Françoise Chaitin–Chatelin

Professeur Université Paris IX Dauphine
et CERFACS
42, Av. G. Coriolis
31057 Toulouse cedex
Tel (33) 61 19 30 26
Email : chatelin@cerfacs.fr

Support de cours pour le DEA intensif
Problèmes inverses en Astrophysique.
Concepts mathématiques et méthodes de résolution.

CIRM–Luminy, 20 – 24 Mars 1995

Le calcul qualitatif.

Comment donner un sens à des résultats faux ?

CERFACS REPORT TR/PA/95/10

Résumé

Pour un ingénieur ou un chercheur, la confiance dans les résultats de grands codes de calcul et dans la robustesse des modèles qu'il utilise est cruciale. En effet, l'émergence des supercalculateurs a permis d'utiliser intensivement la simulation numérique à précision finie comme substitut de l'expérimentation physique, y compris pour des problèmes aux frontières de l'instable. Toutefois, rares sont les outils utilisés dans l'industrie qui permettent d'obtenir des informations sur la qualité et la validité des résultats numériques obtenus sur ordinateur.

Le document qui suit propose un tour d'horizon des méthodes d'analyse de la fiabilité des logiciels. Après avoir montré pourquoi les outils fondés sur les modèles probabilistes sont tombés en désuétude, on dresse un large panorama de l'analyse inverse des erreurs. On montre pourquoi le calcul en précision finie au voisinage de singularités peut être dangereux. Le danger est d'autant plus grand que les singularités peuvent être "rapprochées" par la non normalité des problèmes à traiter. Et lorsque les erreurs d'arrondissage dominent (calculs chaotiques) au point que tous les chiffres soient faux, la quantification (calculer juste) perd son sens. Il ne s'agit plus de contrôler l'erreur qui ne tend plus vers zéro, mais de l'interpréter, de lui donner un **sens**. En effet, le calcul continue à délivrer une certaine information, voilée par le brouillard des erreurs, mais pertinente par rapport à la singularité sous-jacente. Bien loin d'être un handicap, le caractère fini de l'information traitée par l'ordinateur peut être un atout quand il s'agit d'analyser les phénomènes naturels par simulation.

Table des matières

1 Introduction : faut-il avoir peur du calcul sur ordinateur ?

Périodiquement, les magazines scientifiques font paraître un article attirant l'attention du lecteur sur les dangers auxquels l'arithmétique à précision finie exposerait les simulations sur ordinateur (Dinnematin (1987), Colonna (1992)). On évoque un *effet "boule de neige"* : plus les calculs sont longs, plus les erreurs s'accumulent, plus le danger augmente.
La presse spécialisée n'est guère plus rassurante qui parle d'élimination catastrophique, "overflow", "underflow", etc. Tout Cours d'Analyse Numérique débute par un chapitre mettant en garde les étudiants contre les pièges et chausse-trappes du calcul sur ordinateur. Mais l'étudiant sait d'expérience que les avions volent, que les sondes explorent le voisinage de planètes lointaines. Il sait aussi que ces avions, ces sondes ont été calculés sur ordinateur. Alors, catastrophisme d'un côté, efficacité de l'autre ? Où est la vérité ? Peut-on faire confiance aux ordinateurs ?

L'ingénieur, lui, fait confiance aux ordinateurs. Il n'a pas le choix et il les pratique tous les jours. Mais il sait aussi que de temps en temps, de manière semble-t-il imprévisible, des résultats seront douteux, ou même faux, alors que ni l'ordinateur ni le programme ne sont en cause. Il voudrait, face à de tels cas, disposer d'outils d'analyse de la fiabilité (Erhel et Philippe (1989)). Car la question de la confiance qu'il peut accorder aux résultats massivement produits par les super-ordinateurs se pose avec d'autant plus d'acuité que les problèmes à résoudre sont de plus en plus difficiles mathématiquement (non linéaires et mal posés). Il se demande si les architectures massivement parallèles ne menacent pas la fiabilité des logiciels éprouvée sur des architectures traditionnelles. Car il se souvient que, dans les années 80, l'introduction dans les bureaux d'études de stations de travail à l'arithmétique "correcte" (voir la section 3.2) a eu des conséquences inattendues. Les nouveaux résultats, plus précis, contredisaient parfois ceux qui avaient été établis sur les gros ordinateurs (mainframes) dont l'arithmétique était moins bonne.
Dans de très nombreuses applications liées à la sécurité, la question de la fiabilité des calculs est prédominante (Carvallo (1852), Marshall (1992)).

Toutes ces questions n'ont cessé de se poser, sous des formes renouvelées, depuis l'avènement des ordinateurs. Voyons le type de réponses qui ont été successivement apportées depuis un demi-siècle.

