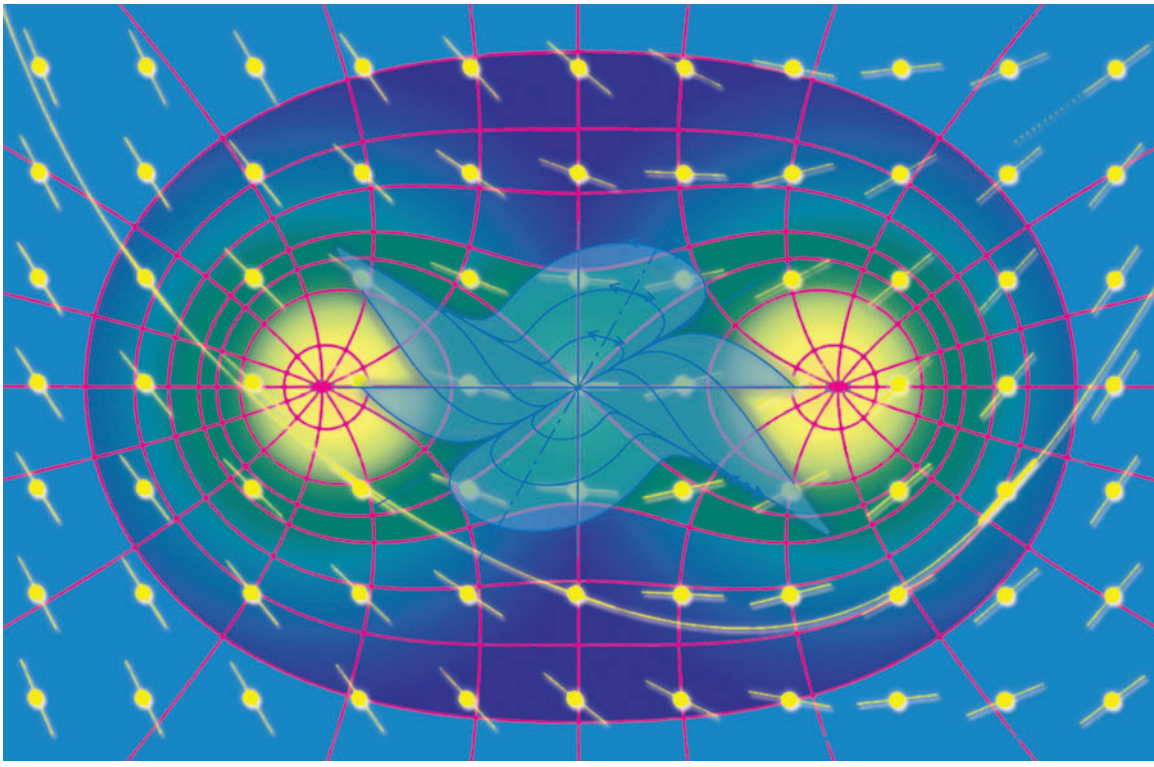
COLLECTION GRENOBLE SCIENCES
DIRIGÉE PAR JEAN BORNAREL

ANALYSE NUMÉRIQUE ET ÉQUATIONS DIFFÉRENTIELLES

Nouvelle édition avec exercices corrigés

Jean-Pierre DEMAILLY



edp sciences

Analyse numérique et équations différentielles

Grenoble Sciences

Grenoble Sciences est un centre de conseil, expertise et labellisation de l'enseignement supérieur français. Il expertise les projets scientifiques des auteurs dans une démarche à plusieurs niveaux (référés anonymes, comité de lecture interactif) qui permet la labellisation des meilleurs projets après leur optimisation. Les ouvrages labellisés dans une collection de Grenoble Sciences correspondent à :

- des projets clairement définis sans contrainte de mode ou de programme,
- des qualités scientifiques et pédagogiques certifiées par le mode de sélection,
- une qualité de réalisation assurée par le centre technique de Grenoble Sciences.

Livres et pap-ebooks

Grenoble Sciences labellise des livres papier (en langue française et en langue anglaise) mais également des ouvrages utilisant d'autres supports. Dans ce contexte, situons le concept de pap-ebook. Celui-ci se compose de deux éléments :

- un **livre papier** qui demeure l'objet central,
- un **site web compagnon** qui propose :
  - des éléments permettant de combler les lacunes du lecteur qui ne posséderait pas les prérequis nécessaires à une utilisation optimale de l'ouvrage,
  - des exercices pour s'entraîner,
  - des compléments pour approfondir un thème, trouver des liens sur internet, etc.

Le livre du pap-ebook est autosuffisant et certains lecteurs n'utiliseront pas le site web compagnon. D'autres l'utiliseront et ce, chacun à sa manière. Un livre qui fait partie d'un pap-ebook porte en première de couverture un logo caractéristique et le lecteur trouvera le site compagnon du présent ouvrage à l'adresse internet suivante :

https://grenoble-sciences.ujf-grenoble.fr/pap-ebook/demailly

Grenoble Sciences bénéficie du soutien de la **région Auvergne-Rhône-Alpes** et du **ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche**. Grenoble Sciences est rattaché à l'**Université Grenoble Alpes**.

