Dynamical strategies using Discrete Stochastic Arithmetic for approximation methods

Fabienne Jezequel

Dynamical strategies using Discrete Stochastic Arithmetic for approximation methods

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

Abstract

Let us consider the converging sequence generated by successively dividing by two the step size used in an approximation method. With an appropriate stopping criterion, we show that in the last approximation obtained, the significant bits which are not affected by round-off errors are in common with the exact result, up to one. This strategy has been successfully applied to several composite quadrature methods. Other strategies, which are not based on "step halving", are also proposed. For approximation methods of a relatively high order, these alternative strategies may sometimes be less costly.

Key words: approximation methods, numerical validation, quadrature methods, trapezoidal rule, Simpson's rule, Gauss-Legendre method, CESTAC method, Discrete Stochastic Arithmetic.

1 Introduction

An approximation method, based on a discretization step, provides a numerical result affected by a global error, which consists of both a truncation error and a round-off error. If the discretization step decreases, the truncation error also decreases, but the round-off error usually increases. The optimal step size, for which the global error is minimal, can be computed dynamically [18]. In this paper, we show how to determine in the corresponding result which digits are affected neither by the truncation error, nor by the round-off error. In section 2, we present theoretical results which enable one to determine, from two approximations computed with step values h and $h/2$, the first digits of the exact result. In section 3, we briefly recall the principles of Discrete Stochastic Arithmetic (DSA) which enables one to estimate round-off error propagation and we present a strategy based on "step halving" to compute the optimal approximation. In section 4, we compare this strategy with other ones, where step reduction is not necessarily regular. In section 5, we present numerical experiments carried out using DSA.

2 Theoretical results on approximation methods

2.1 Preliminary definition

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 1 *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.

2.2 On approximation methods of order p

A numerical method which uses a discretization step h enables one to approximate an exact value L by a value $L(h)$ such that $\lim_{h\to 0} L(h) = L$. The technique of "step halving" consists in computing a sequence of approximations based on several successive divisions of the step by 2. Theorem 1 enables one to determine the number of significant digits in common between two successive approximations and the exact result L.

Theorem 1 *Let us consider a numerical method which provides an approximation $L(h)$ of order p to an exact value L, i.e. $L(h) - L = Kh^p + \mathcal{O}(h^q)$ with $1 \leq p < q,\ K \in \mathbb{R}$. If L_n is the approximation computed with the step $\frac{h_0}{2^n}$, then*

$$C_{L_n,L_{n+1}} = C_{L_n,L} + log_{10}\left(\frac{2^p}{2^p-1}\right) + \mathcal{O}\left(2^{n(p-q)}\right).$$

Proof The truncation error on L_n is

$$L_n - L = K\left(\frac{h_0}{2^n}\right)^p + \mathcal{O}\left(\frac{1}{2^{qn}}\right). \tag{1}$$

Using the same formula for L_{n+1}, one obtains

$$L_n - L_{n+1} = K\ \left(\frac{2^p-1}{2^p}\right)\ \left(\frac{h_0}{2^n}\right)^p + \mathcal{O}\left(\frac{1}{2^{qn}}\right). \tag{2}$$

From equation (1), we deduce

$$\frac{L_n}{L_n - L} = \frac{L_n}{K(\frac{h_0}{2^n})^p\left(1 + \mathcal{O}(2^{n(p-q)})\right)}. \tag{3}$$

$$\frac{L_n}{L_n - L} = \frac{L_n}{K(\frac{h_0}{2^n})^p}\left(1 + \mathcal{O}(2^{n(p-q)})\right). \tag{4}$$

Therefore

$$\frac{L_n}{L_n - L} = \frac{L_n}{K(\frac{h_0}{2^n})^p} + \mathcal{O}(2^{n(2p-q)}). \tag{5}$$

