L'approche probabiliste des erreurs d'arrondi

Le mode d'arrondi alétoire

X_{min} (resp. X_{max}) : le plus petit (resp. le plus grand) nombre flottant:

$$\forall x \notin \mathbb{F} \text{ et } x \in [X_{min}, X_{max}], \exists \left\{X^-, X^+\right\} \in \mathbb{F}$$

tels que

$$X^- < x < X^+ \quad \text{et} \quad \left[X^-, X^+\right] \cap \mathbb{F} = \emptyset$$

Le mode d'arrondi aléatoire : X^- ou X^+ avec probabilité $1/2$.

Qualité numérique des modes d'arrondi

$$X = \varepsilon.M.2^E \text{ avec } X = x - \varepsilon.2^{E-p}.\alpha$$

- en arrondi au plus près, $\alpha \in [-0.5, 0.5[$;
- en arrondi vers zéro, $\alpha \in [0, 1[$;
- en arrondi vers plus ou moins l'infini, $\alpha \in]-1, +1[$;
- arrondi aléatoire, $\alpha \in]-1, +1[$

Tous les mode d'arrondi sont **exacts**.

Les modes d'arrondi sont-ils numériquement de qualité équivalente ?

Modélisation d'un résultat numérique

$$R = r + \sum_{i=1}^{n} u_i(d).2^{E_i - p}.\varepsilon_i.\alpha_i \quad + \quad \mathcal{O}\left(2^{-2.p}\right)$$

R : Résultat informatique,

r : résultat mathématique exact,

n : nombre d'opérations arithmétiques élémentaires,

$u_i(d)$: coefficients ne dépendant que des données,

(E_i, ε_i) : exposants et signes des résultats intermédiaires,

α_i : erreur d'arrondi de la i-ème opérations,

p : nombre de bits contenus dans la mantisse.

Notion de chiffres significatifs exacts

Définition 1 *Soient a et b deux nombres réels, le nombre de chiffres significatifs dans la base d en commun entre a et b est défini dans $[-\infty, +\infty]$ par*

1. *pour* $a \neq b$, $\quad C_{a,b} = \log_d \left| \frac{a+b}{2.(a-b)} \right| \Longleftrightarrow |a-b| = \left| \frac{a+b}{2} \right| .d^{-C_{a,b}}$,
2. $\forall a \in \mathbf{R}, \quad C_{a,a} = +\infty$.

Remarque : Si $|a-b| \ll |a+b|$, on peut prendre $C_{a,b} \approx \log_d \left| \frac{a}{a-b} \right|$.

Le nombre de chiffres significatifs exacts d'un résultat informatique X est donc $C_{X,x}$ où x est le résultat mathématique exact.

Un théorème sur la qualité numérique

Le nombre de chiffres binaires en commun entre R et r est donné par

$$C_R \approx -\log_2 \left| \frac{R-r}{r} \right| = p - \log_2 \left| \sum_{i=1}^{n} u_i(d).2^{E_i}.\varepsilon_i.\frac{\alpha_i}{r} \right|$$

Théorème 1 *La perte de précision lors d'un calcul sur ordinateur est statistiquement indépendante de la précision utilisée pour ce calcul.*

Lié à la validité de l'approximation au premier ordre du modèle.

I. Duff and al., An exeriment with arithmetic precision on linear algebra computations, (1986)

L'approche probabiliste

α_i : modélisées par des variables aléatoires iid.

R est modélisé par :

$$Z = r + \sum_{i=1}^{n} g_i(d).2^{-p} z_i \ ,$$

les z_i étant des variables aléatoires iid sur $[-1, +1]$.

Remarques :

1. En mode d'arrondi aléatoire, α_i est vraiment aléatoire,
2. les α_i sont-elles indépendantes ?
3. les α_i sont-elles centrées ?

