

CERTAINS ASPECTS DU PROBLÈME DE LA PRÉCISION
DANS LES CALCULES NUMÉRIQUES

PAR

TIBERIU POPOVICIU (Cluj)

(Communication au Congrès des mathématiciens roumains, Bucarest, mai-juin 1956)

1. Poussée par des nécessités d'ordre pratique, l'analyse numérique a pris un grand essor dans les dernières décades. Les problèmes de l'analyse numérique sont très variés, mais tous peuvent être caractérisés par ce qu'ils cherchent à obtenir les résultats jusqu'aux derniers calculs numériques.

Dans ce qui suit nous passons en revue, très brièvement, certains aspects tout à fait particuliers de quelques problèmes d'analyse numérique, d'ailleurs présentant une grande importance théorique et pratique, aspects qui rentrent dans le cadre des préoccupations de la Section de Mathématique de la Filiale de Cluj, de l'Académie de la R.P.R.

2. Dans de nombreux problèmes d'analyse numérique on cherche à obtenir un nombre qui dépend d'une fonction f , donc on cherche la valeur d'une fonctionnelle $A[f]$, d'argument f . Pour fixer les idées et pour simplifier, nous supposons pour l'instant, que f est une fonction réelle de variable réelle. La fonctionnelle $A[f]$ peut avoir différentes structures. On rencontre les cas les plus simples dans le calcul de la valeur d'une fonction f ou de sa dérivée d'un ordre quelconque, sur un point donné, ainsi que dans le calcul de l'intégrale de f dans un intervalle donné etc. Ces derniers cas particuliers se rencontrent souvent, par exemple, dans différentes méthodes d'intégration numérique des équations différentielles ou d'autres équations fonctionnelles.

A cause de la trop grande complication ou bien de la connaissance généralement insuffisante de la fonction f on ne peut pas obtenir exactement le nombre $A[f]$. Il faut alors se contenter d'une approximation de ce nombre. Ce qui rend importants ces problèmes de calculs approximatifs c'est qu'en pratique une approximation convenable du nombre $A[f]$ est toujours suffisante.

Il y a différents procédés qui permettent d'obtenir une valeur approximative pour $A[f]$. Un procédé assez général peut être schématisé par l'égalité approximative

$$A[f] \approx B[f], \quad (1)$$

donc on peut prendre comme approximation pour $A[f]$ la valeur $B[f]$ d'une autre fonctionnelle donnée — considérée comme plus simple que $A[f]$ en ce sens

qu'effectivement le calcul du nombre $B[f]$ peut s'effectuer jusqu'au bout et dans certaines conditions imposées d'avance.

Un procédé général d'obtention d'une approximation pour la fonctionnelle $A[f]$ consiste à remplacer la fonction f par une fonction d'approximation φ

$$f \approx \varphi. \quad (2)$$

On peut considérer le nombre $A[\varphi]$ comme une approximation pour $A[f]$ donc

$$A[f] \approx A[\varphi]. \quad (3)$$

D'ailleurs, l'égalité (3) est de la forme (1) si la fonction d'approximation dépend univoquement de la fonction f , donc si φ est la valeur d'un opérateur

$$\varphi(x) = \Phi[f|x] \quad (4)$$

d'argument f . On peut prendre alors

$$B[f] = A[\Phi[f|x]]. \quad (5)$$

Dans ce qui suit, nous nous occuperons seulement de formules d'approximation telles que (5).

3. Dans l'utilisation pratique d'une approximation de la forme (1) il est particulièrement important d'apprécier l'erreur qu'on commet pour cette approximation. Dans la recherche d'une telle évaluation une délimitation satisfaisante du reste

$$R = R[f] = A[f] - B[f] \quad (6)$$

de la formule (1) présente une grande importance. Enfin, pour obtenir une telle délimitation, une étude approfondie de la structure du reste $R(f)$ est très utile.

Si par exemple, $R[f]$ est une fonctionnelle linéaire (additive et homogène) on peut exprimer $R[f]$ — dans des conditions assez générales — à l'aide d'une intégrale (l'intégrale de Stieltjes si l'on se trouve dans les conditions du théorème de F. Riesz).

Il n'est pas suffisant d'exprimer le reste de cette façon, car dans les délimitations pratiques du reste interviennent des propriétés comme celles qui se réfèrent au degré d'exactitude de la formule d'approximation (ou de son reste). Le degré d'exactitude (par rapport à un polynôme) d'une fonctionnelle linéaire $R[f]$ se définit comme étant le plus grand entier n (s'il existe) tel que $R[f] = 0$ pour tout polynôme f de degré n . Dans ce cas le reste peut s'exprimer par une fonctionnelle linéaire de la $(n+1)^{\text{ème}}$ dérivée de f si, bien entendu, certaines conditions d'existence de ces expressions subsistent. Différentes expressions de ce type ont été étudiées par de nombreux auteurs, surtout pour le reste des formules de dérivation et de quadratures numériques.

