Dynamical control of converging sequences using Discrete Stochastic Arithmetic

F. Jézéquel
Laboratoire d'Informatique de Paris 6 - CNRS UMR 7606,
4 place Jussieu, 75252 Paris cedex 05, France
Fabienne.Jezequel@lip6.fr

Abstract

On a computer, the optimal number of iterations of a converging sequence can be determined dynamically using Discrete Stochastic Arithmetic. Computations are performed until the difference between two successive iterates is not significant. If the sequence converges at least linearly, we can estimate the significant digits of the approximation common with the exact limit. This strategy can be used for the computation of integrals with the trapezoidal or Simpson's method. 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 those of the exact value of the integral, up to one bit. Numerical algorithms involving several sequences, such as the approximation of integrals on an infinite interval, can also be dynamically controlled.

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

1 Introduction

Many approximation methods involve the computation of a converging sequence. The limit is then approximated by one of the iterates. It is often difficult to determine the optimal iterate, *i.e.* the approximation for which the global error, consisting of the truncation error and the round-off error, is minimal. In section 2, we recall methods and concepts which enable to estimate round-off error propagation: the CESTAC method, the principles of stochastic arithmetic and finally Discrete Stochastic Arithmetic (DSA). The theorems presented in section 3 enable to control both the truncation error and the round-off error in the computation of a converging sequence. We describe a strategy to compute dynamically the optimal iterate of a converging sequence using DSA. Furthermore we can determine in the approximation obtained which exact significant digits, *i.e.* not affected by round-off errors, are common with the expected limit. Under some assumptions on the speed of convergence of the sequence, we show that the exact significant bits of the result obtained are those of the limit, up to one. In section 4, we present theoretical results established in stochastic arithmetic and their application to the control of arithmetical operations on sequences computed using DSA. 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 Principles of stochastic arithmetic

2.1 The CESTAC method

The numerical quality of a computed result R can be measured by its number of exact significant digits, which is the number of significant digits it has in common with the exact result r, more precisely:

Definition 1 *The number of significant digits in common between two real numbers R and r is defined in $\mathbb{R}$ by*

- *for $R \neq r$, $C_{R,r} = \log_{10} \left| \frac{R+r}{2(R-r)} \right|$,*
- *$\forall r \in \mathbb{R},\ C_{r,r} = +\infty$.*

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

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

The CESTAC (Contrôle et Estimation Stochastique des Arrondis de Calculs) method, which has been developed by La Porte and Vignes [9, 11, 12], enables one to estimate $C_{R,r}$ without any information about r, using a probabilistic approach of round-off errors.
We define below the random rounding mode.

Definition 2 *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 modelized to the first order in 2^{-p} as:

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

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 centered in $[-1, 1]$.

From equation (1), we deduce that:

1. the mean value of the random variable R is the exact result r,
2. the distribution of R is a quasi-Gaussian distribution.

Then to determine the accuracy of R, Student's test can be used. 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.

$$C_{\overline{R}} = \log_{10}\left(\frac{\sqrt{N}\left|\overline{R}\right|}{\sigma \tau_\beta}\right), \tag{2}$$

where

$$\overline{R} = \frac{1}{N}\sum_{i=1}^{N} R_i \quad \text{and} \quad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2. \tag{3}$$

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. 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 (2) 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 $\tau_\beta = 4.4303$.

In order to validate the method, the theoretical study forces to be able to estimate at any time the numerical quality of some intermediate results. This leads to the synchronous implementation of the method, *i.e.* to the parallel computation of the N samples R_i. In practice 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.

2.2 Stochastic arithmetic

Stochastic arithmetic [4, 6, 12] is a modelization 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 *stochastic arithmetic*, applied to these numbers. An equality concept and order relations, which take into account the number of 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 $\frac{\sigma}{m}$ between two independent Gaussian random variables.

