 10^{-(\lambda_a^c - \lambda_\beta^c)}\right|$ are shown. In Line 3, in the right of the same title, the possible values of the product $|man(\alpha_n)man(\beta_n)|$ are presented. The obtained number of correct decimal digits (c. d. d.) of product $\gamma_n = a_n\beta_n$ is shown in bold in each corresponding square.

Case 2: Number of c. d. d. of Product γ_n When $\lambda_{\alpha}^{c} > \lambda_{\beta}^{c}$.		
Let $M_2 = \left\|\alpha_n x + \beta_n y \cdot 10^{-(\lambda_a^c - \lambda_\beta^c)}\right\|$.		
Sub-Cases	**(2.i)** $\|man(\alpha_n)man(\beta_n)\| < 10$	**(2.ii)** $\|man(\alpha_n)man(\beta_n)\| \geq 10$
(a) $100 \leq M_1$	$\lambda_{\beta}^{c} - 2$	$\lambda_{\beta}^{c} - 1$
(b) $10 \leq M_1 < 100$	$\lambda_{\beta}^{c} - 1$	λ_{β}^{c}
(c) $1 \leq M_1 < 10$	λ_{β}^{c}	$\lambda_{\beta}^{c} + 1$
(d) $10^{-1} \leq M_1 < 1$	$\lambda_{\beta}^{c} + 1$	$\lambda_{\beta}^{c} + 2$
(e) $10^{-k} \leq M_1 < 10^{-(k-1)}$, $k = 2, 3, 4$	$\lambda_{\beta}^{c} + k$	$\lambda_{\beta}^{c} + k + 1$

Case 1. Quantities α_n and β_n share the same number of correct decimal digits $\left(\lambda_a^c = \lambda_\beta^c\right)$.

Therefore, according to Definition 2, it holds that

$$|\alpha_n - a^c| = z \cdot 10^{0-(n-\lambda_a)}\ ,\ z \in [1, 10),$$

from which we deduce that we can express quantities α^c and β^c as follows:

$$\alpha^c = \alpha_n + y \cdot 10^{-\lambda_a^c} \qquad \beta^c = \beta_n + x \cdot 10^{-\lambda_\beta^c}, \tag{3.1}$$

where x and y are the signed mantissae of the finite precision error. Taking (3.1) into consideration, we may write:

$$\gamma^c = \alpha^c \cdot \beta^c \overset{(3.1)}{=} \alpha_n\beta_n + \alpha_n x \cdot 10^{-\lambda_\beta^c} + \beta_n y \cdot 10^{-\lambda_a^c} + xy \cdot 10^{-(\lambda_a^c + \lambda_\beta^c)}.$$

Since, by hypothesis, $\lambda_a^c = \lambda_\beta^c = \lambda^c$, the above expression becomes

$$\gamma^c = \alpha_n\beta_n + (\alpha_n x + \beta_n y) \cdot 10^{-\lambda^c} + xy \cdot 10^{-2\lambda^c} \tag{3.2}$$

Thus, according to Definition 2, the finite precision error (f. p. e.) with which product γ_n has been evaluated is

$$\varepsilon_\gamma = (\alpha_n x + \beta_n y) \cdot 10^{-\lambda^c} + xy \cdot 10^{-2\lambda^c}. \tag{3.3}$$

We point out that the subsequent analysis may use (3.3) with slight, straightforward modifications; in fact, in practice, it is sufficient to keep the first-order terms when $\lambda^c \geq 3$, since term $xy \cdot 10^{-2\lambda^c}$ is practically negligible. Should the algorithm tend to fail, i.e., if $\lambda^c < 3$, then, ε_γ of (3.3) can be used in the subsequent analysis, in a very straightforward manner. To compute the number of erroneous decimal digits (e. d. d.) of γ_n, it is absolutely necessary to distinguish the cases $|man(\alpha_n)man(\beta_n)| < 10$ and $|man(\alpha_n)man(\beta_n)| \geq 10$, for reasons that will become evident in the following. In fact:

Case 1.i. It refers to inequality

$$|man(\alpha_n)man(\beta_n)| < 10. \tag{3.4}$$

Immediately below we will show that, in this case, the maximum number of additional erroneous decimal digits generated in the multiplication $\gamma_n = \alpha_n \cdot \beta_n$ is 2. Indeed, here,

since we have assumed that all involved quantities have zero exponents, the product $\alpha_n\beta_n$, is given by $\alpha_n\beta_n = man(\alpha_n)man(\beta_n) = man(\alpha_n\beta_n)$; now, (3.2) becomes

$$\gamma^c = \alpha^c\beta^c = man(\alpha_n\beta_n) + (\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c} \tag{3.5}$$

using the aforementioned first-order approximation. Hence, given that $|man(\alpha_n)man(\beta_n)| < 10$, it is rather straightforward to show that the supremum of quantity $|\alpha_n x + \beta_n y|$ may acquire is $UB = 110$, since all terms, α_n, β_n, x, y, are mantissae. Therefore, we distinguish the following sub-cases:

Case 1.i.a:

$$100 \leq |\alpha_n x + \beta_n y| < UB. \tag{3.6}$$

Then, $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10^2$, which implies that

$$man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c-2)}. \tag{3.7}$$

The above relation (3.7) implies that

$$|man(\alpha^c\beta^c) - man(\alpha_n\beta_n)| = |man(\alpha_n x + \beta_n y)|\cdot 10^{-(\lambda^c-2)},$$

Which according to Definition 2 shows that quantity $\gamma_n = \alpha_n\beta_n$ has been computed with two less correct decimal digits, namely with $(\lambda^c - 2)$ correct decimal digits (c. d. d.) or equivalently with two additional erroneous decimal digits than those of the operands α_n and β_n.

Case 1.i.b:

$$10 \leq |\alpha_n x + \beta_n y| < 100. \tag{3.8}$$

In this sub-case, $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10$, which implies that

$$\gamma^c = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c-1)}. \tag{3.9}$$

Consequently, Definition 2, implies that γ_n has been computed with one less c. d. d. than α_n and β_n.

Case 1.i.c:

$$1 \leq |\alpha_n x + \beta_n y| < 10. \tag{3.10}$$

Now, $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)$, implying that

$$\gamma^c = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c}. \tag{3.11}$$

Together with Definition 2, this means that γ_n has the same number of c. d. d. with α_n and β_n, namely λ^c.

Case 1.i.d:

$$10^{-1} \leq |\alpha_n x + \beta_n y| < 1. \tag{3.12}$$

Here it holds that $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10^{-1}$, implying that

$$\gamma^c = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c+1)}. \tag{3.13}$$

Consequently, one may deduce that the number of γ_n's erroneous decimal digits (e. d. d.) has been reduced by one.

Case 1.i.e. This constitutes a generalization of Case 1.i.d.; in fact, now, we assume that the following inequality holds:

$$10^{-k} \leq |\alpha_n x + \beta_n y| < 10^{-(k-1)},\ k = 2, 3, 4. \tag{3.14}$$

In this, more general case, it holds that $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10^{-k}$, therefore,

$$\gamma^c = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c+k)}. \tag{3.15}$$

Hence, the number of correct decimal digits (c. d. d.) of γ_n has been increased by k. The same approach may be applied for $k \geq 5$; however, we will show that the corresponding probabilities are negligible in practice.

Case 1.ii, which concerns inequality

$$|man(\alpha_n)man(\beta_n)| \geq 10. \tag{3.16}$$

Since $\alpha_n,\ \beta_n,\ x,\ y$ are mantissae, $|\alpha_n x + \beta_n y| < 200$ holds. Therefore, we distinguish the following cases:

Case 1.ii.a:

$$100 \leq\ |\alpha_n x + \beta_n y| < 200. \tag{3.17}$$

In this case, $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10^2$, which implies that

$$\gamma^c = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c-2)}. \tag{3.18}$$

However, $\alpha_n\beta_n = man(\alpha_n)man(\beta_n) = man(\alpha_n\beta_n)\cdot 10 \Rightarrow \alpha^c\beta^c = man(\alpha^c\beta^c)\cdot 10$, if the algorithm has not failed, which means that $E(\alpha^c\beta^c) = E(\alpha_n\beta_n)$. Thus, (3.18) now reads:

$$\begin{array}{l} \gamma^c = \alpha^c\beta^c = man(\alpha^c\beta^c)\cdot 10 = man(\alpha_n\beta_n)\cdot 10 + (\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c} \\ \quad = man(\alpha_n\beta_n)\cdot 10 + man(\alpha_n x + \beta_n y)\cdot 10^2\cdot 10^{-\lambda^c} \Leftrightarrow \\ man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c-1)} \Rightarrow \\ |man(\alpha^c\beta^c) - man(\alpha_n\beta_n)| = |man(\alpha_n x + \beta_n y)|\cdot 10^{-(\lambda^c-1)}. \end{array} \tag{3.19}$$

The above equality (3.19), together with Definition 2 dictates that γ_n has been evaluated with $(\lambda^c - 1)$ correct decimal digits (c. d. d.). Even though (3.6) and (3.17) are quite similar, now, the number of erroneous decimal digits (e. d. d.) of γ_n has been reduced by one, due to the right shift the computing machine has performed, to represent γ_n in its canonical form.

Case 1.ii.b:

$$10 \leq\ |\alpha_n x + \beta_n y| < 100. \tag{3.20}$$

In this case, $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10$ holds. However, now, once more, provided that the algorithm has not failed, one obtains $\alpha_n\beta_n = man(\alpha_n\beta_n)\cdot 10$ and $\alpha^c\beta^c = man(\alpha^c\beta^c)\cdot 10$. Hence,

$$\begin{array}{c} \gamma^c = \alpha^c\beta^c = man(\alpha^c\beta^c)\cdot 10 = \left(man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c}\right)\cdot 10 \\ \Leftrightarrow man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c}. \end{array} \tag{3.21}$$

Definition 2 indicates that γ_n has been evaluated with λ^c c. d. d. (i.e., with no additional finite precision error (f. p. e.)).

Case 1.ii.c:

$$1 \leq\ |\alpha_n x + \beta_n y| < 10. \tag{3.22}$$

Now it holds that $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)$. Supposing that the algorithm has not failed, one deduces

$$\begin{array}{c} \gamma^c = \alpha^c\beta^c = man(\alpha^c\beta^c)\cdot 10 = man(\alpha_n\beta_n)\cdot 10 + man(\alpha_n x + \beta_n y)\cdot 10^{-\lambda^c} \\ \Leftrightarrow man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c+1)} \Rightarrow \\ |man(\alpha^c\beta^c) - man(\alpha_n\beta_n)| = |man(\alpha_n x + \beta_n y)|\cdot 10^{-(\lambda^c+1)}. \end{array} \tag{3.23}$$

The later implies that quantity $\alpha_n\beta_n$ has been computed with an additional correct decimal digit, i.e., that the multiplication operation has relaxed the finite precision error (f. p. e.) by one decimal digit.

Case 1.ii.d:

$$10^{-k} \leq |\alpha_n x + \beta_n y| < 10^{-(k-1)},\ k = 1, 2, 3, 4. \tag{3.24}$$

In this case, it holds that $\alpha_n x + \beta_n y = man(\alpha_n x + \beta_n y)\cdot 10^{-k}$, hence,

$$\begin{gathered}\gamma^c = man(\alpha^c\beta^c)\cdot 10 = man(\alpha_n\beta_n)\cdot 10 + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c+k)} \Leftrightarrow \\ man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man(\alpha_n x + \beta_n y)\cdot 10^{-(\lambda^c+k+1)}.\end{gathered} \tag{3.25}$$

Consequently, the number of correct decimal digits (c. d. d.) of product γ_n has been increased by $k+1$ in this case.

Case 2. α_n and β_n have been calculated with different number of correct decimal digits $\left(\lambda_a^c \neq \lambda_\beta^c \cdot\ \lambda_\alpha \neq \lambda_\beta\right)$ Without any loss of generality, we may assume that $\lambda_\alpha \langle \lambda_\beta \Leftrightarrow \lambda_a^c \rangle \lambda_\beta^c$. Consequently, once more it holds that:

$$\begin{gathered}\alpha^c = \alpha_n + y\cdot 10^{-\lambda_a^c} \qquad \beta^c = \beta_n + x\cdot 10^{-\lambda_\beta^c} \Rightarrow \\ \gamma^c = \alpha^c\cdot\beta^c = \alpha_n\beta_n + \alpha_n x\cdot 10^{-\lambda_\beta^c} + \beta_n y\cdot 10^{-\lambda_a^c} + xy\cdot 10^{-(\lambda_a^c+\lambda_\beta^c)}.\end{gathered} \tag{3.26}$$

As in Case 1, we will use a first-order approximation in (3.26).

Again, the introduced analysis may be extended in a straightforward manner to incorporate the higher order term, too; however, as it will become clear from the subsequent sections, the accuracy improvement is negligible, given also the dramatic increase in complexity. Thus, we may safely assume that $\gamma^c = \alpha_n\beta_n + \alpha_n x\cdot 10^{-\lambda_\beta^c} + \beta_n y\cdot 10^{-\lambda_a^c}\cdot\alpha^c\beta^c = \alpha_n\beta_n + \left(\alpha_n x + \beta_n y\cdot 10^{-(\lambda_a^c-\lambda_\beta^c)}\right)\cdot 10^{-\lambda_\beta^c}$. After setting $\delta = \lambda_a^c - \lambda_\beta^c \geq 1$, we obtain:

$$\alpha^c\beta^c = \alpha_n\beta_n + \left(\alpha_n x + \beta_n y\cdot 10^{-\delta}\right)\cdot 10^{-\lambda_\beta^c}. \tag{3.27}$$

We must now repeat the analysis previously made in connection with Case 1, by letting $\left(\alpha_n x + \beta_n y\cdot 10^{-\delta}\right)$ play the role of $\alpha_n x + \beta_n y$ and λ_β^c play the role of λ^c. Hence, we again distinguish the cases $|man(\alpha_n)man(\beta_n)| < 10$ and $|man(\alpha_n)man(\beta_n)| \geq 10$, thus getting:

Case 2.i: $|man(\alpha_n)man(\beta_n)| < 10$

Case 2.i.a:

$$100 \leq \left|\alpha_n x + \beta_n y\cdot 10^{-\delta}\right|. \tag{3.28}$$

In this case, $man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man\left(\alpha_n x + \beta_n y\cdot 10^{-\delta}\right)\cdot 10^{-(\lambda_\beta^c-2)}$.

Namely, product γ_n is computed with two additional erroneous decimal digits (e. d. d.) than λ_β.

Case 2.i.b:

$$10 \leq \left|\alpha_n x + \beta_n y\cdot 10^{-\delta}\right| < 100. \tag{3.29}$$

Now $man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man\left(\alpha_n x + \beta_n y\cdot 10^{-\delta}\right)\cdot 10^{-(\lambda_\beta^c-1)}$ which means that product γ_n is calculated with one additional erroneous decimal digits (e. d. d.) than β_n.

Case 2.i.c:

$$1 \leq \left|\alpha_n x + \beta_n y\cdot 10^{-\delta}\right| < 10. \tag{3.30}$$

Here, it holds that $man(\alpha^c\beta^c) = man(\alpha_n\beta_n) + man\left(\alpha_n x + \beta_n y\cdot 10^{-\delta}\right)\cdot 10^{-\lambda_\beta^c}$.

