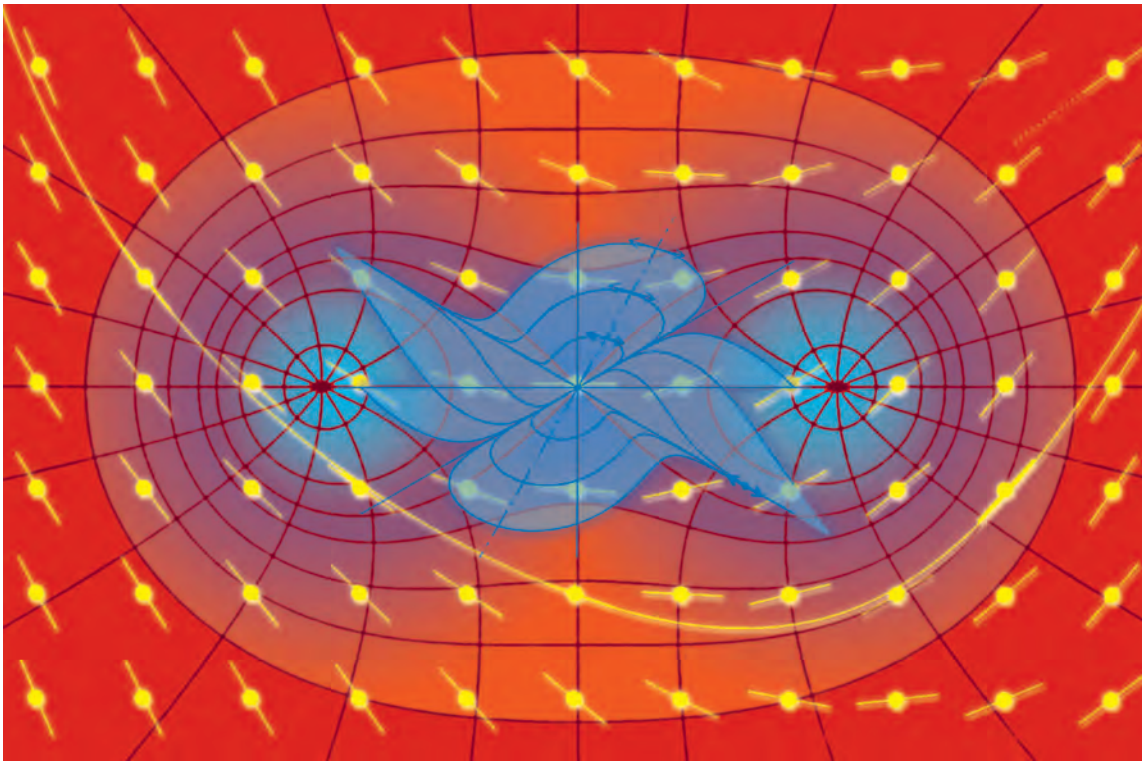
COLLECTION GRENOBLE SCIENCES
DIRIGÉE PAR JEAN BORNAREL

ANALYSE NUMÉRIQUE ET ÉQUATIONS DIFFÉRENTIELLES

Nouvelle édition

Jean-Pierre DEMAILLY



EDP SCIENCES

ANALYSE NUMÉRIQUE
ET
ÉQUATIONS DIFFÉRENTIELLES

Grenoble Sciences

Grenoble Sciences poursuit un triple objectif :

- réaliser des ouvrages correspondant à un projet clairement défini, sans contrainte de mode ou de programme,
- garantir les qualités scientifique et pédagogique des ouvrages retenus,
- proposer des ouvrages à un prix accessible au public le plus large possible.

Chaque projet est sélectionné au niveau de Grenoble Sciences avec le concours de referees anonymes. Puis les auteurs travaillent pendant une année (en moyenne) avec les membres d'un comité de lecture interactif, dont les noms apparaissent au début de l'ouvrage. Celui-ci est ensuite publié chez l'éditeur le plus adapté.

(Contact : Tél. : (33)4 76 51 46 95 - E-mail : Grenoble.Sciences@ujf-grenoble.fr)

Deux collections existent chez EDP Sciences :

- la ***Collection Grenoble Sciences***, connue pour son originalité de projets et sa qualité
- ***Grenoble Sciences - Rencontres Scientifiques***, collection présentant des thèmes de recherche d'actualité, traités par des scientifiques de premier plan issus de disciplines différentes.

Directeur scientifique de Grenoble Sciences

Jean BORNAREL, Professeur à l'Université Joseph Fourier, Grenoble 1

Comité de lecture pour *Analyse numérique et équations différentielles*

- M. ARTIGUE, Professeur à l'IUFM de Reims
- A. DUFRESNOY, Professeur à l'Université Joseph Fourier - Grenoble 1
- J.R. JOLY, Professeur à l'Université Joseph Fourier - Grenoble 1
- M. ROGALSKI, Professeur à l'Université des Sciences et Techniques - Lille 1

Grenoble Sciences bénéficie du soutien du **Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche** et de la **Région Rhône-Alpes**.
Grenoble Sciences est rattaché à l'**Université Joseph Fourier de Grenoble**.

