

50376
1983
285
N° d'ordre : 1111

50376
1983
285

THÈSE

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNIQUES DE LILLE

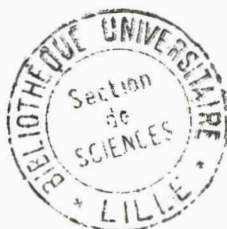
pour obtenir le titre de

DOCTEUR DE 3^{ème} CYCLE

par

Hilaire NKOUNKOU

**CONTRIBUTION A LA MESURE DE LA STABILITE
NUMERIQUE EN MOYENNE**



Soutenue le 25 Novembre 1983 devant la Commission d'Examen

MEMBRES DU JURY :

Président
Rapporteur
Examineurs

G. BREZINSKI
F. CORDELLIER
P. JACOB
J. VIGNES

P R O F E S S E U R S CLASSE EXCEPTIONNELLE

M. CONSTANT Eugène	I.E.E.A.
M. FOURET René	Physique
M. GABILLARD Robert	I.E.E.A.
M. MONTREUIL Jean	Biologie
M. PARREAU Michel	Mathématiques
M. TRIDOT Gabriel	Chimie
M. VIVIER Emile	Biologie
M. WERTHEIMER Raymond	Physique

P R O F E S S E U R S 1ère classe

M. BACCHUS Pierre	Mathématiques
M. BEAUFILS Jean-Pierre (dét.)	Chimie
M. BIAYS Pierre	G.A.S.
M. BILLARD Jean (dét.)	Physique
M. BOILLY Bénoni	Biologie
M. BOIS Pierre	Mathématiques
M. BONNELLE Jean-Pierre	Chimie
M. BOUGHON Pierre	Mathématiques
M. BOURIQUET Robert	Biologie
M. BREZINSKI Claude	I.E.E.A.
M. CELET Paul	Sciences de la Terre
M. CHAMLEY Hervé	Biologie
M. COEURE Gérard	Mathématiques
M. CORDONNIER Vincent	I.E.E.A.
M. DEBOURSE Jean-Pierre	S.E.S.
M. DYMENT Arthur	Mathématiques

PROFESSEURS 1ère classe (suite)

M. ESCAIG Bertrand	Physique
M. FAURE Robert	Mathématiques
M. FOCT Jacques	Chimie
M. GRANELLE Jean-Jacques	S.E.S.
M. GRUSON Laurent	Mathématiques
M. GUILLAUME Jean	Biologie
M. HECTOR Joseph	Mathématiques
M. LABLACHE COMBIER Alain	Chimie
M. LACOSTE Louis	Biologie
M. LAVEINE Jean Pierre	Sciences de la Terre
M. LEHMANN Daniel	Mathématiques
Mme LENOBLE Jacqueline	Physique
M. LHOMME Jean	Chimie
M. LOMBARD Jacques	S.E.S.
M. LOUCHEUX Claude	Chimie
M. LUCQUIN Michel	Chimie
M. MIGEON Michel Recteur à Grenoble	E.U.D.I.L.
M. MIGNOT Fulbert (dét.)	Mathématiques
M. PAQUET Jacques	Sciences de la Terre
M. PROUVOST Jean	Sciences de la Terre
M. ROUSSEAU Jean-Paul	Biologie
M. SALMER Georges	I.E.E.A.
M. SEGUIER Guy	I.E.E.A.
M. SIMON Michel	S.E.S.
M. STANKIEWICZ François	S.E.S.
M. TILLIEU Jacques	Physique
M. VIDAL Pierre	I.E.E.A.
M. ZEYTOUNIAN Radyadour	Mathématiques

P R O F E S S E U R S 2ème classe

M. ANTOINE Philippe	Mathématiques (Calais)
M. BART André	Biologie
Mme BATTIAU Yvonne	Géographie
M. BEGUIN Paul	Mathématiques
M. BELLET Jean	Physique
M. BERZIN Robert	Mathématiques
M. BKOCHE Rudolphe	Mathématiques
M. BODARD Marcel	Biologie
M. BOSQ Denis	Mathématiques
M. BRASSELET Jean-Paul	Mathématiques
M. BRUYELLE Pierre	Géographie
M. CAPURON Alfred	Biologie
M. CARREZ Christian	I.E.E.A.
M. CAYATTE Jean-Louis	S.E.S.
M. CHAPOTON Alain	C.U.E.E.P.
M. COQUERY Jean-Marie	Biologie
Mme CORSIN Paule	Sciences de la Terre
M. CORTOIS Jean	Physique
M. COUTURIER Daniel	Chimie
M. CROSNIER Yves	I.E.E.A.
M. CURGY Jean-Jacques	Biologie
Mlle DACHARRY Monique	Géographie
M. DAUCHET Max	I.E.E.A.
M. DEBRABANT Pierre	E.U.D.I.L.
M. DEGAUQUE Pierre	I.E.E.A.
M. DELORME Pierre	Biologie
M. DELORME Robert	S.E.S.
M. DE MASSON D'AUTUME Antoine	S.E.S.
M. DEMUNTER Paul	C.U.E.E.P.

PROFESSEURS 2ème classe (Suite 1)

M. DENEL Jacques	I.E.E.A.
M. DE PARIS Jean-Claude	Mathématiques (Calais)
Mlle DESSAUX Odile	Chimie
M. DEVRAINNE Pierre	Chimie
M. DHAINAUT André	Biologie
Mme DHAINAUT Nicole	Biologie
M. DORMARD Serge	S.E.S.
M. DOUKHAN Jean-Claude	E.U.D.I.L.
M. DUBOIS Henri	Physique
M. DUBRULLE Alain	Physique (Calais)
M. DUBUS Jean-Paul	I.E.E.A.
M. FAKIR Sabah	Mathématiques
M. FONTAINE Hubert	Physique
M. FOUQUART Yves	Physique
M. FRONTIER Serge	Biologie
M. GAMBLIN André	G.A.S.
M. GLORIEUX Pierre	Physique
M. GOBLOT Rémi	Mathématiques
M. GOSSELIN Gabriel (dét.)	S.E.S.
M. GOUDMAND Pierre	Chimie
M. GREGORY Pierre	I.P.A.
M. GREMY Jean-Paul	S.E.S.
M. GREVET Patrice	S.E.S.
M. GUILBAULT Pierre	Biologie
M. HENRY Jean-Pierre	E.U.D.I.L.
M. HERMAN Maurice	Physique
M. JACOB Gérard	I.E.E.A.
M. JACOB Pierre	Mathématiques
M. JEAN Raymond	Biologie
M. JOFFRE Patrick	I.P.A.

PROFESSEURS 2ème classe (suite 2)

M. JOURNAL Gérard	E.U.D.I.L.
M. KREMBEL Jean	Biologie
M. LANGRAND Claude	Mathématiques
M. LATTEUX Michel	I.E.E.A.
Mme LECLERCQ Ginette	Chimie
M. LEFEVRE Christian	Sciences de la Terre
Mle LEGRAND Denise	Mathématiques
Mle LEGRAND Solange	Mathématiques (Calais)
Mme LEHMANN Josiane	Mathématiques
M. LEMAIRE Jean	Physique
M. LHENAFF René	Géographie
M. LOCQUENEUX Robert	Physique
M. LOSFELD Joseph	C.U.E.E.P.
M. LOUAGE Francis(dét.)	E.U.D.I.L.
M. MACKE Bruno	Physique
M. MAIZIERES Christian	I.E.E.A.
M. MESSELYN Jean	Physique
M. MESSERLIN Patrick	S.E.S.
M. MONTEL Marc	Physique
Mme MOUNIER Yvonne	Biologie
M. PARSY Fernand	Mathématiques
Mle PAUPARDIN Colette	Biologie
M. PERROT Pierre	Chimie
M. PERTUZON Emile	Biologie
M. PONSOLLE Louis	Chimie
M. PORCHET Maurice	Biologie
M. POVY Lucien	E.U.D.I.L.
M. RACZY Ladislas	I.E.E.A.
M. RAOULT Jean François	Sciences de la Terre
M. RICHARD Alain	Biologie

PROFESSEURS 2ème Classe (suite 3)

M. RIETSCH François	E.U.D.I.L.
M. ROBINET Jean-Claude	E.U.D.I.L.
M. ROGALSKI Marc	Mathématiques
M. ROY Jean-Claude	Biologie
M. SCHAMPS Joël	Physique
Mme SCHWARZBACH Yvette	Mathématiques
M. SLIWA Henri	Chimie
M. SOMME Jean	G.A.S.
Mlle SPIK Geneviève	Biologie
M. STAROSWIECKI Marcel	E.U.D.I.L.
M. STERBOUL François	E.U.D.I.L.
M. TAILLIEZ Roger	Institut Agricole
Mme TJOTTA Jacqueline (dét.)	Mathématiques
M. TOULOTTE Jean-Marc	I.E.E.A.
M. TURRELL Georges	Chimie
M. VANDORPE Bernard	E.U.D.I.L.
M. VAST Pierre	Chimie
M. VERBERT André	Biologie
M. VERNET Philippe	Biologie
M. WALLART Francis	Chimie
M. WARTEL Michel	Chimie
M. WATERLOT Michel	Sciences de la Terre
Mme ZINN JUSTIN Nicole	Mathématiques

CHARGES DE COURS

M. ADAM Michel

S.E.S.

CHARGES DE CONFERENCES

M. BAFCOP Joël

I.P.A.

M. DUVEAU Jacques

S.E.S.

M. HOF Lack Jean

I.P.A.

M. LATOUCHE Serge

S.E.S.

M. MALAUSSENA DE PERNO Jean-Louis

S.E.S.

M. NAVARRE Christian

I.P.A.

M. OPIGEZ Philippe

S.E.S.

Je tiens à remercier Monsieur C. BREZINSKI, qui a bien voulu me faire l'honneur de présider le jury de cette thèse.

Je suis très reconnaissant envers Monsieur F. CORDELLIER, qui m'a proposé ce travail, d'en avoir constamment suivi l'évolution ; je suis très heureux de pouvoir lui exprimer ici ma profonde gratitude pour ses nombreux conseils, qui m'ont beaucoup aidé.

Je remercie Monsieur J. VIGNES qui a bien voulu s'intéresser à ce travail.

Je remercie également Monsieur P. JACOB qui a bien voulu participer à ce jury.

Mes remerciements vont aussi à Madame CARON et à Madame DEBOCK qui ont assuré avec gentillesse, compétence et rapidité la réalisation matérielle de cette thèse.

A mes parents,

A mon épouse et aux jumeaux.

TABLE DES MATIÈRES

Introduction	page 1
Chapitre I : Sur le conditionnement en moyenne	5
Introduction	6
I.1 - Notations	8
I.2 - Proposition et Résultat principal	8
Chapitre II : Mesure de la stabilité en borne	16
II.1 - Introduction	17
II.2 - Notations et définitions	18
II.3 - Modèle arithmétique	20
II.4 - Erreur de calcul	22
II.5 - Propositions	24
Chapitre III : Mesure de la stabilité en moyenne	31
Introduction	32
III.1 - Proposition 1	33
III.2 - Proposition 2	33
III.3 - Nombre de stabilité en moyenne	34
Chapitre IV : Calcul d'une valeur moyenne et d'un écart type de l'erreur relative de représentation	38
Introduction	39
IV.1 - Rappels et notations	41
IV.2 - Définition de l'arrondi d'un nombre	41
IV.3 - Erreur relative de représentation	42
IV.4 - Proposition 1	45
IV.5 - Proposition 2	47
Chapitre V : Applications et résultats numériques	51
Introduction	52
V.1 - Applications	54
1.1 - L'effet de cumul dans une somme	54
1.2 - Algorithme de LINZ	56
1.3 - Expressions du Δ^2 d'Aitken	59
1.4 - Calcul de la valeur d'un polynôme par l'algorithme de Horner	62
1.5 - Calcul de la valeur d'un polynôme par le schéma classique	64

V.2 - Modélisation de la simulation	67
V.3 - Analyse des résultats	72
3.1 - Calcul d'une somme	72
3.2 - Calcul de la valeur d'un polynôme	78
3.3 - Calcul de la valeur d'un polynôme	84
3.4 - Calcul de l'expression du Δ^2 d'Aitken	86
Conclusion	96
Annexe : ensemble des programmes	98
Références	118

INTRODUCTION

L'étude de l'erreur de calcul revêt une importance majeure en calcul numérique. Dans cette question deux points de vue se complètent :

- une analyse permet d'avoir des informations sur les erreurs
- on en déduit des techniques qui permettent d'exploiter ces informations pour diminuer les erreurs.

En réalité, la collecte d'informations peut être automatique (soit par l'utilisation de types arithmétiques adéquats tels que les arithmétiques d'intervalles, les arithmétiques non normalisées [16], ou l'usage d'un index de signification, ce qui permet d'avoir une majoration de l'erreur numérique, soit par l'introduction de méthodes de simulation telle que, mode bruité (Sterbenz [30]), méthode de permutation - perturbation (Laporte et Vignes [19]), ou une technique de perturbation des données, ce qui permet d'avoir des informations réalistes sur l'ordre de grandeur de l'erreur.

L'étude de l'erreur de calcul peut aussi reposer sur une analyse formelle du schéma de calcul, analyse qui peut revêtir des formes diverses en fonction de son but ou de sa méthode.

Comme méthode d'analyse on distingue :

- l'analyse directe qui lie l'imprécision du résultat à la précision avec laquelle les calculs sont effectués (Korganoff [20], Esko Ukkonen [17], Henrici [12], Larson [18], [21], Linnainmaa [13])
- l'analyse rétrograde de la progression des erreurs dans un algorithme de calcul est une méthode qui lie la stabilité numérique d'un schéma de calcul au nombre de conditionnement du problème qu'il doit résoudre (Givens [7], Wilkinson [10], Miller [8], Larson [18], Bauer [15]).

Le but de l'analyse peut être d'obtenir une majoration effective de l'erreur. Mais une telle majoration reste généralement très grossière, aussi est-il préférable d'avoir une valeur moyenne de cette erreur, en utilisant une approche probabiliste.

A la suite de Korganoff [20], Henrici [12], l'approche probabiliste a été souvent abordée de façon superficielle (Linnainmaa [13], [14], Esko Ukkonen [17]). Ces auteurs se sont uniquement intéressés au cas des algorithmes fondamentalement linéaires sans pour autant donner une véritable justification à l'introduction de l'approche probabiliste qu'ils font. C'est avec Laporte et Vignes [19], dans leur méthode de perturbation, que l'on commence à voir une justification de l'approche probabiliste.

Le travail que nous proposons dans ce mémoire est une contribution à l'approche probabiliste des erreurs numériques. En justifiant l'introduction de cette approche, nous définissons un nombre de stabilité en moyenne qui mesure la précision avec laquelle un résultat sera connu lorsqu'on conduit les calculs d'une certaine façon. Ce nombre donne à l'approche probabiliste un outil de mesure, et sa connaissance vient compléter celle qui était fournie par le nombre de stabilité en borne [11] qui se révélait un peu trop pessimiste.

Après avoir donné un résultat général sur le conditionnement en moyenne dans le chapitre I, nous introduisons la définition d'un type arithmétique qui donne des résultats aléatoires. Dans ce contexte, l'erreur relative sur le résultat est considérée comme une variable aléatoire fonction de diverses variables aléatoires de moyenne nulle et d'écart type donné qui représentent l'erreur relative de représentation de chaque résultat

intermédiaire. Nous définissons à partir des écarts type de l'erreur relative sur le résultat et de l'erreur relative de représentation, un nombre non attaché à une arithmétique particulière, le nombre de stabilité en moyenne. C'est ce qui fait l'objet des chapitres II et III.

La détermination de l'écart type de l'erreur relative de représentation dans les arithmétiques flottantes de longueur de mantisse déterminée dans une base quelconque est faite dans le chapitre IV en respectant strictement l'hypothèse de Hamming [26] sur la répartition statistique des mantisses des nombres. Le résultat obtenu est identique à celui qu'avait établi R. Alt [28] avec une hypothèse simplifiée.

Il est clair que ces résultats ne sont valables que dans le contexte précis que nous avons défini à savoir une arithmétique fictive donnant des résultats aléatoires selon une loi de probabilité donnée. L'objet du dernier chapitre est d'utiliser une technique de simulation pour vérifier la validité des informations apportées par la connaissance du nombre de stabilité en moyenne.

CHAPITRE I

SUR LE CONDITIONNEMENT EN MOYENNE

INTRODUCTION

Le problème du conditionnement, abordé par plusieurs auteurs [8], [9], peut être présenté de la manière suivante :

Si on considère un problème dépendant des données $(X_1, X_2, \dots, X_n)$ et ayant une solution caractérisée par les résultats R_j , $j = 1, \dots, m$, les nombres R_j sont des fonctions des données X_i , $i = 1, \dots, n$ que nous noterons $R_j = f_j(X_1, \dots, X_n)$, $j = 1, \dots, m$.

Dans le cas où les fonctions f_j sont différentiables, si on note

$R'_j = f_j(X'_1, \dots, X'_n)$, on peut écrire

$$R'_j - R_j = \sum_{i=1}^n \frac{\partial f_j}{\partial X_i} (\lambda X + (1-\lambda)X') \times (X_i - X'_i)$$

$$\lambda \in [0, 1]$$

où $X' = (X'_1, \dots, X'_n)$ est voisin de $X = (X_1, \dots, X_n)$.

La valeur des dérivées partielles $\frac{\partial f_j}{\partial X_i}$ au voisinage de $(X_1, \dots, X_n)$ nous donne alors des informations sur le conditionnement du problème. L'analyse retrograde introduite par Givens [7] et utilisée par la suite par Wilkinson [10] dans le cas des problèmes d'algèbre linéaire consiste à se servir du conditionnement pour analyser l'erreur de calcul. C'est dans ce cadre que nous donnons dans ce chapitre le résultat suivant sur le conditionnement en moyenne.

Soit un problème dépendant des données $(X_1, \dots, X_n)$ qui sont des variables aléatoires indépendantes. Si la solution du problème est caractérisée par la fonction $f(X_1, \dots, X_n)$, qui est une variable aléatoire, $f \in C^2(U)$ où $U \subset \mathbb{R}^n$ est un fermé. Si on suppose que

$$P(|X_i - E(X_i)| > \tau |E(X_i)|) = 0, i = 1, \dots, n$$

où P = Probabilité.

Alors on a le résultat suivant sur le conditionnement en moyenne.

$$\sigma^2(f(X_1, \dots, X_n)) - \sum_{i=1}^n \left[\frac{\partial f}{\partial X_i} (E(X_1), \dots, E(X_n)) \right]^2 \sigma^2(X_i) = O(\tau^3)$$

La valeur des dérivées partielles nous informe sur le conditionnement du problème.

I.1 - NOTATIONS

$X = (X_1, \dots, X_n)$ est un vecteur aléatoire de dimension n , défini sur un espace probabilisé (Ω, A, P) . On note, pour tout $i = 1, \dots, n$, μ_i la loi de probabilité de X_i , c'est à dire :

$$\forall B, \text{ borelien de } \mathbb{R}, \mu_i(B) = P\{X_i \in B\}$$

On suppose que les X_i , $i = 1, \dots, n$, sont des variables aléatoires indépendantes, dans ce cas, la loi de X est la mesure produit $\mu = \mu_1 \otimes \dots \otimes \mu_n$

$$\forall B_1, \dots, B_n \text{ boreliens de } \mathbb{R}^n$$

$$\mu(B_1 \times B_2 \times \dots \times B_n) = \mu_1(B_1) \times \dots \times \mu_n(B_n)$$

faisons l'hypothèse suivante :

$$(H_\tau) \left\{ \begin{array}{l} X = (X_1, \dots, X_n) \text{ variable aléatoire de loi} \\ \mu = \mu_1 \otimes \dots \otimes \mu_n, X_i, \text{ indépendants } i = 1, \dots, n \\ \tau > 0 \\ P\{|X_i - E(X_i)| > \tau | E(X_i)|\} = 0, i = 1, \dots, n \end{array} \right.$$

On posera $E(X_i) = a_i$, $i = 1, \dots, n$

$$E(X) = (E(X_1), \dots, E(X_n)) = (a_1, \dots, a_n) = a$$

$U \subset \mathbb{R}^n$, un fermé.

Enonçons maintenant le résultat principal de ce chapitre sous la forme de la proposition suivante :

I.2 - PROPOSITION

Soit $f \in C^2(U)$ une fonction fixée, $\tau > 0$, X vérifie l'hypothèse H_τ , alors on a :

$$1 - \left| \frac{E(f(X)) - f(E(X))}{f(E(X))} \right| \leq M_1 \tau^2$$

avec $M_1 = M_f \frac{||E(X)||_1^2}{|f(E(X))|}$, où $M_f = \max_{s \in U} \left| \frac{\partial^2 f(s)}{\partial x_i \partial x_j} \right|$
 $i, j \in \{1, \dots, n\}$

$$2 - |\sigma^2(f(X)) - \sum_{i=1}^n \left[\frac{\partial f(E(X))}{\partial x_i} \right]^2 \sigma^2(X_i)| \leq B_{11} \tau^3$$

avec

$$B_{11} = M_f^2 \tau ||E(X)||_2^4 + 2M_f M' ||E(X)||_1^3 + 3M_f^2 \tau ||E(X)||_1^4$$

où

$$M' = \max_{\substack{s \in U \\ i=1, \dots, n}} \frac{\partial f(s)}{\partial x_i}, \quad ||E(X)||_1 = \sum_{i=1}^n |E(X_i)|$$

$$||E(X)||_2 = \left(\sum_{i=1}^n |E(X_i)|^2 \right)^{\frac{1}{2}}$$

□

Avant de passer à la démonstration de la proposition, établissons d'abord les lemmes suivants :

Lemme 1

X vérifiant l'hypothèse H_τ , alors on a :

$$1 - \sigma^2(X_i) = \int_{\mathbb{R}} (x_i - a_i)^2 d\mu_i(x_i) \leq \tau^2 |E(X_i)|^2$$

$$2 - \int_{\mathbb{R}^n} (x_i - a_i)^4 d\mu(x_1, \dots, x_n) \leq \tau^4 |E(X_i)|^4$$

$i = 1, \dots, n$

□

Démonstration

D'après l'hypothèse H_τ : on a

$$(x_i - a_i)^2 \leq \tau^2 a_i^2, \quad \mu_i \text{ presque partout}$$

d'où le résultat en intégrant

$$\int_{\mathbb{R}} (x_i - a_i)^2 d\mu_i(x_i) \leq \tau^2 a_i^2$$

$$\int_{\mathbb{R}} (x_i - a_i)^4 d\mu_i(x_i) \leq \tau^4 a_i^4 \quad i = 1, \dots, n$$

Lemme 2

Soient X qui vérifie H_τ , $U \subset \mathbb{R}^n$, fermé, $f \in C^2(U)$ une fonction fixée, soit

$$T : \mathbb{R}^n \times U \rightarrow \mathbb{R}$$

$$(x_1, \dots, x_n ; s_1, \dots, s_n) \rightarrow T(x_1, \dots, x_n ; s_1, \dots, s_n)$$

avec

$$T(x_1, \dots, x_n ; s_1, \dots, s_n) = \begin{cases} \sum_{i,j=1}^n \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s_1, \dots, s_n)}{\partial x_i \partial x_j} & , \text{ si } a_i \neq x_i \\ 0 & \text{ sinon } , i = 1, \dots, n \end{cases}$$

$$\text{avec } x_i = a_i + h_i \quad s = (s_1, \dots, s_n)$$

$$s_i = a_i + th_i, \quad t \in [0, 1] \quad x = (x_1, \dots, x_n)$$

Alors la fonction T est mesurable et l'intégrale

$$\sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s_1, \dots, s_n)}{\partial x_i \partial x_j} ds_i d\mu(x_1, \dots, x_n)$$

existe et on a

$$\sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \left| \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} \right| ds_i d\mu(x) \leq M_f \tau^2 \|E(X)\|_1^2$$

□

Démonstration

pour $x_i \neq a_i$, $i = 1, \dots, n$

$$(x, s) \rightarrow \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} \text{ est mesurable}$$

=> la fonction T est mesurable, d'autre part

$$\left| \sum_{i,j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i d\mu(x) \right|$$

$$\leq M_f \sum_{i,j=1}^n \int_{\mathbb{R}^n} |x_i - a_i| |x_j - a_j| d\mu(x) = M_f \sum_{i,j=1}^n \int_{\mathbb{R}} |x_i - a_i| d\mu_i(x_i) \int_{\mathbb{R}} |x_j - a_j| d\mu_j(x_j)$$

d'après le théorème de Fubini

donc on a

$$|T(x, s)| \leq M_f \sum_{i=1}^n \sum_{j=1}^n \tau a_i \cdot \tau a_j = M_f \tau^2 \|E(X)\|_1^2$$

Après l'introduction des lemmes 1 et 2 qui vont nous servir par la suite pour faire la preuve de la proposition énoncée auparavant, passons maintenant à la démonstration de la proposition.

1 - La formule de Taylor avec reste intégral est pour une fonction f donnée :

$$\begin{aligned} f(a_1 + h_1, \dots, a_n + h_n) &= f(a_1, \dots, a_n) + \sum_{i=1}^n h_i \frac{\partial f}{\partial a_i}(a_1, \dots, a_n) \\ &\quad + \sum_{i,j=1}^n \int_0^1 (1-t) h_i h_j \frac{\partial^2 f}{\partial a_i \partial a_j}(a_1 + th_1, \dots, a_n + th_n) dt \end{aligned}$$

$$t \in [0, 1]$$

Si $h_i = 0$, $i = 1, \dots, n$, il n'y a pas de développement de Taylor. Par la suite, on supposera toujours $h_i \neq 0$, pour $i = 1, \dots, n$.

