МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2001
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2001
Proceedings of Thirtieth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 8–11, 2001

NUMERICAL ACCURACY OF CONVERGING SEQUENCES

F. Jezequel

If we compute a sequence having a linear convergence until the difference between two successive iterates is not significant, the result obtained has the best numerical accuracy for the computer used. Furthermore its exact significant digits are those of the mathematical value of the limit, up to one bit. This strategy can be used for the trapezoidal or Simpson's method, a sequence is then generated by halving the step value at each iteration. For Romberg's method, which consists in computing a sequence having a super-linear convergence, a similar strategy can also be used.

1. Introduction. Numerical algorithms are often based on the computation of converging sequences. For instance, the approximation of an integral with Romberg's method consists in computing iterates of a sequence. It is often difficult to determine the optimal iterate, i.e. the approximation for which the global error, consisting of the mathematical error and the round-off error, is minimal.

After briefly recalling the dynamical control of the trapezoidal and Simpson's methods, we show how to determine the optimal number of iterations when computing a sequence having a linear convergence. Then we present a similar strategy for Romberg's method, based on the computation of a sequence having a super-linear convergence. Finally a numerical experiment carried out using Discrete Stochastic Arithmetic is described.

2. Dynamical numerical validation of the trapezoidal and Simpson's methods. The computation of an integral with the trapezoidal or Simpson's method uses a step h. The approximation obtained is affected by both the mathematical error and the round-off error. If the step h decreases, the mathematical error also decreases, but the round-off error increases. The optimal step corresponds to a minimal global error. We present a strategy which enables one to determine this optimal step dynamically. It consists in computing a sequence until the difference between two successive iterates is not significant. This strategy is based on the following theorems, proved in [2]. $C_{R,r}$ denotes the number of decimal significant digits common to two real numbers R and r and is defined by $C_{R,r} = \log_{10} \left| \dfrac{R+r}{2.(R-r)} \right|$.

Theorem 1. *We assume that f is a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 2$ and that $f'(a) \neq f'(b)$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using the trapezoidal method with step $\dfrac{b-a}{2^n}$. Then*

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

Theorem 2. *We assume that f is a real function which is $\mathcal{C}^k$ over $[a,b]$ where $k \geq 4$ and that $f^{(3)}(a) \neq f^{(3)}(b)$. Let I_n be the approximation of $I = \int_a^b f(x)dx$ computed using Simpson's method with step $\frac{b-a}{2^n}$. Then*

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

These theorems show that, *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. The strategy described in [2] consists in computing the sequence (I_n) until the difference $I_n - I_{n+1}$ is not significant. Thus we obtain the result of best numerical quality for the computer and the quadrature method used. Furthermore its exact significant digits are those of the mathematical value of the integral.

These theorems can be generalized to the dynamical control of the computation of sequences having a linear convergence.

3. Dynamical numerical validation of sequences converging linearly. Let us consider a sequence (I_n) which converges linearly to I, i.e. which satisfies $I_n - I = C\alpha^n + o(\alpha^n)$ where $C \in \mathbb{R}$ and $0 < \alpha < 1$. The following theorem can apply:

Theorem 3. *Let (I_n) be a sequence converging linearly, then*

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

Proof. $I_n - I = C\alpha^n + o(\alpha^n)$. Using the same formula for I_{n+1}, we obtain $I_n - I_{n+1} = C\alpha^n(1-\alpha) + o(\alpha^n)$ We deduce $I_n - I_{n+1} = (I_n - I)(1-\alpha) + o(\alpha^n)$

Furthermore $I_n + I = 2.I_n + \mathcal{O}\left(\alpha^n\right)$ and $I_n + I_{n+1} = 2.I_n + \mathcal{O}\left(\alpha^n\right)$.

Then

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

and

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

$$C_{I_n,I} = \log_{10}\left|\frac{I_n}{C\alpha^n}\right| + \mathcal{O}\left(\alpha^n\right) \tag{3}$$

and

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

$$\boxed{C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1-\alpha}\right) + \mathcal{O}\left(\alpha^n\right)} \tag{5}$$

If $0 < \alpha < \frac{1}{2}$, $0 < \log_2\left(\frac{1}{1-\alpha}\right) < 1$. Therefore the bits common to I_n and I_{n+1} are those of I up to one.

This assertion is valid if $\mathcal{O}(\alpha^n)$ is negligible: this condition is satisfied when the convergence zone is reached.