Then

$$\frac{L_n + L}{2(L_n - L)} = \frac{L_n}{L_n - L} - \frac{1}{2} = \frac{L_n}{K(\frac{h_0}{2^n})^p} + \mathcal{O}(2^{n(2p-q)}). \tag{6}$$

Similarly, from equation (2), we deduce

$$\frac{L_n + L_{n+1}}{2(L_n - L_{n+1})} = \frac{L_n}{L_n - L_{n+1}} - \frac{1}{2} = \left(\frac{L_n}{K(\frac{h_0}{2^n})^p}\right)\left(\frac{2^p}{2^p - 1}\right) + \mathcal{O}(2^{n(2p-q)}). \tag{7}$$

From definition 1 and equation (6), we deduce

$$C_{L_n,L} = \log_{10}\left|\frac{L_n}{K(\frac{h_0}{2^n})^p}\left(1 + \mathcal{O}(2^{n(p-q)})\right)\right|. \tag{8}$$

$$C_{L_n,L} = \log_{10}\left|\frac{L_n}{K(\frac{h_0}{2^n})^p}\right| + \log_{10}\left|1 + \mathcal{O}(2^{n(p-q)})\right|. \tag{9}$$

Therefore

$$C_{L_n,L} = \log_{10}\left|\frac{L_n}{K(\frac{h_0}{2^n})^p}\right| + \mathcal{O}\left(2^{n(p-q)}\right). \tag{10}$$

Similarly, from definition 1 and equation (7), we deduce

$$C_{L_n,L_{n+1}} = \log_{10}\left|\left(\frac{L_n}{K(\frac{h_0}{2^n})^p}\right)\left(\frac{2^p}{2^p - 1}\right)\right| + \mathcal{O}\left(2^{n(p-q)}\right). \tag{11}$$

Finally

$$C_{L_n,L_{n+1}} = C_{L_n,L} + \log_{10}\left(\frac{2^p}{2^p - 1}\right) + \mathcal{O}\left(2^{n(p-q)}\right). \tag{12}$$

□

If the convergence zone is reached, *i.e.* if the term $\mathcal{O}\left(2^{n(p-q)}\right)$ becomes negligible, the significant digits common to two successive approximations L_n and L_{n+1} are also in common with the exact result L, up to one bit. Indeed the term $\log_{10}\left(\frac{2^p}{2^p-1}\right)$ decreases as p increases and it corresponds to one bit for methods of order 1.

Remark 1 *This assertion can be related to previous works carried out on converging sequences [11, 12]. The sequence (L_n) generated by the technique of "step having" converges linearly to the exact result L. Indeed it satisfies $L_n - L = K\alpha^n + o(\alpha^n)$ with $K \in \mathbb{R}$ and $0 < |\alpha| < 1$. In [11, 12], it has been pointed out that if $0 < \alpha \leq \frac{1}{2}$ (which is the case here), then in the convergence zone, the significant bits common to two successive iterates are also in common with L, up to one.*

2.3 On Newton-Cotes methods

Theorem 1 can apply to Newton-Cotes quadrature rules.
Let $I(h)$ be the approximation to $I = \int_a^b f(x)dx$ by the trapezoidal rule with step h. If $f \in C^4[a,b]$, the truncation error expansion on $I(h)$ up to order 4 is [14]:

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

Let $I(h)$ be the approximation to $I = \int_a^b f(x)dx$ by Simpson's rule with step h. If $f \in C^6[a,b]$, the truncation error expansion on $I(h)$ up to order 6 is [14]:

$$I(h) - I = \frac{h^4}{180}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}(h^6). \tag{14}$$

Equations (13) and (14) are similar to equation (1) which characterizes approximation methods, with $p = 2$ and $q = 4$ for the trapezoidal rule; $p = 4$ and $q = 6$ for Simpson's rule. Therefore the following theoretical results, which had been given in [7] with specific proofs, could have been established from theorem 1.