Si on fait le changement de variable suivant :

$$x_i = a_i + h_i$$

$$s_i = a_i + th_i, \quad i = 1, \dots, n, \quad h_i \neq 0 \quad \text{avec}$$

$$x = (x_1, \dots, x_n), \quad s = (s_1, \dots, s_n), \quad a = (a_1, \dots, a_n)$$

$$h = (h_1, \dots, h_n) \quad s \in U \subset \mathbb{R}^n \text{ fermé.}$$

on a alors la formule

$$\begin{aligned} (1)^* \quad f(x_1, \dots, x_n) &= f(a_1, \dots, a_n) + \sum_{i=1}^n (x_i - a_i) \frac{\partial f}{\partial x_i}(a_1, \dots, a_n) \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f}{\partial x_i \partial x_j}(s_1, \dots, s_n) ds_i \end{aligned}$$

en intégrant les deux membres de cette égalité (1), on a :

$$\begin{aligned} \int_{\mathbb{R}^n} f(x) d\mu(x) &= \int_{\mathbb{R}^n} f(a) d\mu(x) + \sum_{i=1}^n \int_{\mathbb{R}^n} (x_i - a_i) \frac{\partial f}{\partial x_i}(a) d\mu(x) \\ &+ \sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i d\mu(x) \end{aligned}$$

comme $\int_{\mathbb{R}^n} (x_i - a_i) d\mu(x) = 0$ car $a_i = E(X_i)$, $i = 1, \dots, n$

on a :

$$(2)^* \quad E(f(X)) = f(E(X)) + \sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i d\mu(x)$$

donc

$$\left| \frac{E(f(X)) - f(E(X))}{f(E(X))} \right| \leq \frac{1}{|f(E(X))|} \sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \left| \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} \right| ds_i d\mu(x)$$

et d'après les lemmes 2 et 1, on a alors le résultat cherché :

$$\left| \frac{E(f(X)) - f(E(X))}{f(E(X))} \right| \leq M_f \frac{\|E(X)\|_1^2}{|f(E(X))|} \tau^2 = M_1 \tau^2$$

2 - Dans la première partie de cette démonstration, on a établi les formules

(1)* et (2)*. Maintenant si on fait la différence (1)* - (2)*, on obtient :

$$\begin{aligned} f(x) - E(f(X)) &= \sum_{i=1}^n (x_i - a_i) \frac{\partial f(a)}{\partial x_i} + \sum_{i,j=1}^n \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i \\ &- \sum_{i,j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i d\mu(x) \end{aligned}$$

et si on élève chaque membre de cette égalité au carré, on obtient :

$$\begin{aligned} [f(x) - E(f(X))]^2 &= \left[\sum_{i=1}^n (x_i - a_i) \frac{\partial f(a)}{\partial x_i} \right]^2 \quad (1) \\ &+ \left[\sum_{i,j=1}^n \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds_i \right]^2 \quad (2) \end{aligned}$$

$$+2 \left[\sum_{k=1}^n \sum_{i,j=1}^n (x_k^{-a_k}) \frac{\partial f(a)}{\partial x_k} \int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds \right] \quad (3)$$

$$-2 \left[\sum_{k=1}^n (x_k^{-a_k}) \frac{\partial f(a)}{\partial x_k} \sum_{i,j=1}^n \int_{\mathbb{R}^n} \left(\int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds \right) d\mu(x) \right] \quad (4)$$

$$+ \left[\sum_{i,j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds d\mu(x) \right]^2 \quad (5)$$

$$-2 \left[\sum_{i,j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds d\mu(x) \right] \left[\sum_{i,j=1}^n \int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} ds \right] \quad (6)$$

et lorsque l'on intègre sur $\mathbb{R}^n$ les deux membres de cette égalité, on a

pour le membre de gauche :

$$\sigma^2(f(X)) = \int_{\mathbb{R}^n} (f(x) - E(f(X)))^2 d\mu(x)$$

pour le second membre, c'est à dire celui de la droite, on a pour le terme (1),

comme

$$\begin{aligned} \left[\sum_{i=1}^n (x_i^{-a_i}) \frac{\partial f(a)}{\partial x_i} \right]^2 &= \sum_{i=1}^n (x_i^{-a_i})^2 \left[\frac{\partial f(a)}{\partial x_i} \right]^2 \\ &+ \sum_{\substack{i,j=1 \\ i \neq j}}^n (x_i^{-a_i})(x_j^{-a_j}) \frac{\partial f(a)}{\partial x_i} \frac{\partial f(a)}{\partial x_j} \end{aligned}$$

et comme les X_i , $i = 1, \dots, n$ sont des variables aléatoires indépendantes,

on a

$$\int_{\mathbb{R}^n} (x_i^{-a_i})(x_j^{-a_j}) d\mu(x) = 0,$$

donc le terme (1) devient :

$$\int_{\mathbb{R}^n} \left[\sum_{i=1}^n (x_i^{-a_i}) \frac{\partial f(a)}{\partial x_i} \right]^2 d\mu(x) = \sum_{i=1}^n \int_{\mathbb{R}^n} (x_i^{-a_i})^2 \left[\frac{\partial f(a)}{\partial x_i} \right]^2 d\mu(x)$$

$$\int_{\mathbb{R}^n} (1) d\mu(x) = \sum_{i=1}^n \left[\frac{\partial f(a)}{\partial x_i} \right]^2 \sigma^2(X_i)$$

pour le terme (2), on a

$$(2) \leq \left[\sum_{i,j=1}^n M_f \int_{a_i}^{x_i} \frac{(x_i^{-s_i})(x_j^{-a_j})}{(x_i^{-a_i})} ds \right]^2 \leq M_f^2 \left[\sum_{i,j=1}^n (x_i^{-a_i})(x_j^{-a_j}) \right]^2$$

Lorsque l'on intègre sur $\mathbb{R}^n$, comme les variables aléatoires X_i , $i = 1, \dots, n$ sont indépendantes, cela donne :

$$\int_{\mathbb{R}^n} (2) \, d\mu(x) \leq M_f^2 \sum_{i,j=1}^n \int_{\mathbb{R}^n} (x_i - a_i)^2 (x_j - a_j)^2 \, d\mu(x)$$

et en appliquant l'inégalité de Schwarz au second terme de l'inégalité précédente, on a :

$$\int_{\mathbb{R}^n} (2) \, d\mu(x) \leq M_f^2 \sum_{i=1}^n \left(\int_{\mathbb{R}^n} (x_i - a_i)^4 \, d\mu(x) \right)^{\frac{1}{2}} \times \sum_{j=1}^n \left(\int_{\mathbb{R}^n} (x_j - a_j)^4 \, d\mu(x) \right)^{\frac{1}{2}}$$

et d'après le lemme 1, on a alors

$$\int_{\mathbb{R}^n} (2) \, d\mu(x) \leq M_f^2 \tau^4 \|E(X)\|_2^4$$

L'intégrale sur $\mathbb{R}^n$ du terme (4) est nulle. Pour le terme (5) on a d'après le lemme 2 :

$$\sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}^n} \int_{a_i}^{x_i} \frac{(x_i - s_i)(x_j - a_j)}{(x_i - a_i)} \frac{\partial^2 f(s)}{\partial x_i \partial x_j} \, ds_i \, d\mu(x) \leq M_f \tau^2 \|E(X)\|_1^2$$

$$\text{et } \int_{\mathbb{R}^n} (5) \, d\mu(x) \leq M_f^2 \tau^4 \|E(X)\|_1^4$$

et de même, on a

$$\int_{\mathbb{R}^n} (6) \, d\mu(x) \leq 2 M_f^2 \tau^4 \|E(X)\|_1^4$$

d'autre part l'intégrale de (3) existe et est majorée de la façon suivante :

$$\int_{\mathbb{R}^n} (3) \, d\mu(x) \leq 2 M_f M' \sum_{i,j=1}^n \sum_{k=1}^n \int_{\mathbb{R}^n} |x_i - a_i| |x_j - a_j| |x_k - a_k| \, d\mu(x)$$

et l'inégalité de Schwarz donne par une double application de suite :

$$\int_{\mathbb{R}^n} |x_i - a_i| |x_j - a_j| |x_k - a_k| \, d\mu(x) \leq \left(\int_{\mathbb{R}^n} (x_i - a_i)^2 (x_j - a_j)^2 \, d\mu(x) \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} (x_k - a_k)^2 \, d\mu(x) \right)^{\frac{1}{2}}$$

et d'après le lemme 1, on a finalement

$$\int_{\mathbb{R}^n} (3) \, d\mu(x) \leq 2 M_f M' \tau^3 \|E(X)\|_1^3$$

d'où le résultat final

$$\begin{aligned}
 \left| \int_{\mathbb{R}^n} (f(x) - E(f(X)))^2 d\mu(x) - \sum_{i=1}^n \left[\frac{\partial f(a)}{\partial x_i} \right]^2 \sigma^2(X_i) \right| &\leq M_f^2 \tau^4 \|E(X)\|_2^4 \\
 &\quad + 3M_f^2 \tau^4 \|E(X)\|_1^4 \\
 &\quad + 2M_f M' \tau^3 \|E(X)\|_1^3
 \end{aligned}$$

$$\left| \sigma^2(f(X)) - \sum_{i=1}^n \left[\frac{\partial f(E(X))}{\partial x_i} \right]^2 \sigma^2(X_i) \right| \leq B_{11} \tau^3$$

avec

$$B_{11} = M_f^2 \tau \|E(X)\|_2^4 + 2M_f M' \|E(X)\|_1^3 + 3M_f^2 \tau \|E(X)\|_1^4$$

CHAPITRE II

RAPPELS SUR LA MESURE DE LA STABILITE
EN BORNE D'UN SCHEMA DE CALCUL SANS TEST

INTRODUCTION

Dans l'analyse directe de la progression des erreurs, diverses définitions en termes de majoration de l'erreur ont été proposées : Miller[22], Larson[18], Cordellier [11] etc ... Ce chapitre est essentiellement consacré aux rappels des définitions et propositions qui aboutissent à la définition en terme de majoration de l'erreur proposée dans [11]. Nous fournissons aux propositions 1 et 3 de [11] des démonstrations détaillées qui viennent compléter les démonstrations abrégées qui figurent dans [11]. Les majorations en borne étant trop pessimistes (Wilkinson [10]), une approche probabiliste (Larson et Samah [21], [18], Esko Ukkonen [17]) permet de compléter cette analyse de majoration systématique. Nous introduisons dans ce chapitre la définition d'un type arithmétique donnant des résultats aléatoires, ceci pour mieux justifier l'approche probabiliste que nous allons faire.

II.1 - NOTATIONS ET DEFINITIONS [11]

V : Ensemble fini dont les éléments sont appelés variables

$V = R \cup D$; avec

R : Ensemble des résultats

D : Ensemble des données

O : Ensemble fini dont les éléments sont appelés opérateurs

$\Pi_\lambda = \{\alpha_\lambda^{(i)} \mid i = 1, \dots, n_\lambda\}$ est l'ensemble des opérandes de l'opérateur $\alpha_\lambda^{(o)}$

$A_\lambda = \{\mu \in V \mid \mu \rightarrow \lambda\}$ l'ensemble des antécédents stricts de λ

$\bar{A}_\lambda = A_\lambda \cup \{\lambda\}$

Définition 1

Soit S une application qui, à tout $\lambda \in R \subset V$, associe :

$$S_\lambda = \begin{cases} \alpha_\lambda^{(o)} \in O, \text{ opérateur agissant sur } n_\lambda \text{ opérandes} \\ \alpha_\lambda^{(i)} \in V, i = 1, \dots, n, \text{ les opérandes sur lesquels agit } \alpha_\lambda^{(o)} \end{cases}$$

Une telle application est appelée schéma de calcul sans test si elle

induit sur V une relation d'ordre strict notée $\rightarrow$ générée par :

$$\mu = \alpha_\lambda^{(i)} \Rightarrow \mu \rightarrow \lambda, \forall i = 1, \dots, n_\lambda, \forall \lambda \in R.$$

Définition 2

On appellera type arithmétique le couple constitué par :

- un ensemble T dont les éléments sont appelés valeurs
- un ensemble fini d'opérations internes de T noté O_T , ces opérations

pouvant être unaires, binaires, ...

Ce couple sera noté $T_T = (T, O_T)$.

Définition 3

Soit S un schéma de calcul défini sur l'ensemble des variables $V = R \cup D$, et soit T_T un type arithmétique. On appelle mise en oeuvre de S sur le type T_T toute application ϕ de V dans T qui satisfait

$$\phi_\lambda = \alpha_\lambda^{(0)} (\phi_{\alpha_\lambda}(i) \mid i = 1, \dots, n_\lambda), \forall \lambda \in R$$

La restriction de cette application à D sera appelée initialisation du schéma de calcul. On notera

$$\phi_D = \{\phi_\mu \mid \mu \in D\}$$

Le type arithmétique $T_{\mathbb{R}}$ constitué par :

- $\mathbb{R}$ l'ensemble des réels
- $\mathcal{O}_{\mathbb{R}} = \{+, -, \times, /\}$

caractérise le type qui donne des résultats dans $\mathbb{R}$, dans la mise en oeuvre du schéma de calcul S .

Exemple 1

Mais pour éviter les ennuis créés par la présence des résultats indéterminés ou impossibles (division par zéro dans $\mathbb{R}$ ou produit $0 \times \infty$ dans $\bar{\mathbb{R}}$ par exemple), nous nous en tiendrons au type arithmétique suivant :

Le type $T_{\mathbb{R}}^\sim$ sera défini par :

- $\mathbb{R}^\sim = (\mathbb{R} \setminus \{0\}) \cup \{\beta\}$, où β est un élément particulier appelé indéfini.
- $\mathcal{O}_{\mathbb{R}} = \{+, -, \times, /\}$

avec les deux conventions suivantes :

- tout résultat nul est représenté par β
- toute opération dont β est opérande a β pour résultat.

Remarque 1

La limitation de $\mathcal{O}_R$ aux quatre opérations arithmétiques confine les schémas de calcul à l'évaluation des fractions rationnelles. On peut sans dommage y adjoindre des opérations unaires comme l'inversion, la racine carrée, la valeur absolue ou l'exponentielle.

II.2 - MODELE ARITHMETIQUE

Dans la définition des modèles arithmétiques, nous allons envisager deux cas.

1 - Cas discret [11]

Soit $T_F^\sim = (\tilde{F}, \mathcal{O}_F^\sim)$ un type arithmétique avec $\tilde{F} = F \cup \{\beta\}$ où $F \subset \mathbb{R} \setminus \{0\}$, et soit $P_G^\sim$ une projection de $\tilde{F}$ sur $\tilde{G}$ où $\tilde{G} = G \cup \{\beta\}$ avec $G \subset F$. Définissons alors l'ensemble d'opérations $\mathcal{O}_G^\sim$ par :

$$\mathcal{O}_G^\sim = \{\theta \mid * \in \mathcal{O}_F^\sim ; a \theta b = P_G^\sim(a * b), \forall a, b \in \tilde{G}\}$$

prenons par exemple pour type $T_F^\sim$ le type $T_R^\sim$ défini dans l'exemple 1 et supposons que $G \subset \mathbb{R} \setminus \{0\}$ vérifie la propriété suivante :

$$\exists \tau > 0 : \forall x \in \mathbb{R} \setminus \{0\}, \exists g \in G \text{ tel que } |x - g| \leq \tau |g|.$$

Cette propriété implique alors qu'il existe une projection $P_G^\sim$ de $\tilde{\mathbb{R}}$ dans $\tilde{G}$ vérifiant

$$x \in \mathbb{R} \setminus \{0\} \Rightarrow |x - P_G^\sim(x)| \leq \tau |P_G^\sim(x)|$$

et cette procédure permet de définir un type arithmétique unique dès qu'une telle projection est choisie.

Remarque 2

Comme type de projection classique, il y a l'arrondi, la troncature, etc ...

Définition 4

L'ensemble des types arithmétiques ainsi définis pour τ fixé sera noté L_τ et appelé ensemble des arithmétiques de précision relative τ .

2 - Cas continu

On se donne

$$1 - \tilde{R} = (\mathbb{R} \setminus \{0\}) \cup \{\beta\}$$

2 - une loi de probabilité $(\mathbb{L})$

et on fait l'hypothèse suivante :

$$W(\mathbb{L}) \left\{ \begin{array}{l} v ; \text{ variable aléatoire de loi } (\mathbb{L}), \text{ indépendante par rapport aux} \\ \text{autres variables aléatoires, et qui vérifie} \\ \cdot E(v) = 0 \\ \cdot \sigma^2(v) = \rho_2^2 \\ \cdot \text{ toute réalisation } \varepsilon \text{ de } v \text{ est telle que } P(|\varepsilon| > \tau) = 0 \end{array} \right.$$

A cette loi de Probabilité $(\mathbb{L})$ on associe l'ensemble d'opérations définies par :

$$O_{\tilde{R}}^{(\mathbb{L})} = \left\{ \begin{array}{l} \oplus \mid * \in O_{\tilde{R}} ; a \oplus b = (a * b) \times (1 + \varepsilon), \forall a, b \in \tilde{R}, \text{ où } \varepsilon \text{ est} \\ \text{une réalisation d'une variable aléatoire } v \text{ qui vérifie} \\ \text{l'hypothèse } W(\mathbb{L}) \end{array} \right.$$

Soit alors $T_{\tilde{R}}^{(\mathbb{L})} = (\tilde{R}, O_{\tilde{R}}^{(\mathbb{L})})$ le type arithmétique ainsi défini. Il est clair que la mise en oeuvre de chaque opération dans ce type arithmétique correspond à une réalisation d'une variable aléatoire. Autrement dit, dans l'algorithme chaque résultat d'une opération élémentaire est tiré au hasard en respectant la loi $(\mathbb{L})$.

C'est de ce type précis qu'il s'agira par la suite. Si la réalisation pratique de $T_{\tilde{R}}^{(\mathbb{L})}$ peut paraître utopique, son intérêt est de proposer un cadre

dans lequel on saura établir des résultats. Il revient à l'expérimentation d'infirmar ou de confirmer la validité du modèle ainsi défini.

On dira alors que dans $T_R^{(\mathbb{L})}$, les opérations se font de façon aléatoire selon la loi $(\mathbb{L})$.

Remarque 3

La définition de $\mathcal{O}_G^v$ implique

$$\mathcal{O}_G^v = \{\otimes \mid * \in \mathcal{O}_R^v ; a \otimes b = (a * b) \times (1 + \epsilon), \forall a, b \in \tilde{G}\}$$

et contrairement à $\mathcal{O}_R^{(\mathbb{L})}$, l'erreur relative de représentation de module majoré par τ , est une valeur qui ne dépend que du mode de projection utilisé.

Définition 5

L'ensemble des types arithmétiques pour lesquels la mise en oeuvre de toute opération conduit à une erreur relative de représentation dont le module est majoré par τ , sera noté F_τ .

II.3 - ERREUR DE CALCUL

Définition 6

Considérons un schéma de calcul S sur l'ensemble des variables $V = R \cup D$ et mettons le en oeuvre sur :

- 1 - Le type T_R^v , en notant ϕ_λ les valeurs obtenues
- 2 - Le type $T_F^v \in L_\tau$, $T_F^v \in F_\tau$, en notant $\bar{\phi}_\lambda$ les valeurs obtenues
- 3 - Le type $T_R^{(\mathbb{L})}$, en notant ϕ_λ^1 les valeurs obtenues, où ϕ_λ^1 , $\lambda \in V$ est une valeur tirée au hasard selon la loi $(\mathbb{L})$.

- A tout $\lambda \in V$, associons le nombre $\delta\phi_\lambda$ défini par $\delta\phi_\lambda = \bar{\phi}_\lambda - \phi_\lambda$. Par définition, ce nombre est l'erreur numérique sur ϕ_λ induite par le schéma de calcul S mis en oeuvre sur le type T_F^v avec le jeu des données ϕ_D pourvu que l'initialisation satisfasse :

$$\bar{\phi}_\mu = P_F^\gamma(\phi_\mu), \forall \mu \in D$$

- Pour chaque mise en oeuvre de S sur le type $T_R^\gamma(\mathbb{L})$, à tout $\lambda \in V$ on associe $\delta_1 \phi_\lambda$ défini par $\delta_1 \phi_\lambda = \phi_\lambda^1 - \phi_\lambda$. Ce nombre est par définition l'erreur numérique sur ϕ_λ lors d'une mise en oeuvre de S sur ce type, avec le jeu de données ϕ_D pourvu que l'initialisation satisfasse :

$$\phi_\mu^1 = \phi_\mu (1 + \varepsilon_\mu), \forall \mu \in D$$

où ϕ_μ^1 est une valeur tirée au hasard autour de ϕ_μ selon la loi de probabilité $(\mathbb{L})$, et ε_μ l'erreur relative de représentation est une réalisation d'une variable aléatoire v_μ , v_μ vérifiant l'hypothèse $W(\mathbb{L})$.

Donnons une autre notation de la fonction ϕ_λ . Considérons un schéma de calcul S mis en oeuvre sur le type T_R^γ et soient λ et $\mu \in V$ tels que : $\lambda \prec \mu$. Dans les arguments de $\alpha_\lambda^{(0)} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2))$, (en supposant que désormais toutes les opérations qui interviennent dans ces schémas de calcul comportent deux arguments et soient différentiables sur $\mathbb{R}^{*2}$), remplaçons récursivement ϕ_v par $\alpha_v^{(0)} (\phi_{\alpha_v}(1), \phi_{\alpha_v}(2))$, pour tout $v \in (A_\lambda \cap R) \setminus \{\mu\}$. On obtient ainsi une application du p-uple que l'on notera $f_{\lambda,\mu}$.

$$\gamma_p \xrightarrow{R} \underline{f_{\lambda,\mu}} \xrightarrow{R}$$

$$(\phi_\mu ; \phi_v \mid v \in A_\lambda \cap D) \rightarrow (f_{\lambda,\mu}(\phi_\mu ; \phi_v \mid v \in A_\lambda \cap D))$$

où $p = \text{card}(A_\lambda \cap D) + 1$

on écrira donc

$$\phi_\lambda = f_{\lambda,\mu}(\phi_\mu ; \phi_v \mid v \in A_\lambda \cap D)$$

Définition 7

Soit S un schéma de calcul défini sur $V = R \cup D$ que l'on met en oeuvre sur le type T_R^γ . Une initialisation $\phi_D = \{\phi_\mu \mid \mu \in D\}$ sera dite valide pour le résultat ϕ_λ si on a :

$$\mu \in \bar{A}_\lambda \Rightarrow \phi_\mu \neq \beta \text{ et } \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} \neq 0$$

II.4 - RESULTATS

Proposition 1 [11]

Soit S un schéma de calcul sur V mis en oeuvre sur le type T_R^ν avec une initialisation ϕ_D garantissant $\phi_\lambda \neq \beta$, $\forall \lambda \in V$. Il existe $\tau_0 > 0$ tel que la mise en oeuvre de S sur tout type $T_F^\nu \in F_\tau$, $\tau \leq \tau_0$ conduise, pour tout $\lambda \in R$ à une erreur $\delta\phi_\lambda$ vérifiant :

$$\delta\phi_\lambda = \phi_\lambda \varepsilon_\lambda + (1+\varepsilon_\lambda) \left[\sum_{i=1}^2 \frac{\partial \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2))}{\partial \phi_{\alpha_\lambda}(i)} \delta\phi_{\alpha_\lambda}(i) + \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial^2 \alpha_\lambda^{(o)}(\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2))}{\partial \phi_{\alpha_\lambda}(i) \partial \phi_{\alpha_\lambda}(j)} \delta\phi_{\alpha_\lambda}(i) \cdot \delta\phi_{\alpha_\lambda}(j) \right]$$

avec $\phi'_{\alpha_\lambda}(i) \in [\phi_{\alpha_\lambda}(i), \bar{\phi}_{\alpha_\lambda}(i)]$ qui est un intervalle
 $i = 1, 2$.

L'erreur relative de représentation ε_λ vérifie : $|\varepsilon_\lambda| \leq \tau$

□

Démonstration

Par définition, on a $\delta\phi_\lambda = \bar{\phi}_\lambda - \phi_\lambda$

$$\begin{aligned} \delta\phi_\lambda &= \bar{\alpha}_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \\ &= \bar{\alpha}_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) + \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \\ \text{comme } \bar{\alpha}_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) &= \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) \times \varepsilon_\lambda \end{aligned}$$

on a alors

$$\begin{aligned} \delta\phi_\lambda &= \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) \times \varepsilon_\lambda - \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \varepsilon_\lambda + \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \varepsilon_\lambda + \\ &\quad \alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \end{aligned}$$

ce qui donne

$$\delta\phi_\lambda = \phi_\lambda \varepsilon_\lambda + (1+\varepsilon_\lambda) [\alpha_\lambda^{(o)}(\bar{\phi}_{\alpha_\lambda}(1), \bar{\phi}_{\alpha_\lambda}(2)) - \alpha_\lambda^{(o)}(\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2))]$$

Le fait que les opérations qui interviennent dans le schéma de calcul sont différentiables, et le fait que l'on puisse choisir τ arbitrairement petit, et que les fonctions $\alpha_{\mu}^{(o)}$ soient continues permettent de développer en série de Taylor de second ordre la différence $\alpha_{\lambda}^{(o)}(\bar{\phi}_{\alpha_{\lambda}}(1), \bar{\phi}_{\alpha_{\lambda}}(2)) - \alpha_{\lambda}^{(o)}(\phi_{\alpha_{\lambda}}(1), \phi_{\alpha_{\lambda}}(2))$.

on a

$$\alpha_{\lambda}^{(o)}(\bar{\phi}_{\alpha_{\lambda}}(1), \bar{\phi}_{\alpha_{\lambda}}(2)) - \alpha_{\lambda}^{(o)}(\phi_{\alpha_{\lambda}}(1), \phi_{\alpha_{\lambda}}(2)) = \sum_{i=1}^2 \frac{\partial \alpha_{\lambda}^{(o)}(\phi_{\alpha_{\lambda}}(1), \phi_{\alpha_{\lambda}}(2))}{\partial \phi_{\alpha_{\lambda}}(i)} \delta \phi_{\alpha_{\lambda}}(i) + \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial^2 \alpha_{\lambda}^{(o)}(\phi_{\alpha_{\lambda}}(1), \phi_{\alpha_{\lambda}}(2))}{\partial \phi_{\alpha_{\lambda}}(i) \partial \phi_{\alpha_{\lambda}}(j)} \delta \phi_{\alpha_{\lambda}}(i) \cdot \delta \phi_{\alpha_{\lambda}}(j).$$

avec $\phi'_{\alpha_{\lambda}}(i) \in [\phi_{\alpha_{\lambda}}(i), \bar{\phi}_{\alpha_{\lambda}}(i)]$, $i = 1, 2$

d'où le résultat de la proposition 1.

Exemples sur la proposition 1

1 - $\phi = x_0 \times x_1$

on a $D = \{x_0, x_1\}$

$\bar{\phi} = (x_0 + \delta x_0)(x_1 + \delta x_1)(1 + \epsilon)$

$= x_0 x_1 + x_0 \delta x_1 + x_1 \delta x_0 + \delta x_0 \delta x_1 + (\phi + x_0 \delta x_1 + x_1 \delta x_0 + \delta x_0 \delta x_1) \times \epsilon$

$\delta \phi = \bar{\phi} - \phi = \phi \times \epsilon + (1 + \epsilon)(x_0 \delta x_1 + x_1 \delta x_0 + \delta x_1 \delta x_0)$

2 - Calcul de $\phi = x_0 + x_1$

on a $D = \{x_0, x_1\}$, $\mathcal{O} = \{+\}$, $S_{\lambda} = \{\alpha_{\lambda}^{(o)}, \alpha_{\lambda}^{(1)}, \alpha_{\lambda}^{(2)}\}$

λ	ϕ
$\alpha_{\lambda}^{(o)}$	$+$
$\alpha_{\lambda}^{(1)}$	x_0
$\alpha_{\lambda}^{(2)}$	x_1

$\bar{\phi} = [x_0 + \delta x_0 + x_1 + \delta x_1] (1 + \varepsilon_\phi)$, d'où

$\delta\phi = \phi \varepsilon_\phi + (1 + \varepsilon_\phi)(\delta x_0 + \delta x_1)$, sur ces deux exemples, on suppose les données non exactes.

Enonçons ensuite la proposition suivante dont le résultat traduit tout simplement le théorème de dérivation des fonctions composées, appliqué au mode de calcul de la fonction $f_{\lambda,\mu}(\phi_\mu ; \phi_\nu \mid \nu \in A_\lambda \cap D)$.

Proposition 2, [11]

$$\frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu \mid \nu \in A_\lambda \cap D) = \begin{cases} \sum_{\eta \in \Pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_\eta} (\phi_{\alpha_\lambda^{(1)}}, \phi_{\alpha_\lambda^{(2)}}) \times \frac{\partial f_{\eta,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu \mid \nu \in A_\eta \cap D), & \text{si } \mu \rightarrow \lambda \\ & \mu \rightarrow \eta \\ 1, & \text{si } \lambda = \mu \\ 0, & \text{sinon} \end{cases}$$

□

Proposition 3 [11]

Soit S un schéma de calcul mis en oeuvre sur T_R^ν avec une initialisation valide ϕ_D pour le résultat ϕ_λ ; $\exists \tau_0 > 0$, tel que $\forall T_F^\nu \in F_\tau$, $\tau \leq \tau_0$

on a :

$$\delta\phi_\lambda = \sum_{\mu \in \bar{A}_\lambda} \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu \mid \nu \in A_\lambda \cap D) \times \phi_\mu \varepsilon_\mu + \ell_\lambda \tau^2$$

où

* ε_μ , l'erreur relative de représentation du résultat ϕ_μ a son module majoré par τ

* ℓ_λ est une fonction des ε_μ pour $\mu \in \bar{A}_\lambda$ majorée par L_λ uniformément par rapport à $\tau \in]0, \tau_0]$

□

Démonstration

Elle se fait par récurrence sur λ . Nous allons d'abord montrer que c'est vrai pour tout λ tel que $A_\lambda \subset D$, puis en supposant que ce résultat est vrai pour $\mu \in A_\lambda$, $\mu \rightarrow \lambda$ où λ est un élément quelconque de R , nous montrerons alors que le résultat est vrai pour $\lambda \in R$.

D'après la proposition 1, nous avons :

$$\delta\phi_\lambda = \phi_\lambda \varepsilon_\lambda + (1+\varepsilon_\lambda) \left[\sum_{\nu \in \Pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_\nu} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \cdot \delta\phi_\nu + \right. \\ \left. \frac{1}{2} \sum_{\nu \in \Pi_\lambda} \sum_{\nu' \in \Pi_\lambda} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_\nu \partial \phi_{\nu'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \delta\phi_\nu \cdot \delta\phi_{\nu'} \right]$$

$$\phi'_{\alpha_\lambda}(i) \in [\phi_{\alpha_\lambda}(i), \bar{\phi}_{\alpha_\lambda}(i)], \quad i = 1, 2$$

Supposons tout d'abord que $\Pi_\lambda \subset D$ (d'où $A_\lambda \subset D$). Alors

$$\delta\phi_\lambda = \phi_\lambda \varepsilon_\lambda + \sum_{\nu \in \Pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_\nu} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \phi_\nu \times \varepsilon_\nu + \sum_{\nu \in \Pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_\nu} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \phi_\nu \varepsilon_\nu \times \varepsilon_\lambda + \\ \frac{1}{2} \sum_{\nu \in \Pi_\lambda} \sum_{\nu' \in \Pi_\lambda} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_\nu \partial \phi_{\nu'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \phi_\nu \times \phi_{\nu'} \varepsilon_\nu \times \varepsilon_{\nu'} + \\ \frac{\varepsilon_\lambda}{2} \sum_{\nu \in \Pi_\lambda} \sum_{\nu' \in \Pi_\lambda} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_\nu \partial \phi_{\nu'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \phi_\nu \times \phi_{\nu'} \times \varepsilon_\nu \varepsilon_{\nu'},$$

ce qui nous donne :

$$\delta\phi_\lambda = \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu \mid \nu \in A_\lambda \cap D) \times \phi_\mu \varepsilon_\mu + \ell_\lambda \tau^2$$

où

$$\ell_\lambda = \frac{\varepsilon_\lambda}{\tau} \sum_{\nu \in \Pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_\nu} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \phi_\nu \frac{\varepsilon_\nu}{\tau} + \\ \frac{1}{2} \sum_{\nu \in \Pi_\lambda} \sum_{\nu' \in \Pi_\lambda} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_\nu \partial \phi_{\nu'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \phi_\nu \times \phi_{\nu'} \frac{\varepsilon_\nu}{\tau} \times \frac{\varepsilon_{\nu'}}{\tau}$$

Dans ce cas particulier nous avons :

$$|\ell_\lambda| \leq L_\lambda = \sum_{v \in \pi_\lambda} \left| \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_v} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \phi_v \right| + \frac{1}{2} \sum_{v \in \pi_\lambda} \sum_{v' \in \pi_\lambda} \left| \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_v \partial \phi_{v'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \phi_v \phi_{v'} \right|$$

Le résultat est donc vrai pour tout λ tel que $A_\lambda \subset D$.