Illustration de couverture : Alice GIRAUD

ISBN 2-86883-891-X

© EDP Sciences, 2006

ANALYSE NUMÉRIQUE
ET
ÉQUATIONS DIFFÉRENTIELLES

Jean-Pierre DEMAILLY

17, avenue du Hoggar
Parc d'Activité de Courtabœuf - BP 112
91944 Les Ulis Cedex A - France

Ouvrages Grenoble Sciences édités par EDP Sciences

Collection Grenoble Sciences

Chimie. Le minimum à savoir *(J. Le Coarer)* • Electrochimie des solides *(C. Déportes et al.)* • Thermodynamique chimique *(M. Oturan & M. Robert)* • CD de Thermodynamique chimique *(J.P. Damon & M. Vincens)* • Chimie organométallique *(D. Astruc)* • De l'atome à la réaction chimique *(sous la direction de R. Barlet)*

Introduction à la mécanique statistique *(E. Belorizky & W. Gorecki)* • Mécanique statistique. Exercices et problèmes corrigés *(E. Belorizky & W. Gorecki)* • La cavitation. Mécanismes physiques et aspects industriels *(J.P. Franc et al.)* • La turbulence *(M. Lesieur)* • Magnétisme : I Fondements, II Matériaux et applications *(sous la direction d'E. du Trémolet de Lacheisserie)* • Du Soleil à la Terre. Aéronomie et météorologie de l'espace *(J. Lilensten & P.L. Blelly)* • Sous les feux du Soleil. Vers une météorologie de l'espace *(J. Lilensten & J. Bornarel)* • Mécanique. De la formulation lagrangienne au chaos hamiltonien *(C. Gignoux & B. Silvestre-Brac)* • Problèmes corrigés de mécanique et résumés de cours. De Lagrange à Hamilton *(C. Gignoux & B. Silvestre-Brac)* • La mécanique quantique. Problèmes résolus, T. 1 et 2 *(V.M. Galitsky, B.M. Karnakov & V.I. Kogan)* • Description de la symétrie. Des groupes de symétrie aux structures fractales *(J. Sivardière)* • Symétrie et propriétés physiques. Du principe de Curie aux brisures de symétrie *(J. Sivardière)*

Exercices corrigés d'analyse, T. 1 et 2 *(D. Alibert)* • Introduction aux variétés différentielles *(J. Lafontaine)* Mathématiques pour les sciences de la vie, de la nature et de la santé *(F. & J.P. Bertrandias)* • Approximation hilbertienne. Splines, ondelettes, fractales *(M. Attéia & J. Gaches)* • Mathématiques pour l'étudiant scientifique, T. 1 et 2 *(Ph.J. Haug)* • Analyse statistique des données expérimentales *(K. Protassov)* • Nombres et algèbre *(J.Y. Mérindol)*

Bactéries et environnement. Adaptations physiologiques *(J. Pelmont)* • Enzymes. Catalyseurs du monde vivant *(J. Pelmont)* • Endocrinologie et communications cellulaires *(S. Idelman & J. Verdetti)* • Eléments de biologie à l'usage d'autres disciplines *(P. Tracqui & J. Demongeot)* • Bioénergétique *(B. Guérin)* • Cinétique enzymatique *(A. Cornish-Bowden, M. Jamin & V. Saks)* • Biodégradations et métabolismes. Les bactéries pour les technologies de l'environnement *(J. Pelmont)* • Enzymologie moléculaire et cellulaire, T. 1 et 2 *(J. Yon-Kahn & G. Hervé)*

La plongée sous-marine à l'air. L'adaptation de l'organisme et ses limites *(Ph. Foster)* • L'Asie, source de sciences et de techniques *(M. Soutif)* • La biologie, des origines à nos jours *(P. Vignais)* • Naissance de la physique. De la Sicile à la Chine *(M. Soutif)* • Le régime oméga 3. Le programme alimentaire pour sauver notre santé *(A. Simopoulos, J. Robinson, M. de Lorgeril & P. Salen)* • Gestes et mouvements justes. Guide de l'ergomotricité pour tous *(M. Gendrier)* • Science expérimentale et connaissance du vivant. La méthode et les concepts *(P. Vignais, avec la collaboration de P. Vignais)*

Listening Comprehension for Scientific English *(J. Upjohn)* • Speaking Skills in Scientific English *(J. Upjohn, M.H. Fries & D. Amadis)* • Minimum Competence in Scientific English *(S. Blattes, V. Jans & J. Upjohn)*

Grenoble Sciences - Rencontres Scientifiques

Radiopharmaceutiques. Chimie des radiotraceurs et applications biologiques *(sous la direction de M. Comet & M. Vidal)* • Turbulence et déterminisme *(sous la direction de M. Lesieur)* • Méthodes et techniques de la chimie organique *(sous la direction de D. Astruc)* • L'énergie de demain. Techniques, environnement, économie *(sous la direction de J.L. Bobin, E. Huffer & H. Nifenecker)* • Physique et biologie. Une interdisciplinarité complexe *(sous la direction de B. Jacrot)*

Introduction

Le présent ouvrage reprend avec beaucoup de compléments un cours de "Licence de Mathématiques" – ex troisième année d'Université – donné à l'Université de Grenoble I pendant les années 1985-88. Le but de ce cours était de présenter aux étudiants quelques notions théoriques de base concernant les équations et systèmes d'équations différentielles ordinaires, tout en explicitant des méthodes numériques permettant de résoudre effectivement de telles équations. C'est pour cette raison qu'une part importante du cours est consacrée à la mise en place d'un certain nombre de techniques fondamentales de l'Analyse Numérique : interpolation polynomiale, intégration numérique, méthode de Newton à une et plusieurs variables.

L'originalité de cet ouvrage ne réside pas tant dans le contenu, pour lequel l'auteur s'est inspiré sans vergogne de la littérature existante – en particulier du livre de Crouzeix-Mignot pour ce qui concerne les méthodes numériques, et des livres classiques de H. Cartan et J. Dieudonné pour la théorie des équations différentielles – mais plutôt dans le choix des thèmes et dans la présentation. S'il est relativement facile de trouver des ouvrages spécialisés consacrés soit aux aspects théoriques fondamentaux de la théorie des équations différentielles et ses applications (Arnold, Coddington-Levinson) soit aux techniques de l'Analyse Numérique (Henrici, Hildebrand), il y a relativement peu d'ouvrages qui couvrent simultanément ces différents aspects et qui se situent à un niveau accessible pour l'«honnête» étudiant de second cycle. Nous avons en particulier consacré deux chapitres entiers à l'étude des méthodes élémentaires de résolution par intégration explicite et à l'étude des équations différentielles linéaires à coefficients constants, ces questions étant généralement omises dans les ouvrages de niveau plus avancé. Par ailleurs, un effort particulier a été fait pour illustrer les principaux résultats par des exemples variés.

