Dynamical control of converging sequences computation

Fabienne Jézéquel

Laboratoire d'Informatique de Paris 6 - CNRS UMR 7606,
4 place Jussieu, 75252 Paris cedex 05, France

Abstract

Under some assumptions on the speed of convergence of a sequence, the significant digits of one of its iterates in common with the exact limit can be determined by comparing this iterate with the next one. Using a finite precision arithmetic, if computations are performed until the difference between two successive iterates is insignificant, the global error on the last iterate is minimal. Furthermore, for sequences converging at least linearly, we can determine in the result obtained which exact significant digits, *i.e.* not affected by round-off errors, are in common with the exact limit. This strategy can be used for the computation of integrals with the trapezoidal or Simpson's rule. A sequence is then generated by halving the step value at each iteration, while the difference between two successive iterates is a significant value. The exact significant digits of the last iterate are in common with the exact value of the integral, up to one bit. This kind of strategy is then extended to numerical algorithms involving several sequences, such as the approximation of integrals on an infinite interval.

Key words: converging sequences, numerical validation, quadrature methods, trapezoidal rule, Simpson's rule, CESTAC method, Discrete Stochastic Arithmetic

1 Introduction

In a numerical method which involves the computation of a converging sequence, the limit is approximated by one of the iterates. It may be difficult to estimate in the chosen iterate the global error, consisting of the truncation error and the round-off error. The optimal iterate, *i.e.* the approximation for which the global error is minimal, can be computed dynamically [14]. In this

Email address: Fabienne.Jezequel@lip6.fr (Fabienne Jézéquel).

paper, we show that we can determine the significant digits of this optimal iterate, which are affected neither by the truncation error, nor by the round-off error. In section 2, we present theorems established from the truncation error which enable one to determine the significant digits of an iterate in common with the exact limit. As round-off errors must also be taken into account, in section 3, we briefly review methods and concepts which enable one to estimate round-off error propagation with a probabilistic approach: the CESTAC method, the principles of stochastic arithmetic and the implementation provided by Discrete Stochastic Arithmetic (DSA). We also present theoretical results established in stochastic arithmetic for the control of arithmetical operations. In section 4, we describe a strategy to control both the truncation and the round-off error during the computation of a converging sequence. More precisely, under some assumptions on the speed of convergence of the sequence, we can determine in the optimal approximation the exact significant digits, *i.e.* not affected by round-off errors, which are in common with the exact limit. In section 5, we show how the theorems established in the previous sections can be combined to control sequences in which each term is the limit of another sequence. We describe a strategy which can be used for the computation of improper integrals. The last section presents numerical experiments carried out using DSA.

2 Theoretical results on converging sequences

2.1 Preliminary definitions

The theorems presented here have been established for sequences having a linear or an exponential convergence speed. Therefore we recall properties which characterize these two types of convergence speed.

Definition 1 *A sequence (I_n) converges to I with a linear speed if*

$$I_n - I = K\alpha^n + o(\alpha^n), \text{ where } K \in \mathbb{R} \text{ and } 0 < |\alpha| < 1.$$

With a sequence having a linear convergence, the number of iterations required to obtain an approximation of the limit with one more exact digit is quasi-constant.

Definition 2 *A sequence (I_n) converges to I with an exponential speed if*

$$I_n - I = K\,\alpha^{p^n} + o(\alpha^{p^n}), \text{ where } K \in \mathbb{R},\ 0 < |\alpha| < 1 \text{ and } p > 1.$$

With a sequence having an exponential convergence, at each iteration, the number of exact digits is quasi-multiplied by p.

The theoretical results presented in this section require the notion of significant digits common to two real numbers. Therefore we need the following definition.

Definition 3 *Let a and b be two real numbers, the number of significant digits that are common to a and b can be defined in $\mathbb{R}$ by*

(1) for $a \neq b$, $C_{a,b} = \log_{10} \left| \dfrac{a+b}{2(a-b)} \right|$,
(2) $\forall a \in \mathbb{R},\ C_{a,a} = +\infty$.

Then $|a-b| = \left|\frac{a+b}{2}\right| 10^{-C_{a,b}}$. For instance, if $C_{a,b} = 3$, the relative difference between a et b is of the order of 10^{-3} which means that a and b have three significant digits in common.

Remark 4 *The value of $C_{a,b}$ can seem surprising if we consider the decimal notations of a and b. For example, if $a = 2.4599976$ and $b = 2.4600012$, then $C_{a,b} \approx 5.8$. The difference due to the sequences of "0" or "9" is illusive. The significant decimal digits of a and b are really different from the sixth position.*

2.2 On sequences with a linear convergence

Let us consider a sequence (I_n) converging linearly to I. From the number of significant digits common to two successive iterates, I_n and I_{n+1}, the following theorem enables one to determine the number of significant digits common to I_n and the exact limit I.

Theorem 5 *Let (I_n) be a sequence converging linearly to I, i.e. which satisfies $I_n - I = K\alpha^n + o(\alpha^n)$ where $K \in \mathbb{R}$ and $0 < |\alpha| < 1$, then*

$$C_{I_n, I_{n+1}} = C_{I_n, I} + \log_{10}\left(\frac{1}{1-\alpha}\right) + o(1).$$

PROOF.

$$I_n - I = K\alpha^n + o(\alpha^n) \tag{1}$$

By using the same formula for I_{n+1}, one obtains

$$I_n - I_{n+1} = K\alpha^n(1-\alpha) + o(\alpha^n) \tag{2}$$

From equation (1), we deduce

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^n \left(1 + o(1)\right)} \tag{3}$$

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^n} \left(1 + o(1)\right) \tag{4}$$

Therefore

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^n} + o\left(\frac{1}{\alpha^n}\right) \tag{5}$$

Then

$$\frac{I_n + I}{2(I_n - I)} = \frac{I_n}{I_n - I} - \frac{1}{2} = \frac{I_n}{K\alpha^n} + o\left(\frac{1}{\alpha^n}\right) \tag{6}$$

Similarly, from equation (2), we deduce

$$\frac{I_n + I_{n+1}}{2(I_n - I_{n+1})} = \frac{I_n}{I_n - I_{n+1}} - \frac{1}{2} = \frac{I_n}{K\alpha^n} \frac{1}{1-\alpha} + o\left(\frac{1}{\alpha^n}\right) \tag{7}$$