Quelques dates - I

- T.E. Hull et J.R. Swenson, test of probablistic models for propagation of round-off errors, A.C.M., vol. 9, n° 2 (1966).
- D.E. Knuth, The art of computer programming, Addison-Wesley, vol. 2 (1969).
- R.W. Hamming, On the distribution of numbers, The Bell System Tech. Journal (1970).
- M. La Porte et J. Vignes, Error analysis in computing, IFIP congress, Stockholm (1974).
- A. Feldstein et R. Goodman, Convergence estiamtes for the distribution of trailling digits, A.C.M., vol. 23 (1976)

Quelques dates - II

- J.-M. Chesneaux, Modélisation et conditions de validité de la méthode CESTAC, C.R.A.S., tome 307 (1988)
- P. Francois, Contribution à l'étude de méthodes de controle automatique de l'erreur d'arrondi: la méthodologie SCALP, Thèse (1989)
- J. Erhel, Experiments with data perturbations to study conditions numbers and numerical stability, Computing, vol. 51 (1993)
- F. Chatelin et V. Fraissé, Lecture on Finite Precision Computations, SIAM (1995)

La méthode CESTAC - I

Jean Vignes et Michel La Porte - 1972.

Effectue un test statistique sur tout résultat R.

- N exécutions parallèles du programme avec le mode d'arrondi aléatoire
  $\Longrightarrow$ N résultats différents R_i pour tout résultat intermédiaires.
- Résultat informatique : $\overline{R} = \dfrac{1}{N} \cdot \sum_{i=1}^{N} R_i$.

La méthode CESTAC - II

- Nombre de chiffres significatifs exacts de $\overline{R}$:

$$C_{\overline{R}} = \log_{10}\left(\frac{\sqrt{N}.\,|\overline{R}|}{s.\tau_\beta}\right) \quad \text{avec} \quad s^2 = \frac{1}{N-1}.\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2,$$

τ_β : le seuil de confiance de la distribution de Student pour une probabilité β.

En pratique :

$N = 2$ **ou** 3**,** $\beta = 0.95$ **et** $\tau_\beta = 12,706$ **ou** $4,303$**.**

Validitation théorique de la méthode

Validation de l'application du test de Student à Z.

$$\Longrightarrow \qquad \forall \beta \in [0,1], \exists \tau_\beta \in \mathbf{R} \text{ tel que}$$

$$P\left(\left|\overline{Z} - r\right| \leq \frac{\tau_\beta . s}{\sqrt{N}}\right) = \beta$$

donc le nombre de chiffres décimaux significatifs communs à $\overline{Z}$ et à r, est minoré par

$$C_{\overline{Z}} = Log_{10}\left(\frac{\sqrt{N}.\left|\overline{Z}\right|}{s.\tau_\beta}\right).$$

Les hypothèses de validation

1. Les signes et exposants des résultats intermédiaires sont indépendants des perturbations.
2. α_i modélisées par des variables aléatoires indépendantes équiréparties.
3. L'approximation au premier ordre en 2^{-p} dans l'expression de Z.
4. Régularité des coefficients $u_i(d)$.

Discussion des hypothèses - I

Hypothèse 1 : Hypothèse de travail. Quelques coefficients multipliés par 2.

$\Longrightarrow$ Peu d'influence sur le résultat.

Hypothèse 4 :

$$\exists\ a > 0,\ \exists\ A\ >\ 0 \text{ tels que } \forall\ i,\ a\ \leq\ |g_i(d)|\ \leq\ A.$$

Assure une *continuité* de la distribution sinon, risque de rassemblement autour de quelques Dirac donc risque de surévaluation de la précision $\Longrightarrow$ **technique de la "balance".**

L'arrondi du dernier passage est l'opposé de l'avant dernier.

Discussion des hypothèses - II

Hypothèse 2 : Entraine que $E(Z) = r$

C'est un modèle $\Longrightarrow$ **Biais de type 1.**

Résiste à des biais de plusieurs σ.

$$\sigma^2 \;=\; \frac{2}{3}\;.\;\sum_{i=1}^{n}(g_i(d))^2 \quad \text{et} \quad B_I \leq \sum_{i=1}^{n}|g_i(d)|\,.$$

Il faut beaucoup d'opérations **et** des arrondis systématiquement biaisés. Mais on injecte réellement de l'aléatoire dans le calcul.

Discussion des hypothèses - III

Hypothèse 3 : La plus fondamentales.

Entraine que Z représente fidèlement R. CESTAC ne fonctionne que sur $Z \Longrightarrow$ **Biais de type 2.**

Exemple : $U_{n+1} = 111 - \frac{1130}{U_n} + \frac{3000}{U_n.U_{n-1}}$ avec $U_1 = \frac{11}{2}$ et $U_2 = \frac{61}{11}$.

