A dynamical strategy for approximation methods

Fabienne Jézéquel

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

Abstract

The numerical result provided by an approximation method is affected by a global error, which consists of both a truncation error and a round-off error. Let us consider the converging sequence generated by successively dividing by two the step size used. If computations are performed until, in the convergence zone, the difference between two successive approximations is only due to round-off errors, then the global error on the result obtained is minimal. Furthermore its significant bits which are not affected by round-off errors are in common with the exact result, up to one. *To cite this article: F. Jézéquel, C. R. Acad. Sci. Paris, Ser. I 340 (2006).*

Résumé

Une stratégie dynamique pour les méthodes d'approximation. Le résultat numérique fourni par une méthode d'approximation est entaché d'une erreur globale qui comprend à la fois une erreur de troncature et une erreur d'arrondi. Considérons la suite convergente générée en divisant par deux successivement le pas utilisé. Si les calculs sont effectués jusqu'à ce que, dans la zone de convergence, la différence entre deux approximations successives soit uniquement due aux erreurs d'arrondi, alors l'erreur globale sur le résultat obtenu est minimale. De plus, ses bits significatifs non entachés d'erreur d'arrondi sont en commun avec le résultat exact, à un près. *Pour citer cet article : F. Jézéquel, C. R. Acad. Sci. Paris, Ser. I 340 (2006).*

Version française abrégée

Une méthode d'approximation fournit un résultat entaché d'une erreur de troncature inhérente à l'algorithme utilisé et d'une erreur d'arrondi due à la précision finie de l'arithmétique de l'ordinateur. Lorsque le pas de discrétisation

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

d'une telle méthode décroît, l'erreur de troncature diminue, mais l'erreur d'arrondi augmente. Il peut alors être difficile de contrôler ces deux erreurs à la fois. Le théorème 0.1 permet, à partir de deux approximations calculées avec les pas h et $\frac{h}{2}$, de déterminer les premiers chiffres du résultat exact. Il généralise les résultats théoriques qui avaient été établis pour différentes méthodes de quadrature: tout d'abord pour la méthode des trapèzes et celle de Simpson [4], puis pour les méthodes fermées de Newton-Cotes [1] et la méthode de Gauss-Legendre [6].

Théorème 0.1 *Si L_n est une approximation d'ordre p calculée avec le pas $\frac{h_0}{2^n}$ d'une valeur exacte L, dont le développement jusqu'à l'ordre q de l'erreur de troncature est $L_n - L = K\left(\frac{h_0}{2^n}\right)^p + \mathcal{O}\left(\frac{1}{2^{nq}}\right)$ avec $1 \leq p < q,\ p \in \mathbb{N},\ q \in \mathbb{N},\ K \in \mathbb{R}$, alors*

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

où $C_{a,b}$ représente le nombre de chiffres décimaux significatifs communs à deux nombres réels a et b.

Si la zone de convergence est atteinte, c'est-à-dire si le terme $\mathcal{O}\left(2^{n(p-q)}\right)$ devient négligeable, les chiffres significatifs communs à deux approximations successives L_n et L_{n+1} sont aussi en commun avec le résultat exact L, à un bit près. En effet, le terme $log_{10}\left(2^p/(2^p-1)\right)$ décroît lorsque p augmente et correspond à un bit pour les méthodes d'ordre 1. À partir de ce théorème et du développement de l'erreur de troncature due aux méthodes fermées de Newton-Cotes spécifié dans l'équation (7), on peut déduire le corollaire 0.2.