La plupart des méthodes numériques exposées avaient pu être effectivement mises en œuvre par les étudiants au moyen de programmes écrits en Turbo Pascal – à une époque remontant maintenant à la préhistoire de l'informatique. Aujourd'hui, les environnements disponibles sont beaucoup plus nombreux, mais nous recommandons certainement encore aux étudiants d'essayer d'implémenter les algorithmes proposés dans ce livre sous forme de programmes écrits dans des langages de base

comme C ou C++, et particulièrement dans un environnement de programmation libre comme GCC sous GNU/Linux. Bien entendu, il existe des logiciels libres spécialisés dans le calcul numérique qui implémentent les principaux algorithmes utiles sous forme de librairies toutes prêtes – Scilab est l'un des plus connus – mais d'un point de vue pédagogique et dans un premier temps au moins, il est bien plus formateur pour les étudiants de mettre vraiment "la main dans le cambouis" en programmant eux-mêmes les algorithmes. Nous ne citerons pas d'environnements ni de logiciels propriétaires équivalents, parce que ces logiciels dont le fonctionnement intime est inaccessible à l'utilisateur sont contraires à notre éthique scientifique ou éducative, et nous ne souhaitons donc pas en encourager l'usage.

L'ensemble des sujets abordés dans le présent ouvrage dépasse sans aucun doute le volume pouvant être traité en une seule année de cours – même si jadis nous avions pu en enseigner l'essentiel au cours de la seule année de Licence. Dans les conditions actuelles, il nous paraît plus judicieux d'envisager une répartition du contenu sur l'ensemble des deux années du second cycle universitaire. Ce texte est probablement utilisable aussi pour les élèves d'écoles d'ingénieurs, ou comme ouvrage de synthèse au niveau de l'agrégation de mathématiques. Pour guider le lecteur dans sa sélection, les sous-sections de chapitres les plus difficiles ainsi que les démonstrations les plus délicates sont marquées d'un astérisque. Le lecteur pourra trouver de nombreux exemples de tracés graphiques de solutions d'équations différentielles dans le livre d'Artigue-Gautheron : on y trouvera en particulier des illustrations variées des phénomènes qualitatifs étudiés au chapitre X, concernant les points singuliers des champs de vecteurs.

Je voudrais ici remercier mes collègues grenoblois pour les remarques et améliorations constantes suggérées tout au long de notre collaboration pendant les trois années qu'a duré ce cours. Mes plus vifs remerciements s'adressent également à Michèle Artigue, Alain Dufresnoy, Jean-René Joly et Marc Rogalski, qui ont bien voulu prendre de leur temps pour relire le manuscrit original de manière très détaillée. Leurs critiques et suggestions ont beaucoup contribué à la mise en forme définitive de cet ouvrage.

Saint-Martin d'Hères, le 5 novembre 1990

La seconde édition de cet ouvrage a bénéficié d'un bon nombre de remarques et de suggestions proposées par Marc Rogalski. Les modifications apportées concernent notamment le début du chapitre VIII, où la notion délicate d'erreur de consistance a été plus clairement explicitée, et les exemples des chapitres VI et XI traitant du mouvement du pendule simple. L'auteur tient à remercier de nouveau Marc Rogalski pour sa précieuse contribution.

Saint-Martin d'Hères, le 26 septembre 1996

La troisième édition de cet ouvrage a été enrichie d'un certain nombre de compléments théoriques et pratiques : comportement géométrique des suites itératives en dimension 1, théorème des fonctions implicites et ses variantes géométriques dans le chapitre IV ; critère de maximalité des solutions dans le chapitre V ; calcul de géodésiques dans le chapitre VI ; quelques exemples et exercices additionnels dans les chapitres suivants ; notions élémentaires sur les flots de champs de vecteurs dans le chapitre XI.

Saint-Martin d'Hères, le 28 février 2006

CHAPITRE I

CALCULS NUMÉRIQUES APPROCHÉS

L'objet de ce chapitre est de mettre en évidence les principales difficultés liées à la pratique des calculs numériques sur ordinateur. Dans beaucoup de situations, il existe des méthodes spécifiques permettant d'accroître à la fois l'efficacité et la précision des calculs.

1. CUMULATION DES ERREURS D'ARRONDI

1.1. REPRÉSENTATION DÉCIMALE APPROCHÉE DES NOMBRES RÉELS

La capacité mémoire d'un ordinateur est par construction finie. Il est donc nécessaire de représenter les nombres réels sous forme approchée. La notation la plus utilisée à l'heure actuelle est la représentation avec virgule flottante : un nombre réel x est codé sous la forme

$$x \simeq \pm\, m \cdot b^p$$

où b est la *base de numération*, m la *mantisse*, et p l'exposant. Les calculs internes sont généralement effectués en base $b = 2$, même si les résultats affichés sont finalement traduits en base 10.

La mantisse m est un nombre écrit avec virgule fixe et possédant un nombre maximum N de chiffres significatifs (imposé par le choix de la taille des emplacements mémoires alloués au type *réel*) : suivant les machines, m s'écrira

- $m = 0, a_1 a_2 \dots a_N = \displaystyle\sum_{k=1}^{N} a_k b^{-k}, \qquad b^{-1} \leq m < 1$;
- $m = a_0, a_1 a_2 \dots a_{N-1} = \displaystyle\sum_{0 \leq k < N} a_k b^{-k}, \quad 1 \leq m < b.$

Ceci entraîne que la précision dans l'approximation d'un nombre réel est toujours une *précision relative* :

$$\frac{\Delta x}{x} = \frac{\Delta m}{m} \leq \frac{b^{-N}}{b^{-1}} = b^{1-N}.$$

On notera $\varepsilon = b^{1-N}$ cette précision relative.

En Langage C standard (ANSI C), les réels peuvent occuper

– pour le type « float », 4 octets de mémoire, soit 1 bit de signe, 23 bits de mantisse et 8 bits d'exposant (dont un pour le signe de l'exposant). Ceci permet de représenter les réels avec une mantisse de 6 à 7 chiffres significatifs après la virgule, dans une échelle allant de 2^{-128} à 2^{127} soit environ de $10^{-38} = 1\,\mathrm{E}-38$ à $10^{38} = 1\,\mathrm{E}+38$. La précision relative est de l'ordre de 10^{-7}.