Definition 3 *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:*

$$X_1 \ s+ \ X_2 = \left(m_1 + m_2 \ , \ \sigma_1^2 + \sigma_2^2\right) \quad (4)$$

$$X_1 \ s- \ X_2 = \left(m_1 - m_2 \ , \ \sigma_1^2 + \sigma_2^2\right) \quad (5)$$

$$X_1 \ s\times \ X_2 = \left(m_1 \times m_2 \ , \ m_2^2\sigma_1^2 + m_1^2\sigma_2^2\right) \quad (6)$$

$$X_1 \ s/ \ X_2 = \left(m_1/m_2 \ , \ \left(\frac{\sigma_1}{m_2}\right)^2 + \left(\frac{m_1\sigma_2}{m_2^2}\right)^2\right) \ \textit{with} \ \ m_2 \neq 0. \quad (7)$$

If $X = (m, \sigma^2)$, λ_β exists (depending only on β) such that

$$P\left(X \in [m - \lambda_\beta\sigma, m + \lambda_\beta\sigma]\right) = 1 - \beta, \quad (8)$$

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

$$C_{\beta,X} = \log_{10}\left(\frac{|m|}{\lambda_\beta\sigma}\right). \quad (9)$$

When N is a small value (2 or 3), which is the case in practice, the values obtained with equations (2) and (9) are very close. This remark is important for the use of the concept of stochastic arithmetic via the practical use of the CESTAC method.

2.3 Discrete Stochastic Arithmetic

The synchronous implementation of the CESTAC method is essential to control branching statements. Because of round-off errors, if A and B are two computed results 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: unsatisfied stopping criterion, infinite loop in algorithmic geometry... Taking into account the numerical quality of the operands in order relations enables to solve these problems partially [3]. This requires a very important concept: the computational zero, also named informatical zero [10].

Definition 4 *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 not significant.

A computational zero is a value that cannot be differentiated from the mathematical zero because of its round-off error. From this concept, discrete stochastic relations have been defined.

Definition 5 *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* $X - Y = @.0$
- *Discrete stochastic inequalities denoted by ds> and ds≥ are defined as:*
  $X ds{>} Y$ *if* $\overline{X} > \overline{Y}$ *and* $X - Y \neq @.0$,
  $X ds{\geq} Y$ *if* $\overline{X} \geq \overline{Y}$ *or* $X - Y = @.0$.

Discrete Stochastic Arithmetic (DSA) is the joint use on a computer of the synchronous implementation of the CESTAC method, the concept of computational zero and the discrete stochastic relations. 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].

3 A strategy for a dynamical control of converging sequences

When a numerical algorithm requires the evaluation of the limit I of a sequence (I_n), this limit is approximated by one of the iterates. The number of iterations performed depends on the convergence speed of the sequence, which can be either logarithmic, linear or exponential. 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. For sequences which

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

converge at least linearly, the theorems presented in this section enable one by comparing the significant digits common to two successive iterates I_n and I_{n+1} to determine the significant digits common to I_n and the limit I. Furthermore if round-off errors can be estimated, the optimal iterate can be dynamically determined and the number of significant digits it has in common with the limit I can be evaluated.

3.1 Sequences with a linear convergence

Let us consider two successive iterates of a sequence: I_n and I_{n+1}. The number of significant digits common to these iterates, $C_{I_n,I_{n+1}}$, can be easily measured. If the convergence speed of the sequence is linear, the following theorem enables one to deduce the number of significant digits common to I_n and the limit, $C_{I_n,I}$, from $C_{I_n,I_{n+1}}$. This theorem has been established by taking into account the truncation error on two successive iterates, but not the round-off error occurring during their computation.

Theorem 1 *Let (I_n) be a sequence converging linearly to I, i.e. which satisfies $I_n - I = C\alpha^n + o(\alpha^n)$ where $C \in \mathbb{R}$ 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:

$$I_n - I = C\alpha^n + o(\alpha^n) \tag{10}$$

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

$$I_n - I_{n+1} = C\alpha^n(1-\alpha) + o(\alpha^n) \tag{11}$$

From equation (10), we deduce

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^n\,(1 + o(1))} \tag{12}$$

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^n}\,(1 + o(1)) \tag{13}$$

Therefore

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^n} + o\left(\frac{1}{\alpha^n}\right) \tag{14}$$

Then

$$\frac{I_n + I}{2(I_n - I)} = \frac{I_n}{I_n - I} - \frac{1}{2} = \frac{I_n}{C\alpha^n} + o\left(\frac{1}{\alpha^n}\right) \tag{15}$$

Similarly, from equation (11), we deduce

$$\frac{I_n + I_{n+1}}{2(I_n - I_{n+1})} = \frac{I_n}{I_n - I_{n+1}} - \frac{1}{2} = \frac{I_n}{C\alpha^n}\,\frac{1}{1-\alpha} + o\left(\frac{1}{\alpha^n}\right) \tag{16}$$