Corollaire 0.2 *Soit I_n l'approximation de $I = \int_a^b f(x)dx$ par une méthode composite fermée de Newton-Cotes d'ordre p avec le pas $\frac{b-a}{2^n}$. Si $f \in \mathcal{C}^{p+2}[a,b]$ et $f^{(p-1)}(b) \neq f^{(p-1)}(a)$, alors*

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

Le résultat théorique, similaire au corollaire 0.2, qui a été présenté dans [1] est erroné. En effet, le dernier terme de l'équation établie dans [1] devrait être corrigé. En appliquant le corollaire 0.2 à la méthode des trapèzes (d'ordre 2) et à la méthode de Simpson (d'ordre 4), on retrouve les résultats théoriques établis dans [4]. Le théorème 0.1 peut aussi s'appliquer lorsque le domaine d'intégration est découpé en sous-intervalles de longueur h sur lesquels on utilise la méthode classique de Gauss-Legendre à ν points. Le corollaire 0.3, qui se déduit du développement de l'erreur commise spécifié dans l'équation (9) et du théorème 0.1, est en accord avec le théorème 6 établi dans [6].

Corollaire 0.3 *Soit I_n l'approximation de $I = \int_a^b f(x)dx$ obtenue en évaluant*

chaque intégrale sur un sous-intervalle de longueur $\frac{b-a}{2^n}$ *par la méthode de Gauss-Legendre à* ν *points. Si* $f \in \mathcal{C}^{2\nu+1}[a,b]$ *et* $f^{(2\nu-1)}(b) \neq f^{(2\nu-1)}(a)$*, alors*

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

Ces résultats théoriques, établis en tenant compte de l'erreur de troncature, sont particulièrement intéressants si un contrôle des erreurs d'arrondi peut être effectué. Fondée sur une approche probabiliste des erreurs d'arrondi, la méthode CESTAC [8] permet d'estimer le nombre de chiffres significatifs exacts de tout résultat calculé sur ordinateur. Son utilisation dans un code nécessite d'exécuter celui-ci plusieurs fois avec un mode d'arrondi aléatoire. L'Arithmétique Stochastique Discrète (ASD) [9] a été définie à partir de l'implantation synchrone de la méthode CESTAC et du concept de zéro informatique [7]. L'ASD est implantée dans la bibliothèque CADNA, qui fournit le nombre de chiffres significatifs exacts de tout résultat d'un code scientifique, à un près.

En reprenant les notations utilisées dans le théorème 0.1, on considère une suite (L_n) d'approximations de L calculée en ASD avec le pas $\frac{h_0}{2^n}$. Supposons que la zone de convergence soit atteinte. Lorsque la différence $L_n - L_{n+1}$ n'est plus significative, il est inutile de poursuivre le calcul. L'itéré optimal L_{n+1} peut donc être déterminé dynamiquement. De plus, d'après le théorème 0.1, ses bits significatifs non entachés d'erreur d'arrondi sont en commun avec la valeur exacte L, à un près.

Cette stratégie a été appliquée au calcul de l'intégrale I spécifiée dans l'équation (11) en utilisant la méthode des trapèzes et celle de Simpson avec le pas $\frac{1}{2^n}$ et aussi en découpant l'intervalle $[0,1]$ en 2^n sous-intervalles sur lesquels on applique la méthode de Gauss-Legendre à 12 points. Les approximations I_n ont été calculées en ASD jusqu'à ce que la différence $I_n - I_{n+1}$ soit non significative. Le tableau 1 présente les approximations obtenues en simple et en double précision. Pour chaque suite, seuls sont indiqués les chiffres significatifs exacts du dernier itéré I_N. On remarque alors, conformément au théorème 0.1, que ceux-ci sont toujours en commun avec la valeur exacte de l'intégrale, à un près. Le nombre d'itérations nécessaires pour satisfaire le test d'arrêt peut dépendre de la précision choisie, mais aussi de la méthode de quadrature utilisée. En effet, la vitesse de convergence de la suite calculée et la qualité numérique du résultat obtenu varient selon la méthode de quadrature.

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 increases. Therefore it may be a problem to compute the optimal approximation, *i.e.* the result for which the global error is minimal.

In this note, we present a theorem which enables one to determine, from two approximations computed with step values h and $\frac{h}{2}$, the first digits of the exact result. We describe a strategy, based on step halving, to compute dynamically the optimal step size and we show how to determine in the corresponding result which digits are affected neither by the truncation error, nor by the round-off error.