4. Dans le cas d'une fonctionnelle linéaire $R[f]$ la théorie des fonctions convexes d'ordre supérieur nous permet de trouver une forme remarquable pour $R[f]$. Si $R[f]$ est du degré d'exactitude n et si f est une fonction continue dans un intervalle, l'on a

$$R[f] = \alpha [\xi_1, \xi_2, \dots, \xi_{n+2}; f] + \beta [\xi'_1, \xi'_2, \dots, \xi'_{n+2}; f]. \quad (7)$$

Dans cette relation, valable sous des hypothèses très générales, où α, β sont des constantes ne dépendant pas de la fonction f , on a utilisé le symbole bien connu des différences divisées, et $\xi_1, \xi_2, \dots, \xi_{n+2}$ et $\xi'_1, \xi'_2, \dots, \xi'_{n+2}$ sont des groupements de $(n+2)$ points distincts qui dépendent, en général, de la fonction f [5].

Le cas où un des nombres α, β peut être pris égal à zéro est particulièrement intéressant. Ceci arrive si, et seulement si, $R[f] \neq 0$ pour toute fonction (continue et) convexe d'ordre n .

La forme (7) du reste précise beaucoup sa structure en le rattachant aux propriétés différentielles de la fonction f . Le rapport entre les différences divisées et les propriétés différentielles d'une fonction est assuré par toute une série de propriétés « de moyenne » dont les plus simples sont connus, déjà depuis Cauchy [1], tandis que d'autres ont été établies plus récemment [7].

5. Nous n'insisterons plus sur la structure du reste $R[f]$; nous ferons seulement quelques remarques.

a) La structure (7) du reste $R[f]$, que nous supposons une fonctionnelle linéaire, peut être étudiée dans un cadre plus général à l'aide d'une extension convenable de la notion de degré d'exactitude. Cette généralisation revient à supposer que $R[f]$ s'annule sur l'enveloppe linéaire d'un nombre fini de fonctions linéairement indépendantes. On utilise bien entendu la généralisation correspondante de la notion de différence divisée.

b) Jusqu'à présent on ne connaît rien d'analogue pour le cas où $R[f]$ n'est plus une fonctionnelle linéaire. De telles fonctionnelles, comme restes, semblent être importantes. Par exemple, la méthode d'intégration numérique de Runge-Kutta des équations différentielles revient justement à un procédé d'approximation du type général signalé, dans lequel $R[f]$ n'est plus une fonctionnelle linéaire. Dans ce cas f est une fonction à deux ou plusieurs variables indépendantes. Le fait que le reste dans la méthode de Runge-Kutta a été peu étudié jusqu'à présent, est dû à sa structure compliquée.

Il semble que le problème de la structure du reste $R[f]$ dans le cas où $R[f]$ n'est plus une fonctionnelle linéaire de f , est en relation avec la généralisation de la notion de convexité par rapport à un système interpolateur de fonctions. Dans cette dernière direction il faut rappeler les travaux de M. I. Morozov [4], L. Tornheim [11], E. Moldovan [3].

c) Dans l'exemple déjà cité de la méthode de Runge-Kutta le reste dépend d'une fonction de plusieurs variables. Le reste des formules d'approximation de la forme (1) ou (3), dans lesquelles f est une fonction à plusieurs variables réelles a été beaucoup moins étudiée jusqu'à présent. On connaît — bien entendu — de nombreux résultats en ce qui concerne le reste des formules de dérivation et de quadrature (cubature) numériques. Dans notre pays il faut citer les travaux de D. V. Ionescu [2] dans cette direction.

Mais même dans le cas où $R[f]$ est une fonctionnelle linéaire, une formule analogue à (7) n'a pas été encore étudiée dans le cas général. Nous omettons évidemment les cas simples où les résultats s'obtiennent en partant du cas d'une seule variable et en procédant par superposition successives des variables.

6. Pour obtenir une approximation du nombre $A[f]$ il faut calculer effectivement le nombre $B[f]$. Supposons que ce nombre soit donné par la formule (5) par l'intermédiaire de la fonction d'approximation (4) de f . Pour calculer effectivement $B[f]$ il faut faire un certain nombre d'opérations qui consistent généralement en un nombre fini d'opérations arithmétiques élémentaires (addition, soustraction, multiplication et division).

Pour exécuter les calculs, indifféremment s'ils sont effectués à l'aide d'instruments de calcul ou d'une machine, on suit toujours un programme qui, en particulier, précise l'ordre dans lequel sont effectuées les différentes opérations.