From definition 1 and equation (15) we deduce

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^n}\left(1+o(1)\right)\right| \tag{17}$$

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^n}\right| + \log_{10}\left|1+o(1)\right| \tag{18}$$

Therefore

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^n}\right| + o(1) \tag{19}$$

Similarly, from definition 1 and equation (16) we deduce

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n}{C\alpha^n}\ \frac{1}{1-\alpha}\right| + o\left(1\right) \tag{20}$$

Finally

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1-\alpha}\right) + o\left(1\right)} \tag{21}$$

From the number of significant digits in common between I_n and I_{n+1}, we can deduce the number of significant digits in common between I_n and the limit I. If the convergence zone is reached, $o(1) \ll 1$: the last term in equation (21) becomes negligible.
$\forall\, 0 < \alpha < 1$, $\exists k\ 0 < \alpha \leq 1 - \frac{1}{10^k}$ 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.
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 the 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. Further iterations are useless: I_{n+1} is the optimal iterate. In this case, the exact significant digits of I_{n+1} are in common with I_n and they are also in common with I, up to k digits. More concisely, if the sequence (I_n) is computed until the difference between two successive iterates is not significant, then the exact significant digits of the last iterate are those of I, up to k digits.

Remark: if $0 < \alpha \leq \frac{1}{2}$, $0 < \log_2\left(\frac{1}{1-\alpha}\right) \leq 1$, then the significant bits in common between I_n and I_{n+1} are also in common with I, up to one.

3.2 Dynamical control of the trapezoidal or Simpson's method

This strategy can be used for the computation of integrals with the trapezoidal or Simpson's method. 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 $k \geq 2$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using the trapezoidal method 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, 7, 8]:

$$I_n - I = \frac{h^2}{12}\left[f'(b) - f'(a)\right] + \mathcal{O}(h^4) \tag{22}$$

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

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{23}$$

Let f be a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 4$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using Simpson's method 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, 7, 8]:

$$I_n - I = \frac{h^4}{180}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}(h^6). \tag{24}$$

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

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{16}{15}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{25}$$

For both methods, 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. If approximations I_n are computed until $I_n - I_{n+1} = @.0$, the exact significant bits of the last approximation are those of I up to one.

3.3 Sequences with an exponential convergence

The strategy presented for sequences with a linear convergence is also valid for sequences with an exponential convergence. It is based on the following theorem.

Theorem 2 *Let (I_n) be a sequence converging to I in an exponential way, i.e. which satisfies $I_n - I = C\,\alpha^{p^n} + o(\alpha^{p^n})$ where $C \in I\!R$, $0 < \alpha < 1$ 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:

$$I_n - I = C\,\alpha^{p^n} + o(\alpha^{p^n}) \tag{26}$$

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

$$I_n - I_{n+1} = C\,\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right) + o(\alpha^{p^n}) \tag{27}$$

From equation (26), we deduce

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^{p^n}\,(1+o(1))} \tag{28}$$

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^{p^n}}\,(1+o(1)) \tag{29}$$

Therefore

$$\frac{I_n}{I_n - I} = \frac{I_n}{C\alpha^{p^n}} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{30}$$

Then

$$\frac{I_n + I}{2(I_n - I)} = \frac{I_n}{I_n - I} - \frac{1}{2} = \frac{I_n}{C\alpha^{p^n}} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{31}$$

Similarly, from equation (27), we deduce

$$\frac{I_n}{I_n - I_{n+1}} = \frac{I_n}{C\,\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right)\,(1+o(1))} \tag{32}$$

Therefore

$$\frac{I_n}{I_n - I_{n+1}} = \frac{I_n}{C\,\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right)} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{33}$$

Then

$$\frac{I_n + I_{n+1}}{2(I_n - I_{n+1})} = \frac{I_n}{I_n - I_{n+1}} - \frac{1}{2} = \frac{I_n}{C\,\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right)} + o\left(\frac{1}{\alpha^{p^n}}\right) \tag{34}$$

From definition 1 and equation (31) we deduce

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^{p^n}}\,(1+o(1))\right| \tag{35}$$

Therefore

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^{p^n}}\right| + o(1) \tag{36}$$

Similarly, from definition 1 and equation (34) we deduce

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n}{C\left(\alpha^{p^n} - \alpha^{p^{n+1}}\right)}\left(1+o(1)\right)\right| \tag{37}$$

Therefore

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n}{C\ \alpha^{p^n}\ \left(1-\alpha^{p^n(p-1)}\right)}\right| + o(1) \tag{38}$$

Finally

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right) + o(1)} \tag{39}$$

If the convergence zone is reached, the decimal significant digits in common between I_n and I_{n+1} are also common to the limit I, up to $\log_{10}\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right)$. If $0 < \alpha \leq M_n$, with $M_n = (\frac{9}{10})^{\left(\frac{1}{p^n(p-1)}\right)}$, 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.