Supposons maintenant que ce résultat ait été établi $\forall \mu \rightarrow \lambda$, où λ est un élément quelconque de R . Nous avons alors :

$$\begin{aligned} \delta \phi_\lambda = & \phi_\lambda \varepsilon_\lambda + \sum_{v \in \pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_v} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \left[\sum_{\mu \in A_v} \frac{\partial f_{v,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\eta | \eta \in A_v \cap D) \phi_\mu \varepsilon_\mu + \ell_v \tau^2 \right] \\ & + \varepsilon_\lambda \sum_{v \in \pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_v} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \left[\sum_{\mu \in A_v} \frac{\partial f_{v,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\eta | \eta \in A_v \cap D) \phi_\mu \varepsilon_\mu + \ell_v \tau^2 \right] \\ & + \frac{(1+\varepsilon_\lambda)}{2} \sum_{v \in \pi_\lambda} \sum_{v' \in \pi_\lambda} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_v \partial \phi_{v'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \times \left[\sum_{\mu \in A_v} \frac{\partial f_{v,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\eta | \eta \in A_v \cap D) \phi_\mu \varepsilon_\mu + \ell_v \tau^2 \right] \\ & \times \left[\sum_{\mu' \in A_{v'}} \frac{\partial f_{v',\mu'}}{\partial \phi_{\mu'}} (\phi_{\mu'}; \phi_{\eta'} | \eta' \in A_{v'} \cap D) \phi_{\mu'} \varepsilon_{\mu'} + \ell_{v'} \tau^2 \right] \end{aligned}$$

d'après la proposition 2, nous avons :

$$\begin{aligned} \sum_{v \in \pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_v} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \sum_{\mu \in A_v} \frac{\partial f_{v,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\eta | \eta \in A_v \cap D) \times \phi_\mu \varepsilon_\mu = \\ \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_v | v \in A_\lambda \cap D) \times \phi_\mu \varepsilon_\mu \end{aligned}$$

nous avons alors l'expression suivante

$$\begin{aligned} \delta \phi_v = & \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_v | v \in A_\lambda \cap D) \phi_\mu \varepsilon_\mu + \sum_{v \in \pi_\lambda} \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_v} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \times \ell_v \tau^2 \\ & + \frac{\varepsilon_\lambda}{\tau} \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_v | v \in A_\lambda \cap D) \times \phi_\mu \frac{\varepsilon_\mu}{\tau} \tau^2 + \end{aligned}$$

$$\frac{1}{2} \sum_{v \in \pi_\lambda} \sum_{v' \in \pi_\lambda} \sum_{\mu \in A_v} \sum_{\mu' \in A_{v'}} \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_v \partial \phi_{v'}} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \frac{\partial f_{v,\mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\eta | \eta \in A_v \cap D) \times \phi_\mu \frac{\varepsilon_\mu}{\tau} \times$$

$$\frac{\partial f_{\nu', \mu'}}{\partial \phi_{\mu'}} (\phi_{\mu'}; \phi_{\eta'} | \eta' \in A_{\nu'} \cap D) \times \frac{\phi_{\mu'} \epsilon_{\mu'}}{\tau} \tau^2$$

nous avons alors le résultat cherché

$$\delta \phi_{\lambda} = \sum_{\mu \in A_{\lambda}} \frac{\partial f_{\lambda, \mu}}{\partial \phi_{\mu}} (\phi_{\mu}; \phi_{\nu} | \nu \in A_{\lambda} \cap D) \times \phi_{\mu} \epsilon_{\mu} + \ell_{\lambda} \tau^2$$

où

$$\begin{aligned} |\ell_{\lambda}| &\leq \sum_{\nu \in \pi_{\lambda}} L_{\nu} \times \left| \frac{\partial \alpha_{\lambda}^{(0)}}{\partial \phi_{\nu}} (\phi_{\alpha_{\lambda}}(1), \phi_{\alpha_{\lambda}}(2)) \right| + \sum_{\mu \in A_{\lambda}} \left| \frac{\partial f_{\lambda, \mu}}{\partial \phi_{\mu}} (\phi_{\mu}; \phi_{\nu} | \nu \in A_{\lambda} \cap D) \phi_{\mu} \right| + \\ &\frac{1}{2} \sum_{\nu \in \pi_{\lambda}} \sum_{\nu' \in \pi_{\lambda}} \sum_{\mu \in A_{\nu}} \sum_{\mu' \in A_{\nu'}} \left| \frac{\partial^2 \alpha_{\lambda}^{(0)}}{\partial \phi_{\nu} \partial \phi_{\nu'}} (\phi'_{\alpha_{\lambda}}(1), \phi'_{\alpha_{\lambda}}(2)) \frac{\partial f_{\nu, \mu}}{\partial \phi_{\mu}} (\phi_{\mu}; \phi_{\eta} | \eta \in A_{\nu} \cap D) \times \right. \\ &\left. \frac{\partial f_{\nu', \mu'}}{\partial \phi_{\mu'}} (\phi_{\mu'}; \phi_{\eta'} | \eta' \in A_{\nu'} \cap D) \phi_{\mu} \phi_{\mu'} \right| \end{aligned}$$

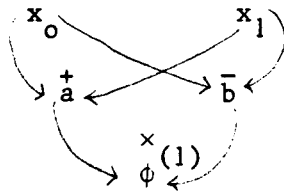
Exemples sur la proposition 3

1 - Calcul de $\phi^{(1)} = (x_0 + x_1)(x_0 - x_1)$

on pose $a = (x_0 + x_1)$; $b = (x_0 - x_1)$

on a $D = \{x_0, x_1\}$, $\bar{A}_{\phi} = \{x_0, x_1, a, b, \phi\}$

Schéma de calcul



si on suppose les données exactes, on a :

$$\bar{a} = (x_0 + x_1)(1 + \epsilon_a) \quad \text{avec} \quad |\epsilon_a| \leq \tau$$

$$\bar{b} = (x_0 - x_1)(1 + \epsilon_b) \quad |\epsilon_b| \leq \tau$$

$$\bar{\phi}^{(1)} = [(x_0 + x_1)(x_0 - x_1)(1 + \epsilon_a)(1 + \epsilon_b)] (1 + \epsilon_{\phi}(1)), \quad |\epsilon_{\phi}(1)| \leq \tau$$

d'où

$$\delta\phi^{(1)} = \bar{\phi}^{(1)} - \phi^{(1)} = \phi^{(1)} (\varepsilon_a + \varepsilon_b + \varepsilon_\phi(1)) + \phi^{(1)} (\varepsilon_a \varepsilon_b + \varepsilon_a \varepsilon_\phi(1) + \varepsilon_b \varepsilon_\phi(1))$$

ce qui est de la forme

$$\delta\phi^{(1)} = \phi^{(1)} (\varepsilon_a + \varepsilon_b + \varepsilon_\phi(1)) + \ell_\phi(1) \tau^2$$

$$\text{avec } |\ell_\phi(1)| \leq 3|\phi^{(1)}|$$

$$2 - \text{Calcul de } \phi^{(2)} = (x_0^2 - x_1^2)$$

$$\text{on pose } a = x_0^2, b = x_1^2$$

On suppose que les données sont exactes.

$$\bar{a} = x_0^2(1+\varepsilon_a), \quad |\varepsilon_a| \leq \tau$$

$$\bar{b} = x_1^2(1+\varepsilon_b), \quad |\varepsilon_b| \leq \tau$$

$$\bar{\phi}^{(2)} = [(x_0^2(1+\varepsilon_a) - x_1^2(1+\varepsilon_b)) (1+\varepsilon_\phi(2))], \quad |\varepsilon_\phi(2)| \leq \tau$$

d'où

$$\delta\phi^{(2)} = \bar{\phi}^{(2)} - \phi^{(2)} = \phi^{(2)} \varepsilon_\phi(2) + x_0^2 \varepsilon_a - x_1^2 \varepsilon_b + x_0^2 \varepsilon_a \varepsilon_\phi(2) - x_1^2 \varepsilon_b \varepsilon_\phi(2)$$

ce qui est de la forme

$$\delta\phi^{(2)} = \phi^{(2)} \varepsilon_\phi(2) + x_0^2 \varepsilon_a - x_1^2 \varepsilon_b + \ell_\phi(2) \tau^2$$

$$\text{avec } |\ell_\phi(2)| \leq x_0^2 + x_1^2$$

Définition 8 Nombre de stabilité en borne [11]

Si une initialisation ϕ_D est valide pour le résultat ϕ_λ du schéma de calcul S,

on introduit le nombre de stabilité en borne comme étant le nombre

$$\Gamma_\lambda(S, \phi_D) = \lim_{\tau \rightarrow 0} \sup_{\substack{T \in L \\ F \in \tau}} \left(\left| \frac{\delta\phi_\lambda}{\phi_\lambda} \right| / \tau \right)$$

CHAPITRE III

MESURE DE LA STABILITE NUMERIQUE EN
MOYENNE D'UN SCHEMA DE CALCUL SANS TEST

INTRODUCTION

Telle qu'elle est définie au chapitre II, la définition du nombre de stabilité en borne précise sur quelle classe de types arithmétiques ce nombre de stabilité est optimal : les types de L_τ , τ tendant vers zéro. Mais il arrive parfois que cette optimalité ne soit plus vérifiée par certains sous-ensembles de L_τ . La majoration donne parfois des bornes trop grossières. C'est pour cela qu'il est nécessaire de compléter cette analyse de majoration par une approche probabiliste qui rendra compte de façon rigoureuse de l'ordre de grandeur effectif de l'erreur. L'étude d'un nombre de stabilité en moyenne fait l'objet de ce chapitre. Partant des notations, et du type arithmétique donnant des résultats aléatoires, du chapitre II, nous allons énoncer des propositions qui vont nous aider pour la définition d'un nombre de stabilité en moyenne.

III.1 - Proposition 1

Soit S un schéma de calcul sur V mis en oeuvre sur le type T_R^γ avec une initialisation ϕ_D garantissant $\phi_\lambda \neq \beta$, $\forall \lambda \in V$. Il existe $\tau_0 > 0$ tel que toute réalisation de S sur le type $T_R^{(\mathbb{L})}$, $\forall \tau \leq \tau_0$, conduise pour $\lambda \in R$ à une erreur $\delta_1 \phi_\lambda$ vérifiant :

$$\delta_1 \phi_\lambda = \phi_\lambda \varepsilon_\lambda + (1 + \varepsilon_\lambda) \sum_{i=1}^2 \frac{\partial \alpha_\lambda^{(0)}}{\partial \phi_{\alpha_\lambda}(i)} (\phi_{\alpha_\lambda}(1), \phi_{\alpha_\lambda}(2)) \delta_1 \phi_{\alpha_\lambda}(i) +$$

$$\frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial^2 \alpha_\lambda^{(0)}}{\partial \phi_{\alpha_\lambda}(i) \partial \phi_{\alpha_\lambda}(j)} (\phi'_{\alpha_\lambda}(1), \phi'_{\alpha_\lambda}(2)) \delta_1 \phi_{\alpha_\lambda}(i) \delta_1 \phi_{\alpha_\lambda}(j)]$$

avec

$$\phi'_{\alpha_\lambda}(i) \in [\phi_{\alpha_\lambda}(i), \phi_{\alpha_\lambda}^1(i)], \quad i = 1, 2$$

L'erreur relative de représentation ε_λ , $\lambda \in V$ est une réalisation d'une variable aléatoire v_λ qui vérifie l'hypothèse $(W(\mathbb{L}))$.

Démonstration

La démonstration est identique à celle de la proposition 1 du chapitre II, il suffit de considérer le cas d'une réalisation de S sur $T_R^{(\mathbb{L})}$, car $T_R^{(\mathbb{L})} \in F_\tau$

III.2 - Proposition 2

Soit S un schéma de calcul mis en oeuvre sur T_R^γ avec une initialisation valide ϕ_D pour le résultat ϕ_λ . Il existe $\tau_0 > 0$ tel que toute réalisation de S sur $T_R^{(\mathbb{L})}$ conduise, $\forall \tau \leq \tau_0$, à une erreur sur le résultat vérifiant

$$\delta_1 \phi_\lambda = \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu |_{\nu \in A_\lambda \cap D}) \times \phi_\mu \varepsilon_\mu + \ell_\lambda \tau^2$$

où

- ε_μ , l'erreur relative de représentation du résultat ϕ_μ est une réalisation d'une variable aléatoire v_μ qui vérifie l'hypothèse $(W(\mathbb{L}))$

- ℓ_λ est une fonction des ε_μ pour $\mu \in \bar{A}_\lambda$ majorée par L_λ uniformément par rapport à $\tau \in]0, \tau_0]$.

□

Remarque 1

La démonstration de cette proposition se fait de la même manière que celle de la proposition 3 du chapitre II, il suffit de considérer le cas d'une réalisation de S sur $T_R^{(L)}$.

III.3 - NOMBRE DE STABILITE EN MOYENNE

Soit S un schéma de calcul mis en oeuvre sur $T_R^{(L)}$ avec une initialisation valide ϕ_D pour le résultat ϕ_λ . Il existe $\tau_0 > 0$, tel que toute réalisation de S sur le type $T_R^{(L)}$, conduite $\forall \tau \leq \tau_0$, à une erreur relative sur le résultat $\frac{\delta_1 \phi_\lambda}{\phi_\lambda}$ qui est une réalisation d'une variable aléatoire dont la variance est donnée par la formule suivante :

$$\sigma^2\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \right)^2 \left(\frac{\phi_\mu}{\phi_\lambda} \right)^2 E(\varepsilon_\mu^2) + m_\lambda \tau^3$$

avec

$$E(\varepsilon_\mu^2) = \rho_2^2 \quad \forall \mu \in \bar{A}_\lambda, \quad \ell_\lambda^1 = \frac{\ell_\lambda}{\phi_\lambda}$$

où

m_λ est une fonction des $E(\ell_\lambda^1 \varepsilon_\mu)$, $\mu \in \bar{A}_\lambda$, majorée par M_λ uniformément par rapport à $\tau \in]0, \tau_0]$

où

$$M_\lambda = 2 \sum_{\mu \in A_\lambda} \left| \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \right| \frac{\phi_\mu}{\phi_\lambda} + \tau |E[\ell_\lambda^1] - E(\ell_\lambda^1)|^2$$

□

Démonstration

D'après la proposition 2, si on considère une réalisation de S, nous avons

$$\frac{\delta_1 \phi_\lambda}{\phi_\lambda} = \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \times \frac{\phi_\mu}{\phi_\lambda} \varepsilon_\mu + \frac{\ell_\lambda}{\phi_\lambda} \tau^2$$

on pose $\frac{\ell_\lambda}{\phi_\lambda} = \ell_\lambda^1$, et on calcule la moyenne de la variable aléatoire $\frac{\delta_1 \phi_\lambda}{\phi_\lambda}$.

$$E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\lambda; \phi_\nu | \nu \in A_\lambda \cap D) \frac{\phi_\mu}{\phi_\lambda} E(\varepsilon_\mu) + \tau^2 E(\ell_\lambda^1)$$

or d'après la définition du type $T_R^{(II)}$ au chapitre II, on a $E(\varepsilon_\mu) = 0$
 $\forall \mu \in \bar{A}_\lambda$ donc $E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \tau^2 E(\ell_\lambda^1)$

pour calculer la variance de $\frac{\delta_1 \phi_\lambda}{\phi_\lambda}$, calculons d'abord l'expression

$$\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)\right)^2, \text{ car } \sigma^2\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = E\left[\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)\right)^2\right], \text{ on a}$$

$$\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \frac{\phi_\mu}{\phi_\lambda} \varepsilon_\mu + \tau^2 (\ell_\lambda^1 - E(\ell_\lambda^1))$$

et

$$\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)\right)^2 = \left(\sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \times \frac{\phi_\mu}{\phi_\lambda} \times \varepsilon_\mu\right)^2 +$$

$$\left[\tau^2 (\ell_\lambda^1 - E(\ell_\lambda^1))\right]^2 +$$

$$2\tau^2 (\ell_\lambda^1 - E(\ell_\lambda^1)) \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \times \frac{\phi_\mu}{\phi_\lambda} \times \varepsilon_\mu$$

comme

$$\left(\sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \times \frac{\phi_\mu}{\phi_\nu} \cdot \varepsilon_\mu\right)^2 = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right)^2 \frac{\phi_\mu}{\phi_\lambda} \times \varepsilon_\mu^2 +$$

$$\sum_{\mu \in A_\lambda} \sum_{\substack{\rho \in A_\lambda \\ \mu \neq \rho}} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu}(\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right) \left(\frac{\partial f_{\lambda, \rho}}{\partial \phi_\rho}(\phi_\rho; \phi_\nu | \nu \in A_\lambda \cap D)\right) \frac{\phi_\mu \times \phi_\rho}{\phi_\lambda^2} \varepsilon_\mu \times \varepsilon_\rho$$

ce qui nous donne en tenant compte du fait que $\varepsilon_\mu, \mu \in \bar{A}_\lambda$ est une réalisation d'une variable aléatoire v_μ qui vérifie $W(\mathbb{L})$ si on cherche $E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)\right)^2$,
 comme les $v_\mu, \mu \in \bar{A}_\lambda$ sont des variables aléatoires indépendantes, on a

$$\begin{aligned}
E\left[\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} - E\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)\right)^2\right] &= \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \left(\frac{\phi_\mu}{\phi_\lambda}\right)^2 E(\varepsilon_\mu^2) + \\
&\quad 2\tau^2 \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \frac{\phi_\mu}{\phi_\lambda} E(\mathcal{L}_\lambda^1 \varepsilon_\mu) + \\
&\quad \tau^4 E[\mathcal{L}_\lambda^1 (\mathcal{L}_\lambda^1 - E(\mathcal{L}_\lambda^1))]^2
\end{aligned}$$

D'après la proposition 2 et la démonstration de la proposition 3 du chapitre II,

on a $E(\mathcal{L}_\lambda^1 \varepsilon_\mu) = 0(\tau)$, on peut alors écrire que :

$$\begin{aligned}
2\tau^2 \sum_{\mu \in A_\lambda} \frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \frac{\phi_\mu}{\phi_\lambda} E(\mathcal{L}_\lambda^1 \varepsilon_\mu) + \\
\tau^4 E[\mathcal{L}_\lambda^1 (\mathcal{L}_\lambda^1 - E(\mathcal{L}_\lambda^1))]^2 &= m_\lambda \tau^3
\end{aligned}$$

ce qui nous donne le résultat cherché.

$$\sigma^2\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right)^2 \left(\frac{\phi_\mu}{\phi_\lambda}\right)^2 E(\varepsilon_\mu^2) + m_\lambda \tau^3$$

D'après la proposition 3, nous avons en notant var = variance

$$\text{var}\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right)^2 \left(\frac{\phi_\mu}{\phi_\lambda}\right)^2 \text{var}(\varepsilon_\mu) + m_\lambda \tau^3$$

comme $\text{var}(\varepsilon_\mu) = \rho_2^2$, $\forall \mu \in \bar{A}_\lambda$, on a

$$\text{var}\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right) = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right)^2 \left(\frac{\phi_\mu}{\phi_\lambda}\right)^2 \rho_2^2 + m_\lambda \tau^3$$

et

$$\frac{\text{var}\left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda}\right)}{\rho_2^2} = \sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D)\right)^2 \left(\frac{\phi_\mu}{\phi_\lambda}\right)^2 + \frac{m_\lambda}{\rho_2^2} \tau^3$$

Définition du nombre de stabilité en moyenne

Soit S un schéma de calcul mis en oeuvre sur le type T_R^γ et sur le type $T_R^{\gamma(L)}$, nous définissons le nombre de stabilité en moyenne du schéma de calcul S comme étant la limite quand τ tend vers zéro de

$$\gamma^2 = \lim_{\tau \rightarrow 0} \frac{\text{var} \left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} \right)}{\rho_2^2}$$

Corollaire

D'après la proposition 3, nous avons

$$\gamma = \sqrt{\lim_{\tau \rightarrow 0} \frac{\text{var} \left(\frac{\delta_1 \phi_\lambda}{\phi_\lambda} \right)}{\rho_2^2}} = \sqrt{\sum_{\mu \in A_\lambda} \left(\frac{\partial f_{\lambda, \mu}}{\partial \phi_\mu} (\phi_\mu; \phi_\nu | \nu \in A_\lambda \cap D) \right)^2 \frac{\phi_\mu^2}{\phi_\lambda^2}}$$

l'expression du nombre de stabilité en moyenne.

CHAPITRE IV

CALCUL D'UNE VALEUR MOYENNE ET D'UN ECART
TYPE DE L'ERREUR RELATIVE DE REPRESENTATION

INTRODUCTION

Afin de pouvoir contrôler le modèle arithmétique que nous avons défini, modèle donnant des résultats aléatoires pour la définition d'un nombre de stabilité en moyenne (chapitre III), il nous faut calculer l'écart type de l'erreur relative de représentation. Nous nous plaçons dans le contexte précis de la représentation des nombres en virgule flottante normalisée, et nous envisageons le cas de l'arrondi d'un nombre avec défaut sur $\frac{1}{2}$. Nous nous intéressons à l'hypothèse de Hamming [26], Knuth [27] sur la répartition statistique des mantisses des nombres. L'objet de ce travail n'est pas neuf et Alt [28] a déjà présenté une réponse. Toutefois en vue d'une simplification des calculs, il a introduit une modification dans l'hypothèse de Hamming [26], à savoir :

[Un nombre machine $\bar{X}$ représente de façon équiprobable tout nombre réel X qui écrit en base b sous forme normalisée et après troncature (ou arrondi) de sa mantisse à n digits nous donne le nombre $\bar{X}$.]

Cette hypothèse revient à choisir une répartition dont la loi est une fonction en escalier. Chacune des marches de cet escalier est de même longueur et centrée sur un nombre machine qui, lui respecte la loi de Hamming. Pour illustrer la différence entre les deux hypothèses prenons l'exemple suivant :

pour la base $b = 2$

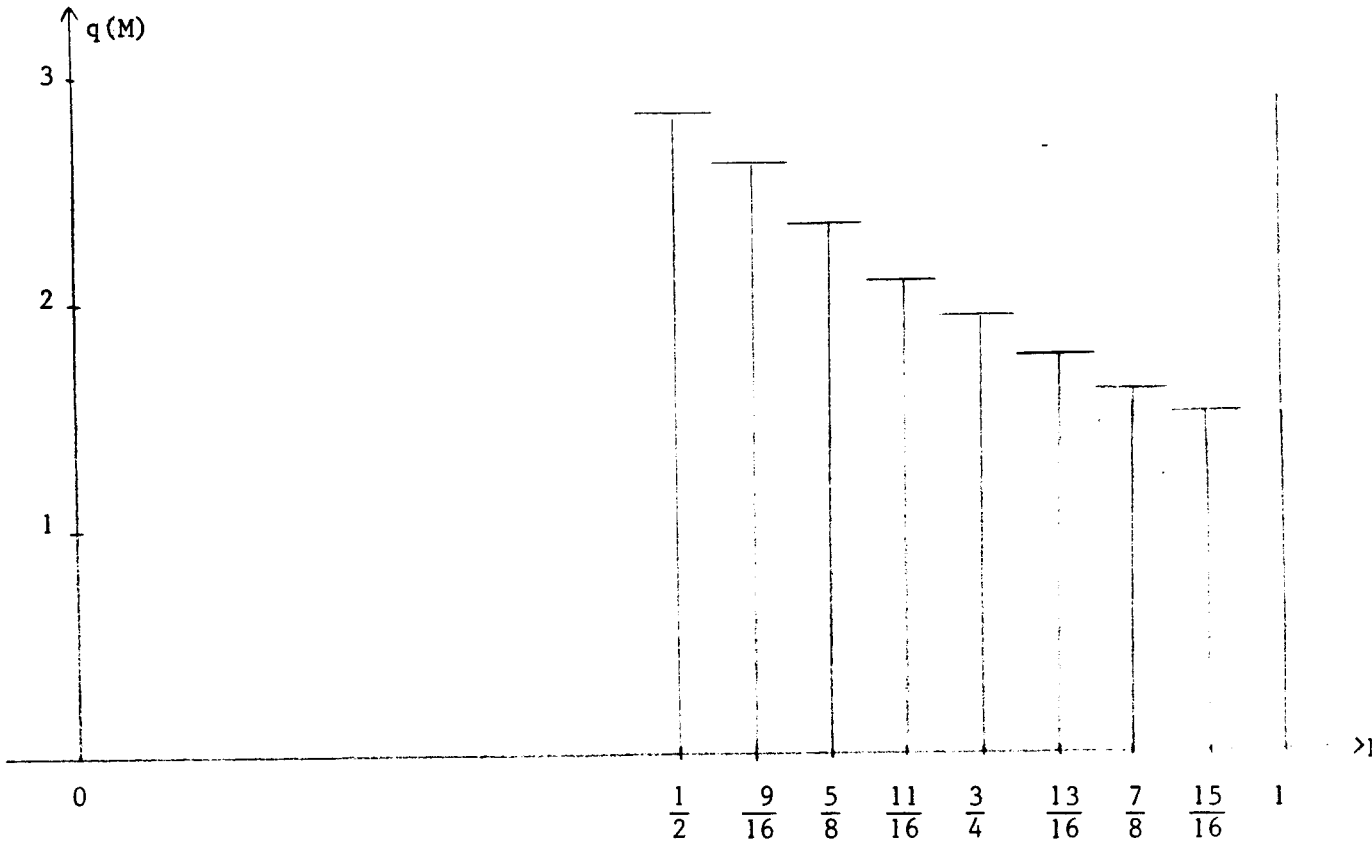
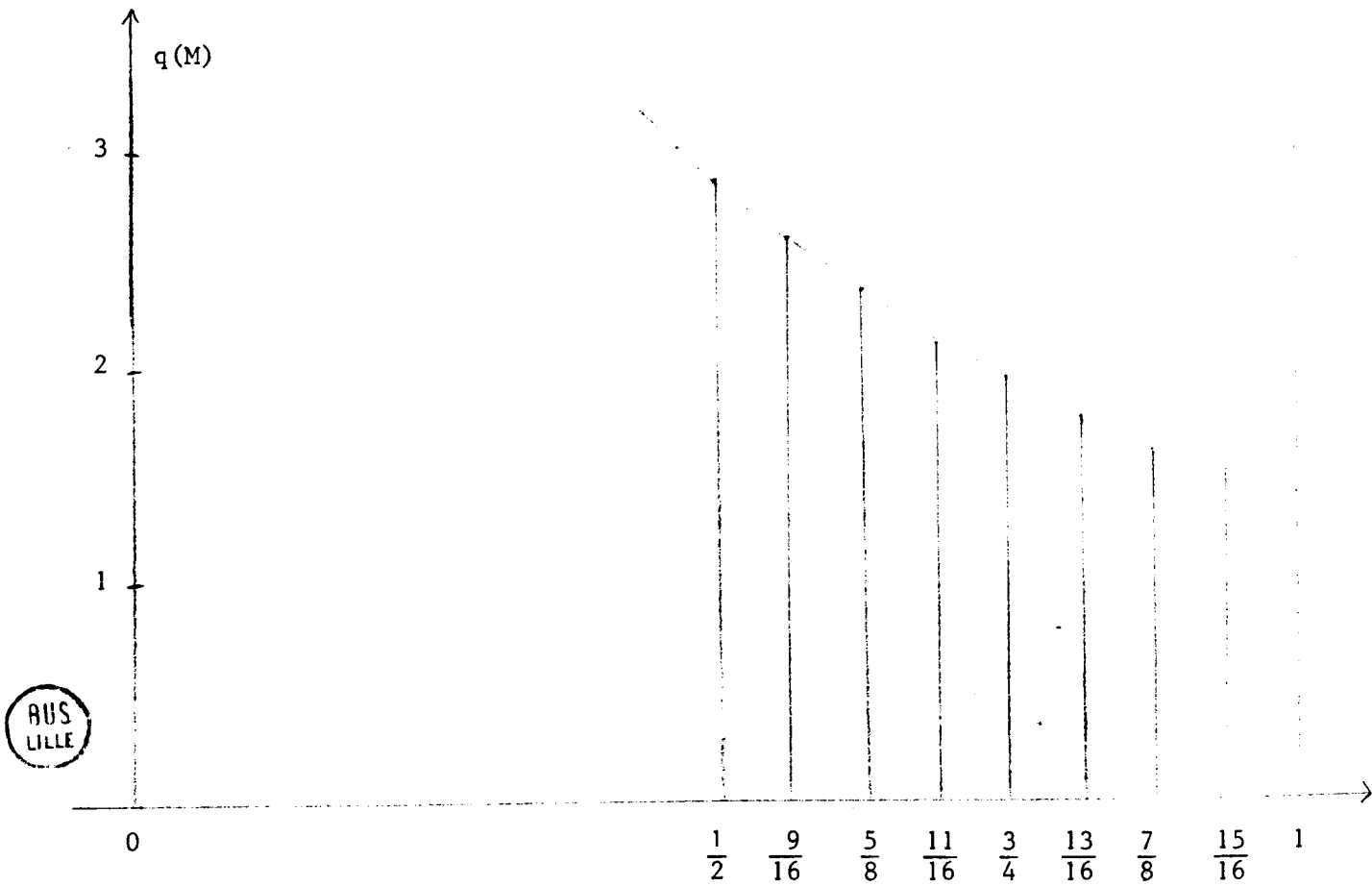
Considérons le cas où la mantisse M d'un nombre machine $\bar{X}$ est représentée sur 4 bits. La loi de distribution des mantisses (de Hamming) étant

$$q(M) = \frac{1}{M \log b}, \quad \frac{1}{b} \leq M < 1$$

Les différentes valeurs que peut prendre M sont :

$$\left(\frac{1}{2} \quad \frac{9}{16} \quad \frac{5}{8} \quad \frac{11}{16} \quad \frac{3}{4} \quad \frac{13}{16} \quad \frac{7}{8} \quad \frac{15}{16} \right)$$

La différence entre les deux hypothèses est illustrée par le graphe de la densité de répartition des mantisses sur l'intervalle $[\frac{1}{2}, 1[$ dans chacun des 2 cas.

