Comptes rendus de l'Académie des sciences. Série 1, Mathématique

VII. APPLICATIONS. — Dans [6] cette technique est appliquée aux équations de Navier-Stokes et de Saint-Venant. Ces équations vérifient évidemment les hypothèses du théorème 2. Par exemple pour le système de Saint-Venant linéarisé en dimension 2 dans le cas subsonique aval nous obtenons (la condition d'ordre 0 dans le cas non visqueux s'écrit $\varphi - cu_1 = 0$) :

ordre 0 :

$$\nu \frac{\partial u_1}{\partial x_1} = \frac{c-U}{c}(-cu_1+\varphi), \qquad \nu \frac{\partial u_2}{\partial x_1} = 0$$

ordre 1 :

$$\nu \frac{\partial u_1}{\partial x_1} = \frac{c-U}{c}(-cu_1+\varphi), \qquad \nu \frac{\partial u_2}{\partial x_1} = -\frac{\nu}{U}\frac{\partial u_2}{\partial t}.$$

Note reçue le 21 avril 1988, acceptée le 7 juin 1988.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] S. ABARBANEL, A. BAYLISS et L. LUSTMAN, *Non reflecting boundary conditions for the compressible Navier-Stokes equations*, ICASE report n° 86.9.

[2] A. BAYLISS et E. TURKEL, Outflow boundary conditions for fluid dynamics, *S.I.A.M. Journal of Stat. Comp.*, 3, n° 2, 1982, p. 250-259.

[3] B. ENGQUIST et A. MAJDA, Absorbing boundary conditions for the numerical simulation of waves, *Math. of Comp*, 31, n° 139, 1977, p. 629-651.

[4] B. GUSTAFSSON et A. SUNDSTROM, Incompletely parabolic problems in fluid dynamics, *S.I.A.M. Journal of App. Math.*, 46, n° 174, 1978, p. 343-358.

[5] L. HALPERN, Artificial boundary conditions for the linear advection diffusion equation, *Math. of Comp.*, 46, n° 174, avril 1986, p. 425-439.

[6] L. HALPERN, *Approximate boundary conditions for incompletely parabolic problems*, internal report U.C.L.A.

[7] L. HALPERN et M. SCHATZMAN, Artificial boundary conditions for incompressible viscous flows, *S.I.A.M.* (à paraître).

[8] D. MICHELSON, Initial boundary value problems for incomplete singular perturbations of hyperbolic systems, *Lectures in Applied Mathematics*, 22, 1985.

[9] J. OLIGER et A. SUNDSTROM, Theoretical and practical aspects of some initial boundary value problems in fluid dynamics, *S.I.A.M. Journal of Appl. Math.*, 35, n° 3, 1978, p. 419-447.

[10] R. STRIKWERDA, Initial boundary value problems for incompletely parabolic systems, *C.P.A.M.*, XXX, 1977, p. 797-822.

Centre de Mathématiques appliquées, École Polytechnique, 91128 *Palaiseau Cedex.*

C. R. Acad. Sci. Paris, t. 307, Série I, p. 417-422, 1988

Analyse numérique/*Numerical Analysis*

Modélisation et conditions de validité de la méthode CESTAC

Jean-Marie CHESNEAUX

Résumé — En se basant sur une modélisation probabiliste de CESTAC, nous montrons, sous des hypothèses généralement vérifiées dans les problèmes réels, l'efficacité de la méthode pour estimer la précision des résultats de calculs sur ordinateurs.

Modelisation and study of CESTAC method

Abstract — *From probabilistic modelisation of CESTAC method and under assumptions generally satisfied in real problems, the efficiency of this method for estimating accuracy of computed results is shown here.*

Abridged English Version — 1. PROBABILISTIC CESTAC METHOD. — The probabilistic CESTAC method of La Porte and Vignes [14] consist in randomly perturbing the last bit of the mantissa of each intermediate result. Then a statistical estimate of the accuracy of the final result is obtained with Student's test. We are faced wich two problems: Is Student's test good estimating the average of the informatical results and what is the relation between the exact result and this average?