From definition 3 and equation (6) we deduce

$$C_{I_n,I} = \log_{10} \left| \frac{I_n}{K\alpha^n} \left(1 + o(1)\right) \right| \tag{8}$$

$$C_{I_n,I} = \log_{10} \left| \frac{I_n}{K\alpha^n} \right| + \log_{10} \left| 1 + o(1) \right| \tag{9}$$

Therefore

$$C_{I_n,I} = \log_{10} \left| \frac{I_n}{K\alpha^n} \right| + o(1) \tag{10}$$

Similarly, from definition 3 and equation (7) we deduce

$$C_{I_n,I_{n+1}} = \log_{10} \left| \frac{I_n}{K\alpha^n} \frac{1}{1-\alpha} \right| + o\left(1\right) \tag{11}$$

Finally

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10} \left(\frac{1}{1-\alpha} \right) + o\left(1\right)} \tag{12}$$

If the convergence zone is reached, $o(1) \ll 1$: the last term in equation (12) becomes negligible. In this case, from the significant digits in common between I_n and I_{n+1}, we can deduce the significant digits in common between I_n and the exact limit I.

If $-1 < \alpha < 0$, then $-\log_{10} 2 < \log_{10}\left(\frac{1}{1-\alpha}\right) < 0$. In this case, if the convergence zone is reached, the significant digits in common between I_n and I_{n+1} are also in common with I.

$\forall \alpha \in]0,1[$, $\exists k$ $0 < \alpha \leq 1 - 10^{-k}$ and therefore $0 < \log_{10}\left(\frac{1}{1-\alpha}\right) \leq k$. If the convergence zone is reached, the significant digits in common between I_n and I_{n+1} are also in common with I, up to k digits. The lower α is, the faster the convergence of the sequence is and the lower k is.

Remark 6 *If $0 < \alpha \leq \frac{1}{2}$, then $0 < \log_2\left(\frac{1}{1-\alpha}\right) \leq 1$. In this case, if the convergence zone is reached, the significant bits in common between I_n and I_{n+1} are also in common with I, up to one.*

2.3 On the trapezoidal and Simpson's rules

Theorem 5 can be used for the evaluation of integrals with the trapezoidal or Simpson's rule. Indeed a sequence which converges linearly can be generated by halving the step value at each iteration.

Let f be a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 3$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using the trapezoidal rule with step $h = \frac{b-a}{2^n}$. If $f'(a) \neq f'(b)$, the development of the error up to order 4 is [1,8,9]:

$$I_n - I = \frac{h^2}{12}\left[f'(b) - f'(a)\right] + \mathcal{O}(h^4) \tag{13}$$

As the sequence (I_n) satisfies $I_n - I = K\alpha^n + \mathcal{O}(\alpha^{2n})$, with $K = \frac{(b-a)^2}{12}\left[f'(b) - f'(a)\right]$ and $\alpha = \frac{1}{4}$, theorem 5 could apply. However the following property has been established in [5]:

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{14}$$

Let f be a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 5$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using Simpson's rule with step $h = \frac{b-a}{2^n}$. If $f^{(3)}(a) \neq f^{(3)}(b)$, the development of the error up to order 6

is [1,8,9]:

$$I_n - I = \frac{h^4}{180}\,[f^{(3)}(b) - f^{(3)}(a)] + \mathcal{O}(h^6). \tag{15}$$

The sequence (I_n) satisfies $I_n - I = K\alpha^n + \mathcal{O}(\alpha^{\frac{3}{2}n})$, with $K = \frac{(b-a)^4}{180}\,[f^{(3)}(b) - f^{(3)}(a)]$ and $\alpha = \frac{1}{16}$. Therefore, as for the trapezoidal rule, theorem 5 could apply. The following property has actually been established in [5]:

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{16}{15}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{16}$$

If the convergence zone is reached, $\mathcal{O}\left(\frac{1}{4^n}\right) \ll 1$. Furthermore $\log_{10}\left(\frac{4}{3}\right)$ and $\log_{10}\left(\frac{16}{15}\right)$ represent at most one bit. Indeed, for both rules, $\alpha < \frac{1}{2}$. Therefore, if the convergence zone is reached, the significant digits common to I_n and I_{n+1} are also common to I, the exact value of the integral, up to one bit.

2.4 On sequences with an exponential convergence

Theoretical results similar to theorem 5 may be established for sequences with an exponential convergence.

Theorem 7 *Let (I_n) be a sequence converging to I with an exponential speed, i.e. which satisfies $I_n - I = K\,\alpha^{p^n} + o(\alpha^{p^n})$ where $K \in \mathbb{R}$, $0 < |\alpha| < 1$ and $p > 1$, then*

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right) + o\,(1)\,.$$

PROOF.

$$I_n - I = K\,\alpha^{p^n} + o(\alpha^{p^n}) \tag{17}$$

By using the same formula for I_{n+1}, one obtains

$$I_n - I_{n+1} = K\,\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right) + o(\alpha^{p^n}) \tag{18}$$

From equation (17), we deduce

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^{p^n}\,(1 + o(1))} \tag{19}$$

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^{p^n}}\,(1 + o(1)) \tag{20}$$

Therefore

$$\frac{I_n}{I_n - I} = \frac{I_n}{K\alpha^{p^n}} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{21}$$

Then

$$\frac{I_n + I}{2(I_n - I)} = \frac{I_n}{I_n - I} - \frac{1}{2} = \frac{I_n}{K\alpha^{p^n}} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{22}$$

Similarly, from equation (18), we deduce

$$\frac{I_n}{I_n - I_{n+1}} = \frac{I_n}{K\ (\alpha^{p^n} - \alpha^{p^{n+1}})\ (1 + o(1))} \tag{23}$$

Therefore

$$\frac{I_n}{I_n - I_{n+1}} = \frac{I_n}{K\ (\alpha^{p^n} - \alpha^{p^{n+1}})} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{24}$$

Then

$$\frac{I_n + I_{n+1}}{2(I_n - I_{n+1})} = \frac{I_n}{I_n - I_{n+1}} - \frac{1}{2} = \frac{I_n}{K\ (\alpha^{p^n} - \alpha^{p^{n+1}})} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{25}$$

From definition 3 and equation (22) we deduce

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{K\alpha^{p^n}}\,(1 + o(1))\right| \tag{26}$$

Therefore

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{K\alpha^{p^n}}\right| + o(1) \tag{27}$$

Similarly, from definition 3 and equation (25) we deduce

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n}{K\ (\alpha^{p^n} - \alpha^{p^{n+1}})}\,(1 + o(1))\right| \tag{28}$$

Therefore

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n}{K\ \alpha^{p^n}\ (1 - \alpha^{p^n(p-1)})}\right| + o(1) \tag{29}$$

Finally

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1 - \alpha^{p^n(p-1)}}\right) + o(1)} \tag{30}$$

If the convergence zone is reached, the decimal significant digits in common between I_n and I_{n+1} are also common to the exact limit I, up to $\log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right)$.