Remark. Let I_n be the approximation 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:

$$I_n - I = \frac{h^2}{12}\left[f'(b) - f'(a)\right] + \mathcal{O}(h^4) \tag{6}$$

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 $\alpha = \frac{1}{4}$.

Let I_n be the approximation 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 8 is:

$$I_n - I = \frac{h^4}{180}\left[f^{(3)}(b) - f^{(3)}(a)\right] + \mathcal{O}(h^8) \tag{7}$$

$I_n - I = C\alpha^n + o(\alpha^n)$ with $C = \frac{(b-a)^4}{180}\left[f^{(3)}(b) - f^{(3)}(a)\right]$ and $\alpha = \frac{1}{16}$.

Consequently theorems 1 and 2 could be established from theorem 3.

4. Dynamical numerical validation of Romberg's method. The same type of strategy can be used for faster convergences.

Romberg's method is based on Richardson's extrapolation on results of the trapezoidal method.

Let f be a real function over $[a, b]$ and I be the exact value of $\int_a^b f(x)dx$. Let $T_1(h)$ be the approximation of I computed using the trapezoidal method with step $h = \frac{b-a}{M}$ ($M \geq 1$). Romberg's method consists in computing the following triangular table:

$$\begin{array}{llcccc}
T_1(h) & T_1\left(\frac{h}{2}\right) & \ldots\ldots & T_1\left(\frac{h}{2^{n-3}}\right) & T_1\left(\frac{h}{2^{n-2}}\right) & T_1\left(\frac{h}{2^{n-1}}\right) \\
T_2(h) & T_2\left(\frac{h}{2}\right) & \ldots\ldots & T_2\left(\frac{h}{2^{n-3}}\right) & T_2\left(\frac{h}{2^{n-2}}\right) & \\
T_3(h) & T_3\left(\frac{h}{2}\right) & \ldots\ldots & T_3\left(\frac{h}{2^{n-3}}\right) & & \\
\vdots & \vdots & & & & \\
T_{n-1}(h) & T_{n-1}\left(\frac{h}{2}\right) & & & & \\
T_n(h) & & & & &
\end{array}$$

The first row of the table represents approximations of I computed using the trapezoidal method with step $\frac{h}{2^j}$ $(j = 0, \ldots, n-1)$. Rows 2 to n are computed using the following formula:

$$\text{For } p = 2, \ldots, n \text{ and } j = 0, \ldots, n-p,\ T_p(\frac{h}{2^j}) = \frac{1}{4^{p-1}-1}\left(4^{p-1}T_{p-1}(\frac{h}{2^{j+1}}) - T_{p-1}(\frac{h}{2^j})\right).$$

The sequence of approximations with Romberg's method $T_1(h), \ldots, T_n(h)$ converges exponentially to I.

In [3], the following equation has been proved:

$$T_n(h) - T_{n+1}(h) = T_n(h) - I + \mathcal{O}\left(\frac{h^{2n+2}}{(2n+2)!\, 2^{n(n-1)}}\right) \tag{8}$$

The following theorem can been established from equation 8 by taking into account the truncation error on $T_n(h)$. See [3] for more details on its proof.

Theorem 4. *We assume that f is a real function which is $\mathcal{C}^k$ over $[a, b]$ where $k \geq 2n+2$ and that $f^{(2n-1)}(a) \neq f^{(2n-1)}(b)$. Let $T_n(h)$ be the approximation of $I = \int_a^b f(x)dx$ computed with n iterations of Romberg's method using the initial step $h = \frac{b-a}{M}$ $(M \geq 1)$. Then*

$$C_{T_n(h),T_{n+1}(h)} = C_{T_n(h),I} + \mathcal{O}\left(\frac{h^{2n}}{(2n)!\, 2^{n(n-1)}}\right).$$

When the convergence zone is reached, the significant digits common to $T_n(h)$ and $T_{n+1}(h)$ are also common to I. If approximations $T_n(h)$ are computed until the difference $T_n(h) - T_{n+1}(h)$ is not significant, the exact significant digits of the last iterate are those of the exact value of the integral.

5. Discrete stochastic arithmetic. The synchronous implementation of the CESTAC method [6] allows to estimate the number of exact significant digits of any computed result. From this, the concept of computed zero has been introduced [4], [5]. A computed zero is either the mathematical zero or a computed result which has no significant digit. In practice, it is a result that the computer can not distinguish from the null value because of round-off errors. Stochastic arithmetic [1], [5] has been developed from this concept. New order relations have been defined, taking into account the accuracy of the operands. Discrete Stochastic Arithmetic is the joint use on a computer of the synchronous CESTAC method and theoretical stochastic arithmetic.