Previous theoretical results had been established for the dynamical control of several quadrature methods: first the trapezoidal rule and Simpson's rule [4], then closed Newton-Cotes methods [1] and the Gauss-Legendre method [6]. The theorem presented here is a generalization of these results and enables one to perform a dynamical control of approximation methods.

2 Theoretical results on approximation methods

2.1 On approximation methods of order p

Let us consider the converging sequence generated by successively dividing by two the step size used in an approximation method of order p. Theorem 2.1 enables one to determine the number of significant digits in common between two successive approximations and the exact result L.

Theorem 2.1 *If L_n is an approximation of order p computed with the step $\frac{h_0}{2^n}$ to an exact value L, such that its truncation error expansion up to order q is $L_n - L = K\left(\frac{h_0}{2^n}\right)^p + \mathcal{O}\left(\frac{1}{2^{nq}}\right)$ with $1 \leq p < q$, $p \in \mathbb{N}$, $q \in \mathbb{N}$, $K \in \mathbb{R}$, then*

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

where $C_{a,b}$ denotes the number of decimal significant digits common to two real numbers a and b.

PROOF.

From the truncation error on L_n, we deduce

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

Then

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

Similarly, from the truncation error on L_n and L_{n+1}, we deduce

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

From equation (3), we deduce

$$C_{L_n,L} = \log_{10}\left|\frac{L_n + L}{2(L_n - L)}\right| = \log_{10}\left|\frac{L_n}{K(\frac{h_0}{2^n})^p}\right| + \mathcal{O}\left(2^{n(p-q)}\right). \tag{5}$$

Similarly, from equation (4), we deduce

$$C_{L_n,L_{n+1}} = \log_{10}\left|\frac{L_n + L_{n+1}}{2(L_n - L_{n+1})}\right| = \log_{10}\left|\left(\frac{L_n}{K(\frac{h_0}{2^n})^p}\right)\left(\frac{2^p}{2^p - 1}\right)\right| + \mathcal{O}\left(2^{n(p-q)}\right). \tag{6}$$

Finally, from equations (5) and (6), we deduce equation (1).

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 $\log_{10}\left(2^p/(2^p-1)\right)$ decreases as p increases and it corresponds to one bit for methods of order 1.

2.2 On Newton-Cotes methods

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 the 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

$$I(h) - I = K_\nu h^p \left[f^{(p-1)}(b) - f^{(p-1)}(a) \right] + \mathcal{O}\left(h^{p+2}\right), \tag{7}$$

where K_ν is a constant which depends on ν.

Corollary 2.2 can be established from theorem 2.1 and equation (7).

Corollary 2.2 *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 the step $\frac{b-a}{2^n}$. If $f \in \mathcal{C}^{p+2}[a, b]$ and $f^{(p-1)}(b) \neq f^{(p-1)}(a)$, then*

$$C_{I_n, I_{n+1}} = C_{I_n, I} + log_{10}\left(\frac{2^p}{2^p - 1}\right) + \mathcal{O}\left(\frac{1}{4^n}\right). \tag{8}$$

The theoretical result, similar to corollary 2.2, which has been presented in [1] is not correct. Indeed the last term $\mathcal{O}\left(\frac{1}{N^{m+1}}\right)$ of the equation established in [1], where $m + 1$ represents the order p of the method and N the number of partitions of the integration interval, should be replaced by $\mathcal{O}\left(\frac{1}{N^2}\right)$.

The application of corollary 2.2 to the trapezoidal rule (of order 2) and to Simpson's rule (of order 4) is consistent with the theoretical results established in [4].

2.3 On the Gauss-Legendre method

Let $I(h)$ be the approximation to $I = \int_a^b f(x)dx$ obtained by evaluating each integral on a sub-interval of length $h = \frac{b-a}{q}$ using the Gauss-Legendre method with ν points. The truncation error on $I(h)$ has been established in [6]. If $f \in \mathcal{C}^{2\nu+1}[a, b]$, then

$$I(h) - I = K_\nu h^{2\nu} \left[f^{(2\nu-1)}(b) - f^{(2\nu-1)}(a) \right] + \mathcal{O}\left(h^{2\nu+1}\right), \tag{9}$$

where the parameter K_ν does not depend on h.