Corollary 1 *Let I_n be the approximation to $I = \int_a^b f(x)dx$ by the trapezoidal rule with step $h = \frac{b-a}{2^n}$. If $f \in C^4[a,b]$ and $f'(b) \neq f'(a)$, then*

$$C_{I_n,I_{n+1}} = C_{I_n,I} + log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{15}$$

Corollary 2 *Let I_n be the approximation to $I = \int_a^b f(x)dx$ by Simpson's rule with step $h = \frac{b-a}{2^n}$. If $f \in C^6[a,b]$ and $f^{(3)}(b) \neq f^{(3)}(a)$, then*

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

The following error expansion for closed Newton-Cotes quadrature rules is given in [1]. Let $I(h)$ be the approximation to $I = \int_a^b f(x)dx$ by the composite closed Newton-Cotes quadrature rule with ν points and step h. Let $p = \nu + 1$ if ν is odd and $p = \nu$ if ν is even. If $f \in \mathcal{C}^{p+2}[a,b]$, then

$$I(h) - I = K_\nu h^p[f^{(p-1)}(b) - f^{(p-1)}(a)] + \mathcal{O}\left(h^{p+2}\right), \tag{17}$$

where K_ν is a constant which depends on ν.
Corollary 3 can be established from theorem 1 and equation (17).

Corollary 3 *Let I_N be the approximation to $I = \int_a^b f(x)dx$ by a composite closed Newton-Cotes quadrature rule of order p with step $h = \frac{b-a}{N}$. If $f \in \mathcal{C}^{p+2}[a,b]$ and $f^{(p-1)}(b) \neq f^{(p-1)}(a)$, then*

$$C_{I_N,I_{2N}} = C_{I_N,I} + log_{10}\left(\frac{2^p}{2^p - 1}\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{18}$$

Assuming $N = 2^n$, corollary 3 is in perfect agreement with corollary 1 (specific to the trapezoidal rule) and with corollary 2 (specific to Simpson's rule).
The theoretical result, similar to corollary 3, which has been established in [1] is not correct. Indeed the following equation given in its proof

$$\frac{I_N + I}{2(I_N - I)} = \frac{I_N N^{m+1}}{K_\nu (b-a)^{m+1}} + \mathcal{O}(1), \tag{19}$$

where $m+1$ represents the order p of the method, should be replaced by

$$\frac{I_N + I}{2(I_N - I)} = \frac{I_N N^{m+1}}{K_\nu (b-a)^{m+1}} + \mathcal{O}\left(N^{m-1}\right). \tag{20}$$

This change generates modifications in all relations of the proof deduced from equation (19) in [1].

2.4 On the Gauss-Legendre method

Theorem 1 can also apply to the Gauss-Legendre method. First let us briefly recall the principles of this quadrature method. The approximation to $\int_{-1}^{1} f(x)dx$ by the Gauss-Legendre method with ν points [9, 10] is $\sum_{i=1}^{\nu} C_i f(x_i)$, where for $i = 1, ..., \nu$, $\{x_i\}$ are the roots of the ν-degree Legendre polynomial P_ν and

$$C_i = \frac{2}{(1 - x_i^2)(P'_\nu(x_i))^2}. \tag{21}$$

For the computation of an integral on another interval such as $I = \int_a^b g(t)dt$, the following change of variable is required.

$$I = \int_a^b g(t)dt = \frac{b-a}{2} \int_{-1}^{1} f(x)dx, \tag{22}$$

with

$$\forall x \in [-1, 1],\ f(x) = g\left(\frac{(b-a)x + b + a}{2}\right). \tag{23}$$

The Gauss-Legendre method with ν points is of order 2ν: it is exact if the integrand is a polynomial of degree r with $r \leq 2\nu - 1$.
Let us assume that the integration domain is partitioned into 2^n subintervals and that the integral on each subinterval is evaluated using the Gauss-Legendre method with ν points. Theorem 2 presents the truncation error on I_n, the sum of the 2^n approximations obtained.

