Journal of Universal Computer Science, vol. 4, no. 1 (1998), 2-10
submitted: 25/9/97, accepted: 1/11/97, appeared: 28/1/98 © Springer Pub. Co.

Dynamical Control of Computations Using the Trapezoidal And Simpson's Rules

J.M. Chesneaux, F. Jézéquel
LIP6 Laboratory, Pierre et Marie Curie University, France
Jean-Marie.Chesneaux@lip6.fr, Fabienne.Jezequel@lip6.fr

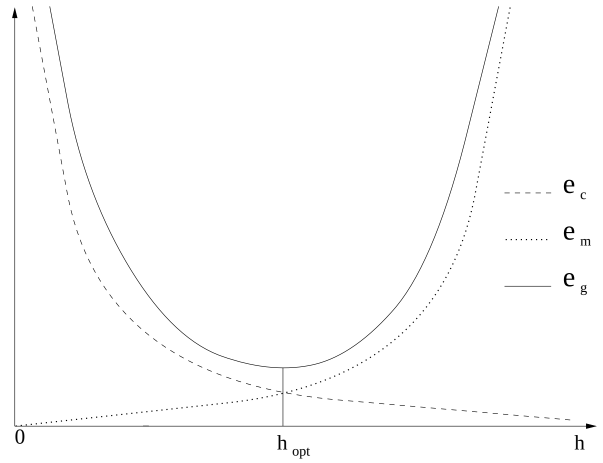
Abstract: If I_n is the approximation of a definite integral $I = \int_a^b f(x)dx$ with step $\frac{b-a}{2^n}$ using the trapezoidal rule (respectively Simpson's rule), if $C_{a,b}$ denotes the number of significant digits common to a and b, we show, in this paper, that $C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right)$ (respectively $C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{16}{15}\right) + \mathcal{O}\left(\frac{1}{16^n}\right)$). According to the previous theorems, using the CADNA library which allows on computers to estimate the round-off error effect on any computed result, we can compute dynamically the optimal value of n to approximate I and we are sure that the exact significant digits of I_n are in common with the significant digits of I.

Key Words: numerical validation, quadrature methods, trapezoidal rule, Simpson's rule

1 Introduction

As for all approximation methods which include a step parameter h, the numerical computation of integrals using the trapezoidal rule or Simpson's rule is affected by two types of errors : the mathematical error e_m(or truncation error) due to the theoretical approximation and the round-off error e_c due to the finite precision of the floating point arithmetic.

For a user, the practical error is a global error e_g which combines e_m and e_c. The well-known behavior of the global error can be described by the following figure :



When h is decreasing, e_m is decreasing but e_c is usually increasing. The classical problem is to find the optimal step h_{opt} which minimizes the global error.

We present in this paper a dynamical strategy which allows simultaneously :

1. to choose at the run time the optimal step from the computer point of view,
2. to guarantee the number of exact significant digits of the computed approximation, that means those in common with the mathematical value of the integral without other analyses.

The last part of the paper presents a numerical example.

2 Numerical accuracy of the trapezoidal rule and Simpson's rule

2.1 Notion of common significant digits

To correctly quantify the accuracy of a computed result, one must estimate the number of its exact significant digits, i.e. the number of significant digits that are common to the computed result and the exact result. 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 $[-\infty, +\infty]$ by*

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

Then, we have $|a-b| = \left| \frac{a+b}{2} \right| .10^{-C_{a,b}}$. 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 : If $|a-b| \ll |a+b|$, one can take $C_{a,b} \approx \log_{10} \left| \frac{a}{a-b} \right|$ which is sometimes more useful.

2.2 Numerical accuracy of the trapezoidal rule

We assume that f is a real function which is $\mathcal{C}^k$ over $[a, b]$ where $k \geq 2$ and that $f'(a) \neq f'(b)$.

Theorem 1 *Let I_n be the value of $\int_a^b f(x)dx$ computed using the trapezoidal rule with step $h = \frac{b-a}{2^n}$. Let I be the exact value of $\int_a^b f(x)dx$. Then*
$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right).$$

Proof : The development up to order 4 of the error due to the trapezoidal rule is [1, 4, 5] :

$$\begin{aligned} I_n - I &= \frac{h^2}{12}\,[f'(b) - f'(a)] + \mathcal{O}(h^4) \\ &= \frac{(b-a)^2}{12\;4^n}\,[f'(b) - f'(a)] + \mathcal{O}\left(\frac{1}{16^n}\right). \end{aligned} \tag{1}$$

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

$$I_n - I_{n+1} = \frac{3}{4}\;\frac{(b-a)^2}{12\;4^n}\,[f'(b) - f'(a)] + \mathcal{O}\left(\frac{1}{16^n}\right) \tag{2}$$