ISBN 978 2 7598 1926 3

Analyse numérique et équations différentielles

Jean-Pierre Demailly

edp sciences

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

Analyse numérique et équations différentielles

Cet ouvrage, labellisé par Grenoble Sciences, est un des titres du secteur Mathématiques de la collection Grenoble Sciences d'EDP Sciences, qui regroupe des projets originaux et de qualité. Cette collection est dirigée par Jean Bornarel, Professeur émérite à l'Université Grenoble Alpes.

Comité de lecture de l'édition précédente :

- 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 Technologies, Lille 1.

Cette nouvelle édition a été suivie par Stéphanie Trine. L'illustration de couverture est l'œuvre d'Alice Giraud.

Autres ouvrages labellisés sur des thèmes proches (chez le même éditeur)

Méthodes numériques appliquées pour le scientifique et l'ingénieur (J.-P. Grivet) • Petit traité d'intégration (J.-Y. Briend) • Introduction aux variétés différentielles (J. Lafontaine) • Nombres et algèbre (J.-Y. Mérindol) • Exercices corrigés d'analyse avec rappels de cours. Tomes I et II (D. Alibert) • Outils mathématiques à l'usage des scientifiques et ingénieurs (E. Belorizky) • Mathématiques pour l'étudiant scientifique. Tomes I et II (P.-J. Haug) • 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) • Mathématiques pour les sciences de la Vie, de la Nature et de la Santé (J.-P. Bertrandias & F. Bertrandias) • Description de la symétrie. Des groupes de symétrie aux structures fractales (J. Sivardière) • Symétrie et propriétés physiques. Des principes de Curie aux brisures de symétrie (J. Sivardière) • Approximation hilbertienne. Splines, ondelettes, fractales (M. Attéia & J. Gaches) • Analyse statistique des données expérimentales (K. Protassov) • Introduction à la mécanique statistique (E. Belorizky & W. Gorecki) • Mécanique statistique. Exercices et problèmes corrigés (E. Belorizky & W. Gorecki) • Magnétisme : I Fondements, II Matériaux (sous la direction d'E. du Trémolet de Lacheisserie) • La mécanique quantique. Problèmes résolus. Tomes I et II (V.M. Galitski, B.M. Karnakov & V.I. Kogan) • Relativité générale et astrophysique, problèmes et exercices corrigés (Denis Gialis & François-Xavier Désert) • Éléments de Biologie à l'usage d'autres disciplines. De la structure aux fonctions (P. Tracqui & J. Demongeot) • Minimum Competence in Scientific English (S. Blattes, V. Jans & J. Upjohn)

et d'autres titres sur le site internet
https://grenoble-sciences.ujf-grenoble.fr

Table des matières

This page intentionally left blank

Introduction

Le présent ouvrage reprend avec beaucoup de compléments un cours de « Licence de Mathématiques » – ce qui autrefois désignait la 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 deuxième 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

La quatrième édition de cet ouvrage se mue en pap-ebook : le lecteur trouvera sur le site compagnon du livre des solutions d'exercices et de multiples compléments théoriques. Consulter pour cela

`https://grenoble-sciences.ujf-grenoble.fr/pap-ebook/demailly`

Par ailleurs, quelques énoncés d'exercices et de problèmes ont été ajoutés au fil des chapitres, et de nombreuses coquilles typographiques présentes dans les versions précédentes ont été expurgées ; l'auteur tient à remercier chaleureusement Mme Stéphanie Trine pour sa relecture très attentive et ses suggestions d'amélioration de la présentation.

Saint-Martin d'Hères, le 8 février 2016

This page intentionally left blank

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 calculés 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 :

$$\begin{aligned}
x &= 0, a_1 a_2 \dots a_N \cdot b^p, \quad b^{-1+p} \leq x < b^p \\
y &= 0, a'_1 a'_2 \dots a'_N \cdot b^q, \quad b^{-1+q} \leq y < b^q
\end{aligned}$$

Notons $\Delta(x+y)$ l'erreur d'arrondi commise sur le calcul de $x + y$. Supposons par exemple $p \geq q$. 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 $\Delta(x+y) \leq b^{-N+p}$, alors que $x + y \geq x \geq b^{-1+p}$.

En cas de débordement $x + y \geq b^p$ (ce qui se produit par exemple si $p = q$ et $a_1 + a'_1 \geq b$), la décimale correspondant à la puissance b^{-N+p} est elle aussi perdue, d'où $\Delta(x + y) \leq b^{1-N+p}$. Dans les deux cas :

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

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. Section 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

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

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

$$\Delta(x_1x_2\ldots x_k) \leq \Delta(x_1\ldots x_{k-1})|x_k| + \varepsilon|x_1\ldots x_{k-1}\cdot x_k|,$$

d'où par une récurrence aisée :

$$\Delta(x_1x_2\ldots x_k) \leq (k-1)\varepsilon|x_1x_2\ldots x_k|.$$

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 :

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

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

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

d'où

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

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 A. 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

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

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

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

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

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

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%.

Index des notations