Au cours de l'exécution de chaque opération (élémentaire) et tout spécialement des multiplications et divisions, on commet en général des erreurs dues au fait qu'en pratique on utilise seulement certains nombres, les soit-disant nombres pratiques, qui s'expriment d'habitude par des fractions décimales limitées ou même limitées à un certain nombre de chiffre décimaux. De cette manière, pendant le calcul de $B[f]$ s'accumulent plusieurs erreurs de calcul qui modifieront le résultat final poursuivi. L'approximation calculée de cette façon sera, en définitive, l'approximation effectivement obtenue du nombre $A[f]$.

7. Dans la pratique du calcul numérique, l'établissement pour chaque cas d'un certain programme de calcul joue un grand rôle. Du point de vue théorique de tels programmes pourrions être étudiés, au moins pour les cas concrets les plus simples, à l'aide des différentes expressions qui représentent $B[f]$.

Pour montrer, dans une certaine mesure, la grande variété de problèmes qui se posent en connexion avec ces considérations nous allons exposer un cas concret simple.

Supposons que (4) soit le polynôme d'interpolation de Lagrange,

$$\varphi(x) = L(x_1, x_2, \dots, x_{n+1}; f | x) \quad (8)$$

relatif à la fonction f et sur les nœuds d'interpolation (distincts) $x_1, x_2, \dots, x_{n+1}$, donc, c'est le polynôme de degré n qui coïncide sur les nœuds x_i avec la fonction f .

Supposons ensuite, qu'il s'agisse du problème de l'interpolation à l'aide du polynôme (8) de Lagrange, donc du calcul de valeur $f(x_i)$ de la fonction f sur le point x_0 (supposé différent des nœuds), connaissant les valeurs de la fonction sur les nœuds d'interpolation. Nous sommes donc, au cas où $A[f] = f(x_0)$ et

$$B[f] = L(x_1, x_2, \dots, x_{n+1}; f | x_0). \quad (9)$$

Le calcul de la valeur (9) sur le point x_0 du polynôme (8) peut se faire de différentes manières, d'après les différents programmes de calcul qui peuvent être indiqués et étudiés par les expressions explicites du polynôme de Lagrange.

8. Soit la formule

$$L(x_1, x_2, \dots, x_{n+1}; f | x_0) = \sum_{i=1}^{n+1} \frac{\prod_{j=1}^{n+1} (i) (x_0 - x_j)}{\prod_{j=1}^{n+1} (i) (x_i - x_j)} f(x_i) \quad (10)$$

où l'indice i au symbole $\prod^{(i)}$ indique que la valeur i de j est exclue. Par l'intermédiaire de la formule (10) on peut indiquer par exemple, qu'il faut effectuer une somme de $n+1$ termes déjà calculés. Chaque terme de la somme se calcule en général par $2n-1$ multiplications et une division. Remarquons que grâce aux résultats partiels, le nombre des multiplications peut être réduit. Ainsi, pour obtenir les produits

$$\prod_{j=1}^{n+1} (i) (x_0 - x_j) \quad (i = 1, 2, \dots, n+1) \quad (11)$$

il n'est pas nécessaire de faire $(n+1)(n-1) = n^2 - 1$ multiplications, mais en général, $3n-3$ multiplications suffisent. Une telle réduction du nombre des opérations à effectuer présente un grand intérêt pratique, évidemment aussi en rapport avec d'autres conditions qu'un programme de calcul doit remplir.

Une autre remarque qu'on peut faire concernant le calcul précédent est que, dans l'obtention de chaque terme la multiplication par $f(x_i)$ peut être antérieure ou postérieure à l'opération de division qui doit être faite, en général, pour trouver ce terme.

La première manière de procéder, qui consiste à calculer en préalable les coefficients de $f(x_i)$, ($i = 1, 2, \dots, n+1$) de la formule (10) est avantageuse si l'on doit appliquer la formule (10) à plusieurs fractions ayant les mêmes nœuds x_i et le même point x_0 . Le second procédé qui consiste à laisser à la fin l'effectuation de la division dans le calcul des termes (éventuellement seulement de certains termes) de la somme (10) peut présenter des avantages dans le cas de données particulières simples, permettant de profiter de certaines simplifications au cours du calcul, comme par exemple, des simplifications de fractions ordinaires etc. et qui ne sont pas avantageuses dans la première façon d'organiser les calculs. Il faut cependant souligner que de telles simplifications ne peuvent pas être prévues théoriquement et n'entrent pas dans le cadre des préoccupations théoriques du calcul numérique.