– pour le type « double », 8 octets de mémoire, soit 1 bit de signe, 51 bits de mantisse et 12 bits d'exposant (dont un pour le signe de l'exposant). Ceci permet de représenter les réels avec une mantisse de 15 chiffres significatifs après la virgule, dans une échelle allant de 2^{-2048} à 2^{2047} soit environ de $10^{-616} = 1\,\mathrm{E}-616$ à $10^{616} = 1\,\mathrm{E}+616$. La précision relative est de l'ordre de 10^{-15}.

1.2. Non-associativité des opérations arithmétiques

Supposons par exemple que les réels soient calculées avec 3 chiffres significatifs et arrondis à la décimale la plus proche. Soit à calculer la somme $x+y+z$ avec

$$x = 8,22, \quad y = 0,00317, \quad z = 0,00432$$

$$\begin{aligned}
(x+y)+z \text{ donne } : \quad & x+y = 8,22317 \simeq 8,22 \\
& (x+y)+z \simeq 8,22432 \simeq 8,22 \\
x+(y+z) \text{ donne } : \quad & y+z = 0,00749 \\
& x+(y+z) = 8,22749 \simeq 8,23.
\end{aligned}$$

L'addition est donc non associative par suite des erreurs d'arrondi !

1.3. Erreur d'arrondi sur une somme

On se propose d'étudier quelques méthodes permettant de *majorer* les erreurs d'arrondi dues aux opérations arithmétiques.

Soient x, y des nombres réels supposés représentés sans erreur avec N chiffres significatifs :

$$x = 0,a_1a_2\ldots a_N \cdot b^p, \quad b^{-1+p} \le x < b^p$$
$$y = 0,a'_1a'_2\ldots a'_N \cdot b^q, \quad b^{-1+q} \le y < b^q$$

Notons $\Delta(x+y)$ l'erreur d'arrondi commise sur le calcul de $x+y$. Supposons par exemple $p \ge q$. S'il n'y a pas débordement, c'est-à-dire si $x+y < b^p$, le calcul de $x+y$ s'accompagne d'une perte des $p-q$ derniers chiffres de y correspondant aux puissances $b^{-k+q} < b^{-N+p}$; donc $\Delta(x+y) \le b^{-N+p}$, alors que $x+y \ge x \ge b^{-1+p}$. En cas de débordement $x+y \ge b^p$ (ce qui se produit par exemple si $p=q$ et $a_1+a'_1 \ge b$), la décimale correspondant à la puissance b^{-N+p} est elle aussi perdue, d'où $\Delta(x+y) \le b^{1-N+p}$. Dans les deux cas :

$$\Delta(x+y) \le \varepsilon(|x|+|y|),$$

où $\varepsilon = b^{1-N}$ est la précision relative décrite au §1.1. Ceci reste vrai quel que soit le signe des nombres x et y.

En général, les réels x, y ne sont eux-mêmes connus que par des valeurs approchées x', y' avec des erreurs respectives $\Delta x = |x' - x|$, $\Delta y = |y' - y|$. A ces erreurs s'ajoute l'erreur d'arrondi

$$\Delta(x' + y') \leq \varepsilon(|x'| + |y'|) \leq \varepsilon(|x| + |y| + \Delta x + \Delta y).$$

Les erreurs Δx, Δy sont elles-mêmes le plus souvent d'ordre ε par rapport à $|x|$ et $|y|$, de sorte que l'on pourra négliger les termes $\varepsilon\Delta x$ et $\varepsilon\Delta y$. On aura donc :

$$\Delta(x + y) \leq \Delta x + \Delta y + \varepsilon(|x| + |y|).$$

Soit plus généralement à calculer une somme $\sum_{k=1}^{n} u_k$ de réels *positifs*. Les sommes partielles $s_k = u_1 + u_2 + \ldots + u_k$ vont se calculer de proche en proche par les formules de récurrence

$$\begin{cases} s_0 = 0 \\ s_k = s_{k-1} + u_k, \qquad k \geq 1. \end{cases}$$

Si les réels u_k sont connus exactement, on aura sur les sommes s_k des erreurs Δs_k telles que $\Delta s_1 = 0$ et

$$\Delta s_k \leq \Delta s_{k-1} + \varepsilon(s_{k-1} + u_k) = \Delta s_{k-1} + \varepsilon s_k.$$

L'erreur globale sur s_n vérifie donc

$$\Delta s_n \leq \varepsilon(s_2 + s_3 + \ldots + s_n),$$

soit

$$\Delta s_n \leq \varepsilon(u_n + 2u_{n-1} + 3u_{n-2} + \ldots + (n-1)u_2 + (n-1)u_1).$$

Comme ce sont les premiers termes sommés qui sont affectés des plus gros coefficients dans l'erreur Δs_n, on en déduit la règle générale suivante (cf. exemple 1.2).

Règle générale – Dans une sommation de réels, l'erreur a tendance à être minimisée lorsqu'on somme en premier les termes ayant la plus petite valeur absolue.

