Contribution à l'étude des erreurs d'arrondi en arithmétique à virgule flottante

Michèle Pichat

THESE

présentée à

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

POUR OBTENIR LE GRADE DE

DOCTEUR ES SCIENCES MATHEMATIQUES

Michèle PICHAT

CONTRIBUTION

A L'ETUDE DES ERREURS D'ARRONDI

EN ARITHMETIQUE A VIRGULE FLOTTANTE

Thèse soutenue en 1976 devant la Commission d'Examen :

Président : J. KUNTZMANN

Examinateurs :
- N. GASTINEL
- A. HOCQUENGHEM
- J. VIGNES
- J.H. WILKINSON

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

Monsieur Michel SOUTIF : Président

Monsieur Gabriel CAU : Vice-Président

MEMBRES DU CORPS ENSEIGNANTS DE L'U.S.M.G.

PROFESSEURS TITULAIRES

MM.	ANGLES D'AURIAC Paul	Mécanique des Fluides
	ARNAUD Paul	Chimie
	AUBERT Guy	Physique
	AYANT Yves	Physique Approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique Expérimentale
	BARBIER Reynold	Géologie Appliquée
	BARJON Robert	Physique Nucléaire
	BARNOUD Fernand	Biosynthèse de la Cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique Chirurgicale
	BEAUDOING André	Clinique de Pédiatrie et Puériculture
	BERNARD Alain	Mathématiques Pures
Mme	BERTRANDIAS Françoise	Mathématiques Pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques Pures
	BEZES Henri	Pathologie Chirurgicale
	BLAMBERT Maurice	Mathématiques Pures
	BOLLIET Louis	Informatique (I.U.T. B)
	BONNET Georges	Electrotechnique
	BONNET Jean-Louis	Clinique Ophtalmologique
	BONNET-EYMARD Joseph	Clinique Gastro-entérologique
Mme	BONNIER Marie-Jeanne	Chimie Générale
MM.	BOUCHERLE André	Chimie et Toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques Appliquées
	BRAVARD Yves	Géographie
	CABANEL Guy	Clinique Rhumatologique et Hydrologique
	CALAS François	Anatomie
	CARLIER Georges	Biologie Végétale
	CARRAZ Gilbert	Biologie Animale et Pharmacodynamie
	CAU Gabriel	Médecine Légale et Toxicologie
	CAUQUIS Georges	Chimie Organique
	CHABAUTY Claude	Mathématiques Pures
	CHARACHON Robert	Clinique Oto-Rhino-Laryngologique
	CHATEAU Robert	Clinique de Neurologie
	CHIBON Pierre	Biologie Animale
	COEUR André	Pharmacie Chimique et Chimie Analytique
	CONTAMIN Robert	Clinique Gynécologique
	COUDERC Pierre	Anatomie Pathologique
	CRAYA Antoine	Mécanique
Mme	DEBELMAS Anne-Marie	Matière Médicale
MM.	DEBELMAS Jacques	Géologie Générale
	DEGRANGE Charles	Zoologie
	DELORMAS Pierre	Pneumo-Phtisiologie
	DEPORTES Charles	Chimie Minérale
	DESRE Pierre	Métallurgie

MM.	DESSAUX Georges	Physiologie Animale
	DODU Jacques	Mécanique Appliquée (I.U.T. A)
	DOLIQUE Jean-Michel	Physique des Plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	DUGOIS Pierre	Clinique de Dermatologie et Syphiligraphie
	GAGNAIRE Didier	Chimie Physique
	GALLISSOT François	Mathématiques Pures
	GALVANI Octave	Mathématiques Pures
	GASTINEL Noël	Analyse Numérique
	GAVEND Michel	Pharmacologie
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques Pures
	GERMAIN Jean-Pierre	Mécanique
	GIRAUD Pierre	Géologie
	JANIN Bernard	Géographie
	KAHANE André	Physique Générale
	KLEIN Joseph	Mathématiques Pures
	KOSZUL Jean-Louis	Mathématiques Pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques Appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie Végétale
Mme	LAJZEROWICZ Janine	Physique
MM.	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie Générale
	LATURAZE Jean	Biochimie Pharmaceutique
	LAURENT Pierre-Jean	Mathématiques Appliquées
	LEDRU Jean	Clinique Médicale B
	LLIBOUTRY Louis	Géophysique
	LOISEAUX Pierre	Sciences Nucléaires
	LONGEQUEUE Jean-Pierre	Physique Nucléaires
	LOUP Jean	Géographie
Melle	LUTZ Elisabeth	Mathématiques Pures
	MALGRANGE Bernard	Mathématiques Pures
	BOUTET DE MONVEL Louis	Mathématiques Pures
	MALINAS Yves	Clinique Obstétricale
	MARTIN-NOEL Pierre	Seméiologie médicale
	MAZARE Yves	Clinique Médicale A
	MICHEL Robert	Minéralogie et Pétrographie
	MICOUD Max	Clinique Maladies Infectieuses
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie Nucléaire
	MULLER Jean-Michel	Thérapeutique (Néphrologie)
	NEEL Louis	Physique du solide
	OZENDA Paul	Botanique
	PAYAN Jean-Jacques	Mathématiques Pures
	PEBAY-PEYROULA Jean-Claude	Physique
	RASSAT André	Chimie Systématique
	RENARD Michel	Thermodynamique
	RINALDI Renaud	Physique
	DE ROUGEMONT Jacques	Neuro-Chirurgie
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SIBILLE Robert	Construction Mécanique (I.U.T. A)
	SOUTIF Michel	Physique Générale
	TANCHE Maurice	Physiologie

MM.	TRAYNARD Philippe	Chimie Générale
	VAILLANT François	Zoologie
	VALENTIN Jacques	Physique Nucléaire
	VAUQUOIS Bernard	Calcul Electronique
Mme	VERAIN Alice	Pharmacie Galénique
MM.	VERAIN André	Physique
	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale
	YOCCOZ Jean	Physique Nucléaire Théorique

PROFESSEURS ASSOCIES

MM.	CLARK Gilbert	Spectrométrie Physique
	CRABBE Pierre	CERMO
	ENGLMAN Robert	Spectrométrie Physique
	HOLTZBERG Frédéric	Basses Températures
	ROST Ernest	Sciences Nucléaires

PROFESSEURS SANS CHAIRE

Melle	AGNIUS-DELORD Claudine	Physique Pharmaceutique
	ALARY Josette	Chimie Analytique
MM.	AMBROISE-THOMAS Pierre	Parasitologie
	BELORIZKY Elie	Physique
	BENZAKEN Claude	Mathématiques Appliquées
	BIAREZ Jean-Pierre	Mécanique
	BILLET Jean	Géographie
	BOUCHET Yves	Anatomie
	BRUGEL Lucien	Energétique (I.U.T. A)
	BUISSON René	Physique (I.U.T. A)
	CONTE René	Physique (I.U.T. A)
	DEPASSEL Roger	Mécanique des Fluides
	GAUTHIER Yves	Sciences Biologiques
	GAUTRON René	Chimie
	GIDON Paul	Géologie et Minéralogie
	GLENAT René	Chimie organique
	GROULADE Joseph	Biochimie Médicale
	HACQUES Gérard	Calcul Numérique
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et Médecine Préventive
	IDELMAN Simon	Physiologie Animale
	JOLY Jean-René	Mathématiques Pures
	JULLIEN Pierre	Mathématiques Appliquées
Mme	KAHANE Josette	Physique
MM.	KUHN Gérard	Physique (I.U.T. A)
	LE ROY Philippe	Mécanique (I.U.T. A)
	LUU DUC Cuong	Chimie Organique
	MAYNARD Roger	Physique du Solide
	PELMONT Jean	Biochimie
	PERRIAUX Jean-Jacques	Géologie et Minéralogie
	PSISTER Jean-Claude	Physique du Solide
Melle	PIERY Yvette	Physiologie Animale
MM.	RAYNAUD Hervé	M.I.A.G.
	REBECQ Jacques	Biologie (CUS)
	REVOL Michel	Urologie
	REYMOND Jean-Charles	Chirurgie Générale
	RICHARD Lucien	Biologie Végétale
Mme	RINAUDO Marguerite	Chimie Macromoléculaire
MM.	ROBERT André	Chimie Papetière

MM.	SARRAZIN Roger	Anatomie et Chirurgie
	SARROT-REYNAUD Jean	Géologie
	SIROT Louis	Chirurgie Générale
Mme	SOUTIF Jeanne	Physique Générale
MM.	STREGLITZ Paul	Anesthésiclogie
	VIALON Pierre	Géologie
	VAN CUTSEM Bernard	Mathématiques Appliquées

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

MM.	AMBLARD Pierre	Dermatologie
	ARMAND Gilbert	Géographie
	ARMAND Yves	Chimie (I.U.T. A)
	BACHELOT Yvan	Endocrinologie
	BARGE Michel	Neuro chirurgie
	BARJOLLE Michel	MIAG
	BEGUIN Claude	Chimie organique
Mme	BERIEL Hélène	Pharmacodynamie
MM.	BOST Michel	Pédiatrie
	BOUCHARLAT Jacques	Psychiatrie adultes
Mme	BOUCHE Liane	Mathématiques (C.U.S.)
MM.	BRODEAU François	Mathématiques (I.U.T. B)
	BUTEL Jean	Orthopédie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et Organogénèse
	CHARDON Michel	Géographie
	CHERADAME Hervé	Chimie Papetière
	CHIAVERINA Jean	Biologie Appliquée (EFP)
	COHEN-ADDAD Jean-Pierre	Spectrométrie Physique
	COLOMB Maurice	Biochimie Médicale
	CONTAMIN Charles	Chirurgie thoracique et cardio-vasculaire
	CORDONNIER Daniel	Néphrologie
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie
	CYROT Michel	Physique du solide
	DELOBEL Claude	M.I.A.G.
	DENIS Bernard	Cardiologie
	DOUCE Roland	Physiologie Végétale
	DUSSAUD René	Mathématiques (C.U.S.)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine Légale
	FAURE Gilbert	Urologie
	FONTAINE Jean-Marc	Mathématiques Pures
	GAUTIER Robert	Chirurgie Générale
	GENSAC Pierre	Botanique
	GIDON Maurice	Géologie
	GROS Yves	Physique (I.U.T. A)
	GUITTON Jacques	Chimie
	HICTER Pierre	Chimie
	IVANES Marcel	Electricité
	JALBERT Pierre	Histologie
	KOLODIE Lucien	Hématologie
	KRAKOWIAK Sacha	Mathématiques Appliquées
	LE NOC Pierre	Bactériologie-virologie
	LEROY Philippe	I.U.T. A
	MACHE Régis	Physiologie Végétale
	MAGNIN Robert	Hygiène et Médecine Préventive
	MALLION Jean-Michel	Médecine du Travail
	MARECHAL Jean	Mécanique (I.U.T. A)
	MARTIN-BOUYER Michel	Chimie (C.U.S.)

M.	MICHOULIER Jean	Physique (I.U.T. A)
Mme	MINIER Colette	Physique (I.U.T. A)
MM.	NEGRE Robert	Mécanique (I.U.T. A)
	NEMOZ Alain	Thermodynamique
	NOUGARET Marcel	Automatique (I.U.T.A)
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (I.U.T. B)
	PEFFEN René	Métallurgie (I.U.T. A)
	PERRET Jean	Neurologie
	PERRIER Guy	Géophysique - Glaciologie
	PHELIP Xavier	Rhumatologie
	RACHAIL Michel	Médecine Interne
	RACINET Claude	Gynécologie et Obstétrique
	RAMBAUD André	Hygiène et Hydrologie
	RAMBAUD Pierre	Pédiatrie
Mme	RENAUDET Jacqueline	Bactériologie
MM.	ROBERT Jean-Bernard	Chimie-Physique
	ROMIER Guy	Mathématiques (I.U.T. B)
	SHOM Jean-Claude	Chimie générale
	STOEBNER Pierre	Anatomie pathologique
	VROUSOS Constantin	Radiologie

MAITRE DE CONFERENCES ASSOCIES

M.	COLE Antony	Sciences Nucléaires

CHARGE DE FONCTIONS DE MAITRE DE CONFERENCES

M.	JUNIEN-LAVILLAVROY Paul	O.R.L.

Fait à SAINT MARTIN D'HERES,
DECEMBRE 1975.

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Président : M. NEEL Louis

Vice-Présidents : M. BENOIT Jean

M. BONNETAIN Lucien

PROFESSEURS TITULAIRES

MM.	BENOIT Jean	Radioélectricité
	BESSON Jean	Electrochimie
	BLOCH Daniel	Physique du solide
	BONNETAIN Lucien	Chimie Minérale
	BONNIER Etienne	Electrochimie et Electrométallurgie
	BRISSONNEAU Pierre	Physique du solide
	BUYLE-BODIN Maurice	Electronique
	COUMES André	Radioélectricité
	FELICI Noël	Electrostatique
	LESPINARD Georges	Mécanique
	MOREAU René	Mécanique
	PARIAUD Jean-Charles	Chimie-Physique
	PAUTHENET René	Physique du solide
	PERRET René	Servomécanismes
	POLOUJADOFF Michel	Electrotechnique
	SILBERT Robert	Mécanique des Fluides

PROFESSEURS ASSOCIES

MM.	RUPPERSBERG Albert, Henner	Chimie
	ROUXEL Roland	Automatique

PROFESSEURS SANS CHAIRE

MM.	BLIMAN Samuel	Electronique
	BOUVARD Maurice	Génie Mécanique
	COHEN Joseph	Electrotechnique
	DURAND Francis	Métallurgie
	FOULARD Claude	Automatique
	LACOUME Jean-Louis	Géophysique
	LANCIA Roland	Electronique
	VEILLON Gérard	Informatique Fondamentale & Appliquée
	ZADWORNY François	Electronique

MAITRES DE CONFERENCES

MM.	ANCEAU François	Mathématiques Appliquées
	BOUDOURIS Georges	Radioélectricité
	CHARTIER Germain	Electronique
	GUYOT Pierre	Chimie Minérale
	IVANES Marcel	Electrotechnique
	JOUBERT Jean-Claude	Physique du solide
	MORET Roger	Electrotechnique Nucléaire
	PIERRARD Jean-Marie	Hydraulique
	ROBERT François	Analyse Numérique
	SABONNADIERE Jean-Claude	Informatique Fondamentale & Appliquée
Mme	SAUCIER Gabrièle	Informatique Fondamentale & Appliquée

MAITRE DE CONFERENCES ASSOCIE

M.	LANDAU Ioan	Automatique

CHERCHEURS DU C.N.R.S. (Directeurs et Maîtres de Recherche)

MM.	FRUCHART Robert	Directeur de Recherche
	ANSARA Ibrahim	Maître de Recherche
	CARRE René	Maître de Recherche
	DRIOLE Jean	Maître de Recherche
	MATHIEU Jean-Claude	Maître de Recherche
	MUNIER Jacques	Maître de Recherche

Je tiens à exprimer ma profonde reconnaissance à :

Monsieur le Professeur GASTINEL, qui a dirigé cette thèse et qui tout au long de ce travail, m'a constamment encouragée, et soutenue de son intéret, de ses conseils. Par ses précieuses indications, ses qualités humaines, son enthousiasme, sa haute compétence, Monsieur GASTINEL m'a permis de mener à bien ce travail. Je voudrais qu'il trouve ici l'expression de ma respectueuse gratitude.

Monsieur le Professeur KUNTZMANN, qui m'a fait détacher à l'Institut de Mathématiques Appliquées de GRENOBLE, qu'il a créé et sans cesse développé, et qui a bien voulu accepter aujourd'hui de présider le Jury de cette thèse.

Monsieur le Professeur WILKINSON, qui a bien voulu examiner ce travail: Que Monsieur WILKINSON me permette de lui exprimer combien je suis sensible à l'honneur qu'il me fait en acceptant de faire partie du Jury.

Monsieur le Professeur HOCQUENGHEM, qui par la bienveillante attention qu'il m'a toujours accordée et par ses conseils et ses encouragements, m'a aidée dans la poursuite de ce travail.

Monsieur le Professeur VIGNES, pour les suggestions apportées à une partie de cette étude, et pour l'attention qu'il a bien voulu m'accorder.

J'adresse enfin mes vifs remerciements à Mademoiselle DAVID et à Mademoiselle LAURENT, du département de Mathématiques du C.N.A.M, pour le soin apporté à la dactylographie de cette thèse, et à Monsieur IGLESIAS, du Service tirage de GRENOBLE, pour la réalisation de ce document.

INTRODUCTION

La représentation, sur un calculateur, d'un nombre réel par un élément d'un sous-ensemble fini S de $\mathbb{R}$, l'utilisation résultante d'une arithmétique approchant l'arithmétique réelle, générent des erreurs d'arrondi. Nous nous sommes attachés au problème de la minimisation de telles erreurs :

- Nous avons d'abord comparé la précision de différents schémas de sommation . Ce premier aspect est un problème de "stratégie". Sa raison d'être en est la non-associativité des opérations de l'arithmétique approchée. Par exemple, il a été établi que :

a) Si dans un calcul de somme de série alternée à terme général décroissant, on effectue la sommation des termes, non dans l'ordre des indices croissants mais dans l'ordre inverse, tous les chiffres de la somme approchée sont significatifs.

b) Si l'on compare les deux schémas suivants de sommation de N nombres flottants positifs x_i ($1 \leqslant i \leqslant N$) :

Schéma 1 : A x_1 est ajouté x_2 , puis au résultat x_3 , et ainsi de suite (schéma usuel).

Schéma 2 : x_1 et x_2 , x_3 et x_4 , ... , x_{N-1} et x_N sont ajoutés 2 à 2 puis ces résultats intermédiaires sont ajoutés les uns aux autres selon le schéma 1.

L'erreur d'arrondi dans le schéma 2 est inférieure ou égale à l'erreur d'arrondi dans le schéma 1 augmentée d'une unité dans son poids le plus faible.

- Nous avons ensuite étudié le problème de la correction des erreurs :

Nous avons, au niveau de l'opération élémentaire -addition, soustraction, multiplication, division- exprimé l'erreur, sous la forme d'un produit de composition sur S , par des opérations de l'arithmétique approchée ; la mise en oeuvre d'une telle représentation pour la correction d'un algorithme général a ensuite été effectuée. Les différentes arithmétiques flottantes ont été successivement étudiées. Nous avons testé les résultats sur des exemples .

- Nous avons enfin étudié le problème du contrôle des erreurs d'arrondi dans le cadre d'une utilisation parallèle de deux arithmétiques flottantes. Cette approche du contrôle d'erreur, plus proche de la réalité que l'arithmétique d'intervalles, a donné des résultats concernant plusieurs algorithmes usuels. Par exemple, deux suites d'opérations flottantes ont été définies pour le calcul des inverses C_1 et C_2 d'une M-matrice A (ou d'une matrice tridiagonale symétrique définie positive) qui vérifient $C_1 \leqslant A^{-1} \leqslant C_2$.

Le chapitre I introduit les définitions nécessaires à notre étude. Est rappelée d'abord la notion d'arrondi de $\mathbb{R}$ sur un sous-ensemble de lui-même, puis sont considérés différents types particuliers de sous-ensembles de $\mathbb{R}$.

Le but du chapitre II est d'exprimer les erreurs d'arrondi élémentaires en arithmétique à virgule flottante, à l'aide de cette même arithmétique.
Les opérations sont successivement étudiées, selon le schéma suivant :
- expressions, par produits de composition sur S , de l'erreur lorsque l'arithmétique flottante satisfait à certains axiomes. Sont ainsi généralisés ou complétés les travaux de T. J. DEKKER [5] , D. E. KNUTH [15], O. MOLLER [28], G. W. VELTKAMP [40] .
- étude des schémas définissant, sur les ordinateurs usuels, les opérations flottantes ; en particulier, sont considérés pour l'addition qui admet une grande variété de définitions, la troncature sans, puis avec, chiffre de garde, et l'arrondi sur chiffre de garde ; dans chaque cas est examinée la double possibilité de représentation des nombres négatifs, par signe et valeur absolue ou en notation de complément à 2.

Les chapitres III et IV traitent des applications des résultats du chapitre précédent à la correction d'algorithmes. La première partie du chapitre III est consacrée à la définition et à l'étude de la convergence d'un algorithme pour le calcul, par des opérations flottantes et avec une précision maximum, d'une somme algébrique ou d'un produit scalaire.
La correction des erreurs d'arrondi dans un algorithme plus général est ensuite abordée sous plusieurs aspects :
- les possibilités d'obtention du problème perturbé dans l'analyse a postériori des erreurs d'arrondi.
- l'utilisation de méthodes de raffinement itératif d'un résultat approché, la détermination exacte d'une fonction de l'erreur étant assurée.
- la résolution numérique de problèmes mal conditionnés.
- l'amélioration de la solution calculée par une méthode itérative d'un système d'équations linéaires.

L'objet du chapitre V est l'étude de la stabilité numérique des schémas de sommation . En particulier ,un théorème général de comparaison des schémas de sommation séquentiel et dichotomique est donné . L'étude du calcul numérique de la somme d'une série alternée à terme général décroissant est effectuée ; deux méthodes sont en outre proposées pour doubler la précision obtenue par l'algorithme de calcul défini plus haut.

Dans le chapitre VI, nous posons le problème suivant :
$(x,y) \in S^2_{\wedge}$ $\star \in \{ + , - , \times , / \}$, l'opération approchée sur S de $\star$, $\textcircled{\star}$, peut légitimement associer à $x \star y$, l'un ou l'autre des 2 éléments de S encadrant ce nombre ; une suite σ de N opérations sur $\mathbb{R}$ étant donnée, il est alors possible d'associer à σ , 2^N suites différentes d'opérations approchées. Peut-on déterminer, dans cet ensemble, deux suites calculant des solutions encadrant la solution exacte ?

Quel ensemble de solutions est obtenu lors de l'application des 2^N suites précédentes ?
Une réponse est apportée à la première question pour les algorithmes de la sommation, du produit scalaire, de l'évaluation numérique d'un polynôme, de l'inversion d'une M-matrice et d'une matrice tridiagonale, symétrique définie positive ainsi que, dans un cadre différent, pour la résolution d'un système triangulaire à matrice inverse positive.
Des théorèmes relatifs à l'ensemble des solutions calculées sont établis pour la somme et le produit scalaire.

Dans le chapitre VII, on présente une étude relative à la répartition de l'erreur d'arrondi élémentaire sur son domaine de définition, lorsqu'on considère cette erreur comme variable aléatoire.

En Annexe est donné un tableau des caractéristiques des arithmétiques à virgule flottante des calculateurs usuels .

CHAPITRE I

ARRONDIS SUR $\mathbb{R}$

Les paragraphes de ce chapitre introduisent successivement les définitions nécessaires pour l'étude des arrondis.

I - ARRONDIS DEFINIS DE $\mathbb{R}$ DANS UN SOUS-ENSEMBLE FINI DE $\mathbb{R}$

Ce paragraphe reprend, dans le cas des ensembles de réels associés à un sous-ensemble fini, les définitions formulées par U. KULISCH [18] pour l'étude générale des arrondis dans les ensembles ordonnés ; la terminologie choisie est toutefois un peu différente de celle de [18] en particulier pour la notion d'arrondi optimal.

Désignons par E, l'ensemble $\mathbb{R}$ ou un intervalle donné de $\mathbb{R}$. Soit S un sous-ensemble fini, non vide, de E.

Définition I. 1.

On dira qu'une application $\square : E \to S$ est un arrondi de E dans S, si les deux axiomes suivants sont vérifiés :

$$x \leqslant y \Longrightarrow \square x \leqslant \square y \qquad \forall\, (x,y) \in E^2 \quad \text{(monotonie)}$$

$$\square x = x \qquad \forall\, x \in S$$

Définition I. 2.

Un arrondi $\square : E \to S$ sera dit G - dirigé (respectivement D - dirigé), si l'axiome suivant est vérifié :

$$\square x \leqslant x \quad , \quad \forall\, x \in E$$

$$\text{(resp. } \square x \geqslant x \quad , \quad \forall\, x \in E \text{)}$$

S étant donné, l'arrondi G - dirigé (respectivement D - dirigé) de E dans S, s'il existe, est unique : Il sera noté ∇ (resp. Δ) [cf. 18]. Par extension, $x \in E$ étant donné, dès que le nombre $\mathrm{Max}\,\{z \mid z \in S \wedge z \leqslant x\}$ (resp. $\mathrm{Min}\,\{z \mid z \in S \wedge z \geqslant x\}$) existe, nous le noterons $\nabla\, x$ (resp. $\Delta\, x$). Remarquons que la définition de l'arrondi entraîne, pour $x \in E$ donné, $\square x = \nabla\, x \vee \Delta\, x$, l'un au moins des nombres $\nabla\, x$, $\Delta\, x$ étant défini, puisque S est non vide.

Compositions d'arrondis

Nous rappelerons le théorème suivant de U. KULISCH, que nous énoncerons uniquement pour des ensembles réels :

Théorème [U. KULISCH].

Si T et S sont deux sous-ensembles finis de E vérifiant $T \subseteq S \subseteq E$ et tels que soient définis les arrondis G - dirigés $\nabla : E \to T$, $\nabla_1 : E \to S$, $\nabla_2 : S \to T$ (resp. les arrondis D - dirigés $\Delta : E \to T$, $\Delta_1 : E \to S$, $\Delta_2 : S \to T$), alors, pour tout élément x de E on a :

$$\nabla x = \nabla_2 (\nabla_1 x) \qquad \Delta x = \Delta_2 (\Delta_1 x)$$

On peut aussi énoncer, ∇ , ∇_1 , ∇_2 , Δ , Δ_1 , Δ_2 ayant les mêmes définitions que ci-dessus :

Lemme I.1.

Si T et S sont deux sous-ensembles finis de E vérifiant $T \subseteq S \subseteq E$, on a, pour $x \in E - T$, l'égalité $\Delta_2(\nabla_1 x) = \nabla_2(\Delta_1 x)$ si et seulement si un et un seul des arrondis G - dirigé et D - dirigé de x dans S , $\nabla_1 x$ ou $\Delta_1 x$, est élément de T .

Soit $x \in E - T$:

a) Si aucun des arrondis dirigés $\nabla_1 x$, $\Delta_1 x$ de x dans S , n'est élément de T , l'inégalité $\nabla x < \nabla_1 x < \Delta_1 x < \Delta x$ entraîne :

$\Delta_2(\nabla_1 x) = \Delta x$

$\nabla_2(\Delta_1 x) = \nabla x$.

b) Si $\nabla_1 x$ et $\Delta_1 x$ sont tous deux éléments de T , $\Delta_2(\nabla_1 x) = \nabla x$, $\nabla_2(\Delta_1 x) = \Delta x$.

c) Si $\nabla_1 x = \nabla x$ et $\Delta_1 x < \Delta x$, l'égalité $\Delta_2(\nabla_1 x) = \nabla_2(\Delta_1 x)$ est vérifiée.

On peut remarquer qu'en particuler, si S se déduit de T par adjonction des demi-sommes de deux éléments consécutifs de T , $\Delta_2(\nabla_1 x)$ et $\nabla_2(\Delta_1 x)$ prennent, pour tout élément x de E , la valeur de T la plus proche de x (ou éventuellement les 2 valeurs optimales).

II - L'ARITHMETIQUE "APPROCHEE"

A un arrondi défini de E dans S , il est possible d'associer une arithmétique sur l'ensemble S.

(Dans les définitions ci-dessous, $y \neq 0$ si l'opération $\star$ représente la division)

Définition I.3.

Un arrondi $\square$ étant défini de E dans S, si $\star$ désigne l'une des opérations +, -, x , / définies sur $\mathbb{R}$, on appellera "opération induite de $\star$ sur S par $\square$ ", la loi $\circledast_{\square}$ de composition interne sur S définie pour $(x,y) \in S^2$ par :

$$x \circledast_{\square} y = \square(x \star y)$$

De façon plus générale, lorsqu'à l'opération $\star$ appartenant à l'ensemble $\{ + , - , \times , / \}$ des opérations élémentaires définies sur $\mathbb{R}$, on associe une loi de composition interne sur S, notée $\circledast$ et qui sera dite l'opération correspondant à $\star$ sur S, on donnera les définitions ci-dessous, énoncées par T.J. DEKKER [5] :

Définition I.4.

L'opération $\circledast : S \times S \to S$ sera dite "correcte" si :

$$\forall (x,y) \in S^2 \quad x \circledast y = \operatorname{Max} \{z \mid z \in S \wedge z \leq x \star y\} \vee \operatorname{Min} \{z \mid z \in S \wedge z \geq x \star y\}$$

Définition I.5.

L'opération $\circledast : S \times S \to S$ sera dite "optimale" si :

$$\forall (x,y) \in S^2 \quad |x \star y - x \circledast y| = \operatorname*{Min}_{z \in S} |x \star y - z|$$

Enfin, nous donnerons la définition suivante :

Définition I.6.

L'opération $\circledast : S \times S \to S$ sera dite "I - dirigée" (respectivement "E - dirigée") pour le couple $(x,y) \in S^2$ si :

$$|x \circledast y| \leq |x \star y|$$

$$(\text{resp. } |x \circledast y| \geq |x \star y|) .$$

Nous utiliserons aussi, relativement à une opération $\circledast : S \times S \to S$, la terminologie d'opération "G - dirigée" (resp. "D - dirigée") dont la définition est semblable à la définition I.2 énoncée pour un arrondi. Cette notion sera aussi considérée relativement à un couple donné de S^2.

Avec les notations précédentes, la définition d'une opération $\circledast$ correcte s'écrit :

$$x \circledast y = \nabla (x \ast y) \underset{\vee}{} \Delta (x \ast y)$$

Il vient immédiatement les énoncés suivants, relatifs aux opérations $\circledast_{\square}$, induites d'une opération $\ast$ par un arrondi :

L'opération $\circledast_{\square} : S \times S \to S$, induite de $\ast$ par un arrondi $\square$, est correcte.

Toute opération $\circledast_{\square}$, induite, par un arrondi, d'une opération $\ast$ commutative, est aussi commutative.

Notation :

Les opérations $\circledast$ ne sont pas généralement associatives. Nous choisirons, pour toute la suite, la convention suivante : L'ordre des opérations $\circledast$ sera défini de la gauche à la droite.

N éléments x_i $(1 \leqslant i \leqslant N)$ de S étant donnés, l'écriture

$$x_1 \circledast x_2 \circledast x_3 \ldots \circledast x_N$$

représente donc, par cette convention, le résultat :

$$(\ldots ((x_1 \circledast x_2) \circledast x_3) \circledast \ldots) \circledast x_N$$

Nous noterons encore, de façon plus concise, ce nombre par :

$\oplus \sum_{i=1}^{N} x_i$ si $\circledast = \oplus$

$\otimes \prod_{i=1}^{N} x_i$ si $\circledast = \otimes$

III - QUELQUES TYPES DE SOUS-ENSEMBLES DE $\mathbb{R}$

S étant un sous-ensemble fini, non vide, de $\mathbb{R}$, S^+ (resp S^-) désignera l'ensemble des éléments positifs (resp. négatifs) de S.

Nous utiliserons la terminologie suivante :

Définition I.7.

S étant un sous-ensemble fini de $\mathbb{R}$ et x un élément de S, on appelle prédécesseur de x dans S (resp. successeur de x dans S), et on note $(x)'$ (resp. $(x)''$) le nombre :

$$(x)' = \max \{ y \mid y \in S \wedge y < x \}$$

$$(\text{resp.}\ (x)'' = \min \{ y \mid y \in S \wedge y > x \})$$

La définition de $(x)'$ (resp. $(x)''$) suppose x différent du plus petit élément (resp. du plus grand élément) de S. Nous désignerons par $\mathbb{R}_S$ l'ensemble des réels compris entre ces deux éléments, et :

Définition I.8.

S étant un sous-ensemble fini de $\mathbb{R}$ contenant 0, et x un élément de $\mathbb{R}_S$, on appelera intervalle de précision de x sur S, et on notera $\varepsilon(x)$, le nombre strictement positif suivant :

$$\varepsilon(x) = \Delta x - \nabla x \qquad \text{si } x \notin S$$
$$\varepsilon(x) = (x)'' - x \qquad \text{si } x \in S^+$$
$$\varepsilon(x) = x - (x)' \qquad \text{si } x \in S^-$$

Pour l'élément 0, on précisera éventuellement si l'intervalle de précision est considéré à droite ou à gauche.

III.1. SOUS-ENSEMBLES DE $[-1, 1]$ A INTERVALLES CONSTANTS.

Un sous-ensemble S de $[-1, 1]$ à intervalles constants, est l'ensemble de $2P + 1$ éléments de la forme :
$x = m\,\varepsilon$ avec $\varepsilon = \frac{1}{P}$ et m entier vérifiant $-P \leqslant m \leqslant P$

Les éléments 1 ou -1 peuvent être exclus de S.
La représentation de $\mathbb{R}$ par un sous-ensemble de $[-1, 1]$ à intervalles constants, implique un cadrage des operandes avant éxécution d'une opération $\circledast$.
L'addition ou la soustraction sur S, de 2 éléments de cet ensemble, n'entraîne aucune erreur s'il ne se produit pas de dépassement de capacité.
Le produit sur $\mathbb{R}$ de 2 éléments de S appartient à l'intervalle $[-1, 1]$ et nous énoncerons une propriété relative au produit, sur S, d'éléments de cet ensemble, par des opérateurs corrects de multiplication :

Une multiplication correcte sur S de 2 éléments x et y de S donne pour résultat l'un des deux arrondis $\nabla(x\,y)$ ou $\Delta(x\,y)$.
Etant donnés N éléments x_i $(1 \leqslant i \leqslant N)$ de S, considérons l'ensemble des résultats obtenus par la multiplication dans l'ordre donné de ces éléments sur S, lorsque, à chaque pas i, la multiplication du produit partiel des i premiers éléments par x_{i+1} est réalisée par l'une ou l'autre des opérations $\otimes_\nabla$ ou $\otimes_\Delta$; on peut énoncer :

Proposition I.1.

Si S est un sous-ensemble de $[-1, 1]$ à intervalles constants, étant donnés N éléments x_i $(1 \leq i \leq N)$ de S, l'un au moins des 2^{N-1} produits :

$$x_1 \otimes_{\square_1} x_2 \otimes_{\square_2} x_3 \otimes_{\square_3} \dots \otimes_{\square_{N-1}} x_N = \otimes_{\square_i} \prod_{i=1}^{N} x_i$$

où $\square_i = \nabla \vee \Delta$,

est égal à $\nabla \left(\prod_{i=1}^{N} x_i\right)$, respectivement $\Delta \left(\prod_{i=1}^{N} x_i\right)$.

Il suffit de raisonner avec $x_i > 0 \quad \forall_i$

Par récurrence :

. La propriété est vraie au rang 2 puisque, par définition :

$$x_1 \otimes_{\square_1} x_2 = \nabla (x_1 x_2) \vee \Delta (x_1 x_2)$$

. Supposons la vérifiée au rang k, c'est-à-dire supposons que dans l'ensemble des éléments $\otimes_{\square_i} \prod_{i=1}^{k} x_i$, il en existe au moins deux, éventuellement de même valeur, que nous noterons $\tilde{\pi}_k$ et $\tilde{\pi}'_k$ vérifiant :

$$\tilde{\pi}_k = \nabla \left(\prod_{i=1}^{k} x_i\right) \quad , \quad \tilde{\pi}'_k = \Delta \left(\prod_{i=1}^{k} x_i\right) .$$

De :

$$\tilde{\pi}'_k x_{k+1} - \tilde{\pi}_k x_{k+1} = x_{k+1} \quad \varepsilon < \varepsilon \qquad \text{si} \quad x_{k+1} \neq 1$$

on déduit qu'il existe au plus un élément de S sur l'intervalle fermé $\left[\tilde{\pi}_k x_{k+1} \ , \ \tilde{\pi}'_k x_{k+1}\right]$.

L'égalité $\otimes_{\square_i} \prod_{i=1}^{k+1} x_i = \nabla \left(\prod_{i=1}^{k+1} x_i\right)$ est obtenue par exemple avec les $\square_i$ $(i = 1 , .. , k-1)$ ayant défini $\tilde{\pi}_k$, et le choix :

$\square_k = \nabla$ s'il n'existe pas d'éléments de S sur $\left[\tilde{\pi}_k x_{k+1} \ , \ \tilde{\pi}'_k x_{k+1}\right]$ ou si cet élément est supérieur strictement à $\prod_{i=1}^{k+1} x_i$,

$\square_k = \Delta$ sinon.

La propriété n'est plus vraie quels que soient les x_i de S, si l'on remplace l'une des opérations de multiplication par la division.

III.2. SOUS-ENSEMBLES DE $\mathbb{R}$ A INTERVALLES NON DECROISSANTS

Définition I.9.

On dira que le sous-ensemble fini S de $\mathbb{R}$ est à intervalles non décroissants, si :

$$|x| \leqslant |y| \Longrightarrow \varepsilon(x) \leqslant \varepsilon(y) \qquad \forall\, (x,y) \in S^2_{\wedge}\ x\,y > 0$$

Les opérateurs corrects d'addition arithmétique vérifient, sur un tel sous-ensemble de $\mathbb{R}$, la propriété suivante :

Proposition I.1'.

Si S est un sous-ensemble de $\mathbb{R}$ à intervalles non décroissants, étant donnés N éléments x_i $(1 \leqslant i \leqslant N)$ de S, de même signe, l'une au moins des 2^{N-1} sommes séquentielles :

$$x_1 \oplus_{\square_1} x_2 \oplus_{\square_2} x_3 \oplus_{\square_3} \cdots \oplus_{\square_{N-1}} x_N = \oplus_{\square_i} \sum_{i=1}^{N} x_i \quad \text{où } \square_i = \nabla_{\vee} \Delta,$$

égale $\nabla \left(\sum_{i=1}^{N}{}_{\frown} x_i\right)$, respectivement $\Delta \left(\sum_{i=1}^{N} x_i\right)$.

Remarque :

Nous donnerons un contre-exemple grossier montrant que la propriété précédente n'est évidemment pas vérifiée pour l'addition arithmétique, sur un sous-ensemble quelconque de $\mathbb{R}$.
Soit le sous-ensemble S de $\mathbb{R} = \{0\ ,\ 3\ ,\ 6\ ,\ 15\ ,\ 21\ ,\ 23\ ,\ 25\ ,\ 29\ ,\ 31\}$; les 4 possibilités de réaliser sur S la somme $3 + 6 + 15 = 24$, par des opérations du type $3 \oplus_{\square_1} 6 \oplus_{\square_2} 15$ sont respectivement :

$$(3 \oplus_{\nabla} 6) \oplus_{\nabla} 15 = 21 = (3 \oplus_{\nabla} 6) \oplus_{\Delta} 15$$

$$(3 \oplus_{\Delta} 6) \oplus_{\nabla} 15 = 29 \text{ et } (3 \oplus_{\Delta} 6) \oplus_{\Delta} 15 = 31$$

III.3. SOUS-ENSEMBLES DE $\mathbb{R}$ DE TYPE FLOTTANT

Définition I.10.

Le sous-ensemble fini S de $\mathbb{R}$ sera dit de type flottant si les deux axiomes suivants sont vérifiés :

1) $o \in S$ et $x \in S \Longrightarrow -x \in S$

2) pour tout x de S^+ tel que $\varepsilon(x)$ et $\varepsilon((x)')$ existent et vérifient $\varepsilon(x) \neq \varepsilon((x)')$, les rapports $\dfrac{x}{\varepsilon(x)}$,et $\dfrac{\varepsilon(x)}{\varepsilon((x)')}$ si $(x)' \neq 0$, sont entiers.

Un système de nombres à virgule flottante, de base b et de nombre de chiffres de mantisse s, est bien, en particulier, de ce type : Tout x positif du système vérifiant $\varepsilon(x) \neq \varepsilon((x)')$ satisfait à

$$\frac{x}{\varepsilon(x)} = b^{s-1} \quad ; \quad \frac{\varepsilon(x)}{\varepsilon((x)')} = b \quad \text{si} \quad (x)' \neq 0.$$

Un sous-ensemble de $\mathbb{R}$ de type flottant est à intervalles non décroissants.
La proposition I.1' est valable pour un tel système.
Tout élément x de S peut s'écrire sous la forme :
$x = \ell\, \varepsilon(x)$ où ℓ est un entier relatif.
L'addition correcte de 2 éléments de S vérifie le :

Lemme I.2.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant, l'erreur entachant l'addition correcte de deux éléments x et y de S tels que $\varepsilon(y) \leqslant \varepsilon(x) \wedge x + y \in \mathbb{R}_S$ vérifie :

$$|x + y - (x \oplus y)| \leqslant \varepsilon(x + y) - \varepsilon(y) \qquad \text{avec}$$
$$= 0 \quad \text{si} \quad \varepsilon(x + y) < \varepsilon(y)$$

Notation :

Etant donné un réel a, nous adopterons les notations suivantes :

$$\lfloor a \rfloor = \max \{n \mid n \in \mathbb{Z} \wedge n \leqslant a\}$$
$$\lceil a \rceil = \min \{n \mid n \in \mathbb{Z} \wedge n \geqslant a\}$$

Démonstration du lemme I.2. :

Par définition, si $u \in \mathbb{R}_S$, l'hypothèse que S est de type flottant entraîne :

$$\nabla u = \left\lfloor \frac{u}{\varepsilon(u)} \right\rfloor \varepsilon(u) \quad , \quad \Delta u = \left\lceil \frac{u}{\varepsilon(u)} \right\rceil \varepsilon(u)$$

Le lemme I.2 s'établit en considérant les 2 valeurs $\nabla(x + y)$ et $\Delta(x + y)$ que peut prendre l'addition correcte de x et y. Avec les notations $x = \ell\, \varepsilon(x)$, $y = m\, \varepsilon(y)$, $\ell, m \in \mathbb{Z}$

$$\nabla (x + y) = \left\lfloor \frac{\ell\, \varepsilon(x) + m\, \varepsilon(y)}{\varepsilon(x + y)} \right\rfloor \varepsilon(x + y)$$

Soit q l'entier relatif défini par :
$\ell\, \varepsilon(x) + m\, \varepsilon(y) = q\, \varepsilon(x + y) + e \qquad 0 \leqslant e < \varepsilon(x + y)$.
L'hypothèse $\varepsilon(y) \leqslant \varepsilon(x)$ entraîne que $\ell\, \varepsilon(x) + m\, \varepsilon(y)$ est multiple de $\varepsilon(y)$.
Il en est de même de $\varepsilon(x + y)$ dès qu'est vérifié $\varepsilon(x + y) \geqslant \varepsilon(y)$-voir note- ($x \oplus y = x + y$ sinon), et e satisfait à :

$$e = \lambda\, \varepsilon(y) \quad \text{avec} \quad \lambda \in \mathbb{Z} \quad \text{et} \quad \begin{cases} |\lambda| \leqslant \dfrac{\varepsilon(x + y)}{\varepsilon(y)} - 1 & \text{si} \quad \varepsilon(x + y) \geqslant \varepsilon(y) \\ \lambda = 0 & \text{sinon} \end{cases}$$

D'où le lemme.

Définition.

Lorsque l'addition de deux éléments x et y de S tels que $\varepsilon(y) \leqslant \varepsilon(x)$, vérifie : $\varepsilon(x + y) < \varepsilon(y)$, il est dit se produire le phénomène de cancellation.

Note: Dans le cas où $0 < |x+y| < \varepsilon(0)$, la propriété $\varepsilon(x+y)$ multiple de $\varepsilon(y)$ peut n'être pas vérifiée. La majoration donnée de l'erreur doit être alors: $|x + y - (x \oplus y)| < \varepsilon(x + y) - \varepsilon(y)$.
-Toutefois la propriété ci-dessus est toujours vérifiée pour les systèmes de nombres à virgule flottante-

III.4. SYSTEMES DE NOMBRES A VIRGULE FLOTTANTE.

Rappel

Un système T_b^s de nombres à virgule flottante, de base b , de nombre de chiffres de mantisse s , est l'ensemble :

$$T_b^s = \{x \mid x = \varepsilon(\sum_{k=1}^{s} a_k \; b^{-k}) \; b^p \wedge \varepsilon = \pm 1 \wedge a_k \in \{0, 1, \ldots, b-1\} \wedge p \in \mathbb{Z}, \; m_2 \leq p \leq m_1\}$$

où m_1 et m_2 sont des entiers relatifs donnés.

Nous considérons, afin de ne pas alourdir inutilement les démonstrations, uniquement une représentation normalisée ($a_1 \neq 0$ pour $x \neq 0$).
Tout élément de $T_b^s - \{0\}$ a alors une écriture unique.
Toutefois, si la représentation normalisée offre le maximum d'avantages, les résultats que nous obtiendrons par la suite, seraient semblables pour une écriture non normalisée des nombres.

Zéros de la représentation : $x = 0$ dès que $a_k = 0$, $\forall\, k$, mais l'exposant du zéro n'est pas défini. Il peut être nécessaire de lever cette ambiguité.
Remarquant qu'on enregistre, dans la plupart des calculateurs, non pas l'exposant p du nombre à virgule flottante, mais le nombre $c = p +$ constante, appelé la "caractéristique" du nombre et permettant de ne pas introduire 2 positions "signe" différentes dans un même mot, nous désignerons dans ce cas par "zéro vrai" la représentation de ce nombre définie par $\varepsilon = +1$, $a_k = 0 \; \forall\, k$, $c = 0$; cette représentation se comporte dans tous les cas comme le nombre réel 0 , contrairement aux autres représentations, dites "zéros anormaux". Lorsqu'une ambiguité pourrait se présenter, nous supposerons toujours qu'a été écrit un zéro vrai.

Notations :

1) La plus petite valeur absolue non nulle des éléments de T_b^s est b^{m_2-1} : nous noterons ξ ce nombre.

2) Si $x \in T_b^s - \{0\}$, on peut écrire ce nombre :
$x = \varepsilon\, \ell\, b^{p-s}$ où l'entier ℓ vérifie : $b^{s-1} \leqslant \ell < b^s$.

Nous noterons : $e(x) = p$ l'exposant de x.

La notation $e(u) = p$ sera étendue à l'entier relatif p associé au réel $u \in \mathbb{R}_{T_b^s} -]-\xi, \xi[$ par : $b^{p-1} \leqslant u < b^p$.
L'exposant d'un réel u tel que $|u| < \xi$ sera pris inférieur à tout autre exposant.

3) Lorsqu'aucune ambiguité n'apparaîtra, nous noterons simplement T le système T_b^s de nombres à virgule flottante de base b , et de nombre de chiffres de mantisse s .

Enfin, on peut énoncer pour un système de nombres à virgule flottante en base 2 :

Proposition I.1".

Si T_2^s est un système de nombres à virgule flottante en base 2, étant donnés N éléments x_i $(1 \leqslant i \leqslant N)$ de T_2^s de même signe et $N-1$ opérations $\star_i$ d'addition , de multiplication ou de division, l'un au moins des éléments :

$$x_1 \ (\star_1)_{\square 1} \ x_2 \ (\star_2)_{\square 2} \ x_3 \cdots (\star_{N-1})_{\square N-1} \ x_N$$

où $\square_i = \nabla \vee \Delta$,
égale $\nabla (x_1 \star_1 x_2 \cdots \star_{N-1} x_N)$, respectivement $\Delta (x_1 \star_1 x_2 \cdots \star_{N-1} x_N)$.

Reprenons le raisonnement par récurrence de la proposition I.1, en supposant au rang k l'existence de deux suites d'arrondis $\square_i$ et $\square_i'$ $(i = 1, \ldots, k-1)$ permettant le calcul sur T_2^s de $\tilde{r}_k = \nabla r_k$ et $\tilde{r}_k' = \Delta r_k$ où $r_k = x_1 \star_1 x_2 \cdots \star_{k-1} x_k$.
Notant $e(\nabla r_k) = \pi_k$ et $x_{k+1} = \ell\, b^{p_{k+1}-s}$, il vient selon la nature de l'opération $\star_k$:

1) $\star_k = \times$

$\tilde{r}'_k\, x_{k+1} - \tilde{r}_k\, x_{k+1} = \ell\, b^{-s}\, \varepsilon(r_{k+1}) < \varepsilon(r_{k+1})$ si $e(r_{k+1}) = \pi_k + p_{k+1}$

$= \ell\, b^{-s+1} \varepsilon(r_{k+1}) < 2\, \varepsilon(r_{k+1})$ si $e(r_{k+1}) = \pi_k + p_{k+1} - 1$ $\wedge$ b=2

Il existe au plus 2 éléments de T_2^s sur l'intervalle fermé $[\tilde{r}_k\, x_{k+1}\ ,\ \tilde{r}'_k\, x_{k+1}]$ et il est toujours possible de définir un choix de $\square_k$ qui, à partir de $\tilde{r}_k$ ou de $\tilde{r}'_k$, calcule $\nabla\, r_{k+1}$ par exemple.

2) $\star_k = /$

$\tilde{r}'_k\, /\, x_{k+1} - \tilde{r}_k\, /\, x_{k+1} = (b^s/\ell)\, \varepsilon(r_{k+1}) < 2\, \varepsilon(r_{k+1})$ si $e(r_{k+1}) = \pi_k - p_{k+1}$ $\wedge$ b=2

$= (b^{s-1}/\ell)\, \varepsilon(r_{k+1}) \leqslant \varepsilon(r_{k+1})$ si $e(r_{k+1}) = \pi_k - p_{k+1} + 1$

On remarquera que $\ell = b^{s-1}$ implique la seconde hypothèse pour $e(r_{k+1})$ - et $e(\tilde{r}_k\, /\, x_{k+1})$ - et on concluera comme précédemment.

Remarque

La proposition I.1" est encore vérifiée si l'expression arithmétique définie à partir des éléments x_i et des opérations $\star_i$ - d'addition arithmétique, de multiplication ou de division - comporte des produits de composition partiels.

Soit en effet :

$r = (x_1 \star_1 x_2 \ldots \star_{p-1} x_p)\ \star_p\ (x_{p+1} \star_{p+1} \ldots \star_{N-1} x_N)$

$= u \star_p v$

D'après ce qui précède, il est possible de choisir des suites d'arrondis, pour $i = 1, \ldots, p-1$, puis, pour $i = p+1, \ldots, N-1$, qui permettent le calcul sur T_2^s des éléments suivants :

$\tilde{u} = \nabla u$, $\tilde{u}' = \Delta u$

$\tilde{v} = \nabla v$, $\tilde{v}' = \Delta v$

En considérant l'intervalle :

$I = [\min(\tilde{u} \star_p \tilde{v}, \tilde{u}' \star_p \tilde{v}, \tilde{u} \star_p \tilde{v}', \tilde{u}' \star_p \tilde{v}'),\ \max(\tilde{u} \star_p \tilde{v}, \tilde{u}' \star_p \tilde{v},\ \tilde{u} \star_p \tilde{v}', \tilde{u}' \star_p \tilde{v}')]$

comme une réunion d'intervalles du type précédemment étudié - c'est-à-dire, par exemple, $[\tilde{u} \star_p \tilde{v}\ ,\ \tilde{u} \star_p \tilde{v}']$ - il suffit alors de remarquer que tous les éléments de T_2^s contenus dans I , sont obtenus au moins une fois pour l'ensemble des choix d'arrondis.

CHAPITRE II

CALCUL AUTOMATIQUE DE L'ERREUR D'ARRONDI ELEMENTAIRE EN ARITHMETIQUE A VIRGULE FLOTTANTE

INTRODUCTION

Soient un système $T_b^s = T$ de nombres à virgule flottante, et les opérations d'addition, soustraction, multiplication, division, définies sur T, que nous noterons respectivement $\oplus$, $\ominus$, $\otimes$, $\oslash$.

Le but de ce chapitre est de déterminer, sous forme de produit de composition sur T par des opérations $\oplus$, $\ominus$, $\otimes$, $\oslash$, des expressions de l'erreur d'arrondi élementaire:

$$X * Y - X \otimes Y \qquad X \in T,\ Y \in T,\ * \in \{+,-,\times,/\}.$$

L'opération de soustraction n'est pas considérée différemment de l'addition : nous supposerons que cette opération vérifie la propriété :

$$X \ominus Y = X \oplus (-Y) \qquad \forall (X, Y) \in T^2$$

Afin de ne pas compliquer inutilement les résultats, nous n'avons pas donné, dans les cas d'overflows où l'erreur d'arrondi reste élément de T, l'écriture du produit de composition représentant l'erreur. Les formules écrites dans ce chapitre, ne sont donc valables que s'il n'y a pas dépassement de capacité, ce dernier phénomène entraînant d'ailleurs généralement, l'arrêt des calculs.

I - L'ERREUR D'ARRONDI DANS L'ADDITION

I. 1. THEOREME PRELIMINAIRE

Proposition II. 1.

Si l'opération $\oplus : T \times T \to T$ est correcte et possède la propriété suivante : $e(X) - e(Y) > s \Longrightarrow X \oplus Y = X$ (C)
alors, $\forall (X, Y) \in T^2$, l'erreur $E = X + Y - (X \oplus Y)$ est élément de $T \cup]-\xi, \xi[$.

Soient, en effet, X et Y deux éléments de T .
Notons p et q leurs exposants respectifs et supposons $p \geqslant q$.

Lorsque $X + Y \in \mathbb{R}_T$, par le chapitre I , on peut affirmer que E vérifie :

$E = \lambda\, \varepsilon(Y) = \lambda\, b^{q-s}$ où λ est un entier relatif qui satisfait à :

$|\lambda| \leqslant \frac{\varepsilon(X+Y)}{\varepsilon(Y)} - 1$ si $\varepsilon(X+Y) \geqslant \varepsilon(Y)$

$= 0$ sinon

. Si $p - q \leqslant s$, il vient $|\lambda| < b^s$

($e(X+Y) \leqslant p+1$ si $p-q < s$, $e(X+Y) \leqslant p$ si $p-q = s$).

Lorsque $\lambda \neq 0$, il existe un entier non négatif unique ν tel que : $b^{s-\nu-1} \leqslant \lambda < b^{s-\nu}$

$E = (\lambda\, b^{\nu})\, .\, b^{q-\nu-s}$ avec $b^{s-1} \leqslant \lambda\, b^{\nu} < b^s$

$q - \nu < m_2 \Longrightarrow |E| < \xi \wedge E \notin T$

$q - \nu \geqslant m_2 \Longrightarrow E \in T$.

. Si $p - q > s$, on a $E = Y \in T$ en raison de la propriété (C) .