Theorem 2 *Let $I = \int_a^b g(t)dt$ and for $i = 1, ..., \nu$, let $\{x_i\}$ be the roots of the ν-degree Legendre polynomial and $\{C_i\}$ the corresponding weights. Let us assume that the integral on each subinterval $[\alpha_{k-1}, \alpha_k]$ with $\alpha_k = a + k\frac{b-a}{2^n}$, for $k = 1, ..., 2^n$, is evaluated using the Gauss-Legendre method with ν points. Let I_n be the sum of the 2^n approximations obtained. If $g \in C^{2\nu+1}[a, b]$, then*

$$I_n - I = \frac{K_\nu}{4^{n\nu}} + \mathcal{O}\left(\frac{1}{2^{n(2\nu+1)}}\right)$$

with $K_\nu = \frac{(b-a)^{2\nu}}{2^{2\nu+1}(2\nu)!}\left(\sum_{i=1}^{\nu} C_i x_i^{2\nu} - \frac{2}{2\nu+1}\right)\left[g^{(2\nu-1)}(b) - g^{(2\nu-1)}(a)\right]$.

In [13], a more general form of theorem 2, where the integration interval is partitioned into q subintervals, is given with its proof.
Corollary 4 can be established from theorems 1 and 2. The same notations and assumptions as in theorem 2 are used.

Corollary 4

$$C_{I_n,I_{n+1}} = C_{I_n,I} + log_{10}\left(\frac{4^\nu}{4^\nu - 1}\right) + \mathcal{O}\left(\frac{1}{2^n}\right).$$

Therefore if the convergence zone is reached, *i.e.* if the term $\mathcal{O}\left(\frac{1}{2^n}\right)$ becomes negligible, the significant digits common to two successive approximations are also in common with the exact value of the integral, up to one bit.

3 A 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 [16, 17, 20], is based on a probabilistic approach of round-off errors and enables one to estimate the number of exact significant digits of any computed result.
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. The computed result is chosen as being the mean value $\overline{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$.
The number $C_{\overline{R}}$ of exact significant digits of $\overline{R}$, *i.e.* its number of significant digits not affected by round-off errors, can be estimated by applying Student's test to the samples of R consisting of the different R_i $(i = 1, ..., N)$:

$$C_{\overline{R}} = \log_{10}\left(\frac{\sqrt{N}\,|\overline{R}|}{\sigma\tau_\beta}\right), \tag{24}$$

with

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2. \tag{25}$$

τ_β is the value of Student's distribution for $N-1$ degrees of freedom at a probability level $1-\beta$. 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 $\tau_\beta = 4.4303$.
The reliability of the CESTAC method has been proved under some hypotheses [4, 6]. Its validation 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 [15].

Definition 2 *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 $\forall i, R_i = 0$ or $C_{\overline{R}} \leq 0$.*

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.

3.2 Stochastic arithmetic

From the synchronous implementation of the CESTAC method and the concept of computational zero, stochastic arithmetic [6, 8, 17] has been defined. Two types of stochastic arithmetic actually exist: it can be either continuous or discrete.
By using the implementation of the CESTAC method, 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 centred 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. Continuous stochastic arithmetic is a modelling of the synchronous implementation of the CESTAC method. Properties established in its theoretical framework can be applied on a computer via the practical use of Discrete Stochastic Arithmetic (DSA) [19].
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 (24). From the concept of computational zero previously introduced, an equality concept and order relations have been defined for DSA.

Definition 3 *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* $X ds= Y$.

Many problems due to branching statements (for example, unsatisfied stopping criteria or infinite loops in algorithmic geometry) are partially solved in DSA, where the numerical quality of the operands in order relations is taken into account [5].
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].

