

## Chapitre 7

# Méthode de contrôle et d'estimation stochastique des arrondis de calcul

Sur un ordinateur les algorithmes numériques ( fini ou itératif ou approché ) donnent des résultats approximatifs dus aux erreurs de codage ou d'arrondi qui peuvent se propager ou à des tests d'arrêt non robustes ( arrêts trop tôt ou trop tard (calculs supplémentaires) qui donnent de mauvais résultats par rapport à la solution mathématique ) ou à l'erreur de méthode.  
Par conséquent, il est important d'étudier et de contrôler ces erreurs lors d'un calcul sur ordinateur en arithmétique flottante.

## 1 Estimation stochastique de la précision sur des résultats de calcul :

Il s'agit d'une méthode mise au point par La Porte et Vignes (1974) pour l'analyse des erreurs informatiques, appelée CESTAC.

principe :

Soit un algorithme numérique défini par :

procédure  $p$  ( $\mathcal{D}, r, +, -, \times, /$ , fonct)

$\mathcal{D} \subset \mathbb{R}$  ensemble des données

$r \in \mathbb{R}$  résultat mathématique

$+, -, \times, /$  opérations dans  $\mathbb{R}$

fonct fonction réelle.

Sur ordinateur, cet algorithme a pour image :

PROCEDURE  $P$  ( $D, R, \oplus, \ominus, \otimes, \oslash$ , Fonct)

$D \subset \mathbb{F}$ , ensemble des nombres flottants

$R \in \mathbb{F}$

résultat informatique

$\oplus, \ominus, \otimes, \oslash$  opérations dans  $\mathbb{F}$

Fonct fonction flottante.

A une procédure informatique  $P$  comportant  $k$  opérations ( par exemple arithmétiques ) correspond au plus  $2^k$  résultats informatiques éventuellement distincts.

Si  $Z = XopY$  avec  $op \in \{\oplus, \ominus, \otimes, \oslash\}$  alors 2 résultats sont possibles  $Z^+$  et  $Z^-$  en considérant le résultat exact réel  $z = XopY$  et  $op \in \{+, -, \times, /\}$  avec  $z \notin \mathbb{F}$ .

Exemple :

$$y = \sqrt{a} + b$$

$lire(A)$

$lire(B)$

$Z := sqrt(A)$

$$Y := Z + B$$

$lire(A)$  donne 2 résultats possibles  $A^+$  et  $A^-$ . De  $A^+$  pour  $lire(B)$  on obtient 2 résultats possibles  $B^+$  et  $B^-$ . De même de  $A^-$ . De  $B^+$  pour l'instruction  $Z := sqrt(A)$  on a 2 résultats possibles  $Z^+$  et  $Z^-$ ... An final on obtient  $2^4$  résultats possibles  $Y_i, i = 1, \dots, 16$ .

principe :

Générer plusieurs résultats  $Y_i$  et en déduire leurs moyenne et variance d'où leur précision.

Donc à partir de la procédure  $P$ , générer tous les résultats  $R_i \quad i = 1, \dots, 2^k = N$  représentant le même résultat mathématique  $r$ .

On considère l'ensemble de ces résultats comme une population  $\mathcal{P}(r)$  de

$$\text{moyenne } \bar{R} = \frac{1}{N} \sum_{i=1}^N R_i$$

$$\text{et de variance } \delta^2 = \frac{1}{N} \sum_{i=1}^N (R_i - \bar{R})^2$$

avec les hypothèses suivantes :

$$r \in [R_{min}, R_{max}]$$

La population  $\mathcal{P}(r)$  est gaussienne

Alors  $r \in [\bar{R} - 2\delta, \bar{R} + 2\delta]$  avec une probabilité de 95%

Si on remplace  $r$  par  $\bar{R}$  on commet une erreur absolue maximum  $2\delta$  avec une probabilité de 95% et une erreur moyenne  $\delta$  avec une probabilité de 95%, d'où une erreur relative moyenne  $\epsilon = \left| \frac{\delta}{\bar{R}} \right|$

Le nombre de chiffres significatifs moyen sur  $\bar{R}$ , noté  $C$  est donné par  $10^{-C} = \epsilon$ , c'est-à-dire  $C = \log_{10} \left| \frac{\bar{R}}{\delta} \right|$

Le nombre de chiffres significatifs minimum sur  $\bar{R}$ , noté  $C_{min}$  est donné par  $10^{-C} = 2\epsilon$ , c'est-à-dire  $C_{min} = \log_{10} \left| \frac{\bar{R}}{\delta} \right| - \log_{10} 2$

#### Remarque

$N = 2^k$  peut être très grand.

Il s'agit de générer peu de résultats  $R_i \quad i = 1, \dots, \nu \ll N$

On peut approximer une loi de Gauss par une loi de Student pour des échantillons avec  $N$  grand.

On obtient les meilleurs estimateurs de  $\bar{R}$  :  $m = \frac{1}{\nu} \sum_{i=1}^{\nu} R_i$

$$\text{et de } \delta^2 : s^2 = \frac{1}{\nu - 1} \sum_{i=1}^{\nu} (R_i - m)^2$$

Alors par la loi de Student on a :

$$Prob \left\{ \bar{R} \in \left[ m - t_{\gamma} \frac{s}{\sqrt{\nu}}, m + t_{\gamma} \frac{s}{\sqrt{\nu}} \right] \right\} = 1 - \gamma\%$$

où  $t_{\gamma}$  est la valeur dans la table de Student pour  $\nu - 1$  degrés de liberté avec le

seuil de  $\gamma\%$

Si  $\nu = 3$  avec un seuil de 5%  $t_\gamma \approx 4,303$

Si on remplace  $\overline{R}$  par  $m$  on commet une erreur absolue  $\xi = t_\gamma \frac{s}{\sqrt{\nu}}$  avec une probabilité de 95%, d'où une erreur relative  $\left| \frac{\xi}{m} \right|$

Le nombre de chiffres significatifs sur  $\overline{R}$ , noté  $C_m$  est donné par  $10^{-C} = \left| \frac{\xi}{m} \right|$ , c'est-à-dire  $C_m = \log_{10} \left| \frac{m}{s} \right| - \log_{10} \frac{t_\gamma}{\sqrt{\nu}}$  avec une probabilité de 95%

Si  $\nu = 3$   $\log_{10} \frac{t_\gamma}{\sqrt{3}} \approx 0,395$  alors  $C_m \approx \log_{10} \left| \frac{m}{s} \right| - 0,4$

Mise en œuvre :

La méthode CESTAC consiste à faire exécuter la procédure trois fois en perturbant de façon aléatoire le résultat de chaque opération arithmétique élémentaire ou les données :

En arithmétique vers 0 ( de troncature ) on ajoute de façon aléatoire au dernier bit de la mantisse la valeur 0 ou 1 avec la probabilité  $p(0) = 1/2$  et  $p(1) = 1/2$ .

En arithmétique au plus près ( d'arrondi ) on ajoute de façon aléatoire au dernier bit de la mantisse la valeur -1, 0, 1 avec les probabilités  $p(-1) = p(1) = 1/4$  et  $p(0) = 1/2$ .