Hence, product γ_n is calculated with no additional e. d. d. when compared to β_n, namely λ_β.

Case 2.i.d:

$$10^{-k} \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10^{-(k-1)},\ k = 1,2,3. \tag{3.31}$$

In this case, $man(\alpha^c \beta^c) = man(\alpha_n \beta_n) + man\left(\alpha_n x + \beta_n y \cdot 10^{-\delta}\right) \cdot 10^{-(\lambda_\beta^c + k)}$.

Then, γ_n is computed with k less e. d. d. than $\lambda_\beta = \max\{\lambda_\alpha, \lambda_\beta\}$. The same approach may be applied for $k \geq 4$, however, the probability that such a case holds, is negligible in practice.

Case 2.ii: $|man(\alpha_n) man(\beta_n)| \geq 10$

For this case, we distinguish the following sub-cases:

Case 2.ii.a:

$$100 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 200. \tag{3.32}$$

In this case we obtain $man(\alpha^c \beta^c) = man(\alpha_n \beta_n) + man\left(\alpha_n x + \beta_n y \cdot 10^{-\delta}\right) \cdot 10^{-(\lambda_\beta^c - 1)}$.

The above equation dictates that product γ_n has been evaluated with $\left(\lambda_\beta^c - 1\right)$ correct decimal digits (c. d. d.).

Case 2.ii.b:

$$10 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 100. \tag{3.33}$$

Then, $man(\alpha^c \beta^c) = man(\alpha_n \beta_n) + man\left(\alpha_n x + \beta_n y \cdot 10^{-\delta}\right) \cdot 10^{-\lambda_\beta^c}$. Consequently, Definition 2 implies that γ_n has exactly the same number of erroneous decimal digits (e. d. d.) as β_n.

Case 2.ii.c:

$$1 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10. \tag{3.34}$$

Now $man(\alpha^c \beta^c) = man(\alpha_n \beta_n) + man\left(\alpha_n x + \beta_n y \cdot 10^{-\delta}\right) \cdot 10^{-(\lambda_\beta^c + 1)}$. Therefore, quantity $\alpha_n \beta_n$ has been computed with an additional c. d. d., as compared to λ_β^c.

Case 2.ii.d:

$$10^{-k} \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10^{-(k-1)},\ k = 1,2,3. \tag{3.35}$$

Here, it holds that $man(\alpha^c \beta^c) = man(\alpha_n \beta_n) + man\left(\alpha_n x + \beta_n y \cdot 10^{-\delta}\right) \cdot 10^{-(\lambda_\beta^c + k + 1)}$. Hence, the number of c. d. d. of γ is greater by $(k+1)$ than $\lambda_\beta^c = \min\left\{\lambda_\alpha^c, \lambda_\beta^c\right\}$.

3.2. Probabilities for Obtaining a Specific Number of Erroneous Digits in the Execution of a Single Multiplication

We once more adopt the distinction in cases made in Section 3.1, which are presented in Tables 1 and 2 below, in a very concise manner. In fact,

Case 1: $\lambda_a = \lambda_\beta = \lambda$.

Moreover, in connection with Case 2 ($\lambda_\alpha^c \neq \lambda_\beta^c$), we cite the following Table 2:

Consider any multiplication of two numbers α_n and β_n sharing the same number λ of e. d. d. Thus, quantity $\gamma_n = \alpha_n \beta_n$ is computed with $\lambda + \xi$ e. d. d. If $\xi > 0$, γ_n is computed with ξ additional e. d. d., while if $\xi < 0$, γ_n is computed with ξ less e. d. d. Then, following Section 3.1, ξ is a random variable, independent of λ. Therefore, the probabilities for obtaining a specific value of ξ are independent of λ; this suggests the use of the following notation:

Notation 2. *Let $\gamma_n = \alpha_n \beta_n$; then, quantity γ_n is computed with $\lambda + \xi$ e. d. d., $\xi = 2, 1, 0, -1, -2, \ldots$. We denote the corresponding probabilities by $P^{EQ}(\xi; \alpha_n, \beta_n)$.*

As before, α_n and β_n are mantissae and x, y are the mantissae of the f. p. e. stochastic part. Hence, for the evaluation of $P^{EQ}(\xi; \alpha_n, \beta_n)$, it is necessary to know the joint probability density function (pdf) of the random variables X, Y, which express the f. p. e. of the mantissae x, y respectively; we shall symbolize this joint pdf as $f_{XY}(x, y)$. We shall give the general formulae of the sought-for probabilities for a generic pdf. Later on, we shall specify

a class of pdfs encountered in practice, we shall calculate the corresponding probabilities and present the associated numeric results. At this point, since $|x|, |y| \in [1, 10)$, we form the square shown in Figure 1, where every mantissae couple (x, y) corresponds to a certain point of the sub-domain

$$J = (A\Lambda_1 T_1 \Lambda_8 A) \cup (B\Lambda_3 T_2 \Lambda_2 B) \cup (\Gamma\Lambda_5 T_3 \Lambda_4 \Gamma) \cup (\Delta\Lambda_7 T_4 \Lambda_6 \Delta)$$

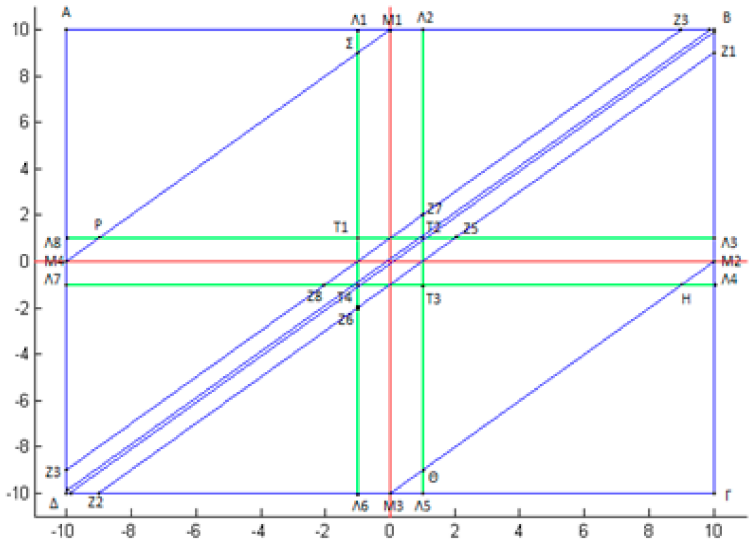

Figure 1. Geometric representation of all pairs of finite precision error mantissae (x, y). Since these pairs do not belong to the cross $\Lambda_1\Lambda_2 T_2\Lambda_3\Lambda_4 T_3\Lambda_5\Lambda_6 T_4\Lambda_7\Lambda_8 T_1\Lambda_1$, the corresponding joint probability function is restricted within $J = (A\Lambda_1 T_1 \Lambda_8 A) \cup (B\Lambda_3 T_2 \Lambda_2 B) \cup (\Gamma\Lambda_5 T_3 \Lambda_4 \Gamma) \cup (\Delta\Lambda_7 T_4 \Lambda_6 \Delta)$.

We point out that the joint probability $f_{XY}(x, y)$ is a conditional pdf, where $(x, y) \in J$, in the sense that it satisfies relation $\iint_J f_{XY}(x, y)dxdy = 1$. If the initial pdf $f^I_{XY}(x, y)$ is defined in a superset of J, then, we restrict it to J by means of the conditional probability rule. Notice that the points of the "cross" $C = \Lambda_1\Lambda_2 T_2\Lambda_3\Lambda_4 T_3\Lambda_5\Lambda_6 T_4\Lambda_7\Lambda_8 T_1\Lambda_1$ do not belong to J, since x and y are mantissae. We again distinguish the sub-cases introduced in Section 3.1.

Case 1.i: $|man(\alpha_n)man(\beta_n)| < 10$, namely relation (3.4).

Case 1.i.a: $100 \leq |\alpha_n x + \beta_n y| < UB$, which is (3.6).

In order to determine the sub-domain of J, where inequality (3.6) holds, we assume, first, that both α_n, β_n are positive mantissae and we draw the straight lines:

$$E_{100} : \alpha_n x + \beta_n y = 100,\ E_{ub} : \alpha_n x + \beta_n y = UB.$$

Let P^i_a be the set of points (x, y) of J that lie between E_{100} and E_{ub}, where superscript i and subscript a express the last two letters of the Case in hand. Further, consider the straight lines:

$$E_{-100} : \alpha_n x + \beta_n y = -100,\ E_{-ub} : \alpha_n x + \beta_n y = -UB.$$

Let N^i_a be the set of points (x, y) of J lying between E_{-100} and E_{-ub} and $D^i_a = P^i_a \cup N^i_a$; D^i_a includes all points of J satisfying (3.6). Then, probability $P^{EQ}\{(x, y) \in D^i_a\} = \iint_{D^i_a} f_{XY}(x, y)dxdy$. However, in this case only, according to the analysis of Section 3.1, $\gamma_n = \alpha_n\beta_n$ is computed with $\xi = 2$ additional e. d. d. than λ. Hence,

$$P^{EQ}(2; \alpha_n, \beta_n) = \iint_{D^i_a} f_{XY}(x, y)dxdy. \tag{3.36}$$

Case 1.i.b: $10 \leq |\alpha_n x + \beta_n y| < 100$, namely inequality (3.8).
For an arbitrary pair of multiplication operands $(\alpha_n,\ \beta_n)$, consider, now, the straight lines:

$$E_{10} : \alpha_n x + \beta_n y = 10 \text{ and } E_{-10} : \alpha_n x + \beta_n y = -10.$$

Let P_b^i be the set of points $(x, y) \in J$ lying between E_{100} and E_{10} and N_b^i be the set of points $(x, y) \in J$ lying between E_{-100} and E_{-10}. $D_b^i = P_b^i \cup N_b^i$ is the entire ensemble of points in J satisfying (3.8), depicted in magenta in Figure 2. Then, probability

$$P^{EQ}\Big\{(x, y) \in D_b^i\Big\} \equiv P^{EQ}(1; \alpha_n, \beta_n) = \iint\limits_{D_b^i} f_{XY}(x, y)dxdy. \tag{3.37}$$

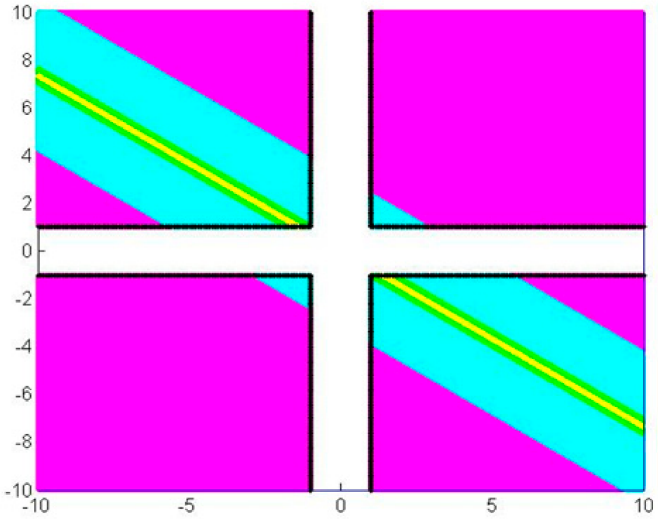

Figure 2. Depiction of the various sub-domains of J, which give rise to different numbers of erroneous decimal digits of $\gamma_5 = \alpha_5\beta_5$, where $|man(\alpha_5)man(\beta_5)| < 10$. In this example, we have selected $\alpha_5 = 2.3912$ and $\beta_5 = 3.2578$; consequently: (a) the sub-region generating one additional e. d. d. is depicted in magenta, (b) the sub-domain that does not increase the f. p. error is shown in cyan (c) the sub-region relaxing the e. d. d. number by one is depicted in green and (d) the one relaxing the number of e. d. d. by two is shown in yellow. Sub-domains that represent an even greater relaxation of the f. p. e. are too small to appear.

Case 1.i.c: $1 \leq |\alpha_n x + \beta_n y| < 10$, that is (3.10).
Next, in accordance with the previous analysis, we draw the straight lines:

$$E_1 : \alpha_n x + \beta_n y = 1 \text{and} E_{-1} : \alpha_n x + \beta_n y = -1.$$

Then, P_c^i is the sub-domain of J bounded by E_1 and E_{10}, while N_c^i is the sub-region bounded by E_{-1} and E_{-10}. Setting $D_c^i = P_c^i \cup N_c^i$ (cyan area in Figure 2), the probability that a pair (x, y) of error mantissae satisfies (3.10) is:

$$P^{EQ}\Big\{(x, y) \in D_c^i\Big\} \equiv P^{EQ}(0; \alpha_n, \beta_n) = \iint\limits_{D_c^i} f_{XY}(x, y)dxdy. \tag{3.38}$$

Finally, concerning the remaining Case 1.i.d, it holds that:
Case 1.i.d: $10^{-k} \leq |\alpha_n x + \beta_n y| < 10^{1-k}$, $k = 1, 2, 3$, namely the condition (3.14).
With a similar reasoning, we define the lines $E_{0.1} : \alpha_n x + \beta_n y = 10^{-1}$, $E_{0.01} : \alpha_n x + \beta_n y = 10^{-2}$, $E_{0.001} : \alpha_n x + \beta_n y = 10^{-3}$, $E_{-0.1} : \alpha_n x + \beta_n y = -10^{-1}$, $E_{-0.01} : \alpha_n x + \beta_n y = -10^{-2}$, $E_{-0.001} : \alpha_n x + \beta_n y = -10^{-3}$ which in turn give rise to the sub-domains D_{d1}^i,

D^i_{d2} and D^i_{d3}. Sub-domains D^i_{d1} and D^i_{d2} are depicted in green and yellow respectively in Figure 2. Eventually, the corresponding probabilities are

$$\begin{aligned} P^{EQ}\{(x,y) \in D^i_{d1}\} &\equiv P^{EQ}(-1;\alpha_n,\beta_n) = \iint_{D^i_{d1}} f_{XY}(x,y)dxdy \\ P^{EQ}\{(x,y) \in D^i_{d2}\} &\equiv P^{EQ}(-2;\alpha_n,\beta_n) = \iint_{D^i_{d2}} f_{XY}(x,y)dxdy \\ P^{EQ}\{(x,y) \in D^i_{d3}\} &\equiv P^{EQ}(-3;\alpha_n,\beta_n) = \iint_{D^i_{d3}} f_{XY}(x,y)dxdy. \end{aligned} \tag{3.39}$$

Case 1.ii: $|man(\alpha_n)man(\beta_n)| \geq 10$, specifically inequality (3.16).

This case may be treated as Case 1.i; however, here, as stated in Section 3.1, the computing machine performs a right shift in order to restore the product $\gamma_n = \alpha_n\beta_n$ in its canonical form. Therefore, product γ_n is computed with a number of e.d.d. reduced by one with respect to the previous Case 1.i. Thus, briefly, we note the following:

Case 1.ii.a: $100 \leq |\alpha_n x + \beta_n y| < 200$, which is (3.17).