From the above relations, we deduce

$$I_n - I_{n+1} = \frac{3}{4}.(I_n - I) + \mathcal{O}\left(\frac{1}{16^n}\right) \tag{3}$$

Furthermore,

$$I_n + I = 2.I_n + \mathcal{O}\left(\frac{1}{4^n}\right) \quad \text{and} \quad I_n + I_{n+1} = 2.I_n + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{4}$$

Then

$$\frac{I_n + I}{2.(I_n - I)} = \frac{I_n}{I_n - I} - \frac{1}{2} = I_n.\frac{4^n}{K_1} + \mathcal{O}(1) \tag{5}$$

and

$$\frac{I_n + I_{n+1}}{2.(I_n - I_{n+1})} = \frac{I_n}{I_n - I_{n+1}} - \frac{1}{2} = I_n.\frac{4^n}{K_2} + \mathcal{O}(1) \tag{6}$$

where $K_1 = \dfrac{[f'(b) - f'(a)].(b-a)^2}{12}$ and $K_2 = \dfrac{3}{4}.K_1$.

According to the previous definition of $C_{a,b}$,

$$C_{I_n,I} = \log_{10}\left|\frac{I_n.4^n}{K_1}\right| + \mathcal{O}\left(\frac{1}{4^n}\right) \tag{7}$$

and

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n.4^n}{K_2}\right| + \mathcal{O}\left(\frac{1}{4^n}\right) \tag{8}$$

Finally, from $K_2 = \dfrac{3}{4}.K_1$,

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{4}{3}\right) + \mathcal{O}\left(\frac{1}{4^n}\right)} \tag{9}$$

One can remark that

1. The first term ($C_{I_n,I}$) of the right hand side is the most important term. Without the others, it means that, if we are able to compute the number of significant digits in common between I_n and I_{n+1}, we are sure that these digits are also in common with the mathematical value of the integral.
2. The second term ($\log_{10}\left(\frac{4}{3}\right)$) is a corrective one. But its value (0.1249..) must be pointed out. In practice, it means that, without the last term, the significant bits in common between I_n and I_{n+1} are also in common with I up to less than 1 bit.
3. The last term ($\mathcal{O}\left(\frac{1}{4^n}\right)$) is also very important. It points out that the two previous remarks are valid if the convergence zone is reached which corresponds, in the formula, to the fact that $\mathcal{O}\left(\frac{1}{4^n}\right) \ll 1$.

2.3 Numerical accuracy of Simpson's rule

We assume that f is a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 4$ and that $f^{(3)}(a) \neq f^{(3)}(b)$. We use the same notations to determine the numerical accuracy of Simpson's rule.

Theorem 2 *Let I_n be the value of $\int_a^b f(x)dx$ computed using the Simpson's rule with step $h = \frac{b-a}{2^n}$. Let I be the exact value of $\int_a^b f(x)dx$. Then*

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

Proof : The development up to order 8 of the error due to Simpson's rule is [1, 4, 5] :

$$\begin{aligned} I_n - I &= \frac{h^4}{180}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}(h^8) \\ &= \frac{(b-a)^4}{180\ 16^n}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}\left(\frac{1}{256^n}\right). \end{aligned} \tag{10}$$

From the corresponding formula for I_{n+1}, we deduce

$$I_n - I_{n+1} = \frac{15}{16}\,\frac{(b-a)^4}{180\ 16^n}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}\left(\frac{1}{256^n}\right) \tag{11}$$

and

$$I_n - I_{n+1} = \frac{15}{16}.(I_n - I) + \mathcal{O}\left(\frac{1}{256^n}\right) \tag{12}$$

Furthermore,

$$I_n + I = 2.I_n + \mathcal{O}\left(\frac{1}{16^n}\right) \quad \text{and} \quad I_n + I_{n+1} = 2.I_n + \mathcal{O}\left(\frac{1}{16^n}\right). \tag{13}$$

Then

$$\frac{I_n + I}{2.(I_n - I)} = I_n.\frac{16^n}{K_1} + \mathcal{O}(1) \quad \text{and} \quad \frac{I_n + I_{n+1}}{2.(I_n - I_{n+1})} = I_n.\frac{16^n}{K_2} + \mathcal{O}(1) \tag{14}$$

where $K_1 = \dfrac{[f^{(3)}(b) - f^{(3)}(a)].(b-a)^4}{180}$ and $K_2 = \dfrac{15}{16}.K_1$.