Si un overflow se produit, notant $|X| = 1\, b^{p-s}$, $|Y| = m\, b^{q-s}$, il vient $|E| = (\, m - (b^s - 1 - 1)\, b^{p-q}\,)\, b^{q-s}$ où $p = m_1$, et $E \in T$ si $|E| \geqslant \xi$ d'où la proposition.

Etant donnés X et Y éléments de T , il sera donc possible si $\oplus$ est correcte et vérifie (C) , de calculer $E = X + Y - (X \oplus Y)$ par des opérations <u>internes</u> sur T .

Remarque

Les résultats obtenus, concernant l'erreur d'addition élémentaire en arithmétique à virgule flottante, seront utiles par la suite :
X et Y étant deux éléments de T tels que $e(X) \geqslant e(Y)_{\wedge}\ X + Y \in \mathbb{R}_T$,
$E = X + Y - (X \oplus Y)$ vérifie :

$$E = \lambda\, b^{e(Y) - s} \qquad \text{avec } \lambda \in \mathbb{Z} \text{ et :}$$

. si $e(Y) \geqslant m_2 + s - 1$

$|\lambda| \leqslant b^{e(X+Y) - e(Y)}{}_{-1}$

$\leqslant b^{e(X \oplus Y) - e(Y)}{}_{-1}$ si $e(X+Y) \geqslant e(Y)_{\vee} e(X \oplus Y) \geqslant e(Y)$

$\lambda = 0$ si $e(X+Y) \leqslant e(Y)_{\vee} e(X \oplus Y) \leqslant e(Y)$

. si $e(Y) < m_2 + s - 1$

Les mêmes formules sont valables, sauf en cas d'apparition d'un underflow ($0 < |X+Y| < \xi$), où l'on peut seulement affirmer $|E| < \xi$.

I. 2. EXPRESSION, PAR UN PRODUIT DE COMPOSITION SUR T, DE L'ERREUR D'ARRONDI

I. 2. 1. Première formule

Soient $X = \varepsilon\, l\, b^{p-s}$ et $Y = \varepsilon'\, m\, b^{q-s}$ deux éléments de T, et soit $E = X + Y - (X \oplus Y)$.
On peut énoncer :

Proposition II. 2.

Si l'opération $\oplus : T \times T \longrightarrow T$ est :
correcte
possède la propriété (C)
I - dirigée pour tout couple $(X, Y) \in T^2{}_{\wedge}\ XY > 0$,

Alors, si $|E| \notin]\, 0, \xi\, [$, E vérifie :

$E = -((X \oplus Y) \ominus X \ominus Y)$ (II · 1) où $e(X) \geqslant e(Y)$

Démonstration :

Posons :

$S = X + Y$

$\tilde{S} = X \oplus Y$

$S = \tilde{S} + E$

Supposons $p \geqslant q$.

Tout d'abord, remarquons que $p-q > s$ entraîne $\tilde{S} = X$ par (C), et (II. 1) conduit bien, dans ce cas, à $E = Y$.

La démonstration dans le cas général s'effectuera en prouvant que $\tilde{S} - X$ est élément de T ; en effet, l'opération $\oplus$ étant correcte, il viendra :

$\tilde{S} \ominus X = \tilde{S} - X$

Alors , $(\tilde{S} \ominus X) - Y = - E$ étant élément de T d'après la proposition II. 1 on aura :

$(\tilde{S} \ominus X) \ominus Y = (\tilde{S} \ominus X) - Y = - E$

Ecrivant :

$S = (\varepsilon\, 1 + \varepsilon'\, m\, b^{q-p})\, b^{p-s}$

envisageons les différentes possibilités pouvant se présenter lors de l'opération $\tilde{S} \ominus X$:

1) $e(S) = e(X)$

Avec la seule hypothèse que $\oplus$ est correcte :

$$\tilde{S} = \lfloor \varepsilon\, 1 + \varepsilon'\, m\, b^{q-p} \rfloor\, b^{p-s} \;\vee\; \lceil \varepsilon\, 1 + \varepsilon'\, m\, b^{q-p} \rceil\, b^{p-s} ,$$

$$\tilde{S} - X = \varepsilon' \left(\lfloor m\, b^{q-p} \rfloor \;\vee\; \lceil m\, b^{q-p} \rceil \right) b^{p-s}$$ est élément de T ;

en effet, on a : $e(\tilde{S} - X) = q \vee q + 1$, ou éventuellement $\tilde{S} - X = 0$.

Il suffit alors de remarquer que $e(\tilde{S} - X) = q + 1$ ne pourrait provoquer d'overflow que pour $p = q = m_1$, mais alors dans ce cas $\tilde{S} = S$.

2) $e(S) < e(X)$

X et Y sont de signes différents.

Mais remarquons alors que :

$q = p \vee p - 1 \Longrightarrow \tilde{S} = S \vee 0$ et dans ce cas $\tilde{S} - X$ est élément de T .

$q \leqslant p - 2 \Longrightarrow e(S) = p - 1$, et :

$$\tilde{S} = \varepsilon \lfloor b1 - mb^{q-p+1} \rfloor\, b^{p-1-s} \;\vee\; \varepsilon \lceil b1 - mb^{q-p+1} \rceil\, b^{p-1-s}$$

Par suite, $\tilde{S} - X = -\varepsilon \ (\lfloor mb^{q-p+1} \rfloor \vee \lceil mb^{q-p+1} \rceil) \ b^{p-1-s}$ est élément de T .

3) $e(S) = e(X) + 1$

$\oplus$ étant I- dirigée pour tout couple (X, Y) tel que $X Y > 0$:

$$\tilde{S} = \varepsilon \ \left\lfloor \frac{1 + m\, b^{q-p}}{b} \right\rfloor \ b^{p+1-s}$$

$$\tilde{S} - X = \varepsilon \left(b \left\lfloor \frac{1 + m\, b^{q-p}}{b} \right\rfloor - 1 \right) b^{p-s}$$

$e(\tilde{S} - X) = q \vee q - 1$, et $\tilde{S} - X$ est élément de T à condition que $| \tilde{S} - X| \geqslant \xi$.

En remarquant que l'éventualité $| \tilde{S} - X | < \xi$ ne peut se produire que si les conditions nécessaires $q = m_2$, $p - q \leqslant s - 2$ sont réalisées, et que dans ce cas l'erreur E entachant $\tilde{S}$ est, en module, inférieure à ξ , on a encore, dans les conditions du théorème : $\tilde{S} - X \in T$.

Une autre proposition est donnée par :

Proposition II. 3.

Si l'opération $\oplus : T \times T \longrightarrow T$ est :
correcte
possède la propriété (C)
si $b = 2$

Alors, si $|E| \notin] 0, \xi [$, E vérifie (II . 1) où $e(X) \geqslant e(Y)$

Pour établir cette proposition, remarquons que l'hypothèse de I- direction de l'opération $\oplus$ pour tout couple $(X, Y) \in T^2_{\wedge} \ X Y > 0$, n'intervient dans la démonstration précédente que lorsque $e(S) = e(X) + 1$, pour éliminer le cas où $\tilde{S}$ est donné par :

$$\tilde{S} = \varepsilon \left\lceil \frac{1 + m\, b^{q-p}}{b} \right\rceil \ b^{p+1-s}$$

Dans ce cas , $\tilde{S} - X = \varepsilon \left(b \left\lceil \frac{1 + m\, b^{q-p}}{b} \right\rceil - 1 \right) b^{p-s}$

$e(\tilde{S} - X) = q \vee q + 1$, et $\tilde{S} - X$ peut ne pas être élément de T lorsque $q = p$.

Or, si $q = p$, $E = - \varepsilon \varphi \, b^{p-s}$ où $\varphi \in \{0, 1, \ldots , b - 1\}$

On vérifie immédiatement que si $b = 2$, $|E| = b^{p-s}$ et $|Y| \leqslant (b^s - 1) \, b^{p-s}$ entrainent $\tilde{S} - X = Y - E \in T$.

Remarque

Les soustractions apparaissant dans la formule (II.1) s'effectuant "sans erreur" dans les hypothèses des propositions II. 2 et II. 3, cette formule peut être remplacée par :

$E = Y \ominus (\tilde{S} \ominus X)$ (II.1')

T. J. DEKKER [5] a montré que la formule (II.1') est valable dans le cas où $b = 3$, à la condition que l'opération $\oplus$ soit optimale. Cet auteur a aussi établi cette formule pour de nombreuses autres arithmétiques à virgule flottante.

I.2.2. Généralisation d'un théorème de D. E. KNUTH

D. E. KNUTH [15] suggère, afin d'éliminer la condition non symétrique $e(X) \geqslant e(Y)$, la formule suivante qu'il établit dans l'hypothèse où $\oplus$ est optimale :

$$E = (Y \ominus (\tilde{S} \ominus X)) \oplus (X \ominus (\tilde{S} \ominus (\tilde{S} \ominus X))).$$

Gardant tout d'abord cette condition $e(X) > e(Y)$, on établit la proposition suivante :

Proposition II. 4.

Si l'opération $\oplus : T \times T \longrightarrow T$ est :

correcte

possède la propriété (C)

Alors, si $|E| \notin]0, \xi[$, E vérifie :

$E = (Y \ominus (\tilde{S} \ominus X)) \oplus (X \ominus (\tilde{S} \ominus (\tilde{S} \ominus X)))$ (II.2) où $e(X) \geqslant e(Y)$

Démonstration :

La formule (II. 2) vient immédiatement en remarquant que :

★ Dans les conditions de la proposition II. 2., on a toujours : $\tilde{S} \ominus X = \tilde{S} - X$, ce qui entraine, l'opération $\oplus$ étant correcte : $X \ominus (\tilde{S} \ominus (\tilde{S} \ominus X)) = 0$

★ Si $e(S) = e(X) + 1$, et que $\oplus$ est E - dirigée pour le couple $(X, Y) \in T^2_{\wedge}$ $X Y > 0$, avec de plus $q = p$, rappelons que :

$$\tilde{S} - X = \varepsilon \left(b \left\lceil \frac{1 + m}{b} \right\rceil - 1\right) b^{p-s}$$

Le cas litigieux $e(\tilde{S} - X) = p + 1$ conduit à l'une des deux valeurs :

$$\tilde{S} \ominus X = \varepsilon \left(\left\lceil \frac{1 + m}{b} \right\rceil - \left\lceil \frac{1}{b} \right\rceil \right) b^{p+1-s} \quad \vee \quad \varepsilon \left(\left\lceil \frac{1 + m}{b} \right\rceil - \left\lfloor \frac{1}{b} \right\rfloor \right) b^{p+1-s}$$

Posons :

$\tilde{S} - X = \hat{S} \ominus X + E'$

$E' = \varepsilon' \psi \, b^{p-s}$ où $\psi \in \{0, 1, \ldots, b-1\}$

D'où :

$Y - (\hat{S} \ominus X) = E + E' = (-\varepsilon \varphi + \varepsilon' \psi) \, b^{p-s}$

ce qui montre que ce nombre est élément de T dès que $E \in T$,

et

$Y \ominus (\hat{S} \ominus X) = E + E'$.

$\tilde{S} - (\hat{S} \ominus X) = \varepsilon \left(\left\lceil \frac{1}{b} \right\rceil \vee \left\lfloor \frac{1}{b} \right\rfloor \right) b^{p+1-s}$ est élément de T,

donc :

$\tilde{S} \ominus (\hat{S} \ominus X) = X + E'$

et E' étant élément de T lorsque $E \in T$, on en déduit (II. 2).

Rappelant la définition suivante, dûe à T. J. DEKKER :

Définition

L'opération $\oplus$: $T \times T \longrightarrow T$ est dite "proprement tronquante" si elle est commutative, et si pour tout couple (X, Y) de T^2 satisfaisant $|X| \geqslant |Y|$, on a :

$$X \oplus Y = \mathrm{Sup} \{Z \mid Z \in T \wedge Z \leqslant X + Y\} \quad \text{si} \quad Y \geqslant 0$$

$$X \oplus Y = \mathrm{Inf} \{Z \mid Z \in T \wedge Z \geqslant X + Y\} \quad \text{si} \quad Y < 0$$

On peut énoncer :

Proposition II. 5

Si l'opération $\oplus$: $T \times T \longrightarrow T$ est proprement tronquante, Alors, si aucun underflow n'apparait lors de la succession d'opérations définie par (II. 2) , E vérifie (II. 2) , quelles que soient les valeurs relatives de X et Y .

Démonstration :

Remarquons tout d'abord que l'hypothèse : $\oplus$ proprement tronquante, entraîne que (C) est vérifiée.

Si donc $e(X) \geqslant e(Y)$, (II. 2) est vérifiée dès que $E \in T$ d'après la proposition précédente.

Démontrons la proposition dans le cas où $e(X) < e(Y)$, ou plutôt, gardant l'hypothèse $e(X) > e(Y)$ pour la commodité des notations, établissons la formule suivante :

$$E = (X \ominus(\tilde{S} \ominus Y)) \oplus (Y \ominus(\tilde{S} \ominus(\tilde{S} \ominus Y))) \qquad (II.3)$$

Posons :

$\tilde{S} - Y = \tilde{S} \ominus Y + E'$

Il vient :

$X - (\tilde{S} \ominus Y) = E + E'$

$\tilde{S} - (\tilde{S} \ominus Y) = Y + E'$

$Y - (\tilde{S} - (\tilde{S} \ominus Y)) = - E'$

La proposition II. 1 entraîne que E' est élément de T à la condition $|E'| \geqslant \xi \vee E' = 0$

Pour prouver (II· 3) , il suffit, l'opération $\oplus$ étant correcte, d'établir que $E + E'$ et $Y + E'$ sont éléments de T .

D'après la démonstration de la proposition II. 1 , E et E' s'écrivent :

$E = M\, b^{q-s}$

$E' = N\, b^{q-s}$

où M et N sont des entiers relatifs.

(en effet, pour $\tilde{S} \neq S$, on a toujours $e(\tilde{S}) \geqslant e(Y)$).

De plus, $\oplus$ étant proprement tronquante, la troncature de la somme $X + Y$ ($|X| \geqslant |Y|$) lorsque $X + Y \notin T$, se fait dans la direction de $- Y$.

E a donc le signe de Y .

E' , lorsque $E' \neq 0$, a le signe de $- Y$.

En effet, l'inégalité $|\tilde{S}| < |Y|$ implique $\tilde{S} = S \vee 0$ et $E' = 0$.

De l'écriture de E et E' , il vient immédiatement que $E + E'$ et $Y + E'$ sont éléments de T à condition que leur formation ne donne pas lieu à l'apparition d'un underflow.

Nous verrons d'ailleurs par la suite qu'une telle éventualité ne peut pas se produire lorsque $E \in T$.

Il est, par contre, possible que E' vérifie $0 < |E'| < \xi$ bien que $|E| > \xi$.

Exemple :

$b = 10$, $s = 6$, $X = 0{,}100000 \cdot 10^{m_2+6}$, $Y = 0{,}234986 \cdot 10^{m_2+3}$

$\tilde{S} = 0{,}100234 \cdot 10^{m_2+6}$, $E = 0{,}986000 \cdot 10^{m_2}$

$\tilde{S} \ominus Y = 0{,}999991 \cdot 10^{m_2+5}$, $E' = -0{,}086000 \cdot 10^{m_2}$

Proposition II. 6.

Si l'opération $\oplus$: $T \times T \to T$ est :

correcte

possède la propriété (C)

Alors, si aucun underflow n'apparait lors de la succession d'opérations définie par (II. 2) , E vérifie (II. 2) , quelles que soient les valeurs relatives de X et Y .

Démonstration :

Démontrons (II. 3) avec les seules hypothèses que $\oplus$ est correcte et possède la propriété (C) .

E' étant défini comme précédemment, nous établirons la proposition en montrant que $E + E'$ et $Y + E'$ sont éléments de T .
La démonstration est une longue suite de cas à envisager.

Tout d'abord, remarquons : $p-q > s \vee E = 0 \Longrightarrow$ (II. 3) vérifiée.

Dans le cas général, écrivons l'égalité :
$\tilde{S} - Y = X - E$
et étudions les différentes possibilités en utilisant les résultats du § I. 2. 1 de ce chapitre.

1) $e(S) = p$
$|E| < b^{p-s} \xrightarrow[\text{si } l \neq b^{s-1}]{} \tilde{S} \ominus Y = X \vee X - \varepsilon'' b^{p-s}$ où ε'' est le signe de E.

Donc :
$E' = -E \vee \varepsilon'' b^{p-s} - E$
ce qui établit que $E + E' \in T$ lorsque $|E| \geqslant \xi$.
Il vient ensuite :
$Y + E' = \varepsilon' (\lfloor m b^{q-p} \rfloor \vee \lceil m b^{q-p} \rceil) b^{p-s}$
et $Y + E'$ est bien élément de T .

Dans le cas particulier $l = b^{s-1}$ et E du signe de X , on vérifie les 2 égalités suivantes :
$E + E' = \lambda b^{p-1-s}$ où λ est un entier relatif avec $0 \leqslant |\lambda| \leqslant b$
$E + E' \in T$ lorsque E est élément de T .
$Y + E' = \varepsilon' (\lfloor m b^{q-p+1} \rfloor \vee \lceil m b^{q-p+1} \rceil) b^{p-1-s}$

2) $e(S) = p-1$

Dans ce cas :

$|E + E'| = 0 \vee b^{p-s} \vee b^{p-1-s} \qquad (1 = b^{s-1})$

et

$E + E' \in T$ lorsque $E \in T$.

On a :

$Y + E' = -\varepsilon \ (\lfloor m\, b^{q-p+1} \rfloor + b\,\theta)\, b^{p-1-s} \vee -\varepsilon \ (\lceil m\, b^{q-p+1} \rceil + b\,\varphi)\, b^{p-1-s}$

avec

$\theta = 0$ ou 1 , $\varphi = 0$ ou -1.

On vérifie bien que $Y + E'$ est élément de T lorsque $|E| \geqslant \xi$: en effet, les conditions $\varphi = -1$ et $q = m_2$ qui , dans certains cas, peuvent entraîner un underflow, impliquerait $|E| < \xi$.

3) $e(S) = p + 1$

On vérifie d'abord, en utilisant l'inégalité $|E| < b^{p+1-s}$ et en considérant le cas particulier $b = 2$, que l'on a toujours $e(X - E) \geqslant p$.

* si $e(X - E) = p$, il vient :

$E + E' = \lambda\, b^{p-s}$

où λ est un entier relatif avec $0 \leqslant |\lambda| \leqslant b$,

et $E + E' \in T$ si $E \in T$.

D'autre part, les résultats vus au début de ce chapitre conduisent à :

$Y + E' = \varepsilon \ (\lfloor m\, b^{q-p} \rfloor \vee \lceil m\, b^{q-p} \rceil)\, b^{p-s}$

et $Y + E'$ est élément de T .

* si $e(X - E) = p + 1$, il vient :

$E + E' = \lambda\, b^{p-s}$

où λ est un entier relatif avec $0 < |\lambda| \leqslant 2\, b - 1$

et

$Y + E' = \varepsilon \ (\lfloor m\, b^{q-p-1} \rfloor \vee \lceil m\, b^{q-p-1} \rceil)\, b^{p+1-s}$

La proposition est donc démontrée.

I. 3. QUELQUES EXEMPLES DE REGLES D'ADDITION EN ARITHMETIQUE A VIRGULE FLOTTANTE

Nous rappelons, dans ce paragraphe, différentes règles permettant de définir l'addition, en arithmétique à virgule flottante normalisée, de deux éléments $X = \varepsilon\ 1\ b^{p-s}$ et $Y = \varepsilon'\ m\ b^{q-s}$ ($e(X) \geqslant e(Y)$) de T .

I. 3. 1. La troncature sans chiffre de garde

Rappel.

La règle définissant l'addition sur T est une règle de troncature sans chiffre de garde , si l'opération d'addition s'effectue suivant le schéma suivant :

1) L'exposant provisoire de la somme $X \oplus Y$ est pris égal à $e(X)$. Si $e(Y) < e(X)$, Y est cadré, ce cadrage s'effectuant en déplaçant de $e(X) - e(Y)$ positions vers la droite la mantisse de ce nombre, et en tronquant les $e(X) - e(Y)$ derniers chiffres de cette mantisse.

2) Les mantisses, qui sont à présent alignées, sont ajoutées algébriquement, puis le résultat obtenu est normalisé, et sa mantisse est tronquée à s chiffres lorsque nécessaire.

I. 3. 1. a. Les nombres négatifs sont représentés par leur signe et leur valeur absolue.

Le cadrage de Y s'est effectué en remplaçant Y par $Y^t = \varepsilon' \lfloor m b^{q-p} \rfloor b^{p-s}$

$$X \oplus Y = (\varepsilon 1 + \varepsilon' \lfloor m\, b^{q-p} \rfloor)\, b^{p-s} \quad \vee \quad 0 \quad \text{si } |X + Y^t| < \xi \quad \text{lorsque } e(S) \leqslant p$$

$$X \oplus Y = \varepsilon \left\lfloor \frac{1 + m\, b^{q-p}}{b} \right\rfloor b^{p+1-s} \qquad \text{lorsque } e(S) = p+1$$

Donc :

Propriété.

1) L'opération $\oplus$ pour la troncature sans chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue, n'est pas correcte, mais l'erreur E vérifie :

$$|E| < b^{P-s} \quad \text{où} \quad P = \max\{e(X \oplus Y)\ ,\ e(X),\ e(Y)\} \quad \text{si} \quad X \oplus Y \neq 0$$
$$|E| < \xi \quad \text{si} \quad X \oplus Y = 0 \qquad \text{(II. 4)}$$

2) E est élément de T dès que $|E| \notin]0, \xi[$

3) $\oplus$ est :

I- dirigée pour tout couple $(X, Y) \in T^2 \wedge X\,Y > 0$

E- dirigée pour tout couple $(X, Y) \in T^2 \wedge X\,Y < 0 \wedge |X + Y^t| \geqslant \xi$

4) La symétrie $(X \to -X)$ est un automorphisme de T .

(II. 4) vient immédiatement.

On a la seconde partie de la propriété en remarquant que $|E|$ est du type $\lambda\, b^{q-s}$ où λ est entier, et :

$$p - q \leqslant s - 1 \Longrightarrow \lambda < b^{s}$$

$$p - q \geqslant s \Longrightarrow E = Y$$

3) et 4) sont évidents.

Rappelons que si l'addition de deux nombres de même signe est correcte, la soustraction arithmétique est de faible précision, puisqu'elle peut engendrer, lorsque se produit le phénomène de cancellation, des erreurs du même ordre de grandeur que le résultat obtenu.

Ainsi :

si $b = 10$, $s = 6$, $X = 0{,}100000$, $Y = 0{,}999999 \cdot 10^{-1}$

on a :

$X \ominus Y = 10^{-6}$ et $E = -9 \cdot 10^{-7}$.

Toutefois $p - q \geqslant 2$ assure $e(X \ominus Y) \geqslant p - 1$, et au plus, deux chiffres du résultat sont entachés d'erreur.

Les théorèmes démontrés au § I. 2. de ce chapitre, ne s'appliquent pas ici, mais on peut énoncer :

<u>Proposition II. 7.</u>

Si la règle définissant l'opération $\oplus$ est une troncature sans chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue, l'erreur E entachant $\tilde{S} = X \oplus Y$ est donnée, si $|E| \notin [0, \xi[$, par la suite d'opérations :

$U = \tilde{S} \ominus X$

$F X = U \ominus \tilde{S} \oplus X$

$E = -(U \ominus b \cdot FX \oplus (b \cdot FX \ominus FX) \ominus Y)$ (II. 5)

avec la condition $|X| \geqslant |Y|$.

<u>Démonstration</u> :

Notation - lors de l'addition $A \oplus B$ de 2 éléments A et B de T, ($|A| \geqslant |B|$) , nous sommes conduits à examiner :

- l'erreur dûe au cadrage de B , et nous noterons :

$$(A + B)_i = A + B^{t}$$

- l'erreur dûe à la troncature après normalisation de ce résultat intermédiaire, c'est-à-dire l'apparition éventuelle d'une retenue.

$\tilde{S}$, si $\tilde{S} \neq 0$, a le signe de X .
Distinguons 2 cas :

1) $e(\tilde{S}) \leqslant p$

Si $\tilde{S} = 0$, l'erreur est, en module, inférieure à ξ , et (II. 5) conduit à $E = 0$, valeur convenable uniquement si $\tilde{S} = S$.

Si $\tilde{S} \neq 0$, la mantisse de $\tilde{S}$ se cadrant sans erreur pour l'exposant p , et l'opération $\tilde{S} - X$ étant une soustraction arithmétique :
$U = \tilde{S} \ominus X = (\tilde{S} - X)_i = \tilde{S} - X = \varepsilon' \lfloor m\, b^{q-p} \rfloor\ b^{p-s}$

Ceci met en évidence $e(U) = q$. (sauf si $p-q \geqslant s$, mais alors par (II. 5), $E = Y$) .
L'opération $U \ominus \tilde{S}$:
les mantisses de U et de $\tilde{S}$ pouvant se cadrer sans erreur pour l'exposant p ($\geqslant$ à $e(\tilde{S})$ et $e(U)$) ,
$(U - \tilde{S})_i = U - \tilde{S} = - \varepsilon\, 1\, b^{p-s}$
$U \ominus \tilde{S} = - X$
$F X = 0$ et $E = - (U \ominus Y)$
$e(U) = e(Y)$ donc $(U - Y)_i = U - Y$, et $U \ominus Y$ s'effectuant sans erreur dès que $E \in T$, la proposition est vérifiée.

2) $e(\tilde{S}) = p + 1$

$U = \tilde{S} \ominus X = \tilde{S} - X^t$ où $X^t = \varepsilon \left\lfloor \frac{1}{b} \right\rfloor b^{p+1-s}$

Remarquons que si $m_2 = \lfloor m\, b^{q-p-1} \rfloor$, $U = \varepsilon\, (m_2 + \theta)\, b^{p+1-s}$ où $\theta = 0$ ou 1 , la deuxième alternative ayant toujours lieu lorsque $p - q = s - 1$, donc : $e(U) = q \vee q + 1$.

$U \ominus \tilde{S} = U - \tilde{S} = - X^t$
$F X = X \ominus X^t$, donc :
$F X$ égale le dernier chiffre x_s de la mantisse de X, $F X = X - X^t$, à la condition que $p - s + 1 \geqslant m_2$, sinon $F X = 0$ même si $X - X^t \neq 0$ mais dans ce cas, $|E| < \xi$. Si ce cas particulier $\mathcal{P}$ n'a pas lieu :
$V = U \ominus b\, .\, F X = U - b\, .\, F X$, d'après l'écriture de U .
Il vient dans tous les cas
$e(V) \leqslant q$ ($e(U) = q + 1$ n'ayant lieu que si $x_s \neq 0$) .

Les mantisses de V et de $W = b\, FX \ominus FX = b\, FX - FX$ se cadrant sans erreur pour l'exposant p :
$(V + W)_i = V + W = \tilde{S} - X$
$|\tilde{S} - X| \leqslant |Y| \Longrightarrow V \oplus W = \tilde{S} - X$
et
$E = - ((\tilde{S} - X) \ominus Y) = X + Y - \tilde{S}$

dès que $E \in T$.
D'où la proposition.

Si $\mathcal{P}$ a lieu, la formule (II. 5) prend la forme : $E = Y \ominus U$

$\star$ $e(U) = q \Longrightarrow (Y - U)_i = [X + Y - \tilde{S}] + [X^t - X]$

les 2 quantités entre crochets étant de signes contraires et de module inférieur à ξ , $Y \ominus U = 0$.

$\star$ $e(U) = q + 1 \Longrightarrow (Y - U)_i = Y^\zeta - U$ où $Y^\zeta = \varepsilon \lfloor \frac{m}{b} \rfloor b^{q+1-s}$

$(Y - U)_i = [X + Y - \tilde{S}] + [X^t - X + Y^\zeta - Y]$

$q \neq p \Longrightarrow Y \ominus U = 0$

mais les conditions $p = q = m_2 + s - 2 \wedge \lfloor \frac{m}{b} \rfloor = b^{s-1} - 1 \wedge x_s + y_s \geqslant b$

où y_s est le dernier chiffre de la mantisse de Y , entraînent :

$Y \ominus U = - \varepsilon \xi$

alors que l'erreur est de module inférieur à ξ .

En conclusion, si $\tilde{E}$ est la valeur calculée par (II. 5) lorsque $|E| < \xi$, on peut seulement affirmer :

$$|E - \tilde{E}| \leqslant 2\xi \left(\frac{b-1}{b}\right)$$

Remarque :

L'écriture $b \cdot FX \oplus (b \cdot FX \ominus FX)$ est imposée par le cas très particulier où $p = q \wedge \lfloor \frac{m}{b} \rfloor = b^{s-1} - 1 \wedge x_s + y_s \geqslant b$.

Ainsi :
$b = 10$, $s = 3$, $X = 0{,}999$, $Y = 0{,}991$ entraîneraient, avec les notations précédentes :
$U \ominus FX \ominus Y = 0{,}01$ introduisant l'erreur b^{p+1-s} , alors que $\tilde{S} = S$

I. 3. 1. b. Les nombres négatifs sont représentés en notation de complément à 2 (base puissance de 2)

Dans la définition donnée, de l'addition pour la règle de troncature sans chiffre de garde, on doit remplacer :

- la mantisse des nombres par la représentation de ces mantisses en complément à 2.
- l'addition algébrique des mantisses par la somme arithmétique, modulo 2, de leurs représentations, sauf dans le cas où se produit un dépassement de capacité de mantisse auquel cas on tronque le résultat du chiffre de poids le plus faible.

Donc :

Lorsque Y est positif , $X \oplus Y$ prend les valeurs écrites au § précédent

Lorsque Y est négatif, le cadrage de Y s'est effectué en remplaçant

Y par $Y^a = - \lceil m\, b^{q-p} \rceil\, b^{p-s}$.

Il vient immédiatement :

$X \oplus Y = (\varepsilon 1 - \lceil m\, b^{q-p} \rceil)\, b^{p-s} \quad \vee\ 0$ lorsque $e(S) \leqslant p$

$X \oplus Y = - \left\lceil \frac{1 + .mb^{q-p}}{b} \right\rceil\, b^{p+1-s}$ lorsque $e(S) = p + 1$

Propriété.

Lorsque l'opération $\oplus$ est définie par la règle de troncature sans chiffre de garde, les nombres négatifs étant représentés en notation de complément à 2 , on peut énoncer :

1) L'erreur E vérifie (II. 4) .

2) Si l'on complète la définition de $\oplus$ par :

$p - q \geqslant s \Longrightarrow X \oplus Y = X$, quels que soient les signes de X et de Y ,

Alors E est élément de T dès que $|E| \notin]0, \xi[$

3) L'opération $\oplus$ est G - dirigée $\forall (X,Y) \in T^2 \wedge X + Y \notin]-\xi, 0[$

4) La symétrie $(X \to - X)$ n'est plus un automorphisme de T .

Expression de E par un produit de composition sur T :

La succession d'opérations de la formule (II. 5) ne définit plus E dans tous les cas. Il suffit, en effet, de considérer par exemple, le cas suivant :

$X > 0$, $Y < 0$, donc si $Y^a = - \lceil m\, b^{q-p} \rceil\, b^{p-s}$, $\tilde{S} = X + Y^a$;

Il vient $U = Y^a$, mais le cas particulier où $\lceil m\, b^{q-p} \rceil = b^{s-p+q}$ entraînant $e(U) = q + 1$, le chiffre de poids le plus faible de la mantisse de Y sera perdu lors du calcul $U \ominus Y$.

O. MOLLER [28] a montré que dans le cas où $b = 2$, les nombres négatifs étant représentés en complément à 2, l'erreur E est donnée par :

$E = (Y \ominus (\tilde{S} \ominus X)) \oplus (X \ominus (\tilde{S} \ominus (\tilde{S} \ominus X))) \oplus (Y \ominus ((\tilde{S} \ominus X) \oplus (Y \ominus (\tilde{S} \ominus X)))$

avec la condition $|X| \geqslant |Y|$,

la partie principale de l'erreur E étant constituée par les deux premiers termes $(Y \ominus (\tilde{S} \ominus X)) \oplus (X \ominus (\tilde{S} \ominus (\tilde{S} \ominus X)))$.

O. MOLLER a établi en outre que cette dernière écriture donne, sauf dans de rares cas particuliers, une bonne approximation de l'erreur commise même si $|X| < |Y|$.

I. 3. 2. La troncature avec chiffre de garde [31].

Définition :

La règle définissant l'addition sur T est une règle de troncature avec chiffre de garde, si l'opération d'addition de deux éléments X et Y de T $(e(X) \geqslant e(Y))$, s'effectue suivant le schéma suivant :

1) L'exposant provisoire de la somme $X \oplus Y$ est pris égal à $e(X)$.
   X est écrit comme un nombre normalisé de $(s + 1)$ chiffres de mantisse, par addition d'un zéro après le chiffre de poids le plus faible de sa mantisse.

Si $e(Y) = e(X)$, il en est de même de Y, sinon Y est cadré, ce cadrage s'effectuant en déplaçant de $e(X) - e(Y)$ positions vers la droite la mantisse de ce nombre, et en tronquant les $e(X) - e(Y) - 1$ derniers chiffres de cette mantisse.

2) Les mantisses qui sont à présent alignées, sont ajoutées algébriquement, puis le résultat obtenu est normalisé, et enfin sa mantisse est tronquée à s chiffres.

I. 3. 2. a. Les nombres négatifs sont représentés par leur signe et leur valeur absolue .

X est exprimé sous la forme $\varepsilon\ (bl)\ b^{p-s-1}$; Y est "approché" sous la forme $\varepsilon'\ \lfloor m\ b^{q-p+1} \rfloor\ b^{p-s-1}$ (cette valeur étant toujours exacte si $q - p + 1 \geqslant 0$) .

Les valeurs prises par $\tilde{S} = X \oplus Y$ sont :

Si $\varepsilon\ \varepsilon' > 0$:

$$\tilde{S} = \varepsilon\ (\ 1 + \lfloor m\ b^{q-p} \rfloor\ )\ b^{p-s} \qquad \text{lorsque } e(\tilde{S}) = p$$

$$\tilde{S} = \varepsilon \left\lfloor \frac{1 + m\ b^{q-p}}{b} \right\rfloor b^{p+1-s} \qquad \text{lorsque } e(\tilde{S}) = p + 1$$

Si $\varepsilon\ \varepsilon' < 0$:

$$\tilde{S} = \varepsilon \left\lfloor \frac{bl - \lfloor mb^{q-p+1} \rfloor}{b} \right\rfloor b^{p-s} \qquad \text{lorsque } e(\tilde{S}) = p$$

$$\tilde{S} = \varepsilon\ (\ bl - \lfloor mb^{q-p+1} \rfloor\ )\ b^{p-s-1} \text{ (ou 0)} \qquad \text{lorsque } e(\tilde{S}) < p$$

(On a exprimé dans ce dernier cas, la valeur de $\tilde{S}$, et non l'écriture après normalisation mettant en évidence l'exposant réel de ce nombre).

Proposition II. 8.

L'opération $\oplus$ pour la troncature avec chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue, est :

1) correcte

2) I - dirigée pour tout couple $(X, Y) \in T^2_\wedge \; X\,Y > 0$, mais n'a pas de direction définie pour l'ensemble des couples $(X , Y) \in T^2_\wedge \; X\,Y < 0$.

3) La symétrie est un automorphisme de T .

L'addition de deux nombres de même signe est correcte.

Pour mettre en évidence ce résultat pour l'addition de deux nombres de signes différents, écrivons, si $q \neq p$, les divisions arithmétiques suivantes :

$m = b^{p-q-1} \quad m_1 + r_1 \wedge r_1 < b^{p-q-1}$

$b1 - m_1 = bq_1 + r_2 \wedge r_2 < b$

On a alors une expression commode de $E = S - \tilde{S}$ lorsque $q \neq p$:
($E = 0$, s'il ne se produit pas d'underflow, pour $q = p$)

. Si $e(\tilde{S}) = p \qquad E = \varepsilon(r_2\, b^{p-q-1} - r_1)\, b^{q-s}$ et $-b^{p-1-s} < \varepsilon E < b^{p-s}$

. Si $e(\tilde{S}) < p \qquad E = -\varepsilon\, r_1\, b^{q-s}$ et $|E| < b^{p-1-s}$

Dans le cas où $e(\tilde{S}) < p$ remarquons :

$$q = p \vee q = p - 1 \Longrightarrow E = 0 \qquad (\text{ou } |E| < \xi)$$

$$q < p - 1 \Longrightarrow e(\tilde{S}) = p - 1$$

La majoration de E est donc, dans tous les cas :

$|E| < b^{P-s}$ où $P = e(\tilde{S})$ si $\tilde{S} \neq 0$

$|E| < \xi$ si $\tilde{S} = 0$

Ceci établit que $\oplus$ est correcte, à la condition de s'assurer que la propriété est bien vérifiée dans le cas particulier $\tilde{S} = \varepsilon b^{p-1}$, ce qui est immédiat par $\varepsilon E > - b^{p-1-s}$.

L'opération $\oplus$ pour la troncature avec chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue, vérifie les hypothèses de la proposition II. 2.

La formule (II. 1) permet donc, pour cette arithmétique, le calcul de E .

(Il est toutefois nécessaire de remarquer que, si E est la valeur théorique de l'erreur, et $\tilde{E}$ la valeur calculée par (II. 1) :

$\tilde{E} = E$ si $E \in T$

$|\tilde{E} - E| < \xi$ si $|E| < \xi$

($\tilde{E} = 0$ en général lorsque $|E| < \xi$, sauf dans le cas très particulier où les conditions nécessaires $e(\tilde{S}) = p + 1$, $q = m_2$, $p - q \leqslant s - 2$ peuvent entraîner $\tilde{E} = Y$)) .

I. 3. 2. b. Les nombres négatifs sont représentés en notation de complément à 2

Désignons par $\tilde{\sigma}_*$ la représentation en complément à 2 de la mantisse $\tilde{\sigma}$ de $\tilde{S} = X \oplus Y$.

σ étant défini par : $S = X + Y = \sigma \cdot b^{e(S)}$, posons :

$\sigma_* = \sigma$ si $\frac{1}{b} \leqslant \sigma < 1$; $\sigma_* = 2 - |\sigma|$ si $-1 < \sigma \leqslant -\frac{1}{b}$.

Il vient :

$\tilde{\sigma}_* = (\sigma_*)_s$

la notation $(\ )_s$ définissant la troncature à s chiffres après le bit de signe.

La propriété est, en effet, évidente si $e(S) = p \vee p + 1 \vee p - 1$, et d'autre part $e(S) < p - 1$ ne peut avoir lieu que pour $q = p \vee p - 1$, et dans ce cas $\tilde{\sigma} = \sigma$ si un underflow n'apparait pas.

Propriété.

L'opération $\oplus$ pour la troncature avec chiffre de garde, les nombres négatifs étant représentés en notation de complément à 2 , est :

1) correcte

2) G - dirigée $\forall (X, Y) \in T^2 \wedge X + Y \notin]-\xi, 0[$.

Si l'on complète la définition de $\oplus$ par :

$p - q > s \Longrightarrow X \oplus Y = X$ quels que soient les signes de X et de Y , afin que (C) ait lieu,

E est élément de T si la formation de $\tilde{S}$ n'a pas donné lieu à un underflow.

Le calcul de E se fera par la formule (II . 1) si $b = 2$, par la formule (II · 2) sinon.

I. 3. 3. L'arrondi sur chiffre de garde

Définition :

La définition de $X \oplus Y$ s'effectue comme en I. 3. 2. , mais en 2) , si la normalisation en fin d'opération n'a pas nécessité un décalage à gauche, le $(s + 1)^{ième}$ chiffre de la mantisse du résultat est non pas tronqué, mais "arrondi" à l'aide de la règle suivante :

a) Si la valeur du $(s + 1)^{ième}$ chiffre de mantisse est inférieure à $b/2$, le résultat est tronqué à s chiffres.

b) Si la valeur de ce $(s + 1)^{ième}$ chiffre est supérieure à $b/2$, on ajoute une unité au $s^{ième}$ chiffre de mantisse, on renormalise éventuellement et on tronque la mantisse à s chiffres.

c) Si la valeur du $(s + 1)^{ième}$ chiffre de mantisse est égal à $b/2$, selon les calculateurs :

c - 1 . Il est opéré comme indiqué en b)

c - 2 . Il est défini un processus d'addition ou de non-addition d'une unité au $s^{ième}$ chiffre de mantisse, évitant le biais introduit par c - 1 .

La définition donnée de l'arrondi sur chiffre de garde s'appliquera à une représentation des nombres par signe et valeur absolue.

(Signalons que certains calculateurs utilisant la notation de complément à 2 pour représenter la mantisse des nombres négatifs traduisent en fin d'opération mais avant arrondi, le résultat sous forme valeur absolue et signe.)

Le mode c - 1 d'arrondi est supposé choisi ci-dessous.

On peut énoncer

Propriété.

L'opération $\oplus$ pour l'arrondi sur chiffre de garde, est :

1) correcte

2) optimale $\forall (X, Y) \in T^2 \wedge X Y > 0$ si b est pair (voir restriction $\mathcal{R}_1$)

3) vérifie la propriété (C) (voir restriction $\mathcal{R}_1$ et $\mathcal{R}_2$)

4) pour $b = 2$ et $b = 3$, l'erreur E vérifie, si $E \in T$, la formule (II. 1)

Les propriétés 1) 2) 3) énoncées viennent immédiatement. Il convient de noter, pour ces résultats, que :

1. L'addition arithmétique pour l'arrondi sur chiffre de garde n'est pas optimale si b est impair.

On remarquera [15], en effet, que lorsque b est impair, un seul chiffre ne suffit pas pour contrôler l'arrondi, sauf si la notation "alternée" est utilisée, auquel cas l'arrondi se ramène simplement à la troncature. Nous excluons cette possibilité dans ce paragraphe.

2. Pour une base puissance de 2, lorsque les nombres négatifs sont représentés en notation de complément à 2 , les propriétés 2) et 3) ne sont plus vérifiées, si précaution n'est pas prise dans les règles définissant représentation et normalisation des mantisses, dans les cas particuliers suivants :

$\mathcal{R}_1$: $p - q > s$ (ou =) ; $X = -(b^s - 1)\, b^{p-s}$; $Y < 0$ pour 2) et 3)

$\mathcal{R}_2$: $p - q > s$; $X = b^{p-1}$; $Y < 0$ pour 3)

qui peuvent conduire à $\tilde{S} = (X)'$.

La formule (II. 1) calculerait, dans ces cas particuliers, E lorsque $E \in T$, la partie principale de E sinon.

Enfin, pour la propriété 4) , on utilisera, d'une part la proposition II. 3 , et d'autre part, on remarquera que pour $b = 3$, l'addition arithmétique pour l'arrondi sur chiffre de garde, de deux nombres de même exposant $q = p$ implique :
$|E| = \varphi \cdot b^{p-s}$ où $\varphi \in \{0, 1\}$, et par suite $\tilde{S} - X = Y - E$ est toujours élément de T .

I. 3. 4. Opérations ⊕ optimales

Rappelons seulement qu'il n'est pas nécessaire d'effectuer l'addition sur un accumulateur à double longueur, pour réaliser une opération ⊕ optimale , lorsque b est pair.

Lorsque les nombres sont représentés avec signe et valeur absolue D. E. KNUTH [15] souligne en effet, que pour un schéma optimal de l'addition de X et Y , il suffit, lors du cadrage de Y , de tronquer la mantisse de ce nombre de la façon suivante :
Si X et Y ont le même signe , Y sera "approché" sous la forme $\varepsilon' \lfloor m\, b^{q-p+2} \rfloor\, b^{p-s-2}$, sinon sous la forme $\varepsilon' \lceil m\, b^{q-p+2} \rceil\, b^{p-s-2}$
(L'arrondi sur le $(s+1)^{\text{ième}}$ chiffre de la mantisse du résultat est alors effectué comme § précédent).

Pour une base puissance de 2 , lorsque les nombres négatifs sont représentés en notation de complément à 2 , il suffira de tronquer, lors du cadrage de Y , la mantisse de ce nombre avec deux chiffres de garde. Toutefois, $\oplus$ ne sera pas optimale dans le cas particulier $\mathcal{R}_1$, lorsque la configuration de mantisse 1 0 ..... 0 est transformée en 1 (b - 1) 0— 0 , avec augmentation d'une unité pour l'exposant correspondant.

I. 3. 5. Une remarque pratique

Nous signalerons le travail de M. A. MALCOLM [25] qui a établi deux algorithmes permettant de déterminer automatiquement, sur les calculateurs disposant d'une arithmétique à virgule flottante correcte, les caractéristiques suivantes de cette arithmétique :

- base de la représentation
- nombre de chiffres de mantisse du système de nombres flottants
- propriété satisfaite par l'opération d'addition :
  $\oplus$ correcte ou $\oplus$ optimale.

Ces algorithmes sont présentés sous forme de procédures FORTRAN.
Dans le cas de calculateurs disposant de registres flottants d'une longueur supérieure à la longueur de la représentation du mot-machine-et effectuant dans ces registres les calculs avec une précision supérieure à la précision avec laquelle les résultats peuvent être représentés en mémoire-il conviendra de se reporter aux modifications des procédures précédentes établies par W.MORVEN GENTLEMAN et S.B.MAROVICH [10] .

Nous mentionnerons enfin la récente étude de S.LINNAINMAA [21]: Cet auteur postérieurement à [33],mais par des démonstrations différentes, établit des conditions nécessaires et suffisantes de validité des formules (II.1) et (II.2).L'arithmétique "paritaire" est aussi étudiée et des résultats expérimentaux sont donnés par l'auteur.

II - L'ERREUR D'ARRONDI DANS LA MULTIPLICATION

II . 1. EXPRESSION, PAR UN PRODUIT DE COMPOSITION SUR T , DE L'ERREUR D'ARRONDI

La valeur sur $\mathbb{R}$ du produit de deux nombres à virgule flottante étant un mot de longueur $2s$, le calcul de l'erreur d'arrondi dans la multiplication reviendra à réaliser une quasi double-précision.

On peut énoncer :

Proposition II.9.

Si s est pair et si les opérations $\otimes$ et $\ominus$ sont correctes, l'erreur $E = X\,Y - (X \otimes Y)$, entachant la multiplication en arithmétique à virgule flottante de deux éléments X et Y de T, vérifie, si aucun underflow n'apparaît lors des opérations définies ci-dessous, la formule suivante :

$$E = -(X \otimes Y \ominus P_1 \ominus P_2 \ominus P_2' \ominus P_3) \qquad \text{(II-6)}$$

où

$P_1 = X_1 \otimes Y_1$; $P_2 = X_1 \otimes Y_2$; $P_2' = X_2 \otimes Y_1$; $P_3 = X_2 \otimes Y_2$

X_1 et X_2 (resp. Y_1 et Y_2) étant définis à partir de $X = \varepsilon\, l\, b^{p-s}$ (resp. $Y = \varepsilon'\, m\, b^{q-s}$) par :

$$\begin{aligned} X_1 &= \varepsilon \lfloor l\, b^{-s/2} \rfloor\, b^{s/2} \,.\, b^{p-s} \\ X_2 &= X - X_1 \qquad \text{(II-7)} \\ Y_1 &= \varepsilon' \lfloor m\, b^{-s/2} \rfloor\, b^{s/2} \,.\, b^{q-s} \\ Y_2 &= Y - Y_1 \end{aligned}$$

Tout d'abord, il vient immédiatement :

Si l'opération $\otimes$: $T \times T \to T$ est correcte, alors $\forall (X,Y) \in T^2$, l'erreur $E = X\,Y - (X \otimes Y)$ est élément de T dès que $|E| \notin]0,\xi[\wedge |X\,Y| < b^{m_1}$.

En effet, si $\xi \leqslant |X\,Y| < b^{m_1}$, $|E|$ est du type $\lambda\, b^{p+q-2s}$ où λ est entier et vérifie $0 \leqslant \lambda < b^s$.

Lorsque $X \otimes Y$ provoque un overflow, hypothèse que nous excluerons, E n'est pas en général élément de T.

(II-6) exprime E à l'aide d'opérations internes définies sur T.

Démonstration :

Posons :

$P = X\ Y$

$\hat{P} = X \otimes Y$

$E = P - \hat{P}$

Si la formation de $\hat{P}$, P_1, P_2, P_2', P_3 et E ne provoque pas d'underflow, il suffit pour établir la proposition, $\ominus$ étant correcte, de démontrer que les trois nombres $\hat{P} - P_1$, $\hat{P} - P_1 - P_2$, $\hat{P} - P_1 - P_2 - P_2'$ sont éléments de T.

Soit $\pi = p + q - \nu$ ($\nu = 0$ ou 1) l'exposant de P.

Si $l_1 = \lfloor l\, b^{-\frac{s}{2}} \rfloor$, $m_1 = \lfloor m\, b^{-\frac{s}{2}} \rfloor$, $l_2 = l - l_1$, $m_2 = m - m_1$,

posons :

$P_2^t = \varepsilon\, \varepsilon' \lfloor l_1\, m_2\, b^{-\frac{s}{2}+\nu} \rfloor\, b^{\pi-s}$ et $R_2 = P_2 - P_2^t$

$P_2'^t = \varepsilon\, \varepsilon' \lfloor l_2\, m_1\, b^{-\frac{s}{2}+\nu} \rfloor\, b^{\pi-s}$ $R_2' = P_2' - P_2'^t$

$P_3^t = \varepsilon\, \varepsilon' \lfloor l_2\, m_2\, b^{-s+\nu} \rfloor\, b^{\pi-s}$ $R_3 = P_3 - P_3^t$

En écrivant :

$R_2 + R_2' + R_3 = \varepsilon\, \varepsilon' (\lambda\, b^{\pi-s} + R)$ où $\lambda \in \{0 , 1 , 2\}$ et $0 \leqslant R < b^{\pi-s}$

il vient :

$\hat{P} = P_1 + P_2^t + P_2'^t + P_3^t + \varepsilon\, \varepsilon'(\lambda + \theta)\, b^{\pi-s}$

avec $\theta = 0$ ou 1 suivant que $\otimes$ est I ou E - dirigée pour le couple (X,Y).

De l'écriture :

$\hat{P} - P_1 = P_2^t + P_2'^t + P_3^t + \varepsilon\, \varepsilon'(\lambda + \theta)\, b^{\pi-s}$

on tire immédiatement que $\hat{P} - P_1$ est de la forme $\varepsilon\, \varepsilon'\, \mu\, b^{\pi-s}$, où μ est entier et vérifie :

$0 \leqslant \mu \leqslant 2(b^{\frac{s}{2}} - 2)\, b^{\nu} + b^{\nu} - 1 + 3 < b^{s}$ dès que $s \geqslant 4$.

$\hat{P} - P_1 \in T$.

Puis :

$\hat{P} - P_1 - P_2 = P_2'^t + P_3^t + \varepsilon\, \varepsilon' (\lambda + \theta)\, b^{\pi-s} - R_2$

- Si $\nu = 1$, $P_2'^t + P_3^t + \varepsilon\, \varepsilon' (\lambda + \theta)\, b^{\pi-s}$ est de la forme $\varepsilon\, \varepsilon'\, \mu_1\, b^{\pi-s}$ où : μ_1 est entier et vérifie $0 \leqslant \mu_1 \leqslant b^{\frac{s}{2}+1} - b + 2 \leqslant b^{\frac{s}{2}+1}$

  En tenant compte de l'écriture de R_2 :

  $\hat{P} - P_1 - P_2 \in T$.

- Si $\nu = 0$, alors $P_3^t = 0$, $\max |P_2'^t| = (b^{\frac{s}{2}} - 2)\, b^{\pi-s}$ cette valeur ne pouvant être obtenue que lorsque $l_2 = m_1 = b^{\frac{s}{2}} - 1$.

La condition $l_2 = m_1 = b^{\frac{s}{2}-1}$ impliquant $\lambda \neq 2$,

$P_2'^{t} + \varepsilon\ \varepsilon'\ (\lambda + \theta)\ b^{\pi-s}$ est de la forme $\varepsilon\ \varepsilon'\ \mu_2\ b^{\pi-s}$

où μ_2 est entier et vérifie $0 \leqslant \mu_2 \leqslant b^{\frac{s}{2}}$.

Par ce résultat et l'écriture de R_2 , il vient encore :

$\tilde{P} - P_1 - P_2 \in T$.

Enfin :

$\tilde{P} - P_1 - P_2 - P_2' = P_3^t + \varepsilon\ \varepsilon'\ (\lambda + \theta)\ b^{\pi-s} - R_2 - R_2'$

implique que ce nombre est du type $N\ b^{p+q-\frac{3s}{2}}$

où N est entier et est composé de $\frac{s}{2} + 2$ chiffres au maximum.

Par suite

$\tilde{P} - P_1 - P_2 - P_2' \in T$ dès que $s \geqslant 4$.

Remarques

1) On ne change pas (II-6) en permutant P_2 et P_2' , sauf en cas d'apparition d'un underflow.

2) En raison des underflows, qui peuvent survenir lors de la formation de P_1 , P_2 , P_2' , P_3 ou lors des soustractions définies en (II-6), on n'est assuré d'obtenir par (II-6) l'erreur E dès que $E \in T$, seulement si est réalisée la condition suffisante suivante :

$$p + q \geqslant m_2 + 2\ s - 1$$

impliquant $E \in T$.

En effet, si cette condition sur les exposants n'est pas vérifiée, on n'a pas les implications suivantes, où $\tilde{E}$ désigne la valeur calculée par (II-6) :

$|E| < \xi \Longrightarrow \tilde{E} = 0$

et

$E \in T \Longrightarrow \tilde{E} = E$

Dans le cas de produits trop "petits", il faudrait pour utiliser (II-6) procéder préalablement à un recadrage.

On peut aussi énoncer :

Proposition II.9'.

La proposition II.9 est encore vraie si l'on remplace X_1 , X_2 , Y_1 , Y_2 par X'_1 , X'_2 , Y'_1 , Y'_2 définis par

$$X'_1 = \varepsilon \left\lceil 1\ b^{-\frac{s}{2}} \right\rceil b^{\frac{s}{2}} \cdot b^{p-s}$$

$$X'_2 = X - X'_1 \qquad \text{(II-7')}$$

$$Y'_1 = \varepsilon' \left\lceil m\ b^{-\frac{s}{2}} \right\rceil b^{\frac{s}{2}} \cdot b^{q-s}$$

$$Y'_2 = Y - Y'_1$$

La démonstration de cette proposition est analogue à la démonstration précédente.