2 Perspective historique

Le souci de tenir compte de l'effet global de **l'addition** d'un très grand nombre d'erreurs expérimentales remonte au moins à Pierre-Simon, Marquis de Laplace (1778). C'est afin de modéliser cet effet par une variable aléatoire qu'il introduisit la loi de distribution dite "normale" (Laplace (1812)). Cette loi des erreurs expérimentales a été l'objet de recherches très actives en probabilités et statistiques. Nous citons pour mémoire les nombreuses versions de la loi des grands nombres et des théorèmes de la limite centrale. Plus récemment, les travaux de Bernstein (1927), von Mises (1936), Levy (1937) et Cramer (1946), se sont attachés, entre autres choses, à affaiblir les hypothèses d'indépendance entre les erreurs individuelles.

Jusqu'au milieu du XX^{e} siècle, l'intérêt des théoriciens s'est surtout porté sur les conséquences des erreurs dans les **Sciences Expérimentales**, et principalement la Physique. Avec pour seule exception notoire les travaux de certains astronomes, tels Newcomb au XIX^{e} siècle, Schlesinger (1917) et Brouwer (1937). Comme les calculs se faisaient à la main, ou avec une machine arithmétique mécanique, le comportement cumulatif des erreurs pouvait être contrôlé à chaque pas de calcul.

Ce n'est qu'avec l'apparition des "calculateurs électroniques" à la fin de la deuxième guerre mondiale, qui donnaient accès à des calculs de taille inconnue jusqu'alors, que certains scientifiques s'inquiétèrent réellement des conséquences possibles des erreurs d'arrondissage sur la précision obtenue en fin de calcul. Mais ils découvrirent assez vite que cet effet n'était pas aussi dramatique qu'ils l'avaient craint.
Les célèbres papiers de von Neumann et Goldstine (1947, 1951), ainsi que Turing (1948) se situent dans la lignée des travaux sur les erreurs expérimentales et inaugurent pour le calcul sur ordinateur une période où le comportement des erreurs d'arrondissage est simulé, afin d'être analysé, à l'aide de *modèles probabilistes*, qu'il s'agisse du résultat d'une longue liste de calcul (von Neumann et Goldstine (1947), Turing (1948), puis Henrici (1962a, 1963), Hull et Swenson (1966)), ou du résultat d'une opération arithmétique élémentaire (Feldstein et Goodman (1976)). Cette approche élégante, que nous détaillerons au paragraphe 4, ne permet pas en pratique d'estimer l'erreur commise. Elle fut donc peu à peu abandonnée à partir des années 1970 par les utilisateurs, à l'exception notable de l'école française de Vignes (La Porte et Vignes (1974),Pichat et Vignes (1993)).

Cet abandon progressif se fit au profit d'une approche déterministe non moins élégante, la méthode dite d'*analyse inverse des erreurs* dans laquelle la solution d'un problème donné calculée sur ordinateur est considérée comme la solution exacte d'un problème perturbé de même type. L'idée figure en germe dans Givens (1954), mais ce fut Wilkinson (1961, 1963) à NPL (National Physical Laboratory, Teddington) qui la développa brillamment durant les trois décennies qui suivirent.
On verra plus loin que l'avantage déterminant de l'approche *inverse* est de permettre de distinguer dans l'estimation de l'erreur calculée la part qui revient à l'algorithme de celle qui revient au problème. Cette approche s'est définitivement imposée dans le domaine de l'Algèbre Matricielle Numérique pour établir la fiabilité arithmétique des logiciels numériques. En témoigne la réalisation de la bibliothèque LAPACK (1991) offrant des logiciels de toute première qualité pour résoudre sur ordinateur des problèmes essentiellement **linéaires** (résolution de systèmes linéaires, inversion de matrices, moindres carrés, valeurs et vecteurs propres). L'analyse inverse permet de garantir la "stabilité inverse", c'est à dire la fiabilité arithmétique des algorithmes.
Si l'erreur obtenue après emploi d'un algorithme fiable est jugée trop grande, cela est dû à ce que le problème posé présente lui-même une trop grande instabilité. Cette instabilité est souvent la conséquence d'une **non-linéarité**, comme on le sait depuis les travaux du météorologue Lorentz (1963) et du biologiste May (1976).