1.4. Erreur d'arrondi sur un produit

Le produit de deux mantisses de N chiffres donne une mantisse de $2N$ ou $2N - 1$ chiffres dont les N ou $N - 1$ derniers vont être perdus. Dans le calcul d'un produit xy (où x, y sont supposés représentés sans erreur) il y aura donc une erreur d'arrondi

$$\Delta(xy) \leq \varepsilon|xy|, \quad \text{où} \quad \varepsilon = b^{1-N}.$$

Si x et y ne sont eux-mêmes connus que par des valeurs approchées x', y' et si $\Delta x = |x' - x|$, $\Delta y = |y' - y|$, on a une erreur initiale

$$\begin{aligned} |x'y' - xy| = |x(y' - y) + (x' - x)y'| &\leq |x|\Delta y + \Delta x|y'| \\ &\leq |x|\Delta y + \Delta x|y| + \Delta x\Delta y. \end{aligned}$$

A cette erreur s'ajoute une erreur d'arrondi

$$\Delta(x'y') \leq \varepsilon|x'y'| \leq \varepsilon(|x| + \Delta x)(|y| + \Delta y).$$

En négligeant les termes $\Delta x \Delta y$, $\varepsilon \Delta x$, $\varepsilon \Delta y$, on obtient la formule approximative

$$\Delta(xy) \leq |x|\Delta y + \Delta x|y| + \varepsilon|xy|. \qquad (*)$$

Soit plus généralement des réels $x_1, \ldots, x_k$, supposés représentés sans erreur. La formule $(*)$ entraîne

$$\Delta(x_1 x_2 \ldots x_k) \leq \Delta(x_1 \ldots x_{k-1})|x_k| + \varepsilon|x_1 \ldots x_{k-1} \cdot x_k|,$$

d'où par une récurrence aisée :

$$\Delta(x_1 x_2 \ldots x_k) \leq (k-1)\varepsilon|x_1 x_2 \ldots x_k|.$$

L'erreur sur un quotient est donnée de même par $\Delta(x/y) \leq \varepsilon|x/y|$. On en déduit pour tous exposants $\alpha_i \in \mathbb{Z}$ la formule générale

$$\Delta(x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}) \leq (|\alpha_1| + \ldots + |\alpha_k| - 1)\varepsilon|x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}| \ ;$$

on observera que $|\alpha_1| + \ldots + |\alpha_k| - 1$ est exactement le nombre d'opérations requises pour calculer $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}$ par multiplications ou divisions successives des x_i.
Contrairement au cas de l'addition, la majoration de l'erreur d'un produit *ne dépend pas de l'ordre des facteurs.*

1.5. Règle de Hörner

On s'intéresse ici au problème de l'évaluation d'un polynôme

$$P(x) = \sum_{k=0}^{n} a_k x^k.$$

La méthode la plus « naïve » qui vient à l'esprit consiste à poser $x^0 = 1$, $s_0 = a_0$, puis à calculer par récurrence

$$\begin{cases} x^k = x^{k-1} \cdot x \\ u_k = a_k \cdot x^k \\ s_k = s_{k-1} + u_k \end{cases} \qquad \text{pour} \quad k \geq 1.$$

Pour chaque valeur de k, deux multiplications et une addition sont donc nécessaires. Il existe en fait une méthode plus efficace :

Règle de Hörner – On factorise $P(x)$ sous la forme :

$$P(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + xa_n)\ldots)).$$

Si l'on pose

$$p_k = a_k + a_{k+1}x + \ldots + a_n x^{n-k},$$

cette méthode revient à calculer $P(x) = p_0$ par récurrence descendante :

$$\begin{cases} p_n = a_n \\ p_{k-1} = a_{k-1} + xp_k, \qquad 1 \leq k \leq n. \end{cases}$$

On effectue ainsi seulement une multiplication et une addition à chaque étape, ce qui économise une multiplication et donc une fraction substantielle du temps d'exécution.

Comparons maintenant les erreurs d'arrondi dans chacune des deux méthodes, en supposant que les réels $x, a_0, a_1, \ldots, a_n$ sont représentés sans erreur.

• **Méthode « naïve ».** On a ici $P(x) = s_n$ avec

$$\begin{aligned} \Delta(a_k \cdot x^k) &\leq k\varepsilon|a_k||x|^k, \\ \Delta s_k &\leq \Delta s_{k-1} + k\varepsilon|a_k||x|^k + \varepsilon(|s_{k-1}| + |u_k|) \\ &\leq \Delta s_{k-1} + k\varepsilon|a_k||x|^k + \varepsilon(|a_0| + |a_1||x| + \ldots + |a_k||x|^k). \end{aligned}$$

Comme $\Delta s_0 = 0$, il vient après sommation sur k :

$$\begin{aligned} \Delta s_n &\leq \sum_{k=1}^{n} k\varepsilon|a_k||x|^k + \varepsilon\sum_{k=1}^{n}(|a_0| + |a_1||x| + \ldots + |a_k||x|^k) \\ &\leq \sum_{k=1}^{n} k\varepsilon|a_k||x|^k + \varepsilon\sum_{k=0}^{n}(n+1-k)|a_k||x|^k. \end{aligned}$$

On obtient par conséquent

$$\Delta P(x) \leq (n+1)\varepsilon\sum_{k=0}^{n}|a_k||x|^k.$$

• **Règle de Hörner.** Dans ce cas, on a

$$\begin{aligned} \Delta p_{k-1} &\leq \Delta(xp_k) + \varepsilon(|a_{k-1}| + |xp_k|) \\ &\leq (|x|\Delta p_k + \varepsilon|xp_k|) + \varepsilon(|a_{k-1}| + |xp_k|) \\ &= \varepsilon(|a_{k-1}| + 2|x||p_k|) + |x|\Delta p_k. \end{aligned}$$

En développant $\Delta P(x) = \Delta p_0$, il vient

$$\Delta p_0 \leq \varepsilon(|a_0| + 2|x||p_1|) + |x|\Big(\varepsilon|a_1| + 2|x||p_2| + |x|\Big(\varepsilon|a_2| + \ldots\Big)\Big)$$

d'où

$$\begin{aligned} \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}|a_k||x|^k + 2\varepsilon\sum_{k=1}^{n}|x|^k|p_k|, \\ \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}|a_k||x|^k + 2\varepsilon\sum_{k=1}^{n}(|a_k||x|^k + \ldots + |a_n||x|^n), \\ \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}(2k+1)|a_k||x|^k. \end{aligned}$$

On voit que la somme des coefficients d'erreur affectés aux termes $|a_k||x|^k$, soit $\varepsilon\sum_{k=0}^{n}(2k+1) = \varepsilon(n+1)^2$, est la même que pour la méthode naïve ; comme $2k+1 \leq 2(n+1)$, l'erreur commise sera dans le pire des cas égale à 2 fois celle de la méthode naïve. Néanmoins, les petits coefficients portent sur les premiers termes calculés, de sorte que la précision de la méthode de Hörner sera nettement meilleure si le terme $|a_k||x|^k$ décroît rapidement : c'est le cas par exemple si $P(x)$ est le début d'une série convergente.