If $0 < |\alpha| \leq M_n$, with $M_n = (\frac{9}{10})^{(\frac{1}{p^n(p-1)})}$, then $0 < \log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right) \leq 1$. The significant digits common to I_n and I_{n+1} are also common to I, up to one. As the number n of iterations increases, M_n also increases and the condition that α must satisfy in order to have $\log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right) \leq 1$ becomes less and less strict. For example, if the sequence (I_n) has a quadratic convergence, which is characterized by $p = 2$, then $M_1 > 0.94$ and $M_5 > 0.99$. Similarly, as p increases, the speed of convergence increases and M_n also increases.

Remark 8 *If the convergence zone is reached, the significant bits in common between I_n and I_{n+1} are also common to the exact limit I, up to* $\log_2\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right)$. *If $0 < |\alpha| \leq 2^{(\frac{1}{p^n(1-p)})}$, then $0 < \log_2\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right) \leq 1$. This condition on α is easily satisfied. Indeed in the case of a quadratic convergence (i.e. for $p = 2$) if $n = 5$, $2^{(\frac{1}{p^n(1-p)})} > 0.97$.*

The theoretical results presented in this section have been established by taking into account only the truncation error on two successive iterates of a sequence. However computed results are also affected by round-off error propagation. The next section describes how round-off errors can be estimated with a probabilistic approach in order to determine the exact significant digits of any computed result.

3 Stochastic approach of round-off errors

3.1 The CESTAC method

The CESTAC (Contrôle et Estimation Stochastique des Arrondis de Calculs) method, which has been developed by La Porte and Vignes [10,12,13], enables one to estimate the number of exact significant digits of any computed result. This method is based on a probabilistic approach of round-off errors using a random rounding mode defined below.

Definition 9 *Each real number x, which is not a floating-point number, is bounded by two consecutive floating-point numbers: X^- (rounded down) and X^+ (rounded up). The random rounding mode defines the floating-point number X representing x as being one of the two values X^- or X^+ with the probability* $1/2$.

With this random rounding mode, the same program run several times provides different results, due to different round-off errors.

It has been proved [2] that a computed result R is modelled to the first order in 2^{-p} as:

$$R \approx Z = r + \sum_{i=1}^{n} g_i(d) 2^{-p} z_i \tag{31}$$

where r is the exact result, $g_i(d)$ are coefficients depending exclusively on the data and on the code, p is the number of bits in the mantissa and z_i are independent uniformly distributed random variables on $[-1, 1]$.

From equation (31), we deduce that:

(1) the mean value of the random variable Z is the exact result r,
(2) under some assumptions, the distribution of Z is a quasi-Gaussian distribution.

Then by identifying R and Z, *i.e.* by neglecting all the second order terms, Student's test can be used to determine the accuracy of R. Thus from N samples R_i, $i = 1, 2, ..., N$, the number of decimal significant digits common to $\overline{R}$ and r can be estimated with the following equation.

$$C_{\overline{R}} = \log_{10} \left(\frac{\sqrt{N} \left| \overline{R} \right|}{\sigma \tau_\beta} \right), \tag{32}$$

where

$$\overline{R} = \frac{1}{N} \sum_{i=1}^{N} R_i \quad \text{and} \quad \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(R_i - \overline{R} \right)^2. \tag{33}$$

τ_β is the value of Student's distribution for $N - 1$ degrees of freedom and a probability level $1 - \beta$.

Thus the implementation of the CESTAC method in a code providing a result R consists in:

- performing N times this code with the random rounding mode, which is obtained by using randomly the rounding mode towards $-\infty$ or $+\infty$; we then obtain N samples R_i of R
- choosing as the computed result the mean value $\overline{R}$ of R_i, $i = 1, ..., N$
- estimating with equation (32) the number of exact decimal significant digits of $\overline{R}$.

In practice $N = 2$ or $N = 3$ and $\beta = 0.05$. Note that for $N = 2$, then $\tau_\beta = 12.706$ and for $N = 3$, then $\tau_\beta = 4.4303$.

Equations (31) and (32) hold if two main hypotheses are verified. These hypotheses are:

(1) the round-off errors α_i are independent, centered uniformly distributed random variables,
(2) the approximation to the first order in 2^{-p} is legitimate.

Concerning the first hypothesis, with the use of the random arithmetic, round-off errors α_i are random variables, however, in practice, they are not rigorously centered and in this case Student's test gives a biased estimation of the computed result. It has been proved [6] that, with a bias of a few σ, the error on the estimation of the number of exact significant digits of $\overline{R}$ is less than one decimal digit. Therefore even if the first hypothesis is not rigorously satisfied, the reliability of the estimation obtained with equation (32) is not altered if it is considered as exact up to one digit.

Concerning the second hypothesis, the approximation to the first order only concerns multiplications and divisions. Indeed the round-off error generated by an addition or a subtraction does not contain any term of higher order. It has been shown [2,4] that, if a computed result becomes insignificant, *i.e.* if the round-off error it contains is of the same order of magnitude as the result itself, then the first order approximation may be not legitimate. In practice the validation of the CESTAC method requires a dynamic control of multiplications and divisions, during the execution of the code. This leads to the synchronous implementation of the method, *i.e.* to the parallel computation of the N samples R_i, and also to the concept of computational zero, also named informatical zero [11].

Definition 10 *During the run of a code using the CESTAC method, an intermediate or a final result R is a computational zero, denoted by @.0, if one of the two following conditions holds:*

- $\forall i, R_i = 0$,
- $C_{\overline{R}} \leq 0$.

Any computed result R is a computational zero if either $R = 0$, R being significant, or R is insignificant. A computational zero is a value that cannot be differentiated from the mathematical zero because of its round-off error.