Hypothèse de Alt [28]Hypothèse de Hamming

IV.1 - RAPPELS ET NOTATIONS

Toute valeur algébrique réelle $X \neq 0$ peut toujours s'écrire sous la forme virgule flottante normalisée dans une base quelconque b : $X = \pm M \times b^E$. M est une mantisse illimitée qui vérifie $|M| \in [\frac{1}{b}, 1[$, et le nombre E , appelé exposant est un entier relatif.

Pratiquement, on se restreint aux nombres dont l'écriture en base b nécessite m chiffres au plus :

m est la longueur de la mantisse. Cela revient à dire que $\pm M = \pm M' / b^m$, où l'entier relatif M' vérifie :

$$|M'| \in [b^{m-1}, b^m[.$$

Notons par :

$VFN(b, m, \xi)$ = Ensemble des nombres représentables en virgule Flottante de base b et de longueur de mantisse m , et dont l'ensemble des exposants est $\xi \in \mathbb{Z}$. Où ξ est un intervalle de $\mathbb{Z}$ contenant l'origine.

Cette valeur algébrique réelle X ne peut être représentée en machine que de façon approchée. Nous nous intéressons au cas où la représentation de X est faite par arrondi.

On notera par (L) , la loi de probabilité

$$\Pr(M \in [\frac{1}{b}, x[) = \frac{1}{\text{Log } b} \int_{1/b}^x \frac{dM}{M} ; \forall x \in]\frac{1}{b}, 1[$$

On appellera *décomposition* A d'une intégrale définie $\int_a^b f(x)dx$; la décomposition suivante :

$$\int_a^b f(x)dx = \int_a^{a+\frac{1}{2}} f(x)dx + \sum_{k=a+1}^{b-1} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f(x)dx + \int_{b-\frac{1}{2}}^b f(x)dx ; a \text{ et } b \in \mathbb{Z}$$

IV.2 - DEFINITION DE L'ARRONDI D'UN NOMBRE AVEC DEF AUT SUR $\frac{1}{2}$

Notons $\bar{X}$ la valeur de $X > 0$ telle qu'elle est représentée en machine après arrondi : $\bar{X} = \bar{M}' b^{E-m}$, $\bar{M}'$ est la mantisse arrondi à m chiffres dans la base b et est définie de la façon suivante :

$$\bar{M}' = \max \{t \in \mathbb{Z}, t \leq M' + \frac{1}{2}\}$$

Montrons comment se présente $\bar{X}$.

a) Détermination de l'exposant E dans $X = M' \times b^{E-m}$

$X = M' \times b^{E-m}$, le logarithme de X dans la base b, donne

$$\text{Log}_b^X = \text{Log}_b M' + E - m \quad ; \quad b^{m-1} \leq M' < b^m$$

$$\text{Log}_b^{X+1} = \text{Log}_b M' + E - m + 1 \quad ; \quad m-1 \leq \text{Log}_b M' < m$$

on a alors $E \leq \text{Log}_b X+1 < E+1$

d'où $E = \text{Entier}(\text{Log}_b X+1)$

où $\text{Entier}(B) = \text{partie entière de } B$.

b) Détermination de $\bar{M}'$

On calcule d'abord le nombre $c = b^{m-E}$, et ensuite on a

$$M' = X \times C \quad \text{dans } X = M' \times b^{E-m}, \quad X > 0$$

$$\text{et } \bar{M}' = \text{Entier}(M' + \frac{1}{2}) = \text{Entier}(Mb^m + \frac{1}{2})$$

est l'expression de la mantisse M arrondi à m chiffres par défaut sur $\frac{1}{2}$,
telle qu'elle est représentée en machine.

La représentation de X en machine est alors $\bar{X}$ tel que

$$\bar{X} = \bar{M}' \times b^{E-m} = \text{Entier}(Mb^m + \frac{1}{2}) \times b^{E-m}$$

IV.3 - ERREUR RELATIVE DE REPRESENTATION

On considère le cas où la répartition statistique des mantisses des nombres obéit à la loi de probabilité (L), et le cas que nous venons de voir, cas où la mantisse d'un nombre positif X est représentée par $\bar{M}' = \text{Entier}(Mb^m + \frac{1}{2})$ l'erreur relative de représentation commise est donnée par :

$$\frac{\delta(X)}{X} = \frac{X - \bar{X}}{X} = \frac{Mb^E - \text{Entier}(Mb^m + \frac{1}{2}) b^{E-m}}{Mb^E} = 1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m}$$

Nous considérons l'erreur relative de représentation comme une quantité aléatoire. Nous pouvons alors calculer sa valeur moyenne et son écart type. Démonstrons d'abord le lemme suivant :

Lemme 1

On a :

$$\sum_{k=b}^{b^m-1} \frac{2}{2k+1} = \text{Log} b - \frac{b^{-2m}(b^2-1)}{24} + \frac{7}{960} b^{-4m}(b^4-1) - \frac{(64\varepsilon^7 - \eta^7)}{2688 \times \varepsilon^7 \times \eta^7}, \quad \varepsilon \text{ et } \eta \in [b^{m-1}, b^m]$$

□

Démonstration

Pour calculer la série $\sum_{k=b}^{b^m-1} \frac{2}{2k+1}$, nous allons utiliser la formule d'Euler Mac-Laurin pour le calcul des intégrales.

Rappelons cette formule [29].

$$\int_a^b f(x) dx = h \left[f\left(\frac{a}{b}\right) + f(a+h) + \dots + f(a+(n-1)h) + f\left(\frac{b}{2}\right) \right] - \sum_{k=1}^{n-1} \frac{h^{2k} B_{2k}}{(2k)!} [f^{(2k-1)}(b) - f^{(2k-1)}(a)] - \frac{h^{2n+1} B_{2n}}{(2n)!} f^{(2n)}(\xi)$$

$\xi \in [a, b]$

posons

$$f_1(x) = \frac{2}{x}, \text{ on a alors } f_1^{(n)}(x) = \frac{2 \times (-1)^n n!}{x^{n+1}}$$

$$f_2(x) = \frac{1}{x}, \text{ on a } f_2^{(n)}(x) = \frac{(-1)^n n!}{x^{n+1}}$$

La formule de Euler Mac-Laurin appliquée à la fonction $f_1(x) = \frac{2}{x}$ sur l'intervalle $[b^{m-1}, b^m]$, avec un pas $h = \frac{1}{2}$ donne :

$$(1) \quad \int_{b^{m-1}}^{b^m} f_1(x) dx = \frac{1}{2} \left[\frac{f_1(b^{m-1})}{2} + \dots + \frac{f_1(b^m)}{2} \right] - \frac{1}{2^2 \times 2!} B_2 [f_1'(b^m) - f_1'(b^{m-1})] - \left(\frac{1}{2}\right)^4 \frac{B_4}{4!} [f_1'''(b^m) - f_1'''(b^{m-1})] - \left(\frac{1}{2}\right)^7 \frac{B_6}{6!} f_1^{(6)}(\varepsilon)$$

$\varepsilon \in [b^{m-1}, b^m]$

en se limitant à la dérivée d'ordre 6.

La même formule appliquée à la fonction $f_2(x) = \frac{1}{x}$ sur $[b^{m-1}, b^m]$ avec un pas $h = 1$ donne :

$$(2) \int_{b^{m-1}}^{b^m} f_2(x) dx = \left[\frac{f_2(b^{m-1})}{2} + \dots + \frac{f_2(b^m)}{2} \right] - \frac{B_2}{2!} [f'_2(b^m) - f'_2(b^{m-1})] \\ + \frac{B_4}{4!} [f''_2(b^m) - f''_2(b^{m-1})] - \frac{B_6}{6!} f_2^{(6)}(\eta) \\ \eta \in [b^{m-1}, b^m]$$

on rappelle les valeurs des nombres de Bernoulli ;

$B_2 = \frac{1}{6}$; $B_4 = -\frac{1}{30}$; $B_6 = \frac{1}{42}$, on peut alors calculer les expressions (1) et (2)

L'expression (1) devient :

$$(a) \int_{b^{m-1}}^{b^m} \frac{2}{x} dx = \frac{1}{2} \left[\frac{1}{b^{m-1}} + \frac{4}{2b^{m-1}+1} + \dots + \frac{4}{2b^m-1} + \frac{1}{b^m} \right] \\ - \frac{1}{48} [f'_1(b^m) - f'_1(b^{m-1})] + \frac{1}{2^4 \times 720} [f_1^{(3)}(b^m) - f_1^{(3)}(b^{m-1})] \\ - \frac{1}{2^7 \times 720 \times 42} f_1^{(6)}(\epsilon)$$

L'expression (2) devient :

$$(b) \int_{b^{m-1}}^{b^m} \frac{dx}{x} = \left[\frac{1}{2b^{m-1}} + \frac{2}{2b^{m-1}+1} + \dots + \frac{1}{2b^m} \right] - \frac{1}{12} [f'_2(b^m) - f'_2(b^{m-1})] \\ + \frac{1}{720} [f_2^{(3)}(b^m) - f_2^{(3)}(b^{m-1})] - \frac{1}{720 \times 42} f_2^{(6)}(\eta)$$

En faisant la différence (a) - (b), on a :

$$\int_{b^{m-1}}^{b^m} \frac{dx}{x} = \sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1} - \frac{1}{48} \left[\frac{2}{b^{2(m-1)}} - \frac{2}{b^{2m}} \right] + \frac{1}{12} \left[\frac{1}{b^{2(m-1)}} - \frac{1}{b^{2m}} \right] \\ + \frac{1}{2^4 \times 720} \left[\frac{12}{b^{4(m-1)}} - \frac{12}{b^{4m}} \right] - \frac{1}{720} \left[\frac{6}{b^{4(m-1)}} - \frac{6}{b^{4m}} \right] \\ + \frac{1}{42 \times \eta^7} - \frac{1}{2^6 \times 42 \times \epsilon^7}$$

ce qui donne finalement :

$$\int_{b^{m-1}}^{b^m} \frac{dx}{x} = \sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1} + \frac{(b^2-1)b^{-2m}}{24} - \frac{7}{960} b^{-4m}(b^4-1) + \frac{(64\varepsilon^7 - \eta^7)}{2688 \times \varepsilon^7 \times \eta^7}$$

et comme

$$\int_{b^{m-1}}^{b^m} \frac{dx}{x} = \text{Log } b, \text{ nous pouvons alors d\'eduire l'expression de la s\'erie.}$$

$$\sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1} = \text{Log } b - \frac{b^{-2m}(b^2-1)}{24} + \frac{7}{960} b^{-4m}(b^4-1) - \frac{(64\varepsilon^7 - \eta^7)}{2688 \times \varepsilon^7 \times \eta^7}; \quad \varepsilon \text{ et } \eta \in [b^{m-1}, b^m]$$

ce qui ach\ev\e la d\'emonstration.

Enon\c\ons maintenant la proposition suivante pour le calcul de la moyenne de l'erreur relative de repr\'esentation.

IV.4 - Proposition 1

Soit X un nombre positif dont la mantisse est repr\'esent\'ee par

$$\bar{M}' = \text{Entier}(Mb^m + \frac{1}{2}), \text{ et dont l'erreur relative de repr\'esentation est } \frac{\delta(X)}{X}.$$

La r\'epartition statistique des mantisses des nombres ob\'eissant \a la loi de probabilit\'e (L), si on consid\ere l'erreur relative de repr\'esentation comme une quantit\'e al\'eatoire, alors la moyenne de l'erreur relative de repr\'esentation est donn\'ee par la formule suivante :

$$E\left(\frac{\delta(X)}{X}\right) = \frac{b^{-2m}(b^2-1)}{24 \times \text{Log } b} - \frac{7b^{-4m}(b^4-1)}{960 \times \text{Log } b} + \frac{(64\varepsilon^7 - \eta^7)}{2688 \times \text{Log } b \times \varepsilon^7 \times \eta^7}$$

$\eta \text{ et } \varepsilon \in [b^{m-1}, b^m]$

□

D\'emonstration

Rappelons que l'erreur relative de repr\'esentation est donn\'ee par :

$$\frac{\delta(X)}{X} = 1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m}$$

La moyenne de cette erreur selon la loi de probabilité (L) sera :

$$E\left(\frac{\delta(X)}{X}\right) = \frac{1}{\text{Log} b} \int_{\frac{1}{b}}^1 \frac{\delta(X)}{X} \frac{dM}{M} = 1 - \frac{1}{\text{Log} b} \int_{\frac{1}{b}}^1 \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \frac{dM}{M}$$

Dans $\int_{\frac{1}{b}}^1 \frac{\text{Entier}(Mb^m + \frac{1}{2})}{M^2 b^m} dM$, faisons le changement de variable suivant :

$$Mb^m = M', \quad dM = \frac{dM'}{b^m}, \quad \text{cela donne}$$

$$\int_{\frac{1}{b}}^1 \frac{\text{Entier}(Mb^m + \frac{1}{2})}{M^2 b^m} dM = \int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM',$$

la décomposition A appliquée à l'intégrale

$$\int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' \quad \text{donne :}$$

$$\begin{aligned} \int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' &= \int_{b^{m-1}}^{b^{m-1} + \frac{1}{2}} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' \\ &\quad + \sum_{k=b^{m-1}+1}^{b^m-1} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' \\ &\quad + \int_{b^m - \frac{1}{2}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' \end{aligned}$$

D'une part on observe que $M' \in [k - \frac{1}{2}, k] \Rightarrow \text{Entier}(M' + \frac{1}{2}) = k$,

d'où

$$\int_{k-\frac{1}{2}}^k \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' = k \int_{k-\frac{1}{2}}^k \frac{dM'}{M'^2} = \frac{1}{2k-1}$$

d'autre part, $M' \in [k, k + \frac{1}{2}[\Rightarrow \text{Entier}(M' + \frac{1}{2}) = k$, d'où

$$\int_k^{k+\frac{1}{2}} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' = \int_k^{k+\frac{1}{2}} \frac{k dM'}{M'^2} = \frac{1}{2k+1}, \quad \text{ce qui donne}$$

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' = \frac{1}{2k+1} + \frac{1}{2k-1}, \quad \text{d'où l'expression de l'intégrale}$$

$$\int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' = \frac{1}{2b^{m-1}+1} + \sum_{k=b^{m-1}+1}^{b^m-1} \left(\frac{1}{2k+1} + \frac{1}{2k-1} \right) + \frac{1}{2b^m-1}$$

Comme

$$\sum_{k=b^{m-1}+1}^{b^m-1} \left(\frac{1}{2k+1} + \frac{1}{2k-1} \right) = \frac{1}{2b^{m-1}+1} + \frac{1}{2b^m-1} + \sum_{k=b^{m-1}+1}^{b^m-2} \frac{2}{2k+1}$$

donc

$$\int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' = \sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1} \quad (4)$$

puisque

$$E\left(\frac{\delta(X)}{X}\right) = 1 - \frac{1}{\text{Log } b} \int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM', \text{ on a alors l'expression de}$$

la moyenne en remplaçant l'intégrale dans (4)

$$E\left(\frac{\delta(X)}{X}\right) = 1 - \frac{1}{\text{Log } b} \left(\sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1} \right), \text{ et d'après le lemme 1 cela donne finalement}$$

$$E\left(\frac{\delta(X)}{X}\right) = \frac{b^{-2m}(b^2-1)}{24 \times \text{Log } b} - \frac{7b^{-4m}(b^4-1)}{960 \times \text{Log } b} + \frac{(64\varepsilon^7 - \eta^7)}{2688 \times \varepsilon^7 \times \eta^7 \times \text{Log } b}$$

$$\eta \text{ et } \varepsilon \in [b^{m-1}, b^m]$$

ce qui nous donne bien le résultat cherché.

Cherchons à évaluer maintenant l'écart type de l'erreur de représentation.

IV.5 - Proposition 2

Soit X un nombre positif dont la mantisse est représentée par $\bar{M}' = \text{Entier}(Mb^m + \frac{1}{2})$ et dont l'erreur relative de représentation est $\frac{\delta(X)}{X}$ et la moyenne de cette erreur $E(\frac{\delta(X)}{X})$. Sous les mêmes hypothèses que dans la proposition 1, on a l'écart type de l'erreur relative de représentation qui est donné sous forme d'un développement en série par l'expression suivante :

$$\sigma\left(\frac{\delta(X)}{X}\right) \approx \sqrt{\frac{b^{-2m}(b^2-1)}{24 \times \text{Log } b} - \frac{b^{-4m}(b^2-1)}{\text{Log } b} \left[\frac{7(b^2+1)}{960} + \frac{(b^2-1)}{(24)^2 \times \text{Log } b} \right]}$$

□

Démonstration

La variance de l'erreur relative $\frac{\delta(X)}{X}$ est :

$$\sigma^2\left(\frac{\delta(X)}{X}\right) = \frac{1}{\text{Log}b} \int_{\frac{1}{b}}^1 \left[\frac{\delta(X)}{X} - E\left(\frac{\delta(X)}{X}\right) \right]^2 \frac{dM}{M}$$

on remplace $\frac{\delta(X)}{X}$ par sa valeur, cela donne

$$\begin{aligned} \sigma^2\left(\frac{\delta(X)}{X}\right) &= \frac{1}{\text{Log}b} \int_{\frac{1}{b}}^1 \left[1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} - E\left(\frac{\delta(X)}{X}\right) \right]^2 \frac{dM}{M} \\ &= \frac{1}{\text{Log}b} \int_{\frac{1}{b}}^1 \left(1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} - \left[E\left(\frac{\delta(X)}{X}\right) \right]^2 \end{aligned}$$

pour calculer la valeur de cette expression, on cherche d'abord la valeur de l'intégrale

$$\begin{aligned} \int_{\frac{1}{b}}^1 \left(1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} &; \text{ on a } \\ \int_{\frac{1}{b}}^1 \left(1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} &= \int_{\frac{1}{b}}^1 \frac{dM}{M} - 2 \int_{\frac{1}{b}}^1 \frac{\text{Entier}(Mb^m + \frac{1}{2})}{M^2 b^m} dM \\ &\quad + \int_{\frac{1}{b}}^1 \left(\frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} \end{aligned}$$

En faisant le changement de variable suivant : $M' = Mb^m$, on a

$$\begin{aligned} \int_{\frac{1}{b}}^1 \left(\frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} &= \int_{b^{m-1}}^{b^m} \left(\text{Entier}(M' + \frac{1}{2}) \right)^2 \frac{dM'}{M'^3} \\ \int_{\frac{1}{b}}^1 \frac{\text{Entier}(Mb^m + \frac{1}{2})}{M^2 b^m} dM &= \int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' \end{aligned}$$

d'où

$$\int_{\frac{1}{b}}^1 \left(1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} = \text{Log}b - 2 \int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM' + \int_{b^{m-1}}^{b^m} \left(\text{Entier}(M' + \frac{1}{2}) \right)^2 \frac{dM'}{M'^3}$$

La décomposition A appliquée à l'intégrale $\int_{b^{m-1}}^{b^m} \frac{\text{Entier}(M' + \frac{1}{2})}{M'^2} dM'$, donne

$$\int_{b^{m-1}}^{b^m} \text{Entier} \left(M' + \frac{1}{2} \right) \frac{dM'}{M'^2} = \sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1}, \text{ et d'après le lemme 1, et la}$$

proposition 1, on a :

$$\int_{b^{m-1}}^{b^m} \frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'^2} dM' = \text{Log } b \left(1 - E \left(\frac{\delta(X)}{X} \right) \right)$$

Pour calculer la seconde intégrale $\int_{b^{m-1}}^{b^m} \left(\frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'} \right)^2 \frac{dM'}{M'}$,

on applique la *décomposition A* à cette intégrale. On pourra alors avoir la valeur de l'expression de l'écart type de l'erreur relative de représentation. On applique donc la *décomposition A* à cette intégrale, ce qui donne

$$\begin{aligned} \int_{b^{m-1}}^{b^m} \left(\frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'} \right)^2 \frac{dM'}{M'} &= \int_{b^{m-1}}^{b^{m-1} + \frac{1}{2}} \left(\frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'} \right)^2 \frac{dM'}{M'} \\ &+ \sum_{k=b^{m-1}+1}^{b^m-1} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \left(\frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'} \right)^2 \frac{dM'}{M'} \\ &+ \int_{b^m - \frac{1}{2}}^{b^m} \left(\frac{\text{Entier} \left(M' + \frac{1}{2} \right)}{M'} \right)^2 \frac{dM'}{M'} \end{aligned}$$

d'une part, $M' \in [k - \frac{1}{2}, k] \Rightarrow \text{Entier} \left(M' + \frac{1}{2} \right) = k$

d'où

$$\int_{k-\frac{1}{2}}^k \left(\text{Entier} \left(M' + \frac{1}{2} \right) \right)^2 \frac{dM'}{M'^3} = \int_{k-\frac{1}{2}}^k \frac{k^2 dM'}{M'^3} = \frac{2k^2}{(2k-1)^2} - \frac{1}{2}$$

d'autre part, $M' \in [k, k + \frac{1}{2}[\Rightarrow \text{Entier} \left(M' + \frac{1}{2} \right) = k$, d'où

$$\int_k^{k+\frac{1}{2}} \left(\text{Entier} \left(M' + \frac{1}{2} \right) \right)^2 \frac{dM'}{M'^3} = \frac{1}{2} - \frac{2k^2}{(2k+1)^2}$$

donc

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \left(\text{Entier} \left(M' + \frac{1}{2} \right) \right)^2 \frac{dM'}{M'^3} = \frac{2k^2}{(2k-1)^2} - \frac{2k^2}{(2k+1)^2}$$

et

$$\begin{aligned}
\int_{b^{m-1}}^{b^m} (\text{Entier}(M' + \frac{1}{2}))^2 \frac{dM'}{M'^3} &= 2 \left(\frac{b^m}{2b^{m-1}} \right)^2 - 2 \left(\frac{b^{m-1}}{2b^{m-1}+1} \right)^2 + \sum_{k=b^{m-1}+1}^{b^m-1} \left(\frac{2k^2}{(2k-1)^2} - \frac{2k^2}{(2k+1)^2} \right) \\
&= \sum_{k=b^{m-1}+1}^{b^m-1} 2 \left(\frac{k}{2k-1} \right)^2 - \sum_{k=b^{m-1}}^{b^m-1} 2 \left(\frac{2}{2k+1} \right)^2 \\
&= \sum_{k=b^{m-1}}^{b^m-1} \frac{2}{2k+1}
\end{aligned}$$

et d'après le lemme 1 et la proposition 1, on a

$$\int_{b^{m-1}}^{b^m} (\text{Entier}(M' + \frac{1}{2}))^2 \frac{dM'}{M'^3} = \text{Log } b \left(1 - E\left(\frac{\delta(X)}{X}\right) \right)$$

d'où finalement

$$\int_{\frac{1}{b}}^1 \left(1 - \frac{\text{Entier}(Mb^m + \frac{1}{2})}{Mb^m} \right)^2 \frac{dM}{M} = \text{Log } b \times E\left(\frac{\delta(X)}{X}\right)$$

ce qui donne pour $\sigma^2\left(\frac{\delta(X)}{X}\right)$ la valeur suivante :

$$\sigma^2\left(\frac{\delta(X)}{X}\right) = E\left(\frac{\delta(X)}{X}\right)^2 - [E\left(\frac{\delta(X)}{X}\right)]^2$$

on remplace alors $E\left(\frac{\delta(X)}{X}\right)$ par la valeur de son expression de

la proposition 1, et en se limitant aux termes en τ^4 , avec $\tau = \frac{b^{1-m}}{2}$, on a l'expression finale de $\sigma^2\left(\frac{\delta(X)}{X}\right)$. On note $\sigma^2\left(\frac{\delta(X)}{X}\right) = \rho_2^2$

$$\sigma^2\left(\frac{\delta(X)}{X}\right) \approx \frac{b^{-2m}(b^2-1)}{24 \times \text{Log } b} - \frac{b^{-4m}(b^2-1)}{\text{Log } b} \left[\frac{7(b^2+1)}{960} + \frac{(b^2-1)}{(24)^2 \times \text{Log } b} \right]$$

ce qui termine la démonstration en prenant la racine carrée de l'expression.

Si on limite le développement aux termes en τ^2 , on a finalement l'expression

$$(B) \quad \sigma^2\left(\frac{\delta(X)}{X}\right) \approx \frac{b^{-2m}(b^2-1)}{24 \times \text{Log } b}$$

On observera que l'expression (B) est la même que celle obtenue par Alt [28]

avec l'hypothèse de Hamming modifiée. Le calcul est alors beaucoup plus simple, ce qui justifie a posteriori cette hypothèse simplificatrice.

L'intérêt de ce travail est donc d'établir la validité de l'hypothèse simplificatrice de Alt.

CHAPITRE V
APPLICATIONS ET RESULTATS NUMERIQUES

INTRODUCTION

Après le calcul de ρ_2 l'écart type de l'erreur relative de représentation dans une arithmétique flottante (chapitre IV), pour contrôler le modèle défini au chapitre II, deux possibilités nous sont offertes à savoir :

- Application directe du modèle donnant des résultats aléatoires, en utilisant une technique dans le genre de la méthode de perturbation de Vignes [19] (méthode qui peut être considérée comme une digitalisation de notre modèle arithmétique).
- Utilisation d'une simulation d'arithmétique.

Nous avons choisi cette dernière possibilité. Pour une base b et une longueur m de mantisse données, notons $ARR(b, m ; \mathbb{Z}, ar)$ l'arithmétique définie à partir de la virgule flottante normalisée $VFN(b, m, \xi)$, dans laquelle toutes les opérations sont faites avec des nombres appartenant à $VFN(b, m, \xi)$, arrondis. $ARR(b, m ; \mathbb{Z}, ar)$ est une arithmétique flottante avec m chiffres en base b , avec arrondi. Pour un schéma de calcul S , avec un jeu de données fixé, notre simulation d'arithmétique consiste à faire varier b et m dans $ARR(b, m ; \mathbb{Z}, ar)$, à chercher l'erreur relative faite sur un résultat donné entre l'arithmétique exacte et l'arithmétique approchée $ARR(b, m ; \mathbb{Z}, ar)$, et à calculer le rapport : $|(erreur\ relative\ sur\ le\ résultat)|/\rho_2$, ρ_2 étant l'écart type de l'erreur relative de représentation. A une valeur de b donnée, et à une valeur de m donnée, correspond une valeur du rapport $|(erreur\ relative\ sur\ le\ résultat)|/\rho_2$, et lorsque b et m varient, nous obtenons alors plusieurs valeurs de ce rapport.

Notre contrôle consistera à vérifier que, en prenant un échantillon de valeurs de b , les valeurs des rapports $|(erreur\ relative\ sur\ le\ résultat)|/\rho_2$

obtenues se distribuent assez bien autour du nombre de stabilité en moyenne Gamma, défini au chapitre III. Pour une base b donnée, on pourra calculer une moyenne des rapports $|(\text{erreur relative sur le résultat})|/\rho_2$ obtenus lorsque m, la longueur de la mantisse varie. Dans les résultats que nous présentons, c'est cette procédure qui est suivie.

La première partie de ce chapitre est consacrée au calcul algébrique des nombres de stabilité en borne et en moyenne dans les exemples que nous avons pris comme application pour contrôler le modèle défini au chapitre II.