Avec CESTAC, sur station SPARC en double précision,
$U(29) = 100.000000000000$.

La distribution s'est déplacé sur le DIRAC 100.

$\Longrightarrow$ invalidation de l'approximation au 1er ordre, déviation totale de la trajectoire mathématique.

L'analyse du biais de type 2

Avec {+,-}, pas de terme d'ordre 2.

Soient $X_i = x_i + \varepsilon_i$ pour $i = 1, 2$:

$$\begin{aligned} X_1.X_2 &= x_1.x_2 + x_1.\varepsilon_2 + \varepsilon_1.x_2 + \varepsilon_1.\varepsilon_2 \\ \frac{X_1}{X_2} &= \frac{x_1}{x_2} + \frac{\varepsilon_1}{x_2} - \frac{x_1.\varepsilon_2}{x_2^2} + \mathcal{O}\left(\frac{\varepsilon_2^2}{x_2^2}\right) \end{aligned}$$

Prépondérance des termes d'ordre 1 si :

Pour "*" : $\varepsilon_1 \ll X_1$ ou $\varepsilon_2 \ll X_2 \Rightarrow$ **multiplications instables**.

Pour "/" : $\varepsilon_2 \ll X_2$ $\Rightarrow$ **divisions instables**.

$\Longrightarrow$ autovalidation possible de la méthode.

Sur la probabilité de l'intervalle de confiance

A un chiffre près $\iff \dfrac{\sqrt{N}.\left|\overline{R}\right|}{10.s.t_\beta} \leq \left|\overline{R}-r\right| \leq \dfrac{10.\sqrt{N}.\left|\overline{R}\right|}{s.t_\beta}$

Soit $P1 = P\left(\dfrac{\sqrt{N}.\left|\overline{R}\right|}{10.s.t_\beta} > \left|\overline{R}-r\right|\right)$ (sous-évaluation).

Soit $P2 = P\left(\dfrac{10.\sqrt{N}.\left|\overline{R}\right|}{s.t_\beta} < \left|\overline{R}-r\right|\right)$ (sur-évaluation).

	P1	P2
Théorie	0.29	0.00054
CESTAC	0.34	0.00025

$N = 3$

Sur le nombre de passage

2 ou 3 sont suffisants.

Expérimentalement, pas d'amélioration avec $N > 3$.

Le nombre de chiffre est en $\log_{10}\left(\frac{1}{\sqrt{N}}\right)$

$\Longrightarrow$ pour gagner 1 chiffre de précision, il faut multiplier N par 100.

On met alors en évidence l'erreur de méthode sans affiner l'estimation de la précision.

L'incohérence de la virgule flottante

Les opérateurs arithmétiques sont une approximation.

Les relations d'ordre sont la traduction exacte des mathématiques.

$\Longrightarrow$ disfonctionnement de l'ensemble.

$$X = Y \quad \not\Rightarrow \quad x = y \quad \text{et} \quad x = y \quad \not\Rightarrow \quad X = Y.$$

$$X \geq Y \quad \not\Rightarrow \quad x \geq y \quad \text{et} \quad x \geq y \quad \not\Rightarrow \quad X \geq Y.$$

Un exemple instructif - I

$$0.3 * x^2 \quad + \quad 2.1 * x \quad + 3.675 \quad = \quad 0$$

- **Arrondi au plus près**

  d = -3.81470E-06

  Il y a deux racines complexes conjuguees.

  z1 = -.3500000E+01 + i * 0.9765625E-03

  z2 = -.3500000E+01 + i * -.9765625E-03

- **Arrondi vers zéro**

  d = 0.

  Le discriminant est nul.

  La racine double reelle est -.3500000E+01

Un exemple instructif - II

$$0.3 * x^2 \quad + \quad 2.1 * x \quad + 3.675 \quad = \quad 0$$

- **Arrondi vers plus l'infini**

  d = 3.81470E-06

  Il y a deux racines reelles distinctes.

  x1 = -.3500977E+01

  x2 = -.3499024E+01

- **Arrondi vers moins l'infini** d = 0.

  Le discriminant est nul.