From the synchronous implementation of the CESTAC method and the concept of computational zero, stochastic arithmetic [4,7,13] has been defined. Two types of stochastic arithmetic actually exist: it can be either continuous or discrete.

3.2 Principles of stochastic arithmetics

3.2.1 Continuous stochastic arithmetic

Continuous stochastic arithmetic is a modelling of the synchronous implementation of the CESTAC method. By using this implementation, so that the N runs of a code take place in parallel, the N results of each arithmetical operation can be considered as realizations of a Gaussian random variable centered on the exact result. One can therefore define a new number, called *stochastic number*, and a new arithmetic, called *(continuous) stochastic arithmetic*, applied to these numbers. An equality concept and order relations, which take into account the number of exact significant digits of stochastic operands, have also been defined.

A stochastic number X is denoted by (m, σ^2), where m is the mean value of X and σ its standard deviation. Stochastic arithmetical operations ($s+$, $s-$, $s\times$, $s/$) correspond to terms to the first order in $\frac{\sigma}{m}$ of operations between two independent Gaussian random variables.

Definition 11 *Let $X_1 = (m_1, \sigma_1^2)$ and $X_2 = (m_2, \sigma_2^2)$. Stochastic arithmetical operations on X_1 and X_2 are defined as:*

$$X_1 \; s+ \; X_2 \; = \; \left(m_1 + m_2 \; , \; \sigma_1^2 + \sigma_2^2\right) \tag{34}$$

$$X_1 \; s- \; X_2 \; = \; \left(m_1 - m_2 \; , \; \sigma_1^2 + \sigma_2^2\right) \tag{35}$$

$$X_1 \; s\times \; X_2 \; = \; \left(m_1 \times m_2 \; , \; m_2^2\sigma_1^2 + m_1^2\sigma_2^2\right) \tag{36}$$

$$X_1 \; s/ \; X_2 \; = \; \left(m_1/m_2 \; , \; \left(\frac{\sigma_1}{m_2}\right)^2 + \left(\frac{m_1\sigma_2}{m_2^2}\right)^2\right) \; with \;\; m_2 \neq 0. \tag{37}$$

An accuracy can be associated to any stochastic number. If $X = (m, \sigma^2)$, λ_β exists (depending only on β) such that

$$P\left(X \in [m - \lambda_\beta\sigma, m + \lambda_\beta\sigma]\right) = 1 - \beta, \tag{38}$$

$I_{\beta,X} = [m - \lambda_\beta\sigma, m + \lambda_\beta\sigma]$ is the confidence interval of m at $1-\beta$. The number of decimal significant digits common to all the elements of $I_{\beta,X}$ and to m is lower bounded by

$$C_{\beta,X} = \log_{10}\left(\frac{|m|}{\lambda_\beta\sigma}\right). \tag{39}$$

The following definition is the modelling of the concept of computational zero, previously introduced.

Definition 12 *A stochastic number X is a stochastic zero, denoted by $\underline{0}$, if and only if*

$$C_{\beta,X} \leq 0 \quad or \quad X = (0,0).$$

In accordance with the concept of stochastic zero, a new equality concept and new order relations have been defined.

Definition 13 *Let $X_1 = (m_1, \sigma_1^2)$ and $X_2 = (m_2, \sigma_2^2)$ be two stochastic numbers.*

- *Stochastic equality, denoted by $s=$, is defined as:*
  $X_1 \; s= \; X_2$ *if and only if* $X_1 \; s- \; X_2 = \underline{0}$.
- *Stochastic inequalities, denoted by $s>$ and $s\geq$ are defined as:*
  $X_1 \; s> \; X_2$ *if and only if* $m_1 > m_2$ *and* $X_1 \; s\neq \; X_2$,
  $X_1 \; s\geq \; X_2$ *if and only if* $m_1 \geq m_2$ *or* $X_1 \; s= \; X_2$.

Continuous stochastic arithmetic is a modelling of the computer arithmetic, which takes into account round-off errors. The properties of continuous stochastic arithmetic [3,4] have pointed out the theoretical differences between the approximative arithmetic of a computer and exact arithmetic.

3.2.2 Discrete Stochastic Arithmetic

Discrete Stochastic Arithmetic (DSA) has been defined from the synchronous implementation of the CESTAC method. With DSA, a real number becomes an N-dimensional set and any operation on these N-dimensional sets is performed element per element using the random rounding mode. The number of exact significant digits of such an N-dimensional set can be estimated from equation (32). From the concept of computational zero previously introduced, an equality concept and order relations have been defined for DSA.

Definition 14 *Let X and Y be N-samples provided by the CESTAC method.*

- *Discrete stochastic equality denoted by $ds=$ is defined as:*
  $X ds= Y$ *if and only if* $X - Y = @.0$.
- *Discrete stochastic inequalities denoted by $ds>$ and $ds\geq$ are defined as:*
  $X ds> Y$ *if and only if* $\overline{X} > \overline{Y}$ *and* $X ds\neq Y$,
  $X ds\geq Y$ *if and only if* $\overline{X} \geq \overline{Y}$ *or* $X ds= Y$.

Order relations in DSA are essential to control branching statements. Because of round-off errors, if A and B are two floating-point numbers and a and b the corresponding exact values,

$$a > b \not\Rightarrow A > B \quad \text{and} \quad A > B \not\Rightarrow a > b.$$

Many problems in scientific computing are due to this dis-correlation: for example, unsatisfied stopping criteria or infinite loops in algorithmic geometry. Taking into account the numerical quality of the operands in order relations enables to partially solve these problems [3].

Therefore DSA enables to estimate the impact of round-off errors on any result of a scientific code and also to check that no anomaly occurred during the run, especially in branching statements. DSA is implemented in the CADNA library [1].

The accuracy of a stochastic number can be related to the number of exact significant digits of an N-sample provided by the CESTAC method. Indeed, when N is a small value (2 or 3), which is the case in practice, the values obtained with equations (32) and (39) are very close. They represent in a computed result the number of significant digits which are not affected by round-off errors. So the two types of stochastic arithmetics are coherent. Properties established in the theoretical framework of continuous stochastic arithmetic can be applied on a computer via the practical use of DSA.

3.3 Theoretical results on stochastic operations

The theoretical results presented here have been established in continuous stochastic arithmetic. They enable one to compare results of arithmetical stochastic operations with those provided by the corresponding classical operations performed on exact values.