2. MODELISATION OF CESTAC METHOD. — Hamming and Knuth ([8], [9]) have shown that the mantissa was logarithmically distributed. Feldstein and Goodman [5] have proved that under this assumption round-off errors are uniformally distributed on [0, 1] for chopping arithmetic and on $[-1/2, +1/2]$ for rounding arithmetic. Hull and Swenson [9] have studied and concluded to the validity of the probabilistic model for round-off errors propagation. An informatical result X before and after rounding and perturbation may be modelise by:

$$X_{after} = X_{before} - 2^{E-p}.\varepsilon.(\alpha - h)$$

where E, ε, α, h, p are respectively the exposant and the sign of X, the round-off errors, the perturbation and the number of the bits of the mantissa. So CESTAC method is more a stimulation than a simulation of round-off errors.

We assume that: (i) the Feldstein and Goodman's conclusions are still true for random perturbed arithmetic;

(ii) neither intermediate result is an informatical zero in the mean of Vignes [15], then we have [1]:

THEOREM. — *At first order in respect to* 2^{-p},

$$R = r + \sum_{i=1}^{n} g_i'(d).2^{-p}.(\alpha_i - h_i) \tag{1}$$

where $g_i'(d)$, α_i, h_i, *and* n *are respectively constants depending only on data, the round-off errors due to computer, perturbations and the number of operations.*

Remark. — We observe as experimentally found in [2] that the loss of accuracy, $\mathrm{Log}_2(R - r/2^{-p}.r)$, is independant of p.

Note présentée par Jacques ARSAC.

0249-6291/88/03070417 $ 2.00

3. STUDY OF BIAS. — The average of R is the exact result if $E[\alpha_i - h_i] = 0$. It depends only on the first hypothesis. This hypothesis will be favoured by a regular distribution of the mantissa of data. In the same way, a sufficient number of operations compared with the number of data leads to a "mixing" of intermediate mantissa and contribute to the validity of the hypothesis 1.

4. APPLICATION OF STUDENT'S TEST IN CESTAC METHOD. — The number of significant digits given by CESTAC method is $C_{\bar{R}} = \mathrm{Log}_{10}(\bar{R} . \sqrt{N}/s . t_0)$ where $\bar{R}$ and s are empirical average and standard deviation.

The effect of non normality on Student's test was studied by Gayen [7], Srivastava [13], Tiku [12]. Let $R = r + \sum_{i=1}^{n} g_i'(d) . 2^{-p} . (\alpha_i - h_i) = r + \sum_{i=1}^{n} Z_i$, let $\mu_{i,k}$, μ_k be the k-th centered moment of Z_i and R, and let λ_3, λ_4 be the third and fourth cumulants of R:

$$\lambda_3 = \left(\sum_{i=1}^{n} \mu_{i,3} \right) / \mu_2^{1.5}, \qquad \lambda_4 = \left(\sum_{i=1}^{n} \mu_{1,4} - 3 . \sum_{i=1}^{n} \mu_{i,2}^2 \right) / \mu_2^2.$$

By the Berry-Essen theorem [6] we deduce that if the intermediate results are approximatively of the same magnitude, R is not far from the normality. Therefore Gayen has shown that: $P(|t| > t_0) = P_0(t_0) + \lambda_3 . P_1(t_0) - \lambda_4 . P_2(t_0) + \lambda_3^2 . P_3(t_0)$ where t is the Student-distribution associated with R, $P_0(t_0) = 1 - \beta$ the theoretical probability of Student and the other $P_i(t_0)$ functions of t_0 depending only on N. For $N = 3$, $\beta = 0.95$ we have $t_0 = 4.303$, $P_2(t_0) \approx 0.0032$. $\lambda_3 = 0$ because of the symmetrical distribution of R and according to (1) we have:

$$\lambda_4 = v . \sum_{i=1}^{n} (g_i'(d))^4 \Big/ \left(\sum_{i=1}^{n} (g_i'(d))^2 \right)^2$$

with $v = -1.2$ and 0.759 respectively for chopping and rounding arithmetic. So the corrections on $1 - \beta$ is at most 0.0076 for chopping and 0.048 for rounding. We can conclude to the validity of CESTAC.

INTRODUCTION. — Nous présentons ici les conditions que doivent satisfaire les algorithmes pour que la méthode de permutation-perturbation due à La Porte et Vignes [11], connue aussi sous le nom de CESTAC, implémentée sur ces algorithmes fournisse correctement la précision des résultats. Nous ne considérons que la méthode de perturbation aléatoire de CESTAC.