According to the previous definition of $C_{a,b}$,

$$C_{I_n,I} = \log_{10}\left|\frac{I_n.16^n}{K_1}\right| + \mathcal{O}\left(\frac{1}{16^n}\right) \tag{15}$$

and

$$C_{I_n,I_{n+1}} = \log_{10}\left|\frac{I_n.16^n}{K_2}\right| + \mathcal{O}\left(\frac{1}{16^n}\right) \tag{16}$$

Finally, from $K_2 = \dfrac{15}{16}.K_1$,

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{16}{15}\right) + \mathcal{O}\left(\frac{1}{16^n}\right)} \tag{17}$$

Of course, the conclusions of the previous section are valid for Simpson's rule, i.e. as with the trapezoidal rule, the significant digits common to the values I_n and I_{n+1} computed with Simpson's rule are also common to the exact value of the integral (up to $\log_{10}\left(\dfrac{16}{15}\right) + \mathcal{O}\left(\dfrac{1}{16^n}\right)$) .

3 The Discrete Stochastic Arithmetic

The synchronous implementation of the CESTAC method allows to estimate the round-off error propagation of the floating point arithmetic. With this method, one can know at any time during the execution of a scientific code the accuracy of any intermediate result. From this point of view the concept of computed zero introduced by J. Vignes [6, 7] is essential. A computed zero is a computed result which has no significant digit or which is the mathematical zero. In practice, this means a result that the computer can not distinguish from the null value because of the round-off error propagation. From this new concept, a new theoretical arithmetic, called stochastic arithmetic [3, 7], has been developed. New definitions for order relations and equality relation have been proposed. All these definitions, in opposition to the classical floating point arithmetic, take into account the accuracy of the operands.

For instance, two computed results are stochastically equal if their difference is a computed zero. In other words, one can say that the computer, because of the round-off errors, is not able to distinguish them. The Discrete Stochastic Arithmetic (DSA) is the use, on computer, of the synchronous CESTAC method associated with the new concepts of the theoretical stochastic arithmetic. The CADNA software [3, 2] is a library which implements automatically the DSA in any code written in Fortran. With CADNA, one can use new numerical types :

the stochastic types. The library contains the definition of all arithmetic operations and order relations for the stochastic types. The control of the accuracy is only performed on variables of stochastic type. When a stochastic variable is printed, only its significant digits appear. For a computed zero, the symbol @.0 is printed. The numerical experiment below has been performed using this software.

4 Numerical experiment

This section presents the computation of integrals with the trapezoidal rule and Simpson's rule using the Discrete Stochastic Arithmetic via the CADNA software [3, 2].

We still denote by I_n an approximation of $\int_a^b f(x)dx$ computed with step $\frac{b-a}{2^n}$ with the trapezoidal or Simpson's rules. The successive values I_n $(n \geq 1)$ are computed using stochastic variables in double precision. This means that, at each iteration, we are able to estimate the number of significant digits of I_n, I_{n+1} and $I_n - I_{n+1}$.

According to the previous results, we decide to stop the computation of the I_n's. Let N be the number of the last iteration. when $I_N - I_{N+1} = @.0$, one can say that

1. before this iteration, $I_n - I_{n+1}$ has exact significant digits. The computation of I_{n+1} gave significant information. But if $I_N - I_{N+1} = @.0$, the transformation of I_N into I_{N+1} is only due to the round-off errors. Further iterations are useless. We can say that, from the computer point of view, the number of iterations has been optimized.
2. Moreover, by definition of the stochastic equality reminded in the previous section, $I_N - I_{N+1} = @.0$ means also that the exact significant digits of I_{N+1} are in common with I_N. Therefore, if the convergence zone is reached, because of the theorems 1 and 2 *the significant digits of the last approximation are in common with the mathematical value of the integral I.*

For instance, the following classical code computes an approximation of

$$I = \int_{-1}^{1} 20.\cos(20x)\ (2.7x^2 - 3.3x + 1.2)\ dx$$

using the trapezoidal rule and the CADNA software [3, 2]. It is written in Fortran 90. The special statements due to CADNA are written in bold. One can remark the simplicity of the stopping criterion !

```
program integral
use cadna ! call to the CADNA library
implicit none
! declaration of the stochastic variables in double precision
type (double_st) :: f, x, sum = 0.d0, a = -1.d0, b = 1.d0
type (double_st) :: integ, h, err, integold, aux
logical :: l =.true.
```

```
integer :: n=1, i=0, j
call cadna_init(0) ! initialization of the CADNA library
aux = f(a) + f(b)
h = b-a
do while(l) ! stopping criterion
    i = i+1
    x = a + h/2.d0
    do j=1, n
       sum = sum + f(x)
       x = x + h
    enddo
    n = 2*n
    h = h/2.d0
    integ = h*(aux + 2.d0*sum)/2.d0
    l = (integold .ne. integ)
    integold = integ
    err = abs(integ - 7.316687747285081d0)
! using the str function, only the exact significant digits
! of the stochastic variables are printed
    write(*,100) i, str(integ), str(err)
enddo
100 format('n = ',i2,', I(f) = ',a25,', err = ',a25)
end

function f(x)
use cadna
implicit none
type (double_st) :: f, x
f = 20.d0*cos(20.d0*x)*((2.7d0*x - 3.3d0)*x + 1.2d0)
end
```

Computations have been performed on a SUN SPARC workstation using the two rules.