3.3 A strategy for a dynamical control of approximation methods

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. Adopting the same notations as in 2.2, let (L_n) be a sequence computed in DSA with an approximation method using the step value $\frac{h_0}{2^n}$. and let us assume that the convergence zone is reached. If discrete stochastic equality is achieved for two successive iterates, *i.e.* $L_n - L_{n+1} = @.0$, the difference between L_n and L_{n+1} is only due to round-off errors and further iterations are useless. The optimal iterate L_{n+1} can therefore be dynamically determined at run time. Furthermore, from theorem 1, the exact significant bits of L_{n+1} are in common with the exact result L, up to one.
Therefore one can dynamically determine the optimal approximation by performing computations until the difference $L_n - L_{n+1}$ has no exact significant digit. If the convergence zone has been reached, then the exact significant bits of the last approximation are in common with L, up to one.

4 Alternative strategies

What results are obtained if one does not strictly apply the strategies described in section 2 which are based on step halving ? In this section, other step reductions, which are not necessarily regular, are proposed.
Theorem 3 applies when an approximation method of order p is used with a step $\frac{h_0}{n}$. From two approximations, which are not necessarily consecutive, it enables one to determine the first digits of the exact result.

Theorem 3 *Let us consider a numerical method which provides an approximation $L(h)$ of order p to an exact value L, i.e. $L(h) - L = Kh^p + \mathcal{O}(h^q)$ with $1 \leq p < q,\ K \in \mathbb{R}$.*
If L_n is the approximation computed with the step $\frac{h_0}{n}$ and $r \in \mathbb{N}^*$, then

$$C_{L_n,L_{n+r}} = C_{L_n,L} + \log_{10}\left(\frac{1}{1-(\frac{n}{n+r})^p}\right) + \mathcal{O}\left(n^{p-q}\right).$$

Proof The truncation error on L_n is

$$L_n - L = K\left(\frac{h_0}{n}\right)^p + \mathcal{O}\left(\frac{1}{n^q}\right). \tag{26}$$

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

Using the same formula for L_{n+r}, one obtains

$$L_n - L_{n+r} = K \left(1 - \left(\frac{n}{n+r}\right)^p\right) \left(\frac{h_0}{n}\right)^p + \mathcal{O}\left(\frac{1}{n^q}\right). \tag{27}$$

From equation (26), we deduce

$$\frac{L_n}{L_n - L} = \frac{L_n}{K(\frac{h_0}{n})^p} + \mathcal{O}(n^{2p-q}). \tag{28}$$

Then

$$\frac{L_n + L}{2(L_n - L)} = \frac{L_n}{L_n - L} - \frac{1}{2} = \frac{L_n}{K(\frac{h_0}{n})^p} + \mathcal{O}(n^{2p-q}). \tag{29}$$

Similarly, from equation (27), we deduce

$$\frac{L_n + L_{n+r}}{2(L_n - L_{n+r})} = \frac{L_n}{L_n - L_{n+r}} - \frac{1}{2} = \left(\frac{L_n}{K(\frac{h_0}{n})^p}\right) \left(\frac{1}{1 - (\frac{n}{n+r})^p}\right) + \mathcal{O}(n^{2p-q}). \tag{30}$$

From definition 1 and equation (29), we deduce

$$C_{L_n,L} = \log_{10} \left| \frac{L_n}{K(\frac{h_0}{n})^p} \right| + \mathcal{O}\left(n^{p-q}\right). \tag{31}$$

Similarly, from definition 1 and equation (30), we deduce

$$C_{L_n,L_{n+r}} = \log_{10} \left| \left(\frac{L_n}{K(\frac{h_0}{n})^p}\right) \left(\frac{1}{1 - (\frac{n}{n+r})^p}\right) \right| + \mathcal{O}\left(n^{p-q}\right). \tag{32}$$

Finally

$$C_{L_n,L_{n+r}} = C_{L_n,L} + \log_{10} \left(\frac{1}{1 - (\frac{n}{n+r})^p}\right) + \mathcal{O}\left(n^{p-q}\right). \tag{33}$$