Once again, lines E_{100} and E_{ub}, confine $P^{ii}_a \subset J$, and lines E_{-100} and E_{-ub}, confine $N^{ii}_a \subset J$. Let, again, $\mathcal{D}^{ii}_a = P^{ii}_a \cup N^{ii}_a$ (shown in magenta in Figure 3, for a specific pair of multiplication operands $(\alpha_n,\ \beta_n)$). When $(x,y) \in \mathcal{D}^{ii}_a$, product γ_n is computed with exactly one additional e. d. d. with probability

$$P^{EQ}\left\{(x,y) \in \mathcal{D}^{ii}_a\right\} \equiv P^{EQ}(1;\alpha_n,\beta_n) = \iint\limits_{\mathcal{D}^{ii}_a} f_{XY}(x,y)dxdy. \tag{3.40}$$

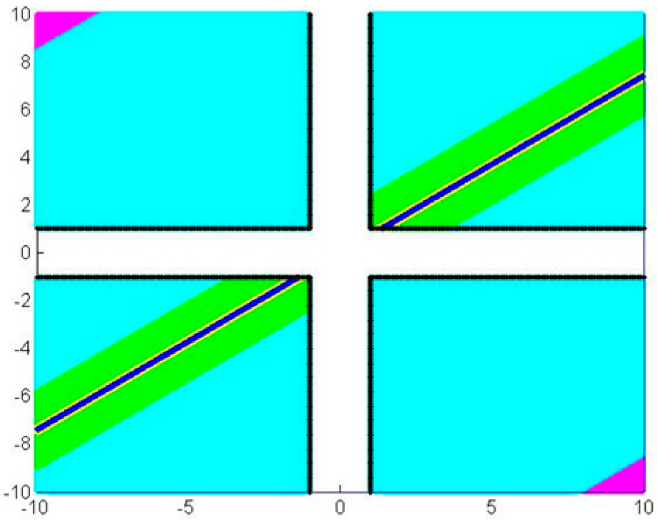

Figure 3. Depiction of the various sub-domains of J, which give rise to different numbers of erroneous decimal digits of $\gamma_5 = \alpha_5\beta_5$, where $|man(\alpha_5)man(\beta_5)| \geq 10$. In this example, we have selected $\alpha_5 = 4.6812$ and $\beta_5 = -6.3178$: therefore (a) the sub-domain generating one additional e. d. d. is depicted in magenta, (b) the sub-region that does not increase the f. p. error is shown in cyan, (c) the sub-domain relaxing the e. d. d. number by one is depicted in green, (d) the one relaxing the number of e. d. d. by two is shown in yellow, while (e) the one relaxing the number of e. d. d. by three is depicted in blue. The sub-domains that represent an even greater relaxation of the f. p. e. are too small to appear.

Case 1.ii.b: $10 \leq\ |\alpha_n x + \beta_n y| < 100$, i.e., (3.20).

We draw the straight lines E_{100}, E_{10} to obtain P^{ii}_b, lines E_{-100}, E_{-10} to confine N^{ii}_b and we let $\mathcal{D}^{ii}_b = P^{ii}_b \cup N^{ii}_b$ (shown in cyan in Figure 3). The probability is

$$P^{EQ}(0;\alpha_n,\beta_n) \equiv P^{EQ}\left\{(x,y) \in \mathcal{D}^{ii}_b\right\} = \iint\limits_{\mathcal{D}^{ii}_b} f_{XY}(x,y)dxdy. \tag{3.41}$$

Case 1.ii.c: $1 \leq |\alpha_n x + \beta_n y| < 10$, that is condition (3.22).

We select points $(x, y) \in J$ lying between E_{10} and E_1, forming P_c^{ii} and points $(x, y) \in J$ lying between E_{-10} and E_{-1}, forming N_c^{ii}; again, we let $\mathcal{D}_c^{ii} = P_c^{ii} \cup N_c^{ii}$ (green area in Figure 3). Now, the probability that f. p. e. is relaxed by one digit is

$$P^{EQ}\left\{(x, y) \in \mathcal{D}_c^{ii}\right\} \equiv P^{EQ}(-1; \alpha_n, \beta_n) = \iint\limits_{\mathcal{D}_c^{ii}} f_{XY}(x, y)dxdy. \tag{3.42}$$

Case 1.ii.d: $10^{-k} \leq |\alpha_n x + \beta_n y| < 10^{-(k-1)}$, $k = 1, 2$, namely the condition (3.24).

Along very similar lines we define sub-domains of J, $\mathcal{D}_{d1}^{ii}$ and $\mathcal{D}_{d2}^{ii}$ (shown in yellow and blue respectively in Figure 3 for a specific pair of operands (α_n, β_n) and $n = 5$). We eventually evaluate

$$\begin{array}{l} P^{EQ}\{(x, y) \in \mathcal{D}_{d1}^{ii}\} \equiv P^{EQ}(-2; \alpha_n, \beta_n) = \iint_{\mathcal{D}_{d1}^{ii}} f_{XY}(x, y)dxdy \\ P^{EQ}\{(x, y) \in \mathcal{D}_{d2}^{ii}\} \equiv P^{EQ}(-3; \alpha, \beta) = \iint_{\mathcal{D}_{d2}^{ii}} f_{XY}(x, y)dxdy. \end{array} \tag{3.43}$$

Case 2: $\lambda_a \neq \lambda_\beta$.

Suppose that without any loss of generality $\lambda_a < \lambda_\beta \left(\Leftrightarrow \lambda_a^c > \lambda_\beta^c\right)$ and that λ_γ is the number of erroneous d. d. with which product $\gamma_n = \alpha_n \beta_n$; let, moreover, $\xi = \lambda_\gamma - \lambda_\beta$.

Notation 3. *In this case, the f. p. error also depends on δ (see Section 3.1). Hence, for the corresponding probability, we use the notation $P^{UN}(\xi, \delta; \alpha_n, \beta_n)$, where as always, without any loss of generality, we assume that α_n and β_n are mantissae and that $\lambda_a < \lambda_\beta$. If the opposite inequality $\lambda_\alpha > \lambda_\beta$ holds, then we use the notation $P^{UN}(\xi, \delta; \beta_n, \alpha_n)$.*

We once more consider straight lines $E_{100}^{\delta}, E_{-100}^{\delta}$, $E_{10}^{\delta}, E_{-10}^{\delta}$, E_1^{δ}, E_{-1}^{δ}, $E_{0.1}^{\delta}$, $E_{-0.1}^{\delta}$, $E_{0.01}^{\delta}$, $E_{-0.01}^{\delta}$, $E_{0.001}^{\delta}, E_{-0.001}^{\delta}$, which confine the corresponding sub-domains of J: $\mathcal{G}_a^i$, $\mathcal{G}_b^i$, $\mathcal{G}_c^i$, $\mathcal{G}_{d1}^i$, $\mathcal{G}_{d2}^i$, $\mathcal{G}_{d3}^i$, $\mathcal{G}_a^{ii}$, $\mathcal{G}_b^{ii}$, $\mathcal{G}_c^{ii}$, $\mathcal{G}_{d1}^{ii}$, $\mathcal{G}_{d2}^{ii}$. The probabilities that a pair of mantissae (x, y) lies in one of the aforementioned domains are:

Case 2.i: $|man(\alpha_n) man(\beta_n)| < 10$, namely condition (3.4).

Case 2.i.a: $100 \leq \left|\alpha_n x + \beta_n y 10^{-\delta}\right|$, i.e., (3.28).

$$P^{UN}\left\{(x, y) \in G_\alpha^i\right\} = \iint\limits_{G_\alpha^i} f_{XY}(x, y)dxdy = P^{UN}(2, \delta; a_n, \beta_n) \tag{3.44}$$

Case 2.i.b: $10 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 100$, that corresponds to (3.29)

$$P^{UN}\left\{(x, y) \in G_b^i\right\} = \iint\limits_{G_b^i} f_{XY}(x, y)dxdy = P^{UN}(1, \delta; a_n, \beta_n). \tag{3.45}$$

Case 2.i.c: $1 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10$, which is the one of (3.30)

$$P^{UN}\left\{(x, y) \in G_c^i\right\} \equiv \iint\limits_{G_c^i} f_{XY}(x, y)dxdy = P^{UN}(0, \delta; \alpha_n, \beta_n). \tag{3.46}$$

Case 2.i.d: $10^{-k} \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10^{-(k-1)}$, $k = 1, 2, 3$, i.e. (3.31)

$$\begin{array}{l} P^{UN}\{(x, y) \in G_{d1}^i\} \equiv P^{UN}(-1, \delta; \alpha_n, \beta_n) = \iint_{G_{d1}^i} f_{XY}(x, y)dxdy \\ P^{UN}\{(x, y) \in G_{d2}^i\} \equiv P^{UN}(-2, \delta; \alpha_n, \beta_n) = \iint_{G_{d2}^i} f_{XY}(x, y)dxdy \\ P^{UN}\{(x, y) \in G_{d3}^i\} \equiv P^{UN}(-3, \delta; \alpha_n, \beta_n) = \iint_{G_{d3}^i} f_{XY}(x, y)dxdy. \end{array} \tag{3.47}$$

Case 2.ii: $|man(\alpha_n) man(\beta_n)| \geq 10$ (3.16).

Case 2.ii.a: $100 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 200$, i.e., the inequality of (3.32).

$$P^{UN}\left\{(x,y) \in \mathcal{G}_a^{ii}\right\} \equiv P^{UN}(1,\delta;\alpha_n,\beta_n) = \iint\limits_{\mathcal{G}_a^{ii}} f_{XY}(x,y)dxdy. \tag{3.48}$$

Case 2.ii.b: $10 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 100$, that corresponds to (3.33)

$$P^{UN}\left\{(x,y) \in \mathcal{G}_b^{ii}\right\} \equiv P^{UN}(0,\delta;\alpha_n,\beta_n) = \iint\limits_{\mathcal{G}_b^{ii}} f_{XY}(x,y)dxdy. \tag{3.49}$$

Case 2.ii.c: $1 \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10$, namely inequality (3.34)

$$P^{UN}\left\{(x,y) \in \mathcal{G}_c^{ii}\right\} \equiv P^{UN}(-1,\delta;\alpha_n,\beta_n) = \iint\limits_{\mathcal{G}_c^{ii}} f_{XY}(x,y)dxdy. \tag{3.50}$$

Case 2.ii.d: $10^{-k} \leq \left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right| < 10^{-(k-1)}$, $k = 1,2,3$, i.e. the case of (3.35)

$$\begin{array}{l} P^{UN}\{(x,y) \in \mathcal{G}_{d1}^{ii}\} \equiv P^{UN}(-2,\delta;\alpha_n,\beta_n) = \iint_{\mathcal{G}_{d1}^{ii}} f_{XY}(x,y)dxdy \\ P^{UN}\{(x,y) \in \mathcal{G}_{d2}^{ii}\} \equiv P^{UN}(-3,\delta;\alpha_n,\beta_n) = \iint_{\mathcal{G}_{d2}^{ii}} f_{XY}(x,y)dxdy. \end{array} \tag{3.51}$$

3.3. Experimental Confirmation of the Previous Theoretical Results

In order to test the validity of the analysis and the results of previous Sections 3.1 and 3.2, we have performed the following experiment:

First, we have chosen a set Σ^{16} consisting of 100,000 couples of randomly chosen mantissae $\left(\alpha_{16}^i, \beta_{16}^i\right)$, having $n = 16$ d. d. in the mantissa. We assume that these numbers are all correct, concerning the first 16 d. d.

We have "contaminated" all $\left(\alpha_{16}^i, \beta_{16}^i\right)$, each one with a different error obtained from a normal population, with various values σ of the std. In fact, for each σ, we have produced 25,000 normally distributed error values $\left(\theta_{\alpha,i}^N, \theta_{\beta,i}^N\right)$ that will play the role of f. p. e. of α_{16}^i and β_{16}^i, should all operations had been made with $n = 16$ d. d. precision and the set of contaminated pairs $\left(\widetilde{\alpha}_{16}^i, \widetilde{\beta}_{16}^i\right) = \left(\alpha_{16}^i + \theta_{\alpha,i}^N, \beta_{16}^i + \theta_{\beta,i}^N\right)$. In addition, we have extended $\widetilde{\alpha}_{16}^i$ and $\widetilde{\beta}_{16}^i$ into a representation of $n = 64$ d. d., by simply zeroing all decimal digits from the seventeenth one up to 64-th digit.

We have performed all multiplications $\widetilde{\gamma}_{16}^i = \widetilde{\alpha}_{16}^i \cdot \widetilde{\beta}_{16}^i$, evidently in 16 d. d. precision, as well as multiplications $\widetilde{\gamma}_{64}^i = \widetilde{\alpha}_{64}^i \cdot \widetilde{\beta}_{64}^i$, in 64 d. d. precision. Then, using Definition 2 and Theorem 5, we have obtained the number of e. d. d. of quantity $\widetilde{\gamma}_{16}^i$, with respect to the e. d. d. of $\widetilde{\alpha}_{16}^i$ and $\widetilde{\beta}_{16}^i$ and the set of f. p. e. differences $\#edd\left(\widetilde{\gamma}_i^{16}\right) - max\left\{\#edd\left(\widetilde{\alpha}_{16}^i\right), \#edd\left(\widetilde{\beta}_{16}^i\right)\right\}$. Using this set, we have compared the corresponding experimental frequencies with the theoretical probabilities predicted in the present section, for various standard deviations of f. p. e. Representative results are shown in Tables 3–6; Table 4 refers to the case where $\#edd\left(\widetilde{\alpha}_{16}^i\right) = \#edd\left(\widetilde{\beta}_{16}^i\right)$, for four arbitrarily chosen pairs (α_i, β_i), shown in Table 3. On the contrary, Tables 5 and 6 refer to the case where $\#edd\left(\widetilde{\beta}_{16}^i\right) - \#edd\left(\widetilde{\alpha}_{16}^i\right) < 0$. Table 5 corresponds to the case in which $\delta = 1$, while Table 6 corresponds to the one in which $\delta = 2$. From both tables, the excellent agreement between theory and experiment is pretty evident.

Table 3. Four arbitrarily chosen pairs of $\left(\alpha_{16}^{i}, \beta_{16}^{i}\right)$.

$\boldsymbol{\alpha_{16}^{i}}$	$\boldsymbol{\beta_{16}^{i}}$
$\alpha_1 = 1.505791937075619$	$\beta_1 = 5.526986816293506$
$\alpha_2 = 2.675404049994100$	$\beta_2 = 2.778498218867048$
$\alpha_3 = 7.946881519204984$	$\beta_3 = 4.557506835434298$
$\alpha_4 = 5.557116785741900$	$\beta_4 = 5.549129305868777$

Table 4. The theoretical probabilities $P^{EQ}\left(\xi; \widetilde{\alpha}_{16}^{i}, \widetilde{\beta}_{16}^{i}\right)$ numerically evaluated, as compared to the actually observed corresponding experimental frequencies of $\xi = \#edd\left(\widetilde{\gamma}_{16}^{\iota}\right) - \#edd\left(\widetilde{\alpha}_{16}^{\iota}\right)$.