Let us consider a sequence (I_n) with an exponential convergence computed using DSA. Computations are performed until, in the convergence zone, $I_n - I_{n+1} =$ @.0. If $\alpha \leq M_n$, the number of exact significant digits of the last iterate are those of the limit I, up to one.

Remark: if the convergence zone is reached, the significant bits in common between I_n and I_{n+1} are also common to the limit I, up to $\log_2\left(\frac{1}{1-\alpha^{p^n(p-1)}}\right)$. If $0 < \alpha \leq 2^{\left(\frac{1}{p^n(1-p)}\right)}$, 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 $n = 5$, $2^{\left(\frac{1}{p^n(1-p)}\right)} > 0.97$.

4 Dynamical control of arithmetical operations on converging sequences

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. In this section, we compare the results obtained using such numerical methods in stochastic arithmetic with the exact values they approximate. The theoretical results presented here have been established in stochastic arithmetic and can be applied to computations performed using DSA.

Theorem 3 *Let X_1 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 those of x_1 up to p.*
Similarly let X_2 be an approximation obtained in stochastic arithmetic of an exact value x_2, such that its exact significant bits are those of x_2 up to q.
Let $\bigcirc$ be an 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 those of the exact value $x_1 \bigcirc x_2$, up to $max(p,q)$.

Proof:
From equation (9), the number of exact significant bits of $X_1 = \left(m_1, \sigma_1^2\right)$, *i.e.* not affected by round-off errors, can be estimated by $\log_2\left(\frac{|m_1|}{\lambda_\beta\sigma_1}\right)$. The number of bits of X_1 in common with the exact value x_1 is therefore $log_2\left(\frac{|m_1|}{\lambda_\beta\sigma_1}\right) - p = log_2\left(\frac{|m_1|}{2^p\lambda_\beta\sigma_1}\right)$. 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 bits of $X_2 = \left(m_2, \sigma_2^2\right)$ in common with the exact value x_2 is $log_2\left(\frac{|m_2|}{\lambda_\beta\sigma_2}\right) - q = log_2\left(\frac{|m_2|}{2^q\lambda_\beta\sigma_2}\right)$.

From equations (4) and (9), the number of exact significant bits of $X_1 s{+}\ X_2$ is $log_2\left(\frac{|m_1+m_2|}{\lambda_\beta\sqrt{\sigma_1^2+\sigma_2^2}}\right)$. 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 $(2^p\sigma_1)^2 + (2^q\sigma_2)^2$. The number of bits of $X_1 s{+}\ X_2$ in common with the exact value $x_1 + x_2$ is therefore $log_2\left(\frac{|m_1+m_2|}{\lambda_\beta\sqrt{(2^p\sigma_1)^2+(2^q\sigma_2)^2}}\right)$, which can be 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 those of $x_1 + x_2$, 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 (6) and (9), the number of exact significant bits of $X_1 s{\times}\ X_2$ is $log_2\left(\frac{|m_1 m_2|}{\lambda_\beta\sqrt{m_2\sigma_1^2+m_1\sigma_2^2}}\right)$. 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 $2^{2p}m_2\sigma_1^2 + 2^{2q}m_1\sigma_2^2$. The number of bits of $X_1 s{\times}\ X_2$ in common with the exact value $x_1 \times x_2$ is therefore $log_2\left(\frac{|m_1 m_2|}{\lambda_\beta\sqrt{2^{2p}m_2\sigma_1^2+2^{2q}m_1\sigma_2^2}}\right)$, which can be 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 those of $x_1 \times x_2$, up to $max(p,q)$.