Lorsque s est impair, un calcul exact du produit de 2 nombres flottants à l'aide d'une arithmétique à mot court, nécessite que l'on partage X (et Y) en utilisant un arrondi afin que les produits P_1 , P_2 , P'_2 , P_3 soient tous éléments de T.
G.W. VELTKAMP [40] a étudié (II-6) dans de telles hypothèses.

T.J. DEKKER [5] a établi un algorithme différent du précédent pour exprimer par un couple de mots courts, la valeur exacte du produit X Y , avec les hypothèses suivantes :

s quelconque
⊕ optimale
⊗ correcte.

Le procédé de cet auteur, calculant directement poids fort et poids faible du produit, nécessite une multiplication de moins et une addition de plus que (II-6).

II.2. FORMULES PRATIQUES

Les nombres X_1 , X_2 , Y_1 , Y_2 (resp. X'_1 , X'_2 , Y'_1 , Y'_2) ont été définis à partir de X et Y par (II.7) ou (II.7') de façon théorique. Lorsqu'on cherche à exprimer ces nombres à l'aide d'opérations internes sur T , une réponse simple à ce problème est donnée dans le cas où $b = 2$ par :

Proposition II.10.

Si les hypothèses de la proposition II.9 sont vérifiées, et si, de plus, $b = 2$ et $\oplus$ est I - dirigée pour tout couple $(U,V) \in T^2_{\wedge}$ $U \, V > 0$, alors l'égalité (II.6) est vérifiée pour x_1, x_2, y_1, y_2 définis à partir de X et Y par :

$$x_1 = Z_1 \oplus X \ominus Z_1 \quad , \quad x_2 = X - x_1$$
$$y_1 = Z_2 \oplus Y \ominus Z_2 \quad , \quad y_2 = Y - y_1 \qquad \text{(II.8)}$$

où

$$Z_1 = B \otimes X \quad , \quad Z_2 = B \otimes Y \quad \text{et} \quad B = b^{s/2}.$$

Démonstration :

Montrons, en effet, que la formule définissant x_1 (respectivement y_1) implique $x_1 = X_1$, sauf dans certains cas particuliers :

Soit x_s le dernier chiffre de la mantisse de X.

. Si $1 \leqslant b^s - b^{s/2} \vee (1 > b^s - b^{s/2} \wedge x_s = 1)$, il vient :

$Z_1 \oplus X = Z_1 + X_1$ d'où $x_1 = X_1$.

. Si $1 > b^s - b^{s/2} \wedge x_s = 0$ (C_P) , il vient :

$x_1 = X_1 - (b - 1)\, b^{s/2} \,.\, b^{p-s}$

$x_2 = X_2 + (b - 1)\, b^{s/2} \,.\, b^{p-s}$

Il suffit, pour établir la proposition II.10, de montrer que l'égalité (II.6) est encore valable lorsque X (ou Y ou X et Y) vérifie la condition particulière C_P :

On remarquera que lorsque l'une de ces éventualités a lieu, les produits $x_1 y_2$, $x_2 y_1$, $x_2 y_2$ sont, sauf en cas d'underflow, éléments de T , et en reprenant la démonstration de (II.6) pour ces 3 cas particuliers, on vérifiera bien la propriété dès que $s \geqslant 6$.

Les formules (II.8) ne conviennent qu'avec les hypothèses précédentes ; en effet, si $b = 2$ mais $\oplus$ est E - dirigée pour des couples d'éléments de T de même signe, ou si $b \neq 2$, on n'est plus assuré dans tous les cas que $x_1 y_2$, $x_2 y_1$ et $x_2 y_2$ appartiennent à T.
Un "découpage" du nombre flottant X est alors une fonction plus compliquée de T dans T.

Ainsi :

Proposition II.11.

Si s est pair , et si l'opération $\oplus$ est définie par une règle de troncature avec chiffre de garde, les nombres négatifs étant représentés par leur signe et leur valeur absolue,

alors , $\forall\ X \in T \wedge X \neq \varepsilon(b^{s-1} + b^{\frac{s}{2}-1})\ b^{p-s}$, l'élément X_1 de T , déduit de X par (II.7) vérifie :

$X_1 = Z \oplus X \ominus Z$

où $Z = b^{s/2} X \ominus X$.

Démonstration :

On établit cette proposition en vérifiant que les hypothèses choisies entraînent l'égalité :

$$Z \oplus X = Z + X_1 \qquad \forall\ X \in T \wedge X \neq \varepsilon(b^{s-1} + b^{\frac{s}{2}-1})\ b^{p-s}$$

En effet, si $X = \varepsilon\ l\ b^{p-s}$, il vient :

$\star\ l > b^{s-1} + b^{\frac{s}{2}-1} \Longrightarrow e(Z) = p + \frac{s}{2}$, avec de plus, en désignant par $x_{\frac{s}{2}+1}$ le $(\frac{s}{2} + 1)^{\text{ième}}$ chiffre de la mantisse de X :

$$Z = b^{\frac{s}{2}} X - X_1 - \varepsilon\, b^{p-\frac{s}{2}} \qquad \text{si} \quad x_{\frac{s}{2}+1} \neq 0$$

$$Z = b^{\frac{s}{2}} X - X_1 \qquad \text{si} \quad x_{\frac{s}{2}+1} = 0$$

Par suite, dans les 2 cas :

$Z \oplus X = Z + X_1$

$\star\ 1 < b^{s-1} + b^{\frac{s}{2}-1} \Longrightarrow e(Z) = p + \frac{s}{2} - 1$, avec

$Z = b^{s/2} X - X_1$ car $x_{\frac{s}{2}+1} = 0$

Il vient encore $Z \oplus X = Z + X_1$

$\star\ 1 = b^{s-1} + b^{\frac{s}{2}-1} \Longrightarrow Z \oplus X \ominus Z = X$

On concluera pour la proposition, l'opération $\oplus$ étant correcte.

Remarquons que, par contre, si $\oplus$ et $\otimes$ sont correctes, et I - dirigées $\forall (U , V) \in T^2_{\wedge}\ U\ V > 0$ pour $\oplus$, et $\forall (U , V) \in T^2$ pour $\otimes$, la fonction :

$\varphi(X) = (\mathcal{B} \otimes X) \oplus X \ominus (\mathcal{B} \otimes X)$ où $\mathcal{B} = b^{s/2-1}$

n'implique pas $\varphi(X) = X_1$, $\forall X \in T_{\wedge}\ X \neq \varepsilon(b^{s-1} + b^{\frac{s}{2}-1})\ b^{p-s}$

Ainsi :

$b = 10$, $s = 6$, $X = 0{,}100034 \longrightarrow \varphi(X) = 0{,}0991.$

Remarque :

Les expressions permettant de calculer X_1 à l'aide d'opérations internes sur T étant souvent compliquées, il sera parfois préférable de calculer X_1 en utilisant des propriétés simples de représentation des nombres et de programmation. (Voir programme ci-dessous).

Pour une arithmétique satisfaisant les hypothèses de la proposition II.11 ,l'application successive ,de la formule établie dans cette proposition et de (II.6), calcule exactement l'erreur élémentaire de multiplication sauf dans le cas de l'élévation au carré d'un nombre de mantisse $\varepsilon(b^{-1} + b^{-s/2-1})$ [±0,100100 sur IBM-360].

II.3. ETUDE EXPERIMENTALE

A titre d'exemple, nous avons effectué le calcul des produits X Y où $X = \frac{1}{10+N}$, $Y = \frac{1}{250+N}$, $0 \leqslant N \leqslant 999$.

Ces calculs ont été effectués sur une IBM 360-30 : Les hypothèses des propositions II.9 et II.11 sont vérifiées.

Trois programmes ont été comparés :

Etant donnés les deux éléments X et Y de T et leur produit sur T , $P = X \otimes Y$:

les programmes 1 et 2 utilisaient respectivement, pour le calcul de $E = X Y - P$, les instructions FORTRAN suivantes : $(b=16, s=6, B=b^{s/2})$

Programme 1.
```
      B = 4096
      Z = B * X - X
     X1 = Z + X - Z
     X2 = X - X1
      Z = B * Y - Y
     Y1 = Z + Y - Z
     Y2 = Y - Y1
      E = - P + X1 * Y1 + X1 * Y2 + X2 * Y1 + X2 * Y2
```

Programme 2.
```
      EQUIVALENCE (J,X1) , (K,Y1)
     X1 = X
     Y1 = Y
      J = (J/4096) * 4096
      K = (K/4096) * 4096
     X2 = X - X1
     Y2 = Y - Y1
      E = - P + X1 * Y1 + X1 * Y2 + X2 * Y1 + X2 * Y2
```

Le programme 3 effectuait en double précision le produit XY.

Les différents temps d'exécution ont été successivement :

Programme 1	1 ' 49
Programme 2	1 ' 50
Programme 3	1 ' 54

note : Le programme 2, ainsi écrit, s'applique à des nombres flottants X et Y de signes positifs. Soulignons en effet que la représentation des nombres négatifs, dans un système IBM 360, utilise :
- la complémentation à 2 pour les nombres entiers
- signe et valeur absolue pour la mantisse des nombres réels.

III - L'ERREUR D'ARRONDI DANS LA DIVISION

Si (X,Y) est un élément de $T \times T - \{0\}$, l'erreur $E = X/Y - X \oslash Y$ n'est pas, en général, élément de T. Cependant E est exprimable comme la somme d'une série d'éléments de T, calculables par opérations internes sur T.

III.1. THEOREME PRELIMINAIRE

Proposition II.12.

Si l'opération $\oslash$ est correcte,
alors, quel que soit le couple (X,Y) de $T \times T - \{0\}$ vérifiant $\left|\frac{X}{Y}\right| < b^{m_1}$,
le nombre $C = E\,Y$ où $E = X/Y - (X \oslash Y)$, est élément de $T \cup]-\xi,\xi[$.

Démonstration :

Soient, en effet, $X = \varepsilon\, l\, b^{p-s}$ et $Y = \varepsilon' m\, b^{q-s}$ $(b^{s-1} \leqslant l,m < b^s)$ deux éléments de T.
Si $X \oslash Y$ provoque un overflow, hypothèse que nous avons exclue, C n'est pas en général élément de T, et dans le cas d'underflow, la propriété est évidente.

Dans le cas général, il vient :

$$X \oslash Y = \begin{cases} \varepsilon\,\varepsilon'\left(\left\lfloor \frac{l\, b^s}{m} \right\rfloor \vee \left\lceil \frac{l\, b^s}{m} \right\rceil\right) b^{p-q-s} & \text{si } l < m \\ \varepsilon\,\varepsilon'\left(\left\lfloor \frac{l\, b^{s-1}}{m} \right\rfloor \vee \left\lceil \frac{l\, b^{s-1}}{m} \right\rceil\right) b^{p-q+1-s} & \text{si } l \geqslant m \end{cases}$$

En posant, dans chacun des cas, la division arithmétique :

$$l\, b^{s-\theta} = m\, d + r \wedge r < m \qquad \theta = 0,1$$

on a :

$$C = \varepsilon\ (r_v r - m)\ b^{p+\theta-2s}$$

d'où immédiatement la proposition.

III.2. CALCUL DE L'ERREUR

III.2.1. Calcul de C

Notons :

$D = X/Y$

$\tilde{D} = X \oslash Y$

On obtiendra C en calculant sans erreur le nombre $X - \tilde{D}\ Y$, c'est-à-dire :

- soit, si l'on travaille en simple précision mais que l'on dispose de la double précision pour la multiplication et l'addition, en réalisant les 2 opérations $\tilde{D}.Y$ et $X - \tilde{D}\ Y$ en double précision - sous réserve que l'arithmétique définie pour la double précision soit correcte -

- soit si l'on calcule uniquement par une arithmétique à mot fixé, en utilisant, par exemple, sous réserve de sa validité dans l'arithmétique flottante employée, la formule suivante :

$$C = X \ominus P_1 \ominus P_2 \ominus P_2' \ominus P_3 \qquad (II\cdot 9)$$

où

$P_1 = \tilde{D}_1 \otimes Y_1$; $P_2 = \tilde{D}_1 \otimes Y_2$; $P_2' = \tilde{D}_2 \otimes Y_1$; $P_3 = \tilde{D}_2 \otimes Y_2$

$\tilde{D}_1$, $\tilde{D}_2$, Y_1 , Y_2 étant définis à partir de $\tilde{D}$ et Y par des formules analogues à (II.7).

Ainsi on établit :

Proposition II.13.

Si s est pair, si les opérations $\ominus$, $\otimes$, $\oslash$ sont correctes, et si $\oslash$ est I - dirigée pour tout couple de $T \times T - \{0\}$, alors, (II.9) est vérifiée à la condition suffisante qu'aucun underflow n'apparaisse lors des opérations définies par cette formule.

Démonstration :

La démonstration est analogue à la démonstration de la proposition II.9.

Posant :

$$m = m_1 b^{s/2} + m_2 \wedge m_2 < b^{s/2}$$

$$d = d_1 b^{s/2} + d_2 \wedge d_2 < b^{s/2}$$

on a la relation en entiers :

$$1\, b^{s-\theta} = m_1 d_1 b^s + (d_1 m_2 + d_2 m_1) b^{s/2} + d_2 m_2 + r \qquad (II.10)$$

traduction de :

$$X = P_1 + P_2 + P_2' + P_3 + C$$

où tous les termes écrits sont du signe ε de X.

Ceci met en évidence :

$$X - P_1 = \varepsilon (1 - m_1 d_1 b^{\theta}) b^{p-s} = \varepsilon \mu_1 b^{p-s}$$

où μ_1 est entier et vérifie : $0 \leqslant \mu_1 < 1$, impliquant $X - P_1 \in T$.

(II.10) exprime que $d_2 m_2 + r$ est divisible par $b^{s/2}$, soit :

$d_2 m_2 + r = \lambda b^{s/2}$,et il vient : $\lambda \leqslant 2 b^{s/2} - 2$.

Par suite :

$$X - P_1 - P_2 = \varepsilon (d_2 m_1 + \lambda) b^{p+\theta-\frac{3s}{2}} = \varepsilon \mu_2 b^{p+\theta-\frac{3s}{2}}$$

où μ_2 est entier et vérifie $0 \leqslant \mu_2 \leqslant b^s - 1$, impliquant $X - P_1 - P_2 \in T$.

Enfin

$$X - P_1 - P_2 - P_2' = \varepsilon \lambda b^{p+\theta-\frac{3s}{2}} \text{ , donc } X - P_1 - P_2 - P_2' \in T.$$

D'où la proposition , Θ étant correcte.

III.2.2. Algorithme d'obtention de E

Notons $\ominus$ et $\otimes$ les opérations de soustraction et de multiplication : en double précision par des opérations correctes

ou en pseudo double précision par des formules du type II.9 , c'est-à-dire vérifiant dans les 2 cas :

$X \otimes Y = X\,Y \qquad X \in T \ , \ Y \in T$

$X \ominus Y\,Z = X - Y\,Z \qquad X \in T \ , \ Y \in T \ , \ Z \in T \ , \ X - Y\,Z \in T$

Si l'opération $\oslash$ est correcte, le quotient $D = X/Y$ s'obtient immédiatement comme la somme de la série $\{\tilde{D}_n\}$ définie, ainsi que la série auxiliaire $\{C_n\}$, par les formules suivantes :

$$\tilde{D}_1 = X \oslash Y$$
$$C_1 = X \ominus \tilde{D}_1 \otimes Y \qquad \text{(II.11)}$$

et pour $n > 1$:

$$\tilde{D}_n = C_{n-1} \oslash Y$$
$$C_n = C_{n-1} \ominus \tilde{D}_n \otimes Y$$

III.2.3. Résultats numériques

Nous avons testé l'algorithme (II.11) sur une IBM 360-30 en simple précision, pour le calcul des trois premiers éléments, de la suite des nombres à virgule flottante, approchant $\frac{1}{X}$ (X = 1 , 2 , ...)

L'écriture de chaque élément du triplet représentant $\frac{1}{X}$ a été réalisée en hexadécimal, afin d'éviter toute erreur de conversion d'écriture des nombres selon des bases différentes. On a ainsi obtenu exactement la représentation de $\frac{1}{X}$ en triple précision.

X=5	$\tilde{D}_1$=40333333	$\tilde{D}_2$=3A333333	$\tilde{D}_3$=34333333
X=6	$\tilde{D}_1$=402AAAAA	$\tilde{D}_2$=3AAAAAAA	$\tilde{D}_3$=34AAAAAA
X=7	$\tilde{D}_1$=40249249	$\tilde{D}_2$=3A249249	$\tilde{D}_3$=34249249

IV - CONCLUSION

Lorsque se pose la question de comparer les 2 méthodes suivantes d'augmentation de la précision d'un résultat :

- utilisation d'un mot long
- ou expression du résultat d'opération élémentaire pour un couple de mots courts (partie principale et erreur) ,

il est possible de faire plusieurs remarques :

1) Il est préférable, lorsque cela est possible, d'exprimer le résultat **R** d'une opération élémentaire sous la forme d'un mot long, plutôt que par 2 mots courts, dans les cas de multiplications, et dans les cas d'additions où la propriété (C) n'intervient pas.
Cela résulte, en effet, de la forme de l'erreur qui vérifie dans les cas ci-dessus :

$$E = \lambda\, b^{e(R) - 2s} \quad \text{où } \lambda \text{ est entier et vérifie } 0 \leqslant |\lambda| < b^{s} .$$

Par contre, lorsque la condition $e(X) - e(Y) > s$ est réalisée lors de l'addition des 2 nombres X et Y , E n'est pas exprimable exactement dans l'écriture par mot long, et c'est l'écriture par couple de mots courts qui permet de garder toute la précision.

Cette forme est également la meilleure lorsqu'un quotient est nécessité avec de nombreux chiffres de mantisse, l'algorithme II. 11 permettant d'approcher un quotient avec une précision arbitraire.

2) Globalement -et non plus au niveau d'opération élémentaire- un résultat calculé en double précision est meilleur qu'un résultat, calculé en simple précision et corrigé par une correction.

Il ne faut, en effet, pas oublier l'importance prépondérante, pour l'obtention d'une précision suffisante, des 2 facteurs suivants :

* nombre suffisant des chiffres de mantisse
* utilisation d'une arithmétique correcte.

Par contre, l'amélioration apportée par une arithmétique, non seulement, correcte, mais optimale, n'est pas décisive.

3) Cependant, l'intérêt des méthodes exposées dans ce travail, est, outre que la double précision n'est pas accessible sur tous les calculateurs, leur possibilité d'itération.

Ces méthodes sont applicables à toute arithmétique, et leur utilisation pour une arithmétique à mot long, assurera une très grande précision dans les cas où cela est nécessaire.

4) Nous soulignerons, pour terminer, que le temps nécessité par les calculs dans la méthode d'expression de l'erreur comme poids faible d'un couple de mots courts, est d'un ordre de grandeur analogue au temps nécessité par la double précision.

V - APPENDICE. SOUS-PROGRAMMES FORTRAN

Nous joignons à ce chapitre, les sous-programmes FORTRAN, relatifs à l'étude précédente, que nous avons utilisé sur le système I.B.M. 360.

Les sous-programmes :

FUNCTION ES (X, Y, S)

" EP (X, Y, P)

" CD (X, Y, D)

calculent respectivement, à partir des données X et Y et des valeurs $S = X \oplus Y$, $P = X \otimes Y$, $D = X \oslash Y$, les erreurs :

$$ES = X + Y - S$$

$$EP = XY - P$$

et le terme correctif $CD = (X/Y - D)\ Y$.

SIMPLE PRECISION

```
      FUNCTION ES(X,Y,S)
      IF(ABS(X)-ABS(Y)) 1,1,2
    1 ES=-S+Y+X
      GOTO 3
    2 ES=-S+X+Y
    3 CONTINUE
      RETURN
      END
```

```
      FUNCTION EP(X,Y,P)
      EQUIVALENCE (J,X1),(K,Y1)
      L=4096
      U=ABS(X)
      V=ABS(Y)
      Q=ABS(P)
      X1=U
      Y1=V
      J=(J/L)*L
      K=(K/L)*L
      X2=U-X1
      Y2=V-Y1
      EP=-Q+X1*Y1+X1*Y2+X2*Y1+X2*Y2
      IF(P.LT.0.0) EP=-EP
      RETURN
      END
```

```
      FUNCTION CD(X,Y,D)
      EQUIVALENCE (J,Y1),(K,D1)
      L=4096
      U=ABS(X)
      V=ABS(Y)
      Q=ABS(D)
      Y1=V
      D1=Q
      J=(J/L)*L
      K=(K/L)*L
      Y2=V-Y1
      D2=Q-D1
      CD=U-D1*Y1-D1*Y2-D2*Y1-D2*Y2
      IF(X.LT.0.0) CD=-CD
      RETURN
      END
```

DOUBLE PRECISION

```
      REAL FUNCTION ES*8(X,Y,S)
      DOUBLE PRECISION X,Y,S
      IF(DABS(X)-DABS(Y)) 1,1,2
    1 ES=-S+Y+X
      GOTO 3
    2 ES=-S+X+Y
    3 CONTINUE
      RETURN
      END
```

```
      REAL FUNCTION EP*8(X,Y,P)
      REAL*8 X,Y,P
      REAL*8 Z,X1,X2,Y1,Y2
      B=268435456.0D0
      Z=B*X-X
      X1=Z+X-Z
      X2=X-X1
      Z=B*Y-Y
      Y1=Z+Y-Z
      Y2=Y-Y1
      EP=-P+X1*Y1+X1*Y2+X2*Y1+X2*Y2
      RETURN
      END
```

```
      REAL FUNCTION CD*8(X,Y,D)
      REAL*8 X,Y,D
      REAL*8 Z,D1,D2,Y1,Y2
      B=268435456.0D0
      Z=B*Y-Y
      Y1=Z+Y-Z
      Y2=Y-Y1
      Z=B*D-D
      D1=Z+D-Z
      D2=D-D1
      CD=X-Y1*D1-Y1*D2-Y2*D1-D2*Y2
      RETURN
      END
```

CHAPITRE III

APPLICATIONS

Introduction

Ce chapitre est relatif aux applications du calcul de l'erreur élémentaire comme poids faible d'un couple de mots en arithmétique à virgule flottante.

T désignera encore dans tout le chapitre le système T_b^s de nombres flottants, et les opérations élémentaires définies sur cet ensemble seront toujours notées $\oplus$, $\ominus$, $\otimes$, $\oslash$.

Le paragraphe I est consacré à un algorithme de correction d'une somme algébrique -ou d'un produit scalaire- .

Les paragraphes II, III et IV examinent l'utilisation des techniques décrites au chapitre précédent dans les trois domaines suivants :

- L'analyse à postériori des erreurs d'arrondi
- L'utilisation de méthodes de raffinement itératif
- La résolution numérique de problèmes mal conditionnés

Ces trois aspects différents sont traités sur des exemples.

I - UN ALGORITHME DE CORRECTION D'UNE SOMME ALGEBRIQUE

I . 1 . LE PROBLEME ETUDIE

Etant donné un système T de nombres à virgule flottante, muni de la loi d'addition $\oplus$, soient N nombres $X_i \in T$, $(1 \leqslant i \leqslant N)$, et $S = \sum_{i=1}^{N} X_i$ la somme de ces nombres sur $\mathbb{R}$.

Nous nous proposons d'obtenir, par des opérations internes sur T , un élément de T dont la distance à S soit la plus petite possible.

Posons : $S^1 = \bigoplus \sum_{i=1}^{N} X_i$

Si e_i désigne l'erreur d'arrondi lors de la ième addition dans le calcul de S^1 , on a immédiatement :

$S = S^1 + \sum_{i=1}^{N-1} e_i$.

Lorsque les erreurs e_i sont éléments de T , un procédé de correction de S^1 sera le suivant : on ajoutera, à cette valeur trouvée S^1 , la somme des e_i calculée sur T ; puis, cette dernière somme étant entachée d'erreur, on pourra effectuer une deuxième correction, et ainsi de suite.

Définition de l'algorithme $\mathcal{A}$

Supposons que la fonction définie sur T^2 par :

$(X,Y) \longrightarrow e = X + Y - (X \oplus Y)$

soit une application de T^2 dans $T \cup]-\xi,\xi[$.

Définissons les suites $\{S^n\}$, $\{e_i^n\}$ $(1 \leqslant i \leqslant N-1)$, $\{E_i^n\}$ $(1 \leqslant i \leqslant N-1)$ et $\{C^n\}$ de la façon suivante :

Les premiers termes de chaque suite sont définis par :

$$S^1 = \bigoplus \sum_{i=1}^{N} X_i$$

$$e_i^1 = \left(\bigoplus \sum_{j=1}^{i} X_j \right) + X_{i+1} - \left(\bigoplus \sum_{j=1}^{i+1} X_j \right) \qquad 1 \leqslant i \leqslant N-1$$

e_i^1 est "l'erreur" commise lors de la ième addition dans le calcul de S^1.

La suite $\{E_i^n\}$ se déduit de $\{e_i^n\}$ par :

$$E_i^n = \begin{cases} e_i^n & \text{si} \quad e_i^n \in T \\ 0 & \text{sinon} \end{cases}$$

Enfin :

$$C^1 = \bigoplus \sum_{i=1}^{N-1} E_i^1 \quad .$$

Puis :

$$S^n = S^{n-1} \oplus C^{n-1}$$

Soient e_i^n $(1 \leqslant i \leqslant N-2)$ définies par :

$$e_i^n = \left(\bigoplus \sum_{j=1}^{i} E_j^{n-1} \right) + E_{i+1}^{n-1} - \left(\bigoplus \sum_{j=1}^{i+1} E_j^{n-1} \right)$$

et soit

$$e_{N-1}^n = S^{n-1} + C^{n-1} - S^n$$

(e_i^n est "l'erreur" commise lors de la ième addition dans le calcul de C^{n-1}).

Enfin :

$$C^n = \bigoplus \sum_{i=1}^{N-1} E_i^n$$

Des définitions, il vient :

$$S = S^1 + \sum_{i=1}^{N-1} e_i^1$$

$$S^{n-1} + \sum_{i=1}^{N-1} E_i^{n-1} = S^n + \sum_{i=1}^{N-1} e_i^n \qquad \text{pour} \quad n > 1.$$

Donc :

$$S = S^n + \sum_{i=1}^{N-1} e_i^n + \sum_{j=1}^{n-1} \sum_{i=1}^{N-1} (e_i^j - E_i^j)$$

ou plus simplement :

si $I_n = \{i \ ; \ e_i^n \neq o\}$

et

$$\Lambda_n = \{(i,j) \; ; \; j < n \wedge o < |e_i^j| < \xi\}$$

vérifiant : Card I_n + Card $\Lambda_n \leqslant N-1$

on a :

$$S = S^n + \sum_{i \in I_n} e_i^n + \sum_{(i,j) \in \Lambda_n} e_i^j \qquad \text{(III.1)}$$

I . 2 . CONVERGENCE DE L'ALGORITHME

Avant d'aborder l'étude du comportement de l'algorithme $\mathcal{A}$, examinons le point suivant :
Lorsque n augmente, les erreurs e_i^n vont tendre à diminuer en valeur absolue. Sera-t-il alors possible de conserver, dans les éléments E_i^n toute l'information sur les termes d'erreur, ou au contraire, l'existence d'inégalités $o < |e_i^n| < \xi$ entraînera-t-elle, par la mise à o de E_i^n, une perte de précision ?
Il vient :

Si l'opération $\oplus$ est correcte et si les N éléments X_i de T vérifient l'inégalité suivante :

$$e(X_i) \geqslant m_2 + s-1 \qquad 1 \leqslant i \leqslant N ,$$

alors, les erreurs e_i^n sont, quel que soit n, éléments de T.

On remarquera pour ce résultat, les deux propriétés suivantes :

1) Si X et Y sont deux éléments de T tels que $e(X) \geqslant e(Y) \wedge X + Y \in R_T$

$$E = X + Y - (X \oplus Y) = \lambda \, b^{e(Y)-s} \quad \text{avec} \quad \lambda \in \mathbb{Z}$$

entraîne $E \in T$ à la condition suffisante $e(Y) \geqslant m_2 + s-1$.

(Remarque du § I.1 chapitre II)

2) La somme $S = X \oplus Y$ de deux éléments de T tels que $e(X) \geqslant e(Y)$ vérifie l'un des deux points a ou b suivants :

a) $e(\tilde{S}) \geqslant e(Y)$

b) Si $e(\tilde{S}) < e(Y)$ et si t est l'entier positif défini par $t = e(Y) - e(\tilde{S})$, alors les t derniers chiffres de la mantisse de $\tilde{S}$ sont nuls. L'erreur entachant la somme de $\tilde{S}$ et d'un élément de T d'exposant supérieur ou égal à $e(\tilde{S}) + t$ satisfait encore à $E = \lambda\, b^{e(Y)-s}, \lambda \in \mathbb{Z}$.

- La plus petite erreur e_i^n, non nulle, possible est $\min_i \{e(X_i)-s\}$ _

Il s'en suit que, dans le cas où les éléments X_i sont d'exposants trop proches de m_2, il conviendra de les multiplier par un facteur de cadrage, afin de ne perdre aucune précision.
Nous ne considérerons donc pas dans la suite l'éventualité d'underflows, mais les modifications de démonstrations dues à ce phénomène sont dans [32] ; $E_i^n = e_i^n$, $\Lambda_n = \emptyset$.

I. 2. 1. Deux théorèmes relatifs à une opération d'addition correcte

On peut énoncer :

Proposition III . 1 .

Si l'opération $\oplus$ est :
correcte
I - dirigée pour tout couple $(X,Y) \in T^2_{\Lambda}$ $X\,Y > 0$,

Alors, si $2\,N \leqslant b^{s-1}$, il existe un entier $k > 0$ tel que :

$$|S - S^k| < 2\, b^{P^k - s} \qquad \text{(III.2)}$$

où $P^k = e(S^k)$.

Démonstration :

La démonstration ci-dessous suppose exclue toute possibilité d'overflow.

Examinons le passage du pas $n-1$ au pas n.

Soit Q^n le plus petit entier tel que :

$$\sum_{i=1}^{N-1} |E_i^{n-1}| < b^{Q^n} \qquad \text{(III.3)}$$

et montrons que nous avons deux possibilités :

* soit $Q^{n+1} < Q^n$
* soit l'algorithme est terminé.

Remarquons que, si $Z_i \in T$, $(1 \leqslant i \leqslant P)$, l'opération $\oplus$ étant I - dirigée pour tout couple $(X,Y) \in T^2_{\wedge}$ $X\,Y > 0$, et de plus correcte, on a :

$$\left| \bigoplus_{i=1}^{l} Z_i \right| \leqslant \sum_{i=1}^{P} |Z_i| \qquad \text{(III.4)} \quad \text{pour } \forall\, l \leqslant P$$

(en effet, $\oplus$ étant correcte, pour tout couple $(X,Y) \in T^2_{\wedge}$ $X\,Y < 0$ on a : $|X \oplus Y| \leqslant \text{Max}\,(|X|\ ,\ |Y|) \leqslant |X| + |Y|$).

$\oplus$ étant correcte et aucun overflow n'étant possible, (III.3) et (III.4) entraînent :

$$|e_i^n| < b^{Q^n - s} \qquad 1 \leqslant i \leqslant N-2$$

Notant $e(S^n) = P^n$:

$$\sum_{i \in I_n} |e_i^n| < (N-2)\, b^{Q^n - s} + b^{P^n - s} < \frac{1}{2} b^{Q^n - 1} + b^{P^n - s}$$

* Si $\frac{1}{2} b^{Q^n - 1} \leqslant b^{P^n - s}$, on a $|S - S^n| < 2\, b^{P^n - s}$, l'algorithme est terminé et seul le dernier bit de la mantisse de S^n peut être entaché d'erreur.

* Sinon, il vient $Q^{n+1} < Q^n$,et si l'inégalité (III.2) n'est pas vérifiée au pas n , on forme le pas suivant, en remarquant que la suite des Q^n est finie.

D'où la proposition.

(Si $S^k = 0$, on fera $P^k = -\infty$ dans l'inégalité (III.2) qui sera prise au sens large).

Notation. Lorsque (III.2) est vérifiée, l'algorithme $\mathcal{A}$ sera dit convergent.

Une autre proposition est donnée par :

Proposition III.2.

Si l'opération $\oplus$ est :

correcte

isotone

une condition suffisante de convergence de l'algorithme $\mathcal{A}$ est que soit réalisée $2N \leqslant b^{s-2}$.

Démonstration :

En effet, (III.4) n'est plus valable, mais sera remplacée par l'inégalité

$$|\oplus \sum_{i=1}^{l} Z_i| \leqslant \oplus \sum_{i=1}^{l} |Z_i| \qquad \text{(III.4')}$$

lorsque la symétrie est un automorphisme de T ,

ou par :

$$|\oplus \sum_{i=1}^{l} Z_i| \leqslant \text{Max} \{\oplus \sum_{i=1}^{l} |Z_i| \ , \ |\ominus \sum_{i=1}^{l} |Z_i|\ |\} \qquad \text{(III.4")}$$

sinon.

On a aussi la double inégalité moins large :

$$\oplus \sum_{i \in J} Z_i \leqslant \oplus \sum_{i=1}^{l} Z_i \leqslant \oplus \sum_{i \in I} Z_i$$

où $I = \{i \ ; \ Z_i > 0\}$, $J = \{i \ ; \ Z_i < 0\}$

valable dès que $\oplus$ est isotone.

On utilisera ensuite le lemme suivant :

Lemme III.1.

Si l'opération $\oplus$ est correcte, et si $2l \leqslant b^{s-1}$, la somme en arithmétique à virgule flottante de l nombres de même signe, $Z_i \in T$ $(1 \leqslant i \leqslant l)$, vérifie :

$$e(\oplus \sum_{i=1}^{l} Z_i) \leqslant e(\sum_{i=1}^{l} Z_i) + 1 \qquad \text{(III.5)}$$

La démonstration du lemme est immédiate :

Soient $S = \sum_{i=1}^{l} Z_i$, $\tilde{S} = \oplus \sum_{i=1}^{l} Z_i$, $p_i = e(\oplus \sum_{j=1}^{i+1} Z_j)$.

Il vient :

$|S - \tilde{S}| \leqslant \sum_{i=1}^{l-1} b^{p_i - s}$ et $b^{p_i - s} \leqslant b^{1-s} \, | \oplus \sum_{j=1}^{i+1} Z_j |$ entraîne :

$$|S - \tilde{S}| \leqslant b^{1-s} \, l \, |\tilde{S}| \qquad \text{(III.6)}$$

d'où le résultat , que l'on déduit aussi des formules (25.5) à (25.10) de [42].

I . 2 . 2 . Un théorème de convergence pour une addition définie par une troncature sans chiffre de garde.

Afin de ne pas compliquer inutilement l'écriture, nous démontrerons la proposition ci-dessous avec l'hypothèse :

$$e(X_i) \geqslant m_2 + s - 1 \qquad 1 \leqslant i \leqslant N$$

pour encore exclure toute possibilité d'underflow.

Si l'hypothèse précédente n'est pas vérifiée, il convient de modifier (III.1) en remarquant que si, par exemple, les nombres flottants sont représentés par leur signe et leur valeur absolue et les erreurs d'arrondi élémentaires calculées par la formule (II.5), la suite $\{E_i^n\}$ se déduit de $\{e_i^n\}$ par :

$$E_i^n = \begin{cases} e_i^n & \text{si } e_i^n \in T - \{0\} \\ 0 \text{ ou } \pm \xi & \text{sinon} \end{cases}$$

Proposition III.3.

Si la règle définissant l'addition est une troncature sans chiffre de garde, une condition suffisante de convergence de l'algorithme $\mathcal{A}$ est que soit réalisée $2N \leqslant b^{s-1}$.

Modifions, avec les hypothèses présentes, la démonstration de la proposition III.1.

Soit encore $Q^n = e(\sum_{i=1}^{N-1} |E_i^{n-1}|)$, et examinons le passage du pas $n-1$ au pas n .

L'inégalité (III.4) reste valable ; en effet :

$\oplus$ est I - dirigée pour tout couple $(X,Y) \in T^2_\wedge$ $X\,Y > 0$

et $|X \oplus Y| \leqslant \max(|X|, |Y|)$ $\forall (X,Y) \in T^2_\wedge$ $X\,Y < 0$.

Le module de l'erreur e_i^n est majoré à l'aide de l'inégalité (II.4) :

$$|e_i^n| < b^{p_i^n - s}$$

où

$$p_i^n = \max \{ e(\oplus \sum_{j=1}^{i+1} E_j^{n-1}) , e(\oplus \sum_{j=1}^{i} E_j^{n-1}) , e(E_{i+1}^{n-1}) \}$$

En utilisant (III.4) , il vient immédiatement :

$$p_i^n \leqslant Q^n$$

et

$$|e_i^n| < b^{Q^n - s} \qquad 1 \leqslant i \leqslant N-2$$

Posant $\pi^n = \max \{ e(S^n) , e(S^{n-1}) , e(C^{n-1}) \}$,

$$|e_{N-1}^n| < b^{\pi^n - s}$$

Distinguons 3 cas :

* Si $\pi^n = e(S^n)$, la démonstration de la proposition s'effectue comme celle de la proposition III.1.

* Si $\pi^n = e(C^{n-1})$, il vient immédiatement $\pi^n \leqslant Q^n$ et la proposition se déduit par : $\sum_{i=1}^{N-1} |e_i^n| < N\, b^{Q^n - s} \Longrightarrow Q^{n+1} < Q^n$.

* Si $\pi^n = e(S^{n-1})$, l'opération $S^n = S^{n-1} \oplus C^{n-1}$ est une soustraction arithmétique ; rappelons que $e(S^{n-1}) - e(C^{n-1}) \geqslant 2$ entraîne $e(S^n) \geqslant e(S^{n-1}) - 1$.

Posons : $\pi^n = Q^n + L$

Si $L \leqslant s-2$, on a :

$\sum_{i=1}^{N-1} |e_i^n| \leqslant (N-2)\, b^{Q^n - s} + b^{Q^n + L - s} < (\frac{1}{2} + \frac{1}{b})\, b^{Q^n - 1}$ donc $Q^{n+1} < Q^n$

Si $L = s - 1$, l'algorithme est terminé, soit au pas n , soit au pas $n+1$.

I . 3 . REMARQUES

Remarque 1. Précision apportée par la première correction :

$$\text{Si} \quad P = \max_{2 \leqslant i \leqslant N} \{ e (\bigoplus \sum_{j=1}^{i} X_j) \}$$

on a :

$$|S - S^2| < (N - 2)\, b^{P+\gamma-2s} + b^{P^2-s}$$

où γ est le plus petit entier tel que $N \leqslant b^{\gamma}$.

Remarque 2. Correction d'un produit scalaire :

Si $X = \{X_i\}$ et $Y = \{Y_i\}$ sont deux vecteurs de T^N , lorsque l'on désigne par e_i les erreurs de multiplication et d'addition suivantes :

$$e_i = X_i\, Y_i - (X_i \otimes Y_i) \qquad i = 1, \ldots , N$$

$$e_{N+i} = \{(\bigoplus \sum_{j=1}^{i} X_j \otimes Y_j) + X_{i+1} \otimes Y_{i+1}\} - \bigoplus \sum_{j=1}^{i+1} X_j \otimes Y_j \qquad i = 1, \ldots , N$$

il vient immédiatement :

$$X\, Y = X \otimes Y + \sum_{i=1}^{2N-1} e_i$$

La correction d'un produit scalaire se ramène, dès que les e_i sont calculés, à la correction d'une somme.

I . 4 . TEST D'ARRET DE L'ALGORITHME $\mathcal{A}$

Il est particulièrement simple lorsque l'opération $\oplus$ est G - dirigée (resp. D - dirigée), puisque l'égalité

$$S^n = S^{n-1}$$

permet d'affirmer la convergence .

S'il n'en est pas ainsi, on adjoindra au calcul de $C^{n-1} = \oplus \sum_{i=1}^{N-1} E_i^{n-1}$,

le calcul de :

$$TEST = \oplus \sum_{i=1}^{N-1} |E_i^{n-1}|$$

et par le lemme III.1 , la convergence sera assurée par la condition suffisante :

$$\frac{3}{2} TEST < \theta \cdot b^{P^n - s} \qquad (III.7)$$

où $\theta = 1$ ou 2 suivant la précision voulue.

On peut aussi adjoindre au calcul de C^{n-1}, le calcul suivant :

Posant $F_i = \oplus \sum_{j=1}^{i+1} E_j^{n-1}$

on calcule

$$TEST = \oplus \sum_{i=1}^{N-2} |F_i|$$

et on utilise la majoration suivante obtenue par (III.6)

$$|S - S^n| < b^{1-s} (1 + b^{1-s}(N-2)) TEST + |E_{N-1}^n|$$

I . 5 ETUDE NUMERIQUE

Nous avons considéré 2 sommes arithmétiques :

$$\sigma_1 = \sum_{i=1}^{N} \frac{1}{i} \quad \text{et} \quad \sigma_2 = \sum_{i=1}^{N} \frac{1}{i^2}$$

et 2 sommes algébriques :

$$\sigma_3 = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i} \quad \text{et} \quad \sigma_4 = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i^2}$$

Nous avons utilisé l'arithmétique à virgule flottante définie sur une IBM 360-30 : $b = 16$, $s = 6$; l'opération $\oplus$ est correcte, I - dirigée pour tout couple $(X,Y) \in T^2_\wedge$ $X\ Y > 0$.

Il vient immédiatement, par la remarque 1 du § I.3, que pour $n < b^3$, une seule correction sera nécessaire pour rendre significatifs tous les chiffres du résultat.

Dans chacun des 4 cas envisagés, soient :
S^1 la valeur de la somme calculée de façon habituelle, en arithmétique à virgule flottante
S^2 la valeur de cette somme corrigée par une seule correction
S D la valeur de la somme effectuée en double précision.

Les calculs ont été faits pour N variant de 100 à 500, par pas de 100. La lecture des résultats montre que tous les chiffres de la somme corrigée S^2 sont significatifs, une éventuelle différence entre les derniers chiffres de S^2 et S D provenant de la conversion des résultats en décimal, comme nous l'avons vérifié en effectuant la sortie S^1, S^2, S D en hexadécimal.

	N	100	200	300	400	500
σ_1	S^1	5,187340	5,877946	6,282538	6,569756	6,792601
	S^2	5,187377	5,878030	6,282662	6,569928	6,792822
	S D	5,187377	5,878030	6,282663	6,569929	6,792822
σ_2	S^1	1,634939	1,639858	1,641470	1,642253	1,642706
	S^2	1,634983	1,639946	1,641605	1,642437	1,642936
	S D	1,634983	1,639946	1,641606	1,642437	1,642936
σ_3	S^1	0,6881702	0,6906486	0,6914758	0,6918888	0,6921354
	S^2	0,6881722	0,6906534	0,6914833	0,6918988	0,6921481
	S D	0,6881722	0,6906534	0,6914833	0,6918987	0,6921482
σ_4	S^1	0,8224151	0,8224491	0,8224530	0,8224527	0,8224512
	S^2	0,8224175	0,8224546	0,8224615	0,8224639	0,8224650
	S D	0,8224175	0,8224545	0,8224614	0,8224639	0,8224650

Enfin, signalons que les temps nécessités pour le calcul des valeurs S^2 et S D, relatives à la somme théorique σ_1 lorsque N = 2000, ont été identiques, le test de comparaison $|X| \geq |Y|$ relatif au calcul de l'erreur élémentaire d'addition par (II.1) étant omis (le test est inutile ici puisque l'on connaît les grandeurs respectives des opérandes).

I . 6 . CONCLUSION

Rappelons que le problème de l'amélioration de la précision d'une somme algébrique, calculée par une arithmétique à mot fini, apparaît dans deux circonstances, éventuellement simultanées :

1) Le nombre de termes de la somme $S = \sum_{i=1}^{N} X_i$ est grand.

Dans de nombreux cas, les X_i sont approximativement du même ordre de grandeur (intégration numérique, résolution d'équations différentielles avec un grand nombre de pas, calcul de somme de séries) mais, N étant grand, les arrondis ou troncatures successives deviennent excessives.

2) Les exposants des termes de la somme sont d'ordres de grandeur très différents, et peuvent ainsi introduire une erreur substantielle, même si N n'est pas très grand : c'est le phénomène de "cancellation", apparaissant en particulier quand une somme intermédiaire est beaucoup plus grande en valeur absolue que la somme finale. De nombreux chiffres significatifs sont alors perdus, et il est nécessaire de corriger les résultats obtenus.

L'obtention de la quasi double-précision d'une somme, à l'aide d'une arithmétique à simple précision, avait déjà été envisagée dès 1965 par W.KAHAN [14] et O.MOLLER [28] qui établissent un schéma de correction au 1er degré (c'est-à-dire en négligeant les erreurs du second ordre), d'une somme $S = \sum_{i=1}^{N} X_i$, différant de celui ci-dessus : au pas i , on ajoute préalablement à X_{i+1} la valeur ou la partie principale de l'erreur commise au pas précédent. O.MOLLER a établi, à cette occasion, une formule exacte d'obtention de l'erreur élémentaire d'addition.

Une autre approche du problème de l'amélioration de la précision d'une somme, a aussi été réalisée par JM.WOLFE [44] qui propose l'utilisation d'accumulateurs en cascade, chaque accumulateur servant à stocker des sommes partielles dont les exposants sont compris entre des limites définies.
M.A.MALCOLM [24] établit une modification de l'algorithme de WOLFE, permettant d'assurer une précision fixée, le coût de la méthode étant augmenté par un découpage préalable, en un nombre q fixé de parties, de chaque terme de la somme.
Enfin D.R.ROSS [36] avait modifié l'algorithme de WOLFE pour permettre son application, non plus seulement à une somme arithmétique, mais à une somme algébrique dont le nombre de chiffres significatifs est fixé à l'avance.

II - OBTENTION DU PROBLEME PERTURBE DANS L'ANALYSE A POSTERIORI DES ERREURS D'ARRONDI

II.1. INTRODUCTION

Soit un problème $(\mathcal{P})$, défini par des paramètres $a_1 , a_2 , \ldots , a_n$, dont la solution x s'exprime comme le résultat d'une succession d'opérations élémentaires sur $\mathbb{R}$.

Notons $x = f(a_1 , a_2 , \ldots , a_n)$. $\qquad (\mathcal{P})$

Lorsque la résolution est effectuée à l'aide d'une arithmétique flottante, le résultat obtenu $\bar{x}$ diffère de x en raison des erreurs d'arrondi ; dans de nombreux cas, il est commode, réalisant une analyse à postériori des erreurs d'arrondi, d'exprimer que $\bar{x}$ est la solution exacte (c'est-à-
calculée sur $\mathbb{R}$) d'un problème $(\mathcal{P}')$ obtenu par perturbation du
me initial, soit

$$\bar{x} = f(a_1 + e_1 , a_2 + e_2 , \ldots , a_n + e_n) \qquad (\mathcal{P}')$$

lérons alors le problème suivant :

Obtenir, par des calculs en arithmétique à virgule flottante,
iture du problème perturbé $(\mathcal{P}')$.

Nous soulignerons plusieurs points :

1) Pour de nombreux algorithmes, dans ce type d'analyse, les erreurs interviennent <u>linéairement</u> et l'on est ramené à des <u>problèmes de sommation d'erreurs.</u>

2) Rappelons que l'écriture du problème perturbé n'est pas, en général, unique, mais que l'intérêt de l'analyse à postériori des erreurs d'arrondi, est certain pour les problèmes expérimentaux dont les paramètres sont entachés d'erreurs de mesure ;
Le problème des erreurs d'entrée des données, peut être aussi considéré de façon analogue.

Nous traiterons, à titre d'exemple, l'algorithme de Gauss.

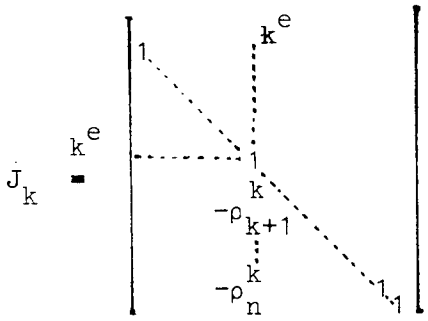
II.2. UN EXEMPLE : L'ALGORITHME DE GAUSS.

II.2.1. Notations

$A \in \mathcal{M}_{n,n}(\mathbb{R})$, non singulière, et $b \in \mathbb{R}^n$ étant donnés, considérons la résolution, par l'algorithme de Gauss général, du système linéaire :

$$A\,x = b \qquad \text{(III.8)}$$

A et b sont multipliés à gauche successivement par les matrices J_k $(k = 1 , \ldots , n - 1)$:



où $\rho_i^k = \dfrac{a_{ik}^k}{a_{kk}^k}$ $(i = k+1, \ldots , n)$

pour former les deux suites $\{A^k\}$ et $\{b^k\}$ $(k = 1 , \ldots , n)$:

$$A^1 = A \quad , \quad b^1 = b$$

$$A^{k+1} = J_k\, A^k \quad , \quad b^{k+1} = J_k\, b^k \qquad (k = 1 , \ldots , n-1)$$

(III.8) est remplacé par le système triangulaire équivalent $Tx = d$, que l'on résoud, et où nous avons posé pour la commodité des notations :

$$T = A^n = J_{n-1} \ldots J_1\, A$$

$$d = b^n = J_{n-1} \ldots J_1\, b$$

II.2.2. Rappel des résultats de l'analyse à postériori des erreurs d'arrondi

Dans ce paragraphe, nous rappelons des résultats formulés et démontrés par N. GASTINEL [8] (Voir aussi J. H. WILKINSON [42]) :

Par une arithmétique à virgule flottante, il est obtenu une suite de matrices $\{\bar{A}^k\}$ et une suite de seconds membres $\{\bar{b}^k\}$ $(k = 1, \ldots , n$, avec $\bar{A}^1 = A$, $\bar{b}^1 = b)$:

Définissant les erreurs absolues e_i^k , e'^k_{ij} , e''^k_{ij} suivantes, commises au cours de la triangularisation pour :

$k = 1 , \ldots , n-1$

$i = k+1 , \ldots , n$

$j = k+1 , \ldots , n+1$ si l'on pose $\bar{a}^k_{i,n+1} = \bar{b}^k_i$:

$$e_i^k = \bar{a}^k_{ik} \,/\, \bar{a}^k_{kk} - \bar{\rho}^k_i \qquad \text{avec} \qquad \bar{\rho}^k_i = \bar{a}^k_{ik} \oslash \bar{a}^k_{kk}$$

$$e'^k_{ij} = \bar{\rho}^k_i \, \bar{a}^k_{kj} - \bar{\rho}^k_i \otimes \bar{a}^k_{kj}$$

$$e''^k_{ij} = \bar{a}^k_{ij} - \bar{\rho}^k_i \otimes \bar{a}^k_{kj} - \bar{a}^{k+1}_{ij}$$

De plus, définissant les erreurs absolues ε'_{ij} , ε''_{ij} , ε_i suivantes, commises lors de la résolution du système triangulaire $\bar{T}\, x = \bar{d}$ pour :

$i = n , \ldots , 1$

$j = n , \ldots , i+1$:

$$\varepsilon'_{ij} = \bar{t}_{ij} \, \bar{x}_j - \bar{t}_{ij} \otimes \bar{x}_j$$

et si $S_{i,n+1} = 0$, $S_{ik} = \bigoplus \sum_{\ell=n}^{k} \bar{t}_{i\ell} \otimes \bar{x}_\ell$

$$\varepsilon''_{ij} = S_{i,j+1} + \bar{t}_{ij} \otimes \bar{x}_j - S_{i,j} \qquad j \neq n$$

$$\varepsilon''_{in} = \bar{d}_i - S_{i,i+1} - (\bar{d}_i \ominus S_{i,i+1})$$

$$\varepsilon_i = (\bar{d}_i \ominus S_{i,i+1}) \,/\, \bar{t}_{ii} - \bar{x}_i$$

et notant :

$\bar{J}_k$ la matrice déduite de J_k en remplaçant dans cette matrice les valeurs ρ^k_i par les valeurs calculées $\bar{\rho}^k_i$

et

$\bar{J} = \bar{J}_{n-1} \ldots \bar{J}_1$

on a :

Théorème (N. GASTINEL [8] , ch V, théorème VII)

La solution approchée du système $A\,x = b$ et obtenue par un calcul en arithmétique à virgule flottante, peut être considérée comme la solution exacte du système :

$$(A + \delta A)\, x = b + \delta b + \bar{\bar{J}}^{-1}\, \delta \bar{d} \qquad \text{(III.9)}$$

où

$$\delta A = \sum_{k=1}^{n-1} M^k \quad , \quad \delta b = \sum_{k=1}^{n-1} z^k \quad ,$$

M^k et z^k étant respectivement la matrice et le vecteur définis par :

$$M^k \begin{cases} m^k_{ik} = - e^k_i\, \bar{a}^k_{kk} & , \quad i > k \\ m^k_{ij} = e'^k_{ij} - e''^k_{ij} & , \quad i,j = k+1, \ldots, n \\ m^k_{ij} = 0 & \text{ailleurs} \end{cases}$$

$$z^k \begin{cases} (z^k)_i = 0 & , \quad i \leqslant k \\ (z^k)_i = e'^k_{i,n+1} - e''^k_{i,n+1} & , \quad i > k \end{cases}$$

et $\delta \bar{d}$ est le vecteur dont la $i^{\text{ème}}$ composante s'écrit :

$$(\delta \bar{d})_i = - \varepsilon_i\, \bar{t}_{ii} + \Big(\sum_{j=i+1}^{n} \varepsilon'_{ij} \Big) + \Big(\sum_{j=i+1}^{n-1} \varepsilon''_{ij} \Big) - \varepsilon''_{in}$$

II.2.3. Etude expérimentale

Nous avons écrit en FORTRAN le programme suivant, calculant au 1er degré (c'est-à-dire en négligeant les erreurs du second ordre), la matrice δA et les vecteurs δb et $\delta \bar{d}$ du problème perturbé (III.9) ; les paramètres D A , D B , J M 1 , D D de la subroutine P G A U S S représentent respectivement δA , δB , $\bar{\bar{J}}^{-1}$ et $\delta \bar{d}$.
L'écriture E S(X,Y,S) , E P(X,Y,P) , C D(X,Y,D) désigne les sous-programmes FUNCTION auxiliaires, de calcul des erreurs élémentaires, définis au chapitre II.

```
      SUBROUTINE PGAUSS(A,B,N,X,IER,DA,DB,JM1,DD)
      REAL JM1
      DIMENSION A(N,N),B(N),X(N),DA(N,N),DB(N),JM1(N,N),DD(N)
      DO 1 I=1,N
      DB(I)=.0
      JM1(I,I)=1.0
      DO 1 J=1,N
      DA(I,J)=.0
    1 CONTINUE
      IER=0
C     TRIANGULARISATION DE A
      NM1=N-1
      DO 2 K=1,NM1
      IF(A(K,K))10,20,10
   20 IER=-1
      GOTO 3
   10 KP1=K+1
      DO 2 I=KP1,N
      R=A(I,K)/A(K,K)
      DA(I,K)=DA(I,K)-CD(A(I,K),A(K,K),R)
      JM1(I,K)=R
      V=R*B(K)
      W=B(I)-V
      Z=-V
      DB(I)=DB(I)+(EP(R,B(K),V)-ES(B(I),Z,W))
      B(I)=W
      DO 2 J=KP1,N
      V=R*A(K,J)
      W=A(I,J)-V
      Z=-V
      DA(I,J)=DA(I,J)+(EP(R,A(K,J),V)-ES(A(I,J),Z,W))
      A(I,J)=W
    2 CONTINUE
C     RETOUR ARRIERE
      X(N)=B(N)/A(N,N)
      DD(N)=-CD(B(N),A(N,N),X(N))
      DO 3 L=2,N
      I=N+1-L
      IP1=I+1
      S=.0
      Z=.0
      DO 4 J=IP1,N
      V=A(I,J)*X(J)
      W=S+V
      Z=Z+EP(A(I,J),X(J),V)+ES(S,V,W)
      S=W
    4 CONTINUE
      S=-S
      V=B(I)+S
      X(I)=V/A(I,I)
      DD(I)=Z-ES(B(I),S,V)-CD(V,A(I,I),X(I))
    3 CONTINUE
      RETURN
      END
```

Nous donnerons à titre d'exemple, le résultat suivant, obtenu par l'activation du sous-programme PGAUSS, pour la résolution de $Ax = b$ avec :

$$A = \begin{vmatrix} 3 & 2 & -1 & 4 & 1 & 1 & 3 & 2 \\ 1 & -3 & 2 & 1 & 0 & -2 & 1 & 1 \\ 0,5 & 1 & 3 & 0,25 & -4 & -1 & 2 & 0 \\ 3 & 2 & -3 & 4 & 1 & 2 & -1 & 1 \\ 2 & -1 & 0,125 & 0,25 & -3 & 2 & 1,25 & -2 \\ -1 & 1 & 0 & 0,5 & 2 & 3 & 4 & 1 \\ -2 & -2 & 1 & 0 & 1 & 3 & 2 & 3 \\ 1 & -1 & 1 & 6 & -3 & 2 & 4 & 3 \end{vmatrix} \qquad b = \begin{vmatrix} 19 \\ 18 \\ 16,25 \\ -2 \\ 7 \\ 8,5 \\ 1 \\ 31 \end{vmatrix}$$

$$\delta A = 10^{-8} \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -6 & 60 & -30 & 24 & 0 & 30 & 0 & 0 \\ -12 & -17 & -95 & 0 & 65 & 96 & -13 & 5 \\ 0 & 0 & 19 & 1 & 25 & 4 & 30 & 30 \\ -12 & 36 & 40 & 56 & -165 & -184 & -1324 & 16 \\ 6 & -74 & -66 & -65 & -101 & -78 & 1860 & -126 \\ 12 & -4 & 3 & -80 & 1561 & -447 & 1470 & -121 \\ -6 & 74 & 13 & -212 & 2379 & -402 & -671 & 704 \end{vmatrix}$$

$$\delta b = 10^{-8} \begin{vmatrix} 0 \\ 89 \\ 42 \\ -75 \\ -1464 \\ -1462 \\ 3016 \\ -8165 \end{vmatrix} \qquad \bar{J}^{-1}\, \delta\bar{d} = 10^{-8} \begin{vmatrix} 489 \\ 322 \\ 47 \\ 447 \\ 1294 \\ 713 \\ 958 \\ 36 \end{vmatrix}$$

Solution théorique x et solution calculée $\bar{x}$:

$$x = \begin{vmatrix} 1 \\ -1 \\ 2 \\ 3 \\ 0 \\ -2 \\ 4 \\ -1 \end{vmatrix} \qquad \bar{x} = \begin{vmatrix} 1,000028 \\ -1 \\ 1,999955 \\ 2,999960 \\ -0,0000275 \\ -2,000015 \\ 4,000010 \\ -0,9999652 \end{vmatrix}$$

les résidus $r_1 = b - A\bar{x}$ et $r_2 = b + \delta b + \bar{J}^{-1}\,\delta\bar{d} - (A + \delta A)\,\bar{x}$ ont été évalués en double précision :

$$r_1 = 10^{-7} \begin{vmatrix} -49 \\ -52 \\ -52 \\ -25 \\ -457 \\ 822 \\ 270 \\ 485 \end{vmatrix} \qquad r_2 = 10^{-14} \begin{vmatrix} 0 \\ -25 \\ -340 \\ -0,3 \\ -21 \\ 110 \\ -2000 \\ -2800 \end{vmatrix}$$

III - UTILISATION DE METHODES DE RAFFINEMENT ITERATIF

III . 1 . INTRODUCTION

Les méthodes de raffinement itératif, de la solution d'un système linéaire ou de l'inverse calculée d'une matrice, ne sont guère utilisables lorsque le calcul numérique du résidu est effectué avec la même précision que celle utilisée pour la résolution du problème initial.