$\boldsymbol{P^{EQ}\left(\xi; \tilde{\alpha}_{16}^{i}, \tilde{\beta}_{16}^{i}\right)}$ **Verification. Normally Distributed Contamination σ = 2.2**					
$\left(\tilde{\alpha}_{16}^{i}, \tilde{\beta}_{16}^{i}\right)$		$\left(\tilde{\alpha}_{16}^{1}, \tilde{\beta}_{16}^{1}\right)$	$\left(\tilde{\alpha}_{16}^{2}, \tilde{\beta}_{16}^{2}\right)$	$\left(\tilde{\alpha}_{16}^{3}, \tilde{\beta}_{16}^{3}\right)$	$\left(\tilde{\alpha}_{16}^{4}, \tilde{\beta}_{16}^{4}\right)$
Generation of 2 Additional e. d. d.	Theoretical Probability	0	0	0	0
	Experimental Frequency	0	0	0	0
Generation of 1 Additional e. d. d.	Theoretical Probability	0.648	0.373	10^{-5}	$3.74 \cdot 10^{-9}$
	Experimental Frequency	0.647	0.389	10^{-5}	0
Generation of No Additional e. d. d.	Theoretical Probability	0.338	0.517	0.702	0.622
	Experimental Frequency	0.339	0.503	0.704	0.623
Relaxation of Product's e. d. d. by 1	Theoretical Probability	0.012	0.099	0.264	0.325
	Experimental Frequency	0.012	0.096	0.262	0.321
Relaxation of Product's e. d. d. by 2	Theoretical Probability	0.001	0.009	0.031	0.046
	Experimental Frequency	0.001	0.010	0.030	0.049

Table 5. The theoretical probabilities $P^{UN}\left(\xi, 1; \widetilde{\alpha}_{16}^{i}, \widetilde{\beta}_{16}^{i}\right)$, namely for the case where $\#edd\left(\widetilde{\beta}_{16}^{i}\right) - \#edd\left(\widetilde{\alpha}_{16}^{i}\right) = 1$, numerically evaluated, as compared to the actually observed corresponding experimental frequencies of $\xi = \#edd\left(\widetilde{\gamma}_{16}^{\iota}\right) - \#edd\left(\widetilde{\beta}_{16}^{\iota}\right)$.

$\boldsymbol{P^{UN}\left(\xi; 1; \tilde{\alpha}_{16}^{i}, \tilde{\beta}_{16}^{i}\right)}$ **Verification. Normally Distributed Contamination σ = 2.2**					
$\left(\tilde{\alpha}_{16}^{i}, \tilde{\beta}_{16}^{i}\right)$		$\left(\tilde{\alpha}_{16}^{1}, \tilde{\beta}_{16}^{1}\right)$	$\left(\tilde{\alpha}_{16}^{2}, \tilde{\beta}_{16}^{2}\right)$	$\left(\tilde{\alpha}_{16}^{3}, \tilde{\beta}_{16}^{3}\right)$	$\left(\tilde{\alpha}_{16}^{4}, \tilde{\beta}_{16}^{4}\right)$
Generation of 2 Additional e. d. d.	Theoretical Probability	0	0	0	0
	Experimental Frequency	0	0	0	0

Table 5. *Cont.*

$P^{UN}\left(\xi;1;\tilde{\alpha}_{16}^{i},\tilde{\beta}_{16}^{i}\right)$ **Verification. Normally Distributed Contamination σ = 2.2**					
$\left(\tilde{\alpha}_{16}^{i},\tilde{\beta}_{16}^{i}\right)$		$\left(\tilde{\alpha}_{16}^{1},\tilde{\beta}_{16}^{1}\right)$	$\left(\tilde{\alpha}_{16}^{2},\tilde{\beta}_{16}^{2}\right)$	$\left(\tilde{\alpha}_{16}^{3},\tilde{\beta}_{16}^{3}\right)$	$\left(\tilde{\alpha}_{16}^{4},\tilde{\beta}_{16}^{4}\right)$
Generation of 1 Additional e. d. d.	Theoretical Probability	0.008	0.127	0	0
	Experimental Frequency	0.008	0.147	0	0
Generation of No Additional e. d. d.	Theoretical Probability	0.870	0.873	0.867	0.631
	Experimental Frequency	0.872	0.853	0.870	0.633
Relaxation of Product's e. d. d. by 1	Theoretical Probability	0.112	$2{\cdot}10^{-4}$	0.133	0.369
	Experimental Frequency	0.110	$1{\cdot}10^{-4}$	0.130	0.367
Relaxation of Product's e. d. d. by 2	Theoretical Probability	0.009	0	0	$4.23{\cdot}10^{-6}$
	Experimental Frequency	0.009	0	0	0

Table 6. The theoretical probabilities $P^{UN}\left(\xi,2;\tilde{\alpha}_{16}^{i},\tilde{\beta}_{16}^{i}\right)$, namely for the case where $\#edd\left(\tilde{\beta}_{16}^{i}\right) - \#edd\left(\tilde{\alpha}_{16}^{i}\right) = 2$, numerically evaluated, as compared to the actually observed corresponding experimental frequencies of $\xi = \#edd\left(\tilde{\gamma}_{16}^{\iota}\right) - \#edd\left(\tilde{\beta}_{16}^{\iota}\right)$.

$P^{UN}\left(\xi,2;\tilde{\alpha}_{16}^{i},\tilde{\beta}_{16}^{i}\right)$ **Verification. Normally Distributed Contamination σ = 2.2**					
$\left(\tilde{\alpha}_{16}^{i},\tilde{\beta}_{16}^{i}\right)$		$\left(\tilde{\alpha}_{16}^{1},\tilde{\beta}_{16}^{1}\right)$	$\left(\tilde{\alpha}_{16}^{2},\tilde{\beta}_{16}^{2}\right)$	$\left(\tilde{\alpha}_{16}^{3},\tilde{\beta}_{16}^{3}\right)$	$\left(\tilde{\alpha}_{16}^{4},\tilde{\beta}_{16}^{4}\right)$
Generation of 2 Additional e. d. d.	Theoretical Probability	0	0	0	0
	Experimental Frequency	0	0	0	0
Generation of 1 Additional e. d. d.	Theoretical Probability	0.004	0.123	0	0
	Experimental Frequency	0.003	0.134	0	0
Generation of No Additional e. d. d.	Theoretical Probability	0.996	0.877	0.865	0.626
	Experimental Frequency	0.997	0.866	0.872	0.637
Relaxation of Product's e. d. d. by 1	Theoretical Probability	0	0	0.135	0.374
	Experimental Frequency	0	0	0.128	0.363
Relaxation of Product's e. d. d. by 2	Theoretical Probability	0	0	0	0
	Experimental Frequency	0	0	0	0

We have repeated the previous step, using uniformly distributed "contamination numbers" θ_i^U, in the interval $[10^{-16}, 10^{-3}]$, producing numbers $\hat{\alpha}_{16}^i = \alpha_{16}^i + \theta_i^U$.

By repeating all actions of previous Step 3 for the case of uniform contamination, the obtained results have confirmed an excellent agreement between theory and practice.

A set of concrete experiments and associated tables.

We have randomly chosen 25,000 couples of mantissae terms $(\alpha_{16}^i, \beta_{16}^i)$, covering all cases referred to in Sections 3.1 and 3.2. We have embedded both α_{16}^i and β_{16}^i in 64 d. d. precision, as described previously in the present sub-section, thus forming corresponding couples $(\alpha_{64}^i, \beta_{64}^i)$. We have contaminated each such pair $(\alpha_{16}^i, \beta_{16}^i)$ with 25,000 normally distributed error values for various distinct values of standard deviation σ. In this way, we have generated 25,000 corresponding contaminated pairs $\left(\widetilde{\alpha}_{16}^i, \widetilde{\beta}_{16}^i\right)$.

We have performed all 25,000 multiplications $\widetilde{\gamma}_{16}^i = \widetilde{\alpha}_{16}^i \cdot \widetilde{\beta}_{16}^i$, as well as the associated products $\gamma_{64}^i = \alpha_{64}^i \cdot \beta_{64}^i$ and, finally, we have evaluated the number of erroneous decimal digits of $\widetilde{\gamma}_{16}^i$ by comparing it with γ_{64}^i.

The results of this experiment for a specific value of σ are shown in Tables 3–6. In Table 3, four arbitrarily chosen different pairs $(\alpha_{16}^i, \beta_{16}^i)$, $i = 1, 2, 3, 4$ are presented. Table 4 refers to the case where $\#edd(\widetilde{\alpha}_{16}^i) = \#edd(\widetilde{\beta}_{16}^i)$, for the corresponding contaminated pairs $\left(\widetilde{\alpha}_{16}^i, \widetilde{\beta}_{16}^i\right)$, while Tables 5 and 6 refer to the cases in which $\delta = 1$ and $\delta = 2$ respectively, where $\delta = \#edd\left(\widetilde{\beta}_{16}^i\right) - \#edd(\widetilde{\alpha}_{16}^i)$.

For all arbitrarily chosen contaminated pairs $\left(\widetilde{\alpha}_{16}^i, \widetilde{\beta}_{16}^i\right)$, we have evaluated the theoretical probabilities introduced in Section 3.1, numerically. From all tables, the excellent agreement between theory and experiment is pretty evident. We would like to point out that this excellent agreement appears in all performed experiments, concerning $10 - ths$ of different values of standard deviation σ.

4. Analysis of the Case of Many Successive Multiplications

In this section, we will compute the probability that M successive multiplications generate λ erroneous d. d. in the final product.

In fact, suppose that any two numbers, γ_n^0 and β_n^0, are multiplied in a computing machine using n decimal digits (d. d.) in the mantissa; let $\gamma_n^1 = \gamma_n^0 \beta_n^0$. Next, γ_n^1 is multiplied by an arbitrary number, say β_n^1, giving rise to $\gamma_n^2 = \gamma_n^1 \beta_n^1$ and so on. The analysis of Section 3 indicates that a different number of erroneous d. d. emerges as it is analytically presented in Tables 1 and 2. Therefore, in order to estimate the number of erroneous decimal digits (e. d. d.) accumulated in a result of many successive multiplications, one may employ the following:

1. The mantissa y of the finite precision error (f. p. e.) accumulated at an arbitrary quantity, say α, is a random variable, already symbolized as Y. Therefore, when two quantities α_n and β_n are multiplied with f. p. e. mantissae x and y, then the f. p. error of the product $\gamma_n = a_n \beta_n$ is itself a random variable.
2. As before, without any loss of generality, suppose that λ_β is the maximum number of e. d. d. between α_n and β_n. Then, reminding that the symbol "#" stands for cardinal number, $\#edd(\gamma_n)$ differ from λ_β by ξ e. d. d. Evidently, ξ is a random variable itself, having integer values $= 2, 1, 0, -1, -2, \ldots$.
3. In Section 3, we have given a method for evaluating the probabilities $P^{EQ}(\xi; \alpha_n, \beta_n)$, namely the probability that product γ_n is computed with a number of erroneous decimal digits (e. d. d.) differing by ξ decimal digits from the common e. d. d. of α_n *and* β_n. In the same section, we have also proposed a method for evaluating the probabilities $P^{UN}(\xi, \delta; \alpha_n, \beta_n)$, i.e., the probability that product γ_n is computed with a number of e. d. d. differing by ξ decimal digits from the maximum number of e. d. d. between α_n *and* β_n. For brevity, in the present section, we will assume that in all successive multiplications the worst case always takes place, namely that the two multiplication operands share the same number of correct decimal digits (c. d.

d.). Moreover, we will momentarily simplify notation by letting $P_2 = P^{EQ}(2; \alpha_n, \beta_n)$, $P_1 = P^{EQ}(1; \alpha_n, \beta_n)$, $P_{-k} = P^{EQ}(-k; \alpha_n, \beta_n)$, $k = 0, 1, 2, \ldots$.

4. We have performed an extensive number of multiplications $\gamma_n^{i+1} = \gamma_n^i \beta_n^i$, $i = 0, 1, 2, \ldots$, where, initially, γ_n^0 and β_n^0 are chosen uniformly from the interval $(-10,\ 10)$ and with n d. d. precision in the mantissa. Then, the f. p. e. mantissa x of the product γ_n^{i+1} follows a normal distribution with zero (0) mean value and standard deviation $\sigma \in [1.1,\ 6.5]$. Hence, probabilities P_j, $j = 2, 1, 0, -1, -2, \ldots$ are immediately obtained via the analysis of Section 3. However, the present analysis is valid for any distribution of error mantissae that gives rise to a set of probabilities P_j.
5. For brevity and simplicity reasons, we shall assume that the ensemble of probabilities P_j remains unaltered throughout the entire successive multiplications process. Should any concern on that arise, as we will explicitly state below, a proper source code may be used in order to compute $P^{EQ}(\xi; \alpha_n, \beta_n)$ dynamically, while the essence of the following analysis remains intact.

Subsequently, we will compute the probability that M successive multiplications generate λ erroneous d. d. in the final product γ_n^M. In fact, suppose that one performs M successive multiplications and that i_1 of them produce two additional e. d. d. ($\xi = 2$), i_2 of them produce one additional e. d. d. ($\xi = 1$), i_3 of them produce no additional e. d. d. ($\xi = 0$) and i_k products "enjoyed" relaxation of the number of e. d. d. by k digits ($\xi = -k$). Then, the number ω of e. d. d. with which the final product of M successive multiplications is obtained, is given by $\omega = \sum_\mu \xi_\mu i_\mu$. We are interested in the mean value and variation of quantity ω. To achieve that, we shall present a set of quite general lemmas and theorems; for this reason, for the present section only, we shall introduce an alternative, equivalent notation described below:

Notation 4. *Let ξ be defined as in the previous analysis above. Then, one may define events $A_1,\ A_2,\ A_3,\ \ldots$ as follows: $A_1 : \xi = 2$, $A_2 : \xi = 1$, $A_3 : \xi = 0$, $A_{3+\kappa} : \xi = -\kappa$, $\kappa = 1, 2, \ldots$, where events $A_{3+\kappa}$ refer to error correction by κ digits.*

Intimately associated with Notation 4 is the following:

Hypothesis 1. *In order to obtain proper bounds of the number λ of e. d. d. accumulated in the final product of M successive multiplications, it is sufficient to assume that at each one of these M successive multiplications, the corresponding probabilities $P^{EQ}(\xi; \gamma_n, \beta_n)$ remain constant. Under this assumption, we let:*

$$P_1 = P^{EQ}(2; \gamma_n, \beta_n) = P(A_1), P_2 = P^{EQ}(1; \gamma_n, \beta_n) = P(A_2),$$
$$P_{3+k} = P^{EQ}(-\kappa; \gamma_n, \beta_n) = P(A_{3+k}),\ k = 0, 1, 2, \ldots.$$

In order to obtain the aforementioned bounds for λ, we shall employ the subsequent quite general results.