The tables 1 and 2 present for each iteration n and for the two methods, the computed approximation I_n and the error $|I_n - I|$, where I is the exact value of the integral, the 16 first digits of which being :

$$I = 0.7316687747285081E + 001.$$

n	I_n	$\|I_n - I\|$
1	0.532202672142963E+002	0.459035794670112E+002
2	-0.233434428466744E+002	0.306601305939594E+002
3	-0.235451792663099E+002	0.308618670135950E+002
4	0.106117380632568E+002	0.32950503159717E+001
5	0.742028156692706E+001	0.1035938196419E+000
6	0.732233719854277E+001	0.564945125769E-002
7	0.731702967403266E+001	0.3419267475E-003
8	0.73167089491443E+001	0.212018592E-004
9	0.73166890697896E+001	0.1322504E-005
10	0.73166878299008E+001	0.826158E-007
11	0.7316687752447E+001	0.5162E-008
12	0.7316687747607E+001	0.322E-009
13	0.7316687747305E+001	0.2E-010
14	0.731668774728E+001	@.0

Table 1: Using Simpson's rule.

n	I_n	$\|I_n - I\|$
1	0.558304008214445E+002	0.485137130741594E+002
2	-0.354998192964467E+001	0.108666696769297E+002
3	-0.185463799321436E+002	0.258630676794287E+002
4	0.332220856440671E+001	0.399447918287836E+001
5	0.639576331629697E+001	0.92092443098810E+000
6	0.709069372798132E+001	0.2259940193037E+000
7	0.72604456875198E+001	0.562420597652E-001
8	0.73026431337381E+001	0.14044613546E-001
9	0.73131775857768E+001	0.3510161508E-002
10	0.7315810268869E+001	0.877478415E-003
11	0.7316468381553E+001	0.219365731E-003
12	0.7316632906094E+001	0.54841190E-004
13	0.731667403700E+001	0.1371028E-004
14	0.731668431971E+001	0.342756E-005
15	0.731668689039E+001	0.85689E-006
16	0.731668753306E+001	0.2142E-006
17	0.73166876937E+001	0.535E-007
18	0.73166877338E+001	0.13E-007
19	0.73166877439E+001	0.3E-008
20	0.7316687747E+001	@.0

Table 2: Using the trapezoidal rule.

Because the convergence zone in this example has been reached, according to the theorems 1 and 2, one can notice that just when $I_n - I_{n+1} = @.0$, we also have $I - I_{n+1} = @.0$. In other words, when the difference between I_n and I_{n+1} becomes no significant, the exact significant digits of the last approximation are those of the mathematical value of the integral.

Remark : The number of iterations performed for the stopping criterion to be satisfied is different for the two methods. This is due to the faster convergence of Simpson's rule. Moreover the last computed value I_{14} has 12 significant digits for Simpson's rule and the last computed value I_{20} has 10 significant digits for the trapezoidal rule. Because the two methods lead to comparable computations, for the same value of n, one can remark that the corresponding I_n's have the same accuracy. In this example, we can verify that Simpson's rule is faster (of course) but also more accurate than the trapezoidal rule.

5 Conclusion

Nowadays, there exists some tools to estimate or to bound the round-off error effect of the floating point arithmetic on computers. In this paper, we have seen that this kind of information allows to find the best approximation from the computer point of view using the trapezoidal rule or Simpson's rule, and, to guarantee simultaneously that the exact significant digits of the approximation are in common with the significant digits of the exact value of the integral. This is very useful in practice and this kind of coding could be extended to Romberg or Gauss integration and more generally to any approximating method having a linear or a super-linear convergence.

References

1. R. L. Burden, J. D. Faires : Numerical analysis, PWS (1993).
2. URL address : http://www-anp.lip6.fr/cadna/
3. J.-M. Chesneaux : L'arithmétique stochastique et le logiciel CADNA, Habilitation à diriger des recherches, Laboratoire MASI-IBP, Université P. et M. Curie (1995).
4. M. K. Jain, S. R. K. Iyengar, R. K. Jain : Numerical methods for scientific and engineering computation, Wiley Eastern (1985).
5. J. H. Mathews : Numerical methods for computer science, engineering and mathematics. Prentice-Hall (1987).
6. J. Vignes, Zéro mathématique et zéro informatique, La vie des Sciences, C.R. Acad. Sci., Paris, n^{o} 1, janvier 1987, pp. 1-13.
7. J. Vignes, A stochastic arithmetic for reliable scientific computation, Math. Comp. Simul., 35, 1993, pp. 233-261.