□

Theorem 3 is equivalent to theorem 1 if $n = r = 2^m$.
If the convergence zone is reached, *i.e.* if $\mathcal{O}\left(n^{p-q}\right) \ll 1$, the significant digits common to L_n and L_{n+r} are also common to the exact value L, up to $\log_{10}\left(\frac{1}{1-(\frac{n}{n+r})^p}\right)$.
$\frac{1}{1-(\frac{n}{n+r})^p} \leq 2$ if and only if $n \leq M_{p,r}$ with $M_{p,r} = \frac{r}{2^{\frac{1}{p}}-1}$. Then the term $\log_{10}\left(\frac{1}{1-(\frac{n}{n+r})^p}\right)$ represents at most one bit. More the order p of the method is high and more the increment r is high, more the value $M_{p,r}$ is high too. It is noticeable that if $r = n$, $\log_2\left(\frac{1}{1-(\frac{n}{n+r})^p}\right)$ represents at most one bit, whatever the order p of the method is.

Consequently, in case that the condition $n \leq M_{p,r}$ is satisfied, if computations are carried out until, in the convergence zone, the difference between two approximations L_n and L_{n+r} has no exact significant digit, then the exact significant bits of the last approximation are in common with the exact value L, up to one. The condition on the number of iterations performed is too restrictive for the composite trapezoidal rule, of order 2. Indeed, in this case, $M_{p,1} \approx 2.4$.
Since the composite Simpson's rule requires an even number of partitions of the integration interval $[a, b]$, a sequence of approximations (I_n) cannot be generated with a step of the form $\frac{b-a}{n}$, but $\frac{b-a}{2n}$. Then the term $\log_{10}(\frac{2^p}{2^p-1})$, with $p = 4$, appears in equation (33). This term was already in theorem 1 and represents at most one bit.
It is actually not advisable to apply to the trapezoidal rule or to Simpson's rule a strategy which is not based on step halving. Indeed the strategy described in section 2 is preferable since points previously computed are always reused.
Theorem 3 can be interesting for methods of a relatively high order, such as the Gauss-Legendre method. Let us consider the case when $r = 1$.

- Using the Gauss-Legendre method with 6 points, if less than 18 iterations are performed with the stopping criterion previously described, on can obtain an approximation in which the exact significant bits are in common with the exact value of the integral, up to one.
- The approximation obtained using the Gauss-Legendre method with 12 points may have the same property if less than 36 iterations are performed.

Let I_n be the approximation obtained with n partitions of the integration domain into subintervals on which the classical Gauss-Legendre method with ν points is applied. As the number n of partitions increases, it is usually not possible to reuse points previously computed. If a dynamical control of the computations is performed, some run time may be saved by increasing by a low value of r the number of partitions of the integration interval. In a numerical experiment described in section 5, using the Gauss-Legendre method with 12 points, in the case when $r = 1$, the result has been obtained with 3 partitions of the integration interval. Using the strategy based on step halving, 4 partitions had to be performed for the stopping criterion to be satisfied.

5 Numerical experiments

First let us consider the integral

$$I = \int_0^1 \frac{\arctan(\sqrt{2+t^2})}{(1+t^2)\sqrt{2+t^2}} dt.$$

The evaluation of this integral is a problem which has been posed in [2]. D.H. Bailey and X.S. Li have indicated in [3] its exact value: $I = \frac{5\pi^2}{96}$. Therefore its 16 first exact digits are: $I \approx 0.5140418958900708$.

This integral has been evaluated using the different strategies described in section 2. Let I_n be the approximation to I computed :

- using the composite trapezoidal rule or the composite Simpson's rule with the step $\frac{1}{2^n}$,
- by partitioning the interval $[0,1]$ into 2^n subintervals on which the Gauss-Legendre method with 12 points is applied.