Lemma 1. *Consider a multinomial distribution with possible outcomes $A_1,\ A_2,\ \ldots, A_N$, with corresponding probability of appearance $P_1,\ P_2,\ \ldots, P_N$. Suppose that one performs an experiment M times, whose outcome is modeled by this distribution. Let the first event with outcome A_1 be observed i_1 times, the second event with outcome A_2 i_2 times and so on. Then, quantity $\omega = A_1 i_1 + \ldots + A_N i_N = \sum_{\mu=1}^N A_\mu i_\mu$ has a mean value $\overline{\omega}$ and a variance S_ω^2:*

$$\overline{\omega} = M \sum_{\mu=1}^N A_\mu P_\mu. \tag{4.1}$$

$$S_\omega^2 = M\Big(\sum_{\mu=1}^N A_\mu^2 P_\mu - \sum_{\mu=1}^N A_\mu^2 P_\mu^2 - 2\sum_{i=1}^N \sum_{j=i+1}^N A_i A_j P_i P_j\Big). \tag{4.2}$$

Proof of Lemma 1. The probability that ω occurs is given by

$$P(\omega) = \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \left(\begin{array}{c} M - i_1 \\ i_2 \end{array} \right) P_2^{i_2} \dots P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})} . \tag{4.3}$$

For the mean value $\overline{\omega}$: By definition:

$$\overline{\omega} = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} [A_1 i_1 + A_2 i_2 + \dots + A_N i_N] P(\omega) \overset{(4.3)}{\Leftrightarrow}$$

$$\overline{\omega} = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} A_1 i_1 \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \left(\begin{array}{c} M - i_1 \\ i_2 \end{array} \right) P_2^{i_2} \left(\begin{array}{c} M - i_1 - i_2 \\ i_3 \end{array} \right) P_3^{i_3} \dots \left(\begin{array}{c} M - i_1 - \dots - i_{N-2} \\ i_{N-1} \end{array} \right) P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})} + \dots$$
$$+ \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} A_N i_N \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \left(\begin{array}{c} M - i_1 \\ i_2 \end{array} \right) P_2^{i_2} \left(\begin{array}{c} M - i_1 - i_2 \\ i_3 \end{array} \right) P_3^{i_3} \dots \left(\begin{array}{c} M - i_1 - \dots - i_{N-2} \\ i_{N-1} \end{array} \right) P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})}$$

We treat each multiple sum separately. Therefore,

$$\overline{\omega_1} = \sum_{i_1} \sum_{i_2} \dots \sum_{i_{N-1}} A_1 i_1 \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \left(\begin{array}{c} M - i_1 \\ i_2 \end{array} \right) P_2^{i_2} \dots P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})}$$
$$= \sum_{i_1=0}^{M} A_1 i_1 \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} \left(\begin{array}{c} M - i_1 - \dots - i_{N-2} \\ i_{N-1} \end{array} \right) P_{N-1}^{i_{N-1}} P_N^{M-i_1-\dots-i_{N-1}}$$
$$= A_1 P_1 M \sum_{i_1=1}^{M-1} \frac{(M-1)!}{(i_1-1)!(M-i_1)!} P_1^{i_1-1} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} \frac{(M-i_1-\dots-i_{N-2})!}{i_{N-1}!(M-i_1-\dots-i_{N-1})!} P_{N-1}^{i_{N-1}} P_N^{M-i_1-\dots-i_{N-1}}$$

i_{N-1}-sum is a version of the identity

$$(P_1 + P_2 + \dots + P_N)^M = \sum_{i_1=0}^{M} \left(\begin{array}{c} M \\ i_1 \end{array} \right) {P_1}^{i_1} (P_2 + \dots + P_N)^{M-i_1}.$$

Hence,

$$\overline{\omega}_1 = A_1 P_1 M . \tag{4.4}$$

By employing the same approach, we obtain the previous relation (4.1)

$$\overline{\omega} = \overline{\omega}_1 + \overline{\omega}_2 + \dots + \overline{\omega}_N \Rightarrow \overline{\omega} = M \sum_{\mu=1}^{N} A_\mu P_\mu .$$

For the variance S_ω^2: By definition:

$$S_\omega^2 = \sum_{\omega} \omega^2 P(\omega) - \overline{\omega}^2 . \tag{4.5}$$

By employing the previously given expression for ω, we obtain:

$$S_\omega^2 = \sum_{i_1=0}^{M} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} [A_1 i_1 + \dots + A_N i_N]^2 \cdot \left(\begin{array}{c} M \\ i_1 \end{array} \right) P_1^{i_1} \dots \left(\begin{array}{c} M - i_1 - \dots - i_{N-2} \\ i_{N-1} \end{array} \right) P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})} -$$
$$M^2 \left(\sum_{\mu=1}^{N} A_\mu P_\mu \right)^2 .$$

Expanding $[A_1 i_1 + \dots + A_N i_N]^2$, we obtain the partial sums:

$$s_1^2 = \sum_{i_1=0}^{M} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} {A_1}^2 {i_1}^2 \binom{M}{i_1} P_1^{i_1} \binom{M-i_1}{i_2} P_2^{i_2} \dots \binom{M-i_1-\dots-i_{N-2}}{i_{N-1}} P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})}$$

$$= {A_1}^2 \sum_{i_1=0}^{M} {i_1}^2 \binom{M}{i_1} P_1^{i_1} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} \binom{M-i_1-\dots-i_{N-1}}{i_N} P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})} \Leftrightarrow$$

$$s_1^2 = \sum_{i_1=0}^{M} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} {A_1}^2 {i_1}^2 \binom{M}{i_1} P_1^{i_1} \binom{M-i_1}{i_2} P_2^{i_2} \dots \binom{M-i_1-\dots-i_{N-1}}{i_N} P_{N-1}^{i_{N-1}} P_N^{M-(i_1+i_2+\dots+i_{N-1})}$$

$$= {A_1}^2 \sum_{i_1=0}^{M} i_1(i_1-1) \binom{M}{i_1} P_1^{i_1} (P_2+\dots+P_N)^{M-i_1} + {A_1}^2 \sum_{i_1=0}^{M} i_1 \binom{M}{i_1} P_1^{i_1} (P_2+\dots+P_N)^{M-i_1}$$

$$= {A_1}^2 M(M-1) P_1^2 \sum_{i_1=2}^{M-2} \frac{(M-2)!}{(i_1-2)!(M-i_1)!} P_1^{i_1-2} (P_2+\dots+P_N)^{M-i_1}$$

$$+ {A_1}^2 P_1 M \sum_{i_1=1}^{M-1} \frac{(M-1)!}{(i_1-1)!(M-i_1)!} P_1^{i_1-1} (P_2+\dots+P_N)^{M-i_1}$$

$$s_1^2 = {A_1}^2 M(M-1) P_1^2 + {A_1}^2 P_1 M$$

Following an analogous process for the other similar terms of quantity S_ω^2, we obtain

$$s_\mu^2 = {A_\mu}^2 M(M-1) P_\mu^2 + {A_\mu}^2 P_\mu M,\ \mu = 1,2,\dots,N$$

We now calculate the cross-product

$$s_{1,2}^2 = \sum_{i_1=0}^{M} \dots \sum_{i_{N-1}=0}^{M-i_1-\dots-i_{N-2}} 2A_1A_2 i_1 i_2 \binom{M}{i_1} P_1^{i_1} \binom{M-i_1}{i_2} P_2^{i_2} \dots \binom{M-i_1-\dots-i_{N-2}}{i_{N-1}} P_{N-1}^{i_{N-1}} P_N^{M-i_1-\dots-i_{N-1}} \Leftrightarrow$$

$$s_{1,2}^2 = 2A_1A_2P_1P_2M(M-1) \sum_{i_1=1}^{M-2} \binom{M-2}{i_1-1} P_1^{i_1-1} \sum_{i_2=1}^{M-i_1} \binom{M-i_1}{i_2-1} P_2^{i_2-1} (P_3+\dots+P_N)^{M-i_1-i_2} \Leftrightarrow$$

$$s_{1,2}^2 = 2A_1A_2P_2P_1M(M-1).$$

Similarly, for the remaining cross-products, we obtain

$$s_{i,j}^2 = 2A_iA_jP_iP_jM(M-1),\ i,j = 1,2,\dots,N,\ i \neq j.$$

Summing up s_μ^2 and $s_{i,j}^2$, we eventually obtain

$$S_\omega^2 = \sum_{\mu=1}^{N} s_\mu^2 + \sum_{\substack{i,j=1 \\ i<j}}^{N} s_{i,j}^2 - \overline{\omega}^2$$

$$\Leftrightarrow S_\omega^2 = M\left[\sum_{\mu=1}^{N} A_\mu^2 P_\mu - \sum_{\mu=1}^{N} A_\mu^2 P_\mu^2 - 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} A_iA_jP_iP_j\right].$$

□

This Lemma along with the central limit theorem, offer the following:

Lemma 2. *Suppose that one executes $M \geq 30$ successive multiplications. Then, the number of erroneous d. d. generated in the product obtained after these multiplications, $\omega = \sum_{\mu=1}^{N} A_\mu i_\mu$, follows a normal distribution with mean value $\overline{\omega}$ and variance S_ω^2 given by (4.1) and (4.2).*

We will apply all the previous results to the three more important cases, described below.
Case 1. The worst case, where in all multiplications $\left|man\left(\gamma_n^i\right) man\left(\beta_n^i\right)\right| < 10$ holds.
Case 2. The most favorable case, where $\left|man\left(\gamma_n^i\right) man\left(\beta_n^i\right)\right| \geq 10$ always holds.
Case 3. The general case, where the distribution of $\left|man\left(\gamma_n^i\right) man\left(\beta_n^i\right)\right|$ is arbitrary.

Case 1. If at each multiplication, inequality (3.4) holds, namely

$$|man(\alpha_n)man(\beta_n)| < 10,$$

then we choose the following values around which the corresponding probabilities are more frequently encountered:

1. $\xi = 2$: $P^{EQ}(2;\gamma_n^i,\beta_n^i) = P_1 \approx O(10^{-9})$ (i.e., almost negligible). We repeat that we use probabilities P^{EQ} only, since we consider the worst case as far as f. p. error generation and accumulation is concerned, namely that $\#edd(\gamma_n^i) = \#edd(\beta_n^i)$.
2. $\xi = 1$: $P^{EQ}(1;\gamma_n^i,\beta_n^i) = P_2 \approx 0.5530$.
3. $\xi = 0$: $P^{EQ}(0;\gamma_n^i,\beta_n^i) = P_3 \approx 0.3934$.
4. $\xi = -1$: $P^{EQ}(-1;\gamma_n^i,\beta_n^i) = P_4 \approx 0.0457$.
5. $\xi = -2$: $P^{EQ}(-2;\gamma_n^i,\beta_n^i) = P_5 \approx 7.8{\cdot}10^{-3}$.
6. $\xi = -3$: $P^{EQ}(-3;\gamma_n^i,\beta_n^i) = P_6 \approx O(10^{-6})$. (i.e., almost negligible).

Therefore, if M such successive multiplications take place, then, the overall number $\omega = \sum_{\mu=1}^{N} A_\mu i_\mu$ of generated e. d. d. follows a multinomial distribution, which may be very well approximated by a normal distribution with $\overline{\omega} = 0.4917 \cdot M$ and $S_\omega^2 = 0.3881 \cdot M$, $M \geq 30$. Hence, quantity $z = \frac{\omega - \overline{\omega}}{S_\omega}$ follows a standard Gauss distribution, i.e., $z \sim N(0,1)$.

However, now, inequality $|man(\alpha_n)man(\beta_n)| < 10$ (3.4) holds, thus, $\overline{\omega}$ is positive and inequality $\frac{\omega - \overline{\omega}}{S_\omega} \geq -4.2649$ holds with confidence 99.999%; coefficient -4.2649 corresponds to the aforementioned confidence level. With this level of significance, the accumulated number ω of erroneous d. d. in the final product, after M successive multiplications, obeying inequality (3.4), satisfies relation

$$\omega \geq -4.2649\, S_\omega + \overline{\omega}. \tag{4.6}$$

However, in this case, the right-hand side of inequality (4.6) is always positive and, moreover, is a monotonically increasing function of M. Consequently, the accumulated number ω of erroneous decimal digits of every product γ_n^i, $i = 1, 2, \ldots, M$, tends to rapidly increase even for a particularly small number of successive multiplications M. This is fully supported by the contents of Tables 7 and 8, below.

Table 7. F_g is a percentage of multiplications in which inequality (3.16) holds. Thus, when $F_g < 50\%$ holds, then γ_n^i becomes completely erroneous after a relatively small number of multiplications, in full accordance with the theoretical predictions. These predictions are based on the results of Theorems 1, 3 and 5, refer to the lower bound of the expected e. d. d. for each F_g and they are presented in the last column for confidence level $1 - 10^{-5}$. For each F_g the experimental and theoretical results manifest an excellent agreement.

Percentage of Successive Multiplications Satisfying $\lvert man(\alpha_n^i)man(\beta_n^i)\rvert \geq 10$.	**Number of Successive Multiplications after Which Product $\gamma_n^i = \alpha_n^i \cdot \beta_n^i$ Was Computed with All 16 d. d. Erroneous.**	**Theoretical Lower Bounds for $\#edd(\gamma_n^i)$ Obtained via Theorems 1, 3 and 5.**
0.0578	92	15.85
0.0718	89	14.00
0.0841	88	12.92
0.0926	90	12.93
0.1160	104	15.15
0.1360	107	14.40
0.1862	108	10.66
0.2310	131	11.58

Table 7. *Cont.*

Percentage of Successive Multiplications Satisfying $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right|\geq$10.	**Number of Successive Multiplications after Which Product γ_n^i=$\alpha_n^i\cdot\beta_n^i$ Was Computed with All 16 d. d. Erroneous.**	**Theoretical Lower Bounds for #$edd\left(\gamma_n^i\right)$ Obtained via Theorems 1, 3 and 5.**
0.2734	175	15.22
0.4053	328	12.90
0.4562	459	10.12
0.4591	489	11.46
0.4826	577	6.39

Table 8. Demonstration of the results of experiment associated with Case 3, described in the Section 4: 3×10^5 successive multiplications $\gamma_n^i=\alpha_n^i\cdot\beta_n^i$ have been performed for various percentages F_g of them satisfying $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right|\geq 10$. The obtained maximum and average numbers of e. d. d. are in full accordance with Theorems 1, 3 and 5. In fact, when $F_g>50\%$ holds, then the evaluated products manifest a considerable resistance to finite precision error. The closest to 1 F_g is, the smaller the number of erroneous digits with which all γ_n^i are computed, exactly as predicted by the theoretical analysis.