From equations (7) and (9), the number of exact significant bits of $X_1 s/X_2$ is $log_2\left(\frac{|\frac{m_1}{m_2}|}{\lambda_\beta\sqrt{(\frac{\sigma_1}{m_2})^2+(\frac{m_1\sigma_2}{m_2^2})^2}}\right)$. To take into account both the truncation error and the round-off error on $X_1 s/X_2$, one has to consider not the variance $(\frac{\sigma_1}{m_2})^2 + (\frac{m_1\sigma_2}{m_2^2})^2$, but $(\frac{2^p\sigma_1}{m_2})^2 + (\frac{2^q m_1\sigma_2}{m_2^2})^2$. The number of bits of $X_1 s/X_2$ in common with the exact value x_1/x_2 is therefore $log_2\left(\frac{|\frac{m_1}{m_2}|}{\lambda_\beta\sqrt{(\frac{2^p\sigma_1}{m_2})^2+(\frac{2^q m_1\sigma_2}{m_2^2})^2}}\right)$, which can be 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 $X_1 s/X_2$ are those of x_1/x_2, up to $max(p,q)$.

From theorem 3, we deduce the following corollary.

Corollary 1 *Let (I_k) be a sequence converging at least linearly to I and let (J_k) be a sequence converging at least linearly to J.*
Let us consider the computation of these sequences in stochastic arithmetic.
Let I_n be an iterate such that the exact significant bits of I_n are in common with I, up to p.
Let J_m be an iterate such that the exact significant bits of J_m are in common with J, up to q.
Let $\bigcirc$ be an arithmetical operator and $s\bigcirc$ the corresponding stochastic operator. Then the exact significant bits of $I_n \; s\bigcirc \; J_m$ are those of the exact value $I \bigcirc J$, up to $max(p,q)$.

This corollary can be used if the sequences (I_k) and (J_k) are computed using DSA. From section 3, as the sequence (I_k) converges at least linearly to I, if it is computed until the difference between two successive iterates is not significant, *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. If an arithmetical operation is performed on I_n and J_m using DSA, the exact significant bits of the result obtained are those of the result of the same operation performed on I and J, up to $max(p,q)$.

Remark: according to section 3, 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 not significant, each limit is approximated by the optimal iterate. According to section 3, 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 approximations, 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 3 and 4 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 not significant.
As for all m, the sequence $(u_{m,n})$ converges at least linearly to u_m, according to section 3, 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 not significant. 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 not significant, the exact significant bits of U_{m+1} are common with u, up to $p + q$.

5.2 Dynamical control of integrals on an infinite domain

Let us consider the computation of an improper integral $g = \int_0^\infty \phi(x)dx$.

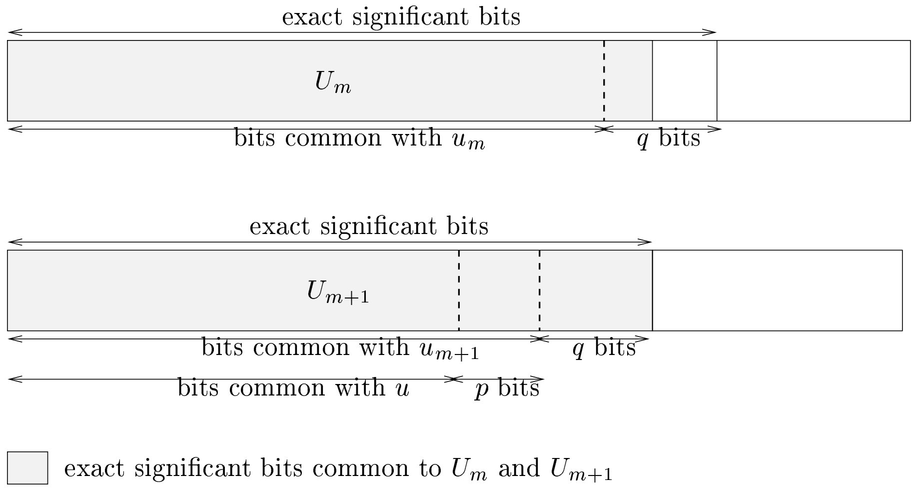

Figure 1: significant bits of U_m and U_{m+1}

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 method with step $\frac{L}{2^n}$.
For all j, the sequence $(F_{j,n})$ is computed until the difference between two successive iterates is not significant. 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 3, for all j, the significant bits of F_{j,n_j} are those of f_j, up to one. Let $G_m = \sum_{j=0}^{m} F_{j,n_j}$. According to corollary 1, the significant bits of G_m are those of 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 not significant. 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 3, if the convergence zone is reached, $C_{g_m,g_{m+1}} = C_{g_m,g} + \delta$ where δ represents p bits. Therefore the bits common to g_m and g_{m+1} are common with g, up to p.
Consequently if the difference $G_m - G_{m+1}$ is not significant, the exact significant bits of G_{m+1} are common with g, up to $p + 1$.