J.H. WILKINSON [42] , [43] a étudié ces méthodes et donné les conditions de convergence du processus itératif.
Cet auteur souligne l'importance de l'accumulation sur 2 s chiffres des produits scalaires dans le calcul du résidu, afin d'éviter que les erreurs entachant ces résidus n'atteignent l'ordre de grandeur de ceux-ci. Les techniques du calcul des erreurs d'arrondi élémentaires permettent, en arithmétique flottante de simple ou de double précision, le calcul presqu'exact des résidus.

Nous avons considéré la méthode la plus simple de raffinement itératif de l'inverse d'une matrice.

III . 2 . UN EXEMPLE : CORRECTION DE L'INVERSE D'UNE MATRICE

III.2.1. Formule théorique

Soit A une matrice inversible, et B_0 une valeur approchée de l'inverse A^{-1} de A . Posant :

$R = I - AB_0$

il vient :

$A^{-1} = B_0 (I - R)^{-1}$

La suite $\{ B_k = B_0 \sum_{j=0}^{k} R^j \}$ que l'on calcule par la récurrence :

$$B_k = B_{k-1} + B_0 R^k \qquad \text{(III. 10)}$$

vérifie :

$$B_k = A^{-1} (I - R^{k+1})$$

et converge vers A^{-1} quand $k \to \infty$, lorsque le rayon spectral de R est inférieur à 1 .

L'erreur au pas k , $A^{-1} - B_k = E_k$ est alors :

$$E_k = A^{-1} R^{k+1} \qquad \text{(III. 11)}$$

III. 2. 2. Calculs en arithmétique à virgule flottante

R définie ci-dessus, est calculée, donc entachée d'erreur, c'est-à-dire que si $\bar{R}$ est la valeur "calculée" :

$$R = \bar{R} + \Omega$$

Notons $B \oplus C$ et $B \otimes C$ la somme et le produit en arithmétique flottante, de 2 matrices B et C définies sur T .

Posant $\bar{B}_0 = B_0$

$$\bar{C}_0 = B_0 \ , \ \bar{C}_k = \bar{C}_{k-1} \otimes \bar{R} \qquad k = 1, \ldots$$

on obtient par l'arithmétique à virgule flottante, la suite de matrices $\{ \bar{B}_k \}$:

$$\bar{B}_k = \bar{B}_{k-1} \oplus \bar{C}_k$$

et soient les suites $\{M_k\}$ et $\{M'_k\}$ de matrices d'erreur suivantes :

$$M_k = \bar{C}_{k-1} \ \bar{R} - \bar{C}_k \qquad k = 1, 2, \ldots$$

$$M'_k = (\bar{B}_{k-1} + \bar{C}_k) - \bar{B}_k$$

Par la sommation des égalités précédentes :

$$\bar{C}_k = B_0 \ \bar{R}^k - \sum_{j=1}^{k} M_j \ \bar{R}^{k-j}$$

$$\bar{B}_k = B_0 + \sum_{j=1}^{k} \bar{C}_j - \sum_{j=1}^{k} M'_j$$

et l'erreur dûe aux calculs $E'_k = B_k - \bar{B}_k$ est :

$$E'_k = B_0 \sum_{j=1}^{k} (R^j - \bar{R}^j) + \sum_{j=1}^{k} \sum_{i=1}^{j} M_i \ \bar{R}^{j-i} + \sum_{j=1}^{k} M'_j \qquad \text{(III. 12)}$$

L'erreur totale commise sera :

$$A^{-1} - \bar{B}_k = E_k + E'_k$$

En utilisant par exemple la norme de matrice suivante :

$$|| A ||_\infty = \max_i \left(\sum_{j=1}^{N} | a_{ij} | \right) \qquad A \in \mathcal{M}_{N,N}(\mathbb{R})$$

il est possible de donner une majoration de l'erreur dûe aux calculs :

Nous rappelerons au préalable les résultats suivants de l'analyse d'erreur (J.H. WILKINSON [42] pp. 18-19 et 83-84) :
$X = (x_1,\ldots,x_N)$ et $Y = (y_1,\ldots,y_N)$ étant 2 vecteurs de $(T_b^s)^N$,
l'erreur entachant le produit scalaire de X et Y, calculé en arithmétique flottante par des opérations $\oplus$ et $\otimes$ correctes, et noté $X \otimes Y$, vérifie :

$$|XY - X \otimes Y| < b^{1-s_1} \left(N|x_1y_1| + N|x_2y_2| + (N-1)\,|x_3y_3| + \ldots + 2|x_Ny_N|\right) \qquad \text{(III.13)}$$

cette écriture supposant $Nb^{1-s} < 0,1$, et s_1 étant défini pour une base b par :

$$s_1 = s - (0{,}08406 / \log_2 b)$$

b^{1-s_1} est à remplacer dans (III. 13) par :

$\frac{1}{2}\, b^{1-s_1}$ si $\oplus$ et $\otimes$ sont optimales

b^{1-s} si $\oplus$ et $\otimes$ sont I-dirigées pour respectivement, tout couple de T^2 de composantes de même signe, et tout couple de T^2,

et la conséquence :

A et B étant 2 matrices à éléments dans T, d'ordre N, l'erreur entachant le produit $A \otimes B$ des 2 matrices, calculé sur T par des opérations correctes, satisfait à :

$$| AB - A \otimes B | \leqslant b^{1-s_1} \; |A| \; D \; |B| \qquad \text{(III. 14)}$$

où $D = (d_i)$ est la matrice diagonale d'éléments $d_1 = N$, $d_i = N-i+2$ $i > 1$

et à fortiori :

$$||AB - A \otimes B||_\infty \leq N\, b^{1-s_1}\, ||A||_\infty\, ||B||_\infty \qquad \text{(III. 14')}$$

Relativement à la somme $A \oplus B$, de 2 matrices A et B, calculée sur T par une addition correcte, on a :

$$||A + B - (A \oplus B)||_\infty \leq b^{1-s}\, ||A + B||_\infty \qquad \text{(III. 15)}$$

Posant alors :

$|| \bar{R} ||_\infty = q$

$|| \Omega ||_\infty = \lambda\, || \bar{R} ||_\infty = \lambda\, q$

majorons les normes des différents termes de (III. 12). Afin de ne pas compliquer inutilement l'écriture, nous donnons les calculs relatifs à des opérations correctes, $\oplus$ et $\otimes$ étant I-dirigées pour respectivement, tout couple de T^2 de composantes de même signe, et tout couple de T^2.
Pour une telle arithmétique, on aura en particulier :

$$||A \otimes B||_\infty \leq || A ||_\infty\, || B ||_\infty \qquad \text{(III. 16)}$$

(III. 16) appliquée aux matrices $\bar{C}_i$ entraîne :

$|| \bar{C}_i ||_\infty \leq || B_0 ||_\infty\, q^i$

et :

$$|| \sum_{j=1}^{k} \sum_{i=1}^{j} M_i\, \bar{R}^{j-i} ||_\infty \leq N\, b^{1-s}\, || B_0 ||_\infty \sum_{j=1}^{k} j\, q^j$$

En prenant les majorations suivantes :

$$||\bar{B}_j||_\infty \leq \sum_{i=0}^{j} ||\bar{C}_i||_\infty \leq ||B_0||_\infty\, (1 + q + \ldots + q^j)$$

et

$$||R^j - \bar{R}^j||_\infty = ||(\bar{R}+\Omega)^j - \bar{R}^j||_\infty \leq q^j\, (C_j^1\, \lambda + \ldots + \lambda^j) = q^j\, [(1+\lambda)^j - 1]$$

et en remarquant que :

$B_0 = A^{-1} (I - R)$ donc :

$$||B_0||_\infty \leqslant ||A^{-1}||_\infty \ (1 + (1+\lambda)\ q)$$

il vient finalement :

$$\frac{||\ E'_k\ ||_\infty}{||\ A^{-1}\ ||_\infty} < (1 + (1+\lambda)\ q) \left[\sum_{j=o}^{k} \{(1+\lambda)^j - 1 + b^{1-s}\ (Nj + k - j + 1)\} q^j\right]$$

En particulier, l'utilisation de l'algorithme $\mathcal{Q}$ pour le calcul de $\bar{R}$ permet de choisir $\lambda = 2b^{1-s}$, et la formule précédente s'écrit :

$$\frac{||\ E'_k\ ||_\infty}{||\ A^{-1}\ ||_\infty} \leqslant (1 + q)\ b^{1-s} \sum_{j=o}^{k} (Nj + 2j + k + 1)\ q^j$$

Cette majoration donne alors, en fonction de B_0 et q , une indication sur l'erreur dûe aux calculs.

Soulignons que si $\bar{R}$ est calculée sur T sans correction, $|\ \Omega\ |$ est de l'ordre de $b^{1-s}\ |\ A\ |\ D\ |\ B_0\ |$ et dans le cas où $|\ A\ |\ |\ B_0\ | > > |\ A\ B_0\ |$ λ a une valeur non négligeable.

III. 2. 3. Exemple numérique

Nous avons effectué les calculs pour la correction de l'inverse de la matrice A d'ordre 6 , d'éléments $a_{ij} = \frac{M}{i + j - 1}$ (M=27720)

Nous avons écrit ci-dessous, en chaque position i, j :

1) La valeur exacte de l'élément de A^{-1}
2) La valeur calculée par la méthode de Gauss
3) La valeur obtenue après une correction
4) La valeur obtenue après deux corrections

(dans les 3 derniers cas, nous n'avons pas réécrit tous les chiffres obtenus, mais seulement ceux qui diffèrent de la valeur exacte).

$q = 0,825$

0,001298701					
3864					
677					
-0,02272727	0,5303030				
57182	265818				
639	2819				
	29				
0,1212121	-3,181818	20,36363			
01084	55308	21120			
056	664	275			
0	6	2			
-0,2727272	7,636363	-50,90909	130,9090		
697844	565578	50089	29 9878		
098	5947	671	38		
1	1	6	1		
0,2727272	-7,954545	54,54545	-143,1818	159,0909	
694311	875210	08720	2,1455	8,0802	
076	076	277	759	852	
1	2	3	.	.	
	3,000000		56,00000	-63,00000	25,20000
-0,1000000	2,968432	-21,00000	5,58675	2,59642	05740
9868890	2,999812	0,81747	5,99762	2,99771	19920
9999210	2,999998	0,99892	5,99997	2,99997	19998
9999985		0,99997			

Rappelons que l'on utilise, de préférence à la méthode précédente, la méthode, de raffinement itératif de l'inverse d'une matrice, à convergence quadratique :

$$B_{k+1} = B_k \ (2\ I - A\ B_k)$$

Les produits $A\ B_k$ pourront être calculés avec précision par les techniques du chapitre II .

IV - RESOLUTION NUMERIQUE DE PROBLEMES MAL CONDITIONNES

IV . 1 . INTRODUCTION

L'impossibilité de résoudre numériquement un problème par arithmétique flottante, est subordonnée à la longueur de la représentation des nombres.
Lorsque la possibilité de se tourner vers la double précision a été épuisée, les techniques de calcul à chaque pas de l'erreur élémentaire, permettront la résolution du problème, la limite imposée à la précision devenant fonction de l'augmentation du coût et de l'encombrement.

Comme lors de la correction d'une somme, on calcule conjointement aux valeurs à déterminer, les erreurs les entachant.

L'accumulation des erreurs ne se fait généralement plus linéairement, comme au paragraphe I.

Nous illustrons la méthode sur deux exemples simples.

IV . 2 . EVALUATION NUMERIQUE D'UN POLYNOME

Si l'on calcule la valeur $P(x)$ d'un polynôme P en un point x voisin d'un zéro mal conditionné, le danger est grand d'obtenir une valeur numérique présentant une perte importante de chiffres significatifs ou même de signe erroné. Ce phénomène, dû à des cancellations successives, est particulièrement génant dans les processus itératifs lorsqu'il génère une décision incorrecte dans le déroulement de l'algorithme (J.H.WILKINSON [42])

Soit P un polynôme défini sur T par ses coefficients a_i $(i = 0 , \dots , N)$ et x un élément de T .
Le calcul de la valeur de P en x :

$$P(x) = a_o x^N + \dots\dots + a_N$$

par le schéma de Horner, sur respectivement $\mathbb{R}$ et T , définit les suites b_i , $\overline{b}_i$ et l'erreur $\varepsilon_i = b_i - \overline{b}_i$ $(i = 0 , \dots , N)$ par :

$$\begin{cases} b_o = a_o \\ b_i = b_{i-1}\, x + a_i & i = 1 , \dots , N \end{cases}$$

$$\begin{cases} \overline{b}_o = a_o \\ \overline{b}_i = \overline{b}_{i-1} \otimes x \oplus a_i & i = 1 , \dots , N \end{cases}$$

Désignant par e_i et e'_i les erreurs absolues commises au pas i :

$$e_i = \overline{b}_{i-1}\, x - (\overline{b}_{i-1} \otimes x) \qquad i = 1 , \dots , N$$

$$e'_i = (\overline{b}_{i-1} \otimes x + a_i) - \overline{b}_i \qquad i = 1 , \dots , N$$

il vient immédiatement que la suite ε_i vérifie la relation de récurrence :

$$\begin{cases} \varepsilon_o = 0 \\ \varepsilon_i = \varepsilon_{i-1}\, x + e_i + e'_i & i = 1 , \dots , N \end{cases}$$

L'adjonction au calcul de $\overline{b}_i$, de l'algorithme :

$$\begin{cases} \overline{\varepsilon}_o = 0 \\ \overline{\varepsilon}_i = \overline{\varepsilon}_{i-1} \otimes x \oplus e_i \oplus e'_i & i = 1 , \dots , N \end{cases}$$

permet, par la correction $\overline{P}(x) = \overline{b}_N \oplus \overline{\varepsilon}_N$, de regagner de nombreux chiffres significatifs.

EXEMPLE : Evaluation au point $x = N - 0{,}01$ du polynôme $P_N(x) = \prod_{j=1}^{N} (x-j)$

Les valeurs obtenues pour l'arithmétique IBM-360 en double précision ont été :

N	VALEUR MACHINE	VALEUR CORRIGEE	VALEUR EXACTE
5	-0.2350349000999650 00	-0.2350349001000000 00	-0.2350349001000000 00
6	-0.1172824151506800 01	-0.1172824151499000 01	-0.1172824151499000 01
7	-0.7025216667152560 01	-0.7025216667479020 01	-0.7025216667479020 01
8	-0.4910626451456940 02	-0.4910626450567830 02	-0.4910626450567830 02
9	-0.3923590534221440 03	-0.3923590534003700 03	-0.3923590534003700 03
10	-0.3527307879189500 04	-0.3527307890069320 04	-0.3527307890069320 04
11	-0.3523780583791440 05	-0.3523780582175260 05	-0.3523780582175250 05
12	-0.3872634824903010 06	-0.3872634859815000 06	-0.3872634859815000 06
13	-0.4643289397177700 07	-0.4643289196918190 07	-0.4643289196918190 07
14	-0.6031623828843690 08	-0.6031632666796720 08	-0.6031632666796720 08
15	-0.8438254919826660 09	-0.8438254100848620 09	-0.8438254100848610 09
16	-0.1264902017411720 11	-0.1264894289717180 11	-0.1264894289717180 11
17	-0.2022596617919380 12	-0.2022565969258120 12	-0.2022565969258120 12
18	-0.3436495415151000 13	-0.3436339581769540 13	-0.3436339581769540 13
19	-0.6191939369715200 14	-0.6181974907603410 14	-0.6181974907603400 14

Nous avons souligné les chiffres inexacts.

On remarque la concordance entre valeurs exacte et corrigée, et la perte progressive de chiffres significatifs dans l'évaluation simple de $P(x)$.

IV . 3 . L'ELIMINATION DE GAUSS

De la même façon qu'en IV . 2 , dans la résolution par l'algorithme de Gauss, du système linéaire :

$$A\ x = b$$

où $A = (a_{ij})$ et $b = (b_i)$ sont définis sur T ,
on associe aux calculs les algorithmes auxiliaires suivants :
- les notations sont celles du § II -

A l'algorithme de triangularisation :

$$\bar{\rho}_i^k = \bar{a}_{ik}^k \oslash \bar{a}_{kk}^k$$

$$\bar{a}_{ij}^{k+1} = \bar{a}_{ij}^k \ominus \bar{\rho}_i^k \otimes \bar{a}_{kj}^k$$

on adjoint l'algorithme suivant, calculant la partie principale de l'erreur $\theta_{ij}^k = a_{ij}^k - \bar{a}_{ij}^k$:

$$(\bar{b}_i^k = \bar{a}_{i,n+1}^k)$$

$$\Delta\bar{\rho}_i^k = (e_i^k \cdot \bar{a}_{kk}^k \oplus \bar{\theta}_{ik}^k \ominus \bar{\rho}_i^k \otimes \bar{\theta}_{kk}^k) \oslash \bar{a}_{kk}^k$$

$$\bar{\theta}_{ij}^{k+1} = \bar{\theta}_{ij}^k \oplus e''^k_{ij} \ominus e'^k_{ij} \ominus \Delta\bar{\rho}_i^k \otimes \bar{a}_{kj}^k \ominus \bar{\theta}_{kj}^k \otimes \bar{\rho}_i^k$$

A l'algorithme de retour-arrière :

$$\bar{x}_i = \left[\bar{d}_i \ominus \{ \oplus \sum_{j=i+1}^{n} \bar{t}_{ij} \otimes \bar{x}_j \}\right] \oslash \bar{t}_{ii}$$

on adjoint le calcul de la partie principale de l'erreur $\xi_i = x_i - \bar{x}_i$:
E C et E P ci-dessous représentent respectivement, l'erreur "générée" dans le calcul de $\bar{x}_i$, et l'erreur "propagée" par les calculs antécédents.

$$E\,P = \bar{\theta}^{n}_{i,n+1} \ominus \Big[\sum_{j=i+1}^{n} \bar{\theta}^{n}_{ij} \otimes \bar{x}_j \oplus \bar{t}_{ij} \otimes \bar{\xi}_j \Big] \ominus \bar{x}_i \otimes \bar{\theta}^{n}_{ii}$$

$$E\,C = \ominus \sum_{j=i+1}^{n} \varepsilon'_{ij} \ominus \sum_{j=i+1}^{n-1} \varepsilon''_{ij} \oplus \varepsilon''_{in} \oplus \varepsilon_i \; \bar{t}_{ii}$$

$$\bar{\xi}_i = (E\,P \oplus E\,C) \oslash \bar{t}_{ii}$$

TECHNIQUE DE CORRECTION : Remarquons que la correction des calculs peut s'effectuer selon les procédés suivants :

1) Correction en fin de résolution uniquement : Au vecteur $\bar{x}$ calculé par l'algorithme de Gauss, on ajoute, à la fin des calculs, le vecteur correction $\bar{\xi}$.

2) La correction 1 est précédée, lorsque la triangularisation de la matrice A est terminée et avant le retour-arrière, de la correction de la matrice triangulaire T et du second membre d obtenus.

3) Correction, à chaque pas, des valeurs calculées par l'algorithme de Gauss :

- par addition de $\bar{\theta}^{k+1}_{ij}$ à $\bar{a}^{k+1}_{ij}$ $\quad \forall\, k, i, j$
- par addition de $\bar{\xi}_i$ à $\bar{x}_i$ $\quad \forall\, i$

l'erreur résultant de l'addition sur T des quantités ci-dessus étant replacée dans respectivement $\bar{\theta}^{k+1}_{ij}$ ou $\bar{\xi}_i$.

Seule la troisième méthode assure, pour le coût, une précision suffisante.

EXEMPLE : Résolution de $H_N\, x = u$, où H_N est la matrice de Hilbert d'ordre N , u le 1er vecteur de la base naturelle de $\mathbb{R}^N$.

Les résultats suivants ont été obtenus :

	Chiffres significatifs	
Solution calculée par l'algorithme de Gauss :	5	4
Solution corrigée par le procédé 1 :	11	7 ou 8
Solution corrigée par le procédé 2 :	12	9 ou 10
Solution corrigée par le procédé 3 :	15	15
	N = 10	N = 11

Ainsi :

N = 11 VALEUR MACHINE	VALEUR CORRIGEE
0.1209949801519040 03	0.1210000000000000 03
-0.7259462091901830 04	-0.7260000000000000 04
0.1415557772889990 06	0.1415700000000000 06
-0.1321158444728510 07	-0.1321320000000000 07
0.6935954572852040 07	0.6936930000000000 07
-0.2219470702487170 08	-0.2219817600000000 08
0.4491725168592840 08	0.4492488000000000 08
-0.5775007009859060 08	-0.5776056000000000 08
0.4571832996706480 08	0.4572711000000000 08
-0.2031907018970570 08	-0.2032316000000000 08
0.3879063217945310 07	0.3879876000000000 07

COUT DE LA METHODE : La correction à chaque pas des valeurs calculées par l'algorithme de Gauss, est coûteuse.
Le temps nécessité par les calculs est environ multiplié par 7 lorsque les erreurs d'opérations élémentaires sont déterminées par les méthodes du chapitre II .
Ce temps de calcul est multiplié par 4 lorsque l'accès au poids faible du résultat d'une opération flottante élémentaire peut être effectué par une instruction arithmétique (exemple : CDC série 6000) .

PROGRAMME FORTRAN : Les sous-programmes FUNCTION auxiliaires utilisés : ES (X, Y, S) , EP (X, Y, P) , CD (X, Y, D) sont ceux du chapitre II.

```
      SUBROUTINE CGAUSS(A,B,N,X,EA,EB,EX)
      REAL*8 A(20,20),EA(20,20),B(20),EB(20),X(20),EX(20)
      REAL*8 ES,EP,CD
      DOUBLE PRECISION R,ER,V,S,Z,W
      NM1=N-1
      DO 7 I=1,N
      EB(I)=0.0D0
      DO 7 J=1,N
      EA(I,J)=0.0D0
    7 CONTINUE
      DO 2 K=1,NM1
      KP1=K+1
      DO 2 I=KP1,N
      R=A(I,K)/A(K,K)
      ER=EA(I,K)-EA(K,K)*R+CD(A(I,K),A(K,K),R)
      ER=ER/A(K,K)
      S=R
      R=S+ER
      ER=ES(S,ER,R)
      V=R*B(K)
      S=B(I)-V
      EB(I)=EB(I)-ER*B(K)-R*EB(K)-EP(R,B(K),V)+ES(B(I),-V,S)
      B(I)=S+EB(I)
      EB(I)=ES(S,EB(I),B(I))
      DO 2 J=KP1,N
      V=R*A(K,J)
      S=A(I,J)-V
      EA(I,J)=EA(I,J)-ER*A(K,J)-EA(K,J)*R-EP(R,A(K,J),V)+ES(A(I,J),-V,S)
      A(I,J)=S+EA(I,J)
      EA(I,J)=ES(S,EA(I,J),A(I,J))
    2 CONTINUE
      X(N)=B(N)/A(N,N)
      EX(N)=EB(N)-EA(N,N)*X(N)+CD(B(N),A(N,N),X(N))
      EX(N)=EX(N)/A(N,N)
      S=X(N)
      X(N)=S+EX(N)
      EX(N)=ES(S,EX(N),X(N))
      DO 3 L=2,N
      I=N+1-L
      IP1=I+1
      S=0.0D0
      Z=0.0D0
      R=0.0D0
      DO 4 J=IP1,N
      V=A(I,J)*X(J)
      W=S+V
      Z=Z+EP(A(I,J),X(J),V)+ES(S,V,W)
      S=W
      R=R+EA(I,J)*X(J)+A(I,J)*EX(J)
    4 CONTINUE
      S=-S
      V=B(I)+S
      X(I)=V/A(I,I)
      Z=-Z+ES(B(I),S,V)+CD(V,A(I,I),X(I))
      R=-R+EB(I)-X(I)*EA(I,I)
      EX(I)=(Z+R)/A(I,I)
      S=X(I)
      X(I)=S+EX(I)
      EX(I)=ES(S,EX(I),X(I))
    3 CONTINUE
      RETURN
      END
```

V - CRITERE D'APPLICATION POUR LES METHODES DES § III et IV

La question essentielle qui se pose, avant l'application à une solution calculée, d'une méthode de correction, est la question de la validité de la solution approchée.

Pour déterminer si une méthode de raffinement est nécessaire et possible, on appliquera, au préalable, au résultat approché, les méthodes de contrôle de résultats d'algorithmes numériques définies par M. LA PORTE et J. VIGNES [19] :

La méthode générale de permutation-perturbation définie par ces auteurs, ou le critère des résidus normés pour la résolution de systèmes linéaires ou d'équations algébriques [20] , permettra d'évaluer la précision de la solution approchée.

Seules les solutions de précision "moyenne" devront et pourront être corrigées .

CHAPITRE IV

L'ERREUR D'ARRONDI DANS LES METHODES ITERATIVES LINEAIRES

INTRODUCTION

Ce chapitre est relatif à la résolution numérique d'un système d'équations linéaires par itérations, en arithmétique à virgule flottante.

Deux problèmes sont abordés :

1) Un système d'équations ayant été résolu par la convergence, sur le système de nombres à virgule flottante, de la suite des itérés d'une méthode itérative linéaire, existe-t-il une possibilité d'amélioration de la précision ?
Une méthode est proposée, utilisant les techniques définies au chapitre II.

2) Possibilité de prévoir à priori, en fonction de la matrice de l'itération linéaire et du vecteur initial, le type de convergence satisfait par la suite des itérés calculés sur le système T de nombres à virgule flottante.

I - RAFFINEMENT DE LA SOLUTION NUMERIQUE D'UN SYSTEME LINEAIRE, CALCULEE PAR UNE METHODE ITERATIVE.

I . 1 . NOTATIONS

T désigne le système T_b^s de nombres à virgule flottante muni des opérations $\oplus$, $\ominus$, $\otimes$, $\oslash$.

$M = (m_{ij})$, $x = (x_i)$ et $y = (y_i)$ étant respectivement une matrice d'ordre N définie sur T et 2 vecteurs de T^N , $M \otimes x$ et $x \oplus y$ représenteront les vecteurs définis sur T^N par :

$$M \otimes x = \left(\bigoplus_{j=1}^{N} m_{ij} \otimes x_j \right)$$

$$x \oplus y = (x_i \oplus y_i)$$

$B = (b_{ij}) \in \mathcal{M}_{N,N}(T)$ et $c = (c_i) \in T^N$ étant donnés, considérons la résolution du système linéaire :

$$x = B\,x + c \qquad \text{(IV.1)}$$

$x^{(o)} = \bar{x}^{(o)}$ étant un vecteur initial sur T^N , soient $\{x^{(n)}\}$ et $\{\bar{x}^{(n)}\}$ les suites définies respectivement sur $\mathbb{R}^N$ et T^N par :

$$x^{(n+1)} = B\,x^{(n)} + c \qquad \text{(IV.2)}$$

$$\bar{x}^{(n+1)} = B \otimes \bar{x}^{(n)} \oplus c \qquad \text{(IV.3)}$$

La convergence de la suite $\{x^{(n)}\}$ vers la solution x de (IV.1) , est assurée à la condition nécessaire et suffisante que le rayon spectral de B soit inférieur à 1 , ce que nous supposerons.

La suite $\{\bar{x}^{(n)}\}$ est dite "numériquement" convergente lorsqu'est réalisée l'une des deux possibilités suivantes :

a) Il existe un entier K vérifiant :
$$\bar{x}^{(K+1)} = \bar{x}^{(K)} \qquad \text{(convergence forte)}$$

b) Il existe un entier K à partir duquel les itérés cyclent sur un nombre fini et petit de valeurs, c'est-à-dire qu'il existe K et $y^1, \ldots, y^p$ vérifiant :

$$\bar{x}^{(K+k)} = y^j \qquad \forall\, k \geq 0 \;\wedge\; k \equiv j - 1 \pmod{p}$$ (convergence faible)

I . 2 . METHODE

Lorsque (IV.1) est résolu numériquement par la convergence sur T de la suite $\{\bar{x}^{(n)}\}$, le vecteur erreur - écart entre les solutions calculées sur $\mathbb{R}$ et T de (IV.1) - est solution d'un système linéaire de même matrice que (IV.1). En effet :

I . 2 . 1 $\{\bar{x}^{(n)}\}$ converge fortement

La limite $\bar{x} = (\bar{x}_i)$ de $\{\bar{x}^{(n)}\}$ vérifie :

$$\bar{x} = B \otimes \bar{x} \oplus c$$

Soit $e = (e_i)$ le vecteur défini par :

$$e = (B\,\bar{x} + c) - \bar{x} \qquad \text{(IV.4)}$$

(IV.1) et (IV.4) entraînent pour le vecteur erreur $\varepsilon = x - \bar{x}$:

$$\varepsilon = B\,\varepsilon + e \qquad \text{(IV.5)}$$

I . 2 . 2 $\{\bar{x}^{(n)}\}$ converge faiblement

On a les égalités :

$$\begin{cases} y^2 = B \otimes y^1 \oplus c \\ y^3 = B \otimes y^2 \oplus c \\ \cdots \\ y^1 = B \otimes y^p \oplus c \end{cases}$$

Particularisant y^1 par exemple, considérons ce vecteur comme la solution de (IV.1) calculée sur T.

Soit e' le vecteur défini par :

$$e' = B\,y^1 + c - y^2 \qquad \text{(IV.6)}$$

il vient pour le vecteur erreur $\varepsilon = x - y^1$:

$$\varepsilon = B\,\varepsilon + e \qquad \text{avec} \qquad e = e' + y^2 - y^1 \qquad \text{(IV.7)}$$

Un procédé de "correction" peut donc être défini par la résolution de (IV.5) - resp. (IV.7) - selon la méthode itérative employée pour la résolution de (IV.1) .

Les ordres de convergence asymptotique, de la méthode de résolution du système et de la méthode de correction, sont égaux, entraînant précisions et vitesses de convergence analogues. Le procédé peut être itéré.

I . 2 . 3 Calcul de e et e' définis en (IV.4) et (IV.6)

e(ou e') représente l'erreur d'arrondi générée par une itération sur T . Rappelons :

Notant u un vecteur de T^N et définissant V et $\overline{V}$ par :

$V = B\,u + c$

$\overline{V} = B \otimes u \oplus c$

si e^P_{ij} et e^S_{ij} $(1 \leq j \leq N)$ sont respectivement les erreurs absolues de multiplication et d'addition dans le calcul de la ième composante de $\overline{V}$:

$$\overline{V}_i = \left(\bigoplus_{j=1}^{N} b_{ij} \otimes u_j\right) \oplus c_i \qquad 1 \leq i \leq N$$

le vecteur $V - \overline{V}$ s'exprime en fonction des erreurs élémentaires définies, que l'on calculera par les méthodes décrites au chapitre II, suivant :

$$(V - \overline{V})_i = \sum_{j=1}^{N} e^P_{ij} + \sum_{j=1}^{N} e^S_{ij} \qquad 1 \leq i \leq N$$

Remarque Lorsque le système linéaire à résoudre n'est pas explicitement donné sous la forme (IV.1) mais par :

$A\,x = b$ où $A \in \mathcal{M}_{N,N}(T)$ et $b \in T^N$

l'expression de e doit être légérement modifiée suivant la méthode itérative employée pour résoudre le système.

Ainsi, pour les méthodes de Jacobi ou de Gauss-Seidel, la convergence, forte par exemple, de la suite numérique s'écrit :

$$\overline{x}_i = \left[\left\{\bigoplus_{\substack{j=1 \\ \neq i}}^{N} (-a_{ij}) \otimes \overline{x}_j\right\} \oplus b_i\right] \oslash a_{ii} \qquad 1 \leq i \leq N$$

Le vecteur erreur ε vérifie :

$$\varepsilon_i = \frac{\sum_{j \neq i} (-a_{ij})\ \varepsilon_j + e_i}{a_{ii}} \qquad 1 \leq i \leq N$$

avec :

$$e_i = \sum_{\substack{j=1 \\ \neq i}}^{N} e^P_{ij} + \sum_{j=1}^{N-1} e^S_{ij} + a_{ii}\ e^D_i$$

et les notations suivantes :

$$e^P_{ij} = (-a_{ij}\ \bar{x}_j) - (-a_{ij} \otimes \bar{x}_j) \qquad j \neq i$$

e^S_{ij} = erreurs absolues commises lors des sommes partielles dans le calcul de :

$$\{ \bigoplus \sum_{j \neq i} (-a_{ij}) \otimes \bar{x}_j \} \bigoplus b_i = s_i$$

$$e^D_i = s_i \ / \ a_{ii} - \bar{x}_i$$

(le nombre $a_{ii}\ e^D_i$ se calcule exactement sur T).

I . 3 . BORNES DE L'ERREUR

I.3.1. Une borne supérieure du vecteur erreur ε

Le vecteur ε vérifie :

$$\varepsilon = (I-B)^{-1}\ e$$

Avec les notations des pages 73 - 74 où $D = (d_i)$ désigne la matrice diagonale d'ordre N , d'éléments $d_1 = N$, $d_i = N-i+2$ $i>1$, il vient :

$$|e| \leq b^{1-s_1} \ (\ |B|\ D + I)\ \ |\bar{x}|$$

$$|e'| \leq b^{1-s_1} \ (\ |B|\ D\ |y^1| + |y^2| \ )$$

et en normes, pour l'une des trois normes de matrices $\| \ \|_1$, $\| \ \|_2$ ou $\| \ \|_\infty$, suivant le type de convergence de $\{ x^{(n)} \}$:

$$\|\varepsilon\| \leq b^{1-s_1} \|(I-B)^{-1}\| \ (N \ \||B|\| + 1) \ \|\bar{x}\|$$

$$\|\varepsilon\| \leq \|(I-B)^{-1}\| \ [b^{1-s_1} (N \ \||B|\| \ \|y^1\| + \|y^2\|) + \|y^2 - y^1\|] \qquad \text{(IV.8)}$$

où l'on peut aussi faire apparaître le conditionnement de $I - B$.

Cas particulier important où $\|B\| < 1$

Les formules (IV.8) deviennent, avec la majoration $\|(I-B)^{-1}\| \leq \frac{1}{1 - \|B\|}$

$$\|\varepsilon\| \leq b^{1-s_1} \frac{N \ \||B|\| + 1}{1 - \|B\|} \ \|\bar{x}\|$$

$$\|\varepsilon\| \leq \frac{b^{1-s_1} (N \ \||B|\| \ \|y^1\| + \|y^2\|) + \|y^2 - y^1\|}{1 - \|B\|}$$

I.3.2. Amélioration de la précision après une correction

Numériquement, la résolution de :

$$\varepsilon = B \varepsilon + e$$

et la correction de la solution $\bar{x}$ ou y^1 , est réalisée par les calculs suivants :

1) calcul sur T de e

L'élément $\bar{e}$ de T est déterminé avec :

$$e = \bar{e} + \eta$$

2) résolution sur T de $\varepsilon = B \varepsilon + \bar{e}$

le vecteur $\bar{\varepsilon}$ de T^N est calculé .

Considérons par exemple le cas de convergence numérique forte des itérés sur T du système initial et du système de correction.

Le nouveau vecteur solution x_c est :

$$x_c = \bar{x} + \bar{\varepsilon}$$

vérifiant avec la notation $e^2 = (B \bar{\varepsilon} + \bar{e}) - \bar{\varepsilon}$

$$x - x_c = (I-B)^{-1} (\eta + e^2) \qquad \text{(IV.9)}$$

Si les opérations $\circledast$ sont correctes, on a au moins les majorations suivantes

$$|\eta_i| \leq (2N-1)\, b^{1-s}\, (1+2N\, b^{1-s})\, (\sum_{j=1}^{N} |e^P_{ij}| + \sum_{j=1}^{N} |e^S_{ij}|)$$

$$\|\eta\| \leq 2N\, b^{2-2s_1}\, (N\, \||B|\| + 1)\, \|\bar{x}\|$$

et

$$\|e^2\| \leq b^{1-s_1}\, (N\, \||B|\| + 1)\, \|\bar{\varepsilon}\|$$

soit, en négligeant les erreurs du 2ème ordre, lorsque $\|B\| < 1$:

$$\|x-x_c\| \leq b^{2-2s_1}\, \frac{N\, \||B|\| + 1}{1 - \|B\|}\, (2N + \frac{N\, \||B|\| + 1}{1 - \|B\|})\, \|\bar{x}\|$$

Les majorations précédentes sont souvent, malheureusement, beaucoup trop larges. Nous considèrons ci-dessous, le problème de l'évaluation précise de l'incertitude sur la solution numérique à corriger.

I. 4. EVALUATION, AVANT CORRECTION, DE L'INCERTITUDE SUR LA SOLUTION NUMERIQUE A RAFFINER

Dans ce paragraphe, nous utilisons les résultats de M. LA PORTE et J. VIGNES [19] , [20] , relatifs au contrôle des erreurs d'arrondi :

Lorsque (IV. 1) a été résolu numériquement par la convergence sur T de la suite $\{ \bar{x}^{(n)} \}$ vers $\bar{x}$, il est, comme pour une méthode directe de résolution, nécessaire, avant d'aborder la correction, de connaitre le nombre de chiffres significatifs des composantes du vecteur solution numérique. En effet, rappelons :

i) De même que précedemment, si tous les chiffres de la solution ont été obtenus significatifs, une correction est évidemment inutile, mais si aucun ne l'a été, la méthode proposée sera insuffisante pour améliorer la précision.

ii) La connaissance du nombre de chiffres significatifs permettra de déterminer combien de fois doit être itérée la méthode de correction : ainsi, si les calculs s'effectuent par exemple sur un système de nombres flottants à 6 chiffres hexadécimaux de mantisse, l'obtention de 2 chiffres hexadécimaux corrects à la 1ère résolution du système, appelera 2 corrections successives.

La détermination du nombre de chiffres significatifs d'un résultat numérique par la méthode de permutation-perturbation est effectuée selon le schéma suivant [20] :

Désignant par :

$$Z^{V} = \{ z^{1} , z^{2} , \ldots , z^{V} \}$$

une population de vecteurs solutions sur T de (IV. 1) , on définit les vecteurs ci-dessous :

$\bar{z}$ de composantes $\bar{z}_j = \frac{1}{v} \sum_{i=1}^{v} z_j^i$

δ^2 de composantes $\delta_j^2 = \frac{1}{v-1} \sum_{i=1}^{v} (z_j^i - \bar{z}_j)^2$

$\hat{\varepsilon}$ de composantes $\hat{\varepsilon}_j = \sqrt{(z_j^1 - \bar{z}_j)^2 + \delta_j^2}$

C^{V} de composantes $C_j^V = \log_{10} \left| \frac{z_j^1}{\hat{\varepsilon}_j} \right|$

L'algorithme d'obtention du nombre de chiffres significatifs de chaque composante du résultat numérique est alors le suivant :
1) On résoud numériquement (IV. 1) sur T : la solution $\bar{x}$ est calculée.
2) Choisissant $\bar{x}$ comme vecteur initial, on calcule par itérations, à l'aide d'une arithmétique flottante différente de celle précédemment utilisée, une seconde solution de(IV.1) sur le système de nombres à virgule flottante, notée z^1.

La nouvelle arithmétique s'obtient en prenant aléatoirement $\circledast_\nabla$ ou $\circledast_\Delta$ pour chacune des opérations définies sur T .

3) On calcule C^2 correspondant à $Z^2 = \{ \bar{x} , z^1 \}$. Deux cas sont alors possibles :

a) si $C^2_j < 1 \quad \forall j$, la solution $\bar{x}$ est à rejeter.

b) si l'inégalité précédente n'est pas vérifiée, on calcule de nouveaux éléments de la population. Chaque nouvel élément s'obtient par résolution itérative de (IV. 1), à l'aide d'opérations flottantes générées aléatoirement comme précédemment ; le vecteur initial est pris égal à $\bar{x}$ ou à l'une des solutions calculées.

A chaque nouvel élément z^i on associe C^{i+1} correspondant à $Z^i = \{ \bar{x} , z^1 , \ldots , z^i \}$ et :

$C^{i+1}_j < 1 \quad \forall_j$ entraîne la même conclusion que précédemment.

Sinon : ou C^{i+1} est stationnaire et chacune de ses composantes donne le nombre de chiffres significatifs de la composante correspondante de $\bar{x}$

ou C^{i+1} n'a pas atteint sa valeur stationnaire et il faut poursuivre le procédé.

Soulignons que la méthode de permutation-perturbation pour la détermination du nombre de chiffres significatifs d'un résultat numérique calculé par méthode itérative, est peu coûteuse, un petit nombre d'itérations étant généralement nécessaire pour calculer, à partir d'un vecteur initial solution de (IV.1) sur le système de nombres à virgule flottante, une autre solution de (IV.1) sur ce système.

I. 5. EXEMPLE NUMERIQUE

Nous avons considéré la résolution, par les méthodes de Jacobi et de Gauss - Seidel, du système linéaire $Ax = b$ pour :

$$A = \begin{vmatrix} 3 & -1 & -0,5 & -0,25 & -0,5 & -0,125 & 0 & -0,5 \\ -2 & 4 & -0,25 & -1 & 0 & -0,25 & 0 & -0,25 \\ -4 & -1 & 8,25 & -0,5 & -0,5 & -1 & -0,5 & -0,25 \\ -1 & -0,5 & -1 & 4 & 0 & 0 & -1 & -0,25 \\ 0 & -1 & -1 & -1,5 & 4,5 & -0,5 & -0,25 & 0 \\ -5 & -1,5 & 0 & -1,5 & 0 & 9 & -0,5 & 0 \\ -0,5 & -2,5 & -0,5 & 0 & -0,5 & 0 & 5 & 0 \\ -3 & -0,5 & -0,5 & 0 & 0 & -0,25 & -0,25 & 5 \end{vmatrix} \qquad B = \begin{vmatrix} 0,125 \\ 0,25 \\ 0,5 \\ 0,25 \\ 0,25 \\ 0,5 \\ 1 \\ 0,5 \end{vmatrix}$$

Les vecteurs x , $\bar{x}$ et x_c désignent respectivement :
la solution exacte du système
la solution obtenue après convergence forte en K_1 itérations
la solution corrigée après K_2 itérations supplémentaires

$$x = \begin{vmatrix} 1 \\ " \\ " \\ " \\ " \\ " \\ " \\ " \end{vmatrix} \qquad \bar{x} = \begin{vmatrix} 0,9999905 \\ " \quad 09 \\ " \quad 11 \\ " \quad 14 \\ " \quad 15 \\ " \quad 12 \\ " \quad 25 \\ " \quad 14 \end{vmatrix} \qquad x_c = \begin{vmatrix} 0,9999999 \\ " \\ " \\ " \\ " \\ " \\ " \\ " \end{vmatrix}$$

avec :

$K_1 = 208$, $K_2 = 65$ pour la méthode de Jacobi

$K_1 = 106$, $K_2 = 33$ pour la méthode de Gauss - Seidel

Le vecteur initial $x^{(0)}$ ayant, dans tous les cas, été choisi égal au vecteur nul.

Une seule correction était nécessaire, puisque la méthode de permutation-perturbation conduisait à la stationnarité sur 5 du nombre de chiffres significatifs de chaque composante, après calcul des deux solutions supplémentaires z^1 et z^2 :

$$z^1 = \begin{vmatrix} 0.9999946E\ 00 \\ 0.9999957E\ 00 \\ 0.9999949E\ 00 \\ 0.9999959E\ 00 \\ 0.9999953E\ 00 \\ 0.9999956E\ 00 \\ 0.9999959E\ 00 \\ 0.9999956E\ 00 \end{vmatrix} \qquad C^2 = \begin{vmatrix} 5,4 \\ 5,4 \\ 5,5 \\ 5,4 \\ 5,5 \\ 5,4 \\ 5,5 \\ 5,4 \end{vmatrix}$$

$$z^2 = \begin{vmatrix} 0.9999997E\ 00 \\ 0.9999986E\ 00 \\ 0.9999998E\ 00 \\ 0.9999988E\ 00 \\ 0.9999996E\ 00 \\ 0.9999991E\ 00 \\ 0.9999995E\ 00 \\ 0.9999990E\ 00 \end{vmatrix} \qquad C^3 = \begin{vmatrix} 5 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{vmatrix} \qquad \text{(partie entière)}$$

II - ITERATIONS LINEAIRES DEFINIES PAR UNE MATRICE POSITIVE (RESP. NEGATIVE)

Des remarques particulières peuvent être faites, lorsque la matrice $B = (b_{ij})$ vérifie $b_{ij} \geqslant 0 \quad \forall\, i, j$ (resp. $b_{ij} \leqslant 0 \quad \forall i, j$).

Considèrons le cas $B \geqslant 0$:

Si le vecteur $x^{(1)} - x^{(0)}$ a toutes ses composantes de même signe, la suite des itérés théoriques sera monotone par l'égalité :

$$x^{(n+1)} - x^{(n)} = B\,(x^{(n)} - x^{(n-1)}) = B^n\,(x^{(1)} - x^{(0)}) \qquad \text{(IV. 10)}$$

D'autre part, par exemple de l'hypothèse :

$$x^{(1)} - x^{(0)} \geqslant 0$$

soit

$$B\,x^{(0)} + c - x^{(0)} \geqslant 0$$

$$B\,x^{(0)} + x - Bx - x^{(0)} \geqslant 0$$

$$(I - B)\,(x - x^{(0)}) \geqslant 0$$

il vient, en remarquant que $(I - B)^{-1}$ est une matrice positive ($\rho\,(B) < 1$) :

$$x \geqslant x^{(0)}$$

Dans l'hypothèse $x^{(1)} \geqslant x^{(0)}$ (resp. $x^{(1)} \leqslant x^{(0)}$), la suite $\{x^{(n)}\}$ est monotone croissante (resp. décroissante), bornée par x .

La propriété de monotonie est conservée par la suite $\{\bar{x}^{(n)}\}$ des itérés calculés sur T , lorsque dans le calcul des itérés successifs, les opérateurs en positions identiques gardent, pour tous les calculs, la même définition (par exemple $\circledast_{\nabla}$ ou $\circledast_{\Delta}$) :

Deux tableaux de N lignes et N colonnes, définissant en effet les opérations de multiplication $\otimes = (\otimes_{ij})$ et d'addition $\oplus = (\oplus_{ij})$, dès que les opérations $\circledast$ sont isotones, il vient :

L'hypothèse :

$$\bar{x}_j^{(n-1)} \leqslant \bar{x}_j^{(n)} \qquad 1 \leqslant j \leqslant N$$

entraînera :

$$b_{ij} \; \bar{x}_j^{(n-1)} \leqslant b_{ij} \; \bar{x}_j^{(n)} \qquad 1 \leqslant i \leqslant N \;,\; 1 \leqslant j \leqslant N$$

$$b_{ij} \otimes_{ij} \bar{x}_j^{(n-1)} \leqslant b_{ij} \otimes_{ij} \bar{x}_j^{(n)} \qquad " \qquad "$$

et par l'isotonie des opérations d'addition :

$$\oplus_{ij} \Big(\sum_{j=1}^{N} b_{ij} \otimes_{ij} \bar{x}_j^{(n-1)} \Big) \oplus_{iN} c_i \leqslant \oplus_{ij} \Big(\sum_{j=1}^{N} b_{ij} \otimes_{ij} \bar{x}_j^{(n)} \Big) \oplus_{iN} c_i$$

d'où :

$$\bar{x}_i^{(n)} \leqslant \bar{x}_i^{(n+1)} \qquad 1 \leqslant i \leqslant N$$

$$\bar{x}^{(0)} \leqslant \bar{x}^{(1)} \;\wedge\; B \geqslant 0 \implies \bar{x}^{(n)} \leqslant \bar{x}^{(n+1)} \qquad \forall\, n$$

$$\bar{x}^{(0)} \geqslant \bar{x}^{(1)} \;\wedge\; " \implies \bar{x}^{(n)} \geqslant \bar{x}^{(n+1)} \qquad "$$

La suite de vecteurs $\{ \bar{x}^{(n)} \}$ est non décroissante si $\bar{x}^{(0)} \leqslant \bar{x}^{(1)}$, non croissante si $\bar{x}^{(0)} \geqslant \bar{x}^{(1)}$:
Si la suite converge, elle converge fortement.

De plus, dans le cas $\bar{x}^{(0)} \leqslant \bar{x}^{(1)}$ par exemple, soit, s'il en existe, $\bar{x} \in T^N$ vérifiant :

$$\bar{x} = B \otimes \bar{x} \oplus c \;\wedge\; \bar{x}^{(0)} \leqslant \bar{x} \qquad \text{(IV. 11)}$$

(IV. 11) implique :

$$\bar{x}^{(1)} \leqslant \bar{x}$$

et par récurrence :

$$\bar{x}^{(n)} \leqslant \bar{x} \qquad \forall\, n$$

d'où la :

Proposition IV. 1.