Les exemples choisis sont les suivants :

1 - somme de n nombres par :

- i) La méthode de cumul simple
- ii) La méthode de Linz

2 - Calcul de l'expression du Δ^2 d'Aitken

- i)
$$X = \frac{(x_0 x_2 - x_1^2)}{x_0 + x_2 - 2x_1}$$
- ii)
$$Y = x_2 - \frac{(x_2 - x_1)^2}{x_0 + x_2 - 2x_1}$$

3 - Calcul de la valeur d'un polynôme par :

- i) L'algorithme de Horner
- ii) Le schéma classique

La deuxième partie de ce chapitre est consacrée au contrôle de la simulation et à l'analyse de tous les résultats. Nous donnons en annexe, tous les programmes de calcul.

V.1 - APPLICATIONS

Rappelons les définitions du nombre de stabilité en borne et du nombre de stabilité en moyenne d'un schéma de calcul S.

Le nombre de stabilité en borne est défini à partir de l'erreur relative sur le résultat ϕ_λ , de la manière suivante (proposition 3 du chapitre II) ; on a

$$\frac{\delta\phi_\lambda}{\phi_\lambda} = \sum_{\mu \in \bar{A}_\lambda} \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu / \nu \in A_\lambda \cap D) \times \phi_\mu \varepsilon_\mu + \ell_\lambda \tau^2$$

et le nombre de stabilité en borne est le plus petit nombre $C_\lambda(S, \phi_D)$

tel que

$$\left| \frac{\delta\phi_\lambda}{\phi_\lambda} \right| \leq (C_\lambda(S, \phi_D)) \times \tau,$$

donc

$$C_\lambda(S, \phi_D) = \sum_{\mu \in \bar{A}_\lambda} \left| \frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu / \nu \in A_\lambda \cap D) \times \frac{\phi_\mu}{\phi_\lambda} \right|$$

D'autre part, le nombre de stabilité en moyenne d'un schéma de calcul S est donné par :

$$\gamma = \sqrt{\sum_{\mu \in \bar{A}_\lambda} \left(\frac{\partial f_{\lambda,\mu}}{\partial \phi_\mu} (\phi_\mu ; \phi_\nu / \nu \in A_\lambda \cap D) \right)^2 \left(\frac{\phi_\mu}{\phi_\lambda} \right)^2}$$

donc il suffit seulement d'avoir l'expression de la propagation des erreurs pour calculer les deux nombres de stabilité. C'est ce que nous ferons dorénavant dans les exemples que nous allons examiner.

1.1 - L'effet de cumul [24]

Pour calculer la somme $S_n = \sum_{i=1}^n a_i$, on utilise souvent l'algorithme suivant :

$$\begin{cases} S_1 = a_1 \\ S_i = S_{i-1} + a_i \quad ; \quad i = 2, \dots, n \end{cases}$$

L'erreur se propage de la façon suivante :

$$\delta S_1 = \delta a_1$$

$$\delta S_i = \delta S_{i-1} + \delta a_i + S_i \epsilon_i$$

avec $|\epsilon_i| \leq \tau$ pour $i = 2, \dots, n$

où

$$\delta S_{i-1} + \delta a_i = \text{Erreur Transmise}$$

Si ϵ_i = Erreur générée

on montre facilement par récurrence que :

$$\delta S_i = \sum_{j=1}^i \delta a_j + \sum_{j=2}^i S_j \epsilon_j$$

et

$$\frac{\delta S_n}{S_n} = \sum_{j=1}^n \frac{\delta a_j}{S_n} + \sum_{j=2}^n \frac{S_j \epsilon_j}{S_n}$$

le nombre de stabilité en borne est :

$$C_n = \sum_{j=2}^n \frac{|S_j|}{|S_n|}$$

Si on suppose que les a_i sont tous de même signe, alors

$$C_n = \frac{1}{S_n} \sum_{j=2}^n \sum_{i=1}^j a_i = \frac{1}{S_n} \left[\sum_{i=1}^n (n+1-i) a_i - a_1 \right]$$

Les a_i interviennent avec un coefficient d'autant plus fort que leur indice est faible. D'un autre côté, si on suppose en plus que les a_i sont tous de même ordre de grandeur alors ;

$$C_n \simeq \frac{1}{na} \left(\sum_{i=1}^n ia - a \right) = \frac{n(n+1)-2}{2n} \simeq \frac{n}{2}$$

où a est un nombre $\neq 0$ tel que $|(a_i - a)/a| \ll 1$ donc dans ce cas particulier,

C_n est de l'ordre de $\frac{n}{2}$. On a alors $\lim_{n \rightarrow \infty} C_n = \infty$; ce qui veut dire que l'erreur

de calcul peut devenir très importante dans une somme par cumul.

Le nombre de stabilité en moyenne est :

$$\gamma = \sqrt{\sum_{i=2}^n \left(\frac{s_i}{s_n}\right)^2}$$

Pour faire une comparaison avec le nombre de stabilité en borne, on se place dans le même contexte particulier que précédemment : on suppose que les a_i sont tous de même signe, et de même ordre de grandeur : $\exists a \neq 0$ tq $|(a_i - a)/a| \ll 1$.

Alors $\frac{\sum_{i=2}^n (s_i)^2}{\left(\sum_{i=1}^n a_i\right)^2}$ a pour ordre de grandeur la quantité

$$\frac{\sum_{i=2}^n \left(\sum_{j=1}^i a\right)^2}{\left(\sum_{i=1}^n a\right)^2} = \frac{\sum_{i=2}^n i^2}{n^2} = \frac{\frac{n(n+1)(2n+1)}{6} - 1}{n^2} \approx \frac{n}{3}$$

donc γ est de l'ordre de $\sqrt{\frac{n}{3}}$

$$\gamma \approx \sqrt{\frac{n}{3}}$$

1.2 - Algorithme de LINZ [23]

La méthode précédente est donc très sensible aux erreurs de calcul puisque l'erreur relative augmente indéfiniment avec le nombre de termes de la somme. On peut atténuer cet effet de cumul en utilisant la méthode suivante proposée par LINZ [23].

Algorithme

Pour calculer la somme de n nombres, on regroupe ces nombres en $[(n+1) \div 2]$ paires de nombres et on effectue les $n \div 2$ additions. On est alors ramené à une somme de $[(n+1) \div 2]$ nombres. Il suffit d'itérer le procédé. ($[u]$ représente ici la partie entière de u et $\div$ la division entière).

Par rapport au cumul simple, le coût opératoire n'est pas changé, mais la programmation est plus compliquée et de plus, on doit disposer ici d'une place mémoire plus importante. (De l'ordre de $\log_2 n$ au lieu de 1 avec le cumul simple). Nous allons utiliser la notion du nombre de stabilité pour montrer que cet algorithme est plus stable. Toutefois pour simplifier l'exposé, nous nous contenterons de l'établir dans le cas où n le nombre de terme de la somme, est une puissance entière de 2. L'algorithme de calcul est le suivant :

$$\begin{cases} v_{0,j} = a_j & ; j = 1, \dots, 2^k \\ v_{i,j} = v_{i-1,2j-1} + v_{i-1,2j} & ; i = 1, \dots, k \\ & j = 1, \dots, 2^{k-i} \\ v_{k,1} = \sum_{i=1}^n a_i, \text{ avec } n = 2^k \end{cases}$$

L'erreur dont $v_{i,j}$ est entaché est :

$$\delta v_{i,j} = \delta v_{i-1,2j-1} + \delta v_{i-1,2j} + \epsilon_{i,j} v_{i,j} \quad \text{avec } |\epsilon_{i,j}| \leq \tau$$

τ étant l'erreur relative maximale de représentation.

$j = 1, \dots, 2^{k-i}$, nous avons

$$\delta v_{i,1} = \delta v_{i-1,1} + \delta v_{i-1,2} + \epsilon_{i,1} v_{i,1}$$

$$\delta v_{i,2} = \delta v_{i-1,3} + \delta v_{i-1,4} + \epsilon_{i,2} v_{i,2}$$

$$\delta v_{i,3} = \delta v_{i-1,5} + \delta v_{i-1,6} + \epsilon_{i,3} v_{i,3}$$

- - - - -

$$\delta v_{i,2^{k-i}} = \delta v_{i-1,2^{k-i+1}} + \delta v_{i-1,2^{k-i+1}} + \epsilon_{i,2^{k-i}} v_{i,2^{k-i}}$$

et en faisant la somme membre à membre, cela donne alors

$$\sum_{j=1}^{2^{k-i}} \delta v_{i,j} = \sum_{j=1}^{2^{k-i+1}} \delta v_{i-1,j} + \sum_{j=1}^{2^{k-i}} \epsilon_{i,j} v_{i,j}$$

$i = 1, \dots, k$, nous avons

$$\sum_{j=1}^{2^{k-1}} \delta v_{1,j} = \sum_{j=1}^{2^k} \delta v_{0,j} + \sum_{j=1}^{2^{k-1}} \epsilon_{1,j} v_{1,j}$$

- - - - -

$$\sum_{j=1}^{2^2} \delta v_{k-2,j} = \sum_{j=1}^{2^3} \delta v_{k-3,j} + \sum_{j=1}^{2^2} \epsilon_{k-2,j} v_{k-2,j}$$

$$\sum_{j=1}^2 \delta v_{k-1,j} = \sum_{j=1}^{2^2} \delta v_{k-2,j} + \sum_{j=1}^2 \epsilon_{k-1,j} v_{k-1,j}$$

$$\delta v_{k,1} = \sum_{j=1}^2 \delta v_{k-1,j} + \epsilon_{k,1} v_{k,1}$$

d'où par récurrence, on montre que

$$\delta v_{k,1} = \sum_{j=1}^{2^k} \delta v_{0,j} + \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \epsilon_{i,j} v_{i,j}$$

Le nombre de stabilité en borne est :

$$C_n = \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \frac{|v_{i,j}|}{|v_{k,1}|}$$

Dans le cas particulier où les a_j sont tous positifs, on a

$$\sum_{j=1}^{2^{k-i}} |v_{i,j}| = \sum_{j=1}^{2^k} a_j, \text{ et le nombre de stabilité en borne } C_n, \text{ devient ;}$$

$$C_n = \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \frac{|v_{i,j}|}{|v_{k,1}|} = \frac{\sum_{i=1}^k \sum_{j=1}^{2^k} a_j}{\sum_{j=1}^{2^k} a_j} = k \frac{\sum_{j=1}^{2^k} a_j}{\sum_{j=1}^{2^k} a_j} = k$$

$$C_n = k$$

Alors que pour le cumul simple, $C_n \approx \frac{n}{2}$, le nombre de stabilité en moyenne est :

$$\gamma = \sqrt{\sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \left(\frac{v_{i,j}}{v_{k,1}} \right)^2}$$

Dans le cas particulier où les a_j sont tous positifs et de même ordre de grandeur, on a

$$v_{i,j} = \sum_{\ell=1}^{2^i} a_{\ell+2^{j-1}} = a \cdot 2^i$$

$$v_{k,1} = a \cdot 2^k$$

donc

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \frac{(v_{i,j})^2}{(v_{k,1})^2} &= \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \frac{(a2^i)^2}{(a2^k)^2} = \sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \frac{2^{2i}}{2^{2k}} \\ &= \frac{\sum_{i=1}^k 2^{k+i}}{2^{2k}} = \frac{\sum_{i=1}^k 2^i}{2^k} = \frac{2^{k+1} - 1}{2^k} \end{aligned}$$

ce qui donne

$$\sum_{i=1}^k \sum_{j=1}^{2^{k-i}} \left(\frac{v_{i,j}}{v_{k,1}} \right)^2 \approx 2$$

donc $\gamma \approx \sqrt{2}$

Le nombre de stabilité en moyenne est indépendant du nombre des termes de la somme. Ce qui montre que l'algorithme de LINZ est plus stable, et donc préférable au cumul simple dans la somme de n nombres.

1.3 - Calcul de l'expression du Δ^2 d'Aitken

Le calcul de l'expression se fait avec les nombres x_0, x_1, x_2 qui sont de même signe, ordonnés, et le plus petit est supérieur à la moitié du plus grand. (Exemple, $x_0 = 0,510, x_1 = 0,500, x_2 = 0,491$).

a) Version stable

$$Y = x_2 - \frac{(x_2 - x_1)^2}{x_0 + x_2 - 2x_1}$$

Pour calculer cette expression, on utilise l'algorithme suivant :

$$\left\{ \begin{array}{l} d_0 = x_1 - x_0 \\ d_1 = x_2 - x_1 \\ d = d_1 - d_0 \\ c = d_1^2 \\ q = \frac{c}{d} \\ Y = x_2 - q \end{array} \right.$$

L'erreur se propage de la manière suivante :

Si on suppose les données x_0, x_1, x_2 exactes, il existe des nombres

$\varepsilon_{d_0}, \varepsilon_{d_1}, \varepsilon_d, \varepsilon_c, \varepsilon_q, \varepsilon_Y$ de module majoré par τ tels que

$$\delta c = 2d_1 \delta d_1 + c \varepsilon_c$$

$$\frac{\delta q}{q} = \frac{\delta c}{c} - \frac{\delta d}{d} + \varepsilon_q$$

$$\delta Y = -\delta q + Y \varepsilon_Y$$

d'où

$$\frac{\delta Y}{Y} = -\frac{q}{Y} (2\varepsilon_{d_1} + \varepsilon_c + \varepsilon_q - (\frac{d_1 \varepsilon_{d_1} - d_0 \varepsilon_{d_0}}{d} + \varepsilon_d)) + \varepsilon_Y$$

$$\frac{\delta Y}{Y} = -\frac{q}{Y} (\varepsilon_c + \varepsilon_q - \varepsilon_d + \frac{(d_1 - 2d_0)}{d} \varepsilon_{d_1} + \frac{d_0}{d} \varepsilon_{d_0}) + \varepsilon_Y$$

d'où on a

$$|\frac{\delta Y}{Y}| \leq [|\frac{q}{Y}| (3 + \frac{|d_1 - 2d_0| + |d_0|}{|d|}) + 1] \tau$$

et

$$C_n = |\frac{q}{Y}| (3 + \frac{|d_1 - 2d_0| + |d_0|}{|d|}) + 1$$

et le nombre de stabilité en moyenne est :

$$\gamma = \sqrt{1 + (\frac{q}{Y})^2 [3 + \frac{(d_1 - 2d_0)^2}{d^2} + \frac{d_0^2}{d^2}]}$$

b) Version instable

$$X = \frac{(x_0 x_2 - x_1^2)}{x_0 + x_2 - 2x_1}$$

Pour le calcul de la valeur de cette expression, on utilise l'algorithme suivant :

$$\left\{ \begin{array}{l} p = x_0 x_2 \\ c = x_1^2 \\ n = p - c \\ d_0 = x_1 - x_0 \\ d_1 = x_2 - x_1 \\ d = d_1 - d_0 \\ X = \frac{n}{d} \end{array} \right.$$

Si on suppose les données x_0, x_1, x_2 exactes, $\exists \epsilon_p, \epsilon_c, \epsilon_n, \epsilon_{d_0}, \epsilon_{d_1}, \epsilon_d, \epsilon_X$ de module majoré par τ tels que :

$$\delta d_0 = d_0 \times \epsilon_{d_0}$$

$$\delta d_1 = d_1 \times \epsilon_{d_1}$$

$$\delta p = p \times \epsilon_p$$

$$\delta c = c \times \epsilon_c$$

$$\delta d = \delta d_1 - \delta d_0 + d \times \epsilon_d$$

$$\delta n = \delta p - \delta c + n \times \epsilon_n$$

$$\frac{\delta X}{X} = \frac{\delta n}{n} - \frac{\delta d}{d} + \epsilon_X, \text{ d'où}$$

$$\frac{\delta X}{X} = \frac{p\epsilon_p - c\epsilon_c}{n} + \epsilon_n + \epsilon_X - \left(\frac{d_1\epsilon_{d_1} - d_0\epsilon_{d_0}}{d} + \epsilon_d \right)$$

$$\left| \frac{\delta X}{X} \right| \leq \left[\frac{|p|+|c|}{|n|} + 3 + \frac{|d_1|+|d_0|}{|d|} \right] \tau$$

$$\text{et } c_n = 3 + \frac{|p|+|c|}{|n|} + \frac{|d_1|+|d_0|}{|d|}$$

et le nombre de stabilité en moyenne est ;

$$\gamma = \sqrt{3 + \frac{p^2+c^2}{n^2} + \frac{d_1^2+d_0^2}{d^2}}$$

1.4 - Calcul de la valeur d'un polynôme par l'algorithme de Horner

Pour calculer la valeur du polynôme

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n$$

l'algorithme de Horner consiste à utiliser la forme

$$P_n(x) = (\dots((a_0 x + a_1)x + a_2)x + \dots + a_{n-1})x + a_n$$

c'est à dire à former la suite

$$\begin{cases} b_0 = a_0 \\ b_i = b_{i-1}x + a_i ; i = 1, \dots, n \end{cases}$$

$$\text{on a } b_n = P_n(x)$$

on suppose les données exactes ; pour l'étude de la propagation des erreurs,

on pose

$$P_i = b_{i-1}x ; i = 1, \dots, n$$

on a alors

$$b_i = P_i + a_i$$

$$\delta P_i = x \delta b_{i-1} + P_i \epsilon_i$$

$$\delta b_i = \delta P_i + b_i \eta_i = x \delta b_{i-1} + P_i \epsilon_i + b_i \eta_i$$

où les ϵ_i , η_i sont des nombres de module majoré par τ , $i = 1, \dots, n$

donc pour l'erreur relative, on a :

$$\frac{\delta b_i}{b_i} = \eta_i + \frac{P_i}{b_i} \left(\frac{\delta b_{i-1}}{b_{i-1}} + \epsilon_i \right)$$

$$\text{on a } \frac{\delta b_0}{b_0} = 0$$

$$\frac{\delta b_1}{b_1} = \eta_1 + \frac{b_0 x}{b_1} \epsilon_1$$

$$\frac{\delta b_2}{b_2} = \eta_2 + \frac{b_1 x}{b_2} \eta_1 + \frac{b_0 x^2}{b_2} \epsilon_1 + \frac{b_1 x}{b_2} \epsilon_2$$

$$\frac{\delta b_3}{b_3} = \eta_3 + \frac{b_1 x^2}{b_3} \eta_1 + \frac{b_2 x}{b_3} \eta_2 + \frac{b_0 x^3}{b_3} \epsilon_1 + \frac{b_1 x^2}{b_3} \epsilon_2 + \frac{b_2 x}{b_3} \epsilon_3$$

$$\frac{\delta b_n}{b_n} = \sum_{i=1}^n \frac{b_i}{b_n} x^{n-i} \times \eta_i + \sum_{i=1}^n \frac{b_{i-1}}{b_n} x^{n-i+1} \times \epsilon_i$$

et le nombre de stabilité en borne est :

$$C_n = \sum_{i=1}^n \left| \frac{b_i}{b_n} \right| |x^{n-i}| + \sum_{i=1}^n \left| \frac{b_{i-1}}{b_n} \right| |x^{n-i+1}|$$

$$C_n = \sum_{j=0}^{n-1} \left| \frac{b_{n-j}}{b_n} \right| |x^j| + \sum_{j=1}^n \left| \frac{b_{n-j}}{b_n} \right| |x^j|$$

posons

$$\tilde{P}(x) = |b_0| x^n + |b_1| x^{n-1} + \dots + |b_n|$$

$${}^2P(x) = b_0^2 x^n + b_1^2 x^{n-1} + \dots + b_n^2$$

on a alors pour le nombre de stabilité en borne

$$C_n = \frac{1}{|b_n|} (2 \tilde{P}(|x|) - |a_0 x^n|) - 1$$

et pour le nombre de stabilité en moyenne

$$\gamma^2 = \sum_{j=0}^{n-1} \left(\frac{b_{n-j} x^j}{b_n} \right)^2 + \sum_{j=1}^n \left(\frac{b_{n-j} x^j}{b_n} \right)^2$$

$$\gamma^2 = \frac{1}{b_n^2} (2 {}^2P(x^2) - a_0^2 x^{2n}) - 1$$

$$\gamma = \sqrt{\frac{1}{b_n^2} (2 {}^2P(x^2) - a_0^2 x^{2n}) - 1}$$

Pour pouvoir faire une comparaison avec le schéma classique, calculons les deux nombres de stabilité sur un exemple simple.

$$P_n(x) = 1 + x + \dots + x^n, \text{ donc } a_i = 1 ; i = 0, \dots, n$$

on calcule les deux nombres de stabilité pour $x = 1$

on a :

$$C_n = \frac{1}{|b_n|} (2 \tilde{P}(|x|) - |a_0 x^n|) - 1$$

Pour le calcul de $\tilde{P}(|x|)$, on a

$$b_0 = 1$$

$$b_1 = 2$$

$$b_2 = 3$$

$$\vdots$$

$$b_n = n+1$$

$$\text{et } \tilde{P}(|x|) = 1 + 2 + \dots + n+1 = \sum_{i=1}^{n+1} i$$

$$\text{Comme } \sum_{i=1}^n i = \frac{(n+1)(n+2)}{2}, \text{ on a alors}$$

$$\tilde{P}(1) = \frac{(n+1)(n+2)}{2} + n+1$$

d'où

$$C_n = \frac{1}{n+1} [(n+1)(n+2) + 2(n+1) - 1] - 1$$

$$C_n = n+3 - \frac{1}{n+1} \approx n+3$$

d'autre part

$${}^2P(1) = 1 + 2^2 + 3^2 + \dots + (n+1)^2 = \sum_{i=1}^n i^2 + (n+1)^2$$

$$\text{or } \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\text{donc } {}^2P(1) = \frac{n(n+1)(2n+1)}{6} + (n+1)^2$$

et

$$\gamma = \sqrt{\frac{1}{(n+1)^2} \left[\frac{n(n+1)(2n+1)}{3} + 2(n+1)^2 - 1 \right] - 1}$$

$$\gamma = \sqrt{\frac{n(2n+1)}{3(n+1)} + 1 - \frac{1}{(n+1)^2}} \approx \sqrt{\frac{2}{3}} n$$

1.5 - Calcul de la valeur d'un polynôme par le schéma classique

Pour calculer la valeur du polynôme

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_n$$

Le schéma classique consiste à former la suite :

$$\begin{cases} x_1 = x \\ S_0 = a_n \\ S_i = S_{i-1} + a_{n-i} x_i \\ x_{i+1} = x_i \cdot x \end{cases} \quad \text{pour } i = 1, \dots, n$$

on a $P_n(x) = S_n$

Comme dans le cas précédent, on suppose les données exactes. La propagation des erreurs de calcul s'effectue de la façon suivante ;

$$\frac{\delta x_{i+1}}{x_{i+1}} = \frac{\delta x_i}{x_i} + \frac{\delta x}{x} + \epsilon_{i+1}, \text{ comme } \frac{\delta x}{x} = 0, \text{ on a alors}$$

$$\frac{\delta x_{i+1}}{x_{i+1}} = \frac{\delta x_i}{x_i} + \epsilon_{i+1}, \text{ ce qui nous donne pour } i = 1, \dots, n$$

$$\frac{\delta x_2}{x_2} = \epsilon_2$$

$$\frac{\delta x_3}{x_3} = \epsilon_2 + \epsilon_3$$

- - - - -

$$\frac{\delta x_n}{x_n} = \epsilon_2 + \epsilon_3 + \dots + \epsilon_n$$

Posons $d_i = a_{n-i} x_i$, on a

$$\frac{\delta d_i}{d_i} = \frac{\delta a_{n-i}}{a_{n-i}} + \frac{\delta x_i}{x_i} + \eta_i, \text{ d'où comme } \frac{\delta a_i}{a_i} = 0 \quad \forall i$$

$$\frac{\delta d_i}{d_i} = \eta_i + \sum_{j=2}^i \epsilon_j$$

par ailleurs, on a $S_i = S_{i-1} + d_i$, $i = 1, \dots, n$ donc

$$\delta S_i = \delta S_{i-1} + \delta d_i + S_i \nu_i$$

et on a alors

$$\delta S_0 = 0$$

$$\delta S_1 = \delta d_1 + S_1 \nu_1$$

$$\delta S_2 = \delta d_1 + \delta d_2 + S_1 v_1 + S_2 v_2$$

$$- \quad - \quad - \quad - \quad - \quad - \quad -$$

$$\delta S_i = \sum_{j=1}^i (\delta d_j + S_j v_j)$$

$$- \quad - \quad - \quad - \quad - \quad - \quad -$$

$$\delta S_n = \sum_{j=1}^n (\delta d_j + S_j v_j)$$

$$\frac{\delta S_n}{S_n} = \frac{1}{S_n} \left[\sum_{j=1}^n d_j \eta_j + \sum_{j=2}^n \sum_{k=2}^j d_j \epsilon_k + \sum_{j=1}^n S_j v_j \right]$$

$$\frac{\delta S_n}{S_n} = \frac{1}{S_n} \left[\sum_{j=1}^n d_j \eta_j + \sum_{k=2}^n \epsilon_k \sum_{j=k}^n d_j + \sum_{j=1}^n S_j v_j \right]$$

où les η_j , v_j ; $j = 1, \dots, n$ et les ϵ_k ; $k = 2, \dots, n$ sont des nombres de module majoré par τ .

On calcule alors le nombre de stabilité en borne qui est :

$$C_n = \frac{1}{|S_n|} \left(\sum_{j=1}^n |d_j| + \sum_{k=2}^n \left| \sum_{j=k}^n (d_j) \right| + \sum_{j=1}^n \left| \sum_{k=0}^j d_k \right| \right)$$

et pour le nombre de stabilité en moyenne, on a

$$\gamma = \sqrt{\frac{1}{S_n^2} \left[\sum_{j=1}^n (d_j)^2 + \sum_{k=2}^n \left(\sum_{j=k}^n d_j \right)^2 + \sum_{j=1}^n \left(\sum_{k=0}^j d_k \right)^2 \right]}$$

Pour avoir un ordre de grandeur de ces nombres de stabilité, reprenons le même exemple simple.

$$P_n(x) = 1 + x + \dots + x^n, \text{ pour la valeur de } x = 1$$

$$\text{on a } S_n = \sum_{k=0}^n d_k = n+1, \text{ et}$$

$$\begin{aligned} C_n &= \frac{1}{n+1} \left[n + \sum_{k=2}^n (n-k+1) + \sum_{j=1}^n (j+1) \right] \\ &= \frac{1}{n+1} \left[n + (n-1)(n+1) - \frac{(n+1)(n+2)}{2} + 1 + \frac{(n+1)(n+2)}{2} \right] \end{aligned}$$

$$C_n \approx n+1$$

pour le nombre de stabilité en moyenne γ , on a

$$\gamma^2 = \frac{1}{(n+1)^2} \left[n + \sum_{k=2}^n (n-k+1)^2 + \sum_{j=1}^n (j+1)^2 \right]$$

$$\gamma^2 = \frac{1}{(n+1)^2} \left[n + (n-1)(n+1)^2 + \sum_{k=2}^n k^2 - 2(n+1) \sum_{k=2}^n k + \sum_{j=1}^n j^2 + 2 \sum_{j=1}^n j + n \right]$$

$$\gamma^2 \approx \frac{2}{3} n, \text{ donc } \gamma \approx \sqrt{\frac{2}{3} n}$$

V.2 - MODELISATION DE LA SIMULATION

Les expériences numériques ont été faites sur l'ordinateur IRIS 80 de CII HB. Tous les calculs sont faits en double précision. Comme sur cet ordinateur l'arithmétique à virgule flottante normalisée est disponible en double précision avec une longueur de mantisse limitée à $m = 14$, la machine travaillant en base $b = 16$. Nous nous sommes limité au cas où la mantisse correspond à une précision comprise entre 3 et 15 chiffres décimaux. La limitation supérieure nous est imposée par le fait que nous avons profité de l'arithmétique double précision pour bâtir notre simulation d'arithmétique flottante.

2.1 - Les opérations arithmétiques de base

Nous allons définir dans cette partie, les opérations de base utilisées dans notre simulation d'arithmétique flottante d'arrondi. Tout nombre réel positif X représenté en virgule flottante normalisée s'écrit dans une base quelconque b :

$$X = M \times b^E$$

Dans la pratique, on se restreint aux nombres dont l'écriture en base b nécessite m chiffres au plus : on a alors

$$X = M' \times b^{E-m}, M' \in [b^{m-1}, b^m[$$

Cette valeur est représentée en machine par :

$$\bar{X} = \bar{M}' \times b^{E-m}$$

Dans notre simulation d'arithmétique, $\bar{M}'$ est la mantisse limitée à m chiffres, arrondie avec défaut sur $\frac{1}{2}$ dans la base b. Nous noterons désormais tout nombre

$\bar{X}$ arrondi par :

$$\bar{X} = M_X \times b^{E_X}, \text{ où}$$

M_X est la mantisse arrondie avec défaut sur $\frac{1}{2}$ à m chiffres.

E_X l'exposant entier

b la base.