Exercice – *Evaluer dans les deux cas l'erreur commise sur les sommes partielles de la série exponentielle*

$$\sum_{k=0}^{n}\frac{x^k}{k!}, \quad x \geq 0$$

en tenant compte du fait qu'on a une certaine erreur d'arrondi sur $a_k = \frac{1}{k!}$.

Réponse. *On trouve* $\Delta P(x) \leq \varepsilon(1+(n+x)e^x)$ *pour la méthode naïve, tandis que la factorisation*

$$P(x) = 1 + x\Big(1 + \frac{x}{2}\Big(1 + \frac{x}{3}\Big(1 + \ldots\Big(1 + \frac{x}{n-1}\Big(1 + \frac{x}{n}\Big)\Big)\ldots\Big)\Big)\Big)$$

donne $\Delta P(x) \leq \varepsilon(1+3xe^x)$, *ce qui est nettement meilleur en pratique puisque* n *doit être choisi assez grand.*

1.6. CUMULATION D'ERREURS D'ARRONDI ALÉATOIRES

Les majorations d'erreurs que nous avons données plus haut pêchent en général par excès de pessimisme, car nous n'avons tenu compte que de la valeur absolue des erreurs, alors qu'en pratique elles sont souvent de signe aléatoire et se compensent donc partiellement entre elles.

Supposons par exemple qu'on cherche à calculer une somme s_n de rang élevé d'une série convergente $S = \sum_{k=0}^{+\infty} u_k$, les u_k étant des réels ≥ 0 supposés représentés sans erreur. On pose donc

$$s_k = s_{k-1} + u_k, \quad s_0 = u_0,$$

et les erreurs Δs_k vérifient

$$\begin{aligned} &\Delta s_k = \Delta s_{k-1} + \alpha_k \\ \text{avec} \quad &\Delta s_0 = 0 \quad \text{et} \quad |\alpha_k| \leq \varepsilon(s_{k-1} + u_k) = \varepsilon s_k \leq \varepsilon S. \end{aligned}$$

On en déduit donc

$$\Delta s_n = \alpha_1 + \alpha_2 + \ldots + \alpha_n$$

et en particulier $|\Delta s_n| \leq n\varepsilon S$. Dans le pire des cas, l'erreur est donc proportionnelle à n. On va voir qu'on peut en fait espérer beaucoup mieux sous des hypothèses raisonnables.

Hypothèses

(1) Les erreurs α_k sont des variables aléatoires globalement indépendantes les unes des autres (lorsque les u_k sont choisis aléatoirement).

(2) L'espérance mathématique $E(\alpha_k)$ est nulle, ce qui signifie que les erreurs d'arrondi n'ont aucune tendance à se faire par excès ou par défaut.

L'hypothèse (2) entraîne $E(\Delta s_n) = 0$ tandis que l'hypothèse (1) donne

$$\mathrm{var}(\Delta s_n) = \mathrm{var}(\alpha_1) + \ldots + \mathrm{var}(\alpha_n).$$

Comme $E(\alpha_k) = 0$ et $|\alpha_k| \leq \varepsilon S$, on a $\mathrm{var}(\alpha_k) \leq \varepsilon^2 S^2$, d'où

$$\sigma(\Delta s_n) = \sqrt{\mathrm{var}(\Delta s_n)} \leq \sqrt{n}\varepsilon S$$

L'erreur quadratique moyenne croît seulement dans ce cas comme $\sqrt{n}$. D'après l'inégalité de Bienaymé-Tchebychev on a :

$$P(|\Delta s_n| \geq \alpha\, \sigma(\Delta s_n)) \leq \alpha^{-2}.$$

La probabilité que l'erreur dépasse $10\sqrt{n}\varepsilon S$ est donc inférieure à 1%.

2. Phénomènes de compensation

Les phénomènes de compensation se produisent lorsqu'on tente d'effectuer des soustractions de valeurs très voisines. Ils peuvent conduire à des pertes importantes de précision.

Les exemples suivants illustrent les difficultés pouvant se présenter et les remèdes à apporter dans chaque cas.

2.1. Exemple : résolution de l'équation $x^2 - 1634x + 2 = 0$

Supposons que les calculs soient effectués avec 10 chiffres significatifs. Les formules habituelles donnent alors

$$\begin{aligned}
&\Delta' = 667\,487, \qquad \sqrt{\Delta'} \simeq 816,9987760\\
&x_1 = 817 + \sqrt{\Delta'} \simeq 1633,998776,\\
&x_2 = 817 - \sqrt{\Delta'} \simeq 0,0012240.
\end{aligned}$$

On voit donc qu'on a une perte de 5 chiffres significatifs sur x_2 si l'on effectue la soustraction telle qu'elle se présente naturellement ! Ici, le remède est simple : il suffit d'observer que $x_1 x_2 = 2$, d'où

$$x_2 = \frac{2}{x_1} \simeq 1,223991125 \cdot 10^{-3}.$$

C'est donc l'*algorithme numérique utilisé qui doit être modifié.*

2.2. EXEMPLE : CALCUL APPROCHÉ DE e^{-10}

Supposons qu'on utilise pour cela la série

$$e^{-10} \simeq \sum_{k=0}^{n} (-1)^k \frac{10^k}{k!},$$

les calculs étant toujours effectués avec 10 chiffres significatifs. Le terme général $|u_k| = 10^k/k!$ est tel que

$$\frac{|u_k|}{|u_{k-1}|} = \frac{10}{k} \geq 1 \quad \text{dès que} \quad k \leq 10.$$