Let us consider a numerical method which aims to approximate an exact value x_1. This method may consist for example in computing an iterate of a sequence (u_n) such that $\lim_{n\to\infty} u_n = x_1$. Even using an arithmetic with infinite precision, the value obtained is not x_1, but an approximation which is affected by a truncation error. We compare here the results obtained using such numerical methods in stochastic arithmetic with the exact values they approximate.

Theorem 15 *Let $X_1 = (m_1, \sigma_1^2)$ be the approximation of an exact value x_1 in stochastic arithmetic. Let us assume that the exact significant bits of X_1, i.e. not affected by round-off errors, are in common with x_1, up to p: the number of significant bits of X_1 in common with x_1 is lower bounded by $log_2\left(\frac{|m_1|}{\lambda_\beta \sigma_1}\right) - p$.*

Similarly let $X_2 = (m_2, \sigma_2^2)$ be an approximation obtained in stochastic arithmetic of an exact value x_2, such that its exact significant bits are in common

[1] URL address: http://www.lip6.fr/cadna/

with x_2, up to q.

Let $\bigcirc$ be an exact arithmetical operator: $\bigcirc \in \{+, -, \times, /\}$ and $s\bigcirc$ the corresponding stochastic operator $s\bigcirc \in \{s+\,, s-\,, s\times\,, s/\}$.

Then the exact significant bits of $X_1\ s\bigcirc\ X_2$ are in common with the exact value $x_1 \bigcirc x_2$, up to $max(p, q)$.

PROOF. From equation (39), the number of exact significant bits of X_1, *i.e.* not affected by round-off errors, is lower bounded by $\log_2\left(\frac{|m_1|}{\lambda_\beta \sigma_1}\right)$. As the number of significant bits of X_1 in common with the exact value x_1 is lower bounded by $log_2\left(\frac{|m_1|}{\lambda_\beta \sigma_1}\right) - p = log_2\left(\frac{|m_1|}{2^p \lambda_\beta \sigma_1}\right)$, to take into account both the truncation error and the round-off error on X_1, one has to consider not the variance σ_1^2, but $(2^p\sigma_1)^2$.

Similarly the number of significant bits of X_2 in common with the exact value x_2 is lower bounded by $log_2\left(\frac{|m_2|}{\lambda_\beta \sigma_2}\right) - q = log_2\left(\frac{|m_2|}{2^q \lambda_\beta \sigma_2}\right)$.

From equations (34) and (39), the number of exact significant bits of $X_1 s+ X_2$ is lower bounded by $log_2\left(\frac{|m_1+m_2|}{\lambda_\beta\sqrt{\sigma_1^2+\sigma_2^2}}\right)$. To take into account both the truncation error and the round-off error on $X_1 s+ X_2$, one has to consider not the variance $\sigma_1^2 + \sigma_2^2$, but $(2^p\sigma_1)^2 + (2^q\sigma_2)^2$. Therefore a lower bound for the number of significant bits of $X_1 s+ X_2$ in common with the exact value $x_1 + x_2$ is $log_2\left(\frac{|m_1+m_2|}{\lambda_\beta\sqrt{(2^p\sigma_1)^2+(2^q\sigma_2)^2}}\right)$, which can be itself lower bounded by $log_2\left(\frac{|m_1+m_2|}{\lambda_\beta\sqrt{\sigma_1^2+\sigma_2^2}}\right) - max(p, q)$. Then the exact significant bits of $X_1 s+ X_2$ are in common with $x_1 + x_2$, up to $max(p, q)$.

As $X_1 s- X_2 = (m_1 - m_2,\ \sigma_1^2 + \sigma_2^2)$, the proof for the subtraction is similar as the one for the addition.

From equations (36) and (39), the number of exact significant bits of $X_1 s\times X_2$ is lower bounded by $log_2\left(\frac{|m_1 m_2|}{\lambda_\beta\sqrt{m_2\sigma_1^2+m_1\sigma_2^2}}\right)$. To take into account both the truncation error and the round-off error on $X_1 s\times X_2$, one has to consider not the variance $m_2\sigma_1^2 + m_1\sigma_2^2$, but $2^{2p}m_2\sigma_1^2 + 2^{2q}m_1\sigma_2^2$. Therefore a lower bound for the number of significant bits of $X_1 s\times X_2$ in common with the exact value $x_1 \times x_2$ is $log_2\left(\frac{|m_1 m_2|}{\lambda_\beta\sqrt{2^{2p}m_2\sigma_1^2+2^{2q}m_1\sigma_2^2}}\right)$, which can be itself lower bounded by $log_2\left(\frac{|m_1 m_2|}{\lambda_\beta\sqrt{m_2\sigma_1^2+m_1\sigma_2^2}}\right) - max(p, q)$. Then the exact significant bits of $X_1 s\times X_2$ are in common with $x_1 \times x_2$, up to $max(p, q)$.

From equations (37) and (39), the number of exact significant bits of X_1s/X_2 is lower bounded by $log_2\left(\frac{|\frac{m_1}{m_2}|}{\lambda_\beta\sqrt{(\frac{\sigma_1}{m_2})^2+(\frac{m_1\sigma_2}{m_2^2})^2}}\right)$. To take into account both the truncation error and the round-off error on X_1s/X_2, one has to consider not the variance $(\frac{\sigma_1}{m_2})^2+(\frac{m_1\sigma_2}{m_2^2})^2$, but $(\frac{2^p\sigma_1}{m_2})^2+(\frac{2^qm_1\sigma_2}{m_2^2})^2$. Therefore a lower bound for the number of significant bits of X_1s/X_2 in common with the exact value x_1/x_2 is $log_2\left(\frac{|\frac{m_1}{m_2}|}{\lambda_\beta\sqrt{(\frac{2^p\sigma_1}{m_2})^2+(\frac{2^qm_1\sigma_2}{m_2^2})^2}}\right)$, which can be itself lower bounded by $log_2\left(\frac{|\frac{m_1}{m_2}|}{\lambda_\beta\sqrt{(\frac{\sigma_1}{m_2})^2+(\frac{m_1\sigma_2}{m_2^2})^2}}\right)-max(p,q)$. Then the exact significant bits of X_1s/X_2 are in common with x_1/x_2, up to $max(p,q)$.