Si $b_{ij} \geqslant 0 \quad \forall\, i, j$
si les opérateurs $\circledast_{ij}$ sont isotones et fixés $\forall\, i, j$,
si $\bar{x}^{(1)} \geqslant \bar{x}^{(0)}$ ou $\bar{x}^{(1)} \leqslant \bar{x}^{(0)}$

alors :
si la suite $\{ \bar{x}^{(n)} \}$ converge, elle converge fortement vers :

$$\min \{\bar{x} : \bar{x} = B \otimes \bar{x} \oplus c \wedge \bar{x} \geqslant \bar{x}^{(0)}\} \qquad \text{si } \bar{x}^{(1)} \geqslant \bar{x}^{(0)}$$

$$\max \{\bar{x} : \bar{x} = B \otimes \bar{x} \oplus c \wedge \bar{x} \leqslant \bar{x}^{(0)}\} \qquad \text{si } \bar{x}^{(1)} \leqslant \bar{x}^{(0)}$$

On remarquera en effet que si les hypothèses de la proposition IV. 1 sont vérifiées avec $\bar{x}^{(0)} \leqslant \bar{x}^{(1)}$ par exemple, et si la suite $\{\bar{x}^{(n)}\}$ est convergente, donc fortement convergente, vers une limite $\bar{\bar{x}}$, on peut affirmer :

$$\bar{\bar{x}} = \inf \{\bar{x} : \bar{x} = B \otimes \bar{x} \oplus c \wedge \bar{x} \geqslant \bar{x}^{(0)}\}$$

puisque :
Tout vecteur $\bar{x}$ défini par :
$\bar{x} = B \otimes \bar{x} \oplus c \wedge \bar{x} \geqslant \bar{x}^{(0)}$
satisfaisant, relativement aux itérés $\bar{x}^{(n)}$:
$\bar{x}^{(n)} \leqslant \bar{x} \qquad \forall\, n$
on a l'implication :

$$\bar{\bar{x}}_i \leqslant \bar{x}_i \qquad 1 \leqslant i \leqslant N$$

On peut d'ailleurs, étant donnés deux tableaux $\otimes$ et $\oplus$ définissant les opérations de multiplication et d'addition, caractériser l'ensemble des vecteurs $\bar{x}$ de T^N satisfaisant $\bar{x} = B \otimes \bar{x} \oplus c$; par exemple :

Proposition IV. 2.

Si B est une matrice à éléments non négatifs, de rayon spectral inférieur à 1 ,
si les opérations $\circledast_{ij}$ sont G-dirigées (resp. D-dirigées).
alors :
l'ensemble $E = \{\bar{x} : \bar{x} \in T^N_{\wedge} \quad \bar{x} = B \otimes \bar{x} \oplus c\}$ est un sup demi-treillis (resp. inf demi-treillis).

Exemple (grossier) : Soient $T = T^2_{10}$ et :

$$B = \begin{vmatrix} 0,99 & 0 \\ 0 & 0,99 \end{vmatrix} \qquad c = \begin{vmatrix} 0,011 \\ 0,011 \end{vmatrix}$$

Si $\circledast_{ij} = \circledast_{\nabla}$, alors E contient les vecteurs $\bar{x}$, de composantes $\bar{x}_1$ et $\bar{x}_2$ définies par :

$\bar{x}_1 \in T_{\wedge} \quad 0,11 \leqslant \bar{x}_1 \leqslant 1$

$\bar{x}_2 \in T_{\wedge} \quad 0,11 \leqslant \bar{x}_2 \leqslant 1$

La démonstration de la proposition IV.2 s'établit par ($\star_{ij} = \star_{\nabla}$) :
Si $\bar{x} = (\bar{x}_i)$ et $\bar{y} = (\bar{y}_i)$ sont 2 éléments de E ne vérifiant pas la relation d'ordre, soit $\bar{z}^{(0)} = (\bar{z}^{(0)}_i)$ défini par :

$$\bar{z}^{(0)}_i = \max\{\bar{x}_i, \bar{y}_i\} \qquad i = 1, \ldots, N.$$

et soit pour $n \geq 1$: $\bar{z}^{(n)} = B \otimes_{\nabla} \bar{z}^{(n-1)} \oplus_{\nabla} c$.

$\bar{z}^{(0)} \geq \bar{x} \wedge \bar{x} \in E \Longrightarrow \bar{z}^{(1)} \geq \bar{x}$ et $\bar{z}^{(1)} \geq \bar{z}^{(0)}$
$\bar{z}^{(0)} \geq \bar{y} \wedge \bar{y} \in E \Longrightarrow \bar{z}^{(1)} \geq \bar{y}$

La suite $\{\bar{z}^{(n)}\}$ est non décroissante.
Elle est, d'autre part, bornée par x car les opérations $\circledast_{ij}$ étant G-dirigées, on a : $\bar{z}^{(1)} \leq B\,\bar{z}^{(0)} + c$ et :

$\bar{z}^{(0)} \leq B\,\bar{z}^{(0)} + c \wedge \rho(B) < 1 \Longrightarrow \bar{z}^{(0)} \leq x$, ce qui implique $\bar{z}^{(n)} \leq x$, $\forall\, n$
(En remarquant que $\bar{z}^{(n)} \leq x \Longrightarrow \bar{z}^{(n+1)} \leq B\,\bar{z}^{(n)} + c \leq x$).
$\{\bar{z}^{(n)}\}$ converge fortement vers $\bar{t} = \min\{\bar{x} : \bar{x} \in E \wedge \bar{x} \geq \bar{z}^{(0)}\}$, qui est borne supérieure de la paire $\{\bar{x}, \bar{y}\}$.

CONSEQUENCES : Une conséquence des résultats précédents est :

Pour des opérations $\circledast$ G-dirigées (resp. D-dirigées) alors,
le choix du vecteur initial tel que :
$\bar{x}^{(0)} \geqslant x \wedge \bar{x}^{(0)} \geqslant \bar{x}^{(1)}$ (resp. $\bar{x}^{(0)} \leqslant x \wedge \bar{x}^{(0)} \leqslant \bar{x}^{(1)}$)
assure, lorsque la convergence a lieu, la précision maximum.

En effet, dans l'hypothèse $\circledast_{ij} = \circledast_{\nabla}$ $\forall i, j$ par exemple, tout élément $\bar{x}$ de T^N vérifiant $\bar{x} = B \otimes_{\nabla} \bar{x} \oplus_{\nabla} c$, satisfait la relation d'ordre :

$$\bar{x} \leqslant x$$

Pour le choix énoncé du vecteur initial, la convergence des itérés $\bar{x}^{(n)}$ a lieu vers :

$$\max \{ \bar{x} : \bar{x} = B \otimes_{\nabla} \bar{x} \oplus_{\nabla} c \}$$

L'effet des erreurs d'arrondi, pour l'arithmétique considérée, est minimisé.

Nous énoncerons enfin le corollaire suivant de la proposition IV. 1 :

Corollaire IV. 1.
Si la matrice B du système linéaire $x = B x + c$ est positive, et si le vecteur solution x est élément de T^N et vérifie $x = B \otimes_{\nabla} x \oplus_{\nabla} c$ (resp. $x = B \otimes_{\Delta} x \oplus_{\Delta} c$)

alors :
la suite $\{\bar{x}^{(n)}\}$ des itérés, calculés sur T par des opérations $\circledast$ G-dirigées (resp. D-dirigées), à partir d'un vecteur initial $\bar{x}^{(0)}$ vérifiant $\bar{x}^{(0)} \geqslant x \wedge \bar{x}^{(0)} \geqslant x^{(1)}$ (resp. $\bar{x}^{(0)} \leqslant x \wedge \bar{x}^{(0)} \leqslant \bar{x}^{(1)}$) converge fortement vers x .

EXEMPLES :

1) La résolution, par la méthode de Jacobi, du système linéaire $A\,x = b$ du § I. 5 , avec : $\circledast_{ij} = \circledast_{\nabla}$, $\bar{x}_i^{(0)} = 1{,}5 \quad \forall\, i$,

conduit, par convergence forte, à la solution exacte du système :

$x_i = 1 \quad \forall\, i$, en 160 itérations.

Soulignons l'effet d'accélération de convergence produit par les erreurs d'arrondi dès que $\bar{x}^{(1)} \leqslant \bar{x}^{(0)}$, $\circledast_{ij} = \circledast_{\nabla}$, puisque la résolution en double précision du système, à partir du même vecteur initial, détermine avec le test d'arrêt suivant :

$$| \bar{x}_i^{(n+1)} - \bar{x}_i^{(n)} | \leqslant 10^{-7} \quad \forall\, i$$

le vecteur de composantes $(\bar{x}_{DP})_i = 1{,}000001$ en 190 itérations.

2) Soit le problème $(\mathcal{P})$ aux dérivées partielles :

$$(\mathcal{P}) \begin{cases} -\Delta u + u = 0 & \text{dans le carré (C) : } 0 < x < 1 \text{ , } 0 < y < 1 \\ u = ye^{-x} & \text{sur les bords du carré} \end{cases}$$

où $u(x,y)$ est une fonction des 2 variables x et y

et soient, avec $h = \frac{1}{N+1}$,

la grille de points de coordonnées $x_i = ih$, $y_j = jh$ $(i,j = 0,1\ldots,N+1)$

et le problème $(\mathcal{P}')$ "approchant" $(\mathcal{P})$:

$$(\mathcal{P}') \begin{cases} u_{ij} = \dfrac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}}{4 + h^2} \\ u_{ij} = (ye^{-x})_{x = x_i \text{ , } y = y_j} \quad \text{sur les bords de (C)} \end{cases}$$

La résolution de $(\mathcal{P}')$ par la méthode de Gauss-Seidel a, pour $N = 10$, donné les résultats suivants où u_{DP} désigne la solution de $(\mathcal{P}')$ calculée en double précision, et u^1, u^2, u^3, u^4 désignent respectivement les solutions obtenues par convergence pour :

$\circledast = \circledast_{\nabla}$ $\quad \bar{u}^{(0)} < \bar{u}^{(1)}$

$\circledast = \circledast_{\nabla}$ $\quad \bar{u}^{(0)} > \bar{u}^{(1)}$

$\circledast = \circledast_{\Delta}$ $\quad \bar{u}^{(0)} < \bar{u}^{(1)}$

$\circledast = \circledast_{\Delta}$ $\quad \bar{u}^{(0)} > \bar{u}^{(1)}$

		u^1	u^2	u_{DP}
i=5 (extraits de tables)	j=1	0.5770592E-01	0.5770653E-01	0.5770702D-01
	=2	0.1154117E 00	0.1154129E 00	0.1154139D 00
	=3	0.1731173E 00	0.1731190E 00	0.1731206D 00
	=4	0.2308224E 00	0.2308246E 00	0.2308269D 00
	=5	0.2885270E 00	0.2885296E 00	0.2885327D 00
	=6	0.3462309E 00	0.3462340E 00	0.3462376D 00
	=7	0.4039344E 00	0.4039375E 00	0.4039415D 00
	=8	0.4616367E 00	0.4616398E 00	0.4616438D 00
	=9	0.5193383E 00	0.5193407E 00	0.5193443D 00
	=10	0.5770385E 00	0.5770399E 00	0.5770420D 00

u^3	u^4	ye^{-x}
0.5770778E-01	0.5770835E-01	0.5770330E-01
0.1154155E 00	0.1154166E 00	0.1154066E 00
0.1731229E 00	0.1731244E 00	0.1731099E 00
0.2308300E 00	0.2308320E 00	0.2308131E 00
0.2885365E 00	0.2885389E 00	0.2885165E 00
0.3462422E 00	0.3462446E 00	0.3462198E 00
0.4039464E 00	0.4039490E 00	0.4039230E 00
0.4616485E 00	0.4616511E 00	0.4616264E 00
0.5193481E 00	0.5193505E 00	0.5193297E 00
0.5770445E 00	0.5770457E 00	0.5770330E 00

On remarquera que dans le cas étudié, le vecteur u^1, bien qu'entaché d'une plus forte erreur que u^2 par rapport à u_{DP}, est une meilleure approximation de la solution exacte u de $(\mathcal{P})$:

L'erreur commise en substituant $(\mathcal{P}')$ à $(\mathcal{P})$ et l'erreur dûe aux calculs, se compensent en effet davantage dans le 1er cas, la relation $u \leqslant u_{DP}$ résultant du signe de $u^{(4)}_{x^4}(x, y)$ dans le carré unité.

Lorsque la matrice B de l'itération linéaire est négative :

Le choix d'un vecteur initial $x^{(0)}$ tel que $x^{(0)} \leqslant x^{(2)}$ (resp. $x^{(0)} \geqslant x^{(2)}$) implique :

$x \geqslant x^{(0)}$

$\{ x^{(2k)} \}$ monotone non décroissante

$\{ x^{(2k+1)} \}$ monotone non croissante

Le calcul de la suite $\{ \bar{x}^{(n)} \}$:
par des opérations D-dirigées pour les itérés d'indice pair
par des opérations G-dirigées pour les itérés d'indice impair
et l'arrêt des calculs dès que la relation d'ordre entre les éléments des 2 suites $\{ \bar{x}^{(2k)} \}$ et $\{ \bar{x}^{(2k+1)} \}$ n'est plus vérifiée, permet :

- l'accélération de la convergence
- l'obtention d'une bonne précision.

III - RESOLUTION ITERATIVE EN ARITHMETIQUE A VIRGULE FLOTTANTE DE L'EQUATION LINEAIRE $x = ax + c$

Nous supposons a et c éléments de T_b^s dans l'équation :

$$x = ax + c \qquad \text{(IV. 13)}$$

L'étude de cette équation, inintéressante du point de vue pratique, permet de souligner l'influence des erreurs d'arrondi suivant les valeurs de a et d'établir des comparaisons entre différentes arithmétiques, relativement à la résolution numérique de (IV. 13) par itérations.
Nous énoncerons trois théorèmes :

Proposition IV. 3.

Si les opérations $\otimes$ et $\oplus$ sont correctes, alors :

si $-1 < a \leqslant 0$ $\quad$ Card $\{\bar{x} \mid \bar{x} = a \otimes \bar{x} \oplus c\} \leqslant 2$

si $0 < a \leqslant \frac{1}{2}$ $\quad$ Card $\{\bar{x} \mid \bar{x} = a \otimes \bar{x} \oplus c\} \leqslant 4$, sauf peut-être si la mantisse de c est de valeur absolue supérieure à $1 - 2b^{1-s}$ (condition C_m).

En effet, si les opérations $\otimes$ et $\oplus$ sont correctes, et si $\bar{x}$ est un élément de T vérifiant $\bar{x} = a \otimes \bar{x} \oplus c$, alors :

$$-1 < a \leqslant \frac{1}{2} \Longrightarrow |x - \bar{x}| < \frac{1}{1-a}\varepsilon(\bar{x}) \qquad \text{(IV. 14)}$$, sauf peut-être si C_m a lieu

$$\frac{1}{2} < a < 1 \Longrightarrow |x - \bar{x}| < \frac{2}{1-a}\varepsilon(\bar{x}) \qquad \text{(IV. 15)}$$

(IV. 15) vient immédiatement quel que soit a satisfaisant $|a| < 1$ par l'écriture :

$$a\bar{x} = a \otimes \bar{x} + e_1 \quad \wedge \quad |e_1| < \varepsilon(a\bar{x}) \leqslant \varepsilon(\bar{x})$$

$$a \otimes \bar{x} + c = \bar{x} + e_2 \quad \wedge \quad |e_2| < \varepsilon(\bar{x})$$

entraînant avec $e = e_1 + e_2$:

$$|x - \bar{x}| = \frac{|e|}{1-a} < \frac{2}{1-a}\ \varepsilon(\bar{x})$$

Ecrivons :

$$|a\bar{x}| \leqslant \frac{|a|}{1-a}\left[|c| + |e|\right]$$

Si $-1 < a \leqslant 0$, on aura :

$$\frac{|a|}{1-a} \leqslant \frac{1}{2}$$

$$|x| \leqslant |c|$$

$$|a\bar{x}| < \frac{1}{2}\left[|c| + 2b\ \varepsilon(c)\right] < |c|$$

Si $0 < a \leqslant \frac{1}{2}$, on aura :

$$|a\bar{x}| < |c| + (2b - 1)\ \varepsilon(c)$$

assurant

$\varepsilon(a \otimes \bar{x}) \leqslant \varepsilon(c)$ si C_m n'a pas lieu.

Les 2 cas ci-dessus permettant d'écrire :

$\varepsilon(a \otimes \bar{x}) \leqslant \varepsilon(c)$

on restreindra la borne de e_2 par :

$|e_2| \leqslant \varepsilon(\bar{x}) - \varepsilon(a \otimes \bar{x})$

et

$|e| < \varepsilon(\bar{x})$

implique (IV. 14)

La proposition IV. 3. se déduit de (IV. 14) par :

$-1 < a \leqslant 0 \Longrightarrow \bar{x} = \nabla x \vee \Delta x$

$0 < a \leqslant \frac{1}{2} \Longrightarrow \bar{x} = (\nabla x)' \vee \nabla x \vee \Delta x \vee (\Delta x)''$ (sauf peut-être cas C_m :

ainsi pour $x = 0{,}50x + 0{,}92$ et $T = T_{10}^2$, $\{\bar{x}\} = \{1{,}7;\ 1{,}8;\ 1{,}9;\ 2{,}0;\ 2{,}1\}$)

Remarque :

Le résultat $\bar{x} = \nabla x \vee \Delta x$ lorsque $-1 < a \leqslant 0$, n'implique évidemment pas que ∇x (resp. Δx) vérifie $\nabla x = a \otimes \nabla x \oplus c$ (resp. $\Delta x = a \otimes \Delta x \oplus c$) pour l'un au moins des choix "corrects" d'opérations : $(\otimes_\nabla , \oplus_\nabla)$, $(\otimes_\nabla , \oplus_\Delta)$, $(\otimes_\Delta , \oplus_\nabla)$, $(\otimes_\Delta , \oplus_\Delta)$, ni que la suite des itérés $\{ \bar{x}^{(n)} \}$ converge -fortement ou faiblement- vers des valeurs satisfaisantes.

Ainsi pour $x = -0{,}99x + 0{,}43$ et $T = T^2_{10}$:

$\nabla x = 0{,}21$ ne vérifie pas $\nabla x = a \otimes \nabla x \oplus c$

La suite $\{\bar{x}^{(n)}\}$, obtenue à partir de $\bar{x}^{(0)} = 0{,}21$ par $\bar{x}^{(n+1)} = a \otimes_\Delta \bar{x}^{(n)} \oplus_\Delta c$ si n est pair, $\bar{x}^{(n+1)} = a \otimes_\nabla \bar{x}^{(n)} \oplus_\nabla c$ si n est impair, converge faiblement en cyclant sur 0 et $0{,}43$.

On a cependant :

Proposition IV. 4.

Si $-1 < a < 0$, alors quel que soit $\bar{x}^{(0)} \neq x$, la suite $\{ \bar{x}^{(n)} \}$ des itérés définis par :

$\bar{x}^{(n+1)} = a \otimes_\nabla \bar{x}^{(n)} \oplus_\nabla c$ si n pair quand $\bar{x}^{(0)} < x$,
si n impair quand $\bar{x}^{(0)} > x$

$\bar{x}^{(n+1)} = a \otimes_\Delta \bar{x}^{(n)} \oplus_\Delta c$ si n impair quand $\bar{x}^{(0)} < x$,
si n pair quand $\bar{x}^{(0)} > x$,

détermine avec le test d'arrêt suivant :

$\bar{x}^{(P+1)} - \bar{x}^{(P)} \leqslant 0$ si P pair $\wedge\ \bar{x}^{(0)} < x$ ou P impair $\wedge\ \bar{x}^{(0)} > x$

$\bar{x}^{(P+1)} - \bar{x}^{(P)} \geqslant 0$ si P impair $\wedge\ \bar{x}^{(0)} < x$ ou P pair $\wedge\ \bar{x}^{(0)} > x$

l'un des deux nombres ∇x ou Δx (égal à $\bar{x}^{(P)}$).

Nous supposons, afin de ne pas alourdir la démonstration, que les éléments de T, au voisinage de x, vérifient une condition de régularité relative à leur espacement, c'est-à-dire que nous supposons que la mantisse $m(x)$ de x vérifie :

$$\frac{1}{b} + b^{1-s} \leqslant |m(x)| < 1 - b^{1-s} \quad \text{(condition } c_x\text{)}$$

A chaque itéré $\bar{x}^{(n+1)}$, associons l'itéré $x^{(n+1)}$, calculé sur $\mathbb{R}$ à partir de $\bar{x}^{(n)}$, par :

$$x^{(n+1)} = a\,\bar{x}^{(n)} + c$$

la disposition de $x^{(n+1)}$ par rapport à x, à partir de la position de $\bar{x}^{(n)}$ par rapport à x est :

$\bar{x}^{(n)} < x \Longrightarrow x^{(n+1)} > x$

$\bar{x}^{(n)} > x \Longrightarrow x^{(n+1)} < x$

avec de plus par $|a| < 1$:

$| x^{(n+1)} - x | < | \bar{x}^{(n)} - x |$

Etudions l'écart entre $x^{(n+1)}$ et $\bar{x}^{(n+1)}$, avec les notations :

$a\,\bar{x}^{(n)} = a \otimes \bar{x}^{(n)} + e_1$

$a \otimes \bar{x}^{(n)} + c = \bar{x}^{(n+1)} + e_2$

$e = e_1 + e_2$

$x^{(n+1)} = \bar{x}^{(n+1)} + e$

Plusieurs cas peuvent se présenter :

i) $a\,\bar{x}^{(n)}$ et c sont de même signe

lorsque $\varepsilon(a \otimes \bar{x}^{(n)}) \leqslant \varepsilon(c)$ alors $|e| < \varepsilon(\bar{x}^{(n+1)})$

lorsque $\varepsilon(a \otimes \bar{x}^{(n)}) > \varepsilon(c)$ alors $|e| < 2\varepsilon(\bar{x}^{(n+1)})$, ce cas impliquant $\varepsilon(\bar{x}^{(n+1)}) > \varepsilon(x)$ par : $\varepsilon(c) \geqslant \varepsilon(x)$.

ii) a $\bar{x}^{(n)}$ et c sont de signes contraires et une cancellation ne se produit pas lors de la formation de $\bar{x}^{(n+1)}$

$|a \otimes \bar{x}^{(n)}| > |c|$ implique que $x^{(n+1)}$ et $\bar{x}^{(n+1)}$ n'ont pas le signe de x .

$|a \otimes \bar{x}^{(n)}| \leqslant |c|$ assure $|e| < \varepsilon\ (\bar{x}^{(n+1)})$

iii) a $\bar{x}^{(n)}$ et c sont de signes contraires et une cancellation se produit lors de la formation de $\bar{x}^{(n+1)}$

$|a \otimes \bar{x}^{(n)}| > |c|$ entraîne la même implication que ci-dessus

$|a \otimes \bar{x}^{(n)}| \leqslant |c|$ assure $|e| < \varepsilon\ (c) \leqslant b\ \varepsilon\ (x)$ mais :

lorsque $a < -1+\frac{1}{b}$, la condition c_x entraîne $|x^{(n+1)}| < x - b\ \varepsilon(x)$

lorsque $a \geqslant -1+\frac{1}{b}$, de $|c| < (2-\frac{1}{b})\ |x|$ et $|a \otimes \bar{x}^{(n)}| > |x|$ on tire

$|e - a \otimes \bar{x}^{(n)}| \leqslant (1 - \frac{1}{b})\ |x|$

En conclusion, l'erreur e est inférieure à $\varepsilon(\bar{x}^{(n+1)})$ dès que $x^{(n+1)}$ est "suffisament près" de x .

Il résulte de l'étude précédente que :

$\bar{x}^{(n+1)}$ a la même disposition par rapport à x que $x^{(n+1)}$

$| \bar{x}^{(n+1)} - x | < | \bar{x}^{(n)} - x |$

tant que $x^{(n+1)}$ vérifie :

$x^{(n+1)} \leqslant \nabla x \vee x^{(n+1)} \geqslant \Delta x$

Si $x \in T$, $\bar{x}^{(n)}$ converge fortement vers x .

Si $x \notin T$, avec $\bar{x}^{(0)} < x$ par exemple, il existe un indice $P = 2Q$ (ou $P = 2Q + 1$) satisfaisant : $\nabla x < x^{(2Q)} < x$ (ou $x < x^{(2Q+1)} < \Delta x$) .

Considèrons le 1^{er} cas : $\bar{x}^{(2Q)} = \Delta x$

le test d'arrêt est alors réalisé :

au rang P si $\bar{x}^{(P-1)} = \Delta x$

au rang P + 1 sinon.

Remarques :

1) la suite $\{ \bar{x}^{(n)} \}$ converge plus rapidement que la suite $\{ x^{(n)} \}$.

2) L'étude de l'erreur e établit que l'élément $\bar{x}$ de T , le plus proche de x -ou les deux s'il y a équidistance- vérifie certainement $\bar{x} = a \otimes \bar{x} \oplus c$ pour $\circledast = \circledast_\Delta$ si $\bar{x} = \Delta x$, $\circledast_\nabla$ si $\bar{x} = \nabla x$.

Nous donnerons aussi à titre d'exemple, la :

Proposition IV. 5.

Si $a = 1 - b^{-s}$, alors

1) si les opérations $\otimes$ et $\oplus$ sont correctes et I-dirigées,

$\forall \bar{x} \in T_b^s \cap]b^{e(x)-2}, b^{e(x)-1}] \times$ signe(x) si le 1er chiffre de mantisse de x est 1, pour $\bar{x} = b^{e(x)-1} \times$ signe(x) sinon, on a:

$\bar{x} = a \otimes \bar{x} \oplus c$

2) si les opérations $\otimes$ et $\oplus$ sont correctes et E-dirigées, $\forall \bar{x}^{(0)} \in [0, b^{e(x)}]$ la suite $\{ \bar{x}^{(n)} \}$ converge vers $b^{e(x)} \times$ signe(x). (voir r) $\times$ signe (x)

$x = b^s c$

Nous considérons $c > 0$.

1) s'établit par :

$a \otimes_\nabla \bar{x} = (\bar{x})'$

$(\bar{x})' \oplus_\nabla c = \bar{x}$ dans les conditions indiquées .

2) s'établit par :

$a \otimes_\Delta y = y$ sauf si $a \otimes_\Delta y = a\, y$

(r): Si la condition (C) du ch.II n'est pas vérifiée par l'opération $\oplus$ si mantisse$(x) < \frac{1}{b} + \frac{1}{b^2}$

il convient de modifier légérement la proposition.

IV - CONCLUSION

Il est certain que la possibilité de disposer, sur un calculateur, des 2 opérations approchées $\circledast_{\nabla}$ et $\circledast_{\Delta}$ de chacune des opérations élémentaires $*$ définies sur $\mathbb{R}$, est un élément important pour le contrôle des erreurs d'arrondi : Ce fait est particulièrement vrai dans le cas de méthodes itératives de résolution, les opérateurs étant fréquemment monotones.

D'un point de vue différent, nous mentionnerons dans ce chapitre, l'étude effectuée par M. TIENARI [37] pour le contrôle de la longueur de mantisse, dans les méthodes itératives en arithmétique à virgule flottante.

CHAPITRE V

STABILITE NUMERIQUE DE DIFFERENTS SCHEMAS DE SOMMATION.

CALCUL DE SOMMES DE SERIES

INTRODUCTION

Les problèmes étudiés dans ce chapitre, relatifs à l'algorithme de la sommation, sont des problèmes de "stratégie" :

Les opérations $\circledast$ définies sur les sous-ensembles finis de $\mathbb{R}$ ne sont pas généralement associatives. La non-associativité de l'opération $\oplus$ pose, pour la sommation, d'éléments donnés, par l'arithmétique approchée, le problème de la disposition des opérations en vue de l'obtention de la meilleure précision.
Les résultats qui suivent sont des théorèmes de comparaison de précision entre les algorithmes les plus usuels réalisant la sommation de N nombres par $N - 1$ additions (coût minimum).

Les théorèmes sont établis pour des sous-ensembles de $\mathbb{R}$ de type flottant, ou des systèmes de nombres à virgule flottante.

Une application particulière des résultats obtenus, est leur utilisation au calcul de sommes de séries numériques.

I - ETUDE COMPARATIVE DE SCHEMAS DE SOMMATION [34]

I. 1. PREMIER THEOREME

Une conséquence directe du lemme I. 2 du chapitre I , relatif aux sous-ensembles de $\mathbb{R}$ de type flottant, est la proposition suivante :

Proposition V. 1.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant,
si l'addition sur S , notée $\oplus$, est correcte,
si x_1 , x_2 , ... , x_N sont N éléments de S^+ vérifiant les inégalités :

$$\varepsilon \left(\bigoplus_{i=1}^{p} x_i \right) \leqslant \varepsilon (x_{p+1}) \qquad p = 2,\ldots,N-1 \qquad (V.\ 1)$$

alors, si l'on désigne par :

$\sigma = \sum_{i=1}^{N} x_i$ la somme des x_i sur $\mathbb{R}$

$\bar{\sigma}_s = \bigoplus \sum_{i=1}^{N} x_i$ la somme des x_i sur S par $\oplus$ dans l'ordre séquentiel,

il vient :

$$\bar{\sigma}_s = \nabla \sigma \vee \Delta \sigma$$

Désignant par e_i ($i = 1, \ldots , N - 1$) l'erreur élémentaire commise lors de la ième addition dans le calcul de $\bar{\sigma}_s$, on peut affirmer par le lemme I. 2 (chapitre I) :

$$|e_i| \leqslant \varepsilon \Big(\bigoplus \sum_{j=1}^{i+1} x_j \Big) - \varepsilon \Big(\bigoplus \sum_{j=1}^{i} x_j \Big) \qquad i=1,\ldots,N-1$$

Par la majoration :

$$| \sigma - \bar{\sigma}_s | \leqslant \sum_{i=1}^{N-1} | e_i | < \varepsilon (\bar{\sigma}_s)$$

on déduit alors la proposition.

- Pour conclure à l'égalité $\bar{\sigma}_s = \nabla \sigma \vee \Delta \sigma$ dans le cas particulier où $\bar{\sigma}_s$ vérifie $\varepsilon ((\bar{\sigma}_s)') \neq \varepsilon (\bar{\sigma}_s)$, on remarquera que :

a) si $\Big(\bigoplus \sum_{i=1}^{N-1} x_i \Big) + x_N < \bar{\sigma}_s$ alors :

$$|\sigma - \bar{\sigma}_s| < \varepsilon \Big(\Big(\bigoplus \sum_{i=1}^{N-1} x_i \Big) + x_N \Big) \leqslant \varepsilon ((\bar{\sigma}_s)')$$

b) si $\Big(\bigoplus \sum_{i=1}^{N-1} x_i \Big) + x_N \geqslant \bar{\sigma}_s$ alors $e_{N-1} \geqslant 0$.

$$\sigma \geqslant \bar{\sigma}_s \implies \bar{\sigma}_s = \nabla \sigma$$

$$\sigma < \bar{\sigma}_s \implies \bar{\sigma}_s - \sigma < \sum_{i=1}^{N-2} e_i < \varepsilon \Big(\bigoplus \sum_{i=1}^{N-1} x_i \Big) \leqslant \varepsilon ((\bar{\sigma}_s)')$$

Remarque

Dans les hypothèses suivantes :
S est un sous-ensemble de $\mathbb{R}$ de type flottant,
l'addition sur S , notée $\oplus$, est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ) ,

$x_1, x_2, \ldots, x_N$ sont N éléments de S^+ vérifiant (V. 1),

si σ et $\overline{\sigma}_s$ sont définies comme précédemment et si l'on note $\overline{\sigma}$ une somme quelconque des x_i sur S par $\oplus$, $\overline{\sigma}_s$ est une meilleure approximation de σ dans l'ensemble $\sum = \{ \overline{\sigma} \}$:

$$| \sigma - \overline{\sigma}_s | = \underset{\overline{\sigma} \in \sum}{\text{Min}} | \sigma - \overline{\sigma} |$$

L'ordre séquentiel croissant est optimal pour la sommation des suites $\{x_i\}$ à croissance rapide (vérifiant V. 1) .

Exemple : Sommation par $\oplus_\nabla$ des nombres suivants écrits en base 10 avec 2 chiffres de mantisse : 0,030 ; 0,10 ; 0,20 ; 0,30 ; 0,67 ; 1,0 ; 1,7 ; 14.
$\sigma = 18$
Somme calculée par sommation dans l'ordre séquentiel croissant : 18
Somme calculée par sommation dans l'ordre séquentiel décroissant : 16
Somme calculée par sommation suivant un schéma dichotomique : 16

Le corollaire suivant de la proposition V. 1, relatif à la sommation de trois nombres, sera utile pour la suite :

Corollaire V. 1.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant, pour tout triplet $(x, y, z) \in (S^+)^3$ et vérifiant l'inégalité :

$$\varepsilon (x + y) \leqslant \varepsilon (z) \qquad \text{(V. 2)}$$

on a :

$$(x \oplus_\nabla y) \oplus_\nabla z \geqslant \max \{ (x \oplus_\nabla z) \oplus_\nabla y \ , \ (y \oplus_\nabla z) \oplus_\nabla x \}$$

$$(x \oplus_\Delta y) \oplus_\Delta z \leqslant \min \{ (x \oplus_\Delta z) \oplus_\Delta y \ , \ (y \oplus_\Delta z) \oplus_\Delta x \}$$

I. 2. THEOREMES GENERAUX DE COMPARAISON DES SOMMES SEQUENTIELLE ET DICHOTOMIQUE

I. 2. 1. Sommation d'éléments à valeurs positives et décroissantes

Lorsque les éléments à sommer sont classés dans l'ordre décroissant, on peut énoncer (l'écriture ci-dessous supposant que N est une puissance de 2 : $N = 2^k$, afin de ne pas compliquer inutilement les notations) :

Proposition V. 2.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant,
si l'addition sur S , notée $\oplus$, est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ),
si $x_1, x_2, \ldots, x_N$ sont N éléments de S^+ rangés dans l'ordre décroissant,

alors, les sommes $\sigma = \sum_{i=1}^{N} x_i$, $\bar{\sigma}_s = \oplus \sum_{i=1}^{N} x_i$ et $\bar{\sigma}_d = \sigma_{1k}$ où

σ_{1k} est défini par la récurrence suivante :

$$\sigma_{io} = x_i \qquad i = 1, \ldots, N$$

$$\sigma_{ij} = \sigma_{2i-1,\, j-1} \oplus \sigma_{2i,\, j-1} \qquad i = 1, \ldots, N/2^j$$

$$j = 1, \ldots, k$$

vérifient :

$$|\sigma - \bar{\sigma}_d| \leqslant |\sigma - \bar{\sigma}_s|$$

Démonstration :

L'algorithme de sommation définissant $\bar{\sigma}_d$ est le schéma de sommation dichotomique .

Etablissons la proposition V. 2 dans le cas où $\oplus = \oplus_\nabla$

Il suffit de montrer $\bar{\sigma}_d \geqslant \bar{\sigma}_s$.

De l'inégalité :

$$\sigma_{2i-1,\,o} + \sigma_{2i,\,o} \geqslant \sigma_{2i+1,\,o} + \sigma_{2i+2,\,o}$$

on déduit par la monotonie de l'arrondi

$$\sigma_{i,\,1} \geqslant \sigma_{i+1,\,1} \qquad 1 \leqslant i < N/2$$

et pour tout j :

$$\sigma_{i,\,j} \geqslant \sigma_{i+1,\,j} \qquad 1 \leqslant i < N/2^j$$

Montrons alors l'inégalité suivante :

$$\oplus_\nabla \sum_{i=1}^{N/2^{j+1}} \sigma_{i,\,j+1} \geqslant \oplus_\nabla \sum_{i=1}^{N/2^j} \sigma_{i,\,j} \qquad \forall\, j \qquad (V.\,3)$$

Ecrivons :

$$\oplus_\nabla \sum_i \sigma_{i,\,j+1} = \sigma_{1,\,j+1} \oplus_\nabla (\sigma_{3,j} \oplus_\nabla \sigma_{4,j}) \oplus_\nabla (\sigma_{5,j} \oplus_\nabla \sigma_{6,j}) \oplus_\nabla \cdots$$

$$\oplus_\nabla (\sigma_{N/2^j-1,\,j} \oplus_\nabla \sigma_{N/2^j,j})$$

En remarquant :

$$\varepsilon(\sigma_{3,j} + \sigma_{4,j}) \leqslant \varepsilon(\sigma_{1,\,j+1})$$

il vient, par le corollaire V. 1,

$$\sigma_{1,\,j+1} \oplus_\nabla (\sigma_{3,j} \oplus_\nabla \sigma_{4,j}) \geqslant \sigma_{1,j} \oplus_\nabla \sigma_{2,j} \oplus_\nabla \sigma_{3,j} \oplus_\nabla \sigma_{4,j} = A$$

puis

$$A \oplus_\nabla (\sigma_{5,j} \oplus_\nabla \sigma_{6,j}) \geqslant A \oplus_\nabla \sigma_{5,j} \oplus_\nabla \sigma_{6,j}$$

et ainsi de suite, ce qui entraîne, en tenant compte de la propriété de monotonie de l'arrondi, l'inégalité (V. 3).

Par l'application répétée de l'inégalité (V. 3), on déduit :

$$\sigma_{1k} \geqslant \oplus_{\nabla} \sum_{i} \sigma_{i,j} \qquad \forall\, j$$

et en particulier :

$$\overline{\sigma}_d \geqslant \overline{\sigma}_s$$

Remarques :

Lorsque N n'est pas une puissance de 2, pour certains j, le nombre d'éléments de la couche j est du type $2\lambda - 1$, impair, définissant $\sigma_{\lambda,\, j+1} = \sigma_{2\lambda - 1,\, j}$.

Les hypothèses de la proposition V. 2 ne sont pas suffisament fortes pour assurer l'optimalité d'un schéma de sommation particulier. A titre d'exemple, la sommation des nombres décimaux, de 2 chiffres de mantisse :

9,4 8,5 7,5 6,3 5,8 4,7 4,5 4,4

donne les résultats suivants :

$\sigma = 51,1$

si $\oplus = \oplus_{\nabla}$:

$$x_1 \oplus (x_2 \oplus x_3) \oplus (x_4 \oplus x_5) \oplus (x_6 \oplus x_7) \oplus x_8 = 50$$

$$\overline{\sigma}_d = 48 \qquad \overline{\sigma}_s = 47 \qquad \oplus \sum_{i=N}^{1} x_i = 48$$

si $\oplus = \oplus_{\Delta}$:

$$\overline{\sigma}_d = 52 \qquad \overline{\sigma}_s = 54 \qquad \oplus \sum_{i=N}^{1} x_i = 54$$

I. 2. 2. Sommation d'éléments à valeurs positives quelconques

Etablissons le lemme préliminaire suivant :

Lemme V. 1.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant, pour tout triplet $(x, y, z) \in (S^+)^3$, les inégalités suivantes sont vérifiées :

$$(x)' \oplus_{\nabla} (y \oplus_{\nabla} z) \geqslant ((x \oplus_{\nabla} y) \oplus_{\nabla} z)' \qquad (V. 4)$$

$$(x)'' \oplus_{\Delta} (y \oplus_{\Delta} z) \leqslant ((x \oplus_{\Delta} y) \oplus_{\Delta} z)'' \qquad (V. 4')$$

Démonstration :

Démontrons (V. 4) .

Il suffit d'établir $(x)' \oplus_{\nabla} (y \oplus_{\nabla} z) \geqslant (\nabla (x + y + z))'$.

L'égalité :

$$x + y + z - \left[(x)' \oplus_{\nabla} (y \oplus_{\nabla} z) \right] = \varepsilon ((x)') + e_1 + e_2$$

où :

$$e_1 = y + z - (y \oplus_{\nabla} z)$$

$$e_2 = (x)' + (y \oplus_{\nabla} z) - \left[(x)' \oplus_{\nabla} (y \oplus_{\nabla} z) \right]$$

permet, avec les inégalités déduites du lemme I. 2 :

$$| e_1 | < \varepsilon (y + z)$$

$$| e_2 | < \varepsilon \left[(x)' + (y \oplus_{\nabla} z) \right] - \varepsilon (y \oplus_{\nabla} z) \qquad \text{si } \varepsilon(y \oplus_{\nabla} z) \leqslant \varepsilon((x)')$$

$$| e_2 | < \varepsilon \left[(x)' + (y \oplus_{\nabla} z) \right] - \varepsilon ((x)') \qquad \text{si } \varepsilon(y \oplus_{\nabla} z) > \varepsilon((x)')$$

d'affirmer dans les 2 cas, avec la notation $\tilde{\sigma} = (x)' \oplus_{\nabla} (y \oplus_{\nabla} z)$

$$(x + y + z) - \tilde{\sigma} < 2\,\varepsilon\,(\tilde{\sigma})$$

donc :

$$\tilde{\sigma} \geqslant (\nabla (x + y + z))'$$

Il vient alors :

Proposition V. 3.

Si S est un sous-ensemble de $\mathbb{R}$ de type flottant,
si l'addition sur S, notée $\oplus$, est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ),
alors, étant donnés N éléments x_i $(1 \leqslant i \leqslant N)$ de S^+, les deux sommes définies sur S par :

$$\bar{\sigma}_s = \bigoplus \sum_{i=1}^{N} x_i$$

$$\bar{\bar{\sigma}} = \begin{cases} \bigoplus \sum_{j=1}^{N/2} (x_{2j-1} \oplus x_{2j}) & \text{si } N \text{ pair} \\ \left\{ \bigoplus \sum_{j=1}^{\lfloor N/2 \rfloor} (x_{2j-1} \oplus x_{2j}) \right\} \oplus x_N & \text{si } N \text{ impair} \end{cases}$$

vérifient :

$$\bar{\bar{\sigma}} \geqslant (\bar{\sigma}_s)' \qquad \text{si } \oplus = \oplus_{\nabla}$$

$$\bar{\bar{\sigma}} \leqslant (\bar{\sigma}_s)'' \qquad \text{si } \oplus = \oplus_{\Delta}$$

Démonstration :

La démonstration est immédiate par récurrence.

Prenons le cas où $\oplus = \oplus_\nabla$ et posons :

$$\bar{\sigma}_s^{2i} = \oplus_\nabla \sum_{j=1}^{2i} x_j$$

$$\bar{\bar{\sigma}}^{2i} = \oplus_\nabla \sum_{j=1}^{i} (x_{2j-1} \oplus_\nabla x_{2j})$$

- Pour $i = 2$, la propriété $\bar{\bar{\sigma}}^{2i} \geqslant (\bar{\sigma}_s^{2i})'$ est bien vraie.

- Pour i fixé $\geqslant 2$, supposons $\bar{\bar{\sigma}}^{2i} \geqslant (\bar{\sigma}_s^{2i})'$

$$\bar{\bar{\sigma}}^{2i+2} = \bar{\bar{\sigma}}^{2i} \oplus_\nabla (x_{2i+1} \oplus_\nabla x_{2i+2})$$

$$\bar{\sigma}_s^{2i+2} = \bar{\sigma}_s^{2i} \oplus_\nabla x_{2i+1} \oplus_\nabla x_{2i+2}$$

Si $\bar{\bar{\sigma}}^{2i} \geqslant \bar{\sigma}_s^{2i}$, on a évidemment, S étant à intervalles non décroissants et l'arrondi monotone, $\bar{\bar{\sigma}}^{2i+2} \geqslant (\bar{\sigma}_s^{2i+2})'$.

Si $\bar{\bar{\sigma}}^{2i} = (\bar{\sigma}_s^{2i})'$, la propriété résulte du lemme précédent.

I. 3. THEOREMES RELATIFS AUX SYSTEMES DE NOMBRES A VIRGULE FLOTTANTE

Ce paragraphe établit des théorèmes plus particuliers de comparaison d'algorithmes de sommation, pour des nombres x_i, écrits en virgule flottante et vérifiant des hypothèses assez restrictives.

Relativement aux systèmes de nombres à virgule flottante de base 2, il vient :

Proposition V. 4

N éléments x_i $(1 \leqslant i \leqslant N)$ de $(T_2^s)^+$ étant donnés, si l'addition sur T_2^s est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ), si les N éléments x_i ont même exposant

alors, les sommes $\bar{\sigma}_s$ et $\bar{\sigma}_d$ calculées respectivement sur T_2^s par les schémas de sommation séquentiel et dichotomique, vérifient l'inégalité :

$$| \sigma - \bar{\sigma}_d | \leqslant | \sigma - \bar{\sigma}_s |$$

La proposition précédente se démontre de façon analogue à la proposition V. 2 :
Si $\oplus = \oplus_\nabla$, on établira l'inégalité $\bar{\sigma}_d \geqslant \bar{\sigma}_s$, en remarquant que si p désigne l'exposant commun des x_i,

$$e(x_i) = p \qquad i = 1, \ldots, N$$

par les hypothèses, les éléments σ_{ij} ont, pour j fixé, tous le même exposant :

$$e(\sigma_{ij}) = p + j \qquad j = 0, \ldots, k$$

les inégalités du type :

$$\varepsilon(\sigma_{3,j} + \sigma_{4,j}) \leqslant \varepsilon(\sigma_{1,j+1})$$

sont alors assurées.

Pour un système de nombres à virgule flottante de base b quelconque, la démonstration de la proposition V. 4 reste valable, si l'on remplace la sommation dichotomique par la sommation b à b d'éléments de T_b^s ayant même exposant.
(Le schéma de sommation b à b se définit pour $N = b^k$ par la récurrence :

$$\sigma_{io} = x_i \qquad i = 1, \ldots, N$$

$$\sigma_{ij} = \bigoplus \sum_{l=b(i-1)+1}^{bi} \sigma_{l,j-1} \qquad \begin{matrix} i = 1, \ldots, b^{k-j} \\ j = 1, \ldots, k \end{matrix} \quad)$$

Proposition V. 4'.

Pour un système de nombres flottants de base b, lorsque l'addition flottante est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ), la sommation b à b de flottants de même signe et de même exposant, calcule une meilleure approximation de la somme de ces nombres sur $\mathbb{R}$ que la sommation séquentielle.

En pratique, cette méthode de sommation pour une base $b \neq 2$, n'est pas usitée.

Si, dans le système T_b^s, $b \neq 2$, on compose la sommation par le schéma dichotomique à la sommation b à b, de flottants de même exposant, la première sommation se comporte moins "régulièrement", par rapport à la sommation séquentielle, que la seconde ; toutefois, dans la plupart des cas, le schéma dichotomique est meilleur que la sommation b à b.

Un cas d'application courante (surtout lorsque p, ci-dessous, est nul) où il est possible de classer ces deux procédés, indépendamment des x_i, est donné par la :

Proposition V. 5.

Si l'addition sur $T = T_b^s$ $(b > 3)$ est induite de l'addition sur $\mathbb{R}$ par ∇ (resp. Δ), alors, étant donnés les éléments x_i de T vérifiant :

$$\frac{1}{2}\, b^p \leqslant x_i < b^p \qquad \forall\, i \qquad (p \text{ entier relatif fixé}),$$

la sommation par le schéma dichotomique de ces nombres donne une meilleure approximation de $\sum_i x_i$, que la sommation b à b.

La démonstration de la proposition V. 5, pour $\oplus = \oplus_\nabla$, lorsque b est une puissance de 2, $b = 2^\alpha$ $(\alpha > 1)$, s'établit en remarquant l'inégalité suivante :

$$\sigma_{i,j}^{(b)} \leqslant \sigma_{i,\alpha j}^{(2)} \qquad \forall\, i$$

où les notations $\sigma_{ij}^{(b)}$ et $\sigma_{ij}^{(2)}$ désignent les éléments des couches j dans les calculs par $\oplus$ suivant les schémas de sommation respectivement b à b et 2 à 2.

La proposition V. 5 n'est pas valable lorsque $b = 3$. Ainsi :

Exemple 1 : $b = 3$, $s = 2$, $\oplus = \oplus_\nabla$, $N = 6$

$x_1 = x_4 = 0{,}12$ $\qquad x_2 = x_5 = 0{,}21$ $\qquad x_3 = x_6 = 0{,}20$ $\qquad \sigma = 11$

Par les sommations : b à b $\bar{\sigma}_b = 11$ $\qquad 2$ à 2 : $\bar{\sigma}_d = 10$

Lorsque $b = 10$, la proposition V. 5 se démontre par la succession d'inégalités du type suivant, écrites pour $N = 16$:

$$\sigma_{1,1}^{(10)} \leqslant \sigma_{1,3}^{(2)} \oplus \sigma_{5,1}^{(2)}$$

$$\oplus \sum_{i=11}^{16} x_i \leqslant \sigma_{6,1}^{(2)} \oplus \sigma_{7,1}^{(2)} \oplus \sigma_{8,1}^{(2)}$$

Par les hypothèses sur les éléments x_i :

$$\sigma_{6,1}^{(2)} \oplus \sigma_{7,1}^{(2)} \oplus \sigma_{8,1}^{(2)} = \sigma_{6,1}^{(2)} + \sigma_{7,1}^{(2)} + \sigma_{8,1}^{(2)} = \sigma_{6,1}^{(2)} \oplus \sigma_{4,2}^{(2)}$$

et

$$\bar{\sigma}_b \leqslant (\sigma_{1,3}^{(2)} \oplus \sigma_{5,1}^{(2)}) \oplus (\sigma_{6,1}^{(2)} \oplus \sigma_{4,2}^{(2)}) \leqslant \sigma_{1,3}^{(2)} \oplus (\sigma_{5,1}^{(2)} \oplus (\sigma_{6,1}^{(2)} \oplus \sigma_{4,2}^{(2)}))$$

$$\leqslant \sigma_{1,3}^{(2)} \oplus ((\sigma_{5,1}^{(2)} \oplus \sigma_{6,1}^{(2)}) \oplus \sigma_{4,2}^{(2)}) = \bar{\sigma}_d$$

Nous donnons enfin ci-dessous un exemple où la base est une puissance de 2 , mais où la condition $1/2\ b^p \leqslant x_i < b^p \quad \forall\ i$, n'est pas satisfaite :

Exemple 2 :

b = 4 s = 4 N = 16 avec :

$x_1 = 0,1103$	$x_2 = 0,1012$	$x_3 = 0,2013$	$x_4 = 0,1330$
$x_5 = 0,1003$	$x_6 = 0,1100$	$x_7 = 0,1131$	$x_8 = 0,2320$
$x_9 = 0,1323$	$x_{10} = 0,1330$	$x_{11} = 0,2131$	$x_{12} = 0,2222$
$x_{13} = 0,1323$	$x_{14} = 0,2002$	$x_{15} = 0,1323$	$x_{16} = 0,2200$

La somme exacte de ces nombres est $\sigma = 0,132122 \times 4^2$
Les sommes calculées en arithmétique flottante par $\oplus_\nabla$ suivant les schémas de sommation, respectivement séquentiel, b à b et dichotomique sont :

$\bar{\sigma}_s = 0,1312 \times 4^2$
$\bar{\sigma}_b = 0,1321 \times 4^2$
$\bar{\sigma}_d = 0,1313 \times 4^2$

I. 4. CONCLUSION

1) On peut remarquer qu'étant donné un algorithme, succession d'opérations élémentaires, l'étude de la minimisation des erreurs d'arrondi pour un coût fixé, c'est-à-dire, suivant la définition de I. BABUSKA [1] pour un nombre d'opérations fixé, la nécessité de mémoires supplémentaires ou une plus grande difficulté de programmation étant négligées, peut être approchée sous plusieurs aspects ; un algorithme étant défini et son coût fixé, on peut, en effet, envisager :

* une étude maximaliste, c'est-à-dire la recherche de la stratégie minimisant la borne supérieure de l'erreur d'arrondi, ce maximum étant calculé lorsque les données initiales parcourent leurs espaces de définition.

* une étude probabiliste, afin de minimiser, dans les mêmes hypothèses que pour l'étude précédente, la moyenne de l'erreur d'arrondi, considérée comme variable aléatoire.

* une étude arithmétique : recherche de la stratégie minimisant l'erreur d'arrondi pour un coût donné et pour des données initiales fixées.

C'est ce dernier point de vue que nous avons adopté, auquel répondent les propositions V. 1 à V. 5.

Nous constatons que les résultats obtenus par l'étude arithmétique rejoignent les conclusions établies par l'étude maximaliste. En effet :

J. H. WILKINSON [42 p.17] a établi que la borne supérieure de l'erreur d'arrondi entachant une somme algébrique, est minimum si les termes sont sommés dans l'ordre des valeurs absolues croissantes.

P. LINZ [22] a démontré que les bornes supérieures de l'erreur d'arrondi entachant une somme de N nombres, sont, pour les schémas séquentiel et dichotomique dans le rapport $N/2 \log_2 N$, lorsque les hypothèses suivantes sont satisfaites : les N nombres sont écrits en virgule flottante normalisée, la base étant 2 et la règle définissant l'addition étant une règle de troncature.

I. BABUSKA [1] établit que le schéma dichotomique est optimal, du point de vue maximaliste, pour un coût fixé à N - 1 opérations.

2) Cette première partie du chapitre V comporte peu d'exemples numériques, portant sur des sommes à nombre d'éléments faible : la comparaison expérimentale entre les schémas de sommation séquentiel et dichotomique a été réalisée par plusieurs auteurs ([1] , [22]).
Nous signalerons l'étude expérimentale de J. GREGORY [11] , comparant précision, rapidité d'éxécution et encombrement relatifs des schémas de sommation, séquentiel, dichotomique, et d'une méthode de "correction".
Enfin, pour la programmation du schéma dichotomique, on pourra voir R. J. WALKER [41].

3) Nous soulignerons l'application des résultats précédents au calcul de sommes de séries positives : la recherche d'une meilleure précision dans ce calcul, implique de ne pas utiliser systématiquement l'algorithme usuel, de la sommation séquentielle de N termes de la série dans l'ordre des indices croissants, particulièrement dans le cas de séries monotones.

II - CALCUL DE LA SOMME D'UNE SERIE NUMERIQUE ALTERNEE DONT LE TERME GENERAL DECROIT

II . 1 . INTRODUCTION ET NOTATIONS

T désignant le système T_b^s de nombres à virgule flottante, considérons le problème suivant :

Une série numérique alternée de terme général décroissant étant donnée sur T , soit à calculer la somme des N premiers termes de la série :

$$S = x_1 - x_2 + x_3 - \ldots + (-1)^{N-1} x_N$$

avec : $0 < x_{i+1} < x_i \qquad 1 \leqslant i \leqslant N-1$

Le calcul sur T de la somme S définie s'effectue généralement par la sommation séquentielle des termes x_i , dans l'ordre des indices croissants.

Le but de l'étude ci-dessous est d'établir que l'algorithme de sommation sur T des termes dans l'ordre des modules croissants, calcule une meilleure approximation de S sur T .