Corollary 2.3, which can be deduced from theorem 2.1 and equation (9), is in agreement with a theoretical result established in [6].

Corollary 2.3 *Let I_n be the approximation to $I = \int_a^b f(x)dx$ obtained by evaluating each integral on a sub-interval of length $\frac{b-a}{2^n}$ using the Gauss-Legendre*

method with ν points. If $f \in \mathcal{C}^{2\nu+1}[a,b]$ and $f^{(2\nu-1)}(b) \neq f^{(2\nu-1)}(a)$, then

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

2.4 A strategy for a dynamical control of approximation methods

Based on a probabilistic approach, the CESTAC method [8] enables one to estimate round-off error propagation in a code by running it several times with a random rounding mode. Continuous stochastic arithmetic [5] is a modelling of its synchronous implementation. Discrete Stochastic Arithmetic (DSA) [9] has been defined from the synchronous implementation of the CESTAC method and the concept of computational zero [7]. DSA is implemented in the CADNA library [1], which provides the number of exact significant digits of any result of a scientific code, up to one.

Adopting the same notations as in 2.1, 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 the difference between L_n and L_{n+1} is only due to round-off errors, further iterations are useless. Therefore if computations are performed until the difference $L_n - L_{n+1}$ has no exact significant digit, the optimal iterate L_{n+1} can be dynamically determined at run time. Furthermore, from theorem 2.1, its significant bits which are not affected by round-off errors are in common with the exact result L, up to one.

3 Numerical experiment

The evaluation of the integral I defined in equation (11) is a problem which has been posed in [2].

$$I = \int_0^1 \frac{\arctan(\sqrt{2+t^2})}{(1+t^2)\sqrt{2+t^2}} dt \tag{11}$$

Its exact value has been indicated in [3]: $I = \frac{5\pi^2}{96}$. Let I_n be the approximation to I computed by using the composite trapezoidal rule or the composite Simpson's rule with the step $\frac{1}{2^n}$, or by partitioning the interval $[0,1]$ into 2^n sub-intervals on which the Gauss-Legendre method with 12 points is applied. Starting from I_0 (the approximation obtained with no partition of the integration interval), approximations I_n have been computed in DSA until the

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

difference $I_n - I_{n+1}$ is a computational zero (has no exact significant digit). Table 1 presents the approximations obtained in single and in double precision. In every sequence, only the exact significant digits (*i.e.* not affected by round-off errors) of the last iterate, estimated using DSA, are reported.

method	in single precision	in double precision
trapezoidal	I_8 = 5.1404E-01	I_{19} = 5.140418958899E-001
Simpson	I_8 = 5.14041E-01	I_{10} = 5.1404189589007E-001
Gauss-Legendre	I_1 = 5.140419E-01	I_1 = 5.14041895890070E-001

Table 1
Approximations to I, its 16 first exact digits being 0.5140418958900708, Approximations de I, dont les 16 premiers chiffres exacts sont 0.5140418958900708.

In agreement with theorem 2.1, the exact significant digits of each approximation I_N obtained are in common with I, up to one. 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.

4 Conclusion

Using the dynamical strategy described in this note, one can compute with an approximation method the result for which the global error is minimal. Furthermore its significant digits which are not affected by round-off errors are in common with the exact result, up to one. This strategy has been successfully used for the computation of integrals arising in computational physics, for instance a multiple integral involved in the neutron star theory [6].

References

[1] S. Abbasbandy, 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, X.S. Li, A comparison of three high-precision quadrature schemes, In: Proc. 5th Real Numbers and Computers conference, Lyon, France, 2003, pp 81–95.

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

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

[6] F. Jézéquel, F. Rico, J.-M. Chesneaux, M. Charikhi, Reliable computation of a multiple integral involved in the neutron star theory, Mathematics and Computers in Simulation, 71(1) 44–61, 2006.

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

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

[9] J. Vignes, Discrete Stochastic Arithmetic for Validating Results of Numerical Software, Num. Algo., 37(1–4) 377–390, 2004.