On a donc 2 termes de valeur absolue maximale

$$|u_9| = |u_{10}| = \frac{10^{10}}{10!} \simeq 2,755 \cdot 10^3$$

tandis que $e^{-10} \simeq 4,5 \cdot 10^{-5}$. Comparons u_{10} et e^{-10} :

u_{10} :	2	7	5	5,	·	·	·	·	·	·
e^{-10} :				0,	0	0	0	0	4	5

Ceci signifie qu'au moins 8 chiffres significatifs vont être perdus par compensation des termes u_k de signes opposés. Un remède simple consiste à utiliser la relation

$$e^{-10} = 1/e^{10} \quad \text{avec} \quad e^{10} \simeq \sum_{k=0}^{n} \frac{10^k}{k!}.$$

On essaiera dans la mesure du possible d'*éviter les sommations dans lesquelles des termes de signes opposés se compensent.*

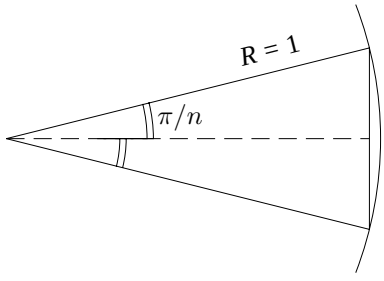
2.3. EXEMPLE : CALCUL APPROCHÉ DE π PAR LES POLYGÔNES INSCRITS

Soit P_n le demi-périmètre du polygone régulier à n côtés inscrit dans un cercle de rayon 1.

Le côté de ce polygone vaut $2 \cdot R \sin \pi/n = 2 \sin \pi/n$, d'où

$$P_n = n \sin \frac{\pi}{n},$$
$$P_n = \pi - \frac{\pi^3}{6n^2} + o\Big(\frac{1}{n^3}\Big).$$

On obtient donc une approximation de π avec une précision de l'ordre de $6/n^2$.



Pour évaluer P_n, on utilise une méthode de dichotomie permettant de calculer par récurrence

$$x_k = P_{2^k} = 2^k \sin \frac{\pi}{2^k}.$$

Si α est un angle compris entre 0 et $\frac{\pi}{2}$ on a

$$\sin \frac{\alpha}{2} = \sqrt{\frac{1}{2}(1 - \cos \alpha)} = \sqrt{\frac{1}{2}(1 - \sqrt{1 - \sin^2 \alpha})}. \qquad (*)$$

En substituant $\alpha = \pi/2^k$, on en déduit les formules

$$\begin{cases} x_{k+1} = 2^k \sqrt{2(1 - \sqrt{1 - (x_k/2^k)^2})} \\ x_1 = 2, \end{cases}$$

et d'après ce qu'on a dit plus haut $\lim\limits_{k \to +\infty} x_k = \pi$.

Ce n'est pourtant pas du tout ce qu'on va observer sur machine ! Dès que $(x_k/2^k)^2$ sera inférieur à la précision relative des calculs, l'ordinateur va donner

$$\sqrt{1 - (x_k/2^k)^2} = 1 \quad \text{d'où} \quad x_{k+1} = 0.$$

Pour éviter cette difficulté, il suffit de remplacer $(*)$ par

$$\sin \frac{\alpha}{2} = \sqrt{\frac{1}{2} \, \frac{1 - \cos^2 \alpha}{1 + \cos \alpha}} = \frac{\sin \alpha}{\sqrt{2(1 + \cos \alpha)}} \qquad (**)$$

d'où

$$\sin \frac{\alpha}{2} = \frac{\sin \alpha}{\sqrt{2(1 + \sqrt{1 - \sin^2 \alpha})}}.$$

On obtient alors la formule de récurrence

$$x_{k+1} = \frac{2x_k}{\sqrt{2(1 + \sqrt{1 - (x_k/2^k)^2})}}$$

qui évite le phénomène de compensation précédent, de sorte que le calcul des x_k peut être poussé beaucoup plus loin.

On obtiendra une méthode plus efficace encore en observant qu'on peut évaluer $\cos\alpha$ dans $(**)$ par la formule $\cos\alpha = \frac{\sin 2\alpha}{2\sin\alpha}$. Ceci donne

$$\sin\frac{\alpha}{2} = \sin\alpha\sqrt{\frac{\sin\alpha}{2\sin\alpha+\sin 2\alpha}}, \quad \text{d'où}$$
$$x_{k+1} = x_k\sqrt{\frac{2x_k}{x_k+x_{k-1}}}.$$

Deux valeurs d'initialisation sont alors requises pour démarrer, par exemple $x_1 = 2$ et $x_2 = 2\sqrt{2}$.

3. PHÉNOMÈNES D'INSTABILITÉ NUMÉRIQUE

Il s'agit de phénomènes d'amplification des erreurs d'arrondi. Une telle amplification se produit assez fréquemment dans le cas de calculs récurrents ou itératifs.

3.1. CAS D'UN CALCUL RÉCURRENT

Supposons à titre d'exemple qu'on cherche à évaluer numériquement l'intégrale

$$I_n = \int_0^1 \frac{x^n}{10+x}, \qquad n\in\mathbb{N}.$$

Un calcul immédiat donne

$$I_0 = \int_0^1 \frac{dx}{10+x} = \Big[\ln(10+x)\Big]_0^1 = \ln\frac{11}{10},$$
$$I_n = \int_0^1 \frac{x}{10+x}\cdot x^{n-1}dx = \int_0^1 \Big(1-\frac{10}{10+x}\Big)x^{n-1}dx$$
$$= \int_0^1 x^{n-1}dx - 10\int_0^1 \frac{x^{n-1}}{10+x}\,dx = \frac{1}{n} - 10\,I_{n-1}.$$

Ceci permet de calculer I_n par récurrence avec

$$\begin{cases} I_0 = \ln\frac{11}{10} \\ I_n = \frac{1}{n} - 10\,I_{n-1}. \end{cases}$$