Nous donnerons des conditions sous lesquelles, tous les chiffres de la somme calculée par ce second algorithme, sont significatifs.

Pour la commodité des notations, nous poserons :

$$X_i = (-1)^{N-i} x_{N-i+1}$$

soit

$$S = \sum_{i=1}^{N} X_i$$

Notons :

$$F_i = \oplus \sum_{j=1}^{i} X_j \qquad i = 1, \ldots, N \text{ (sommation dans l'ordre séquentiel)}$$

$E_i = F_i + X_{i+1} - F_{i+1}$ $\qquad i = 1, \ldots, N-1$

$p_i = e(F_i)$ $\qquad i = 1, \ldots, N$ (exposant de F_i)

L'écart entre S et la somme $\overline{S} = F_N$, calculée sur T dans l'ordre des modules croissants, est :

$$E = S - F_N$$

vérifiant :

$$E = \sum_{i=1}^{N-1} E_i$$

II . 2 . ETUDE PRELIMINAIRE

Nous excluons les possibilités d'underflow en admettant qu'est réalisée la condition suffisante $p_1 \geqslant m_2 + s-1$.

Supposons par exemple X_1 positif.

La suite des sommes partielles $\sum_{j=1}^{i} X_j$, i impair, est positive et croissante.

La suite des sommes partielles $\sum_{j=1}^{i} X_j$, i pair, est négative et décroissante.

Ces propriétés sont conservées par les sommes partielles F_i calculées sur T ; en effet :

Si les conditions suivantes sont réalisées :

- L'opération $\oplus$ est correcte
- $\forall_i$, $|X_{i+1}| \geqslant (|X_i|)''$ où $(|X_i|)''$ désigne le successeur de $|X_i|$

dans T , on peut affirmer que, pour tout i :

F_i a le signe de $(-1)^{i-1}$

$|F_i| < |X_{i+1}|$

$|F_{i+2}| > |F_i|$

Les deux premiers résultats s'établissent conjointement par récurrence, la troisième inégalité est en conséquence de ceux ci.

Remarque : La nécessité de l'hypothèse: $|X_{i+1}| \geqslant (|X_i|)$" est illustrée par l'exemple suivant :

$$b = 10,\ s = 2\ ,\ \oplus = \oplus_\nabla\ ,\ S = 0{,}26 - 1{,}4 + 1{,}4 - 11 + 11$$

où $F_3 = 0{,}20 \quad F_5 = 0$.

D'autre part, rappelant :

Si X et Y sont deux éléments de T , d'exposants respectifs $e(X)$ et $e(Y)$ avec $e(X) \geqslant e(Y)$, l'erreur $\varepsilon = X + Y - (X \oplus Y)$ entachant l'addition de ces deux nombres sur T par une opération $\oplus$ correcte, vérifie :

$$\varepsilon = \lambda\, b^{e(Y)-s} \quad \text{où} \quad \lambda \quad \text{est un entier tel que} \quad 0 \leqslant |\lambda| \leqslant b^{e(X+Y)-e(Y)}-1$$

λ est nul dans les cas de soustractions arithmétiques où $e(X+Y) < e(Y)$ (cancellation).

nous remarquerons que, dans le problème étudié, le phénomène de cancellation ne peut se produire pour 2 additions successives $F_{i-1} + X_i$ et $F_i + X_{i+1}$, en raison de la non-décroissance des exposants p_i de même parité, et nous énoncerons ici :

Lemme V . 2 .

Si, dans le calcul des termes F_i , une cancellation se produit au rang i , c'est-à-dire si $p_i < p_{i-1}$, l'erreur E_i au rang suivant vérifie :

$$E_i = \lambda\, b^{p_{i-1}-s} \quad \text{où} \quad \lambda \quad \text{est un entier tel que} \quad 0 \leqslant |\lambda| \leqslant b^{p_{i+1}-p_{i-1}}-1 \ .$$

Soit, en effet :

$$F_i = F_{i-1} \oplus X_i$$

$$F_{i+1} = F_i \oplus X_{i+1}$$

Une cancellation lors du calcul de F_i (la condition $p_{i-1} = e(X_i) \vee e(X_i)-1$ est nécessaire pour cela) implique, si t est l'entier positif défini par $p_i = p_{i-1} - t$, que les t derniers chiffres de la mantisse de F_i sont nuls.

L'erreur E_i entachant le calcul de F_{i+1} s'écrit alors

$$E_i = \lambda\, b^{p_i - s + t} \quad \text{avec} \quad 0 \leqslant |\lambda| \leqslant b^{p_{i+1} - p_i - t} - 1$$

Extrayons de la suite $(p_1, p_2, \ldots, p_N)$ la sous-suite $(p_{i_1}, p_{i_2}, \ldots, p_{i_K})$ définie par les conditions suivantes :

$$\begin{cases} p_{i_1} = p_1 \\ p_{i_j} > p_{i_{j-1}} \quad \text{(strict)} \qquad j \geqslant 2 \end{cases}$$

L'étude précédente permet d'écrire les erreurs élémentaires E_{i_j-1} et l'erreur globale E sous la forme :

$$E_{i_j-1} = \lambda_{i_j-1}\, b^{p_{i_j} - s}_{i_{j-1}} \quad \text{où} \quad 0 \leqslant |\lambda_{i_j-1}| \leqslant b^{p_{i_j} - p_{i_{j-1}} - 1} - 1 \qquad 2 \leqslant j \leqslant K$$

$$E = \sum_{i=1}^{N-1} E_i = \sum_{j=2}^{K} E_{i_j-1}$$

II . 3 . THEOREME

On peut énoncer

Proposition V . 6 .

Si $X_1, X_2, \ldots, X_N$ sont N éléments de T , de signes alternés et de modules strictement croissants,

Si l'addition sur T , notée $\oplus$, est correcte,

Alors, si $\bar{S} = \bigoplus \sum_{i=1}^{N} X_i$ et $S = \sum_{i=1}^{N} X_i$ sont respectivement, la somme des X_i sur T et $\mathbb{R}$, l'une au moins des égalités suivantes est vérifiée :

$$\bar{S} = \nabla S \vee \Delta S \quad \text{ou} \quad \bar{S} - X_N = \nabla(S-X_N) \vee \Delta(S-X_N)$$

Si $\text{Max}\{e(S), e(S-X_N)\} \leqslant e(X_1) + s$ (V . 5)

l'erreur $E = S - \bar{S}$ est élément de T .

Si l'inégalité (V.5) est stricte et si l'une des conditions suivantes est réalisée :

1) $e(\sum_{j=1}^{2i} X_j) \leqslant e(\sum_{j=1}^{2i-1} X_j)$ $\forall i$, et $\bigoplus$ I-dirigée
(resp.$\geqslant$)

2) $k_1 + k_2 \leqslant s-1$ où $k_1 = \max\{e(S), e(S-X_N)\} - e(X_1)$ et k_2 désigne le nombre maximum de chiffres de cancellation lors des sommes partielles $\bigoplus \sum_{j=1}^{i} X_j$, $2 \leqslant i \leqslant N$

on a :

$$E = -(\bar{S} \ominus X_N \ominus X_{N-1} \ominus \dots \ominus X_1) \qquad (V.\ 6)$$

Démonstration :

$$E = \sum_{j=2}^{K} E_{i_j - 1}$$

La majoration :

$$|E| \leqslant \sum_{j=2}^{K} b^{p_{i_j}-s} - b^{p_{i_{j-1}}-s} = b^{p_{i_K}-s} - b^{p_1-s} < b^{p_{i_K}-s}$$

entraîne :

l'égalité $\bar{S} = F_N = \nabla S \vee \Delta S$ si $p_{i_K} = p_N$

l'égalité $F_{N-1} = \nabla(S-X_N) \vee \Delta(S-X_N)$ sinon,

d'où la première partie de la proposition.

D'autre part, E s'écrit :

$$E = \left(\sum_{j=2}^{K} \lambda_{i_j-1} \, b^{p_{i_{j-1}} - p_1} \right) b^{p_1 - s} = M \, b^{p_1 - s}$$

où

$M = \sum_{j=2}^{K} \lambda_{i_j-1} \, b^{p_{i_{j-1}} - p_1}$ est un entier satisfaisant, si (V . 5) est vérifiée à l'inégalité :

$$|M| < b^s$$

E est donc élément de T .

Pour la troisième partie de la proposition, notons pour $1 \leqslant i \leqslant N$:

$$F_i = l_i \, b^{p_i - s} \qquad b^{s-1} \leqslant |l_i| < b^s$$

Remarquons que E_i , erreur entachant l'addition correcte de deux nombres de signes contraires F_i et X_{i+1} où $e(F_i) \leqslant e(X_{i+1})$ vérifie :

$$F_i - E_i = \left(\lfloor l_i \, b^{p_i - p_{i+1}} \rfloor \vee \lceil l_i \, b^{p_i - p_{i+1}} \rceil \right) b^{p_{i+1} - s} \quad \text{si} \quad p_{i+1} \geqslant p_i \qquad (V.7)$$

Le calcul de E par la formule (V.6), lorsque l'une des conditions suffisantes 1) ou 2) est réalisée, s'établit en montrant que, pour $i = N - 1, \ldots, 0$ les nombres :

$$U_i = F_N - \sum_{j=N}^{j=i+1} X_j$$

sont éléments de T .

Ces nombres s'écrivent :

$$U_i = F_i - \left(\sum_{j=i}^{N-1} E_j \right) \qquad (V.8)$$

Nous distinguerons deux cas, suivant les valeurs relatives de p_i et p_{i+1} :

a) $p_{i+1} > p_i$

$p_{i+1} = p_i + u$ où u est un entier positif.

On décomposera :

$$U_i = F_i - E_i - \left(\sum_{j=i+1}^{N-1} E_j\right)$$

Un calcul semblable à celui effectué sur E pour montrer que E est élément de T, permet d'écrire :

$$\sum_{j=i+1}^{N-1} E_k = M\, b^{p_{i+1}-s} \quad \text{où } M \text{ est entier et vérifie } |M| < b^{s-1}$$

(L'inégalité V.5 est stricte)

D'autre part, par la formule (V.7) :

$F_i - E_i = L\, b^{p_{i+1}-s}$ où L est entier et vérifie $b^{s-1-u} \leq |L| \leq b^{s-u}$

De l'écriture :

$U_i = (L - M)\, b^{p_{i+1}-s}$

où l'entier relatif $L - M$ satisfait à $|L - M| < b^s$

on déduit $U_i \in T$.

b) $p_{i+1} \leq p_i$

Nécessairement $e(X_{i+1}) = p_i \vee p_i + 1$. Distinguons les 2 cas :

$b-1) e(X_{i+1}) = p_i + 1$

On a donc :

$e(X_j) \geq p_i + 1$ pour $j \geq i + 1$.

Soit λ, s'il existe, le plus petit indice supérieur à i vérifiant $p_{\lambda+1} > p_i$

S'il n'existe pas un tel indice dans l'ensemble $\{ i + 1 , \ldots , N - 1 \}$ le problème est terminé puisque $U_i = F_i$, sinon décomposons (V.8) sous la forme :

$$U_i = (F_\lambda - \sum_{j=\lambda}^{N-1} E_j) - (\sum_{j=i+1}^{\lambda} X_j)$$

Pour le premier terme :

$p_{\lambda+1} > p_i \geqslant p_\lambda$ implique par le cas a) :

$$F_\lambda - \sum_{j=\lambda}^{N-1} E_j = m\, b^{p_{\lambda+1}-s}$$

et en notant $p_{\lambda+1} = p_i + t \qquad t > 0$

$$F_\lambda - \sum_{j=\lambda}^{N-1} E_j = m\, b^{t-1} \,.\, b^{p_i+1-s} \quad \text{avec } |m\, b^{t-1}| < b^s$$

- Cette dernière majoration résulte de l'étude faite en a) , la notation $p_{\lambda+1} = p_\lambda + u \quad (u \geqslant t)$ permettant d'écrire :

$|m| < b^{s-u} + b^{s-1-t} < b^{s-t} + b^{s-1-t}$ si $F_\lambda - \sum_{j=\lambda}^{N-1} E_j$ a le signe de F_λ

$|m| < b^{s-1-t}$ si " est du signe contraire à F_λ -

Pour le second terme, remarquons :

$e(X_j) = p_i + 1 \qquad i + 1 \leqslant j \leqslant \lambda$

donc :

$$\sum_{j=i+1}^{\lambda} X_j = n\, b^{p_i+1-s} \quad \text{où } n \text{ est entier}$$

De l'écriture :

$$\sum_{j=i+1}^{\lambda} X_j = (\sum_{j=i+1}^{\lambda-1} X_j) + X_\lambda$$

où le terme entre parenthèse est du signe de $X_{\lambda-1}$ et de module inférieur à X_λ ,

on tire :

$|n| < b^s - b^{s-1}$

$U_i \in T$

$$\underline{b - 2)\ e(X_{i+1}) = p_i}$$

U_i étant considéré sous la forme (V.8), on remarquera que l'on peut écrire :

$\sum_{j=i}^{N-1} E_j = M\, b^{p_i - s}$ où M est entier et satisfait à $|M| < b^{s-1}$

Il faut alors vérifier que dans la soustraction :

$$F_i - \sum_{j=i}^{N-1} E_j = (l_i - M)\, b^{p_i - s} \qquad (V.9)$$

l'entier relatif $l_i - M$ satisfait à $|l_i - M| < b^s$.

Distinguons deux cas :

$$b - 2 - 1)\quad p_{i+1} = p_i$$

De :

$F_{i+1} = F_i + X_{i+1}$, $p_{i+1} = p_i = e(X_{i+1})$ on déduit :

$$|l_i| < (b-1)\, b^{s-1}$$

d'où le résultat $U_i \in T$.

$$b - 2 - 2)\ p_{i+1} < p_i$$

Suivant les conditions 1) ou 2) de la proposition V.6 :

* Si la condition 1) est vérifiée avec par exemple :

$$e\left(\sum_{j=1}^{2i} X_j\right) \leqslant e\left(\sum_{j=1}^{2i-1} X_j\right)$$

alors l'inégalité $p_{i+1} < p_i$ ne peut avoir lieu que pour des valeurs impaires de l'indice i .

Pour tout i ,

$E_{2i-1} = 0$

E_{2i} a le signe de F_{2i+1}

Dans la formule (V.9), les éléments F_i et $\sum_{j=i}^{N-1} E_j$ sont de même signe et $U_i \in T$.

* Si la condition 2) est vérifiée, on a :

$|M| < b^{k_1}$

D'autre part :

$|F_i| = |X_{i+1} - F_{i+1}|$

et

$|l_i| < b^s - b^{s-k_2-1}$

La condition $k_1 + k_2 \leqslant s - 1$ assure alors $|l_i - M| < b^s$ dans (V.9)

Pour tout i , $F_N - \sum_{j=N}^{j=i+1} X_j$ est élément de T , donc, $\oplus$ étant correcte, on obtient successivement :

$$F_N \ominus X_N = F_N - X_N$$

puis

$$(F_N \ominus \sum_{j=N}^{i+1} X_j) \ominus X_i = F_N - \sum_{j=N}^{i} X_j \qquad i = N - 1, \ldots, 1$$

d'où le résultat.

REMARQUES :

1) Les conditions suffisantes 1) et 2) de la proposition V.6 sont assez restrictives. Relativement à ce point, nous donnerons les deux exemples suivants pour lesquels la formule (V.6) n'est pas exacte.

Exemple 1.

$b = 10$, $s = 2$, $S = 0{,}29 - 0{,}63 + 0{,}64 - 0{,}71 + 1{,}6 - 2{,}2 + 12$, $\oplus$ I-dirigée.

La condition suffisante 1) est vérifiée mais l'inégalité V.5 n'est pas stricte.

$\overline{S} = 10$ $\quad E = 0{,}99$ $\qquad \overline{S} \ominus \sum_{i=N}^{1} X_i = -0{,}96$

Exemple 2.

$b = 10$, $s = 3$, $\oplus = \oplus_{\Delta}$, $S = 0{,}984 - 0{,}995 + 0{,}998 - 2{,}23 + 25$

$k_1 = 2$, $k_2 = 1$

$\overline{S} = 23{,}8$ $\quad E = -0{,}043$ $\qquad \overline{S} \ominus \sum_{i=N}^{1} X_i = 0{,}046$

2) Si la formule (V.6) n'est pas applicable dans tous les cas, nous remarquerons que dès que E est élément de T , cette erreur est directement calculable par la formule :

$$E = \sum_{j=2}^{K} E_{i_j - 1}$$

Il sera fait à chaque pas i une comparaison des exposants p_{i+1} et $\max_{j \leqslant i} p_j$, les erreurs élémentaires, lorsqu'elles seront non nulles, seront calculées par la formule - chapitre II - :

$$E_i = -(F_{i+1} \ominus X_{i+1} \ominus F_i)$$

II. 4. EXEMPLES NUMERIQUES ET MODE OPERATOIRE

1) Le calcul des 2 sommes $\sigma_3 = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i}$ et $\sigma_4 = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i^2}$, sur une I.B.M. 360, a donné les résultats joints (à comparer à ceux du paragraphe I. 5 du chapitre III, p.62):

Dans les 2 tableaux de nombres, A -nombres écrits en base 16- et B -nombres écrits en base 10- :

$\bar{\sigma}_3$ et $\bar{\sigma}_4$ sont les sommes calculées sur T dans l'ordre des modules croissants

E_{σ_3} et E_{σ_4} sont les erreurs entachant $\bar{\sigma}_3$ et $\bar{\sigma}_4$, calculées par la formule (V. 6)

σ_{3DP} et σ_{4DP} sont les valeurs exactes de σ_3 et σ_4, calculées en double précision.

(Les tableaux A et B, réprésentant les mêmes résultats dans 2 bases différentes, ont été tous deux reproduits pour conserver précision et compréhension des résultats. Relativement au tableau B nous soulignerons que, bien que les valeurs $\bar{\sigma}_3$ et $\bar{\sigma}_4$ soient des éléments de T, leur sortie est effectuée en double précision pour éviter une catastrophique erreur de changement de base).

On vérifie bien :

- que tous les chiffres de mantisse de $\bar{\sigma}_3$ et $\bar{\sigma}_4$ sont exacts
- que E_{σ_3} et E_{σ_4} donnent la valeur exacte des erreurs $\sigma_3 - \bar{\sigma}_3$ et $\sigma_4 - \bar{\sigma}_4$: nous avons souligné les chiffres de mantisse d'ordre supérieur à $s = 6$ dans la double précision ainsi que la mantisse du terme d'erreur.

TABLEAU A

N	$\overline{\sigma}_3$	E_{σ_3}	σ_{3DP}	$\overline{\sigma}_4$	E_{σ_4}	σ_{4DP}
100	40B02C0E	80000000	40B02C0E00000000	40D289F4	3A88F000	40D289F488F00000
200	40B0CEAB	BAC00000	40B0CEAA40000000	40D28C62	3A59F000	40D28C6259F00000
300	40B1050D	BA1B0000	40B1050CE5000000	40D28CD6	3A1F2000	40D28CD61F200000
400	40B12048	BAFA0000	40B1204706000000	40D28CFE	3AB85100	40D28CFEB8510000
500	40B130A0	BA1F0000	40B1309FE1000000	40D28D11	3A87C200	40D28D1187C2CCCC
600	40B13B87	B9E00000	40B13B86F2000000	40D28D1B	3AC13800	40D28D1BC1380000
700	40B14352	BADC0000	40B1435124000000	40D28D21	3AEC2900	40D28D21EC29CCCC
800	40B1492A	BAE70000	40B1492919000000	40D28D25	3AED4600	40D28D25ED460000
900	40B14DB5	BA2E0000	40B14DB4D2000000	40D28D28	3AAC4400	40D28D28AC44CCCC
1000	40B15158	BA190000	40B15157E7000000	40D28D2A	3AA33600	40D28D2AA3360000

TABLEAU B

N	$\overline{\sigma}_3$	E_{σ_3}	σ_{3DP}	$\overline{\sigma}_4$	E_{σ_4}	σ_{4DP}
100	0.68817222118378	0.0	0.68817222118378	0.82241749763489	0.3258174E-07	0.82241753021663
200	0.69065350294113	-0.4470348E-07	0.69065345823765	0.82245457172394	0.2094021E-07	0.8224545926641 4
300	0.69148331880569	-0.6286427E-08	0.69148331251927	0.82246148586273	0.7246854E-08	0.82246149310959
400	0.69189882278442	-0.5820766E-07	0.69189876457676	0.82246387004852	0.4291451E-07	0.82246391296303
500	0.69214820861816	-0.7217750E-08	0.69214820140041	0.82246500253677	0.3160858E-07	0.82246503414535
600	0.69231456518173	-0.3259629E-08	0.69231456192210	0.82246559858322	0.4498725E-07	0.8224656435 7C47
700	0.69243347644806	-0.5122274E-07	0.69243342522532	0.82246595621109	0.5498532E-07	0.82246601119641
800	0.69252264499664	-0.5378388E-07	0.69252259121276	0.82246619462967	0.5524453E-07	0.82246624987420
900	0.69259196519852	-0.1071021E-07	0.69259195448831	0.82246637344360	0.4010872E-07	0.82246641355232
1000	0.69264745712280	-0.5820766E-08	0.69264745130204	0.82246649265289	0.3800051E-07	0.8224665306534C

2) Le sous-programme reproduit ci-dessous:

SOMSER (U, N, S, E) réalise, à partir du tableau U des N premiers termes d'une série numérique alternée de terme général décroissant, les calculs de $\bar{S} = \oplus \sum_{i=N}^{1} U(i)$ et de $E = \left(\sum_{i=1}^{N} U(i) \right) - \bar{S}$.

Ce dernier calcul est effectué par la méthode de comparaison à chaque pas des exposants p_{i+1} et $\max_{j \leqslant i} p_j$. La fonction auxiliaire NEXP (Z) isole l'exposant du nombre flottant Z .

```
      SUBROUTINE SOMSER(U,N,S,E)
      DIMENSION U(N)
      S=0.0
      E=0.0
      NO=NEXP(U(N))
      DO 1 I=1,N
      L=N+1-I
      T=S+U(L)
      NP=NEXP(T)
      IF(NP.EQ.NO)GOTO 2
      NO=NP
      E=E-(T-U(L)-S)
    2 CONTINUE
      S=T
    1 CONTINUE
      RETURN
      END


      INTEGER FUNCTION NEXP(Z)
      EQUIVALENCE(M,U)
      L=16777216
      U=ABS(Z)
      M=M/L
      NEXP=M
      RETURN
      END
```

Mode opératoire

La recherche d'une meilleure précision dans le calcul de la somme d'une série numérique alternée, par la sommation des termes dans l'ordre croissant, entraîne une modification au niveau de la programmation : Il est en effet nécessaire, avant de procéder à la sommation proprement dite, de déterminer l'ordre N du plus petit terme, en valeur absolue, à calculer.

Ceci pourra se faire par un balayage rapide, à l'aide d'un indice k augmentant à partir de 1, par "pas" large de P (par exemple P = 25 ou 50), une comparaison de x_k à la précision cherchée étant établie pour chaque pas. Ce balayage et, pour une précision donnée, le nombre d'additions supplémentaires de termes de la série que peut entraîner par rapport à la méthode classique, l'utilisation du pas P pour déterminer l'arrêt des calculs, est largement compensé par la précision obtenue.

Relativement au critère d'arrêt lors du calcul numérique de la limite d'une suite, mention doit être faite du critère de convergence numérique établi par K. NICKEL et K. RITTER [30] : Le critère donné par ces auteurs, pour la détermination,en fonction de la longueur s de la représentation du mot-machine,de l'indice N(s) du dernier terme à calculer, permet d'assurer la convergence numérique sous des conditions suffisantes assez faibles.

CHAPITRE VI

UTILISATION CONJOINTE DE DEUX ARITHMETIQUES CORRECTES

RODUCTION

Soit S un sous-ensemble fini de $\mathbb{R}$ contenant 0.
A l'opération $\star$ appartenant à l'ensemble $\{+ , - , \times , /\}$ des opérations élémentaires définies sur $\mathbb{R}$, il a été associé les 2 opérations correctes, définies sur S, $\circledast_{\nabla}$ et $\circledast_{\Delta}$. $\forall (x, y) \in S^2_{\wedge}\ x \star y \in \mathbb{R}_S$:

$$x \circledast_{\nabla} y = \text{Max} \{ z \mid z \in S \wedge z \leqslant x \star y \}$$

$$x \circledast_{\Delta} y = \text{Min} \{ z \mid z \in S \wedge z \geqslant x \star y \}$$

Une suite $\mathcal{G} = \{\star_i , i=1, \ldots , L\}$ d'opérations sur $\mathbb{R}$ étant donnée, il est possible de faire correspondre à $\mathcal{G}$, 2^L suites d'opérations correctes sur S :

$O = \{ (\star_i)_{\square_i} \}$ où $\square_i$ prend la valeur ∇ ou Δ.

Deux problèmes se posent alors :

- Etude de l'ensemble des résultats obtenus lors de l'application des 2^L suites précédentes à une même suite finie de $L + 1$ opérandes.

- Possibilité de déterminer, parmi les suites d'opérations sur S, 2 suites particulières calculant, à partir de la suite donnée d'opérandes, deux valeurs d'encadrement de la valeur définie sur $\mathbb{R}$ par l'application de $\mathcal{G}$ aux données initiales.

I - DEFINITIONS

$N \in \mathbb{N}$, étant donnés un vecteur $x = (x_1, x_2, \ldots, x_N)$ de S^N et une suite d'opérations sur $\mathbb{R}$, $\mathcal{G} = \{*_1, *_2, \ldots, *_{N-1}\}$ définissant le réel, noté $i_{\mathcal{G}}(x)$, $r = r_N$ par : $r_1 = x_1$

$$r_{i+1} = r_i *_i x_{i+1} \quad i=1,\ldots,N-1,$$

soit $O = \{\circledast_1, \circledast_2, \ldots, \circledast_{N-1}\}$ une suite d'opérations sur S, non nécessairement correctes correspondant aux opérations de $\mathcal{G}$.

On appelera "image de x par O" et on notera $i_O(x)$ le nombre $\bar{r} = \bar{r}_N$ défini par : $\bar{r}_1 = x_1$

$$\bar{r}_{i+1} = \bar{r}_i \circledast_i x_{i+1} \qquad i=1,\ldots,N-1.$$

Nous poserons :

Définition VI. 1.

La suite d'opérations sur S, $O = \{\circledast_1, \ldots, \circledast_{N-1}\}$ sera dite une G - suite (resp. D - suite) de $\mathcal{G}$ sur la partie $\mathcal{P}$ de S^N si :

$$i_O(x) \leqslant i_{\mathcal{G}}(x) \quad , \forall\, x \in \mathcal{P}$$
$$(\text{resp. } i_O(x) \geqslant i_{\mathcal{G}}(x) \quad , \forall\, x \in \mathcal{P})$$

Définition VI. 2.

Une G - suite $\tilde{O} = \{\tilde{\circledast}_1, \ldots, \tilde{\circledast}_{N-1}\}$ de $\mathcal{G}$ sur $\mathcal{P}$, sera dite minimisante (resp. maximisante) sur $\mathcal{P}$ si :

- les opérations $\tilde{\circledast}_i$ sont correctes
- pour toute autre G - suite $O = \{\circledast_1, \ldots, \circledast_{N-1}\}$ de $\mathcal{G}$ sur $\mathcal{P}$ d'opérations correctes, on a :

$$i_O(x) \leqslant i_{\tilde{O}}(x) \quad , \forall\, x \in \mathcal{P}$$
$$(\text{resp. } i_O(x) \geqslant i_{\tilde{O}}(x) \quad , \forall\, x \in \mathcal{P})$$

De façon analogue, on définit une D - suite minimisante (resp. maximisante) de $\mathcal{G}$ sur $\mathcal{P}$.

Définition VI. 3.

$O = \{ \circledast_i \}$ étant une suite d'opérations correctes sur S, correspondant à une suite $\mathcal{G} = \{ * _i \}$ d'opérations définies sur $\mathbb{R}$, on appelle suite conjuguée de O, et on note $\bar{O} = \{ \bar{\circledast}_i \}$ la suite définie par :

$$\bar{\circledast}_i \Big| = \begin{cases} (*_i)_\Delta & \text{si } \circledast_i = (*_i)_\nabla \\ (*_i)_\nabla & \text{si } \circledast_i = (*_i)_\Delta \end{cases}$$

Les notations et définitions précédentes s'étendent lorsque la relation d'ordre définie dans l'espace image n'est pas une relation d'ordre totale -succession d'opérations élémentaires faisant correspondre, à un vecteur de S^N, un élément d'une puissance de S, vecteur ou matrice-.

Etant donnée une suite $\mathcal{G}$ d'opérations sur $\mathbb{R}$, l'existence, dans l'ensemble des suites d'opérations correctes sur S correspondant à $\mathcal{G}$ d'une G - suite (resp. D - suite) de $\mathcal{G}$ n'est pas assurée.

II - SUITES MAXIMISANTES D'OPERATIONS SUR S RELATIVES AUX ALGORITHMES USUELS [35]

II. 1. SOMME ALGEBRIQUE ET PRODUIT SCALAIRE

2 N éléments x_i et y_i de S étant donnés (i=1,...,N), soient :

$$r = \sum_{i=1}^{N} x_i \qquad , \quad r' = \sum_{i=1}^{N} x_i \; y_i$$

Désignant par $\mathcal{G}_1$ et $\mathcal{G}_2$ les suites d'opérations définissant r et r', on a immédiatement le résultat suivant :

Proposition VI. 1.

Les deux suites $\overset{\sim}{\delta}_G = \{ (*_i)_\nabla \}$ et $\overset{\sim}{\delta}_D = \{ (*_i)_\Delta \}$ sont, respectivement, une G - suite et une D - suite maximisantes de $\mathcal{G}_1$ sur S^N, de $\mathcal{G}_2$ sur S^{2N}.

Conséquences pratiques :

Un cas où la mise en oeuvre des suites maximisantes d'opérations précédentes, sur un système $T_b^s = T$ de nombres à virgule flottante, est très simplement résolue, est celui de la somme algébrique lorsqu'on dispose de l'opération

$$\oplus_\nabla \text{ (resp. } \oplus_\Delta) :$$

Considèrons le cas où l'opération $\oplus$, d'addition sur T, est définie par $\oplus_\nabla$.

La symétrie ($z \to -z$) n'étant pas un automorphisme de T, on formera :

$$r^1 = \oplus \sum_{i=1}^{N} x_i$$

$$r^2 = - \{ \oplus \sum_{i=1}^{N} (- x_i) \}$$

Alors :

$r^1 = i \overset{\sim}{\delta}_G (x)$, $r^2 = i \overset{\sim}{\delta}_D (x)$ $\qquad x = (x_1, \ldots, x_N)$.

Dans les autres cas, il sera préférable de générer directement les opérations $(*)_\nabla$ et $(*)_\Delta$. Nous notons ci-dessous, à titre d'exemple, les encadrements des sommes $\sigma_1 = \sum_{i=1}^{N} \frac{1}{i}$ et $\sigma_3 = \sum_{i=1}^{N} \frac{(-1)^{i-1}}{i}$, calculés

sur une IBM 360 ($\sigma_{1\ D\ P}$ et $\sigma_{3\ D\ P}$ désignent les valeurs obtenues en double précision).

	N	100	200	300	400	500
σ_1	$\sigma_{1\ G}$	5,187340	5,877946	6,282538	6,569756	6,792601
	$\sigma_{1\ D}$	428	8129	816	70128	3069
	$\sigma_{1\ D\ P}$	377	8030	663	69929	2822
σ_3	$\sigma_{3\ G}$	0,6881702	0,6906486	0,6914756	0,6918882	0,6921345
	$\sigma_{3\ D}$	42	579	904	9088	610
	$\sigma_{3\ D\ P}$	22	534	833	8987	482

II. 2. EVALUATION NUMERIQUE D'UN POLYNOME

Un polynôme P étant défini sur S par ses coefficients a_i ($i = 0, \ldots, N$), soit r la valeur de ce polynôme en un point $x \in S$:

$$r = P(x) = a_0 x^N + \ldots + a_N$$

Désignant par $\mathcal{G}$ la suite d'opérations

$\mathcal{G} = \{ *_{2i-1} = x , *_{2i} = + , i = 1, \ldots, N \}$ qui permet le calcul de r par le schéma de Horner, on peut énoncer :

Proposition VI. 2

Pour le calcul de la valeur d'un polynôme en un point par le schéma de Horner, une G - suite et une D - suite maximisantes de $\mathcal{G}$ sont respectivement :

. sur $S^{N+1} \times S^{+}$

$$\gamma_G = \{ \circledast_i = \circledast_{i\,\nabla} \quad i = 1, \ldots, 2N \} \qquad \gamma_D = \overleftarrow{\gamma}_G$$

. sur $S^{N+1} \times S^{-}$

$$\tilde{\gamma}_G = O_1 \text{ si } N \text{ impair, } O_2 \text{ si } N \text{ pair} \qquad \tilde{\gamma}_D = \overleftarrow{\tilde{\gamma}}_G$$

avec :

$$O_1 = \{ \circledast_i = \circledast_{i\,\nabla} \text{ si } i \equiv 1 \text{ ou } 2 \pmod 4$$
$$= \circledast_{i\,\Delta} \text{ si } i \equiv 3 \text{ ou } 0 \pmod 4 \}$$

$$O_2 = \{ \circledast_i = \circledast_{i\,\Delta} \text{ si } i \equiv 1 \text{ ou } 2 \pmod 4$$
$$= \circledast_{i\,\nabla} \text{ si } i \equiv 3 \text{ ou } 0 \pmod 4 \}$$

Exemple : valeurs d'encadrement du polynôme $P_N(x) = \prod_{j=1}^{N} (x-j)$ au point $x = N - 0{,}01$ (r et $\bar{r}$ désignent respectivement, la valeur exacte et la valeur obtenue par l'arithmétique d'une IBM 360 - 30).

N	r_G	r_D	r	$\bar{r}$
5	- 0,2359009	- 0,2335205	- 0,2350349	- 0,2339172
6	- 1,233643	- 1,117676	- 1,172824	- 1,171875
7	- 9,113281	- 4,878906	- 7,025215	- 6,410156
8	- 108,4023	- 8,023438	- 49,10625	- 88,42188
9	- 2594	2417,438	- 392,3589	- 88

La mise en évidence du mauvais conditionnement de la plus grande racine de $P_N(x)$, est ainsi particulièrement nette.

II. 3. CALCUL DE LA RACINE CARREE D'UN NOMBRE PAR LA METHODE DE NEWTON

Etant donné un élément c de S^+ , la détermination de la racine carrée de c par la méthode de Newton, est réalisée, à partir d'une valeur initiale x_o , à l'aide de la formule de récurrence. :

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{c}{x_n} \right) \qquad \text{(VI. 1)}$$

Désignons par G la suite d'opérations qui permet, à partir des données c et x_o , le calcul de x_{n+1} , et supposons, sans restreindre le problème, $x_o \geqslant \sqrt{c}$.
On peut énoncer :

Proposition VI. 3.

Si S est un système de nombres à virgule flottante de base 2, pour la détermination du $(n+1)$ ième itéré de (VI.1) à partir d'une valeur initiale $x_o \geqslant \sqrt{c}$, une G - suite et une D - suite d'opérations sur S , maximisantes de la suite des opérations sur $\mathbb{R}$, sont respectivement :

$$\tilde{o}_G = \{ \textcircled{*}_i = \textcircled{x_i}_{\nabla} \} \quad , \quad \tilde{o}_D = \{ \textcircled{*}_i = \textcircled{x_i}_{\Delta} \}$$

La proposition VI. 3. s'établit par récurrence sur l'ordre n . Nous ferons au préalable la remarque suivante :

Si les opérations $\oplus$ et $\oslash$ définies sur $S = T_2^s$ sont correctes, les éléments v et $\bar{v}$ définis respectivement sur $\mathbb{R}$ et T_2^s , à partir d'un élément $u \geqslant \sqrt{c}$ de T_2^s par :

$$v = \frac{1}{2} \left(u + \frac{c}{u} \right) \qquad \bar{v} = \frac{1}{2} \left(u \oplus c \oslash u \right)$$

vérifient :

$$\bar{v} = \nabla v \vee \Delta v \qquad \text{(VI. 2)}$$

(VI. 2) est obtenue par :

a) $| c/u - c \textcircled{/} u | < \varepsilon (c/u)$

b) $u \geq \sqrt{c} \Rightarrow c \textcircled{/} u \leqslant u$ donc $|(u + c \textcircled{/} u) - (u \textcircled{\oplus} c \textcircled{/} u)| \leqslant \varepsilon(u + c \textcircled{/} u) - \varepsilon(c \textcircled{/} u)$

assurant ainsi :

$$| v - \bar{v} | < \varepsilon (\bar{v})$$

- Pour conclure à l'égalité $\bar{v} = \nabla v \vee \Delta v$ dans le cas particulier où $\bar{v}$ vérifie $\varepsilon ((\bar{v})') \neq \varepsilon (\bar{v})$, la démonstration est analogue à la démonstration établie lors du cas particulier similaire de la proposition V. 1 du chapitre V -

Il vient alors, pour la suite $\tilde{\mathcal{O}}_G$ par exemple :

Le couple $\tilde{\mathcal{O}}_G = \{ \textcircled{/}_\nabla , \textcircled{\oplus}_\nabla \}$ définit, à l'ordre 1, une G - suite maximisante d'opérations de $\mathcal{G}$, et l'image $\tilde{x}_1$ de x_o par $\tilde{\mathcal{O}}_G$ vérifie $\tilde{x}_1 \geq \nabla(\sqrt{c})$.

Supposons qu'à l'ordre n , la suite $\tilde{\mathcal{O}}_G = \{ \textcircled{*_i}_\nabla \}$ soit une G - suite maximisante de $\mathcal{G}$ et que l'inégalité $\tilde{x}_n \geqslant \nabla (\sqrt{c})$ soit satisfaite : les nièmes itérés $\tilde{x}_n$ et $\bar{x}_n$ de (VI. 1) , calculés sur T_2^s par respectivement, $\tilde{\mathcal{O}}_G$ et une suite O d'opérations correctes sur T_2^s correspondant à $\mathcal{G}$, vérifient $\bar{x}_n \geqslant \tilde{x}_n$

1) Si $\tilde{x}_n > \sqrt{c}$, les valeurs $\bar{\bar{x}}_{n+1} = \frac{1}{2} (\bar{x}_n + \frac{c}{\bar{x}_n})$ et $\tilde{\bar{x}}_{n+1} = \frac{1}{2}(\tilde{x}_n + \frac{c}{\tilde{x}_n})$

satisfont :

$$\bar{\bar{x}}_{n+1} \geq \tilde{x}_{n+1} > \sqrt{c} ,$$

en raison du sens de variation de la fonction $y = \frac{1}{2}(x + \frac{c}{x})$ qui de plus admet, au point $x = \sqrt{c}$, un minimum égal à $\sqrt{c}$.

(VI. 2) assure alors :

$$\bar{x}_{n+1} \geqslant \tilde{x}_{n+1} \geqslant \nabla(\sqrt{c})$$

2) Si $\tilde{x}_n = \nabla(\sqrt{c}) \leqslant \sqrt{c}$, le processus itératif de Newton est terminé. En effet, montrons que les conditions considérées entraînent l'une des deux possibilités :

$\tilde{x}_{n+1} = \nabla(\sqrt{c})$ donc convergence forte de $\{ \tilde{x}_n \}$ vers $\nabla(\sqrt{c})$.

$\tilde{x}_{n+1} = \Delta(\sqrt{c}) > \sqrt{c}$ et convergence faible de $\{ \tilde{x}_n \}$, les itérés cyclant sur les 2 valeurs $\nabla(\sqrt{c})$ et $\Delta(\sqrt{c})$ - la propriété de G - suite de $\tilde{O}_G$ n'est plus vérifiée dans ce cas limite -

De :

$$\frac{1}{2}\left(\nabla\sqrt{c} + c/\nabla\sqrt{c} \right) - \Delta\sqrt{c} = \left[c - \left(\Delta\sqrt{c} + \varepsilon(\sqrt{c}) \right) \nabla\sqrt{c} \right] / 2\nabla\sqrt{c} \leqslant 0$$

on conclut :

$$\tilde{x}_{n+1} = \nabla\sqrt{c} \text{ , } \Delta\sqrt{c} \text{ ,}$$

la condition $c < \left(\Delta\sqrt{c} + \varepsilon(\sqrt{c}) \right) \nabla\sqrt{c}$ étant une condition nécessaire et suffisante de convergence forte.

Une conséquence de la démonstration précédente est alors pour la suite $\{ \tilde{x}_n \}$, des itérés calculés par $\tilde{O}_G$, décroissante tant que $\tilde{x}_n > \nabla(\sqrt{c})$:

Proposition VI. 4.

Un élément c étant donné sur T_2^s , la suite des itérés du calcul de la racine carrée de c sur T_2^s par la méthode de Newton lorsque les opérations d'addition et de division sur T_2^s sont respectivement $\oplus_\nabla$ et $/_\nabla$,converge, quelle que soit la valeur initiale $x_o \in T_2^s$:

soit fortement vers $\nabla(\sqrt{c})$

soit faiblement en cyclant sur les valeurs $\nabla(\sqrt{c})$ et $\Delta(\sqrt{c})$.

J. M. YOHE [45] a établi la convergence forte vers $(\nabla \sqrt{c})$" de la suite des itérés de Newton, calculés à partir d'une valeur initiale $x_o \neq \sqrt{c}$, lorsque les opérations d'addition et de division sur T_2^s sont respectivement $\oplus_\Delta$ et $\oslash_\Delta$.

Exemples :

1) $b = 2 \quad s = 3 \quad c = 0{,}100 \quad \sqrt{c} = 0{,}10110\ldots \quad 0 = \tilde{\mathcal{O}}_G$
Convergence forte de $\{ \tilde{x}_n \}$ vers $0{,}101$.

2) $b = 2 \quad s = 3 \quad c = 0{,}110 \quad \sqrt{c} = 0{,}11011\ldots \quad 0 = \tilde{\mathcal{O}}_G$
Convergence faible de $\{ \tilde{x}_n \}$ qui cycle sur $0{,}110$ et $0{,}111$.

Remarque

Si le calcul de la racine carrée d'un nombre par la méthode de Newton, est effectué en arithmétique flottante de base b différente de 2 :

La proposition VI. 3 reste valable, si le nième itéré de (VI. 1) calculé par $\tilde{\mathcal{O}}_G$ satisfait à $\tilde{x}_n \geq \sqrt{c}$.

* Il suffit en effet, de reprendre le raisonnement par récurrence de la proposition, en remarquant que l'hypothèse :

$$\sqrt{c} \leq \tilde{x}_n \leq \bar{x}_n$$

qui implique $\tilde{x}_n + c/\tilde{x}_n \leq \bar{x}_n + c/\bar{x}_n$,

assure $\tilde{x}_n \oplus_\nabla c \oslash_\nabla \tilde{x}_n \leq \bar{x}_n \oplus c \oslash \bar{x}_n$

par l'égalité $u \oplus c \oslash u = \nabla(u + c/u) \vee \Delta(u + c/u) \qquad \forall u \in T_b^s \wedge u \geq \sqrt{c}$
la monotonie de l'arrondi permet alors de conclure $\tilde{x}_{n+1} \leq \bar{x}_{n+1}$ *

La convergence des itérés de Newton vers les 2 nombres flottants entourant $\sqrt{c}$ n'est par contre plus assurée, la formule (VI. 2) étant à remplacer, avec les mêmes notations, pour une base b multiple de 2, par l'inégalité plus large : $|v - \bar{v}| < \frac{b}{2}\ \varepsilon(\bar{v})$.

La propriété de convergence forte de la suite des itérés $\tilde{x}_n$ calculés par $\tilde{\mathcal{O}}_D$ est toujours satisfaite lorsque b est multiple de 2 .

*En effet, $\tilde{x}_n > \sqrt{c}$ assurant $\tilde{x}_n + c / \tilde{x}_n < 2\, \tilde{x}_n$, 2 cas peuvent se présenter:

$\Delta\ (\tilde{x}_n + c/\tilde{x}_n) < 2\tilde{x}_n$ impliquant $\tilde{x}_{n+1} \leq \tilde{x}_n$

$\Delta\ (\tilde{x}_n + c/\tilde{x}_n) = \Delta(2\tilde{x}_n)$ impliquant $\tilde{x}_{n+2} = \tilde{x}_{n+1} = \frac{1}{2}\ \Delta\ (2\tilde{x}_n)$

$(\frac{1}{2}\ \Delta\ (2u) \in T\ ,\ \forall u \in T \wedge \tilde{x}_{n+1} \geq \tilde{x}_n \Rightarrow \tilde{x}_n + c/\tilde{x}_n \leq \tilde{x}_{n+1} + c/\tilde{x}_{n+1} < 2\, \tilde{x}_{n+1})$ *

II.4.DECOMPOSITION LR ET INVERSE D'UNE M - MATRICE

II.4.1. Décomposition LR d'une M - matrice

Soit $A = (a_{ij})$ une M-matrice donnée sur S Rappelons :

Définition. $A \in \mathcal{M}_{n,n}(\mathbb{R})$ est une M-matrice si les deux axiomes suivants sont vérifiés :
- les éléments hors diagonaux de A sont négatifs ou nuls.
- tous les mineurs principaux de A sont positifs.

La décomposition d'une M-matrice A en un produit L R où $L = (\ell_{ij})$ est triangulaire inférieure ne comportant que des "1" sur la diagonale, et $R = (r_{ij})$ triangulaire supérieure, est toujours possible par les formules classiques :

$$r_{p+1,i} = a_{p+1,i} - \sum_{k=1}^{p} \ell_{p+1,k}\ r_{k,i} \qquad , p+1 \leq i \leq n$$

$$\ell_{i,p+1} = (a_{i,p+1} - \sum_{k=1}^{p} \ell_{i,k}\ r_{k,p+1}) / r_{p+1,p+1} \qquad , p+2 \leq i \leq n \qquad \text{(VI. 3)}$$

$p = 0 , 1 , \ldots , n-1.$

On a le résultat suivant [7]

Si A est une M - matrice, alors dans la décomposition LR de A :
- les éléments diagonaux de R sont positifs.
- les éléments hors-diagonaux de L et R sont négatifs ou nuls.

Désignant par σ la suite d'opérations (VI.3) définissant L et R, considérons les deux suites particulières suivantes de calcul sur S de L et R, de résultats notés $\tilde{L}$ et $\tilde{R}$:

$$\tilde{r}_{p+1,i} = a_{p+1,i} \ominus_{\Delta} \left\{ \bigoplus_{\nabla} \sum_{k=1}^{p} \tilde{\ell}_{p+1,k} \otimes_{\nabla} \tilde{r}_{k,i} \right\}$$

$$\tilde{\ell}_{i,p+1} = \left[a_{i,p+1} \ominus_{\Delta} \left\{ \bigoplus_{\nabla} \sum_{k=1}^{p} \tilde{\ell}_{i,k} \otimes_{\nabla} \tilde{r}_{k,p+1} \right\} \right] \oslash_{\Delta} \tilde{r}_{p+1,p+1} \qquad (\check{\sigma}_D)$$

et

$$\sigma_G = \bar{\sigma}_D$$

On peut énoncer :

Proposition VI. 5.
La suite $\tilde{O}_D$ est une D - suite maximisante de $\mathcal{G}$ sur son domaine de définition.
La suite $\tilde{O}_G$ est une G - suite maximisante de $\mathcal{G}$ sur son domaine de définition si la condition suivante est réalisée :

$$\tilde{r}_{pp} > 0 \qquad \forall\, p$$

Soit par exemple, à établir que $\tilde{O}_D$ est une D - suite maximisante de $\mathcal{G}$: Désignant par $O = \{\circledast\}$ une suite d'opérations correctes sur S correspondant à $\mathcal{G}$ et donnant pour image de A , $\bar{L}$ et $\bar{R}$, il suffit de raisonner par récurrence :
Supposant que les p premières lignes de R, $\tilde{R}$ et $\bar{R}$ et les p premières colonnes de L , $\tilde{L}$ et $\bar{L}$ vérifient les inégalités suivantes :

$$r_{ij} \leqslant \tilde{r}_{ij} \leqslant 0 \quad , \quad \bar{r}_{ij} \leqslant \tilde{r}_{ij} \quad , \quad 1 \leqslant i \leqslant p \quad , \quad i+1 \leqslant j \leqslant n$$

$$o < r_{ii} \leqslant \tilde{r}_{ii} \quad , \quad \bar{r}_{ii} \leqslant \tilde{r}_{ii} \quad , \quad 1 \leqslant i \leqslant p$$

$$\ell_{ij} \leqslant \tilde{\ell}_{ij} \leqslant 0 \quad , \quad \bar{\ell}_{ij} \leqslant \tilde{\ell}_{ij} \quad , \quad 1 \leqslant j \leqslant p \quad , \quad j+1 \leqslant i \leqslant n \ ,$$

on établit que ces inégalités sont encore vérifiées pour les $(p+1)^e$ lignes de R , $\tilde{R}$ et $\bar{R}$ et les $(p+1)^e$ colonnes de L , $\tilde{L}$ et $\bar{L}$:
Ainsi, de :

$$\bar{\ell}_{p+1,k}\ \bar{r}_{k,i} \geqslant \tilde{\ell}_{p+1,k}\ \tilde{r}_{k,i} \geqslant 0 \qquad , \ 1 \leqslant k \leqslant p$$

on déduit par la monotonie de l'arrondi :

$$\oplus \sum_{k=1}^{p} \bar{\ell}_{p+1,k} \otimes \bar{r}_{k,i} \geqslant \oplus \sum_{k=1}^{p} \tilde{\ell}_{p+1,k} \otimes \tilde{r}_{k,i}$$

ce qui entraîne, $\{ (\star_i) = (\star_i)_\nabla \}$ étant une G - suite maximisante pour le produit scalaire :

$$\oplus \sum_{k=1}^{p} \bar{\ell}_{p+1,k} \otimes \bar{r}_{k,i} \geqslant \oplus_\nabla \sum_{k=1}^{p} \tilde{\ell}_{p+1,k} \otimes_\nabla \tilde{r}_{k,i}$$

Il vient alors :

$$\bar{r}_{p+1,i} \leqslant \tilde{r}_{p+1,i} \leqslant 0 \qquad , \ p+2 \leqslant i \leqslant n$$

$$\bar{r}_{p+1,p+1} \leqslant \tilde{r}_{p+1,p+1}$$

et, sous la condition $\bar{r}_{p+1,p+1} > 0$ (nous ne considérons que les suites O pour lesquelles cette condition $\bar{r}_{pp} > 0 \ \forall\, p$ est satisfaite) :

$$\bar{\ell}_{i,p+1} \leqslant \tilde{\ell}_{i,p+1} \leqslant 0 \qquad , \ p+2 \leqslant i \leqslant n \ .$$

Ces inégalités sont vérifiées pour la première ligne de R, $\tilde{R}$ et $\bar{R}$ et la première colonne de L, $\tilde{L}$ et $\bar{L}$.

II. 4. 2. Résolution du système linéaire $AX = B$, A étant une M - matrice, B un second membre positif (resp. négatif)

Soient respectivement, $B = (b_i)$ et $X = (x_i)$, le second membre, donné sur S , et la solution du système linéaire $AX = B$.
La décomposition de A en un produit LR étant effectuée, le calcul de X est ramené à la résolution des deux systèmes triangulaires :

$$LY = B$$

$$RX = Y$$

définissant la suite d'opérations Θ .

Soient, sur S :
la suite $(\tilde{\Theta}_G)$ suivante, associant à L, B et R les images $\tilde{Y}$ et $\tilde{X}$:

$$\tilde{y}_i = b_i \ominus_\nabla \{ \oplus_\Delta \sum_{j=1}^{i-1} \ell_{ij} \otimes_\Delta \tilde{y}_j \} \qquad i = 1,\dots, n$$

$$\tilde{x}_i = [\tilde{y}_i \ominus_\nabla \{ \oplus_\Delta \sum_{j=i+1}^{n} r_{ij} \otimes_\Delta \tilde{x}_j \}] \oslash_\nabla r_{ii} \quad i = n,\dots,1 \qquad (\tilde{\Theta}_G)$$

et $\tilde{\Theta}_D = \bar{\Theta}_G$.

Il vient :

Proposition VI. 6.

A étant une M - matrice,

- Si $B \geqslant 0$,

$(\overset{\sim}{\delta}_D , \overset{\sim}{\theta}_G)$ est une G - suite de $(\mathcal{G} , \Theta)$

$(\overset{\sim}{\delta}_G , \overset{\sim}{\theta}_D)$ est une D - suite de $(\mathcal{G} , \Theta)$ si la condition $\tilde{r}_{pp} > 0 \ \forall \ p$, est réalisée.

- Si $B \leqslant 0$,

$(\overset{\sim}{\delta}_G , \overset{\sim}{\theta}_G)$ est une G - suite de $(\mathcal{G} , \Theta)$ si la condition $\tilde{r}_{pp} > 0 \ \forall \ p$, est réalisée.

$(\overset{\sim}{\delta}_D , \overset{\sim}{\theta}_D)$ est une D - suite de $(\mathcal{G} , \Theta)$.

L'application des résultats précédents à l'algorithme de détermination de l'inverse A^{-1} de A par résolution de n systèmes linéaires, permet d'énoncer en particulier lorsque $S = T_b^s$ -les résultats précédents ne sont pas modifiés par le fait que $\oplus$ ci-dessous est non correcte- :

Proposition VI. 6'.

Si, sur $T_b^s = T$,

- l'opération $\oplus$ est définie par une troncature sans chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue,
- les opérations $\otimes$ et $\oslash$ sont correctes et I - dirigées $\forall (x,y) \in T^2$,

alors les inverses C et $\tilde{C}$ d'une M - matrice A , calculées par décomposition LR de A et résolution de n systèmes linéaires, sur, respectivement $\mathbb{R}$ et T, vérifient :

$$\tilde{C} \leqslant C$$

Exemple de résolution de système linéaire :

Nous avons écrit ci-dessous, les solutions $\tilde{X}$, $\bar{X}$ et X obtenues respectivement, par l'algorithme (γ_D , γ_G) , par l'arithmétique IBM 360 - 30, et par la résolution exacte sur $\mathbb{R}$ de $AX = B$, pour la matrice A et le second membre B de l'exemple p.94 du chapitre IV :

$\tilde{X}$	$\bar{X}$	X
0,9999908	0,9999927	1,000000
" 17	" 28	"
" 18	" 31	"
" 20	" 34	"
" 18	" 33	"
" 14	" 30	"
" 26	" 39	"
" 18	" 33	"

II. 4. 3. Remarque : Algorithme classique de Gauss appliqué à la résolution de $AX = B$, A M - matrice, $B \geqslant 0$.

La résolution, sur $\mathbb{R}$, du système linéaire $AX = B$

- par la méthode précédente,
- ou par l'algorithme classique de Gauss, consistant à former 2 suites $\{ A^k \}$ et $\{ B^k \}$ $(1 \leqslant k \leqslant n)$ de matrices et de vecteurs pour remplacer le système donné par le système triangulaire $RX=Y$

détermine dans les 2 cas l'unique solution du système donné.