2.1.1 - Addition et Soustraction de deux nombres arrondis $\bar{X}$ et $\bar{Y}$

Pour la somme de deux nombres $\bar{X}$ et $\bar{Y}$ arrondis, deux cas se présentent

1 - Cas où les deux nombres arrondis ont des exposants différents

$$\bar{X} = M_X b^{E_X}, M_X \in [b^{m-1}, b^m[$$

$$\bar{Y} = M_Y b^{E_Y}, M_Y \in [b^{m-1}, b^m[$$

a) Si $E_X < E_Y$

On cherche la mantisse non normalisée dans la b de $S = \bar{X} + \bar{Y}$

$a_1 = +1$, on a affaire à une addition

$a_1 = -1$, on a affaire à une soustraction

Calculons la mantisse non normalisée dans la base b de la somme de $\bar{X}$ et de $\bar{Y}$

M_N = mantisse non normalisée

$$M_N = M_X \cdot b^{(E_X - E_Y)} + \text{Signe de } (a_1) \times M_Y$$

on normalise ensuite la mantisse M_N pour l'écrire sous la forme

$$M_N = M_A b^M, M_A \in [b^{m-1}, b^m[.$$

La somme de $\bar{X}$ et $\bar{Y}$ est alors donnée dans ce cas par :

$$S = M_S \cdot b^{E_S}, \text{ avec}$$

$$M_S = M_A$$

$$\text{et } E_S = E_M + E_Y$$

b) Si $E_Y \leq E_X$

On cherche la mantisse non normalisée dans la base b

$$M_N = M_X + \text{Signe de } (a_1) \cdot M_Y \times b^{(E_Y - E_X)}$$

On normalise la mantisse M_N pour l'écrire sous la forme $M_N = M_A b^{E_M}$,

et la somme est telle que

$$S = M_S \cdot b^{E_S}, \text{ avec } M_A \in [b^{m-1}, b^m[$$

$$M_S = M_A \text{ et } E_S = E_X + E_M$$

2 - Cas où les deux nombres arrondis $\bar{X}$ et $\bar{Y}$ ont les mêmes exposants

On cherche la mantisse non normalisée

$$M_N = M_X + \text{Signe de } (a_1) \times M_Y$$

on normalise ensuite la mantisse dans la base b , pour l'écrire sous la forme $M_N = M_A b^{E_M}$.

La somme de X et Y est donnée par

$$S = M_S \times b^{E_S} \text{ avec } M_S \in [b^{m-1}, b^m[$$

$$M_S = M_A, \text{ et } E_S = E_X + E_M$$

Pour l'addition on posera $a_1 = +1$ et pour la soustraction on posera $a_1 = -1$

2.1.2 - La multiplication

Le produit P de deux nombres arrondis avec défaut $\bar{X}$ et $\bar{Y}$ est

défini par :

$$\text{Si } \bar{X} = M_X b^{E_X}, \bar{Y} = M_Y b^{E_Y}$$

on cherche la mantisse non normalisée du produit.

$$M_N = M_X \times M_Y$$

On normalise la mantisse dans la base b pour l'écrire sous la forme

$$M_N = M_A b^{E_M}, \text{ et le produit } P = M_P \cdot b^P \text{ est tel que}$$

$$M_P = M_A, \quad M_A \in [b^{m-1}, b^m[$$

$$E_P = E_X + E_Y + E_M$$

On fait appel à la phase de normalisation chaque fois qu'une mantisse M est telle que $M \notin [b^{m-1}, b^m[$.

2.1.3 - La division

Soient $\bar{X} = M_X b^{E_X}$ et $\bar{Y} = M_Y b^{E_Y}$, deux nombres arrondis avec défaut sur $\frac{1}{2}$, le quotient de $\bar{X}$ par $\bar{Y}$ est défini par :

$$Q = M_q \times b^{E_q}$$

tel que $M_q = M_A$

$$E_q = E_X - E_Y + E_M$$

où M_A est la mantisse normalisée dans la base b de $M_N = \frac{M_X}{M_Y}$, et E_M l'exposant dans la forme

$$M_N = M_A \times b^{E_M}, \quad M_A \in [b^{m-1}, b^m[$$

2.2 - Mode d'emploi des sous-programmes

2.2.1 - Les sous-programmes de normalisation et de sortie des résultats

*** CREATION ***

Appel : CALL CREATION (Z, AM, L)

Paramètre en entrée : Z

Paramètres de sortie : AM, L

Description : ce sous-programme calcule la mantisse normalisée AM et l'exposant L du nombre Z sous la forme $Z = AM * B ** L$

*** SORTIE ***

Appel : CALL SORTIE (Z, AM, L, X)

Paramètres en entrée : Z, AM, L

Paramètre de sortie : X

Description : ce sous-programme donne l'arrondi avec défaut sur $\frac{1}{2}$ du nombre Z.

2.2.2 - Les sous-programmes utilisateurs

Nous présentons ici l'ensemble des sous-programmes concernant les quatre opérations arithmétiques.

*** SOMME ***

Appel : CALL SOMME (IT, AM1, LX1, AM2, LX2, AM, LS)

Paramètres en entrée : IT, AM1, LX1, AM2, LX2

Paramètres de sortie : AM, LS

Sous-programme utilisé : CREATION

Description : Ce sous-programme fait l'addition ou la soustraction des mantisses AM_1 et AM_2 de deux nombres X et Y, tout en pondérant les mantisses suivant que les exposants de X et Y sont ou ne sont pas égaux.

IT = +1, on fait une addition

IT = -1, on fait une soustraction.

*** MULTIP ***

Appel : CALL MULTIP (AM1, LX, AM2, LY, MA, LP)

Paramètres en Entrée : AM1, LX, AM2, LY

Paramètres de sortie : AM, LP

Sous-programme utilisé : CREATION

Description : Ce sous-programme effectue la multiplication des mantisses AM1 et AM2 de X et Y, et fait la somme des exposants.

*** DIVISION ***

Appel : CALL DIVISION (AM1, LX1, AM2, LX2, AM, LQ)

Sous-programme utilisé : CREATION

Description : Ce sous-programme fait la division de la mantisse AM1 par la mantisse AM2 des nombres X1 et X2 et effectue la différence des exposants.

Notations

$\text{PREC}(M, B) = |(\text{erreur relative sur le résultat})|/\rho_2$

B = base

M = nombre de chiffres de la mantisse

$\text{PMOY}(B)$ = moyenne des $\text{PREC}(M, B)$.

V.3 - ANALYSE DES RESULTATS

3.1 - Calcul d'une somme

Présentation des résultats

Nous avons calculé la somme $\sum_{i=1}^n a_i$, avec $a_i = 0,333$, $i = 1, \dots, N$ pour différentes valeurs de N , ceci par les deux méthodes suivantes :

- La méthode de LINZ
- La méthode du cumul simple

Le tableau I donne pour chaque méthode, le nombre de termes N de la somme, le nombre de stabilité en moyenne Gamma, et le nombre de stabilité en borne C_n . Les tableaux II et III représentent pour chacune des deux méthodes les rapports $\text{PREC}(M, B)$. L'échantillon de bases choisies va de la base 2 à la base 16. Ces tableaux nous montrent pour chaque base et pour une longueur de mantisse donnée, les rapports $\text{PREC}(M, B)$ qui constituent les différentes colonnes. Au bas de chaque tableau, on note les valeurs de $\text{PMOY}(B)$. Les tableaux II_a et III_b présentent les valeurs de $\text{PMOY}(B)$ pour les deux méthodes, avec le nombre de termes N correspondant, et la base B correspondante.

Méthode de LINZ			Cumul simple	
N	Gamma	C _n	Gamma	C _n
16	1,4142	4	2,417	8
32	1,4142	5	3,2660	16
64	1,4142	6	4,618	32
128	1,4142	7	6,570	64
512	1,4142	9	13,06	128

Tableau I

Calcul du nombre de stabilité en moyenne Gamma pour la méthode de LINZ,
et pour la méthode du cumul simple.

Calcul du nombre de stabilité en borne pour les deux méthodes.

Tableau II
Méthode de LINZ

N	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3	***	***	***	***	***	***	***	***	***	***	1.24	1.94	.99	2.35	1.62 .00
4	***	***	***	***	.90	2.04	.00	1.30	2.80	1.92	.48	3.83	1.60	.94	1.53
5	***	***	***	.06	1.25	.55	.00	.61	.00	2.83	.48	3.01	.00	5.48	.00
6	***	***	***	.43	2.91	.09	.00	.46	.00	.36	3.39	.57	3.32	5.48	1.53
7	***	***	.00	1.43	.62	2.22	.00	4.50	.00	.23	1.94	3.10	.75	5.48	1.53
8	***	1.08	.00	.43	4.75	4.17	2.67	4.57	.00	.65	.48	3.01	1.82	5.48	3.07
9	***	1.69	.00	1.43	2.91	.56	2.67	4.50	.00	1.17	.00	1.08	4.06	5.48	1.53
10	***	.27	.00	.43	1.66	2.22	2.67	5.19	.00	3.15	1.45	.23	2.46	5.48	1.53
11	.00	.31	.00	1.43	2.29	4.17	2.67	.84	.00	1.35	3.39	3.58	2.25	5.49	1.53
12	.00	1.02	.00	.48	3.54	2.75	2.67	.08	.00	2.11	3.39	.86	1.83	***	***
13	.00	1.25	.00	1.43	3.54	1.30	2.67	1.93	.00	2.43	***	***	***	***	***
14	.00	1.93	.00	.43	.21	.18	.00	3.27	.03	***	***	***	***	***	***
15	.00	2.34	.00	1.43	1.25	.09	2.67	***	***	***	***	***	***	***	***
16	.00	.17	.00	.43	.42	.65	***	***	***	***	***	***	***	***	***
17	.00	1.25	.00	1.43	4.78	***	***	***	***	***	***	***	***	***	***
18	.00	.30	.00	.43	1.69	***	***	***	***	***	***	***	***	***	***
19	.00	.74	.00	1.43	***	***	***	***	***	***	***	***	***	***	***
20	.00	.40	.00	.48	***	***	***	***	***	***	***	***	***	***	***
21	.00	.62	.00	***	***	***	***	***	***	***	***	***	***	***	***
22	.00	.24	.00	***	***	***	***	***	***	***	***	***	***	***	***
23	.00	1.09	.00	***	***	***	***	***	***	***	***	***	***	***	***
24	.00	1.65	.30	***	***	***	***	***	***	***	***	***	***	***	***
25	.00	.33	***	***	***	***	***	***	***	***	***	***	***	***	***
26	.00	2.80	***	***	***	***	***	***	***	***	***	***	***	***	***
27	.00	3.59	***	***	***	***	***	***	***	***	***	***	***	***	***
28	.00	3.04	***	***	***	***	***	***	***	***	***	***	***	***	***
29	.00	2.21	***	***	***	***	***	***	***	***	***	***	***	***	***
30	.00	.31	***	***	***	***	***	***	***	***	***	***	***	***	***
31	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
32	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
33	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
34	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
35	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
36	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
37	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
38	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
39	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
40	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
41	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
42	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
43	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
44	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
45	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
46	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
47	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
48	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
49	.00	***	***	***	***	***	***	***	***	***	***	***	***	***	***
50	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Phcy: .00 1.22 .00 .87 2.12 1.64 1.56 2.48 .26 1.59 1.70 2.03 2.04 4.55 1.36															



Tableau III

Cumul simple

[illegible]

base \ N	16	32	64	128	512
	PMOY	PMOY	PMOY	PMOY	PMOY
2	0	0	0	0	0
3	0,95	1,13	1,13	1,22	1,92
4	0	0	0	0	0
5	0,95	0,95	0,95	0,87	0,98
6	1,22	1,64	1,64	2,12	2,12
7	1,25	1,64	1,64	1,64	1,62
8	1,56	1,56	1,56	1,56	1,56
9	1,44	2,48	2,48	2,48	2,41
10	0	0,26	0,26	0,26	0,58
11	1,51	1,51	1,59	1,59	2,02
12	1,55	1,55	1,70	1,70	1,58
13	1,73	1,73	2,03	2,03	3,29
14	1,37	1,37	2,04	2,04	2,04
15	2,20	2,70	4,55	4,55	4,55
16	1,36	1,36	1,36	1,36	1,36

Tableau IIa

Méthode de LINZ

N base	16	32	64	128	512
	PMOY	PMOY	PMOY	PMOY	PMOY
2	1,51	1,51	1,51	1,51	1,51
3	1,93	3,56	3,56	3,56	3,56
4	3,04	3,54	9,04	9,04	9,04
5	1,25	2,32	5,62	5,75	5,75
6	1,25	2,23	6,75	14,21	14,21
7	2,11	2,81	7,39	19,39	19,68
8	1,39	3,12	5,12	13,14	26,15
9	2,15	2,88	9,19	18,83	30,74
10	0	0,39	0,20	0,40	6,76
11	1,82	4,21	3,24	10,78	54,52
12	1,98	4,34	4,93	10,22	46,95
13	1,26	5,66	6,69	12,16	78,35
14	1,11	5,41	4,32	8,15	86,53
15	1,71	2,38	10,03	24,25	34,91
16	1,82	3,99	7,22	12,39	62,63

Tableau IIIb

Cumul simple

Analyse des résultats

Pour la méthode de LINZ, le nombre de stabilité en moyenne GAMMA est indépendant du nombre de termes de la somme. Nous obtenons de très bon résultats pour cette méthode, car les valeurs de PREC (M, B) et de PMOY(B) se distribuent bien autour de Gamma. Cette méthode est plus stable que la méthode du cumul simple comme le montrent les tableaux II_a et III_b. Le nombre de stabilité en moyenne Gamma pour la méthode de cumul dépend du nombre de termes N de la somme, et plus ce nombre N est grand, plus la valeur de Gamma augmente. L'erreur de calcul peut devenir très importante dans une somme par cumul. Car on est amené au bout d'un certain temps à faire la somme de deux nombres dont l'un, cumul des termes précédents est beaucoup plus grand que le terme ajouté. L'erreur relative augmente régulièrement avec N. Les essais numériques nous ont montré que selon la valeur de la base B et de la longueur de la mantisse rencontrées, il existe un nombre N₀ tel que pour N ≥ N₀, le résultat de la somme reste stationnaire.

Exemples :

Pour la base 10, pour N ≥ 331, le résultat de la somme reste toujours égal à 100, pour la valeur de M = 3. Pour la base 2, pour N ≥ 22, le résultat est stationnaire à 8, pour la valeur de M = 4. En base 6, avec N ≥ 648, le résultat reste toujours égal à 216, pour m = 4.

3.2 - Calcul de la valeur d'un polynôme

Présentation des résultats

Le calcul de la valeur du polynôme $\sum_{i=0}^N a_i x^{n-i}$ est fait pour les valeurs suivantes :

$$a_i = 0,999, \text{ pour } i = 1, \dots, N$$

$$x = 0,999$$

$$N = 30$$

Les méthodes de calcul utilisées sont : le schéma classique et la méthode de Horner. Le tableau VI regroupe pour les deux méthodes, les nombres de stabilité en moyenne et en borne. Les tableaux VII et VIII rassemblent pour chacune des deux méthodes, les rapports $\text{PREC}(M, B)$. Le Tableau VII_a reprend les valeurs de $\text{PMOY}(B)$ pour les deux méthodes de calcul.

Méthode de Horner		Schéma classique	
Gamma	C_n	Gamma	C_n
4,5295	33	4,4466	31

Tableau VI

Calcul des nombres de stabilité en moyenne et en borne pour le

polynôme $\sum_{i=0}^N a_i x^{n-i}$

bases	Méthode de Horner	Schéma Classique
	PMOY	PMOY
2	9,79	3,31
3	8,76	2,52
4	5,99	4,55
5	6,62	3,74
6	4,38	1,12
7	4,39	1,38
8	2,69	1,50
9	4,51	1,38
10	0,74	3,30
11	2,55	2,78
12	2,34	4,92
13	2,55	4,42
14	2,04	4,40
15	1,80	4,20
16	4,42	6,50

Tableau VIIa

Tableau VII
Méthode de HORNER

[illegible]

Tableau VIII
Schéma classique

M	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3	***	***	***	***	***	***	***	***	***	8.47	17.98	16.29	20.48	18.27	24.61
4	***	***	***	***	1.61	.36	1.92	.39	3.54	1.57	3.41	1.70	3.39	1.54	1.47
5	***	***	***	10.24	.15	.13	.24	.06	11.29	1.25	1.46	2.99	3.22	2.42	9.96
6	***	***	3.03	2.93	.45	1.83	.45	.07	3.54	1.33	2.35	5.95	.30	5.71	5.33
7	***	3.68	5.89	2.96	1.05	1.51	3.07	3.05	1.38	2.17	.09	2.27	1.24	.88	2.06
8	***	5.69	20.72	5.23	1.27	3.17	1.05	.83	1.53	.22	5.60	1.35	1.17	5.41	6.03
9	***	4.01	3.88	5.07	1.35	2.61	1.68	3.07	3.02	.60	4.93	.78	.56	2.22	.06
10	***	.13	11.53	3.62	2.16	.10	.87	.97	.54	.53	4.44	.74	9.76	.93	2.05
11	.07	4.61	3.43	1.09	.12	1.42	1.31	1.29	.53	6.18	5.04	10.17	.20	.38	6.91
12	2.33	.16	3.04	.79	1.49	1.71	1.76	3.46	5.00	1.39	3.96	2.00	5.67	***	***
13	7.13	1.01	.25	7.12	.02	.08	2.88	1.11	5.69	6.83	***	***	***	***	***
14	9.32	2.01	.24	5.77	.03	.03	.52	.85	.29	***	***	***	***	***	***
15	16.55	2.06	4.80	2.61	.24	3.09	3.29	***	***	***	***	***	***	***	***
16	.65	6.57	7.42	2.05	2.77	1.89	***	***	***	***	***	***	***	***	***
17	3.58	3.31	.54	.89	2.52	***	***	***	***	***	***	***	***	***	***
18	6.13	6.86	6.03	4.64	1.68	***	***	***	***	***	***	***	***	***	***
19	14.11	.52	1.81	1.11	***	***	***	***	***	***	***	***	***	***	***
20	1.03	2.65	.27	4.58	***	***	***	***	***	***	***	***	***	***	***
21	4.22	.55	2.25	***	***	***	***	***	***	***	***	***	***	***	***
22	1.98	1.94	4.86	***	***	***	***	***	***	***	***	***	***	***	***
23	4.52	.06	5.24	***	***	***	***	***	***	***	***	***	***	***	***
24	.13	1.91	1.17	***	***	***	***	***	***	***	***	***	***	***	***
25	.42	1.53	***	***	***	***	***	***	***	***	***	***	***	***	***
26	2.07	2.34	***	***	***	***	***	***	***	***	***	***	***	***	***
27	7.92	3.26	***	***	***	***	***	***	***	***	***	***	***	***	***
28	.33	1.89	***	***	***	***	***	***	***	***	***	***	***	***	***
29	.20	1.78	***	***	***	***	***	***	***	***	***	***	***	***	***
30	2.30	1.83	***	***	***	***	***	***	***	***	***	***	***	***	***
31	2.59	***	***	***	***	***	***	***	***	***	***	***	***	***	***
32	1.85	***	***	***	***	***	***	***	***	***	***	***	***	***	***
33	1.46	***	***	***	***	***	***	***	***	***	***	***	***	***	***
34	4.03	***	***	***	***	***	***	***	***	***	***	***	***	***	***
35	.43	***	***	***	***	***	***	***	***	***	***	***	***	***	***
36	2.20	***	***	***	***	***	***	***	***	***	***	***	***	***	***
37	1.54	***	***	***	***	***	***	***	***	***	***	***	***	***	***
38	2.03	***	***	***	***	***	***	***	***	***	***	***	***	***	***
39	2.69	***	***	***	***	***	***	***	***	***	***	***	***	***	***
40	2.04	***	***	***	***	***	***	***	***	***	***	***	***	***	***
41	2.93	***	***	***	***	***	***	***	***	***	***	***	***	***	***
42	1.09	***	***	***	***	***	***	***	***	***	***	***	***	***	***
43	1.13	***	***	***	***	***	***	***	***	***	***	***	***	***	***
44	2.20	***	***	***	***	***	***	***	***	***	***	***	***	***	***
45	3.64	***	***	***	***	***	***	***	***	***	***	***	***	***	***
46	1.60	***	***	***	***	***	***	***	***	***	***	***	***	***	***
47	3.59	***	***	***	***	***	***	***	***	***	***	***	***	***	***
48	3.09	***	***	***	***	***	***	***	***	***	***	***	***	***	***
49	3.70	***	***	***	***	***	***	***	***	***	***	***	***	***	***
50	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***

Préc: 5.51 2.22 4.55 5.74 1.12 1.38 1.50 1.36 3.30 2.78 4.92 4.42 4.40 4.20 6.50



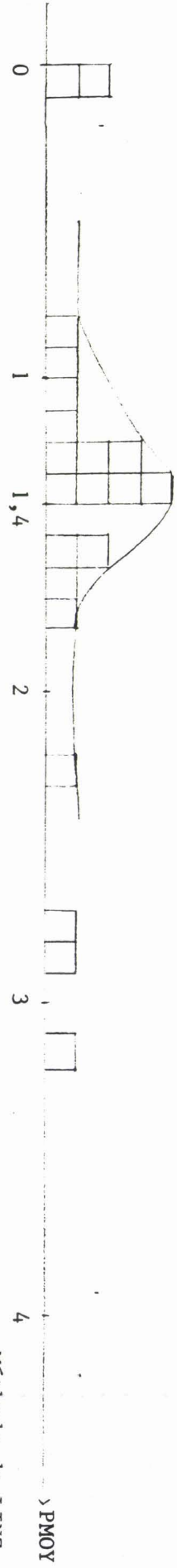


figure a

Méthode de LINZ
N = 16

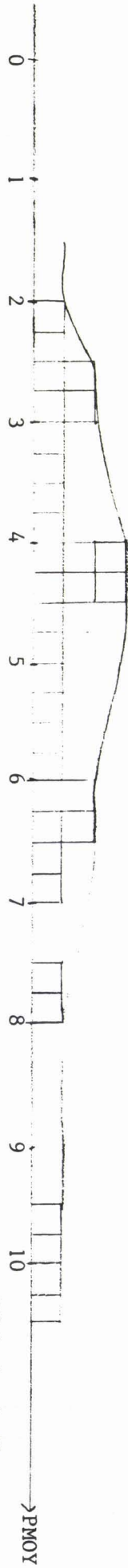


figure b₂

Méthode de Horner
pour $\sum_{i=0}^N a_i x^{n-i}$

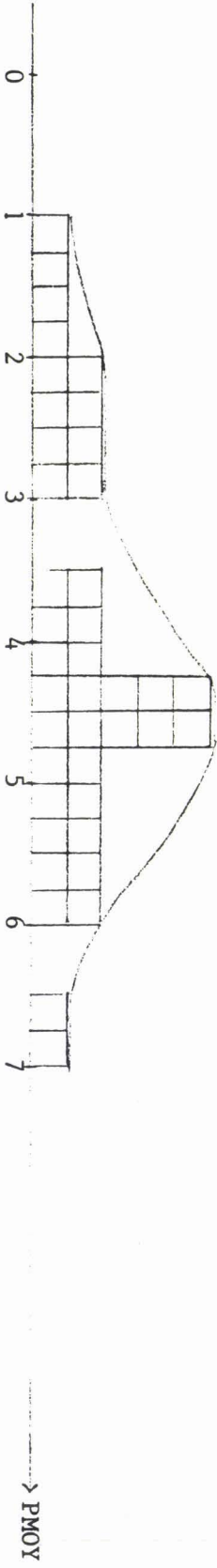


figure b₁

Schéma classique



Analyse des résultats

Les deux schémas de calcul : schéma classique et schéma de Horner sont stables dans les deux cas. Les nombres de stabilité en moyenne obtenus pour les deux méthodes sont presque du même ordre de grandeur. Pour cet exemple, le nombre de stabilité en moyenne pour le schéma de calcul de la méthode classique est légèrement plus faible que celui du schéma de Horner. La méthode de Horner donne ici des résultats moins bons par rapport au schéma classique. C'est ce que nous montrent les figures b_1 et b_2 .

3.3 - Calcul de la valeur du polynôme

$$P(x) = 12192x^3 - 32257x^2 - 85344x + 225799$$

pour $x = 2,645752$.

Pour cet exemple, la méthode de Horner paraît meilleure que le schéma classique tant du point de vue du nombre de stabilité que celui des résultats effectifs. Toutefois, on observe que le nombre de stabilité en moyenne est particulièrement pessimiste. Ce phénomène qu'on peut attribuer à la présence de la cancellation sera expliqué dans l'exemple suivant.

Le tableau VIII_b donne le nombre de stabilité en moyenne pour les deux méthodes. Le tableau VIII_a rassemble les valeurs de PMOY(B) pour chacune des deux méthodes.

Méthode de Horner	Schéma Classique
Gamma	Gamma
0,1281 D+14	0,1653 D+14

Tableau VIII_b

Calcul du nombre de stabilité en moyenne pour le polynôme $P(x)$.



bases	Méthode de Horner		Schéma Classique	
	PMOY		PMOY	
2	0,127	D+13	0,250	D+13
3	0,159	D+13	0,290	D+13
4	0,761	D+12	0,628	D+12
5	0,135	D+13	0,167	D+13
6	0,411	D+12	0,132	D+13
7	0,897	D+12	0,541	D+12
8	0,421	D+12	0,312	D+13
9	0,617	D+12	0,827	D+12
10	0,882	D+12	0,260	D+13
11	0,169	D+13	0,236	D+13
12	0,101	D+12	0,634	D+11
13	0,180	D+12	0,349	D+12
14	0,471	D+12	0,426	D+11
15	0,555	D+12	0,883	D+13
16	0,170	D+13	0,420	D+13

Tableau VIII_a

3.4 - Calcul de l'expression du Δ^2 d'Aitken

$$a) \quad X = x_2 - \frac{(x_2 - x_1)^2}{x_0 + x_2 - 2x_1}$$

Présentation des résultats

Le calcul de cette expression a été fait avec des données x_0, x_1, x_2 qui forment une suite géométrique de la forme

$$x_n = \hat{x} + \lambda^n (x_0 - \hat{x}), \quad n = 0, 1, 2$$

où $\hat{x}$ est la valeur exacte de X , et $1 > \lambda > \frac{1}{2}$.

Trois jeux de données distinctes ont été retenus, à savoir

$$1) \begin{cases} x_0 = 0,510 \\ x_1 = 0,500 \\ x_2 = 0,491 \end{cases} \quad 2) \begin{cases} x_0 = 0,310 \\ x_1 = 0,300 \\ x_2 = 0,291 \end{cases} \quad 3) \begin{cases} x_0 = 0,210 \\ x_1 = 0,200 \\ x_2 = 0,191 \end{cases}$$

et suivant les valeurs des données, le nombre de stabilité en moyenne Gamma est donné dans le tableau IV. Dans le tableau V_a on a repris les 3 jeux de données et on a tabulé en fonction des bases, la moyenne des rapports PREC (M, B) obtenus pour les diverses longueurs de mantisse.

	Expression a	Expression b
données	Gamma	Gamma
$x_0 = 0,510$ $x_1 = 0,500$ $x_2 = 0,491$	3,12	863,14
$x_0 = 0,310$ $x_1 = 0,300$ $x_2 = 0,291$	5,85	606,81
$x_0 = 0,210$ $x_1 = 0,200$ $x_2 = 0,191$	11,06	514,97

Tableau IV

Calcul du nombre de stabilité en moyenne pour les expression a et b du Δ^2 d'aitken.



données bases	$x_0 = 0,510$ $x_1 = 0,500$ $x_2 = 0,491$	$x_0 = 0,310$ $x_1 = 0,300$ $x_2 = 0,291$	$x_0 = 0,210$ $x_1 = 0,200$ $x_2 = 0,191$
	PMOY	PMOY	PMOY
2	0,1421	0,1965	0,1458
3	0,6459	0,9881	2,4098
4	0,1355	0,1059	0,1203
5	1,7168	0,7016	2,9135
6	0,4343	1,4928	1,3498
7	0,7164	0,8564	2,0119
8	0,0687	0,2154	0,4598
9	0,4994	0,3283	1,5854
10	0	0	0
11	0,9649	1,6424	0,7027
12	1,4305	0,81790	5,3667
13	0,4793	0,5283	4,2402
14	0,8329	2,5054	1,6621
15	0,4603	0,6426	0,3592
16	0,0814	0,1594	0,2447

Tableau V_a

Analyse des résultats

On observe que les valeurs moyennes sont toujours inférieures au nombre de stabilité Gamma. Notre modèle consiste à pénaliser chaque résultat d'une erreur aléatoire systématique y compris dans le cas de la soustraction de nombres voisins alors qu'on sait que la soustraction de deux nombres voisins n'introduit que très exceptionnellement des erreurs en arithmétique flottante.

Cet exemple illustre donc le caractère pessimiste de notre modèle dans le cas des soustractions de nombres voisins. On peut toutefois y remédier en introduisant un nombre de stabilité simplifié [24].

$$b) Y = \frac{x_0 x_2 - x_1^2}{x_0 + x_2 - 2x_1}$$

Présentation des résultats et analyse

L'étude des différents résultats numériques sur cet exemple montre que selon le type de données que l'on utilise, PMOY(B) semble être une fonction décroissante par morceaux.