  La racine double reelle est -.3500000E+01

Un mauvais choix

$$\begin{pmatrix} 21 & 130 & 0 & 2.1 \\ 13 & 80 & 4.74\ 10^8 & 752 \\ 0 & -0.4 & 3.9816\ 10^8 & 4.2 \\ 0 & 0 & 1.7 & 9\ 10^{-9} \end{pmatrix} .X = \begin{pmatrix} 153.1 \\ 849.74 \\ 7.7816 \\ 2.6\ 10^{-8} \end{pmatrix}$$

$$x^t_{sol.ex.} \;=\; \left(1, 1, 10^{-8}, 1\right).$$

Sur station Sun, en simple précision arrondie au plus près :

$$x^t_{sol.cal.} \;=\; (62.61987, -8.953979, 0.0, 1.0)\,.$$

Lors de la réduction de la troisième colonne, le pivot $A(3,3)$ vaut 4864. Mais la valeur mathématique de $a(3,3)$ est zéro.

La notion de zéro informatique

Introduite par Jean Vignes en 1982.

Définition 2 *Soient* $(X_1, .., X_N)$ N *représentants obtenus par la méthode CESTAC d'un résultat informatique* X, X *est un* **zéro informatique**, *noté @.0, si et seulement si*

$$C_{\overline{X}} \leq 0 \quad ou \quad X_1 = .. = X_N = 0.$$

Cela signifie que 0 est dans l'intervalle de confiance.

Les définitions stochastiques

Définition 3 *Soient X et Y deux résultats informatiques (N-uplet), X est stochastiquement égal à Y, noté X s= Y, si et seulement si*

$$X - Y = @.0.$$

Définition 4 *Soient X et Y deux résultats informatiques (N-uplet), X est stochastiquement strictement supérieur à Y, noté X s> Y, si et seulement si*

$$\overline{X} > \overline{Y} \quad et \quad X \ s\neq \ Y.$$

On appelle **arithmétique stochastique discrète** l'utilisation conjointe sur ordinateur de la méthode CESTAC et des définitions stochastiques.

Quelques propriétés

Propriété 1 $x = 0 \implies X = @.0$.

Propriété 2 $X \; s\neq \; Y \implies x \neq y$.

Propriété 3 $X \; s> \; Y \implies x > y$.

Propriété 4 $x \geq y \implies X \; s\geq \; Y$.

Propriété 5 *La relation s> est transitive.*

Propriété 6 *La relation s= est reflexive, symétrique mais non transitive.*

Propriété 7 *La relation s≥ est reflexive, anti-symétrique mais non transitive.*

Exemple 1

$$0.3 * x^2 \quad + \quad 2.1 * x \quad + 3.675 \quad = \quad 0$$

- **En ASD**

  d = @.0

  Le discriminant est nul.

  La racine double reelle est -.3500000E+01

Exemple 2-a

Soit l'algorithme $\mathcal{A}$ appliqué à des a_i positifs :

$$
\begin{array}{l}
max = 0; \\
Pour\ i = 1, ..., n\ faire \\
\quad si\ max < a_i\ alors\ max = a_i;
\end{array}
$$

Théorème 2 *L'algorithme $\mathcal{A}$, appliqué à une famille de résultats informatiques a_i, fournit une valeur max vérifiant :*

$$\forall\ 1 \leq i \leq n, \quad a_i\ s\!\leq\ max.$$

Appliqué à la recherche d'un pivot max $\Longrightarrow$ le pivot est choisi parmi les valeurs **significatives**.

Exemple 2-b

$$\begin{pmatrix} 21 & 130 & 0 & 2.1 \\ 13 & 80 & 4.74\ 10^{8} & 752 \\ 0 & -0.4 & 3.9816\ 10^{8} & 4.2 \\ 0 & 0 & 1.7 & 9\ 10^{-9} \end{pmatrix} .X = \begin{pmatrix} 153.1 \\ 849.74 \\ 7.7816 \\ 2.6\ 10^{-8} \end{pmatrix}$$

$x^{t}_{sol.ex.} \;=\; \big(1, 1, 10^{-8}, 1\big).$

x_ sol(1) = 0.100E+01

x_ sol(2) = 0.9999E+00

x_ sol(3) = 0.1000000E-07

x_ sol(4) = 0.9999999E+00

car $A(3,3)$ est écarté du choix du pivot.

Aujourd'hui

- La maitrise de la méthode est acquise
- Etude de la notion de convergence en ASD
- Recherche d'algorithmes prouvés reposant uniquement sur les propriétés de l'ASD.