Percentage of Successive Multiplications Satisfying $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right|\geq$10	**Maximum Number of e. d. d. with Which Product γ_n^i=$\alpha_n^i\cdot\beta_n^i$ Has Been Computed, $i\leq3\cdot10^5$**	**Average Number of e. d. d. Accumulated in All Products γ_n^i=$\alpha_n^i\cdot\beta_n^i$, i=1,2,…,$3\cdot10^5$**
0.5014	13	7.7761
0.5029	12	7.2217
0.5051	11	6.9437
0.5063	11	6.8764
0.5076	11	6.6435
0.5131	10	6.1244
0.5143	10	6.0928
0.5154	10	6.0165
0.5171	10	5.9189
0.5204	10	5.7784
0.5260	9	5.5140
0.5300	9	5.4266
0.5404	9	5.1647
0.5481	9	5.0070
0.5588	8	4.8098
0.5899	8	4.4619
0.6182	8	4.1856
0.7742	6	3.4101
0.8422	6	3.0055
0.8702	5	2.9013
0.8913	5	2.8306
0.9195	5	2.7292
0.9412	5	2.6463
0.9519	4	2.6105

Theorem 1. *Let us assume that a number of M successive multiplications $\gamma_n^{i+1}=\gamma_n^i\beta_n^i, i=0,\ldots,M-1$ is performed and that for every multiplication, inequality $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right|<10$ holds. In this case, the product γ_n^i, $i=1,\ldots,M$ is prone to serious finite precision error accumulation. We also assume that Hypothesis 1 now holds. Let, in the N-th iteration the number ω, $\omega=\sum_{\mu=1}^{N}A_\mu i_\mu$, be the number of erroneous d. d. with which quantity γ_n^N has been evaluated. Then, it holds that*

$$\omega\geq C(\alpha)\cdot S_\omega+\overline{\omega}, \tag{4.7}$$

where α is the desired level of significance and $C(\alpha)$ is the lower bound of the corresponding confidence interval.

The theorem holds for any desired level of significance α. Due to the fact that quantity $(C(\alpha)\cdot S_\omega + \overline{\omega})$ *is always positive in this case and it is a monotonically increasing function of* M*, quantity* $\omega = \#edd\left(\gamma_n^N\right)$ *tends to increase rapidly, even for particularly small numbers of* M*.*

Hypothesis 2 and Associated Notation 5. *Suppose that probabilities* $P_i = P^{EQ}\left(\xi; \gamma_n^i, \beta_n^i\right)$ *do not remain constant throughout the successive multiplications, but on the contrary, they depend on the current* $i-th$ *multiplication. In this case, we consider the following events and the corresponding probabilities for an arbitrary multiplication, say the* $i-th$ *one:*
$A_{1,i} = \left\{2 \text{ e.d.d. added to } \gamma_n^i\right\}; P_{1,i} = P^{EQ}\left(2; \gamma_n^i, \beta_n^i\right),$
$A_{2,i} = \left\{1 \text{ e.d.d. added to } \gamma_n^i\right\};\ P_{2,i} = P^{EQ}\left(1; \gamma_n^i, \beta_n^i\right),,$
$A_{3,i} = \left\{\text{no e.d.d. added to } \gamma_n^i\right\}; P_{3,i} = P^{EQ}\left(0; \gamma_n^i, \beta_n^i\right),$
$A_{4,i} = \left\{\text{reduction of } \gamma_n^i \text{ e.d.d. by one}\right\}; P_{4,i} = P^{EQ}\left(-1; \gamma_n^i, \beta_n^i\right)$ *etc.*

If one adopts the above Hypothesis 2, the following result holds:

Theorem 2. *Under the conditions imposed by Hypothesis 2, one may dynamically compute the exact (up to* ± 1 *d. d.) number of erroneous decimal digits, which are accumulated at the* $i-th$*, arbitrary, product* γ_n^i*,* $i = 1, 2, \ldots, M$ *by applying the method introduced in Section 3. This dynamic computation of* $\#edd\left(\gamma_n^i\right)$ *can be made by a rather straightforward code based on the results of Section 3.*

Case 2. Now, we assume that at each one of the M successive multiplications, inequality (3.16) holds, i.e., that

$$|man(\gamma_n)man(\beta_n)| \geq 10.$$

Then, consider the following associated, quite representative probabilities, in accordance with the analysis of Section 3.2, Case 1.ii:

1. Probability that A_1 $(\xi = 2)$ occurs is $P_1 = 0$, since in this case equality $\xi = 2$ can never occur.
2. Probability that A_2 $(\xi = 1)$ occurs is $P_2 = 6.146 \cdot 10^{-4}$.
3. Probability that A_3 $(\xi = 0)$ occurs is $P_3 = 0.7468$.
4. Probability that A_4 $(\xi = -1)$ occurs is $P_4 = 0.2224$.
5. Probability that A_5 $(\xi = -2)$ occurs is $P_5 = 0.02720$.
6. Probability that A_6 $(\xi = -3)$ occurs is $P_6 = 0.002979$.

As a rule, the probabilities of events A_6, $A_7, \ldots$ are pretty small, practically zero; however, the entire analysis is absolutely valid if one incorporates the (very small) corresponding probabilities in it. Hence, according to Lemma 1, $\overline{\omega} = -0.2851 \cdot M$ and $S_\omega^2 = 0.2783 \cdot M$.

We would like to emphasize that in this case, the mean value $\overline{\omega}$ of generated e. d. d. is negative.

Now, quantity $z = \frac{\omega - \overline{\omega}}{S_\omega}$ follows a standard Gauss distribution, i.e., $z \sim N(0,1)$. Hence, inequality $\frac{\omega - \overline{\omega}}{S_\omega} \leq 4.2649$ holds with confidence 99.999%. With this confidence level, the accumulated number ω of e. d. d. after M successive multiplications obeying (3.16), satisfies

$$\omega \leq 4.2649 S_\omega + \overline{\omega}. \tag{4.8}$$

Here, $4.2649 S_\omega + \overline{\omega}$ is a monotonically decreasing function of M. Thus, the accumulated number ω of e. d. d. remains very close to zero, even for a very large number of multiplications. This has been fully experimentally verified as described in Section 4. Hence, the following holds:

Theorem 3. *Suppose that a number of* M *successive multiplications* $\gamma_n^{i+1} = \gamma_n^i \beta_n^i, i = 0, \ldots, M-1$ *is performed. For every such multiplication, let inequality* $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right| \geq 10$ *(3.16) holds. Then, for all practical purposes, these multiplications accumulate a negligible amount of f. p. error on the product* $\gamma_n^{i+1} = \gamma_n^i \beta_n^i$ *for all* $i = 0, \ldots, M$*.*

Moreover, the number of erroneous decimal digits (e. d. d.) accumulated in the arbitrary γ_n^i product, is, as a rule, a decreasing function of i.

If, in addition, Hypothesis 1 is adopted, then the number ω of e. d. d. accumulated in the i-th multiplication satisfies inequality (4.8).

The theorem holds for any desired level of significance α, the only difference being the coefficient of S_ω.

By a complete analogy with Theorem 2, one may adopt Hypothesis 2, in which case the following result holds:

Theorem 4. *Under the conditions imposed by Hypothesis 2, one may dynamically compute the exact (up to ± 1 d. d.) number of erroneous decimal digits, which are accumulated at the $i-th$, arbitrary, product γ_n^i, $i = 1, 2, \ldots, M$ by applying the method introduced in Section 3. This dynamic computation of $\#edd\left(\gamma_n^i\right)$ can be made by a rather straightforward code based on the results of Section 3.*

Case 3. In the general case, either inequality $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right| \geq 10$ (3.16) or inequality $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right| < 10$ (3.4) arbitrarily holds. Then, in order to obtain a rigorous estimation of the number of e. d. d. in each multiplication, together with the corresponding probability, one must know the statistical distribution of $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right|$, as compared to ten (10). In general, these distributions may highly depend on the algorithm in hand. However, in order to obtain an estimation of the corresponding generated f. p. error, we will state the very interesting example where both $\left|man\left(\gamma_n^i\right)\right|$ and $\left|man\left(\beta_n^i\right)\right|$ follow a uniform distribution in the interval $[1, 10)$. In fact, in this case, the set of $\left(\gamma_n^i, \beta_n^i\right)$ in the $\gamma_n^i\beta_n^i$-plain satisfying (3.4), is the 2D domain bounded by the straight lines $\gamma_n^i = 1$ and $\beta_n^i = 1$ and the hyperbola $\gamma_n^i\beta_n^i = 10$. Dually, the 2D domain for which the alternative inequality (3.16) holds, is the one limited by the straight lines $\gamma_n^i = 10$, $\beta_n^i = 10$ and the same hyperbola. Then, we follow the results of Section 3 and we use the graphical representation associated with the square of Figure 1 for the probability density function $f_{XY}^{UN}(x, y) = \frac{1}{81}$, defined on this square except the cross. Consequently, in a rather straightforward manner, we obtain $P\left\{\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right| < 10\right\} = \frac{1}{81}\int_1^{10}\frac{10}{\beta_n^i}d\beta_n^i = \frac{1}{81}\left[10\ln\left(\beta_n^i\right)\right]_1^{10} \cong 0.7157$.

In case that there is no discernible distribution of $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right|$ within the course of the algorithm, we may dynamically calculate the finite precision error accumulation for every product in order to estimate the accumulation of the finite precision error in the algorithm in general, as described in Theorems 2 and 4; we remind that Theorem 2 refers to the worst case in which $\left|man\left(\gamma_n^i\right)man\left(\beta_n^i\right)\right| < 10$ always holds, while Theorem 4 is connected to the dual inequality (3.16), which is most favorable from the point of view of generation of finite precision error during multiplication. In any case, the following holds:

Theorem 5. *Suppose again that during M successive multiplications $\gamma_n^{i+1} = \gamma_n^i\beta_n^i, i = 0, \ldots, M-1$ and that for a fraction, say F_g, of these multiplications, inequality (3.16) holds, while for the other fraction $F_s = 1 - F_g$ of them inequality (3.4) holds. Then, concerning the f. p. error accumulation in the products γ_n^i, $i = 1, \ldots, M$, the following two cases hold:*

(i) if $F_g > 0.5$, product γ_n^i tends to behave as described in Case 2, i.e., the overall number of e. d. d. of γ_n^i, $i = 1, \ldots, M$ is restrained. The closer to 1 fraction F_g is, the greater the restriction of the number of e. d. d. accumulated in products γ_n^i (see Table 8).

(ii) If $F_g < 0.5$, the accumulated f. p. error in the products γ_n^i is amplified. The closer to 0 F_g is, the more rapidly the f. p. e. accumulated in products γ_n^i grows (Table 7).

The theoretical approach and the associated results introduced in the present Section 4, have been fully confirmed experimentally, as it will be described in Section 6 of the present work.

5. Comparing the Finite Precision Error Generation and Accumulation during Execution of the Same Algorithm Including Successive Multiplications, with Different Finite Word Length/Precision

We shall begin by giving a brief description of the goal of the present section: consider an algorithm $\mathcal{A}$, involving multiplications at each iteration. We execute $\mathcal{A}$ first with n decimal digits in the mantissa (say $n \geq 7$) and simultaneously with m decimal digits (d. d.) in the mantissa, where we assume that $m \geq 2n + 7$, using exactly the same input in both cases. Consider any quantity γ of $\mathcal{A}$ and let γ_n^i be the value of this quantity at the $i-$th iteration of $\mathcal{A}$, where all calculations are made with precision of n d. d. in the mantissa. Similarly, let γ_m^i be the value of this quantity at the same iteration of $\mathcal{A}$, when all operations are made with precision of m d. d. In the present section, we will compare the number of erroneous d. d. with which any such two quantities γ_n^i, γ_m^i are calculated and, in particular, for the difference $\Delta = \left|\#edd\left(\gamma_n^i\right) - \#edd\left(\gamma_m^i\right)\right|$.

In Section 4 we have concluded that, independently of the finite word length, the number of e. d. d. of any product γ follows a normal distribution if the number of successive multiplications which generated γ, is greater than or equal to 30. Thus, the difference in the number of e. d. d. between γ_n^i and γ_m^i also follows a normal distribution with mean value zero and a variance that can be immediately estimated from the results of Section 4. Hence, one may deduce:

Theorem 6. *Suppose that an algorithm $\mathcal{A}$ including an arbitrary number of successive multiplications, is executed in parallel with two different finite word lengths corresponding to n and m decimal digits (d. d.). Let the two representations of an arbitrary quantity γ of $\mathcal{A}$ be γ_n and γ_m respectively, in these two finite word lengths. Consider the random variable*

$$\Lambda = \{number\ of\ e.\ d.\ d.\ accumulated\ in\ \gamma_m\} - \{number\ of\ e.\ d.\ d.\ accumulated\ in\ \gamma_n\}. \tag{5.1}$$

Λ follows a normal distribution with mean value zero and variance $2S_\omega^2$, where S_ω^2 is given in (4.2). Let $F_{m,n}(t)$ be the cumulative distribution function of Λ's normal distribution. Then, the probability that $\Delta = |\Lambda|$ is greater than ζ d. d. (where, clearly, $\zeta \geq 0$) is given by

$$P(\Delta > \zeta) = 2(1 - F_{m,n}(\zeta)). \tag{5.2}$$

Corollary 1. *Based on the analysis introduced in Sections 3 and 4, one may deduce in a quite straightforward manner that the probabilities that γ_n^i and γ_m^i differ in absolute value by $\Delta \geq 7$ decimal digits, is practically zero. This holds true for arbitrarily large number M of successive multiplications executed in A.*

Theorem 7. *As in Theorem 6, we let $\mathcal{A}$ be executed in parallel with the two different finite word lengths n and m, where $m > 2n + 7$. Then, for an arbitrary quantity γ of $\mathcal{A}$, the following hold:*

1. *We project γ_m to n d. d. in the mantissa, obtaining a restricted representation $\widetilde{\gamma}_n$ of γ_m. We compare γ_n and $\widetilde{\gamma}_n$ by means of Definitions 1 and 2. If the obtained result is κ e. d. d. $(\kappa < n)$, then we deduce that precisely the last κ digits of γ_n are erroneous.*
2. *We also deduce that γ_m has at most $\kappa + 7$ e. d. d. or, equivalently, that the first m–κ –7 d. d. of γ_m are correct.*
3. *As long as $\kappa < n$ holds, then, $\widetilde{\gamma}_n$ is a fully correct representation of γ with n d. d.*

6. Experiments That Fully Support the Theoretical Analysis

In this section, we shall introduce a number of experiments that have been specifically designed by the authors, in order to test the validity and the reliability of the theoretical analysis and results presented in the previous sections.

6.1. Description of a First Class of Experiments That Confirm the Theoretical Approach

Aiming at testing methodology and the associated theoretical results introduced in Sections 3 and 4, we have proceeded as follows: first, we have selected a set S_n of

10^6 randomly chosen floating point numbers having 16 decimal digits (d. d.) in the mantissa (subscript n stands for 16); the elements of this set come from a uniform distribution. All numbers were expressed in scientific form.

Next, we have extended each number a_n of S_n into a 40 d. d. representation, in scientific form, setting the last 24 d. d. of each number's mantissa to zero. Thus, we have obtained floating point numbers a_m forming set S_m ($m = 40$).

Subsequently, we have chosen an arbitrary, momentarily fixed value of F_g in the interval $[0, 1]$. We have performed $F_g \cdot 10^9$ multiplications with $n = 16$ d. d., for which the multiplication operands $\alpha_n,\ \beta_n$ satisfied (3.16). Next, we have performed $\left(1 - F_g\right) \cdot 10^9$ multiplications with 16 d. d. word length, where the opposite inequality, (3.4), namely that the mantissa of the product terms have absolute value smaller than 10, holds. We have ensured that no repetition of any multiplication occurred.