Pour simplifier, nous considérons un algorithme fini qui, en n opérations arithmétiques fournit un résultat unique $r \in \mathbb{R}$. Nous nous proposons d'estimer le lien entre un résultat informatique R et le résultat mathématique r découlant de la même suite d'opérations effectuée avec une précision infinie.

1. RAPPEL DE LA MÉTHODE CESTAC. — La méthode CESTAC (Contrôle et Estimation Stochastique des Arrondis de Calcul) [14] basée sur une approche stochastique de la propagation des erreurs d'arrondi consiste

— à perturber aléatoirement le dernier bit de la mantisse après chaque opération arithmétique en virgule flottante; le résultat informatique de tout algorithme exécuté sur ordinateur devient alors une variable aléatoire;

— à générer de la sorte une population statistique limitée à deux ou trois résultats informatiques R_i;

— à prendre $\bar{R}$, moyenne des R_i, comme résultat informatique;

— puis à appliquer le test de Student afin de déterminer le nombre de chiffres significatifs $C_{\bar{R}}$ exacts de $\bar{R}$ (nombre de chiffres communs à $\bar{R}$ et à r).

2. Exposé du problème. — La validité de la méthode soulève deux problèmes totalement indépendants:

(*a*) Peut-on utiliser le test de Student afin d'obtenir un intervalle de confiance pour la moyenne de la population totale des résultats.

(*b*) Où se situe le résultat exact r vis-à-vis de cette population?

3. Modélisation de la méthode CESTAC. — Dans l'approche probabiliste on considère que les erreurs d'arrondi sont des quantités aléatoires, Hamming et Knuth ([8], [10]) ont expliqué empiriquement que les mantisses étaient logarithmiquement réparties sur [1/2, 1].

Feldstein et Goodman [5], sous cette conjecture, ont démontré que les erreurs d'arrondi étaient approximativement uniformément distribuées sur [0, 1] ou $[-1/2, +1/2]$ suivant l'arithmétique (troncature ou arrondi). Hull et Swenson [9] ont montré la validité de l'approche stochastique en simulant les erreurs d'arrondi en simple précision à l'aide de la double préxision.

La méthode CESTAC consiste à ajouter une quantité aléatoire h sur la mantisse après arrondi. Si X est un résultat informatique intermédiaire avant arrondi et perturbation, X_{per} ce même résultat après arrondi et perturbation:

$$X_{per} = X - 2^{E-p}.\varepsilon.(\alpha - h)$$

où E, p, ε, α, h représentent respectivement l'exposant de X, le nombre de bits de la mantisse, le signe de X, l'arrondi de la machine ramené à [0, 1], et la perturbation aléatoire.

La méthode CESTAC n'est donc pas une simulation mais plutôt une stimulation des erreurs d'arrondi. Si l'on pose comme problème le choix de la distribution de h pour stimuler de façon optimale α (minimiser l'augmentation d'erreur) on démontre [1] que $h(x)=0$ ou 1 avec une probabilité 1/2 pour l'arithmétique tronquée, $h(x)=-1$ ou $+1$ avec probabilité 1/4 et 0 avec probabilité 1/2 pour l'arithmétique arrondie constituent les meilleures solutions. Remarquons que les perturbations en troncature ont permis de recentrer les erreurs d'arrondi.

Hypothèse 1. — Nous supposons que les conclusions de Feldstein et Goodman sont encore valables dans le cas de l'arithmétique perturbée et que les erreurs d'arrondi ainsi équidistribuées sont indépendantes.

Hypothèse 2. — Nous supposons qu'aucun résultat intermédiaire n'est un zéro informatique au sens de Vignes [15].

Cette hypothèse équivaut à supposer que les exposants et signes des résultats intermédiaires ne dépendent pas des perturbations. On démontre alors [1]:

Théorème. — *Au premier ordre en* 2^{-p},

$$R = r + \sum_{i=1}^{n} g_i'(d).2^{-p}.(\alpha_i - h_i) \tag{3}$$

avec $g_i'(d)$, α_i, h_i, p *et* n *étant respectivement des constantes ne dépendant que des données et de l'algorithme, les arrondis de machines et les perturbations aléatoires lors des calculs intermédiaires, le nombre de bits de la mantisse et le nombre d'opérations.*