Ce problème apparemment bien posé mathématiquement conduit numériquement à des résultats catastrophiques. On a en effet

$$\Delta I_n \simeq 10\,\Delta I_{n-1},$$

même si on néglige l'erreur d'arrondi sur $1/n$. L'erreur sur I_n explose donc exponentiellement, l'erreur initiale sur I_0 étant multipliée par 10^n à l'étape n. Comment faire alors pour calculer par exemple I_{36} ? La suite x^n étant décroissante

pour $x \in [0,1]$, on voit que la suite I_n est elle-même décroissante. Comme $10 \leq 10+x \leq 11$, on a de plus

$$\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{10(n+1)}.$$

L'approximation $I_n \simeq \frac{1}{11(n+1)}$ donne une erreur absolue $\leq \frac{1}{110(n+1)}$ et donc une erreur relative $\frac{\Delta I_n}{I_n} \leq \frac{1}{10}$. Ceci donne l'ordre de grandeur mais n'est pas très satisfaisant. L'idée est alors de *renverser la récurrence* en posant

$$I_{n-1} = \frac{1}{10}\Big(\frac{1}{n} - I_n\Big).$$

En négligeant l'erreur sur $\frac{1}{n}$, on a donc cette fois $\Delta I_{n-1} \simeq \frac{1}{10}\,\Delta I_n$, estimation qui va dans le bon sens. Si l'on part de $I_{46} \simeq \frac{1}{11.47}$, on obtiendra pour I_{36} une erreur relative sans doute meilleure que 10^{-10}.

Exercice – *Montrer que* $0 \leq I_n - I_{n+1} \leq \frac{1}{10(n+1)(n+2)}$, *et en déduire à partir de la formule exprimant* I_n *en fonction de* I_{n+1} *que l'on a en fait l'estimation*

$$\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{11(n+1)} + \frac{1}{110(n+1)(n+2)},$$

donc $\Delta I_n \simeq \frac{1}{11(n+1)}$ *avec erreur relative* $\leq \frac{1}{10(n+2)}$.

On voit donc le rôle fondamental joué par le coefficient d'amplification de l'erreur, 10 dans le premier cas, 1/10 dans le second. En général, si on a un coefficient d'amplification $A > 1$, il est impératif de limiter le nombre n d'étapes en sorte que $A^n\varepsilon$ reste très inférieur à 1, si ε est la précision relative des calculs.

3.2. Calculs itératifs

Soit à calculer une suite (u_n) définie par sa valeur initiale u_0 et par la relation de récurrence

$$u_{n+1} = f(u_n),$$

où f est une fonction donnée. On a donc $u_n = f^n(u_0)$ où $f^n = f \circ f \circ \ldots \circ f$ est la n-ième itérée de f. On considère par exemple la suite (u_n) telle que

$$u_0 = 2, \qquad u_{n+1} = |\ln(u_n)|,$$

dont on cherche à évaluer le terme u_{30}. Un calcul effectué à la précision 10^{-9} sur un ordinateur nous a donné $u_{30} \simeq 0,880833175$.

A la lumière de l'exemple précédent, il est néanmoins légitime de se demander si ce calcul est bien significatif, compte tenu de la présence des erreurs d'arrondi. En partant de valeurs de u_0 très voisines de 2, on obtient en fait les résultats suivants

(arrondis à 10^{-9} près, sur la même implémentation de calcul que ci-dessus) :

u_0	2,000000000	2,000000001	1,999999999	$5 \cdot 10^{-10}$
u_5	5,595485181	5,595484655	5,595485710	$9 \cdot 10^{-8}$
u_{10}	0,703934587	0,703934920	0,703934252	$5 \cdot 10^{-7}$
u_{15}	1,126698502	1,126689382	1,126707697	$8 \cdot 10^{-6}$
u_{20}	1,266106839	1,266256924	1,265955552	10^{-4}
u_{24}	1,000976376	1,001923276	1,000022532	10^{-3}
u_{25}	0,000975900	0,001921429	0,000022532	100%
u_{26}	6,932150628	6,254686211	10,700574400	50%
u_{30}	0,880833175	0,691841353	1,915129896	100%

La dernière colonne donne l'ordre de grandeur de l'écart relatif $\Delta u_n/u_n$ observé entre la deuxième ou troisième colonne et la première colonne. On voit que cet écart augmente constamment pour atteindre environ 10^{-3} sur u_{24}. Pour le calcul de u_{25}, il se produit une véritable catastrophe numérique : l'écart relatif devient voisin de 100% ! Il en résulte que toutes les valeurs calculées à partir de u_{25} sont certainement non significatives pour une précision des calculs de 10^{-9}.

Pour comprendre ce phénomène, il suffit d'observer qu'une erreur Δx sur la variable x entraîne une erreur $\Delta f(x)$ sur $f(x)$, approximativement donnée par

$$\Delta f(x) = |f'(x)|\,\Delta x.$$

Ceci se voit bien sûr en approximant $f(x+\Delta x) - f(x)$ par sa différentielle $f'(x)\Delta x$, lorsque f est dérivable au point x. Le coefficient d'amplification de l'erreur absolue est donc donné par la valeur absolue de la dérivée $|f'(x)|$; ce coefficient peut être parfois assez grand. Souvent dans les calculs numériques (et ici en particulier), il est plus pertinent de considérer les erreurs relatives. La formule

$$\frac{\Delta f(x)}{|f(x)|} = \frac{|f'(x)||x|}{|f(x)|}\,\frac{\Delta x}{|x|}$$

montre que le coefficient d'amplification de l'erreur relative est $|f'(x)||x|/|f(x)|$. Dans le cas $f(x) = \ln(x)$ qui nous intéresse, ce coefficient vaut $1/|\ln x|$; il devient très grand lorsque x est proche de 1, comme c'est le cas par exemple pour u_{24}.

4. PROBLÈMES

4.1. Soit $x \geq 0$; on note $F(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x e^{-t^2}\,dt$.

(a) Encadrer $F(x)$ par deux entiers consécutifs.

(b) En remplaçant e^{-t^2} par un développement en série entière de x, exprimer $F(x)$ comme somme d'une série. On choisit $x = 3$; calculer les 10 premiers termes