Le tableau IX regroupe pour différentes bases, et les différentes valeurs de M, les rapports PREC(M,B), ceci pour les données $x_0 = 0,210$; $x_1 = 0,200$; $x_2 = 0,191$. Le tableau IX_a quant à lui, regroupe les valeurs de PMOY(B), pour les bases considérées.

bases \ données	$x_0 = 0,510$ $x_1 = 0,500$ $x_2 = 0,491$	$x_0 = 0,310$ $x_1 = 0,300$ $x_2 = 0,291$	$x_0 = 0,210$ $x_1 = 0,200$ $x_2 = 0,191$
	PMOY	PMOY	PMOY
2	627,1	486,6	505,0
3	518,4	355,0	731,3
4	1035,9	665,0	271,5
5	1122,8	370,0	591,6
6	656,8	293,2	448,8
7	820,9	259,2	355,9
8	725,3	201,2	367,0
9	750,9	158,5	349,3
10	16,7	0,2	0
11	758,7	79,1	254,8
12	392,8	624,6	305,6
13	563,9	776,9	130,1
14	477,8	751,2	134,5
15	422,5	804,5	84,9
16	357,9	492,7	72,2

Tableau IX_a

Tableau IX
Expression b

	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
1	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
2	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
3	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
4	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
5	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
6	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
7	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
8	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
9	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
10	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
11	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
12	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
13	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
14	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
15	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
16	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
17	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
18	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
19	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
20	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
21	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
22	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
23	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
24	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
25	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
26	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
27	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
28	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
29	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
30	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
31	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
32	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
33	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
34	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
35	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
36	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
37	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
38	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
39	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
40	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
41	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
42	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
43	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
44	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
45	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
46	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
47	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
48	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
49	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
50	3	2	1	0	4	3	2	1	0	5	6	7	8	9	10	11	12	13	14	15	16
Play:	505.0	751.5	271.5	591.6	448.3	355.9	367.0	349.3	.0	254.8	305.6	130.1	134.5	84.9	72.2						

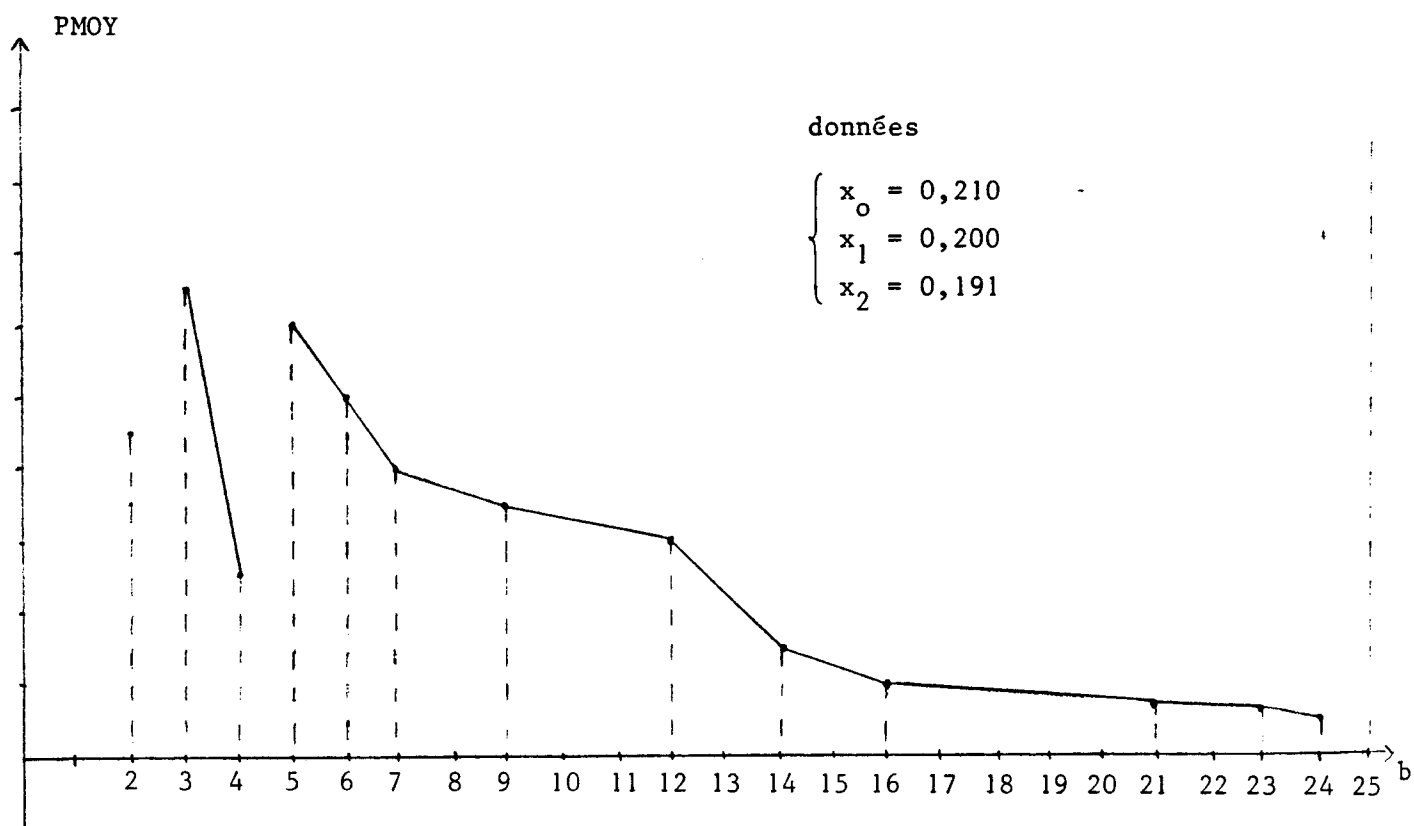


figure c

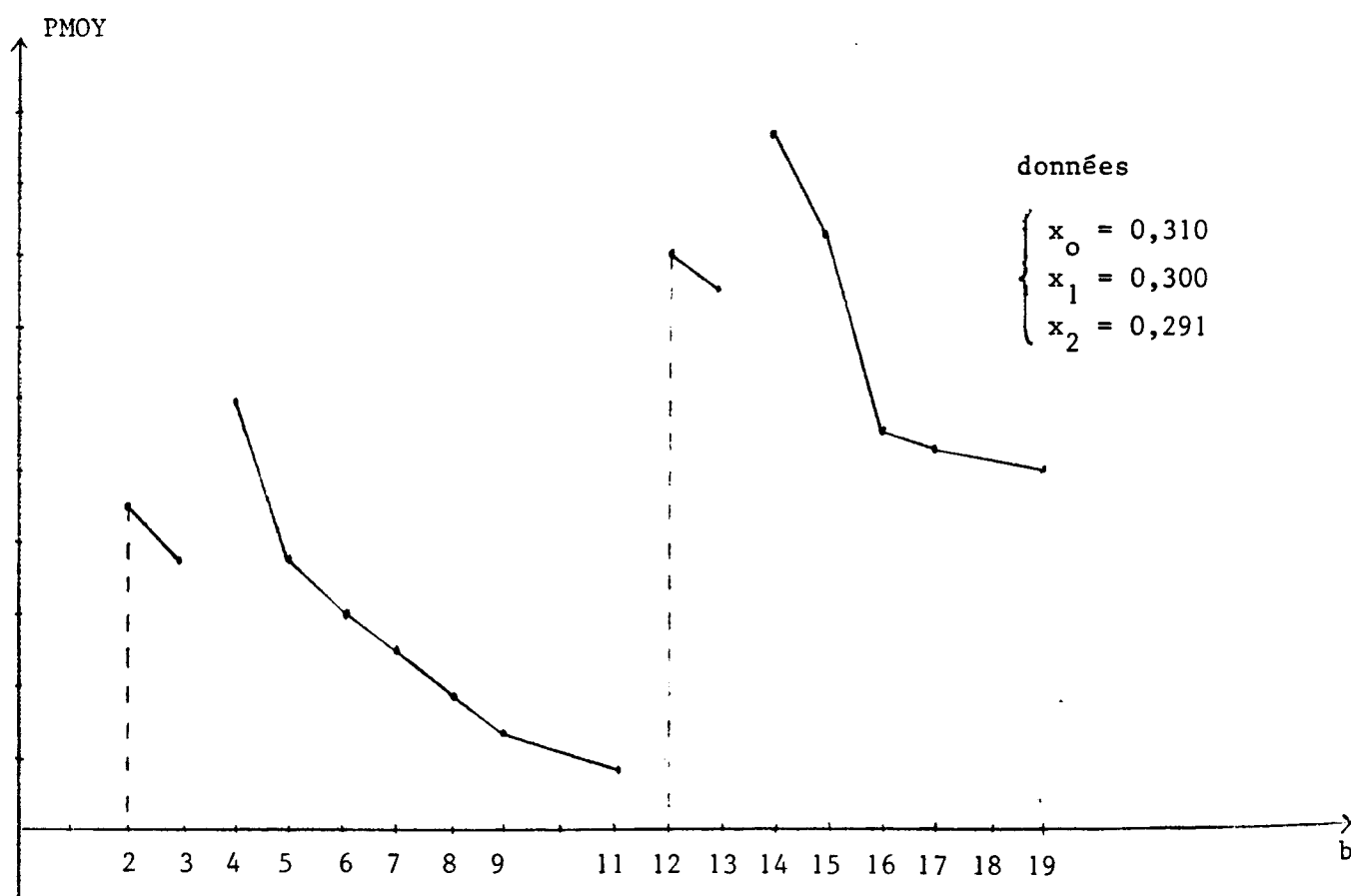


figure d



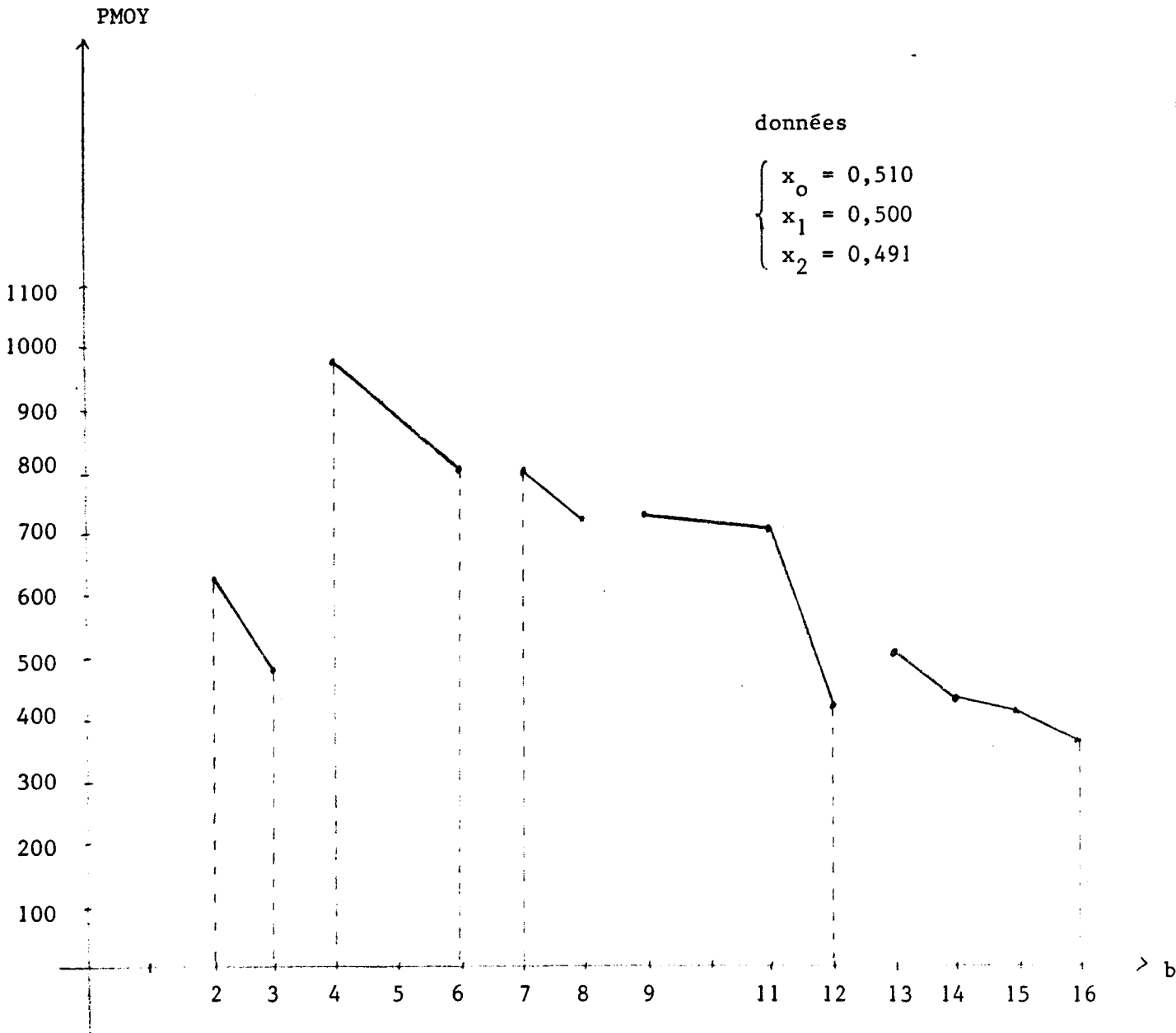


figure e

Prenons par exemple les données

$$x_0 = 0,210$$

$$x_1 = 0,200$$

$$x_2 = 0,191$$

La fonction PMOY(B) obtenue est une fonction décroissante par morceaux, les discontinuités étant situées

- 1) entre les bases 2 et 3
- 2) entre les bases 4 et 5
- 3) entre les bases 24 et 25.

Ce résultat mérite une explication.

On vérifie aisément que dans l'algorithme de calcul de cette expression, les opérations qui engendrent la plus grande erreur sont : le carré c et le produit p.

$$p = x_0 x_2$$

$$c = x_1^2$$

$$n = p - c$$

$$d_0 = x_1 - x_0$$

$$d_1 = x_2 - x_1$$

$$d = d_1 - d_0$$

$$Y = \frac{n}{d}$$

$$\text{on a en effet } \left| \frac{\partial Y}{\partial p} \right| = \left| \frac{\partial Y}{\partial c} \right| = \left| \frac{1}{d} \right|$$

or cette erreur relative engendrée par le carré ou le produit est d'autant plus grande que la mantisse de c ou de p est voisine de 1 par valeur supérieure, selon la base considérée. Il est alors aisé de voir que les discontinuités correspondent précisément aux cas où la mantisse de c ou de p est voisine de 1 par valeur supérieure.

Dans l'exemple précédent, on a $c \approx p \approx \frac{1}{25}$, et il y a changement d'exposant dans l'écriture de c pour les bases 25, 5, 3, 2, car

$$\frac{1}{25} = 0.1 \text{ en base } 25$$

$$\frac{1}{25} = 0.01 \text{ en base } 5$$

$$\frac{1}{25} \geq 0.001001 \text{ en base } 3$$

$$\frac{1}{25} \approx 0.0001 \text{ en base } 2$$

Ce phénomène se rencontre dans le calcul de cette expression, avec ce type de données. C'est ainsi que pour les données

$$x_0 = 0,310$$

$$x_1 = 0,300$$

$$x_2 = 0,291$$

on a $p \approx c \approx \frac{1}{11}$, et les discontinuités de la fonction PMOY(B) se situent ici entre les bases 3 et 4, entre les bases 11 et 12, etc ...

Pour les données $x_0 = 0,510$, $x_1 = 0,500$, $x_2 = 0,491$ on peut reprendre le même raisonnement. Les figures c, d, e montrent les courbes des fonctions PMOY(B) suivant les données.

CONCLUSION

L'approche probabiliste que nous venons de faire dans ce travail, reste une contribution à l'étude des erreurs de calcul dans la mise en oeuvre des algorithmes numériques. Cette étude probabiliste permet de compléter la politique de majoration systématique dans la recherche de l'erreur de calcul, et permet aussi de rendre compte de façon effective l'ordre de grandeur effectif de l'erreur, grâce à la définition d'un nombre de stabilité en moyenne.

La définition du nombre de stabilité en moyenne précise sur quel type arithmétique elle s'applique : le type arithmétique donnant des résultats aléatoires. L'étude théorique faite, il revenait à l'expérimentation d'infirmier ou de confirmer la validité de notre modèle arithmétique. Même s'ils ont permis d'apprécier les limites du modèle, comme dans le cas du procédé Δ^2 d'Aitken, les tests numériques confirment dans l'ensemble le bien fondé de ce modèle pour mesurer la stabilité des schémas de calcul.

ANNEXE

Nous donnons dans cette partie tous les programmes et sous-programmes utilisés pour la mise en oeuvre de notre simulation d'arithmétique d'arrondi, pour tous les exemples testés. Notre but en présentant l'ensemble de ces programmes n'est pas de vouloir fournir une bibliothèque de programmes, mais tout simplement une référence pour ceux qui voudront contrôler les résultats numériques donnés dans ce travail.



```

1. C*****
2. C  CALCUL DE LA SOMME DE N NOMBRES ; METHODE DE LINZ
3. C  SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
4. C  DELTX= ERREUR RELATIVE DU RESULTAT FINAL
5. C  LL=NOMBRE DE SIG CALCULES
6. C  PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
7. C  FINAL SUR L ECART TYPE DE L ERREUR RELATIVE D AFFECTATION
8. C  PMOY= MOYENNE DE PREC
9. C*****
10.      COMMON B,RLB,M,IS
11.      DIMENSION VAR(60,20),PREC(60,20),SIG(60,20),PMOY(20)
12.      DIMENSION V(10,650),R(10,650),A(10,650),T(10,650),L(10,650)
13. C    LECTURE DES DONNEES
14.      READ(105,22) IS
15.      22  FORMAT(I3)
16.      DO 11 KK=1,10
17.      B=FLOAT(KK+1)
18.      RLB=ALOG(B)
19.      IB=INT(R)
20.      LL=0
21.      PMOY(IB)=0.
22.      K=7
23.      N=128
24.      DO 33 M=5,60
25.      SR=B**M
26.      PT=16.**13
27.      IF(SR.GT.PT) GOTO 53
28.      WW=2*M
29.      DO 1 J=1,N
30.      V(0,J)=0.333
31.      CALL CREATION (V(0,J),T(0,J),L(0,J))
32.      CALL SORTIE (V(0,J),T(0,J),L(0,J),R(0,J))
33.      1  CONTINUE
34.      DO 3 I=1,K
35.      L1=K-I
36.      L2=(2)**L1
37.      DO 4 J=1,L2
38.      V(I,J)=R(I-1,2*J-1)+R(I-1,2*J)
39.      IT=+1
40.      CALL SOMME (IT,T(I-1,2*J-1),L(I-1,2*J-1),T(I-1,2*J),L(I-1,2*
41.      S,T(I,J),L(I,J))
42.      CALL SORTIE (V(I,J),T(I,J),L(I,J),R(I,J))
43.      4  CONTINUE
44.      3  CONTINUE
45. C*****
46. C  CALCUL DE LA VALEUR EXACTE DE LA SOMME S(N)
47. C*****
48.      DO 5 J=1,N
49.      A(0,J)=R(0,J)
50.      5  CONTINUE
51.      DO 6 I=1,K

```



```

52.      L1=K-I
53.      L2=(2)**L1
54.      DO 7 J=1,L2
55.      A(I,J)=A(I-1,2*J-1)+A(I-1,2*J)
56.      7 CONTINUE
57.      6 CONTINUE
58.      Y=A(K,1)
59.      C*****
60.      C  CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE REPRESENTATION
61.      C  ;SIG
62.      C*****
63.      D6=24.**2
64.      D5=(B**2-1.)
65.      D7=B**(-WW)*D5/(24.*RLB)
66.      D2=B**(-2*WW)*D5/RLB
67.      D3=7.*(B**2+1.)/960.
68.      D4=D5/(D6*RLB)
69.      VAR(M,IB)=D7-D2*(D3+D4)
70.      SIG(M,IB)=SQRT(VAR(M,IB))
71.      DD=10.**(-16)
72.      IF(SIG(M,IB),LE.DD) GOTO 53
73.      C*****
74.      C  CALCUL DE L ERREUR RELATIVE DU RESULTAT FINAL : DELTX
75.      C*****
76.      DELTX=ABS(R(K,1)-Y)/Y
77.      C  CALCUL DU RAPPORT DELTX/SIG
78.      P1=10.*(ALOG(2.)/RLB)
79.      MP1=INT(P1)
80.      IF(MP1.LT.M) GOTO 12
81.      53 PREC(M,IB)=10**6
82.      GOTO 33
83.      12 P2=50.*(ALOG(2.)/RLB)
84.      MP2=INT(P2)
85.      IF(M.LT.MP2) GOTO 34
86.      PREC(M,IB)=10**6
87.      GOTO 33
88.      34 PREC(M,IB)=DELTX/SIG(M,IB)
89.      LL=LL+1
90.      PMOY(IB)=PMOY(IB)+PREC(M,IB)
91.      33 CONTINUE
92.      C*****
93.      C  CALCUL DE LA MOYENNE DES VALEURS DE PREC(M)
94.      C*****
95.      PMOY(IB)=PMOY(IB)/FLOAT(LL)
96.      11 CONTINUE
97.      WRITE(108,18) (IB,IB=2,16)
98.      18 FORMAT(2X,'M',1X,15(1X,I4))
99.      DO 26 M=3,50
100.     WRITE(108,25) M,(PREC(M,IB),IB=2,16)
101.     25 FORMAT(2X,I2,2X,15(1X,F4.2))
102.     26 CONTINUE
103.     WRITE(108,42) (PMOY(IB),IB=2,16)
104.     42 FORMAT(1X,(/),6X,15(1X,F4.2))
105.     STOP
106.     END

```

```

1. C*****
2. C  METHODE DE CUMUL SIMPLE  UTILISANT LES SUBROUTINES D ARRONDI
3. C  SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
4. C  DELTX= ERREUR RELATIVE DU RESULTAT FINAL
5. C  LL=NOMBRE DE SIG CALCULES
6. C  PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
7. C  FINAL SUR L ECART TYPE DE L ERREUR RELATIVE D AFFECTATION
8. C  PMOY= MOYENNE DE PREC
9. C*****
10.      COMMON B,RLB,M,IS
11.      DIMENSION U(1500),V(1500),S(1500),T1(1500),L1(1500),R1(1500)
12.      DIMENSION T4(1500),L4(1500),R4(1500),T2(1500),L2(1500),R2(1500)
13.      DIMENSION VAR(60,20),SIG(60,20),PREC(60,20),PMOY(20),D(1500)
14. C    LECTURE DES DONNEES
15.      READ(105,22) IS
16.      22  FORMAT(I3)
17.      DO 11 K=1,15
18.      B=FLOAT(K+1)
19.      RLB=ALOG(B)
20.      IB=INT(B)
21.      LL=1
22.      PMOY(IB)=0.
23.      N=128
24.      DO 33 M=3,60
25.      SR=B**M
26.      PT=16.**13
27.      IF(SR.GT.PT) GOTO 53
28.      WW=2**M
29.      DO 4 I=1,N
30.      U(I)=0.333
31.      CALL CREATION (U(I),T1(I),L1(I))
32.      CALL SORTIE (U(I),T1(I),L1(I),R1(I))
33.      4  CONTINUE
34. C*****
35. C  CALCUL DE LA VALEUR EXACTE DE LA  SOMME  S(N)
36. C  AVEC DES DONNEES ARRONDIES
37. C*****
38.      S(1)=R1(1)
39.      R2(1)=R1(1)
40.      T2(1)=T1(1)
41.      L2(1)=L1(1)
42.      DO 2 I=2,N
43.      S(I)=S(I-1)+R1(I)
44.      V(I)=R2(I-1)+R1(I)
45.      IT=+1
46.      CALL SOMME (IT,T2(I-1),L2(I-1),T1(I),L1(I),T2(I),L2(I))
47.      CALL SORTIE (V(I),T2(I),L2(I),R2(I))
48.      IF(R2(I).EQ.R2(I-1)) GOTO 3
49.      2  CONTINUE
50.      Y=S(N)
51. C*****

```

```

52. C   CALCUL DE L ERREUR RELATIVE DU RESULTAT FINAL : DELTX
53. C*****
54.     DELTX=ABS(R2(N)-Y)/Y
55.     GOTO 7
56. 3   CONTINUE
57.     N=I-1
58.     R2(N)=R2(I-1)
59.     Y=S(N)
60.     DELTX=ABS(R2(N)-Y)/Y
61. 7   CONTINUE
62. C*****
63. C   CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE REPRESENTATION
64. C   ;SIG
65. C*****
66.     D6=24.**2
67.     D5=(B**2-1.)
68.     D7=B**(-WW)*D5/(24.*RLB)
69.     D2=B**(-2*WW)*D5/RLB
70.     D3=7.*(B**2+1.)/960.
71.     D4=D5/(D6*RLB)
72.     VAR(M,IB)=D7-D2*(D3+D4)
73.     SIG(M,IB)=SQRT(VAR(M,IB))
74.     DD=10.**(-16)
75. C   CALCUL DU RAPPORT DELTX/SIG
76.     IF(SIG(M,IB).LE.DD) GOTO 53
77.     P1=10.*(ALOG(2.)/RLB)
78.     MP1=INT(P1)
79.     IF(MP1.LT.M) GOTO 12
80. 53   PREC(M,IB)=10**6
81.     GOTO 33
82. 12   P2=50.*(ALOG(2.)/RLB)
83.     MP2=INT(P2)
84.     IF(M.LT.MP2) GOTO 34
85.     PREC(M,IB)=10**6
86.     GOTO 33
87. 34   PREC(M,IB)=DELTX/SIG(M,IB)
88.     ON=FLOAT(N)/3.
89.     IF(PREC(M,IB).LT.ON) GOTO 55
90.     PREC(M,IB)=ON/B
91. 55   LL=LL+1
92.     PMOY(IB)=PMOY(IB)+PREC(M,IB)
93. 33   CONTINUE
94. C*****
95. C   CALCUL DE LA MOYENNE DES VALEURS DE PREC
96. C*****
97.     PMOY(IB)=PMOY(IB)/FLOAT(LL)
98. 11   CONTINUE
99.     WRITE(108,18) (IB,IB=2,16)
100. 18   FORMAT(2X,'M',1X,15(1X,I6))
101.     DO 26 M=3,50
102.     WRITE(108,25) M,(PREC(M,IB),IB=2,16)

103. 25   FORMAT(2X,I2,2X,15(1X,F6.2))
104. 26   CONTINUE
105.     WRITE(108,42) (PMOY(IB),IB=2,16)
106. 42   FORMAT(1X,(/),6X,15(1X,F6.2))
107.     STOP
108.     END

```

```

1. C*****
2. C  CALCUL DE LA VALEUR D UN POLYNOME EN UTILISANT LE SCHEMA
3. C  CLASSIQUE AVEC LES DONNEES ARRONDIES
4. C  AVEC LES DONNEES  A(I) ; X : I=0,...,N
5. C  SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
6. C  DELTX= ERREUR RELATIVE DU RESULTAT FINAL
7. C  L= NOMBRE DE SIG CALCULES
8. C  PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
9. C  FINAL SUR LA VARIANCE DE L ERREUR RELATIVE D AFFECTATION
10. C  PMOY= MOYENNE DE PREC
11. C*****
12.      COMMON B,RLB,M,IS
13.      DIMENSION VAR(60,20),SIG(60,20)
14.      DIMENSION PREC(50,20),PMOY(20),A(50),B1(50),T1(50),L1(50),S
15.      DIMENSION R1(50),D(50),X1(50),T4(50),L4(50),R4(50)
16.      DIMENSION T2(50),L2(50),R2(50),T3(50),L3(50),R3(50)
17. C  LECTURE DES DONNEES
18.      READ(105,22) IS
19.      22  FORMAT(I3)
20.      DO 11 K=1,15
21.      B=FLOAT(K+1)
22.      RLB=ALOG(B)
23.      IB=INT(B)
24.      L=0
25.      PMOY(IB)=0.
26.      X=0.999
27.      N=30
28.      DO 33 M=5,60
29.      SR=B**M
30.      PT=16.**15
31.      IF(SR.GT.PT) GOTO 53
32.      W=2*M
33.      CALL CREATION (X,P,LP)
34.      CALL SORTIE (X,P,LP,R)
35.      DO 2 I=0,N
36.      A(I)=0.999
37.      CALL CREATION (A(I),T1(I),L1(I))
38.      CALL SORTIE (A(I),T1(I),L1(I),R1(I))
39.      2  CONTINUE
40.      T4(1)=P
41.      L4(1)=LP
42.      R4(1)=R
43.      T3(0)=T1(N)
44.      L3(0)=L1(N)
45.      R3(0)=R1(N)
46.      DO 3 I=1,N
47.      D(I)=R1(N-I)*R4(I)
48.      CALL MULTIP (T1(N-I),L1(N-I),T4(I),L4(I),T2(I),L2(I))
49.      CALL SORTIE (D(I),T2(I),L2(I),R2(I))
50.      S(I)=R3(I-1)+R2(I)
51.      IT=+1