The very same multiplications have been repeated with 40 d. d. precision, among the corresponding numbers $a_m \in S_m$. Suppose that two numbers $\alpha_n, \beta_n \in S_n$, when multiplied, generate γ_n^1 with finite precision error (f. p. e.) $x_n^1 \geq 0$, while γ_m^1 is generated with f. p. e. $x_m^1 \geq 0$. These errors have been computed via Definition 2 and Theorems 5 and 6 introduced in Section 5. More specifically:

(i) We have restricted γ_m^1 into $n = 16$ d. d., thus obtaining the number $\widetilde{\gamma}_n^1$.
(ii) According to Theorem 6, $\widetilde{\gamma}_n^1$ is a correct representation of product γ having $n = 16$ decimal digits in its mantissa.
(iii) We have compared γ_n^1 and $\widetilde{\gamma}_n^1$ using Definitions 1 and 2, i.e., by forming their difference $\left|\gamma_n^1 - \widetilde{\gamma}_n^1\right|$. In this way, we have obtained the exact number of erroneous decimal digits (e. d. d.) with which quantity γ_n^1 has been evaluated.

By merging the obtained products γ_n^1 and γ_m^1 in two distinct ensembles, we have formed two new sets, $S_n^1,\ S_m^1$, being in a natural biunivocal relation $(\gamma_n^1,\ \gamma_m^1)$.

Moreover, for the same value F_g, we have performed $F_g \cdot 10^9$ multiplications between $\alpha_n^1,\ \beta_n^1 \in S_n^1$, satisfying (3.16), as well as $\left(1 - F_g\right) \cdot 10^9$ multiplications where (3.4) holds, obtaining 10^9 products γ_n^2. Again, during the aforementioned process, no repetition of any multiplication occurred. The very same multiplications have been performed with 40 decimal digits precision, between corresponding elements of set S_m^1, obtaining products γ_m^2. The erroneous d. d. of γ_n^1 have been computed using γ_m^1, as described above based on the results of Sections 2 and 5. We let products γ_n^2 and γ_m^2 form sets S_n^2 and S_m^2 respectively, maintaining the natural biunivocal relation $(\gamma_n^2,\ \gamma_m^2)$.

We continued in this way, forming sets $(S_n^3, S_m^3), \ldots,\ \left(S_n^i, S_m^i\right)$, etc. with the same factor F_g. In all these cases we evaluated the number of e. d. d. with which products γ_n^i, are computed as it has been previously described in connection with γ_n^1 and γ_m^1. In addition, whenever an exponent exceeded a large absolute value (e.g., 50) during the previous process, it was set to zero, since the exponent $10^\tau,\ \tau \in \mathbb{Z}$, of the scientific form plays no role in the f. p. e. generation and accumulation in the multiplication process in general. We have taken this action, in order to avoid possible effects of overflow or underflow in consecutive multiplications, since these easily spotted problems have nothing to do with the present study. However, we have kept the overall exponent of each product by simple recursive additions.

We have repeated the aforementioned experiment for various values of F_g, where always $F_g \in [0,1]$. At this point, we have distinguished two additional sub-cases: (a) $F_g < 0.5$ and (b) $F_g > 0.5$.

Sub-case (a) is quite analogous to Case 1, for which inequality (3.4) holds permanently. Specifically, the obtained products $\gamma_n^i = \alpha_n^i \beta_n^i$, have been calculated with all digits erroneous after a relatively small number of iterations, as shown in Table 7. The smaller fraction F_g, the more serious the f. p. error is.

On the contrary, Sub-case (b) is quite similar to Case 2, in the sense that products $\gamma_n^i = \alpha_n^i \beta_n^i$ manifested substantially smaller f. p. e. accumulation, as Table 8 manifests. In full accordance with the theoretical analysis, the smaller F_g, the smaller the accumulated f. p. e. in γ_n^i is.

6.2. A Second Class of Experiments for Testing the Theoretical Analysis Concerning Successive Multiplications

Case 1. All Successive Multiplications Satisfy Inequality (3.4), $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right| < 10 \ \left(\Leftrightarrow F_g \cong 0\right)$.

In connection to it, we have performed the following experiment: we have implemented an artificial algorithm, which forces all successive multiplications to satisfy (3.4). The flow chart of this algorithm is the following:

Starting from an arbitrary number $\beta_0 \in [1, 10)$, we express it with a certain number n of decimal digits (d. d.), as well as with $m = 2n + 10$ d. d. We then multiply β_0 by itself in both precisions. In case β_0^2 exceeds ten, then we subtract a properly selected positive integer c_0, from β_0 in both precisions; we do so, in order that $1 \leq \beta_0^2 < 10$ now holds. We stress that c_0 is adequately selected to be an integer in order that its subtraction from the initial β_0 does not add any e. d. d. to β_0; in all performed experiments, we ensured this by checking the number of e. d. d. of the difference $(\beta_0 - c_0)$, via Definition 2. By comparing product β_0^2 in both precisions, we calculate the erroneous decimal digits (e. d. d.) of β_0^2 in the n digits precision. Next, we set the exponent of β_0^2 equal to zero, in order to avoid overflow or underflow and we let the obtained mantissa of β_0^2 be a new number, β_1, expressed in both precisions. Then, we repeat the previous actions by letting β_1 play the role of β_0 and we evaluate and store the number of e. d. d. with which β_1^2 is computed, after ensuring that $\beta_1^2 \in [1, 10)$, via a proper subtraction $\beta_1 - c_1$, as before. We continue this process until the obtained β_i^2 is calculated in the n digits precision with all its digits erroneous, while we have ensured that $\beta_i^2 \in [1, 10)$.

We have executed this algorithm for 1000 different initial values of β_0, always belonging to the interval $[1, 10)$. The obtained maximum number of iterations for which β_i^2 was totally erroneous is shown in Table 9 for various values of precision n. Thus, we obtain the particularly important result that β_i^2 is totally erroneous after an impressively small number of iterations, in comparison to the employed precision, in full accordance with the theory and in particular with Theorem 1 of Section 4.

Table 9. Table demonstrating the number of iterations after which the output of the algorithm described in Case 1 of the present Section, offered totally erroneous results, for various employed finite word lengths n. The results are in full accordance with the theoretical analysis presented in the previous sub-sections. The experimentally observed results are in excellent agreement with the content of Theorem 1 of Section 4.

Employed Precision in Decimal Digits	Number of Iterations after Which All Digits of β_i^2 Were Erroneous, Independently of the Choice of β_0
16	61
64	249
128	484
256	967
512	1915
1024	3843
2048	7670
4096	15,285

Case 2. All Successive Multiplications Satisfied $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right| \geq 10 \ \left(\Leftrightarrow F_g \cong 1\right)$.

We have, again, performed an additional experiment, in which we have written an artificial algorithm, that forces all successive multiplications to satisfy $\left|man\left(\alpha_n^i\right)man\left(\beta_n^i\right)\right| \geq 10$. Indeed, this algorithm is quite similar to the one described in connection with Case 1 above and it has the following flow chart:

Starting, again, from an arbitrary number $\beta_0 \in [1, 10)$, we express it in both n and $m = 2n + 10$ d. d. precision. We then execute $\beta_0 \cdot \beta_0$ in both precisions. In case β_0^2 is smaller than ten, then we add a properly selected positive integer c_0 to β_0 in both precisions, so as $\beta_0^2 \geq 10$. We stress that $\beta_0 + c_0$ never manifests any e. d. d. By comparing product β_0^2 in both precisions, we calculate and store the n-precision number's e. d. d., again by means of Definition 2 and Theorem 5. Next, we set the exponent of β_0^2 to zero, once more to avoid overflow or underflow and we let the obtained mantissa of β_0^2 be a new number β_1 expressed in both precisions. Next, we repeat the previous actions by letting β_1 play the role of β_0 and we store and evaluate the number of e. d. d. with which number β_1^2 is computed, after ensuring that $\beta_1^2 \in [1, 10)$ by adding a proper c_1 to β_1, if necessary. We repeated this process for an arbitrarily large number of iterations, while monitoring the f. p. error of β_1^2.

We have executed this algorithm 10^{10} times in 16 and 42 d. d. precision for 1000 initial values of β_0, always belonging to the interval $[1, 10)$. The experiment has shown that the number of erroneous decimal digits with which β_1^2 has been calculated never exceeded two (2), while the mean value of these e. d. d. remained always pretty close to zero, even for the larger numbers of iterations of the algorithm, in full accordance with Theorem 3.

6.3. Description of a Third Experiment That Fully Supports the Theoretical Results regarding the Case of Successive Multiplications with a Varying Word Length

We have experimentally tested the correctness of Theorem 7 of Section 5, by performing M successive multiplications as described in Section 4. However, now, each multiplication has been executed three times with 16, 40 and 128 d. d. in the mantissa. In this way for each product γ we have obtained three representations, γ_{16}, γ_{40} and γ_{128} in parallel. Next, we have restricted γ_{40} to 16 d. d., obtaining representation $\widetilde{\gamma}_{16}$ as described before. Similarly, we have projected γ_{128} to both 16 and 40 d. d., obtaining the corresponding representations $\hat{\gamma}_{16}$ and $\hat{\gamma}_{40}$. Eventually, we have compared γ_{16} with $\widetilde{\gamma}_{16}$ and $\hat{\gamma}_{16}$ by means of Definitions 1 and 2; we have also compared $\widetilde{\gamma}_{16}$ with $\hat{\gamma}_{16}$ and γ_{40} with $\hat{\gamma}_{40}$ via the same method. The obtained results are shown in Table 10 and fully justify the aforementioned Theorems of Section 5, but also of Section 4.

Table 10. Comparison of the number of erroneous decimal digits (e. d. d.) accumulated in all the intermediate results of 10^8 successive multiplications. All these multiplications have been executed in parallel, with 16, 40 and 128 d. d. in the mantissa. All obtained experimental results fully support the theoretical analysis introduced in Section 5 and in particular the content of Theorems 6 and 7.

Minimum Erroneous Decimal Digits Difference Between 16 and 40 Decimal Digits Representation.	-3
Maximum Erroneous Decimal Digits Difference Between 16 and 40 Decimal Digits Representation	3
Mean Erroneous Decimal Digits Difference Between 16 and 40 Decimal Digits Representation	-0.0945
Maximum Number of Erroneous Decimal Digits in the 16 Decimal Digits Representation	12
Maximum Number of Erroneous Decimal Digits in the 40 Decimal Digits Representation	32

7. Eventual Applications Associated with the Present Work

In the section in hand, we shall present and highlight an ensemble of possible and probable applications, which will be based in the analysis and methodology introduced here. Thus:

A. In certain applications, like the ones that will be described below, it is preferable and/or necessary to use finite elements methods, which employ polynomials of high order to approximate the considered function on each element, usually called

"higher order basis functions". In this approach, if the higher order basis functions are of order n, then one must use elements consisting of $(n+1)$ nodes; moreover, one frequently uses the following basis functions ([20]):

$$\hat{\psi}_i(\xi) = \frac{(\xi-\xi_1)\dots(\xi-\xi_j)\dots(\xi-\xi_{i-1})(\xi-\xi_{i+1})\dots(\xi-\xi_{n+1})}{(\xi_i-\xi_1)\dots(\xi_i-\xi_j)\dots(\xi_i-\xi_{i-1})(\xi_i-\xi_{i+1})\dots(\xi_i-\xi_{n+1})}, \tag{7.1}$$

where (a) i is the cardinal number of the node in hand, $i = 1,2,\dots,n+1$, (b) j represents the cardinal number of the other nodes of the specific element, hence $j = 1,2,3,\dots,n+1$, $j \neq i$, (c) ξ is the independent variable of the polynomial basis function and (d) evidently ξ_j, $j = 1,2,3,\dots,n+1$ is the value that this variable acquires on the $j-th$ element of the node in hand.

It is rather clear that both the nominator and the denominator in relation (7.1) are results of successive multiplications.

However, even in the case of second order basis functions, one employs the basis functions:

$$\hat{\psi}_1(\xi) = \frac{1}{2}\xi(\xi-1),\ \hat{\psi}_2(\xi) = 1-\xi^2,\ \hat{\psi}_3(\xi) = \frac{1}{2}\xi(\xi+1), \tag{7.2}$$

which includes multiplications. Consequently, the entire previous analysis may be applied immediately, so that together with $\hat{\psi}_i(\xi)$, $i = 1,2,\dots,n+1$ computation, the user may know the exact number of erroneous decimal digits with which this quantity has been evaluated, each time. Clearly, in case that the numerical value of such a basis function for a certain ξ is highly or even totally "contaminated", then the user may immediately receive a corresponding signal.

Therefore, more specifically, this method can be applied to the subsequent applications:

1. In research associated with the modelling of the fatigue of materials employed in the rail-wheel system ([21,22]).
2. In the study of rail corrugation ([23,24]).
3. In the study of the influence of bending on the value of friction coefficient ([25]).
4. In tackling important classes of contact problems in elatostatics ([26,27]).
5. In the investigation and analysis of the spatial stress-strain states of a pipe with respect to its corrosion damage, taking into account various types of complex loading ([28]).
6. In real time analysis of local damage in wear-and-fatigue tests, whenever finite elements methods are required/applied ([29]).

It is worthwhile noticing that in many of the aforementioned studies the involved models frequently include multiplications; consequently the approach introduced in the present manuscript may also be proved helpful in associated numerical experiments.

B. The sequence of powers of a real number.

Consider a single real number, say $\beta > 0$. Moreover, consider the sequence of powers of β, usually computed recursively by means of the following succession of multiplications:

$$\begin{gathered} z_0 = \beta\cdot\beta \\ z_1 = z_0\cdot z_0 \\ z_2 = z_1\cdot z_1 \\ \dots \\ z_n = z_{n-1}\cdot z_{n-1},\ n \in \mathbb{N}. \end{gathered}$$

Suppose that the numerical value of $\beta > 0$ is such that, statistically, multiplication $z_{n-1}\cdot z_{n-1}$, $n \in \mathbb{N}$, satisfies inequality (3.4)

$$|man(z_{n-1})\cdot man(z_{n-1})| < 10,$$

more frequently than the opposite one (3.16), namely

$$|man(z_{n-1})\cdot man(z_{n-1})| \geq 10$$

Then, according to the previous analysis, one expects that z_n will continually be evaluated with a larger number of erroneous decimal digits (e. d. d.), as n grows. To verify/demonstrate that, we have employed $\beta = 1.12$, we have generated sequence z_n of the powers of β by means of the aforementioned sequence of successive multiplications and we have evaluated the exact number of e. d. d. with which z_n is calculated each time; the determination of the exact number of erroneous d. d. has been made as described in Sections 5 and 6, using $n = 16$ decimal digits word length and $m = 40$ decimal digits precision. The associated results are depicted in Figure 4, from which it is evident that after the impressively small number of 55 iterations, the power z_n is computed with all its digits erroneous.