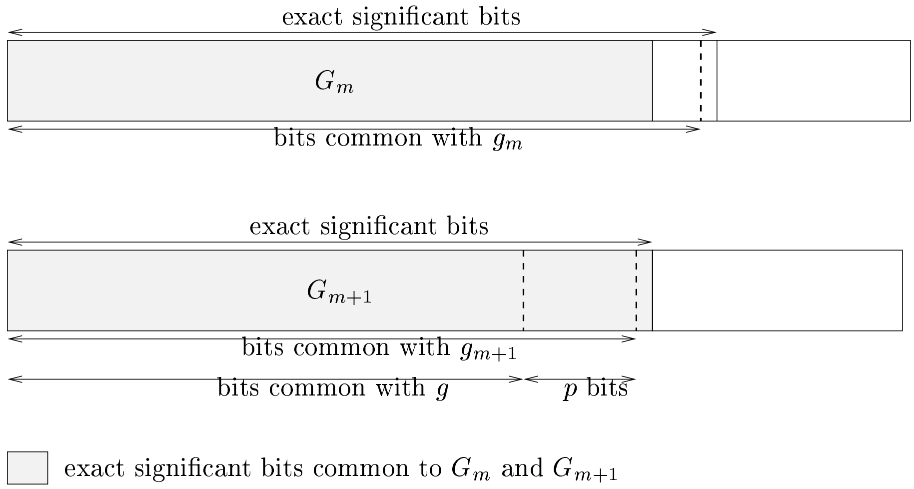

Figure 2: significant bits of G_m and G_{m+1}

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 methods using the strategy described in section 3. Approximations I_n have been computed with step $\frac{1}{2^n}$ until the difference $I_n - I_{n+1}$ is not significant. From section 3.2, 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 methods 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.

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

Table 1: Approximations of I

We can notice that the exact significant digits of each approximation obtained 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 method than with the trapezoidal method. This is due to the different convergence speeds of the sequences computed. Indeed the approximation of I is of order 2 with the trapezoidal method and of order 4 with Simpson's method. For each method, the error on the last iterate $|I_N - I|$ is not significant. Because of round-off error propagation, the computer can not distinguish I_N from I.

6.2 Computation of an improper integral

The strategy described in section 5.2 is used to compute the improper integral $g = \int_0^\infty e^{-ax}\, dx = \frac{1}{a}$, where $a > 0$.
Using the same notations as in section 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 method using DSA. For every j, a sequence is computed until the difference between two successive iterates is not significant.
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 1 can apply.
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 not significant. 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 C, where $C = log_{10}(\frac{2}{1-\alpha})$.
Table 2 (respectively 3) presents for $a = 1$ (respectively $a = 10^{-5}$) and different values of L the exact significant decimal digits of the approximation G_M.
The number of exact significant digits of G_M not in common with g is approximated by C.
As L increases, the number M of integrals f_j to be approximated decreases.
We notice that the exact significant digits of G_M are those of g up to $\lceil C \rceil$.

L	$C \approx$	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 2: Results obtained with Simpson's method for $a = 1$

L	$C \approx$	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

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

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, J. D. Faires, Numerical analysis, PWS (1993).

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

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

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

[5] J.-M. Chesneaux, F. Jézéquel, *Dynamical control of computations using the Trapezoidal and Simpson's rules.* Journal of Universal Computer Science, Vol. 4 (1), (1998) 2-10.

[6] J.-M. Chesneaux, J. Vignes, *Les fondements de l'arithmétique stochastique.* C.R.A.S., Paris, t. 315, série 1, 1992, pp. 1435-1440.

[7] M. K. Jain, S. R. K. Iyengar, R. K. Jain, Numerical methods for scientific and engineering computation, Wiley Eastern (1985).

[8] J. H. Mathews, Numerical methods for computer science, engineering and mathematics. Prentice-Hall (1987).

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

[10] J. Vignes, Zéro mathématique et zéro informatique, La vie des Sciences, C.R. Acad. Sci., Paris, 1, 1-13, 1987.

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

[12] J. Vignes, A stochastic arithmetic for reliable scientific computation, Math. Comp. Simulation, 35, 233-261, 1993.