Approximations I_n have been computed in DSA, using the CADNA library, until the difference $I_n - I_{n+1}$ has no exact significant digit. From theorem 1, the exact significant bits (*i.e.* not affected by round-off errors) of the last approximation I_N are in common with I, up to one.
Table 1 presents the approximations to I obtained in single and in double precision. In every sequence, only the exact significant digits of the last iterate, estimated using DSA, are reported.

method	in single precision	in double precision
trapezoidal	$I_8 = 0.51404E+00$	$I_{19} = 0.5140418958899E+000$
Simpson	$I_8 = 0.514041E+00$	$I_{10} = 0.51404189589007E+000$
Gauss-Legendre	$I_1 = 0.5140419E+00$	$I_1 = 0.514041895890070E+000$

Table 1: Approximations to $I \approx 0.5140418958900708$

It is noticeable that the exact significant digits of each approximation I_N obtained are in common with I, up to one. The error $I_N - I$ is always a computational zero. Because of round-off errors, the computer cannot distinguish the approximation obtained from the exact value of the integral.
The number of iterations required for the stopping criterion to be satisfied may depend on the precision chosen, but also on the quadrature method used. Indeed the convergence speed of the computed sequence and the numerical quality of the result obtained vary according to the quadrature method. Starting from I_0 (the approximation obtained with no partition of the integration interval), the sequence generated by the Gauss-Legendre method with 12 points converges particularly quickly: in two iterations, a result with an excellent numerical quality is obtained whatever the precision chosen is.

Let us now consider the integral

$$J = \int_0^{20} \sin(t)dt.$$

$J = 1 - \cos(20)$. Its 16 first exact digits are: $J \approx 0.5919179381866080$.
This integral has been evaluated using strategies described in sections 2 and 4. Let J_n be the approximation to J computed :

- using the composite trapezoidal rule or the composite Simpson's rule with the step $\frac{20}{2^n}$,

- by partitioning the integration domain into 2^n subintervals on which the Gauss-Legendre method with 12 points is applied,
- by partitioning the integration domain into n subintervals on which the Gauss-Legendre method with 12 points is applied.

Approximations J_n have been computed until the difference $J_n - J_{n+1}$ has no exact significant digit. From theorems 1 and 3, the exact significant bits of the last approximation J_N are in common with J, up to one.
Table 2 presents the approximations to J obtained in double precision. In every sequence, only the exact significant digits of the last iterate, estimated using DSA, are reported.

method	approximation in double precision
trapezoidal	$J_{23} = 0.591917938186E + 000$
Simpson	$J_{15} = 0.5919179381866E + 000$
Gauss-Legendre (2^n partitions)	$J_2 = 0.59191793818660E + 000$
Gauss-Legendre (n partitions)	$J_3 = 0.59191793818660E + 000$

Table 2: Approximations to $J \approx 0.5919179381866080$

Like in the first experiment, the exact significant digits of each approximation J_N obtained are in common with J, up to one and the error $J_N - J$ is always a computational zero.
The number of iterations performed for the stopping criterion to be satisfied and the numerical quality of the result obtained depend on the quadrature method used. Using the Gauss-Legendre method with 12 points and the strategy based on step halving described in section 2, the last approximation J_2 has been obtained with 4 partitions of the integration interval. Applying the same quadrature method to n partitions of the integration domain, the last approximation J_3 has been obtained with 3 partitions. The results obtained with the two strategies have the same numerical quality.
Therefore run time may be saved by applying the Gauss-Legendre method to n partitions rather than 2^n partitions, as long as n remains relatively low. Since the sequence of approximations generated with 2^n partitions converges faster than the one obtained with n partitions, when relatively high values of n are reached, a strategy based on step halving is less costly. It is actually advisable to use an hybrid strategy:

- the strategy based on n partitions of the integration domain for relatively low values of n,
- and then the strategy based on step halving.