Theorem 15 enables one to control arithmetical operations performed on computed results of numerical methods. This theorem has been proved for stochastic arithmetical operations, which are a modelling of the operations performed in the synchronous implementation of the CESTAC method. In practice, theorem 15 is used, according to 3.2.2, for results obtained in DSA. In the next section, we present, in accordance with theorem 15 and the theoretical results presented in section 2, a strategy to dynamically control converging sequences computed in DSA.

4 A strategy for a dynamical control of converging sequences

When a numerical algorithm requires the evaluation of the limit of a sequence, this limit is approximated by one of the iterates. As the number of iterations increases, the truncation error usually decreases, but the round-off error increases. Therefore the choice of the optimal iterate may be problematic.

DSA enables one to estimate the number of exact significant digits of any computed result, *i.e.* its significant digits which are not affected by round-off error propagation. Let us consider the computation of a sequence (I_n) in DSA and let us assume that the convergence zone is reached. If discrete stochastic equality is achieved for two successive iterates, *i.e.* $I_n - I_{n+1} = @.0$, the difference between I_n and I_{n+1} is only due to round-off errors and further iterations are useless. The optimal iterate I_{n+1} can therefore be dynamically determined at run time. Furthermore, if the sequence (I_n) converges at least linearly to I, from section 2, the exact significant digits of I_{n+1} are in common with I, up to k digits. The value k, which depends on the convergence speed of (I_n), can be determined from theorem 5 or 7.

Let us consider a sequence generated using the trapezoidal or Simpson's rule with the technique of step halving previously described. If the convergence zone is reached and computations are performed until the difference between two successive iterates is insignificant, then, from section 2, the exact significant bits of the last iterate are in common with the exact value of the integral, up to one.

More generally, if a sequence (I_n) converging at least linearly to I is computed using DSA, the optimal iterate can be dynamically determined and the number of significant digits it has in common with the exact limit I can be evaluated. If operations on limits of sequences are required in a numerical algorithm, a similar strategy, based on the following theorem, can be used.

Theorem 16 *Let us consider the computation in DSA of two sequences (I_k) and (J_k) converging at least linearly to I and J respectively.*

Let I_n (respectively J_m) be an iterate such that its exact significant bits are in common with I up to p (respectively J up to q).

If we denote by $\bigcirc$ an exact arithmetical operator, then the exact significant bits of $I_n \bigcirc J_m$ are in common with the exact value $I \bigcirc J$, up to $max(p, q)$.

PROOF. From section 2, as the sequence (I_k) converges at least linearly to I, if it is computed until the difference between two successive iterates is insignificant, *i.e.* $I_{n-1} - I_n = @.0$, then we can determine the value p such that the exact significant bits of I_n are in common with I, up to p. Similarly if the sequence (J_k) is computed until $J_{m-1} - J_m = @.0$, then we can determine the value q such that the exact significant bits of J_m are in common with J, up to q. According to the application of theorem 15 in DSA, if an arithmetical operation is performed on I_n and J_m, the exact significant bits of the result are those obtained with the same operation performed on I and J, up to $max(p, q)$.

Remark 17 *According to section 2, if the convergence of the sequences (I_k) and (J_k) is sufficiently fast, then $p = q = 1$. In this case, the exact significant bits of the result obtained are those provided by the same operation on the limits, up to one.*

More generally, in a numerical algorithm involving the computation of several sequences, if each sequence is computed until the difference between two successive iterates is insignificant, each limit is approximated by the optimal iterate. According to section 2, if each sequence converges at least linearly, we can evaluate the number of significant digits common between the limit and its approximation. If arithmetical operations are performed on these approxi-

mations, we can determine the significant digits of the result obtained which are common with the result of the same operations performed on the limits.

5 Dynamical control of combined sequences

This section shows how to approximate the limit of a sequence by its optimal iterate, this iterate being itself the limit of another sequence. The theorems presented in sections 2 and 3 can be combined to determine the number of digits of the approximation obtained which are in common with the exact result. In the strategies described in this section, small letters denote exact values and capital letters the corresponding approximations computed using DSA.

5.1 A strategy to compute combined sequences

We consider a sequence in which each term u_m is the limit of another sequence. More precisely, let (u_m) be a sequence converging at least linearly to u and, for all m, let $(u_{m,n})$ be a sequence converging at least linearly to u_m.

For all m, let U_m be the approximation of u_m computed using DSA. U_m is obtained by computing the sequence $(u_{m,n})$ until, in the convergence zone, the difference between two successive iterates is insignificant.

As for all m, the sequence $(u_{m,n})$ converges at least linearly to u_m, according to section 2, one can determine the value q such that the exact significant bits of U_m are common to u_m, up to q.

Figure 1 represents the significant bits of U_m and U_{m+1} if the difference $U_m - U_{m+1}$ is insignificant. In this case, the exact significant bits of U_{m+1} are common to U_m and are also common to u_m and u_{m+1}, up to q.

As the sequence (u_m) converges at least linearly to u, one can determine the value p such that the bits common to u_m and u_{m+1} are common with u, up to p.

Consequently if the difference $U_m - U_{m+1}$ is insignificant, the exact significant bits of U_{m+1} are common with u, up to $p + q$.