Il n'en est pas de même pour la résolution sur S , suivant une même arithmétique, de $AX = B$ par les 2 méthodes considérées, lorsque, comme il en est l'usage, le calcul des éléments $r_{p+1,i}$ et $\ell_{i,p+1}$ inclut le calcul des produits scalaires $\oplus \sum_k \ell_{p+1,k} \otimes r_{k,i}$ et $\oplus \sum_k \ell_{i,k} \otimes r_{k,p+1}$.

Nous avons étudié la résolution de $AX = B$, A M − matrice, $B \geqslant 0$, relativement à l'algorithme de décomposition LR de A.
Pour l'algorithme classique de Gauss, on aura immédiatement :

Désignant par $\mathcal{G}'$ la suite d'opérations :

$$a_{ij}^{k+1} = a_{ij}^{k} - \frac{a_{ik}^{k}}{a_{kk}^{k}} a_{kj}^{k} \qquad i, j = k+1, \ldots, n$$

$$b_{i}^{k+1} = b_{i}^{k} - \frac{a_{ik}^{k}}{a_{kk}^{k}} b_{k}^{k} \qquad i = k+1, \ldots, n$$

$k = 1, \ldots, n-1$

et définissant la suite $(\tilde{O}'_1)$ d'opérations sur S, suivante :

$$\tilde{a}_{ij}^{k+1} = \tilde{a}_{ij}^{k} \ominus_{\Delta} \left(\tilde{a}_{ik}^{k} \oslash_{\Delta} \tilde{a}_{kk}^{k} \right) \otimes_{\nabla} \tilde{a}_{kj}^{k}$$

$$\tilde{b}_{i}^{k+1} = \tilde{b}_{i}^{k} \ominus_{\nabla} \left(\tilde{a}_{ik}^{k} \oslash_{\Delta} \tilde{a}_{kk}^{k} \right) \otimes_{\Delta} \tilde{b}_{k}^{k} \qquad (\tilde{O}'_1)$$

ainsi que $\tilde{O}'_2 = \tilde{\tilde{O}}'_1$, il vient de même que précédemment :

$\tilde{O}'_1$ est une suite maximisante de $\mathcal{G}'$ sur son domaine de définition

$\tilde{O}'_2$ est une suite maximisante de $\mathcal{G}'$ sur son domaine de définition, si la condition de stabilité $\tilde{a}_{kk}^{k} > 0$ est réalisée,

et les images (R, Y) et $(\tilde{R}, \tilde{Y})$ de (A, B) par $\mathcal{G}'$ et $O' = \tilde{O}'_1 \vee \tilde{O}'_2$ vérifient :

$$R \leqslant \tilde{R} \wedge \tilde{Y} \leqslant Y \qquad \text{si} \quad O' = \tilde{O}'_1$$

$$R \geqslant \tilde{R} \wedge \tilde{Y} \geqslant Y \qquad \text{si} \quad O' = \tilde{O}'_2$$

et aussi :

Dans les hypothèses relatives à l'arithmétique flottante de la proposition VI.6'. l'inégalité :

$$\tilde{C} \leqslant C$$

est vérifiée lorsque $\tilde{C}$ est calculée sur T par l'algorithme de Gauss classique .

II . 5 . INVERSE D'UNE MATRICE TRIDIAGONALE, SYMETRIQUE, DEFINIE POSITIVE

Soit A une matrice d'ordre n, donnée sur S, tridiagonale, symétrique, que nous écrivons :

$$A = \begin{vmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \ddots & b_{n-1} & \\ & & & b_{n-1} & a_n & \end{vmatrix}$$

et soit $C = A^{-1}$

Les formules 6 de calcul de la $j^{\text{ème}}$ colonne de C, par la résolution de $A X = e_j$, sont :

- Pour le calcul des élèments diagonaux α_i de la décomposition triangulaire de A, et des éléments ε_i du second membre correspondant :

$$\alpha_1 = a_1$$

$$\alpha_i = a_i - b_{i-1}^2 \; / \; \alpha_{i-1} \qquad i = 2, \ldots, n$$

$$\varepsilon_i = 0 \qquad i = 1, \ldots, j-1$$

$$\varepsilon_j = 1$$

$$\varepsilon_i = -(b_{i-1} / \alpha_{i-1}) \; \varepsilon_{i-1} \qquad i = j+1, \ldots, n$$

- Pour le retour-arrière :

$$x_i = (\varepsilon_i - b_i \, x_{i+1}) \; / \; \alpha_i \qquad i = n, \ldots, 1$$

Notons $\tilde{\alpha}_i$, $\tilde{\varepsilon}_i$, $\tilde{x}_i$ les élèments calculés sur S par l'algorithme suivant :

$$\tilde{\alpha}_i = a_i \ominus_\Delta (b_{i-1} \oslash_I \tilde{\alpha}_{i-1}) \otimes_I b_{i-1} \qquad i = 2, \ldots, n$$

$$\tilde{\varepsilon}_i = -(b_{i-1} \oslash_I \tilde{\alpha}_{i-1}) \otimes_I \tilde{\varepsilon}_{i-1} \qquad i = j+1, \ldots, n \quad (\tilde{0}_I)$$

$$\tilde{x}_i = (\tilde{\varepsilon}_i \ominus_I b_i \otimes_I \tilde{x}_{i+1}) \oslash_I \tilde{\alpha}_i \qquad i = n, \ldots, 1$$

où l'opération $\overset{(*)}{}_I$ a la définition : $x \overset{(*)}{}_I y = \begin{cases} x \overset{(*)}{}_\nabla y & \text{si } x * y \geqslant 0 \\ x \overset{(*)}{}_\Delta y & \text{si } x * y \leqslant 0 \end{cases}$

On peut énoncer, relativement à $\tilde{O}_I$, et plus généralement en excluant l'hypothèse de correction de la soustraction arithmétique.

Proposition VI.7.

Si, sur S :

- l'opération $\oplus$ est E - dirigée $\forall (x,y) \in S^2_\wedge \; x\,y < 0$, et correcte et I - dirigée $\forall (x,y) \in S^2_\wedge \; x\,y \geqslant 0$,
- les opérations $\otimes$ et $/$ sont correctes et I - dirigées $\forall (x,y) \in S^2$,

alors les inverses C et $\tilde{C}$ d'une matrice tridiagonale, symétrique définie positive, obtenues par l'algorithme de Gauss, sur, respectivement $\mathbb{R}$ et S , vérifient :

$$|\tilde{c}_{i\,j}| \leqslant |c_{i\,j}| \qquad i,j = 1, \ldots , n$$

Démonstration

On vérifie successivement les inégalités suivantes :

$$\tilde{\alpha}_i \geqslant \alpha_i \qquad i = 2, \ldots , n$$

$$|\tilde{\varepsilon}_i| \leqslant |\varepsilon_i| \qquad i = j + 1, \ldots , n$$

L'inégalité :

$$|\tilde{x}_i| \leqslant |x_i| \qquad i = n, \ldots , 1$$

s'établit alors par récurrence sur i , en remarquant que pour $i = n - 1, \ldots , j$, les nombres ε_i et $b_i\, x_{i+1}$ sont de signes contraires (la récurrence permet d'établir que $\tilde{x}_i$ a le signe $\tilde{\varepsilon}_i$ qui reste toujours celui de ε_i , les opérations I - dirigées étant supposées correctes, et on tient compte de l'égalité $\varepsilon_{i+1} = -(b_i / \alpha_i)\, \varepsilon_i$ où $\alpha_i > 0$).

Remarques

1) Sous la condition de stabilité $\tilde{\alpha}_i > 0 \;\; \forall\, i$, $\tilde{O}_E = \tilde{O}_I$ donne pour image de A , $\tilde{C}$ vérifiant , $\forall\, i, j$: $|\tilde{c}_{i\,j}| \geqslant |c_{i\,j}|$.

2) La proposition VI.7 est encore valable avec les hypothèses suivantes sur A :

A est tridiagonale, ses mineurs principaux sont positifs et tout couple d'éléments de A , disposés symétriquement par rapport à la diagonale principale, est formé d'éléments de même signe.

Exemple : Les inverses $\tilde{C}$ et C de la matrice A définie par : $a_i = 3$ $(1 \leqslant i \leqslant 6)$, $b_i = 1$ $(1 \leqslant i \leqslant 5)$, sont respectivement :

$$C = \frac{1}{377} \begin{vmatrix} 144 & -55 & 21 & -8 & 3 & -1 \\ -55 & 165 & -63 & 24 & -9 & 3 \\ 21 & -63 & 168 & -64 & 24 & -8 \\ -8 & 24 & -64 & 168 & -63 & 21 \\ 3 & -9 & 24 & -63 & 165 & -55 \\ -1 & 3 & -8 & 21 & -55 & 144 \end{vmatrix} \qquad [6 \text{ p.152}]$$

$$\tilde{C} = \frac{1}{377} \begin{vmatrix} 144 & -54,99995 & 20,99997 & -7,999987 & 2,999993 & -0,9999985 \\ -54,99995 & 164,9999 & -62,99991 & 23,99995 & -8,999985 & 2,999995 \\ 20,99997 & -62,99994 & 167,9998 & -63,99994 & 23,99995 & -7,999990 \\ -7,999989 & 23,99995 & -63,99994 & 167,9999 & -62,99994 & 20,99997 \\ 2,999993 & -8,999984 & 23,99995 & -62,99994 & 164,9999 & -54,99995 \\ -0,9999979 & 2,999993 & -7,999987 & 20,99997 & -54,99995 & 143,9999 \end{vmatrix}$$

II . 6 . INVERSION PAR LA METHODE DES SOUS-MATRICES, D'UNE MATRICE A BLOCS DIAGONAUX M-MATRICES, BLOCS EXTRA-DIAGONAUX POSITIFS.

Soient A une matrice d'ordre n , d'inverse Z , et un partitionnement des deux matrices effectué selon le schéma ci-dessous :

$$A = \begin{array}{|c|c|} \hline B & S \\ \hline R & C \\ \hline \end{array} \quad \begin{matrix} \} p \\ \} m \end{matrix} \qquad Z = \begin{array}{|c|c|} \hline D & U \\ \hline T & E \\ \hline \end{array} \quad \begin{matrix} \} p \\ \} m \end{matrix}$$

Rappelons [9] , qu'avec les notations :

$Q = C - R\,B^{-1}\,S \qquad F = B^{-1}\,S \qquad G = R\,B^{-1}$

l'algorithme d'obtention de Z , sous la condition d'inversibilité de B et de Q , est :

$$\begin{aligned} E &= Q^{-1} \\ T &= -\,E\,G \\ U &= -\,F\,E \\ D &= B^{-1} - F\,T \end{aligned} \qquad \text{(VI.4)}$$

Définissant, si $A = (a_{ij})$ et $B = (b_{ij})$ sont 2 matrices éléments de $\mathcal{M}_{n,n}(S)$, les matrices $A \underset{\nabla}{\oplus} B$, $A \underset{\Delta}{\oplus} B$, $A \underset{\nabla}{\otimes} B$, $A \underset{\Delta}{\otimes} B$ par les égalités :

$$A \underset{\nabla}{\oplus} B = (a_{ij} \underset{\nabla}{\oplus} b_{ij}) \qquad A \underset{\Delta}{\oplus} B = (a_{ij} \underset{\Delta}{\oplus} b_{ij})$$

$$A \underset{\nabla}{\otimes} B = \Big(\underset{\nabla}{\oplus} \sum_{k=1}^{n} a_{ik} \underset{\nabla}{\otimes} b_{kj} \Big) \qquad A \underset{\Delta}{\otimes} B = \Big(\underset{\Delta}{\oplus} \sum_{k=1}^{n} a_{ik} \underset{\Delta}{\otimes} b_{kj} \Big)$$

on peut énoncer par exemple , $\tilde{\theta}_D$ et $\tilde{\theta}_G$ ayant les définitions du § II.4 :

<u>Proposition VI.8.</u>

Si dans le partitionnement de A les conditions suivantes sont réalisées :

R et S matrices positives, B et Q M-matrices,

alors, une G - suite de la suite d'opérations définissant $(D , E , -T , -U)$ sur $\mathbb{R}$ par (VI.4) est :

$\tilde{B}^{-1} = \mathcal{J}_{\tilde{\theta}_D, \tilde{\theta}_G}(B)$

$\tilde{Q} = C \underset{\Delta}{\ominus} R \underset{\nabla}{\otimes} \tilde{B}^{-1} \underset{\nabla}{\otimes} S \qquad \tilde{F} = \tilde{B}^{-1} \underset{\nabla}{\otimes} S \qquad \tilde{G} = R \underset{\nabla}{\otimes} \tilde{B}^{-1}$

$\tilde{E} = \mathcal{J}_{\tilde{\theta}_D, \tilde{\theta}_G}(\tilde{Q})$

$\tilde{T} = -\,\tilde{E} \underset{\nabla}{\otimes} \tilde{G}$

$\tilde{U} = -\,\tilde{F} \underset{\nabla}{\otimes} \tilde{E}$

$\tilde{D} = \tilde{B}^{-1} \underset{\nabla}{\oplus} \tilde{F} \underset{\nabla}{\otimes} (-\tilde{T})$

Un cas particulier où Q a la propriété de M - matrice est donné par le :

Lemme VI . 1

Si les conditions suivantes sont réalisées :

- la sous-matrice B de A est diagonale
- la matrice A est diagonalement dominante et vérifie :

$a_{ij} \leqslant 0 \quad i \neq j \quad , \quad p+1 \leqslant i,j \leqslant n$

$a_{ij} \geqslant 0$ sinon ,

alors Q définie en (VI.4) est une M - matrice.

Le lemme VI.1 s'établit en montrant que Q vérifie les deux propriétés suivantes :

1) $q_{ij} \leqslant 0 \qquad i \neq j$

2) Q est à diagonale positive dominante.

Ces 2 propriétés impliquent alors pour Q la propriété de M - matrice [7].

L'inégalité $q_{ij} \leqslant 0$ pour $i \neq j$ vient immédiatement.

D'autre part, de $q_{ij} = c_{ij} - \sum_{k=1}^{p} \frac{r_{ik}\ s_{kj}}{b_k}$ on déduit :

$$q_{ii} - \sum_{j \neq i} |q_{ij}| = c_{ii} - \sum_{\substack{j=1 \\ \neq i}}^{m} |c_{ij}| - \sum_{j=1}^{m} \sum_{k=1}^{p} \frac{r_{ik}\ s_{kj}}{b_k} =$$

$$c_{ii} - \sum_{\substack{j=1 \\ \neq i}}^{m} |c_{ij}| - \sum_{k=1}^{p} \frac{|r_{ik}|}{b_k} \sum_{j=1}^{m} |s_{kj}|$$

La matrice A étant à diagonale positive dominante :

$$b_k > \sum_{j=1}^{m} |s_{kj}| \qquad k = 1, \ldots, p$$

donc :

$q_{ii} - \sum_{j \neq i} |q_{ij}| > c_{ii} - \left(\sum_{\substack{j=1 \\ \neq i}}^{m} |c_{ij}| + \sum_{k=1}^{p} |r_{ik}| \right) > 0$ par hypothèse, d'où le résultat.

Relativement aux systèmes de nombres à virgule flottante, nous énoncerons :

Proposition VI . 9 .

Si

1) les conditions du lemme VI.1 sont réalisées

2) $S = T_b^s$, avec l'arithmétique suivante :

- l'opération $\oplus$ est définie par une troncature sans chiffre de garde, les nombres négatifs étant représentés avec signe et valeur absolue,

- les opérations $\otimes$ et $\oslash$ sont correctes et I - dirigées $\forall (X,Y) \in T^2$,

alors, les inverses Z et $\tilde{Z}$ de A , calculées par la méthode des sous-matrices, sur, respectivement $\mathbb{R}$ et T_b^s , vérifient :

$$|\tilde{z}_{ij}| \leq |z_{ij}| \qquad i,j = 1, \ldots , n$$

Cette proposition vient immédiatement : on remarquera en particulier que pour l'arithmétique considérée, l'hypothèse $c_{ij} \leq 0 \quad i \neq j$ est nécessaire pour conclure à $\tilde{q}_{ij} \geq q_{ij} \quad \forall\, i,j$.

II . 7 . REMARQUES

1) Les suites maximisantes d'opérations sur S que nous avons mises en évidence aux paragraphes II.1 à II.6 précédents, exclusion faite du paragraphe II.3 , définissent 2 schémas de calcul identiques aux schémas générés, pour les algorithmes considérés, par l'arithmétique d'intervalle définie par R.E. MOORE [29] :

a, b, c, d étant éléments de S :

$$[a,b] \oplus [c,d] = [\nabla(a+c) , \Delta(b+d)]$$

$$[a,b] \ominus [c,d] = [\nabla(a-d) , \Delta(b-c)]$$

$$[a,b] \otimes [c,d] = [\nabla(\min\{ac,ad,bc,bd\}) , \Delta(\max\{ac,ad,bc,bd\})]$$

et si $0 \notin [c,d]$:

$$[a,b] \oslash [c,d] = [\nabla(\min\{a/c,a/d,b/c,b/d\}) , \Delta(\max\{a/c,a/d,b/c,b/d\})]$$

Pour les algorithmes cités ci-dessus, les bornes, de la solution, calculées par l'arithmétique d'intervalle peuvent être éffectivement atteintes par le calcul sur S.

Il n'en est pas ainsi, sauf cas particulier de la valeur x_o initiale, pour l'algorithme d'obtention du nième itéré du calcul de la racine carrée d'un nombre par la méthode de Newton.

2) Nous donnons ci-dessous un exemple très simple de non - existence de G - suite d'opérations correctes sur S correspondant à une suite donnée d'opérations sur $\mathbb{R}$:

Pour $S = T^2_{10}$, $r = (x\,y - z\,u)\,(x\,y - v\,w)$, $x = 0{,}57$, $y = 0{,}63$, $z = 0{,}58$, $u = 0{,}62$, $v = 0{,}65$, $w = 0{,}55$

les valeurs $\bar{r}$ calculées par les différentes suites correctes d'opérations sur T^2_{10} ont été :

$$0 \quad \text{et} \quad 0{,}0001$$

alors que $r = -\,0{,}0000008$.

III - UN EXEMPLE DE SUITES NON MAXIMISANTES D'OPERATIONS SUR S GENERANT DES IMAGES ENCADRANT LA SOLUTION REELLE.

$N , M \in \mathbb{N}$, étant donnés un vecteur x de S^N et une suite $\mathcal{O}$ d'opérations sur $\mathbb{R}$, faisant correspondre à x un vecteur $r = (r_i)$ de $\mathbb{R}^M$, introduisons la définition suivante :

Définition VI.4.

Deux suites O^1 et O^2 d'opérations sur S , correspondant à $\mathcal{O}$, sont dites <u>suites d'encadrement</u> de $\mathcal{O}$ sur la partie $\mathcal{P}$ de S^N , si $\forall\, x \in \mathcal{P}$, les images r, r^1 et r^2 de x par $\mathcal{O}, O^1$ et O^2 satisfont :

$$r^1_i \leqslant r_i \leqslant r^2_i \quad \vee \quad r^1_i \geqslant r_i \geqslant r^2_i \qquad i = 1 , \ldots , M$$

La définition précédente s'applique lorsque $\mathcal{O}$ fait correspondre à x de S^N, une matrice de M lignes et P colonnes, par l'isomorphisme de $\mathcal{M}_{M,P}(\mathbb{R})$ sur $\mathbb{R}^{M\times P}$.

III.1. RESOLUTION D'UN SYSTEME TRIANGULAIRE - NOTATIONS

Soient $A = (a_{ij})$ une matrice triangulaire d'ordre n, non singulière, donnée sur S, et $b = (b_i)$ un vecteur donné de S^n. Supposons par exemple, A triangulaire inférieure, et considérons la résolution du système linéaire :

$$A\,x = b \qquad (VI.5)$$

et la suite $\mathcal{O}$ d'opérations sur $\mathbb{R}$ définissant x.

Désignons respectivement par $x = (x_i)$ la solution exacte de (VI.5), $\overline{x} = (\overline{x}_i)$ la solution calculée sur S, et $\varepsilon = (\varepsilon_i)$ le vecteur erreur $x - \overline{x}$.

Définissant pour $i = 1, \ldots, n$, les nombres e_i et $\overline{\overline{x}}_i$ suivants :

$$\overline{\overline{x}}_i = \left(b_i - \sum_{j=1}^{i-1} a_{ij}\,\overline{x}_j\right) / a_{ii} \qquad 1 \leqslant i \leqslant n$$

$$e_i = \overline{\overline{x}}_i - \overline{x}_i \qquad 1 \leqslant i \leqslant n$$

(e_i est l'erreur "générée au rang i" et soit $e = (e_i)$), rappelons l'égalité suivante vérifiée par $\overline{x}$ ([8] , [42]) :

$$A\,\overline{x} = b - (\mathrm{Diag}\ A)\,e$$

soit encore avec la notation $A^{-1} = C = (c_{ij})$:

$$\varepsilon_i = \sum_{j=1}^{i} c_{ij}\ a_{jj}\ e_j \qquad (VI.6)$$

III.2. DETERMINATION DE SUITES D'ENCADREMENT DE $\mathcal{O}$

III . 2 . 1 L'erreur e_i

L'expression de e_i en fonction des erreurs élémentaires de multiplication, d'addition et de division, entachant le calcul de $\overline{x}_i$ sur S, a été rappelée au § II.2.2 du chapitre III .

$\mathcal{O}_i$ étant la suite d'opérations sur $\mathbb{R}$ définissant $\overline{\overline{x}}_i$ par :

$$\overline{\overline{x}}_i = \left[b_i - \left(\sum_{j=1}^{i-1} a_{ij} \ \overline{x}_j \right) \right] / a_{ii}$$

deux suites maximisantes d'encadrement de $\mathcal{O}_i$ sont respectivement générées par les choix suivants d'opérations sur S :

$$C_E = \{ \otimes_\Delta , \oplus_\Delta , \ominus_\nabla , \mathcal{O} = \left| \begin{array}{ll} \mathcal{O}_\nabla & \text{si } a_{ii} > 0 \\ \mathcal{O}_\Delta & \text{si } a_{ii} < 0 \end{array} \right. \} , \ C_E' = \overline{C}_E$$

($C_E \longrightarrow a_{ii} \ e_i \geqslant 0$; $C_{E'} \Longrightarrow a_{ii} \ e_i \leqslant 0$)

III . 2 . 2 . Théorème

Proposition VI . 10

Si dans la résolution d'un système triangulaire $A\,x = b$ donné sur S , la matrice inverse de A , $C = A^{-1}$ est de la forme :

$$C = D \ P \ \Delta$$

où :

P est une matrice non négative

$D = (d_i)$ et $\Delta = (\delta_i)$ sont 2 matrices diagonales d'éléments égaux à 1 ou -1

alors, deux suites O^1 et O^2 d'encadrement de $\mathcal{O}$ sont définies par les choix C_i^1 et C_i^2 suivants, d'opérations sur S pour le calcul de $\overline{x}_i$, $i = 1 , \ldots , n$:

$$C_i^1 = \left| \begin{array}{ll} C_E & \text{si } \delta_i > 0 \\ C_{E'} & \text{si } \delta_i < 0 \end{array} \right. , \quad C_i^2 = \overline{C_i^1}$$

En effet, avec la notation $P = (p_{ij})$, (VI.6) s'écrit :

$$\varepsilon_i = \sum_{j=1}^{i} d_i \ p_{ij} \ \delta_j \ a_{jj} \ e_j = d_i \sum_{j=1}^{i} p_{ij} \ \delta_j \ a_{jj} \ e_j$$

ce qui avec :

$\delta_j \; a_{jj} \; e_j \geqslant 0 \quad \forall\, j \quad$ pour $\quad 0^1$

$\delta_j \; a_{jj} \; e_j \leqslant 0 \quad \forall\, j \quad$ pour $\quad 0^2$

implique la proposition.

III . 2 . 3 Exemple et contre-exemple

1) Considérons la résolution sur T^2_{10} du système triangulaire :

$$\begin{vmatrix} 3 & & & \\ -4 & 5 & & \\ 1 & -2 & 1 & \\ -2 & 1 & -1 & 1 \end{vmatrix} \quad \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{vmatrix} = \begin{vmatrix} 1 \\ 0 \\ 0 \\ 0 \end{vmatrix}$$

La condition d'application de la proposition VI.10 est vérifiée par la matrice A d'inverse A^{-1} non négative.

Les 2 suites 0^1 et 0^2 d'encadrement de $\mathcal{O}$, définies par cette proposition, calculent respectivement les solutions approchées :

$\bar{x}^1 = (\,0{,}33 \;;\; 0{,}26 \;;\; 0{,}19 \;;\; 0{,}59)$ et $\bar{x}^2 = (0{,}34 \;;\; 0{,}28 \;;\; 0{,}22 \;;\; 0{,}62)$

de la solution exacte du système $x = (0{,}3333.. \;;\; 0{,}2666.. \;.\; 0{,}2 \;;\; 0{,}6)$

On remarquera que 0^1 et 0^2 ne sont pas maximisantes relativement au calcul de la composante x_3 :

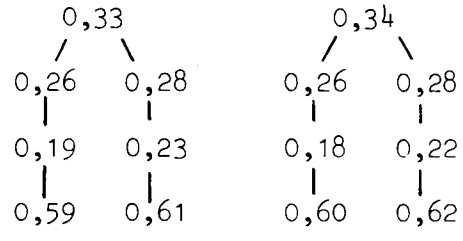
La résolution sur T^2_{10} , du système donné, par l'ensemble des suites correctes d'opérations, s'explicite en effet suivant :

$\bar{x}_1 = 1 \oslash 3$

$\bar{x}_2 = (4 \otimes \bar{x}_1) \oslash 5$

$\bar{x}_3 = -\bar{x}_1 \oplus 2 \otimes \bar{x}_2$

$\bar{x}_4 = 2 \otimes \bar{x}_1 \ominus \bar{x}_2 \oplus \bar{x}_3$



définissant 4 solutions approchées différentes $\bar{x}$ de x .

On constate que $\bar{x}^1$ satisfait à :

$$\| x - \bar{x}^1 \| = \min_{\bar{x}} \| x - \bar{x} \|$$

pour les 3 normes de Holder usuelles .

2) $\mathcal{O}$ étant la suite d'opérations sur $\mathbb{R}$ définie par :

$$\begin{vmatrix} 7 & & & & \\ 9 & -2 & & & \\ -2 & 1 & -1 & & \\ 1 & 0 & 0 & -1 & \\ 0 & 2 & -3 & 5 & -1 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{vmatrix} = \begin{vmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{vmatrix}$$

il n'existe pas, dans l'ensemble des suites correctes d'opérations sur T_{10}^2 correspondant à $\mathcal{O}$, de suites d'encadrement de $\mathcal{O}$.

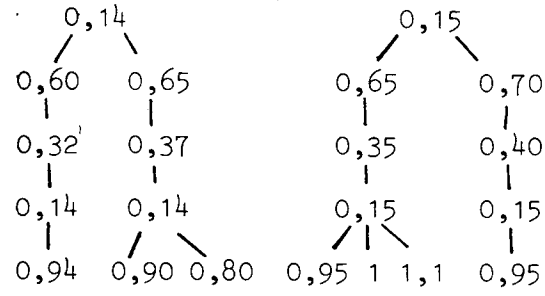
$\bar{x}_1 = 1 \oslash 7$

$\bar{x}_2 = (9 \otimes \bar{x}_1) \oslash 2$

$\bar{x}_3 = -2 \otimes \bar{x}_1 \oplus \bar{x}_2$

$\bar{x}_4 = \bar{x}_1$

$\bar{x}_5 = 2 \otimes \bar{x}_2 \ominus 3 \times \bar{x}_3 \oplus 5 \otimes \bar{x}_4$



la solution exacte du système étant :

x = (0,1428.. ; 0,6428.. ; 0,3571.. ; 0,1428.. ; 0,9285..)

III . 2. 4. Une autre formulation des hypothèses

Une forme différente de (VI.6) peut être donnée en exprimant les éléments de C en fonction des mineurs de A .

Notant Δ_{ij} le mineur correspondant à a_{ij} , $D(A)$ le déterminant de A , et $D(A_{\{j+1,\ldots,i\}\ ,\ \{j,\ldots,i-1\}})$ le déterminant de la sous-matrice de A formée par les lignes $j+1$, ... , i et les colonnes j, ... , $i-1$ pour $j \leq i - 1$:

$$c_{ii} = 1/a_{ii}$$

$$c_{ij} = \frac{(-1)^{i+j}\,\Delta_{ji}}{D(A)} = \frac{(-1)^{i+j}\ \prod_{l=1}^{j-1} a_{ll}\ D(A_{\{j+1,\ldots i\},\{j,\ldots,i-1\}})\ \prod_{l=i+1}^{n} a_{ll}}{D(A)} \qquad j<i$$

Soient :

M_{i-1} la matrice quasi-triangulaire d'ordre $i-1$, définie pour $i = 2, \ldots , n$ par :

$$M_{i-1} = \left(\frac{a_{lj}}{a_{ll}} \ ; \ l = 2 , \ldots , i \ ; \ j = 1, \ldots , i-1\right)$$

Δ^k_{i-1} le déterminant d'ordre k , défini pour $k = 1, \ldots , i-1$ par :

$$\Delta^k_{i-1} = \left|\frac{a_{lj}}{a_{ll}} \ ; \ l = i , \ldots , i-k+1 \ ; \ j = i-1 , .. , i-k\right|$$

(Δ^k_{i-1} est le kième mineur principal de la matrice N_{i-1} , déduite de M_{i-1} par symétrie centrale)

$$M_{i-1} = \begin{vmatrix} \frac{a_{21}}{a_{22}} & 1 & & \\ \frac{a_{31}}{a_{33}} & \frac{a_{32}}{a_{33}} & 1 & \\ \vdots & \vdots & & \\ \frac{a_{i1}}{a_{ii}} & \frac{a_{i2}}{a_{ii}} & \cdots\cdots\cdots & \frac{a_{i,i-1}}{a_{ii}} \end{vmatrix}$$

Une autre formulation de (VI.6) est donc :

$$\varepsilon_i = e_i + \sum_{j=i-1}^{1} (-1)^{i+j}\ \Delta^{i-j}_{i-1}\ e_j \qquad (VI.7)$$

et l'énoncé correspondant à la proposition VI.10 est, avec la notation

$$C_G = \{ \begin{matrix} \otimes_\Delta \ , \ \oplus_\Delta \ , \ \ominus_\nabla \ , \text{ si } a_{ii} > 0 \\ \otimes_\nabla \ , \ \oplus_\nabla \ , \ \ominus_\Delta \ , \text{ si } a_{ii} < 0 \end{matrix} , \ \emptyset_\nabla \ \} \ , \ C_D = \overline{C}_G \ :$$

Si dans la résolution d'un système triangulaire $A\,x = b$ donné sur S, la matrice quasi-triangulaire d'ordre $n-1$ définie par :

$$M_{n-1} = \left(\frac{a_{ij}}{a_{ii}} \; ; \; i = 2, \ldots, n \; ; \; j = 1, \ldots, n-1\right)$$

est de la forme :

$$M_{n-1} = D'\,P'\,\Delta'$$

où

$P' = (p'_{ij})$ est une matrice vérifiant $D(P'_{\{p,..,q\},\{p,...,q\}}) > 0 \quad 1 \leq p,q \leq n-1$

$D' = (d'_i)$ et $\Delta' = (\delta'_i)$ sont 2 matrices diagonales d'éléments égaux à 1 ou -1

alors, deux suites 0^1 et 0^2 d'encadrement de $\mathcal{O}$ sont définies par les choix C_i^1 et C_i^2 suivants, d'opérations sur S pour le calcul de $\bar{x}_i$, $i = 1, \ldots, n$:

$C_1^1 = C_G$

$$C_i^1 = \begin{cases} C_G & \text{si } (-1)^{i-1} \prod_{j=1}^{i-1} d'_j\,\delta'_j > 0 \\ C_D & \text{" } \quad \text{" } \quad < 0 \end{cases} \qquad i>1$$

, $C_i^2 = \bar{C}_i^1$

III . 2 . 5 . Cas où les suites d'encadrement de $\mathcal{O}$ sont maximisantes.

Un cas particulier où les 2 suites d'encadrement de $\mathcal{O}$ définies précédemment sont maximisantes, est donné par la :

Proposition VI . 11

Si tous les termes du développement de $D(P'_{\{p,\ldots,q\},\{p,\ldots,q\}})$ $(1 \leq p,q \leq n-1)$ sont de même signe, alors 0^1 et 0^2 sont 2 suites maximisantes de $\mathcal{O}$.

Il suffit de raisonner pour $M_{n-1} = P'$.

$\bar{x} = (\bar{x}_i)$ désignant une solution du système calculée sur S par une suite 0 d'opérations correctes correspondant à $\mathcal{O}$, la double inégalité :

$$\begin{cases} \bar{x}_i^1 \leqslant \bar{x}_i \leqslant \bar{x}_i^2 & i \text{ impair} \\ \bar{x}_i^1 \geqslant \bar{x}_i \geqslant \bar{x}_i^2 & i \text{ pair} \end{cases} \qquad \text{(VI.8)}$$

s'établit par récurrence sur l'indice i .

Pour $i = 1$, (VI.8) est satisfaite.

Pour $i = 2$, les hypothèses impliquent $\frac{a_{21}}{a_{22}} > 0$ ce qui, avec les définitions de O^1 et O^2 , assure (VI.8) .

Supposant alors :

$$\begin{cases} \dfrac{a_{ij}}{a_{ii}} \text{ du signe de } (-1)^{i+j-1} & (2 \leqslant i \leqslant l \ \wedge \ 1 \leqslant j \leqslant i-1) \\ \text{(VI.8) vérifiée } \forall\, i \leqslant l \end{cases}$$

on vérifie que ces 2 propriétés sont encore satisfaites pour $i = l + 1$.

Le développement, par rapport à sa dernière ligne de $D(M_l)$:

$$D(M_l) = (-1)^{l+1} \frac{a_{l+1,1}}{a_{l+1,l+1}} + \sum_{j=2}^{l} (-1)^{l+j} \frac{a_{l+1,j}}{a_{l+1,l+1}} D(M_{j-1})$$

assure la 1ère propriété.

Ce résultat, les définitions de O^1 et O^2 et la 2ème hypothèse de récurrence, impliquant alors (VI.8) pour $i = l + 1$.

Remarque : La matrice triangulaire A satisfaisant $M_{n-1} = \left(\frac{a_{ij}}{a_{ii}}\right) = P'$ avec tous les termes du développement de $D(P'_{\{p,\ldots q\},\{p,\ldots,q\}})$ $(1 \leqslant p,q \leqslant n-1)$ de même signe, vérifie :

$$Z = D\ A\ \Delta$$

où :

Z est une M - matrice triangulaire

$D = (d_i)$ et $\Delta = (\delta_i)$ sont 2 matrices diagonales d'élément égaux à :

$d_i = (-1)^i \times \text{signe } (a_{ii})$ $\qquad$ $\delta_i = (-1)^i$

Les 2 suites O^1 et O^2 sont maximisantes dans le cas où A est une M - matrice, résultat que nous rapprocherons des résultats énoncés au § II.4 (Résolution d'un seul système triangulaire donné sur S à second membre quelconque dans ce paragraphe, et résolution successive, après décomposition L R d'une matrice donnée sur S , de deux systèmes triangulaires à second membre de signe constant dans le § II.4.2).

III . 2 . 6 . Exemple numérique

Nous avons considéré la résolution du système $A\,x = b$ pour :

$$A = \begin{vmatrix} 3 & & & & & & & \\ 7 & 5 & & & & & & \\ 1 & 5 & 3 & & & & & \\ 1 & 1 & 12 & 7 & & & & \\ 3 & 2 & 2 & 16 & 7 & & & \\ 3 & 5 & 5 & 1 & 20 & 5 & & \\ 1 & 2 & 1 & 3 & 2 & 15 & 3 & \\ 2 & 0 & 1 & 3 & 0 & 1 & 10 & 1 \end{vmatrix} \qquad b = \begin{vmatrix} 1 \\ 4 \\ 3 \\ 7 \\ 10 \\ 13 \\ 9 \\ 6 \end{vmatrix}$$

Les éléments de A vérifient

$$a_{i,i-1} > \sum_{\substack{j=1 \\ \neq i-1}}^{i} |a_{ij}| \qquad i = 2, \ldots, n$$

Par le théorème de Gerschgorin-Hadamard, les valeurs propres réelles de toute sous-matrice $M_{\{p,..,q\},\{p,..,q\}}$ $(1 \leqslant p,q \leqslant n-1)$ de M_{n-1} , sont positives.

M_{n-1} est une matrice de type P'.

La solution exacte X du système, et les solutions X^1 et X^2 calculées sur T_{16}^6 par les 2 suites d'opérations O^1 et O^2 sont respectivement ($\overline{X}$ est la solution calculée par l'arithmétique IBM 360) :

$$X = \begin{vmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 1/3 \\ 1/3 \\ 1/3 \\ 1/3 \\ 1/3 \end{vmatrix} \quad X^1 = \begin{vmatrix} 0,3333333 \\ \text{..} \ 34 \\ \text{..} \ 27 \\ \text{..} \ 47 \\ \text{..} \ 03 \\ \text{..} \ 464 \\ \text{..} \ 2682 \\ \text{..} \ 9691 \end{vmatrix} \quad X^2 = \begin{vmatrix} 0,3333334 \\ \text{..} \ 2 \\ \text{..} \ 40 \\ \text{..} \ 22 \\ \text{..} \ 59 \\ \text{..} \ 225 \\ \text{..} \ 871 \\ 28085 \end{vmatrix} \quad \tilde{X} = \begin{vmatrix} 0,3333333 \\ \text{..} \ 4 \\ \text{..} \ 6 \\ \text{..} \ 0 \\ \text{..} \ 41 \\ \text{..} \ 05 \\ \text{..} \ 476 \\ \text{..} \ 1947 \end{vmatrix}$$

IV - THEOREMES GENERAUX

Les théorèmes qui suivent sont relatifs à l'ensemble des 2^{N-1} images que les suites d'opérations correctes sur S, correspondant à la suite $\sigma = \{ *_i , i=1 , \dots , N-1 \}$ d'opérations sur $\mathbb{R}$, associent à une même suite finie de N opérandes donnés sur S.

Dans le cas de la sommation, c'est-à-dire $\sigma = \{+ , i=1 , \dots , N-1\}$, on peut énoncer :

Proposition VI . 12 .

Etant donné le vecteur $X = (x_1, \dots , x_N)$ de $(S^+)^N$, si S est un sous-ensemble de $\mathbb{R}$ à intervalles non décroissants, alors :

- L'ensemble E, des images de X par les 2^{N-1} suites O d'additions correctes sur S, $O = \{ \oplus_{\square_i} , i = 1 , \dots , N-1 ; \square_i = \nabla , \Delta \}$,

est formé d'au plus N éléments consécutifs de S.

- Lorsque Card E = N, le $j^{\text{ième}}$ élément s_j de E vérifie :

$$n(s_j) = \text{Card} \{ O \mid i_O (X) = s_j \} = C_{N-1}^{j-1} \quad j = 1, \dots , N$$

Démonstration :

Il suffit de raisonner par récurrence.

Pour la 1ère partie de la proposition VI. 12. :

. La propriété est vraie au rang 2

. Supposons la vérifiée au rang k , c'est-à-dire supposons que l'ensemble E_k des images de $X_k = (x_1 , \dots , x_k)$ par les 2^{k-1} suites d'additions correctes sur S , soit formé de λ éléments consécutifs de S : $s_1^k , \dots , s_\lambda^k$ avec $\lambda \leqslant k$.

Lorsqu'on forme les nombres $s_i^k + x_{k+1}$, il vient :

$$\mu = \text{Card} \{ z \mid z \in \left[s_1^k + x_{k+1} , s_\lambda^k + x_{k+1} \right] \wedge z \in S \} \leqslant \lambda ,$$

l'égalité $\mu = \lambda$ impliquant $s_1^k + x_{k+1} \in S \wedge s_\lambda^k + x_{k+1} \in S$.

Dans ce dernier cas, l'égalité $\text{Card } E_{k+1} = \lambda$, et dans le cas général, l'inégalité $\text{Card } E_{k+1} \leqslant \mu + 2 \leqslant \lambda + 1$ assurent bien la propriété au rang $k + 1$, les éléments de E_{k+1} étant, par ailleurs, consécutifs sur S .

La 2ème partie de la proposition s'établit alors :

. Pour $N = 2$ et $\text{Card } E_2 = 2$, on a bien $n(s_1^2) = n(s_2^2) = 1 = C_1^0 = C_1^1$

. Supposons alors $\text{Card } E_k = k$ et, $n(s_j^k) = C_{k-1}^{j-1}$ $(j = 1, \dots , k)$.

(L'égalité $\text{Card } E = N$ entraîne en effet, $\text{Card } E_k = k , \forall\, k$, d'après l'inégalité $\text{Card } E_{k+1} \leqslant \text{Card } E_k + 1$ obtenue précédemment).

Raisonnant par l'absurde, et utilisant à nouveau la propriété de non décroissance des longueurs $z_{i+1} - z_i$ pour l'ensemble des éléments z_i de S^+ classés dans l'ordre naturel, on établit : $\text{Card } E_{k+1} = k + 1$ implique qu'il existe un et un seul élément de S , s_j^{k+1} dans tout intervalle $\left[s_{j-1}^k + x_{k+1} , s_j^k + x_{k+1} \right[$, $j = 2, \dots , k$ et qu'aucun des nombres $s_j^k + x_{k+1}$ n'est élément de S .

On a alors immédiatement :

$$n\,(s_j^{k+1}) = n\,(s_{j-1}^{k}) + n\,(s_j^{k}) = C_{k-1}^{j-2} + C_{k-1}^{j-1} = C_k^{j-1} \qquad j = 2, \ldots, k$$

$$n\,(s_1^{k+1}) = n\,(s_1^{k}) = 1$$

$$n\,(s_{k+1}^{k+1}) = n\,(s_k^{k}) = 1$$

On peut, pour une somme algébrique, énoncer :

Proposition VI . 13 .

Etant donnés le vecteur $X = (x_1, \ldots, x_N)$ de S^N et l'ensemble E des images de X par les suites O d'additions correctes sur S, $O = \{ \oplus_{\square_i}, i = 1, \ldots, N-1 ; \square_i = \nabla \text{ ou } \Delta \}$, alors si la condition suivante est réalisée :

$$\varepsilon\Big(\oplus_{\nabla} \sum_{j=1}^{i} x_j\Big) = \varepsilon\Big(\oplus_{\Delta} \sum_{j=1}^{i} x_j\Big) = \text{constante indépendante de } i \quad \forall\, i$$

E est formé d'éléments consécutifs de S et, avec la notation $P = \text{Card } E$, le $j^{\text{ième}}$ élément s_j de E vérifie :

$$n(s_j) = \text{Card}\,\{ O \mid i_O(X) = s_j \} = 2^{N-P}\, C_{P-1}^{j-1} \qquad j = 1, \ldots, P$$

Démonstration :

Raisonnons encore par récurrence :

. La propriété est vraie pour $N = 2$

. Supposons la vérifiée au rang k, c'est-à-dire, en reprenant les notations de la démonstration précédente, supposons que E_k soit formé d'éléments consécutifs de S, et que, si $\lambda = \text{Card } E_k$, on ait :

$$n(s_j^k) = 2^{k-\lambda}\, C_{\lambda-1}^{j-1} \qquad j = 1, \ldots, \lambda$$

Les nombres $s_j^k + x_{k+1}$ ayant le même intervalle de précision que les nombres s_j^k, deux possibilités seules peuvent se présenter :

1) $s_j^k + x_{k+1} \in S \qquad \forall\, j = 1, \ldots, \lambda$

alors, Card $E_{k+1} = \lambda$ et $n(s_j^{k+1}) = 2\, n\,(s_j^k)$

2) il existe un et un seul élément de S intérieur à tout intervalle $\left[s_{j-1}^k + x_{k+1}\ ,\ s_j^k + x_{k+1}\right]$

alors, Card $E_{k+1} = \lambda + 1$ et la propriété vient immédiatement par :

$$n(s_j^{k+1}) = n(s_{j-1}^k) + n(s_j^k) \qquad j = 2, \ldots, \lambda$$

Il vient aussi :

Proposition VI . 14 .

Etant donnés le vecteur $X = (x_1, \ldots, x_N)$ de $(T_b^s)^N$, une suite O de $N-1$ additions correctes sur T_b^s et la suite $\bar{O}$ conjuguée de O , alors si la condition suivante est réalisée :

$$\varepsilon\Big(\bigoplus_{\nabla} \sum_{j=1}^{i} x_j\Big) = \varepsilon\Big(\bigoplus_{\Delta} \sum_{j=1}^{i} x_j\Big) \leqslant \varepsilon\Big(\sum_{j=1}^{i-1} x_j\Big) \qquad \forall\, i \geqslant 2$$

on peut affirmer :

$i_O(X) + i_{\bar{O}}(X) =$ constante indépendante de O .

Avec la notation $X_k = (x_1, \ldots, x_k)$, supposons qu'au rang $k < n$, toute suite O_k de $k-1$ additions correctes vérifie :

$$i_{O_k}(X_k) + i_{\bar{O}_k}(X_k) = C \text{ indépendante de } O_k$$

Les conditions énoncées dans la proposition :

$$\varepsilon(i_{O_k}(X_k)) = \alpha \text{ indépendante de } O_k$$

$$\varepsilon(i_{O_k}(X_k) + X_{k+1}) = \beta \leqslant \alpha \qquad \forall\, O_k$$

impliquent que tous les nombres $i_{O_k}(X_k) + X_{k+1}$ sont les images par une translation d'amplitude e avec $0 \leqslant e < \beta$ d'éléments de T_b^s .

Si $e \neq 0$, on a :

$$i_{O_k}(X_k) \oplus x_{k+1} = i_{O_k}(X_k) + x_{k+1} + \begin{cases} e & \text{si } \oplus = \oplus_\Delta \\ -(\beta - e) & \text{si } \oplus = \oplus_\nabla \end{cases}$$

et par suite :

$$i_{O_{k+1}}(X_{k+1}) + i_{\bar{O}_{k+1}}(X_{k+1}) = C + 2x_{k+1} + 2e - \beta$$

Pour le produit scalaire, on démontre des théorèmes analogues aux théorèmes précédents, ainsi :

Si S est un sous-ensemble de $\mathbb{R}$ à intervalles non décroissants, l'ensemble des images des vecteurs $X = (x_1,\ldots,x_N)$ et $Y = (y_1,\ldots,y_N)$ de $(S^+)^N$ par les 2^{2N-1} produits scalaires "corrects" différents, est formé d'au plus $2N$ éléments consécutifs de S .

Enfin, nous signalerons que si S est un sous-ensemble de $[-1,1]$ à intervalles constants, la proposition VI . 12 est vérifiée pour tout vecteur de S^N lorsqu'on remplace l'opération $+$ par l'opération $\times$. On a aussi :

Si S est un sous-ensemble de $[-1,1]$ à intervalles constants, si $\mathcal{A}$ est un algorithme ne comportant que des opérations élémentaires d'addition, de soustraction ou de multiplication, l'ensemble des solutions de $\mathcal{A}$ calculées sur S par des opérations correctes, est formée d'éléments consécutifs de S .

(Exemple : Résolution d'un système triangulaire à éléments diagonaux unité).

V - CONCLUSION

La détermination sur T_b^S de suites d'encadrement, de la suite $\mathcal{G}$ d'opérations sur $\mathbb{R}$ associée à un algorithme, n'est pas aisée en général. De plus, l'existence de telles suites n'est pas assurée.

Les résultats énoncés dans ce chapitre auront essentiellement leur application :

- dans la connaissance de la stabilité numérique du problème étudié, lorsque les suites d'encadrement de $\mathcal{G}$ sont maximisantes.
- dans la connaissance, pour certaines arithmétiques flottantes courantes, des positions respectives des solutions calculées sur $\mathbb{R}$ et T .

CHAPITRE VII

N E E T U D E S U R L A R E P A R T I T I O N D E L ' E R R E U R D ' A R R O N D I

INTRODUCTION

Nous considérons dans ce chapitre des systèmes de nombres à virgule flottante.

Dans la première partie, est étudiée l'hypothèse de l'uniforme répartition de l'erreur d'arrondi élémentaire sur son domaine de définition, lorsqu'on considère cette erreur comme variable aléatoire :
En arithmétique à virgule flottante de base b et de nombre de chiffres de mantisse s, cette hypothèse généralement admise s'exprime, avec les notations suivantes :

$$x \in T_b^s \quad , \quad y \in T_b^s \quad , \quad p = e(x \star y) \quad , \quad E = x \star y - (x \circledast y)$$

en écrivant que E, considérée comme variable aléatoire, est uniformément répartie :

- dans l'intervalle $]-b^{p-s}, b^{p-s}[$ si $\circledast$ est correcte
- dans l'intervalle $[0, b^{p-s}[$ si $\circledast = \circledast_\nabla$

Nous proposons quelques remarques sur cette hypothèse, tenant compte de la répartition discrète - et non continue - des nombres.

Dans la seconde partie de ce chapitre, est mentionné le problème du meilleur choix, de la base d'une représentation des nombres en virgule flottante, relativement au critère de la précision.

I - FONCTION DE REPARTITION DE L'ERREUR D'ARRONDI CONSIDEREE COMME VARIABLE ALEATOIRE

I . 1 . FORMULATION DU PROBLEME

Nous étudierons l'hypothèse d'équirépartition sur son domaine de définition, de l'erreur d'arrondi entachant une opération élémentaire en arithmétique à virgule flottante, en considérant le problème suivant :

Etant donnés un système $T_b^s = T$ de nombres à virgule flottante, une opération $\circledast$ appartenant à l'ensemble $\{ \oplus , \ominus , \otimes , \oslash \}$, et un exposant p , on désigne par $\overline{\Sigma}$ l'ensemble :

$$\overline{\Sigma} = \{(x,y) \mid (x,y) \in T^2 \wedge e(x \circledast y) = p \}$$

Soit à étudier l'image $\overline{\mathcal{E}}$ de $\overline{\Sigma}$ par l'application de T^2 dans $\mathbb{R}$:

$$(x,y) \rightarrow E = x \star y - (x \circledast y) .$$

Nous traiterons ce problème pour l'addition et la soustraction arithmétiques, dans le cas de la TRONCATURE avec ou sans chiffre de garde. Cette règle, souvent utilisée, est en effet mal connue du point de vue que nous adoptons (influence des cas de cancellation, dissymétrie apportée à l'erreur par la présence du chiffre de garde dans l'addition de 2 nombres de signes contraires).

Le cas de la multiplication est abordé.

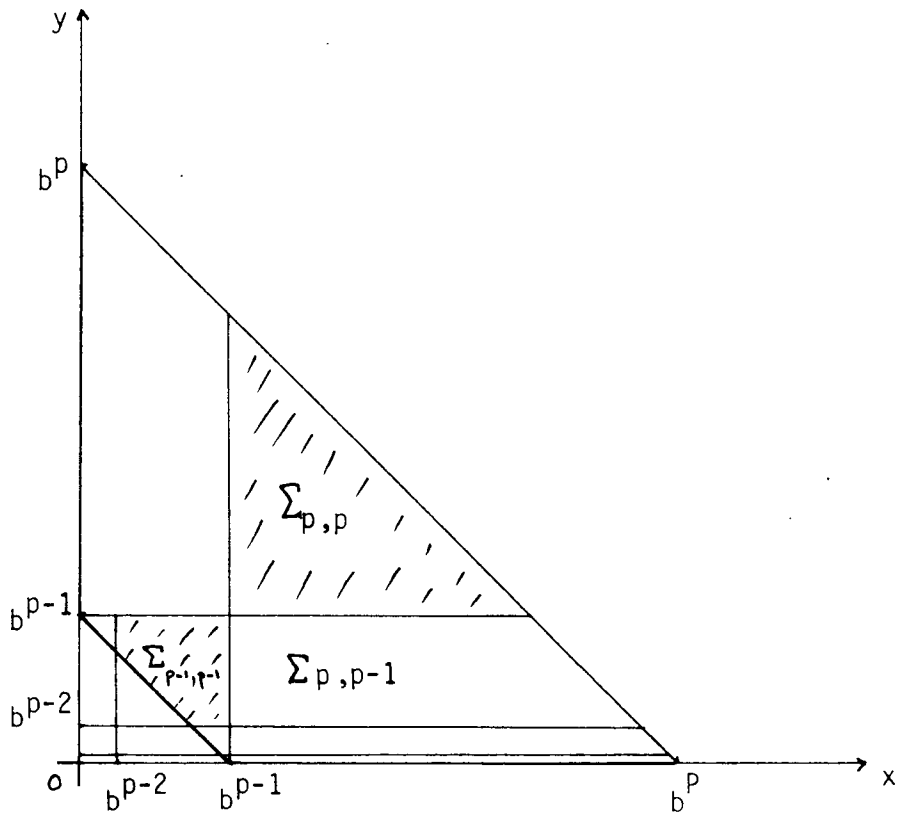
En raison des symétries, pour une représentation des nombres par signe et valeur absolue ou en notation de complément à 2, nous travaillerons avec la représentation des nombres par signe et valeur absolue, et nous considérerons uniquement les couples (x,y) tels que $x \star y$ est positif. Nous noterons respectivement Σ et $\mathcal{E}$ les sous-ensembles de $\overline{\Sigma}$ et $\overline{\mathcal{E}}$ considérés.

I . 2 . L'ADDITION ARITHMETIQUE

Dans le cas d'une règle de troncature, les chiffres de garde, même s'ils existent, ne participent pas à l'opération d'addition de deux nombres de même signe. Un seul cas est donc à considérer dans ce paragraphe.