The CADNA library [1], [7] automatically implements Discrete Stochastic Arithmetic in any scientific code written in Ada, C or Fortran. It allows the use of new numerical types: the stochastic types. At any time the numerical quality of any stochastic variable can be controlled. When a stochastic variable is printed only its exact significant digits appear. In case of a computed zero, the symbol @.0 is printed. The numerical experiment described in next section has been carried out using the CADNA library.

6. Numerical experiment. Let us consider the integral

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

The 16 first exact decimal digits of I are: $I = 0.7316687747285081E + 001$.

I has been estimated with the trapezoidal and Simpson's methods using the CADNA library in single and double precision. Approximations I_n have been computed with step $\frac{1}{2^n}$ until the difference $I_n - I_{n+1}$ is not significant. From theorems 1 and 2, we can guarantee that the exact significant digits of the last iterate I_N are in common with the exact value of I up to one bit.

I has also been estimated with Romberg's method. Using the initial step $h = 2$, approximations $T_n(h)$ have been computed until $T_n(h) - T_{n+1}(h)$ is not significant. From theorem 4, the exact significant digits of the last iterate $I_N = T_N(h)$ are those of the exact value of I.

The number of exact significant digits of the different approximations of I has been estimated by the CADNA library. For each sequence, the exact significant digits of the last iterate are reported below.

	in single precision	in double precision
trapezoidal method :	$I_{12} = 0.7317E + 01$	$I_{21} = 0.7316687747E + 001$
Simpson′s method :	$I_9 = 0.73167E + 01$	$I_{15} = 0.731668774729E + 001$
Romberg′s method :	$I_9 = 0.731669E + 01$	$I_{11} = 0.73166877472851E + 001$

For each method, all the exact significant digits of I_N are in common with I. The number of iterations performed and the number of exact significant digits of the last iterate depend on the method used. This is due to the different type of convergence of the sequences computed: linear for the trapezoidal and Simpson's methods, exponential for Romberg's method. Furthermore 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.

7. Conclusion. A strategy to dynamically control converging sequences has been established. Thanks to Discrete Stochastic Arithmetic, it is possible to determine the approximation of the limit which has the best numerical accuracy. Its exact significant digits are those of the mathematical value expected. This strategy can be used for the computation of an integral with the trapezoidal, Simpson's or Romberg's method.

The studies carried out for single definite integrals could be extended to multiple integrals, integrals over infinite intervals or singular integrals. The sequences examined in this note all converge to a scalar value. Another perspective to this work could be the numerical validation of sequences of vectors involved for example in iterative methods for linear systems.

REFERENCES

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

[2] J.-M. Chesneaux, F. Jezequel. Dynamical control of computations using the trapezoidal and Simpson's rules. *J. Univ. Comp. Sci.*, **4**, No 1 (1998), 2–10.

[3] F. Jézéquel. Dynamical control of computations using approximation methods. 16th IMACS world congress, Lausanne (Switzerland), august 2000.

[4] J. Vignes. Zéro mathématique et zéro informatique. La vie des Sciences, *C.R. Acad. Sci., Paris*, **1** (1987), 1–13.
[5] J. Vignes. A stochastic arithmetic for reliable scientific computation. *Math. Comp. Simul.*, **35** (1993), 233–261.
[6] J. Vignes, M. La Porte. Error analysis in computing. Information Processing, vol 74, North Holland, 1974.
[7] URL address : `http://www-anp.lip6.fr/cadna/`

Fabienne Jezequel
Laboratoire d'Informatique de Paris 6
CNRS UMR 7606
4, place Jussieu
75252 Paris Cedex 05
France
e-mail: `Fabienne.Jezequel@lip6.fr`

ИЗЧИСЛИТЕЛНА ТОЧНОСТ НА СХОДЯЩИ РЕДИЦИ

Ф. Жезекил

Ако пресмятаме редица, с линеен порядък на сходимост, докато разликата на два последователна члена стане по-малка от най-малкото представимо в компютъра число, полученият резултат е с най-голяма изчислителна точност за използвания компютър. Такъв подход може да се използва за апроксимация с метода на трапеците или на Симпсън, като стъпката се разполовява на всяка итерация. Подобен подход може да се използва и за пресмятане по метода на Ромберг, при който получената редица има свръх линеен ръст на сходимост.