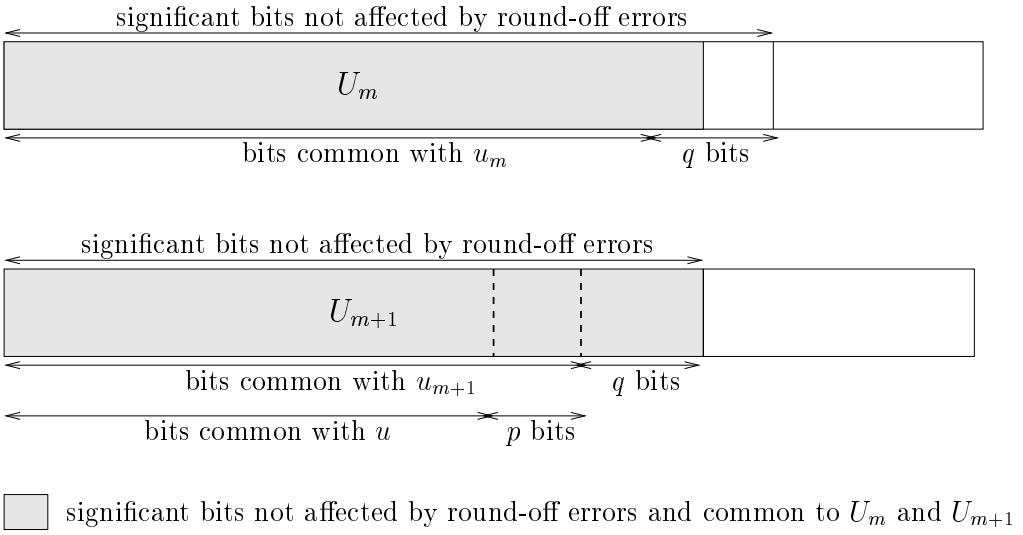

Fig. 1. Significant bits of U_m and U_{m+1}

5.2 Dynamical control of integrals on an infinite domain

Let us consider the computation of an improper integral $g = \int_0^\infty \phi(x)dx$. The infinite interval of integration is partitioned into finite intervals of length L. Let $f_j = \int_{jL}^{(j+1)L} \phi(x)dx$ and $g_m = \sum_{j=0}^{m} f_j$, $\lim_{m\to\infty} g_m = g$.

g can be numerically approximated by an iterate g_m, m being sufficiently high. The optimal number of iterates to compute can be determined dynamically using DSA.

Let $F_{j,n}$ be the approximation of f_j computed using the trapezoidal or Simpson's rule with step $\frac{L}{2^n}$. For all j, the sequence $(F_{j,n})$ is computed until the difference between two successive iterates is insignificant. This is not achieved at the same iteration of all values of j. Let n_j be the iteration at which $F_{j,n_j-1} - F_{j,n_j} = @.0$.

According to section 2, for all j, the exact significant bits of F_{j,n_j} are in common with f_j, up to one. Let $G_m = \sum_{j=0}^{m} F_{j,n_j}$. According to theorem 16, the exact significant bits of G_m are in common with g_m, up to one.

Figure 2 represents the significant bits of G_m and G_{m+1} if the difference $G_m - G_{m+1}$ is insignificant. In this case, the exact significant bits of G_{m+1} are common to G_m and are also common to g_m and g_{m+1}, up to one.

We assume that the sequence (g_m) converges at least linearly to g. According to section 2, if the convergence zone is reached, $C_{g_m,g_{m+1}} = C_{g_m,g} + \gamma$ where γ represents p bits. Therefore the bits common to g_m and g_{m+1} are common with g, up to p.

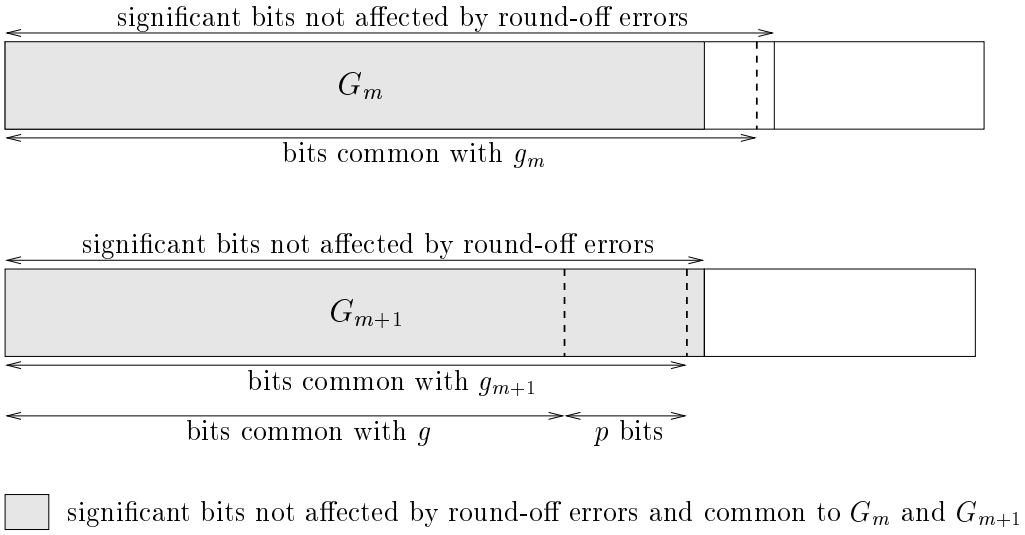

Fig. 2. Significant bits of G_m and G_{m+1}

Consequently if the difference $G_m - G_{m+1}$ is insignificant, the exact significant bits of G_{m+1} are common with g, up to $p + 1$.

6 Numerical experiments

Numerical experiments have been carried out using DSA implemented in the CADNA library. Two examples are presented: the computation of a definite integral and the computation of an integral on an infinite interval.

6.1 Computation of a definite integral

Let us consider the integral $I = \int_0^1 \frac{6x^3 - 15x^2 - 28x + 22}{9x^2 + 12x + 4} dx = 1.$

I has been estimated with the trapezoidal and Simpson's rules using the strategy described in section 2. Approximations I_n have been computed with step $\frac{1}{2^n}$ until the difference $I_n - I_{n+1}$ is insignificant. From section 2, we can guarantee that the exact significant bits of the last iterate I_N are in common with the exact value of I, up to one.

Table 1 presents for both rules the approximations of I obtained in single and double precision. The number of exact significant digits of each result has been estimated using DSA. For each sequence, the exact significant digits of the last iterate are reported in table 1.

We can notice that the exact significant digits of each approximation obtained

Table 1
Approximations of I

rule	in single precision	in double precision
trapezoidal	$I_9 = 0.10000E+01$	$I_{21} = 0.100000000000E+001$
Simpson	$I_8 = 0.100000E+01$	$I_{13} = 0.1000000000000E+001$

are in common with I. The number of iterations requested for the stopping criterion to be satisfied depends of course on the precision chosen, but also on the quadrature method used. Whatever the precision is, less iterations are performed with Simpson's rule than with the trapezoidal rule. This is due to the different convergence speeds of the computed sequences. Indeed the approximation of I is of order 2 with the trapezoidal rule and of order 4 with Simpson's rule. For each rule, the error on the last iterate $|I_N - I|$ is insignificant. Because of round-off error propagation, the computer can not distinguish I_N from I.