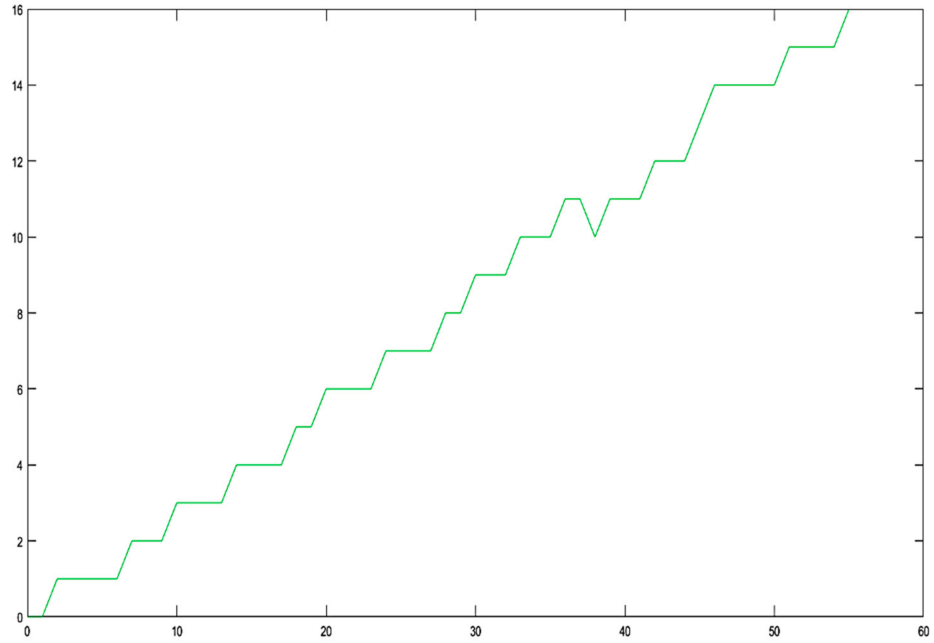

Figure 4. The evolution of the number of the erroneous decimal digits (e. d. d.) accumulated in the power β^{2n}, $\beta = 1.12$, due to finite precision error. The abscissa represents the recursions' cardinal number, while $y'y$ axis represents the number of e. d. d. Number β has been chosen in such a way, so as inequality $|man(z_{n-1})\cdot man(z_{n-1})| < 10$ holds more frequently than the dual one (3.16). As a consequence, the number of e. d. d. grows rapidly, in full accordance with the analysis and the results of Sections 4 and 6.

We must emphasize that, in order to circumvent the effects of overflow, each time we have multiplied the mantissae of z_{n-1} only and not the entire number z_{n-1}. Equivalently, whenever the exponent of z_n exceeded a rather large number, say $E(z_n) = 50$, then we have divided with $10^{E(z_n)}$. However, we have registered the power's exponent each time by simple recursive additions. It is important to stress that, in both these approaches the number of erroneous decimal digits accumulated in z_n were identical.

C. Continual multiplication of contaminated numbers.

Exactly the same analysis holds true, in the case that instead of multiplying z_{n-1} with itself to produce z_n, we instead perform the sequence of multiplications:

$$\begin{gathered} z_0 = x_0 \cdot y_0 \\ z_1 = z_0 \cdot y_1 \\ \cdots \\ z_n = z_{n-1} \cdot y_{n-1},\ n \in \mathbb{N}, \end{gathered}$$

where $y_0, y_1, \ldots, y_{n-1},\ n \in \mathbb{N}$ is an arbitrary sequence of contaminated numbers, to which erroneous digits are accumulated probably due to another procedure. The application (D) that follows, we believe that it will clarify the content of these statements.

D. Finite Precision Error Accumulated in Various Fast Kalman Algorithms.

One of the most widely used filtering procedures is the Kalman one [30]. In many of these algorithmic schemes a certain scalar quantity, say $\alpha_m^b(n+1),\ m, n \in \mathbb{N}$, is updated at the $(n+1) - th$ time instant by means of a formula of the type

$$a_m^b(n+1) = \lambda \cdot \alpha_m^b(n) \cdot J_m(n+1), \tag{7.3}$$

where (i) $\lambda \in \mathbb{R}$ is the so-called "forgetting factor" almost always belonging to the interval $[0.97, 0.99]$ and (ii) $J_m(n+1)$ is another quantity of the algorithm, which is also computed recursively. In many applications [30], quantities $\alpha_m^b(n)$ and $J_m(n+1)$ have values such that inequality (3.4)

$$|man(z_{n-1}) \cdot man(z_{n-1})| < 10,$$

holds very frequently, statistically. Hence, every formula of the type (7.3), tends to generate one additional erroneous decimal digit in a relatively small number of recursions; this erroneous digit is added to the value of $a_m^b(n+1)$. Subsequently, since $a_m^b(n+1)$ enters directly or indirectly, in all other formulae of the corresponding Kalman algorithms, including $J_m(n+1)$, it follows that these schemes are very frequently destroyed due to this successive-multiplication-based finite precision error, in an impressively small number of iterations [30].

Thus, for example, the faster existing Kalman algorithm (the FAEST [31]) can never converge in practice due to this type of f. p. e.

In general, the methodology introduced here allows for both the evaluation of the number of erroneous decimal digits with which all quantities in any fast Kalman algorithm are computed, as well as for finding methods of stabilizing various algorithms of this class ([32]).

8. Conclusions

In this paper, we have presented a new approach to the study of the finite precision error generation and accumulation in the multiplication process. We have initially given a strict mathematical definition of the number of correct digits of a real quantity expressed in any finite word length. We emphasize that although the analysis introduced here is made in the decimal radix, it offers accurate results and prediction of the f. p. e. generated and accumulated in any computing machine that performs an arbitrary number of multiplications, successively.

Along this new approach, we have shown the following fundamental result: suppose that one executes an arbitrary multiplication $\gamma_n = \alpha_n \beta_n$ in a computing environment employing the equivalent of n decimal digits in the mantissa. Moreover, let operands α_n and β_n have λ erroneous decimal digits at most in their mantissae. Then, the number of e. d. d. with which product γ_n is calculated depends on the value of $|man(\alpha_n) man(\beta_n)|$. In fact, if inequality $|man(\alpha_n) man(\beta_n)| < 10$ holds, then product γ_n is calculated with at most $\lambda + 2$ erroneous d. d. or with $\lambda,\ \lambda - 1,\ \lambda - 2,\ \lambda - 3$ e. d. d. In case the complementary inequality holds, then product γ_n may be calculated with up to $\lambda + 1$, or with $\lambda,\ \lambda - \kappa,\ \kappa = 1,\ 2,\ 3,\ 4$ e. d. d.

We have also shown that the chance of encountering one of the aforementioned cases heavily depends on the exponent of quantity $\left|\alpha_n x + \beta_n y \cdot 10^{-\delta}\right|$, where x and y are the multiplication operands' f. p. e. mantissae and $\delta = \#edd(\beta_n) - \#edd(\alpha_n)$.

In order to calculate the probabilities that each one of the aforementioned cases holds, we have introduced the rectangular shaped set of points of Figure 1 and we have defined the sub-domains in which the values of the random variables x and y correspond, in order that product γ_n is computed with a specific number of e. d. d. Then, by integration on the corresponding sub-domains, we have calculated the associated probabilities.

We have also given exact formulae for the mean value and standard deviation of the number of e. d. d. accumulated in the results of successive multiplications.

Moreover, we have established that if we perform the exact same set of successive multiplications using n and $m > 2n + 7$ d. d., then we may easily track the number of e. d. d. accumulated in the n precision results.

Finally, in order to test the validity of the introduced theoretical analysis, we have performed a number of specially developed experiments. The results of these experiments fully supported the theoretical analysis introduced here.

We emphasize that the developed novel methodology is expandable, so as to tackle the finite precision error generation and accumulation in any arithmetic operation; this will be the subject of forthcoming manuscripts.

Author Contributions: Conceptualization, C.P., D.A., F.G., C.C. and A.R.M.; Funding acquisition; Investigation, A.R.M. and C.C.; Project administration, C.P.; Resources, F.G. and C.C.; Software, C.P., C.C., A.R.M., F.G. and D.A.; Supervision, C.P. and D.A.; Validation, C.P., D.A., F.G., A.R.M. and C.C.; Writing—original draft, C.P. and F.G.; Writing—review & editing, A.R.M., C.P., D.A. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study did not employ any data sets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Caraiscos, C.; Liu, B. A roundoff error analysis of the LMS adaptive algorithm. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 34–41. [CrossRef]
2. Moustakides, G.V. Correcting the instability due to finite precision of the fast Kalman identification algorithms. *Signal Process.* **1989**, *18*, 33–42. [CrossRef]
3. Steele, G.L.; White, J.L. How to Print Floating-Point Numbers Accurately. In Proceedings of the ACM SIGPLAN 1990 Conference on Programming Language Design and Implementation, New York, NY, USA, 20–22 June 1990; pp. 112–126. [CrossRef]
4. Bai, Z. Error Analysis of the Lanczos Algorithm for the Nonsymmetric Eigenvalue Problem. *Math. Comput.* **1994**, *62*, 209–226. [CrossRef]
5. Arioli, M.; Fassino, C. Roundoff error analysis of algorithms based on Krylov subspace methods. *Bit Numer. Math.* **1996**, *36*, 189–205. [CrossRef]
6. Lowenstein, J.H.; Vivaldi, F. Anomalous transport in a model of Hamiltonian round-off. *Nonlinearity* **1998**, *11*, 1321–1350. [CrossRef]
7. Allen, E.; Burns, J.; Gilliam, D.; Hill, J.; Shubov, V. The impact of finite precision arithmetic and sensitivity on the numerical solution of partial differential equations. *Math. Comput. Model.* **2002**, *35*, 1165–1195. [CrossRef]
8. Gelb, A. Parameter Optimization and Reduction of Round Off Error for the Gegenbauer Reconstruction Method. *J. Sci. Comput.* **2004**, *20*, 433–459. [CrossRef]
9. Martel, M. Semantics of roundoff error propagation in finite precision calculations. *High. Order Symb. Comput.* **2006**, *19*, 7–30. [CrossRef]
10. Wang, P.; Huang, G.; Wang, Z. Analysis and application of multiple-precision computation and round-off error for nonlinear dynamical systems. *Adv. Atmos. Sci.* **2006**, *23*, 758–766. [CrossRef]
11. Papakostas, G.; Karras, D.; Boutalis, Y.; Mertzios, B. Fast numerically stable computation of orthogonal Fourier–Mellin moments. *IET Comput. Vis.* **2007**, *1*, 11–16. [CrossRef]

12. Kountouris, A. A randomized algorithm for controlling the round-off error accumulation in recursive digital frequency synthesis (DFS). *Digit. Signal Process.* **2009**, *19*, 534–544. [CrossRef]
13. Linderman, M.D.; Ho, M.; Dill, D.L.; Meng, T.H.; Nolan, G.P. Towards program optimization through automated analysis of numerical precision. In Proceedings of the 8th annual IEEE/ACM International Symposium on Code Generation and Optimization, Toronto, ON, Canada, 24–28 April 2010; pp. 230–237. [CrossRef]
14. Turchetti, G.; Vaienti, S.; Zanlungo, F. Relaxation to the asymptotic distribution of global errors due to round off. *EPL Europhys. Lett.* **2010**, *89*, 40006. [CrossRef]
15. Cheng, A.-D. Multiquadric and its shape parameter—A numerical investigation of error estimate, condition number, and round-off error by arbitrary precision computation. *Eng. Anal. Bound. Elem.* **2012**, *36*, 220–239. [CrossRef]
16. Deng, A.-W.; Wei, C.-H.; Gwo, C.-Y. Stable, fast computation of high-order Zernike moments using a recursive method. *Pattern Recognit.* **2016**, *56*, 16–25. [CrossRef]
17. Das, A.; Briggs, I.; Gopalakrishnan, G.; Krishnamoorthy, S.; Panchekha, P. Scalable yet Rigorous Floating-Point Error Analysis. In Proceedings of the SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 9–19 November 2020; pp. 1–14. [CrossRef]
18. Papaodysseus, C.; Koukoutsis, E.; Vassilatos, C. Error propagation and methods of error correction in LS FIR filtering and l-step ahead linear prediction. *IEEE Trans. Signal Process.* **1994**, *42*, 1097–1108. [CrossRef]
19. Papaodysseus, C.; Koukoutsis, E.; Triantafyllou, C. Error sources and error propagation in the Levinson-Durbin algorithm. *IEEE Trans. Signal Process.* **1993**, *41*, 1635–1651. [CrossRef]
20. Becker, E.B.; Carey, G.F.; Oden, J.T.; Belytschko, T. Finite Elements, An Introduction. *J. Appl. Mech.* **1982**, *49*, 682. [CrossRef]
21. Bendikiene, R.; Bahdanovich, A.; Cesnavicius, R.; Ciuplys, A.; Grigas, V.; Jutas, A.; Marmysh, D.; Nasan, A.; Shemet, L.; Sherbakov, S.; et al. Tribo-fatigue Behavior of Austempered Ductile Iron MoNiCa as New Structural Material for Rail-wheel System. *Mater. Sci.* **2020**, *26*, 432–437. [CrossRef]
22. Iannitti, G.; Ruggiero, A.; Bonora, N.; Masaggia, S.; Veneri, F. Micromechanical modelling of constitutive behavior of austempered ductile iron (ADI) at high strain rate. *Appl. Fract. Mech.* **2017**, *92*, 351–359. [CrossRef]
23. Liu, Q.; Zhang, B.; Zhou, Z. An experimental study of rail corrugation. *Wear* **2003**, *255*, 1121–1126. [CrossRef]
24. Ahlbeck, D.R.; Daniels, L.E. Investigation of rail corrugations on the Baltimore Metro. *Wear* **1991**, *144*, 197–210. [CrossRef]
25. Trzepiecinski, T.; Lemu, H.G. Effect of Lubrication on Friction in Bending under Tension Test-Experimental and Numerical Approach. *Metals* **2020**, *10*, 544. [CrossRef]
26. Campos, L.; Oden, J.; Kikuchi, N. A numerical analysis of a class of contact problems with friction in elastostatics. *Comput. Methods Appl. Mech. Eng.* **1982**, *34*, 821–845. [CrossRef]
27. Migórski, S.; Gamorski, P. A new class of quasistatic frictional contact problems governed by a variational–hemivariational inequality. *Nonlinear Anal. Real World Appl.* **2019**, *50*, 583–602. [CrossRef]
28. Sherbakov, S. Three-Dimensional Stress-Strain State of a Pipe with Corrosion Damage Under Complex Loading. *Tribol. lubr. Lubr.* **2011**. [CrossRef]
29. Sosnovskiy, L.; Bogdanovich, A.; Yelovoy, O.; Tyurin, S.; Komissarov, V.; Sherbakov, S. Methods and main results of Tribo-Fatigue tests. *Int. J. Fatigue* **2014**, *66*, 207–219. [CrossRef]
30. Papaodysseus, C.; Koukoutsis, E.; Stavrakakis, G.; Halkias, C. Exact analysis of the finite precision error generation and propagation in the FAEST and the fast transversal algorithms: A general methodology for developing robust RLS schemes. *Math. Comput. Simul.* **1997**, *44*, 29–41. [CrossRef]
31. Carayannis, G.; Manolakis, D.; Kalouptsidis, N. A fast sequential algorithm for least-squares filtering and prediction. *IEEE Trans. Acoust. SpeechSignal Process.* **1983**, *31*, 1394–1402. [CrossRef]
32. Boutalis, Y.; Papaodysseus, C.; Koukoutsis, E. A New Multichannel Recursive Least Squares Algorithm for Very Robust and Efficient Adaptive Filtering. *J. Algorithms* **2000**, *37*, 283–308. [CrossRef]