Such a hybrid strategy has been used in [13] for the evaluation of three-dimensional integrals involved in the neutron star theory. The multiple integrals were decomposed into one-dimensional integrals which were computed using a hybrid

strategy based on the Gauss-Legendre method with 12 points. As more than 5000 three-dimensional integrals had to be computed, saving run time was an important issue.

6 Conclusion and perspectives

Using an approximation method, a converging sequence can be generated by halving the step size at each iteration. If computations are carried out until, in the convergence zone, the difference between two successive approximations has no exact significant digit, then the last approximation is the optimal iterate. Furthermore its significant bits which are not affected by round-off errors are in common with the exact result, up to one. Under some assumptions, one can obtain with other step reductions an approximation with the same property. These alternative strategies may sometimes be less costly than the strategy based on "step halving".
All the strategies proposed in this paper are based on the computation of a sequence of approximations, each iterate corresponding to a constant step size. In the case of the approximation to an integral, the integration domain is partitioned into subintervals of the same length. An adaptive strategy can sometimes save run time. With a quadrature method, it consists in increasing the step size where the integrand does no vary much. It would be interesting to propose an adaptive strategy in DSA and to guarantee the result obtained.
Similar theoretical results as corollary 4 for the Gauss-Legendre method could be established for other Gaussian quadrature methods. In this paper, the dynamical control of the Gauss-Legendre method is based on successive partitions of the integration domain on which the classical Gauss-Legendre method is applied. A drawback of this method lies in the fact that, from one iteration to another, the sets of abscissas to consider have no points in common. With the Gauss-Kronrod method [14], a sequence of approximations can be computed where previous function evaluations can be reused. The dynamical control of this method would be an interesting perspective.

References

[1] S. Abbasbandy and M.A. Fariborzi Araghi. A stochastic scheme for solving definite integrals. *Applied Numerical Mathematics*, 55(2):125–136, 2005.

[2] Z. Ahmed. Definitely an integral. *American Mathematical Monthly*, 109(7):670–671, 2002.

[3] D.H. Bailey and X.S. Li. A comparison of three high-precision quadrature schemes. In *Proc. 5th Real Numbers and Computers conference*, pages 81–95, Lyon, France, September 2003.

[4] J.-M. Chesneaux. Study of the computing accuracy by using probabilistic approach. In C. Ullrich, editor, *Contribution to Computer Arithmetic and*

Self-Validating Numerical Methods, pages 19–30, IMACS, New Brunswick, New Jersey, USA, 1990.

[5] J.-M. Chesneaux. The quality relations in scientific computing. *Num. Algo.*, 7:129–143, 1994.

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

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

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

[9] S.D. Conte and C. de Boor. *Elementary numerical analysis*. McGraw-Hill, International Student edition, 1980.

[10] H. Engels. *Numerical quadrature and cubature*. Academic Press, 1980.

[11] F. Jézéquel. Dynamical control of converging sequences computation. *Applied Numerical Mathematics*, 50(2):147–164, 2004.

[12] F. Jézéquel. *Contrôle dynamique de méthodes d'approximation*. Habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, February 2005.

[13] F. Jézéquel, F. Rico, J.-M. Chesneaux, and M. Charikhi. Reliable computation of a multiple integral involved in the neutron star theory. *Mathematics and Computers in Simulation*, to appear.

[14] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12 of *Texts in applied mathematics*. Springer, 3rd edition, 2002.

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

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

[17] J. Vignes. A stochastic arithmetic for reliable scientific computation. *Math. Comput. Simulation*, 35:233–261, 1993.

[18] 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*, pages 233–251, Marseille, France, 1996.

[19] J. Vignes. Discrete stochastic arithmetic for validating results of numerical software. *Num. Algo.*, 37(1–4):377–390, December 2004.

[20] J. Vignes and M. La Porte. Error analysis in computing. In *Information Processing 1974*, pages 610–614. North-Holland, 1974.