6.2 Computation of an improper integral

Let us consider the improper integral $g = \int_0^\infty e^{-ax}\,dx = \frac{1}{a}$, where $a > 0$.

g has been estimated using the strategy described in 5.2. Using the same notations as in 5.2, let $g_m = \sum_{j=0}^{m} f_j$, where $f_j = \int_{jL}^{(j+1)L} e^{-ax}\,dx$. The approximations of the integrals f_j are computed with Simpson's rule using DSA. For every j, a sequence is computed until the difference between two successive iterates is insignificant.

As $g_m - g = \int_{(m+1)L}^{\infty} e^{-ax}\,dx = \frac{\alpha^{m+1}}{a}$, where $\alpha = e^{-aL}$, the sequence (g_m) converges linearly to g. Therefore theorem 5 can apply: if the convergence zone is reached, the significant bits common to two successive iterates are also common to g, up to $log_2(\frac{1}{1-\alpha})$.

Let G_m be the approximation of g_m computed using DSA. The sequence (G_m) is computed until the difference between two successive iterates is insignificant. We denote by M the iteration at which $G_{M-1} - G_M = @.0$. According to section 5.2, the exact significant bits of G_M are in common with g, up to $log_2(\frac{1}{1-\alpha}) + 1$. Therefore the exact significant decimal digits of G_M are in common with g up to δ, where $\delta = log_{10}(\frac{2}{1-\alpha})$.

Table 2 presents for $a = 1$ and different values of L the approximations G_M obtained in double precision. The number of exact significant digits of G_M not in common with g is approximated by δ. As the length L increases, the number M of integrals f_j to be approximated decreases. Only the exact significant

digits of G_M are reported: the other significant digits are affected by round-off error propagation. We notice that the number of exact significant digits obtained (from thirteen to fifteen) is satisfying for computations carried out in double precision. The exact significant digits which are not in common with the exact value $g = 1$ can easily be identified. For example, if $L = 10^{-1}$, among the fourteen exact significant digits of G_M, the two last digits are not in common with g. We notice that, for every approximation G_M reported in table 2, its exact significant digits are in common with g up to $\lceil \delta \rceil$.

Table 2
Results obtained with Simpson's rule for $a = 1$

L	$\delta \approx$	M	G_M
10^{-2}	2.3	2335	0.9999999999276E+000
10^{-1}	1.3	284	0.99999999999953E+000
1	0.5	33	0.999999999999996E+000
10	0.3	4	0.99999999999999E+000
50	0.3	2	0.10000000000004E+001

Table 3 presents for $a = 10^{-5}$ and different values of L the exact significant digits of the approximations G_M obtained in double precision. As in table 2, we notice that if the length L increases, the number M of integrals f_j to be approximated decreases. For each approximation G_M obtained, we can easily identify its exact significant digits which are in common with the exact value $g = 10^5$. As in table 2, we notice that the exact significant digits of G_M are in common with g up to $\lceil \delta \rceil$.

Table 3
Results obtained with Simpson's rule for $a = 10^{-5}$

L	$\delta \approx$	M	G_M
10^2	3.3	19136	0.999999995109E+005
10^3	2.3	2346	0.9999999999352E+005
10^4	1.3	279	0.99999999999923E+005
10^5	0.5	33	0.999999999999995E+005
10^6	0.3	5	0.99999999999999E+005

7 Conclusion

Discrete Stochastic Arithmetic can be used to dynamically determine the optimal iterate of a converging sequence. Furthermore, if the sequence converges

at least linearly, the number of significant digits of this iterate common with the limit can be estimated. This number depends on the speed of convergence of the sequence.

If an arithmetical operation is performed on the optimal iterates of two sequences, we can determine the significant digits of the computed result common with the exact result of the same operation performed on the two limits. This allows a dynamical control of numerical algorithms involving the computation of several sequences. Integrals on an infinite interval can be approximated by computing several converging sequences. By controlling dynamically each sequence, we can determine the significant digits of the approximation common with the exact value of the integral.

The sequences examined in this paper all converge to a scalar value. A perspective to this work could be the numerical validation of sequences of vectors involved for example in iterative methods for solving linear systems.

References

[1] R. L. Burden and J. D. Faires, Numerical analysis, 7th ed., Brooks-Cole Publishing, 2001.

[2] J.-M. Chesneaux, Study of the computing accuracy by using probabilistic approach, in: *Contribution to computer arithmetic and self-validating numerical methods*, C. Ullrich ed., IMACS, New Brunswick, NJ, 1990, pp. 19-30.

[3] J.-M. Chesneaux, The equality relations in scientific computing, *Num. Algo.* 7 (1994) 129-143.

[4] J.-M. Chesneaux, L'arithmétique stochastique et le logiciel CADNA, Habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, 1995.

[5] J.-M. Chesneaux and F. Jézéquel, Dynamical control of computations using the Trapezoidal and Simpson's rules, *J. Univ. Comput. Sci.* 4 (1) (1998) 2-10.

[6] J.-M. Chesneaux and J. Vignes, Sur la robustesse de la méthode CESTAC, *C. R. Acad. Sci. Paris Sér. I Math.* 307 (1988) 855-860.

[7] J.-M. Chesneaux and J. Vignes, Les fondements de l'arithmétique stochastique, *C. R. Acad. Sci. Paris Sér. I Math.* 315 (1992) 1435-1440.

[8] M. K. Jain, R. K. Jain and S. R. K. Iyengar, Numerical methods for scientific and engineering computation, Halsted Press, 1985.

[9] J. H. Mathews, Numerical methods for mathematics, science and engineering, 2nd ed., Prentice-Hall, 1992.

[10] J. Vignes and M. La Porte, Error analysis in computing, in: *Information Processing 74*, North-Holland, 1974.

[11] J. Vignes, Zéro mathématique et zéro informatique, *C. R. Acad. Sci. Paris Sér. I Math.* 303 (1986) 997-1000; also: *La Vie des Sciences* 4 (1) (1987) 1-13.

[12] J. Vignes, Estimation de la précision des résultats de logiciels numériques, *La Vie des Sciences* 7 (2) (1990) 93-145.

[13] J. Vignes, A stochastic arithmetic for reliable scientific computation, *Math. Comput. Simulation* 35 (1993) 233-261.

[14] J. Vignes, A stochastic approach to the analysis of round-off error propagation. A survey of the CESTAC method. in: *Proc. 2nd Real Numbers and Computers conference*, Marseille, France, 1996, pp. 233-251.