9. On connaît les avantages que présente dans le problème de l'interpolation par polynômes, la formule de Newton :

$$L(x_1, x_2, \dots, x_{n+1}; f | x_0) = \sum_{i=0}^n (x_0 - x_1)(x_0 - x_2) \dots (x_0 - x_i) [x_1, x_2, \dots, x_{i+1}; f] \quad (12)$$

(le premier terme étant $f(x_1)$), où $[x_1, x_2, \dots, x_{i+1}; f]$ est la différence divisée de la fonction f sur les nœuds $x_1, x_2, \dots, x_{i+1}$. Le programme de calcul indiqué d'habitude par la formule (12) consiste à obtenir successivement les nombres $y_1, y_2, \dots, y_n$ où

$$y_i = [x_1, x_2, \dots, x_{n-i+1}; f] + (x_0 - x_{n-i+1}) y_{i-1} \quad (i = 1, 2, \dots, n), \quad (13)$$

$$y_0 = [x_1, x_2, \dots, x_{n+1}; f] \quad y_n = L(x_1, x_2, \dots, x_{n+1}; f | x_0)$$

qui comporte en général n multiplications, donc moins (pour n assez grand) que le programme indiqué pour la formule (10).

On ne doit pas oublier qu'ici ont été calculées au préalable les différences divisées $[x_1, x_2, \dots, x_{i+1}; f]$ ($i = 1, 2, \dots, n$). Le calcul de ces différences divisées comporte généralement des erreurs de la nature de celles signalées à cause des divisions, par les différences des nœuds qui interviennent. Il faut alors, comme nous l'avons fait dans une autre occasion [9], examiner l'influence des erreurs de calcul commises dans la formation du tableau des différences divisées sur le programme de calcul indiqué pour la formule (12).

10. Pour conclure, et concernant le programme de calcul signalé pour la formule (12) on peut faire les remarques suivantes:

a) En recherchant à réduire au minimum les erreurs de calcul qui s'accumulent par ce programme, on trouve une justification à l'emploi des différentes formules particulières d'interpolation d'après la position du point x_0 , parmi les nœuds. De cette façon, et par cette voie, se justifient comme nous l'avons montré

auparavant [6], l'application pratique de la formule de Newton à différences ascendantes ou descendantes, des formules d'Euler, Stirling, Bessel etc. suivant la position du point x_0 près des extrémités ou près du centre etc, du tableau des différences (les nœuds sont supposés équidistants et arrangés en ordre croissant ou décroissant).

b) Dans le cas des nœuds équidistants, au lieu des différences divisées on utilise les différences habituelles dont la formation ne nécessite pas de divisions. On peut donc admettre en général, que le tableau des différences ne présente pas des erreurs de calcul. Nous avons montré dans un autre travail [8], que pour le programme (13) modifié en ce sens, on peut préciser une délimitation des erreurs de calcul accumulées, établissant aussi des tableaux suffisants en pratique pour trouver rapidement de telles délimitations.

BIBLIOGRAPHIE

1. A. Cauchy, *Sur les fonctions interpolaires*, C. R. de l'Acad. Sci. Paris **11** (1840), 775.
2. D. V. Ionescu, *Formule de cubatură în care domeniul de integrare este un triunghi oarecare*, Stud. și Cerc. Științ., Filiala Cluj, VI (1955), 7.
3. E. Moldovan, *O generalizare a noțiunii de convexitate*, Stud. și Cerc. Științ., Filiala Cluj, VI (1955), 65.
4. М. И. Морозов, *О некоторых вопросах равномерного приближения непрерывных функций посредством функций интерполяционных классов*, Изв. АН СССР **16** (1952), 75.
5. T. Popoviciu, *Asupra formei restului în unele formule de aproximare ale analizei*, Lucrările Sesiunii generale științifice din 2—12 iunie 1950, Ed. Acad. R.P.R., București 1951, p. 183.
6. — *Considerații teoretice asupra utilizării practice a unor formule de interpolare*, Bul. științ. Acad. R.P.R., Sect. Șt. Mat. Fiz., XII, 2 (1951), 441.
7. — *Folytonos függvények k öpértékteleiről*, A Magyar Tud. Akad. II Oszt. Közl. IV (1954), 353.
8. — *Despre precizia calculului numeric în interpolarea prin polinomul lui Newton pe noduri echidistante*, Stud. și Cerc. Științ. Filiala Cluj VI (1956), 20.
9. — *Despre precizia calculului numeric în interpolarea prin polinoame*, Bul. Științ. Acad. R.P.R., Sect. Șt. Mat. Fiz., VII, 4 (1955), 953.
10. J. Radon, *Restausdrücke bei Interpolation und Quadraturformeln durch bestimmte Integrale*, Monatshefte für Math. u. Physik **42** (1935), 389.
11. L. Tornheim, *On n-parameter Families of Functions and Associated Convex Functions*, Trans. Amer. Math. Soc. **69** (1950), 457.