```

```

52.      CALL SOMME (IT,T3(I-1),L3(I-1),T2(I),L2(I),T5(I),L3(I))
53.      CALL SORTIE (S(I),T3(I),L3(I),R3(I))
54.      X1(I+1)=R4(I)*R
55.      CALL MULTIP (T4(I),L4(I),P,LP,T4(I+1),L4(I+1))
56.      CALL SORTIE (X1(I+1),T4(I+1),L4(I+1),R4(I+1))
57.      3 CONTINUE
58. C*****
59. C   CALCUL DE LA VALEUR DU POLYNOME AVEC DES  DONNEES
60. C   ARRONDIES
61. C*****
62.      B1(0)=R1(N)
63.      DO 4 I=1,N
64.      4   B1(I)=B1(I-1)+D(I)
65. C*****
66. C   CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE  REPRESENTATION
67. C   ,SIG
68. C*****
69.      D6=24.**2
70.      D5=(B**2-1.)
71.      D7=B**(-W)*D5/(24.*RLB)
72.      D2=B**(-2*W)*D5/RLB
73.      D3=7.*(B**2+1.)/960.
74.      D4=D5/(D6*RLB)
75.      VAR(M,IB)=D7-D2*(D3+D4)
76.      SIG(M,IB)=SQRT(VAR(M,IB))
77.      DD=10.**(-16)
78.      IF(SIG(M,IB),LE,DD) GOTO 53
79. C*****
80. C   CALCUL DE L ERREUR RELATIVE SUR LE RESULTAT : DELTX
81. C*****
82.      DELTX=ABS(R3(N)-B1(N))/B1(N)
83. C   CALCUL DU RAPPORT DELTX/SIG
84.      P1=10.*(ALOG(2.)/RLB)
85.      MP1=INT(P1)
86.      IF(MP1.LT.M) GOTO 12
87.      53 PREC(M,IB)=10**6
88.      GOTO 33
89.      12 P2=50.*(ALOG(2.)/RLB)
90.      MP2=INT(P2)
91.      IF(M.LT.MP2) GOTO 34
92.      PREC(M,IB)=10**6
93.      GOTO 33
94.      34 PREC(M,IB)=DELTX/SIG(M,IB)
95.      L=L+1
96.      PMOY(IB)=PMOY(IB)+PREC(M,IB)
97.      33 CONTINUE
98. C*****
99. C   CALCUL DE LA  VALEUR MOYENNE DE PREC
100. C*****
101.      PMOY(IB)=PMOY(IB)/FLOAT(L)
102.      11 CONTINUE

```

```
103.      WRITE(108,18) (IB,IB=2,16)
104.      18  FORMAT(2X,'M',1X,15(1X,I5))
105.      DO 26 M=3,50
106.      WRITE(108,25) M,(PREC(M,IB),IB=2,16)
107.      25  FORMAT(2X,I2,2X,15(1X,F5.2))
108.      26  CONTINUE
109.      WRITE(108,42) (PMUY(IB),IB=2,16)
110.      42  FORMAT(1X,(/),6X,15(1X,F5.2))
111.      STOP
112.      END
```

```

1. C*****
2. C   CALCUL DE LA VALEUR D UN POLYNOME PAR LA METHODE DE HORNER
3. C   AVEC LES SOUS PROGRAMMES D ARRONDI
4. C   AVEC LES DONNEES A(I) ; X : I=0,...,N
5. C   SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
6. C   DELTX= ERREUR RELATIVE DU RESULTAT FINAL
7. C   L= NOMBRE DE SIG CALCULES
8. C   PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
9. C   FINAL SUR LA VARIANCE DE L ERREUR RELATIVE D AFFECTATION
10. C   PMOY= MOYENNE DE PREC
11. C*****
12.     COMMON B,RLB,M,IS
13.     DIMENSION VAR(60,20),SIG(60,20),PREC(60,20),PMOY(20)
14.     DIMENSION A(150),T1(150),L1(150),R1(150),T4(150),L4(150)
15.     DIMENSION R4(150),D(150),T2(150),L2(150),R2(150),S(150)
16.     DIMENSION B1(150),B2(150),T3(150),I3(150),R3(150)
17. C   LECTURE DES DONNEES
18.     READ(105,22) IS
19. 22  FORMAT(I3)
20.     DO 11 K=1,15
21.         B=FLOAT(K+1)
22.         RLB=ALOG(B)
23.         IB=INT(B)
24.         L=0
25.         PMOY(IB)=0.
26.         X=0.999
27.         N=30
28.         DO 33 M=3,60
29.             SR=B**M
30.             PT=16.**13
31.             IF(SR.GT.PT) GOTO 53
32.             W=2*M
33.             DO 2 I=0,N
34.                 A(I)=0.999
35.                 CALL CREATION (A(I),T1(I),L1(I))
36.                 CALL SORTIE (A(I),T1(I),L1(I),R1(I))
37. 2     CONTINUE
38. C   CALCUL DES B2(I) ; I=0,...,N
39.         B2(0)=A(0)
40.         DO 7 I=1,N
41.             D(I)=B2(I-1)*X
42.             B2(I)=D(I)+A(I)
43.             CALL CREATION (D(I),T2(I),L2(I))
44.             CALL SORTIE (D(I),T2(I),L2(I),R2(I))
45.             S(I)=R2(I)+R1(I)
46.             IT=+1
47.             CALL SOMME (IT,T2(I),L2(I),T1(I),L1(I),T4(I),L4(I))
48.             CALL SORTIE (S(I),T4(I),L4(I),R4(I))
49. 7     CONTINUE
50. C*****
51.     CALL CREATION (X,P,LP)

```

```

52.      CALL SORTIE (X,P,LP,R)
53.      C   CALCUL DE LA VALEUR DU POLYNOME AVEC DES  DONNEES
54.      C   ARRONDIES
55.      C*****
56.      B1(0)=R1(0)
57.      DO 4 I=1,N
58.      4   B1(I)=B1(I-1)*R+R1(I)
59.      C*****
60.      C   CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE REPRESENTATIO
61.      C   ,SIG
62.      C*****
63.      D6=24.**2
64.      D5=(B**2-1.)
65.      D7=B**(-W)*D5/(24.*RLB)
66.      D2=B**(-2*W)*D5/RLB
67.      D3=7.*(B**2+1.)/960.
68.      D4=D5/(D6*RLB)
69.      VAR(M,IB)=D7-D2*(D3+D4)
70.      SIG(M,IB)=SQRT(VAR(M,IB))
71.      DD=10.**(-16)
72.      IF(SIG(M,IB).LE.DD) GOTO 53
73.      C*****
74.      C   CALCUL DE L ERREUR RELATIVE SUR LE RESULTAT : DELTX
75.      C*****
76.      DELTX=ABS(R4(N)-B1(N))/B1(N)
77.      C   CALCUL DU RAPPORT DELTX/SIG
78.      P1=10.*(ALOG(2.)/RLB)
79.      MP1=INT(P1)
80.      IF(MP1.LT.M) GOTO 12
81.      53 PREC(M,IB)=10**6
82.      GOTO 33
83.      12 P2=50.*(ALOG(2.)/RLB)
84.      MP2=INT(P2)
85.      IF(M.LT.MP2) GOTO 34
86.      PREC(M,IB)=10**6
87.      GOTO 33
88.      34 PREC(M,IB)=DELTX/SIG(M,IB)
89.      L=L+1
90.      PMOY(IB)=PMOY(IB)+PREC(M,IB)
91.      33 CONTINUE
92.      C*****
93.      C   CALCUL DE LA VALEUR MOYENNE DE PREC
94.      C*****
95.      PMOY(IB)=PMOY(IB)/FLOAT(L)
96.      11 CONTINUE
97.      WRITE(108,18) (IB,IB=2,16)
98.      18 FORMAT(2X,'M',1X,15(1X,I5))
99.      DO 26 M=3,50
100.     WRITE(108,25) M,(PREC(M,IB),IB=2,16)
101.     25 FORMAT(2X,I2,2X,15(1X,F5.2))
102.     26 CONTINUE
103.     WRITE(108,42) (PMOY(IB),IB=2,16)
104.     42 FORMAT(1X,(/),6X,15(1X,F5.2))
105.     STOP
106.     END

```



```

1. C*****
2. C  CALCUL DU PROCEDE DELTA 2 D AITKEN AVEC LES SOUS
3. C  PROGRAMME D ARRONDI;EXPRESSION  $Y=X2-(X2-X1)**2/(X0+X2-2*X1)$ 
4. C  AVEC LES DONNEES X ,X0 ,X1 ,X2
5. C  X= VALEUR EXACTE DU RESULTAT
6. C  SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
7. C  DELTX= ERREUR RELATIVE DU RESULTAT FINAL
8. C  L= NOMBRE DE SIG CALCULES
9. C  PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
10. C  FINAL SUR LA VARIANCE DE L ERREUR RELATIVE D AFFECTATION
11. C  PMOY= MOYENNE DE PREC
12. C*****
13.      COMMON B,RLB,M,IS
14.      DIMENSION PREC(60,20),VAR(60,20),SIG(60,20),PMOY(20)
15. C  LECTURE DES DONNEES
16.      READ(105,22) IS
17.      22  FORMAT(I3)
18.      DO 11 K=1,15
19.          B=FLOAT(K+1)
20.          RLB=ALOG(B)
21.          IB=INT(B)
22.          L=0
23.          PMOY(IB)=0.
24.          X0=0.510
25.          X1=0.500
26.          X2=0.491
27.          DO 3 M=3,60
28.              SR=B**M
29.              PT=16.**13
30.              IF(SR.GT.PT) GOTO 53
31.              W=2*M
32.              CALL CREATION (X0,T0,L0)
33.              CALL SORTIE (X0,T0,L0,Y0)
34.              CALL CREATION (X1,T1,L1)
35.              CALL SORTIE (X1,T1,L1,Y1)
36.              CALL CREATION (X2,T2,L2)
37.              CALL SORTIE (X2,T2,L2,Y2)
38. C*****
39. C  CALCUL DE LA VALEUR EXACTE DE X AVEC DES DONNEES ARRONDIES
40. C*****
41.      DN=Y0+Y2-2*Y1
42.      DT=(Y2-Y1)**2
43.      Y11=Y2-DT/DN
44.      D0=Y1-Y0
45.      IT=-1
46.      CALL SOMME (IT,T1,L1,T0,L0,T3,L3)
47.      CALL SORTIE (D0,T3,L3,R1)
48.      D1=Y2-Y1
49.      IT=-1
50.      CALL SOMME (IT,T2,L2,T1,L1,T4,L4)
51.      CALL SORTIE (D1,T4,L4,R2)

```

```

52.      D=R2-R1
53.      IT=-1
54.      CALL SOMME (IT,T4,L4,T3,L3,T5,L5)
55.      CALL SORTIE (D,T5,L5,R3)
56.      RD=ABS(R3)
57.      DD=10.**(-16)
58.      IF(RD.LE,DD) GOTO 53
59.      C=R2*R2
60.      CALL MULTIP (T4,L4,T4,L4,T6,L6)
61.      CALL SORTIE (C,T6,L6,R4)
62.      Q=R4/R3
63.      CALL DIVISION (T6,L6,T5,L5,T7,L7)
64.      CALL SORTIE (Q,T7,L7,R5)
65.      Y=Y2-R5
66.      IT=-1
67.      CALL SOMME (IT,T2,L2,T7,L7,T8,L8)
68.      CALL SORTIE (Y,T8,L8,R6)
69.      C*****
70.      C  CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE REPRESENTATION
71.      C  ;SIG
72.      C*****
73.      D6=24.**2
74.      D5=(B**2-1.)
75.      D7=B**(-W)*D5/(24.*RLB)
76.      D2=B**(-2*W)*D5/RLB
77.      D3=7.*(B**2+1.)/960.
78.      D4=D5/(D6*RLB)
79.      VAR(M,IB)=D7-D2*(D3+D4)
80.      SIG(M,IB)=SQRT(VAR(M,IB))
81.      IF(SIG(M,IB).LE,DD) GOTO 53
82.      C*****
83.      C  CALCUL DE L ERREUR RELATIVE DU RESULTAT FINAL : DELTX
84.      C*****
85.      DELTX=ABS(R6-Y11)/Y11
86.      C  CALCUL DU RAPPORT DELTX/SIG
87.      P1=10.*(ALOG(2.)/RLB)
88.      MP1=INT(P1)
89.      IF(MP1.LT,M) GOTO 12
90.      53  PREC(M,IB)=10**6
91.      GOTO 3
92.      12  P2=50.*(ALOG(2.)/RLB)
93.      MP2=INT(P2)
94.      IF(M.LT,MP2) GOTO 34
95.      PREC(M,IB)=10**6
96.      GOTO 3
97.      34  PREC(M,IB)=DELTX/SIG(M,IB)
98.      L=L+1
99.      PMOY(IB)=PMOY(IB)+PREC(M,IB)
100.      3  CONTINUE
101.      C*****
102.      C  CALCUL DE LA MOYENNE DES VALEURS DE PREC(M)

```

```
103. C*****
104.      PMOY(IB)=PMOY(IB)/FLOAT(L)
105.      11  CONTINUE
106.      18  WRITE(108,18) (IB,IB=2,16)
107.      18  FORMAT(2X,'M',1X,15(1X,I6))
108.      DO 26 M=3,50
109.      25  WRITE(108,25) M,(PREC(M,IB),IB=2,16)
110.      25  FORMAT(2X,I2,2X,15(1X,F6.1))
111.      26  CONTINUE
112.      42  WRITE(108,42) (PMOY(IB),IB=2,16)
113.      42  FORMAT(1X,(/),6X,15(1X,F6.1))
114.      STOP
115.      END
```

```

1. C*****
2. C   CALCUL DU PROCEDE DELTA 2 D AITKEN AVEC LES SOUS
3. C   PROGRAMMES D ARRONDI  EXPRESSION  $X=(X0*X2-X1**2)/(X0+X2-2*X1$ 
4. C   AVEC LES DONNEES X ,X0 ,X1 ,X2
5. C   X= VALEUR EXACTE DU RESULTAT
6. C   SIG= VARIANCE DE L ERREUR RELATIVE D AFFECTATION
7. C   DELTX= ERREUR RELATIVE DU RESULTAT FINAL
8. C   L= NOMBRE DE SIG CALCULES
9. C   PREC= RAPPORT DE L ERREUR RELATIVE SUR LE RESULTAT
10. C   FINAL SUR L ECART TYPE DE L ERREUR RELATIVE D AFFECTATION
11. C   PMOY= MOYENNE DE PREC
12. C*****
13. C   COMMON B,RLB,M,IS
14. C   DIMENSION PREC(60,20),VAR(60,20),SIG(60,20),PMOY(20)
15. C   LECTURE DES DONNEES
16. C   READ(105,22) IS
17. C   22  FORMAT(I3)
18. C   DO 11 K=1,15
19. C   B=FLOAT(K+1)
20. C   RLB=ALOG(B)
21. C   IB=INT(B)
22. C   L=0
23. C   PMOY(IB)=0.
24. C   X0=0.210
25. C   X1=0.200
26. C   X2=0.191
27. C   DO 3 M=3,60
28. C   SR=B**M
29. C   PT=16.**15
30. C   IF(SR.GT.PT) GOTO 53
31. C   W=2*M
32. C   CALL CREATION (X0,T0,L0)
33. C   CALL SORTIE (X0,T0,L0,Y0)
34. C   CALL CREATION (X1,T1,L1)
35. C   CALL SORTIE (X1,T1,L1,Y1)
36. C   CALL CREATION (X2,T2,L2)
37. C   CALL SORTIE (X2,T2,L2,Y2)
38. C   P=Y0*Y2
39. C   CALL MULTIP (T0,L0,T2,L2,T3,L3)
40. C   CALL SORTIE (P,T3,L3,R1)
41. C   C1=Y1*Y1
42. C   CALL MULTIP (T1,L1,T1,L1,T4,L4)
43. C   CALL SORTIE (C1,T4,L4,R2)
44. C*****
45. C   CALCUL DE LA VALEUR EXACTE DE X AVEC DES DONNEES ARRONDIES
46. C*****
47. C   DN=Y0+Y2-2*Y1
48. C   DT=P-C1
49. C   R=DT/DN
50. C   D0=Y1-Y0
51. C   IT=-1

```

```

52.      CALL SOMME (IT,T1,L1,T0,L0,T5,L5)
53.      CALL SORTIE (D0,T5,L5,R3)
54.      D1=Y2-Y1
55.      IT=-1
56.      CALL SOMME (IT,T2,L2,T1,L1,T6,L6)
57.      CALL SORTIE (D1,T6,L6,R4)
58.      D=R4-R3
59.      IT=-1
60.      CALL SOMME (IT,T6,L6,T5,L5,T7,L7)
61.      CALL SORTIE (D,T7,L7,R5)
62.      DD=10.**(-16)
63.      RD=ABS(R5)
64.      IF(RD,LE,DD) GOTO 3
65.      C*****
66.      C  CALCUL DE L ECART TYPE DE L ERREUR RELATIVE DE REPRESENTATION
67.      C  ;SIG
68.      C*****
69.      D6=24.**2
70.      D5=(B**2-1.)
71.      D7=B**(-W)*D5/(24.*RLB)
72.      D2=9**(-2*W)*D5/RLB
73.      D3=7.*(B**2+1.)/960.
74.      D4=D5/(D6*RLB)
75.      VAR(M,IB)=D7-D2*(D3+D4)
76.      SIG(M,IB)=SQRT(VAR(M,IB))
77.      IF(SIG(M,IB).LE,DD) GOTO 53
78.      S=R1-R2
79.      IT=-1
80.      CALL SOMME (IT,T3,L3,T4,L4,T8,L8)
81.      CALL SORTIE (S,T8,L8,R6)
82.      Y=R6/R5
83.      CALL DIVISION (T8,L8,T7,L7,T9,L9)
84.      CALL SORTIE (Y,T9,L9,R7)
85.      DELTX=ABS(R7-R)/R
86.      C  CALCUL DU RAPPORT DELTX/SIG
87.      P1=10.*(ALOG(2.)/RLB)
88.      MP1=INT(P1)
89.      IF(MP1.LT.M) GOTO 12
90.      53  PREC(M,IB)=10**6
91.      GOTO 3
92.      12  P2=50.*(ALOG(2.)/RLB)
93.      MP2=INT(P2)
94.      IF(M.LT.MP2) GOTO 34
95.      PREC(M,IB)=10**6
96.      GOTO 3
97.      34  PREC(M,IB)=DELTX/SIG(M,IB)
98.      L=L+1
99.      PMOY(IB)=PMOY(IB)+PREC(M,IB)
100.      3  CONTINUE
101.      C*****
102.      C  CALCUL DE LA VALEUR MOYENNE DE PREC

```



```

103. C*****
104.      PMOY(IB)=PMOY(IB)/FLOAT(L)
105.      11  CONTINUE
106.      WRITE(108,18) (IB,IB=2,16)
107.      18  FORMAT(2X,'M',1X,15(1X,I6))
108.      DO 26 M=3,50
109.      WRITE(108,25) M,(PREC(M,IB),IB=2,16)
110.      25  FORMAT(2X,I2,2X,15(1X,F6.1))
111.      26  CONTINUE
112.      WRITE(108,42) (PMOY(IB),IB=2,16)
113.      42  FORMAT(1X,(/),6X,15(1X,F6.1))
114.      STOP
115.      END

```

```

1.      SUBROUTINE CREATION (Z,AM,L)
2.      C*****
3.      C   Z;  PARAMETRE EN ENTREE
4.      C   AM, L ; PARAMETRES DE SORTIE
5.      C   CE SOUS PROGRAMME CALCULE LA MANTISSE AM ET L EXPOSANT
6.      C   L DU NOMBRE Z  DANS LA BASE B
7.      C*****
8.      COMMON B,RLR,M
9.      JNT(X)=INT(X+100.)-100
10.     Z1=ABS(Z)
11.     IF(Z1.GT.1.E-40) GOTO 10
12.     L=0
13.     AM=0.
14.     RETURN
15. 10  CONTINUE
16.  C   CALCUL DE L EXPOSANT LZ  DE Z DANS LA  FORMULE  $Z=M1*R^{*(LZ-M)}$ 
17.     Z2=ALOG(Z1)/RLR
18.     LZ=JNT(Z2+1)
19.     C=B** (M-LZ)
20.  C   CALCUL DE LA MANTISSE DE Z  DANS LA BASE B
21.     Y11=Z+C*SIGN(0.5,Z)
22.     AM=AIN7(Y11)
23.  C   CALCUL DE L EXPOSANT DE Z DANS LA BASE B
24.     L=L7-M
25.     RETURN
26.     END

```

```

1.      SUBROUTINE DIVISION (AM1,LX1,AM2,LX2,AM,L)
2.      C*****
3.      C   AM1,LX1,AM2,LX2,  PARAMETRES EN ENTREE
4.      C   AM,L  PARAMETRES DE SORTIE
5.      C   CE SOUS PROGRAMME EFFECTUE LA DIVISION  DES MANTISSES
6.      C   ; AM1/AM2,  PUIS LA SOUSTRACTION DES EXPOSANTS ; LX1-LX2
7.      C*****
8.      COMMON B,RLR,M
9.      DM=AM1/A12
10.     LD=LX1-LX2
11.  C   NORMALISATION DE LA MANTISSE OBTENUE
12.     CALL CREATION (DM,BM,LEX)
13.     AM=BM
14.     L=LD+LEX
15.     RETURN
16.     END

```

```

1.      SUBROUTINE SORTIE (A,AM,L,X)
2.      C*****
3.      C   CE SOUS PROGRAMME DONNE L'ARRONDI DU NOMBRE A, DE
4.      C   MANTISSE AM, ET DE L'EXPOSANT L DANS LA BASE B
5.      C   A,AM,L   PARAMETRES EN ENTREE
6.      C   X,   PARAMETRE DE SORTIE
7.      C   IS =1 ,ON FAIT DES IMPRESSIONS INTERMEDIAIRES
8.      C   IS =0  IL N'Y A PAS D'IMPRESSIONS INTERMEDIAIRES
9.      C*****
10.     COMMON B,RIB,M,IS
11.     IF(IS.EQ.0) GOTO 2
12.     WRITE(108,5) A
13.     WRITE(108,6) AM
14.     WRITE(108,7) L
15.     5   FORMAT(1X,'A=',D22.15)
16.     6   FORMAT(1X,'AM=',D22.15)
17.     7   FORMAT(1X,'L=',I4)
18.     2   CONTINUE
19.     Y3=AM*B**(L)
20.     IF(AM.EQ.0) GOTO 12
21.     U1=Y3/A
22.     U=ABS(U1-1)
23.     IF(U.LT.16.E-13) GOTO 13
24.     X=Y3
25.     IF(IS.EQ.0) GOTO 3
26.     WRITE(108,8) X
27.     8   FORMAT(1X,'X=',D22.15)
28.     3   CONTINUE
29.     RETURN
30.     13  CONTINUE
31.     X=Y3
32.     IF(IS.EQ.0) GOTO 4
33.     WRITE(108,8) X
34.     4   CONTINUE
35.     RETURN
36.     12  CONTINUE
37.     X=0.
38.     IF(IS.EQ.0) GOTO 9
39.     WRITE(108,8) X
40.     9   CONTINUE
41.     RETURN
42.     END

```


SUBROUTINE MULTIP (AM1,LX1,AM2,LX2,AM,L)

```

1.
2. C*****
3. C  AM1,LX1,AM2,LX2,  PARAMETRES EN ENTREE
4. C  AM,L  PARAMETRES DE SORTIE
5. C*****
6.     COMMON B,RLR,M
7.     PM=AM1*AM2
8.     LP=LX1+LX2
9. C  NORMALISATION DE LA MANTISSE OBTENUE
10.    CALL CREATION (PM,PM,LEX)
11.    AM=PM
12.    L=LP+LEX
13.    RETURN
14.    END

```

SUBROUTINE SOMME (IT,AM1,LX1,AM2,LX2,AM,L)

```

1.
2. C*****
3. C  IT, AM1, LX1, AM2, LX2,  PARAMETRES EN ENTREE
4. C  IT=-1  ON A AFFAIRE A UNE DIFFERENCE
5. C  IT=+1  ON A AFFAIRE A UNE SOMME
6. C  AM,L,  PARAMETRES DE SORTIE
7. C*****
8.     COMMON B,RLR,M
9.     IF(LX1.NE.LX2) GOTO 4
10.    SM=AM1+SIGN(1,IT)*AM2
11.    LS=LX1
12.    GOTO 3
13. 4  CONTINUE
14.    IF(LX1.LT.LX2) GOTO 7
15.    C=B**(LX2-LX1)
16.    AM3=AM2*C
17.    SM=AM1+SIGN(1,IT)*AM3
18.    LS=LX1
19.    GOTO 3
20. 7  CONTINUE
21.    C=B**(LX1-LX2)
22.    AM3=AM1*C
23.    SM=AM3+SIGN(1,IT)*AM2
24.    LS=LX2
25. 3  CONTINUE
26. C  NORMALISATION DE LA MANTISSE OBTENUE
27.    CALL CREATION (SM,SM,LEX)
28.    AM=SM
29.    L=LS+LEX
30.    RETURN
31.    END

```



REFERENCES

- [1] B.V. GNEDENKO, A.N. KOLMOGOROV
Limit distributions for sums of independent random variables
Addison Wesley (1964)
- [2] J. NEVEU
Bases mathématiques du calcul des probabilités
Masson (1970)
- [3] W.A. GRANVILLE, P.F. SMITH et W.R. LONGLEY
Eléments de calcul différentiel et intégral.
Vuibert, Paris (1962).
- [4] Y. CHOQUET
Topologie
Masson, Paris (1969).
- [5] P.L. HENNEQUIN et A. TORTRAT
Théorie des probabilités et quelques applications
Masson, Paris (1965)
- [6] Ch. PISOT et M. ZAMANSKY
Mathématiques générales - Algèbre et Analyse
Dunod, Paris (1966)
- [7] W. GIVENS
A method of computing eigenvalues and eigenvectors suggested by
classical results on symmetric matrices
Appl. Math. Ser. N.B.S. 29 (1953) 117-122.
- [8] W. MILLER
Software for Roundoff analysis
ACM Trans. Math. Softw. 1 (1975) 108-128.
- [9] J.R. RICE
A theory of condition
SIAM Journ. Num. Anal. 3 (1966) 287-310.

- [10] J.H. WILKINSON
Rounding error on algebraic processes
Her Majesty's stationery office, London (1963)
- [11] F. CORDELLIER
Mesure de la stabilité numérique d'un schéma de calcul sans test.
dans "les mathématiques de l'informatique", Colloque de l'AFCET,
Paris, (1982) 505-513
- [12] P. HENRICI
Elements of numerical analysis (ch. 16)
Wiley, New-York (1964).
- [13] S. LINNAINMAA
Taylor expansion of the accumulated rounding error
B.I.T. 16 (1976) 146-160.
- [14] S. LINNAINMAA
Towards accurate statistical estimation of rounding errors in
floating-point computations
BIT 15 (1975) 165-173.
- [15] F.L. BAUER
Computational graphs and rounding error
SIAM Journ. Num. Anal. 11 (1974) 87-96.
- [16] R.L. ASHENHURST et N. METROPOLIS
Unnormalized floating-point arithmetic
Journ. A.C.M. 6 (1959) 415-428.
- [17] E. UKKONEN
On the calculation of the effects of roundoff errors
A.C.M. Trans. Math. Softw. 7 (1981) 259-271.
- [18] J. LARSON and A. SAMEH
Algorithms for roundoff error analysis. A relative approach.
Computing 24 (1980) 277-297.

- [19] M. LA PORTE ET J. VIGNES
Algorithmes numériques, analyse et mise en oeuvre.
Technip, Paris (1974)
- [20] A. KORGANOFF
Méthodes de calcul numérique - Tome 1
Dunod, Paris (1961)
- [21] J. LARSON and A. SAMEH
Efficient calculation of the effects of roundoff errors
A.C.M. Trans. Math. Softw. 4 (1978) 228-236.
- [22] W. MILLER and D. SPOONER
Software for roundoff analysis II
A.C.M. Trans. Math. Softw. 4 (1978) 369-387.
- [23] P. LINZ
Accurate floating point summation
C.A.C.M. 13 (1970) 361-362.
- [24] F. CORDELLIER
Cours d'Analyse Numérique
Institut Industriel du Nord, Lille (1979)
- [25] M. PICHAT
Contribution à l'étude des erreurs d'arrondi en arithmétique à virgule flottante.
Thèse d'Etat, Grenoble (1976)
- [26] R.W. HAMMING
On the distribution of numbers
Bell syst. Techn. Journ. (1970) 1609-1625.
- [27] D.E. KNUTH
The art of computer programming Vol II
Addison Wesley (1969)

- [28] R. ALT
Etude statistique de l'erreur numérique d'affectation sur un ordinateur d'arithmétique en base quelconque. Application à l'erreur commise dans le calcul d'une somme de produits de nombres.
Rapport I.P. 76.5 (1976)
- [29] J. BASS
Cours de Mathématiques - Tome I.
Masson, Paris (1964).
- [30] P.H. STERBENZ
Floating-point computation
Prentice-Hall, Englewood cliffs, N.J. (1974).