CARACTERISATION DES ENSEMBLES Σ ET $\mathfrak{E}$.

Σ est l'ensemble des points de T^2 intérieurs au trapèze défini par les inégalités $\begin{cases} b^{p-1} \leqslant x + y < b^p \\ x \geqslant 0 \quad , \quad y \geqslant 0 \end{cases}$ où seule la grande base est exclue.



Pour la caractérisation de $\mathcal{E}$:

Posons $e(x) \geqslant e(y)$.

En remarquant qu'une condition nécessaire, pour que l'exposant de $x \oplus y$ prenne la valeur fixée p , est $e(x) = p \vee p-1$, étudions ces deux possibilités :

1) $e(x) = p$

Représentons x et y sous la forme :

$$x = l\, b^{p-s}$$

$$b^{s-1} \leqslant l, m < b^s \qquad (\text{ou } m = 0)$$

$$y = m\, b^{q-s}$$

E se déduit de (x,y) par :

$$E = r\, b^{q-s}$$

où r est défini par la division arithmétique $m = b^{p-q} m' + r \wedge r < b^{p-q}$.

La partie Σ_p de Σ à considérer ici, est l'union des $s+1$ domaines $\Sigma_{p,q}$ définis,

pour $q = p , p-1 , \ldots , p-s+1$ par :

$$\Sigma_{p,q} = \{ (x,y) \mid (x,y) \in T^2 \wedge e(x) = p \wedge e(y) = q \wedge x + y < b^p \}$$

pour $q = p - s$ par :

$$\Sigma_{p,p-s} = \{(x,y) \mid (x,y) \in T^2 \wedge e(x) = p \wedge e(y) \leqslant p - s \}$$

Parcourons Σ_{pq} $(p - s < q \leqslant p)$ sur les droites $x =$ constante. Les entiers m caractérisant les ordonnées des éléments de Σ_{pq} situés sur la droite $x =$ constante $= l\, b^{p-s}$ (avec $(b^s - 1)\, b^{p-q} > b^{s-1}$) sont les entiers successifs vérifiant l'inégalité :

$$b^{s-1} \leqslant m < \min (b^s, (b^s - 1)\, b^{p-q})$$

c'est-à-dire une suite d'entiers successifs compris entre deux multiples de b^{p-q} , le premier multiple seul étant inclus.
Le parcours de $\sum_{pq}$ à l fixé implique donc l'obtention avec la même fréquence de chacune des valeurs de r .

L'image $\mathcal{E}_{pq} = \{ r\, b^{q-s} \mid r \in \mathbb{N} \wedge 0 \leqslant r < b^{p-q} \}$ de $\sum_{pq}$ admet donc des éléments de valeurs équiprobables si l'on affecte une probabilité uniforme à l'ensemble des points de T^2 .
L'image $\mathcal{E}_p$ de $\sum_p$ est l'union :

$$\left(\bigcup_{q=p}^{p-s+1} \mathcal{E}_{pq} \right) \cup \{ y \mid y \in T \wedge y < b^{p-s} \}$$

2) $e(x) = p - 1$

$$x = l\, b^{p-1-s}$$

$$y = m\, b^{q-s}$$

Par la division arithmétique $b^{p-q-1}\, l + m = b^{p-q}\, m' + r \wedge r < b^{p-q}$, on a la représentation de E sous la forme $E = r\, b^{q-s}$.

Ayant défini pour $q = p - 1, \ldots , p - s$ les domaines :

$$\sum_{p-1,q} = \{(x,y) \mid (x,y) \in T^2 \wedge e(x) = p - 1 \wedge e(y) = q \wedge x + y \geqslant b^{p-1}\},$$

parcourons ces domaines sur les droites $x =$ constante.

Le parcours de $\sum_{p-1,q}$ à l fixé implique

* si $b^s(1 - b^{q-p}) < l < b^s$ l'équifréquence des restes
* si $b^s(1 - b^{q-p+1}) < l \leqslant b^s(1 - b^{q-p})$ l'obtention, dès que ρ défini par :

$$l = b\, l' + \rho \wedge p < b$$

est non nul, d'une majoration de 1 pour la fréquence des restes

$0 , 1 , 2 , \ldots , \rho\, b^{p-1-q} - 1$ sur la fréquence uniforme.

L'image de $\sum_{p-1} = \bigcup_{q=p-1}^{p-s} \sum_{p-1,q}$ est $\mathcal{E}_{p-1} = \bigcup_{q=p-1}^{p-s} \mathcal{E}_{p-1,q}$

VALEUR MOYENNE DE E

Les hypothèses : Considérer E comme variable aléatoire implique une hypothèse sur la probabilité d'obtention des différents points de T^2. Nous choisirons l'hypothèse d'équiprobabilité des points de T^2 . Cette hypothèse de première approximation serait, pour une étude plus fine, à remplacer [12] par l'hypothèse, réalisant une meilleure approche du schéma réel, d'une distribution logarithmique des mantisses des points de T d'exposant fixé (Voir aussi [38]).
Soulignons toutefois que dans ce dernier cas, l'obtention des restes sur une droite x = constante s'effectue avec des différences négligeables de probabilité de ces différents restes pour les ensembles $\sum_{p,q}$, $\sum_{p-1,q}$ lorsque q est proche de p ; par contre, pour l'ensemble $\{ y \mid y \in T_{\wedge} y < b^{p-s} \}$ il existe une différence significative des valeurs moyennes calculées suivant les deux hypothèses.

Si le calcul qui suit - valeur moyenne de E quand (x,y), de composantes positives, parcourt $\sum$ - peut paraître d'un intérêt limité, dans le cas où les nombres sont représentés avec signe et valeur absolue, il n'en est pas de même dans le cas où les nombres négatifs sont représentés en notation de complément à 2:

Pour l'ensemble des couples (x,y) satisfaisant à $e(x \oplus y) = p$ avec des composantes de signes quelconques , il vient en effet, pour l'erreur entachant l'addition définie par une règle de troncature, une moyenne nulle dans le 1er cas, la symétrie étant un automorphisme de T .
Le calcul ci-dessous s'applique donc uniquement aux sommes arithmétiques dans la première hypothèse ; joint aux calculs du paragraphe suivant, il s'applique aux sommes algébriques dans la seconde hypothèse.

Le dénombrement des éléments des différents sous-ensembles entraîne :

$$\text{Card } (\sum_{p,p}) = \frac{1}{2}(b^s - 2\,b^{s-1})\,(b^s - 2\,b^{s-1} + 1) \sim \frac{1}{2}\,b^{2s-2}\,(b-2)^2$$

et pour $t = 1\ ,\ \ldots\ ,\ s-1$:

$$\text{Card } (\sum_{p,p-t}) = (b^s - b^{s-1})\,(b^s - b^{s-t} - b^{s-1}) + \frac{1}{2}\,(b^s - b^{s-1})\,(\frac{b^s - b^{s-1}}{b^t} + 1)$$

$$\sim b^{2s-2}(b-1)\,(b - 1 - b^{1-t}) + \frac{1}{2}\,b^{2s-2-t}\,(b-1)^2$$

où

le 1er terme représente le nombre d'éléments de la partie rectangulaire de $\sum_{p,p-t}$ et le second terme le nombre d'éléments de la partie triangulaire de cet ensemble.

$$\text{Card } (\sum_{p-1,p-1}) = (b^s - b^{s-1})^2 - \frac{1}{2}\,(b^s - 2\,b^{s-1})\,(b^s - 2\,b^{s-1} + 1)$$

$$\sim b^{2s-2}\,(\frac{b^2}{2} - 1)$$

et pour $t = 2\ ,\ \ldots\ ,\ s$

$$\text{Card } (\sum_{p-1,p-t}) = (b^s - b^{s-1})\,(b^{s-t} - 1) + \frac{1}{2}\,(b^s - b^{s-1})\,(\frac{b^s - b^{s-1}}{b^{t-1}} + 1)$$

$$\sim \frac{1}{2}\ b^{2s-1-t}(b^2-1)$$

Nous pouvons remarquer que l'existence des restes en surnombre associés aux domaines $\sum_{p-1,q}$ est négligeable.

Ecrivons la valeur moyenne de E sous la forme :

$$\overline{E} = \frac{1}{2}\ b^{p-s}\ (1 - \sum_{t=o}^{s} \alpha_t)$$

avec :

$$\alpha_o = \frac{\text{Card } \Sigma_{pp}}{\text{Card } \Sigma} \quad , \quad \alpha_1 = b^{-1} \frac{\text{Card}\Sigma_{p-1,p-1} + 2\ \text{Card}\Sigma_{p,p-1}}{\text{Card } \Sigma}$$

$$\alpha_t = 2\ b^{-t} \frac{\text{Card } \Sigma_{p,p-t} + \text{Card } \Sigma_{p-1,p-t}}{\text{Card } \Sigma} \qquad 1 < t < s$$

Le biais apporté par l'ensemble $\mathcal{E}_{p,p-s}$ est fonction de la distance de l'exposant $p - s$ à m_2. Les calculs ci-dessous ne considèrent pas cet ensemble.

Il vient :

$$\text{Card } (\Sigma - \Sigma_{p,p-s}) \sim (2s - 1)\ (b - 1)^2\ b^{2s-2}$$

et:

$$\tilde{E}_{\mathcal{E}-\mathcal{E}_{p,p-s}} = \frac{1}{2}\ b^{p-s}(1-\omega) \quad \text{où} \quad \omega = \frac{b^2+b-2}{2(b-1)^2(2s-1)} \qquad \text{(VII.1)}$$

On constate que la différence, entre les moyennes théorique et pratique, est faible.

Ainsi :

pour $b = 2$, $s = 22$ $\qquad \omega = 0{,}046$

et

pour $b = 16$, $s = 6$ $\qquad \omega = 0{,}054$

Lemme VII.1.

Pour l'ensemble des couples (x,y) de T^2 satisfaisant aux conditions :

$$e(x \oplus y) = p \wedge x > 0 \wedge y > 0 \wedge e(x) - e(y) < s$$

la moyenne de l'erreur entachant l'addition , définie par une règle de troncature, des deux composantes du couple ,est égale à $\frac{1}{2}b^{p-s}(1-\omega)$ où ω a la valeur définie par (VII.1)

I . 3 . L'ADDITION DE DEUX NOMBRES DE SIGNES CONTRAIRES

Nous supposerons $x > o$, $y \leqslant o$.

I . 3 . 1 . Le cas de la troncature sans chiffre de garde

Dans ce cas, l'opération $\oplus$ est E-dirigée et non correcte. La valeur absolue de l'erreur d'arrondi élémentaire n'est évidemment plus contenue dans l'intervalle $[\, o, b^{p-s}[$. Il est cependant intéressant d'évaluer l'importance relative des cas de "cancellation".

CARACTERISATION DE Σ ET $\mathcal{E}$

Σ est l'ensemble des points $(x = l\, b^{p+t-s}$ $t \geqslant o$, $y = -\, m\, b^{q-s})$ de T^2 dont les coordonnées vérifient l'inégalité :

$$b^{s-1} \leqslant (l - \lfloor m\, b^{q-p-t} \rfloor)\, b^t < b^s \qquad \text{(VII . 2)}$$

Définissant les domaines $\Sigma_{p+t,q}$ suivants :

$$\Sigma_{p+t,q} = \{(x,y) \mid x \in T^+_\wedge \; y \in T^-_\wedge \; e(x) = p+t \wedge e(y) = q \wedge b^{p-1} \leqslant x \oplus y < b^p\} \quad (q > p-s) ,$$

avec dans le cas $t = o$ $q = p-s$

$$\Sigma_{p,p-s} = \{(x,y) \mid x \in T^+_\wedge \; y \in T^-_\wedge \; e(x) = p \wedge e(y) \leqslant p-s \}$$

on a :

$$\Sigma = \bigcup_{t=o}^{s-1} \Big(\bigcup_{q \in I_t} \Sigma_{p+t,q} \Big)$$

les ensembles I_t d'indices étant respectivement :

$I_o = \{p , p-1 , \ldots p-s\}$

$I_1 = \{p+1 , \ldots , p+2-s\}$

$I_t = \{p+t , p+t-1\}$ pour $t = 2 , \ldots , s-1$

L'ensemble $\mathcal{E}_{p+t,q}$ se déduit de $\Sigma_{p+t,q}$ par les égalités :

$$E = - r\, b^{q-s}$$

où r est défini par la division $m = b^{p+t-q}\, m' + r \wedge r < b^{p+t-q}$

De (VII.2) on déduit immédiatement que $\Sigma_{p+t,q}$ est l'ensemble des points $(x = l\, b^{p+t-s}\ ,\ y = - m\, b^{q-s})$ de T^2 dont les coordonnées vérifient :

$$\max\,(b^{s-1}\ ,\ b^{p+t-q}(1 - b^{s-t} + 1)) \leqslant m \leqslant \min\,(b^{s}-1, b^{p+t-q}(1 - b^{s-1-t}) + b^{p+t-q}-1)$$

Le parcours de $\Sigma_{p+t,q}$ sur les droites x = constante, implique alors l'équifréquence des éléments de $\mathcal{E}_{p+t,q}$.

VALEUR MOYENNE DE E .

Il vient :

$$\text{Card}\,(\Sigma_{p,p}) = \frac{1}{2}\, b^{2s-2}\ \ (b-2)^2$$

et pour $u = 1\ ,\ \ldots\ ,\ s-1$:

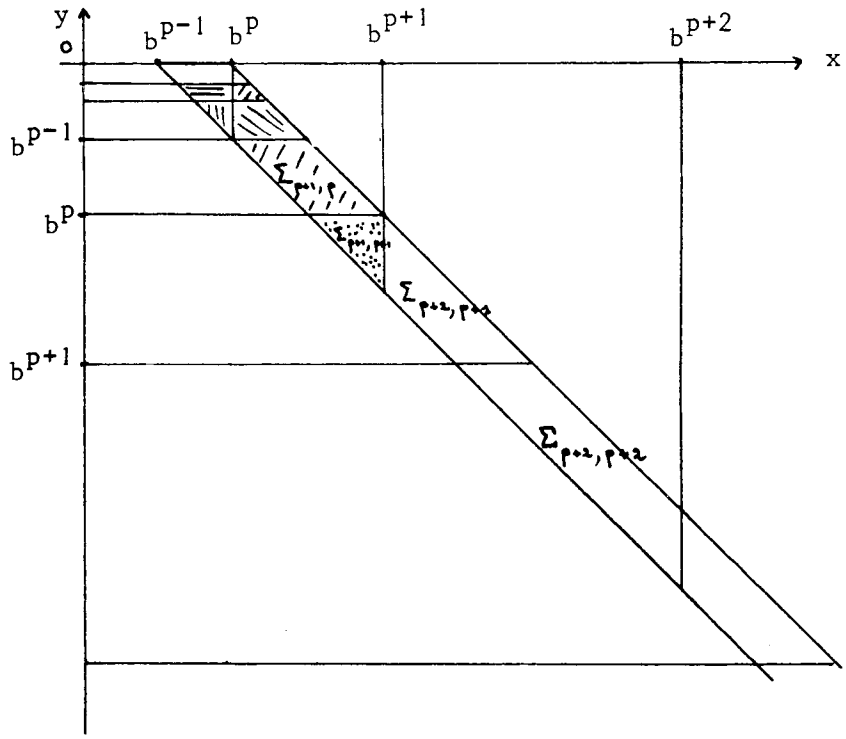
$$\text{Card}\,(\Sigma_{p,p-u}) = b^{s-1}\,(b-1)\,(b^{s} - b^{s-1} - \frac{1}{2}\, b^{s-u} - \frac{1}{2}\, b^{s-1-u} + \frac{1}{2})$$

$$\text{Card}\,(\Sigma_{p+1,p}) = \frac{1}{2}\, b^{2s-3}\,(b^2 - 2) + \frac{1}{2}\, b^{s-1}\,(2 - b)$$

$$\text{Card}\,(\Sigma_{p+1,p+1-u}) = \frac{1}{2}\,(b^2 - 1)\, b^{2s-2-u} - \frac{1}{2}\,(b - 1)\, b^{s-1} \qquad 2 \leqslant u \leqslant s-1$$

$$\text{Card}\,(\Sigma_{p+t,p+t}) = b^{s-t-1}\,(b-1)\,(b^{s} - b^{s-1} - \frac{1}{2}\, b^{s-t} - \frac{1}{2}\, b^{s-t-1} + \frac{1}{2}) \qquad t \geqslant 1$$

$$\text{Card}\,(\Sigma_{p+t,p+t-1}) = \frac{1}{2}\,(b-1)\,b^{s-t}\,(b^{s-t} + b^{s-t-1} - 1) \qquad t \geqslant 2$$



Notons $\bar{E}_{p+t,q}$ la valeur moyenne de E sur le sous-ensemble $\mathcal{E}_{p+t,q}$.
La valeur moyenne $\bar{E}$ de E sur l'ensemble $\mathcal{E}$ s'exprime sous la forme :

$$\bar{E} = \Big(\sum_{t=o}^{s-1} \sum_{q \in I_t} \bar{E}_{p+t,q} \times \text{Card}\,\Sigma_{p+t,q} \Big) \,/\, \text{Card}\,\Sigma$$

$$\mathcal{E}_{p+t,q} = \{E = -\,r\,b^{q-s} \mid r = o\,,\ \ldots\ ,\ b^{p+t-q} - 1 \ \text{ équiprobables}\}$$

$$\bar{E}_{p+t,q} = -\,\frac{1}{2}\,b^{p+t-s}\,(1 - b^{q-p-t})$$

et :

$$\bar{E} = -\,\frac{1}{2}\,b^{p-s}\Big[\Big(\sum_{t=o}^{s-1} \sum_{q} (1 - b^{q-p-t})\,b^{t}\,\text{Card}\,\Sigma_{p+t,q} \Big) \,/\, \text{Card}\,\Sigma \Big]\,.$$

On obtient successivement :

$$\sum_t \sum_q (1 - b^{q-p-t})\, b^t \,\mathrm{Card} \sum_{p+t,q} \sim (s - \tfrac{1}{2})\,(b-1)^2\, b^{2s-2}$$

et

$$\mathrm{Card}\,(\sum - \sum_{p,p-s}) \sim (s - \tfrac{1}{2})\,(b-1)^2\, b^{2s-2}$$

soit :

$$\boxed{\bar{E}_{\mathfrak{C}\cdot\mathfrak{C}_{p,p-s}} = -\frac{1}{2}\, b^{p-s}} \qquad \text{(VII . 3)}$$

Lemme VII . 2 .

Pour l'ensemble des couples (x,y) de T^2 satisfaisant aux conditions :

$$e(x \oplus y) = p \wedge x > 0 \wedge y \leqslant 0 \wedge |x| > |y| \wedge e(x) - e(y) < s$$

la moyenne de l'erreur entachant l'addition, définie par une règle de troncature sans chiffre de garde, des deux composantes du couple, est, en valeur absolue, exactement égale à $\frac{1}{2}\, b^{p-s}$.

I . 3 . 2 Le cas de la troncature avec un chiffre de garde

L'addition de deux nombres de signes contraires est correcte mais n'a pas de direction définie.

CARACTERISATION DE Σ ET $\mathfrak{C}$

Avec des définitions analogues à celles du paragraphe précédent :

$$\Sigma = \bigcup_{t=o}^{s} \left(\bigcup_{q \in I_t} \sum_{p+t,q} \right)$$

où :

$I_o = \{ p, p-1 , \ldots , p-s-1 \}$

$I_1 = \{ p+1 , \ldots , p+1-s \}$

$I_t = \{ p+t , p+t-1 \}$ pour $t = 2 , \ldots , s-1$

$I_s = \{ p+s-1 \}$

L'ensemble $\mathcal{E}$:

★ pour $2 \leqslant t \leqslant s$

$$\mathcal{E}_{p+t,q} = \{o\} \qquad \forall_q$$

★ pour $t = 1$

$$\mathcal{E}_{p+1,p+1} = \mathcal{E}_{p+1,p} = \{o\}$$

L'addition pour $q < p$ de $x = l\, b^{p+1-s}$ et $y = - m\, b^{q-s}$ est E - dirigée : $x \oplus y = (b\, l - \lfloor m\, b^{q-p} \rfloor)\, b^{p-s}$

$\sum_{p+1,q}$ $(q<p)$ est l'ensemble des points $(x = l\, b^{p+1-s}, y = - m\, b^{q-s})$ de T^2 dont les coordonnées vérifient :

$$\max (b^{s-1} , b^{p-q} (b\, l - b^s + 1)) \leqslant m \leqslant \min (b^s - 1 , b^{p-q}(b\, l - b^{s-1} + 1)-1)$$

$$= b^s - 1$$

et dans ce cas :

$$\mathcal{E}_{p+1,q} = \{ E = - r\, b^{q-s} \mid r = 0,1, \ldots , b^{p-q} - 1 \text{ équiprobables } \}$$

★ pour $t = 0$

$$\mathcal{E}_{p,p} = \{0\} , \mathcal{E}_{p,p-s-1} = \{ y | y \in T^- \wedge e(y) \leqslant p-s-1 \}$$

Pour $p-s \leqslant q \leqslant p-1$, $\sum_{p,q}$ est l'ensemble des points $(x = l\, b^{p-s} , y = - m\, b^{q-s})$ de T^2 dont les coordonnées vérifient :

$$\left\lfloor \frac{b\,l - \lfloor m\, b^{q-p+1} \rfloor}{b} \right\rfloor \geqslant b^{s-1}$$

ou

$$b^{s-1} \leqslant m \leqslant \min\,(b^s - 1\ ,\ b^{p-q-1}\,(b\,l - b^s + 1) - 1)$$

En rappelant :

$E = (r_2\, b^{p-q-1} - r_1)\, b^{q-s}$ avec $m = b^{p-q-1}\, m_1 + r_1 \wedge r_1 < b^{p-q-1}$

$b\,l - m_1 = bq_1 + r_2 \wedge r_2 < b$,

le parcours de $\sum_{p,q}$ à l fixé, implique pour les éléments de $\mathcal{E}_{pq}$:

- si $b^s \leqslant b^{p-q-1}(bl - b^s + 1)$ une fréquence uniforme
- si $b^s > b^{p-q-1}(bl - b^s + 1)$ l'obtention pour l fixé, des erreurs $0\ ,\ -b^{q-s}\ ,\ \ldots\ ,\ -(b^{p-q-1} - 1)\, b^{q-s}$ en surnombre par rapport à la fréquence uniforme.

$$\mathcal{E}_{p,p-1} = \{\ E = r\, b^{p-1-s}\ \mid\ r = 0,\ \ldots\ ,\ b-1\ \}$$

$$\mathcal{E}_{p,q} = \{\ E = r\, b^{q-s}\ \mid\ r = -(b^{p-q-1}-1),\ \ldots,\ 0,1,\ \ldots, b^{p-q} - b^{p-q-1}\} \quad 1<p-q<s$$

$$\mathcal{E}_{p,p-s} = \{\ E = r\, b^{p-2s}\ \mid\ r = 1\ ,\ \ldots\ ,\ b^s - b^{s-1}\ \}$$

VALEURS MOYENNES DE E

Le nombre d'éléments d'un sous-ensemble $\sum_{p+t,q}$ est égal, au terme $O(b^s)$ près, au nombre d'éléments du sous-ensemble correspondant du § I.3.1 précédent.

L'unique formule de dénombrement à modifier est relative au nombre d'éléments de l'ensemble $\sum$, qui inclut Card $\sum_{p,p-s}$:

$$\text{Card}\ (\textstyle\sum - \sum_{p,p-s-1}) \sim (s + \frac{1}{2})\ (b - 1)^2\ b^{2s-2}$$

Dans les sous-ensembles $\mathcal{E}_{p,q}$, l'écart entre les fréquences des erreurs positives, et négatives ou nulles, est négligeable.

Nous avons étudié l'ensemble $\mathcal{E}$ image de l'ensemble Σ :

$$\Sigma = \{ (x,y) \mid (x,y) \in T^2 \wedge e(x \oplus y) = p \wedge x > o \wedge y \leqslant o \wedge |x| > |y| \}$$

avec l'hypothèse de la représentation des nombres négatifs par signe et valeur absolue. Le calcul de la valeur moyenne de E quand (x,y) parcourt Σ sera effectué dans les deux hypothèses de représentation des nombres négatifs.

1) Les nombres négatifs sont représentés par leur signe et leur valeur absolue.

La valeur moyenne de E quand (x,y) parcourt $\Sigma - \Sigma_{p,p-s-1}$ s'écrit :

$$\bar{E}\,\mathcal{E} - \mathcal{E}_{p,p-s-1} = \frac{1}{2}\, b^{p-s}\,(1-\omega) \quad \text{où} \quad \omega = \frac{4s(b-1) + (b-2)^2}{(2s+1)\,(b-1)\,b} \qquad \text{(VII.4)}$$

soit, pour :

b	2	4	8	16
ω	$\frac{2s}{2s+1}$	$\frac{3s+1}{6s+3}$	$\frac{7s+9}{28s+14}$	$\frac{15s+49}{120s+60}$

2) Les nombres négatifs sont représentés en notation de complément à 2

L'opération $\oplus$ est G - dirigée.

Il vient pour les sous-ensembles de $\mathcal{E}$ à éléments non tous nuls :

$$\mathcal{E}_{p+1,q} = \mathcal{E}_{p,q} = \{ E = r\, b^{q-s} \mid r = 0,1, \ldots , b^{p-q} - 1 \} \quad p-1 \leq q \leq p+1-s$$

$$\mathcal{E}_{p,p-s} = \{ E = r\, b^{p-2s} \mid r = 1, \ldots , b^{s} - b^{s-1} \}$$

$$\bar{E}_{\mathcal{E}-\mathcal{E}_{p,p-s-1}} = \frac{1}{2} b^{p-s} (1 - \omega) \quad \text{où} \quad \omega = \frac{b^2 + 4b - 2}{(2s+1)(b-1)b} \qquad \text{(VII.5)}$$

pour $b = 2$, $s = 22$ $\qquad \omega = 0,11$

pour $b = 16$, $s = 6$ $\qquad \omega = 0,10$

Lemme VII . 3 .

Pour l'ensemble des couples (x,y) de T^2 satisfaisant aux conditions

$$e(x \oplus y) = p_{\wedge}\, x>o \wedge y<o \wedge |x| > |y| \wedge e(x) - e(y) \leq s$$

la moyenne de l'erreur entachant l'addition, définie par une règle de troncature avec chiffre de garde, des deux composantes du couple, est égale à $\frac{1}{2} b^{p-s} (1 - \omega)$ où , suivant la représentation des nombres négatifs, par signe et valeur absolue ou en notation de complément à 2 , ω prend les valeurs définies respectivement par (VII.4) et (VII.5).

I . 4 CONCLUSION RELATIVE A L'ADDITION

Les calculs de moyennes effectués aux paragraphes précédents pour respectivement, l'addition et la soustraction arithmétiques, permettent d'évaluer la valeur moyenne de l'erreur entachant l'addition de deux nombres de signes quelconques.

(Nous rappelerons que les éléments des sous-ensembles $\mathcal{E}_{p,p-s}$ - resp. $\mathcal{E}_{p,p-s-1}$ - sont égaux en valeur absolue et de signes opposés pour l'addition et la soustraction arithmétiques dans le cas d'une troncature sans chiffre de garde - resp. avec chiffre de garde, avec de plus dans ce cas une moyenne de $\frac{1}{2} b^{p-s}$ pour les couples (x,y) tels que $e(x) = p_{\wedge} e(y) = p - s$.

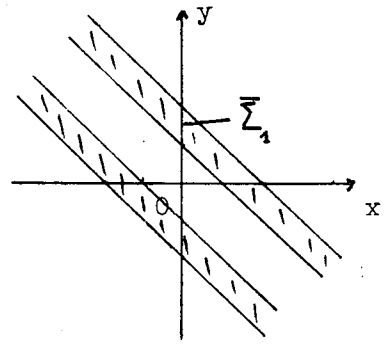
Etant donné le domaine $\overline{\Sigma}$:

$\overline{\Sigma} = \{ (x,y) \mid (x,y) \in T^2 \wedge e(x \oplus y) = p \}$

et l'image $\overline{\mathcal{E}}$ de $\overline{\Sigma}$ par l'application :

$(x,y) \rightarrow E = x + y - (x \oplus y)$

nous ferons les remarques suivantes :



1) L'écart entre les valeurs moyennes, théorique et calculée exactement, de E sur $\overline{\mathcal{E}}$, est assez faible.
Toutefois, l'existence de ce biais dans le cas d'une représentation des nombres négatifs en complément à 2 , se révelera lorsque les sommations s'effectueront sur un grand nombre de termes.

2) Pour des calculs effectués sur des composantes de couples (x,y) appartenant à des sous-ensembles particuliers de $\overline{\Sigma}$, par exemple

$\overline{\Sigma}_1 = \{ (x,y) \mid (x,y) \in \overline{\Sigma} \wedge x \oplus y > o \}$,

l'hypothèse théorique de répartition de l'erreur d'arrondi ne donnera pas des résultats satisfaisants.

Il est à noter que, dans certains cas, le raffinement apporté par le chiffre de garde, s'il diminue la moyenne de l'erreur élémentaire, peut paradoxalement nuire à la précision d'une suite de calculs : Ainsi, les études statistiques de H. KUKI et W.J. CODY [17] établissent, relativement à la somme algébrique, une précision meilleure pour une troncature sans chiffre de garde que pour une troncature avec un chiffre de garde, dans l'hypothèse de la représentation des nombres négatifs avec signe et valeur absolue ; le phénomène de compensation entre les différentes erreurs élémentaires, a diminué plus fortement l'erreur totale dans le 1er cas.

Un cas similaire de "rupture de symétrie" est présenté par la règle d'arrondi - définissant $x \circledast y$ par l'égalité :

$| x \bigstar y - x \circledast y | = \underset{z \in T}{\text{Min}} | x \bigstar y - z |$ - quand l'arrondi se fait systématiquement loin du zéro (resp. vers le zéro) lorsque l'égalité $x \bigstar y = \frac{\Delta x + \nabla x}{2}$ a lieu. La perte de précision entrainée par la dissymétrie de la définition, a été soulignée par M. URABE [39] et H. KUKI et W.J. CODY [17], qui ont proposé des modifications de la règle simple précédente.

Enfin, dans le cadre de cette étude, nous mentionnerons l'obtention d'une valeur moyenne strictement nulle de l'erreur d'entrée, dans l'arithmétique paritaire définie par S. LINNAINMAA [21].

I. 5. LE CAS DE LA MULTIPLICATION

L'étude de la répartition de l'erreur d'arrondi élémentaire est, dans le cas de la multiplication en arithmétique à virgule flottante, plus complexe que dans le cas de l'addition, et nous poserons ici seulement le problème.

Considèrons par exemple, le cas où l'opération $\otimes$ de multiplication définie sur T, est correcte et I - dirigée.

Au couple (x,y) de $(T^+)^2$, de représentation :

$$x = lb^{e(x)-s}$$

$$b^{s-1} \leqslant l, m < b^s$$

$$y = mb^{e(y)-s}$$

est associée l'erreur élémentaire E de multiplication :

$$E = rb^{e(x) + e(y) - 2s}$$

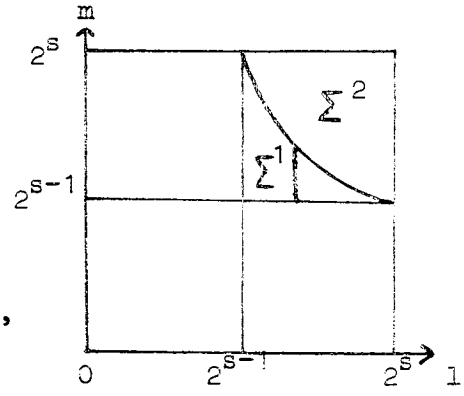
où r est défini par la division arithmétique suivante :

$$\left| \begin{array}{l} lm = b^{s-\theta} q + r \ \wedge \ r < b^{s-\theta} \\ \text{avec} \quad \theta = 1 \quad \text{si} \quad lm < b^{2s-1} \\ \qquad\quad \theta = 0 \quad \text{si} \quad lm \geqslant b^{2s-1} \end{array} \right. \qquad \text{(VII. 6)}$$

On est ainsi amené à considèrer les 2 ensembles $\sum^1$ et $\sum^2$ de couples (l,m) d'entiers, d'images intérieures au carré $\begin{cases} b^{s-1} \leqslant l < b^s \\ b^{s-1} \leqslant m < b^s \end{cases}$,

et vérifiant respectivement :
$lm < b^{2s-1}$ pour $\sum^1$, $lm \geqslant b^{2s-1}$ pour $\sum^2$.
Pour chaque ensemble $\sum^1$ et $\sum^2$, il faut étudier les valeurs prises par le reste r de (VII. 6) lorsque le couple (l,m) parcourt l'ensemble.

Dans l'hypothèse par exemple d'une base égale à 2, le parcours de $\sum^1$ (de manière analogue $\sum^2$) sur les droites l = constante, entraîne les résultats suivants [13] :

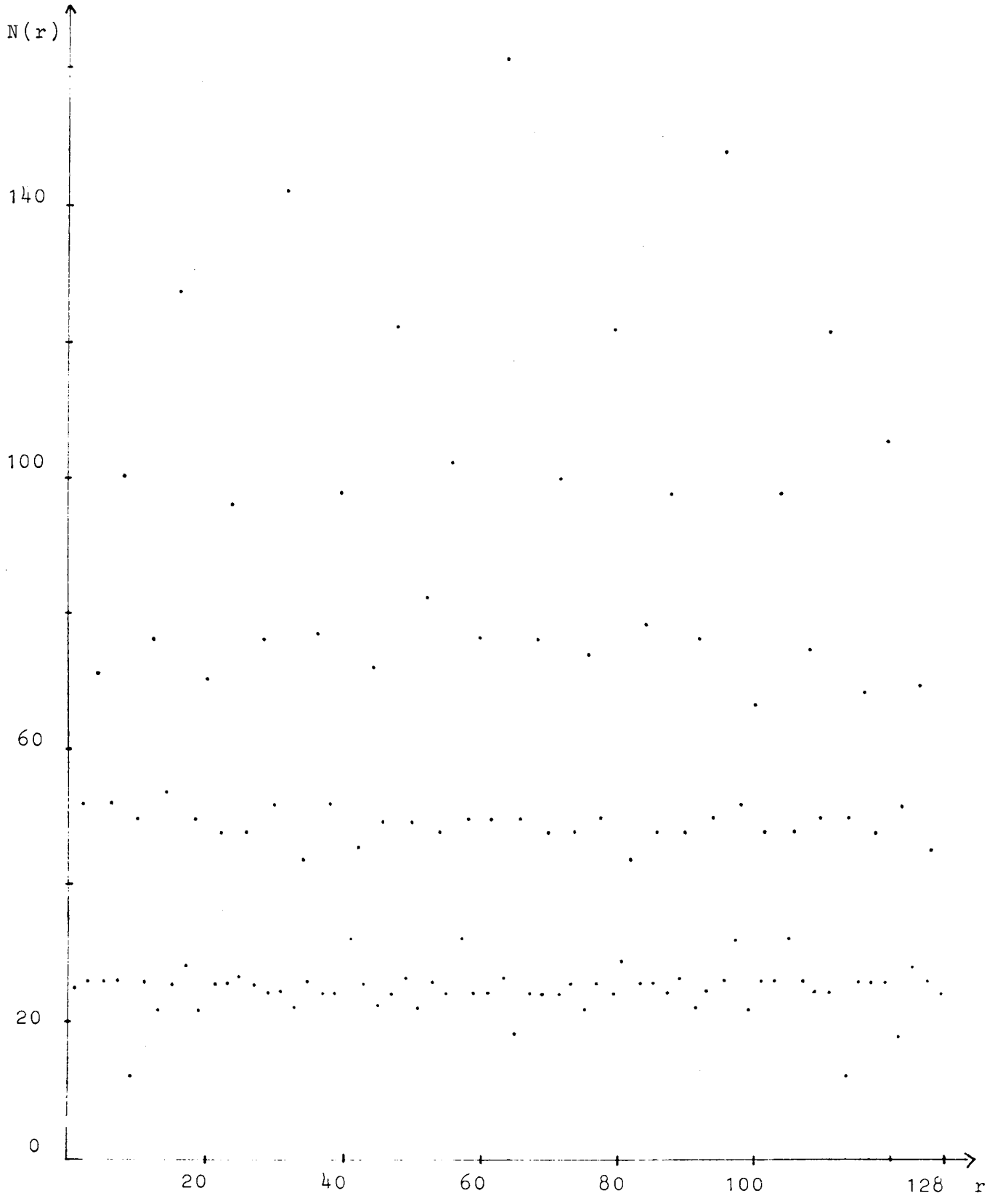


- <u>Pour l fixé impair</u>, m variant de 2^{s-1} à $\lfloor 2^{2s-1}/l \rfloor$, les restes des divisions de lm par 2^{s-1} , prennent leurs valeurs parmi $0, 1, 2, \ldots , 2^{s-1}-1$ et forment un système de restes <u>non congrus</u> (mais incomplet).

- <u>Pour l fixé pair, $l = 2^k\, l'$</u> , l' impair, m variant de 2^{s-1} à $\lfloor 2^{2s-1} / l \rfloor$, les restes des divisions prennent leurs valeurs parmi $0, 2^k, 2 \cdot 2^k, \ldots , (2^{s-k-1} - 1)\, 2^k$.

L'image $\mathcal{E}^1_{s-1}$ de $\sum^1$ par l'application : $(l,m) \longrightarrow r$ est l'union $\bigcup_{k=0} \mathcal{E}^1_k$, où $\mathcal{E}^1_k$ désigne l'ensemble image de la partie de $\sum^1$ obtenue pour les valeurs de l multiples de 2^k .
Malheureusement, montrer que la répartition des éléments de $\mathcal{E}^1_k$, $\forall\, k$, est analogue à une répartition uniforme, se révèle difficile en raison de la limitation $lm < 2^{2s-1}$ (En arithmétique fixe par contre, on a exactement l'équirépartition des éléments des sous-ensembles de type $\mathcal{E}^2_k$).

Nous donnons ci-dessous à titre d'exemple , pour b=2 et s=8 ,
les fréquences des différents restes obtenus lorsque (l,m) parcourt $\sum^{-1}$:



II - LE PROBLEME D'UN MEILLEUR CHOIX DE LA BASE

Il est nécessaire de mentionner dans ce travail, le problème du choix optimal de la base de la représentation des nombres à virgule flottante.
Nous en rappelerons ci-dessous les points essentiels.
Ce problème peut être posé sous plusieurs formes.

Si le nombre de bits affecté à la représentation des nombres à virgule flottante est fixé, on comparera les différentes représentations obtenues lorsque varient, d'une part le choix de la base, d'autre part le nombre de bits affectés respectivement à la mantisse et à l'exposant.
La comparaison entre les représentations doit s'effectuer en tenant compte de :

- la largeur du domaine de $\mathbb{R}$ représenté
- la précision de la représentation.

Ce dernier point de vue est assez délicat et différentes mesures de la précision ont été proposées par plusieurs auteurs :

Une mesure de la précision est donnée par l'erreur relative de représentation d'un réel par un nombre à virgule flottante.
Cette erreur est, suivant la terminologie de W.J. CODY, une caractéristique statique de la représentation.

L'étude de W.M. Mc KEEMAN ([23] , 1967) sur la valeur moyenne de cette erreur relative, pour une distribution logarithmique des mantisses, établit qu'il n'existe pas, relativement à ce critère, de différence significative entre les performances des bases puissances de 2, dans un calculateur binaire.

L'étude de W.S. BROWN et P.L. RICHMAN ([3], 1969) sur la borne supérieure de l'erreur relative définie ci-dessus, établit qu'un meilleur choix de la base est donné par $b = r$, r désignant le nombre d'états possibles du dispositif élémentaire composant le mot-mémoire.

[23] et [3] établissent la supériorité de $b = 2$ pour un calculateur binaire, lorsqu'est utilisée, pour une représentation normalisée des nombres, l'écriture implicite du bit le plus significatif de la mantisse -bit toujours égal à 1 sauf pour le nombre nul-.

D.W. MATULA ([27], 1972) a mis en évidence une perte de précision liée au choix d'une large base et entrainée par les 2 points suivants :
- la distribution du chiffre de tête de mantisse étant non uniforme mais logarithmique, accentue le défaut de précision provoqué par les mantisses à bits de tête nuls.
- il existe des intervalles réels, non petits, sur lesquels pour $b > d$, l'espacement entre des éléments adjacents de T_b^s est plus grand que l'espacement entre des éléments adjacents de T_d^s .

Cependant, comme le souligne W.J. CODY, il est utile d'élargir la discussion et de considèrer les caractéristiques dynamiques de tout le système arithmétique à virgule flottante,défini par la représentation des nombres et par l'arithmétique utilisée.

R.P. BRENT ([2], 1973) étudiant :
- l'écart-type de l'erreur relative d'arrondi d'ure part
- les résultats, pour différentes bases et différentes arithmétiques, de la simulation de sommes, résolutions de systèmes linéaires, calculs de valeurs propres, d'autre part,

établit, pour ces critères, que le choix $b = 2$ avec le bit le plus significatif implicite, est le meilleur, suivi par le choix $b = 4$, les opérations flottantes étant optimales.

Une conclusion semblable est donnée par W.J. CODY ([4], 1973), à savoir que compte-tenu des contraintes de l'écriture implicite d'un bit en base 2, une représentation des nombres en base 4 et une arithmétique d'arrondi sur deux chiffres de garde, assurera un meilleur système arithmétique.

Mentionnons aussi, en faveur de la base 4, le travail de T. KREIFELTS ([16] , 1973) .

Enfin, à coté de l'étude statistique, relative à la précision des systèmes de nombres à virgule flottante, de H. KUKI et W.J. CODY ([17], 1973) , nous mentionnerons le travail de J.D. MARASA et D.W. MATULA ([26], 1973) :
Ces auteurs, dans une étude simulative de la propagation de l'erreur d'arrondi au cours de longs calculs entre variables corrélées, ont établi que la croissance de l'erreur dépend essentiellement du type des opérations impliquées dans le calcul : pourcentages respectifs d'additions, soustractions, multiplications et divisions ; la soustraction en particulier, conduit en raison des cancellations, aux plus fortes croissances de l'erreur ; l'arrondi, préférable à la troncature en général, n'introduit qu'un facteur multiplicatif, variant peu avec le nombre d'opérations réalisées.

Il apparait enfin que le choix de la base est moins décisif que la possibilité de disposer sur un calculateur de 2 arithmétiques flottantes, l'utilisation parallèle de ces arithmétiques étant particulièrement significative lorsque les opérateurs sont monotones.

(Voir [46] pour la réalisation d'un tel calculateur).
L'importance de la considèration attentive de la "stratégie" des calculs sera à nouveau soulignée, en rappelant, par exemple les résultats relatifs à la résolution itérative des systèmes linéaires ou le fait qu'une sommation particulière de termes de signes opposés au chapitre V, a permis d'éviter l'effet traditionnellement néfaste de ces opérations.

ANNEXE

CARACTERISTIQUES DE QUELQUES ARITHMETIQUES A VIRGULE FLOTTANTE

A titre d'exemple, nous avons noté ci-dessous les caractéristiques de quelques arithmétiques à virgule flottante fréquemment utilisées, et considèrées pour des représentations normalisées.

En relation avec le chapitre II, nous avons rappelé, pour chaque mode d'opération, les théorèmes permettant le calcul des erreurs d'arrondi élémentaires.

Nous soulignerons que sur certains calculateurs (CDC 6000-7000, UNIVAC 1108), l'accès direct à l'erreur d'arrondi entachant une opération élémentaire, est prévue par instruction de l'arithmétique flottante : le résultat d'une opération élémentaire est dans ce cas, écrit sur deux nombres à virgule flottante.

Notation : Nous désignerons par $\mathcal{C}$, le mode de troncature réalisant pour $X \in T_b^s$, $Y \in T_b^s$ l'opération d'addition-soustraction ou de multiplication selon la succession d'étapes suivantes :

1) Pour l'addition :

On décale vers la droite la mantisse du nombre de plus petit exposant, sans perte de précision

On ajoute les 2 mantisses alors cadrées.

Pour la multiplication :

On multiplie les mantisses de X et Y .

Soit $m(R)$ le produit de composition de mantisses obtenu.

2) On normalise $m(R)$ uniquement dans le cas où s'est produit un dépassement de capacité de mantisse, ce qui ne peut avoir lieu qu'en cas d'addition de 2 nombres de même signe.

3) $m(R)$ est tronquée à s chiffres.

le calcul est achevé de manière habituelle.

type de calculateur	base	nombre de chiffres de mantisse	représentation de la mantisse	Mode Addition - Soustractio
PDP 11/45	2	23 bits + 1 bit muet 55 bits + 1 bit muet	Signe et valeur absolue	Deux modes sont possible: Opération correct Opération correct
IBM 360 370	16	6 14 28 poss.	Signe et valeur absolue	Troncature avec un chiffre de garde
cii 10 070	16	6 14	Nombres négatifs en notation de complément à 2	Troncature avec un chiffre de garde Troncature sans chiffre de garde
IRIS 80	16	6 14	Nombres négatifs en notation de complément à 2	Deux modes 1) Troncature avec un chiffre de garde 2) Arrondi sur chiffre de garde Deux modes 1) Troncature sans chiffre de garde 2) Arrondi sur chiffre de garde

'opérations ultiplication - Division	Calcul des erreurs élémentaires suivant :
ans chacun des formats t I - dirigée t E - dirigée	Propositions II.3 II.9, II.9', II.10 II.13
Opération correcte et I - dirigée	Propositions II.2 II.9, II.11 II. 13
)p. correcte et I - dirigée)p. correcte et I - dirigée	Propositions II.6 II.9, II.9', II.13
sont possibles :)p. correcte et I - dirigée Arrondi sont possibles :)p. correcte et I - dirigée Arrondi	Propositions II.6 II.9, II.9'. II.13 Voir [5] Voir [5]

type de calculateur	base	nombre de chiffres de mantisse	représentation de la mantisse	Mode Addition - Soustractic
CDC 6600 CDC 7600	2	48	Nombres négatifs en notation de complément à 1	Deux modes sont possible: 1) Ecriture en DOUBLE sur 2 nombres flottants Détermination du poid Troncature σ 2) Ecriture en SIMPLE selon des processus
UNIVAC 1108	2	27 60	Nombres négatifs en notation de complément à 1	Ecriture en DOUBLE sur 2 nombres flottants Détermination du poid Troncature sans chiffre de garde Troncature sans chiffre de garde

Note : Lorsque l'erreur d'arrondi élémentaire est déterminée à l'aide d'une instruction
- Pour une opération d'addition - soustraction, cette erreur est mise à 0 en cas
- Pour une opération de division, la quantité déterminée est non pas l'erreur E

opérations ltiplication - Division	Calcul des erreurs élémentaires suivant :
ivant la précision désirée : ECISION du résultat exposants appropriés (sauf division) rt du résultat : . correcte et I - dirigée ECISION du résultat rticuliers d'arrondi	Directement par instructions de l'arithmétique flottante pour l'addition, la soustraction, la multiplication (mode 1 d'opérations)
ECISION du résultat exposants appropriés. rt du résultat : ération correcte et I - dirigée) : Troncature) : correcte et I - dirigée	Directement par instructions de l'arithmétique flottante Voir paragraphe I. 3. 1. b

e l'arithmétique flottante :
l'écart trop grand entre les exposants des 2 opérandes.
lle-même, mais la quantité E Y .

B I B L I O G R A P H I E

[1] BABUSKA, I. : Numerical Stability in Mathematical Analysis.
IFIP congr. 68, Invited papers, 11-23 (1968).

[2] BRENT, R.P : On the Precision Attainable with Various Floating-Point Number Systems.
IEEE Trans. Computers, C-22, 6 (June 1973), 601-607.

[3] BROWN, W.S and RICHMAN, P.L : The Choice of Base.
Comm. ACM 12,10 (Oct.1969), 560-561.

[4] CODY, W.J, Jr : Static and Dynamic Numerical Characteristics of Floating-Point Arithmetic.
IEEE Trans. Computers, C-22, 6 (June 1973), 598-601.

[5] DEKKER , T.J : A Floating-Point Technique for Extending the Available Precision.
Numer. Math. 18, 224-242 (1971).

[6] DURAND, E : Solutions numériques des équations algébriques.
Tome II. Masson (1972).

[7] FIEDLER, M and PTÀK, V : On matrices with non-positive off-diagonal elements and positive principal minors.
Чехословацкий математический журнал T.12 (87)
1962, Prague pp. 382-400.

[8] GASTINEL, N : Matrices du second degré et normes générales en analyse numérique linéaire.
Publications scientifiques et techniques du Ministère de l'Air, Paris, N.T.110-S.D.I.T (1960).

[9] GASTINEL, N : Analyse Numérique linéaire.
Hermann, Paris 1966.

[10] GENTLEMAN, W. MORVEN and MAROVICH, S.B : More on Algorithms that Reveal Properties of Floating Point Arithmetic Units.
Short Comm. ACM 17, 5 (May 1974), 276-277.

[11] GREGORY, J. : A Comparison of Floating-Point Summation Methods.
Short Comm. ACM 15, 9 (Sept. 1972), 838.

[12] HAMMING, R.W : On the Distribution of Numbers.
Bell Syst. Tech. Journal, Vol 49 (October 1970)
pp. 1609-1625.

[13] HARDY, G.H and WRIGHT, E.M : An Introduction to the theory of numbers. 4th edition. OXFORD University Press 1962, theorem 56 p 51.

[14] KAHAN, W : Further remarks on reducing truncation errors. Comm. ACM 8,1 (January 1965), 40.

[15] KNUTH, D.E : The Art of Computer Programming. Vol 2 . Addison Wesley (1969).

[16] KREIFELTS, T : Optimal choise of basis for a floating-point arithmetic Computing 11,4 (1973), 353-363 (German) .

[17] KUKI, H and CODY, W.J : A Statistical Study of the Accuracy of Floating Point Number Systems. Comm. ACM 16, 4 (April 1973), 223-230.

[18] KULISCH, U : An Axiomatic Approach to Rounded Computations. Numer. Math. 18, 1-17 (1971).

[19] LA PORTE, M et VIGNES, J : Etude statistique des erreurs dans l'arithmétique des ordinateurs; application au contrôle des résultats d'algorithmes numériques. Numer. Math. 23, 63-72 (1974).

[20] LA PORTE, M et VIGNES, J : Algorithmes numériques, analyse et mise en oeuvre. Technip Paris, 1974.

[21] LINNAINMAA, S : Analysis of some known methods of improving the accuracy of floating-point sums. BIT 14 (1974), 167-202.

[22] LINZ, P : Accurate Floating-Point Summation. Comm. ACM 13, 6 (June 1970), 361-362.

[23] Mc KEEMAN, W.M. : Representation Error for Real Numbers in Binary Computer Arithmetic. IEEE Trans. on Electronic Computers (October 1967), 682-683.

[24] MALCOLM, M.A : On Accurate Floating-Point Summation. Comm. ACM 14, 11 (November 1971), 731-736.

[25] MALCOLM, M.A : Algorithms To Reveal Properties of Floating-Point Arithmetic.
Comm. ACM 15, 11 (november 1972), 949-951.

[26] MARASA, J.D and MATULA, D.W : A Simulative Study of Correlated Error Propagation in Various Finite-Precision Arithmetics.
IEEE Trans. Computers, C-22, 6 (june 1973), 587-597.

[27] MATULA, D.W : Significant digits : Numerical analysis or numerology.
IFIP congr. 71, vol 2, pp. 1278-1283 (1972).

[28] MOLLER, O : Quasi double-precision in floating point addition.
BIT 5 (1965), 37-50.

[29] MOORE, R.E : Interval analysis.
Prentice Hall 1966.

[30] NICKEL, K and RITTER, K : Termination Criterion and Numerical Convergence.
Siam J. Numer. Anal. Vol. 9 , No 2, June 1972, pp 277-283.

[31] PICHAT, M : Correction d'une somme en arithmétique à virgule flottante.
Numer. Math. 19, 400-406 (1972).

[32] PICHAT, M : Un algorithme, en arithmétique à virgule flottante, pour augmenter la précision d'une somme.
C.R.A.S. t. 274 (1972) série A, pp. 1639-1642.

[33] PICHAT, M : Correction d'une somme calculée en arithmétique à virgule flottante, à l'aide de l'arithmétique utilisée.
Colloque d'Analyse Numérique Epinal Juin 1972.

[34] PICHAT, M : Stabilité numérique de différents schémas de sommation dans un sous-ensemble de $\mathbb{R}$ de type flottant.
C.R.A.S t. 277 (1973) série A, pp. 315-317.

[35] PICHAT, M : Utilisation conjointe de deux arithmétiques flottantes
Séminaire A.N. Grenoble I (Janv. 1975) n° 216.

[36] ROSS, D.R : Reducing Truncation Errors Using Cascading Accumulators.
Comm. ACM 8 , 1 (January 1965), 32-33.

[37] TIENARI, M : On the control of floating-point mantissa length in iterative computations.
International Computing Symposium 1973. A. Günther et al.(eds.) 1974, 315 - 322.

[38] TSAO, N.K : On the Distributions of Significant Digits and Roundoff Errors.
Comm. ACM 17, 5 (May 1974), 269-271.

[39] URABE, M : Roundoff Error Distribution in Fixed-Point Multiplication and a Remark about the Rounding Rule.
SIAM J. Numer. Anal., Vol. 5, n° 2 (june 1968) pp 202-210.

[40] VELTKAMP, G.W : Private communicaticns.
Technological University, Eindhoven (1968).

[41] WALKER, R.J. : Binary Summation.
Short Comm. ACM 14, 6 (june 1971), 417.

[42] WILKINSON, J.H : Rounding errors in algebraic processes.
Her Majesty's Stationary Office (1963).

[43] WILKINSON, J.H : The algebraic eigenvalue problem.
Clarendon Press, Oxford, 1965.

[44] WOLFE, J.M : Reducing truncation errors by programming.
Comm. ACM 7, 6 (June 1964), 355-356.

[45] YOHE, J.M : Interval Bounds for Square Roots and Cube roots.
Computing, Vol. 11, Fasc. 1, 1973.

[46] YOHE, J.M : Roundings in Floating-Point Arithmetic.
IEEE Trans. Computers, C-22, 6 (June 1973), 577-586.

INDEX DES TERMES ET NOTATIONS

T A B L E D E S M A T I E R E S