Jusqu'à présent, l'approche classique qui consiste à établir la convergence numérique des méthodes d'approximation en supposant que l'arithmétique soit exacte, s'est révélée globalement suffisante pour expliquer les comportements constatés en précision finie. Dans de tels cas, la perturbation introduite par le passage sur ordinateur n'est pas suffisante pour exciter l'instabilité potentielle de l'équation.
Mais les nouveaux problèmes à la limite de l'instable rencontrés par la physique et la technologie rendent l'approche ci-dessus souvent insuffisante. Il devient alors impératif de tenir compte du fait que la stabilité de certains processus de calculs peut devenir **conditionnelle à l'arithmétique** (Chaitin-Chatelin et Frayssé (1995), Trefethen et Trummer (1987)).
De telles situations critiques se rencontrent lorsque le calcul est effectué au voisinage d'un point **singulier**, c'est-à-dire d'un point **non régulier**, d'un point où la solution ne dépend pas de manière continuement dérivable des données du problème. Le calcul en un point régulier représente la situation classique, là où les calculs se passent bien : il n'y a pas de différence significative entre résultat exact et résultat calculé sur un ordinateur de précision suffisante.
En un point singulier, par contre, la différence entre arithmétique exacte et arithmétique à précision

finie peut devenir très importante, d'une manière qui dépend de la nature de la singularité. Or une singularité n'est pas générique : elle disparaît par perturbation en général. C'est ce qui se passe lors du calcul sur ordinateur sous l'effet de perturbations engendrées par la précision finie. Ceci explique que l'on résolve toujours un problème régulier sur ordinateur mais que, néanmoins, les calculs soient guidés par la singularité sous-jacente, tant que cette singularité exerce une influence. C'est alors que les résultats obtenus sur ordinateur peuvent être faux.

Les physiciens ont longtemps cru que les singularités n'étaient que des curiosités mathématiques puisque les quantités physiques qui ont un sens ne peuvent être infinies. Mais la physique contemporaine a fermement établi leur rôle essentiel, par exemple dans la réduction des théories en physique fondamentale (Berry (1991)). Assez souvent, une théorie générale fondée sur des principes plus élémentaires se réduit à une théorie particulière lorsqu'un paramètre sans dimension tend vers zéro. La question cruciale est la nature de la limite quand le paramètre tend vers zéro, et bien souvent, cette limite est fortement **singulière**, révélant par là un nouvel état de la matière.

Nous terminerons ce tour d'horizon sur le calcul à précision finie en présentant la théorie de Calcul Qualitatif, développée au CERFACS depuis plusieurs années. Lorsque les erreurs d'arrondi dominent (calculs chaotiques) cette théorie vise à utiliser les erreurs comme un révélateur des propriétés mathématiques du problème hors de portée du calcul à précision finie. Dans ce cas où les calculs sont déclarés "impossibles" au sens classique (Ekeland (1984), Chatelin (1989b)) car tous les chiffres sont faux, la **quantification** (calculer juste, par exemple) perd évidemment son sens.
Néanmoins la réalité du calcul sur ordinateur indique qu'il existe un large spectre de situations intermédiaires significatives entre le juste et le faux. Dans le Calcul Qualitatif, il n'est plus question de contrôler l'erreur comme cela se fait classiquement puisque elle ne tend pas vers zéro. Il s'agit de lui donner un **sens** : lorsque le calcul ne peut plus donner une information complète sur la solution exacte, il *continue néanmoins de donner une certaine information* voilée par le brouillard des erreurs, mais pertinente par rapport aux propriétés de l'équation à résoudre. Le Calcul Qualitatif consiste à tirer de calculs faux des informations qualitatives sur la nature de la singularité sous-jacente. Ainsi dans les calculs météorologiques, tous les calculs sont entâchés de grandes erreurs mais les tendances sont justes : la prévision du temps sur 3 à 5 jours est fiable.
Selon le regard porté sur elles, les erreurs de calcul peuvent être considérées négativement parce qu'elles impliquent un résultat faux, ou bien positivement parce qu'elles informent. Chemin faisant, nous argumenterons que l'ordinateur à précision finie, parce qu'il traite toujours d'informations finies, peut être un laboratoire représentant plus fidèlement la réalité physique que ne le ferait un ordinateur calculant exactement. En effet, tout physicien sait qu'aucune équation de la physique ne peut être exacte.

Cela nous permettra de répondre par la négative à la question posée en ouverture : **"Non, il ne faut pas avoir peur des ordinateurs"**.

3 L'arithmétique des ordinateurs

Historiquement, les scientifiques et les ingénieurs ont presque toujours réalisés leurs calculs vraiment complexes de manière approchée, en ne conservant qu'un nombre fini de chiffres significatifs grâce à un procédé appelé l'arrondissage[1] (round-off). Traditionnellement le calcul à la main utilise l'arrondi au plus près (round to the nearest) ou l'arrondi dirigé (par excès ou par défaut, directed rounding). Que fait l'ordinateur ? Il utilise une arithmétique dite à virgule flottante.

[1]Lorsqu'il n'y a pas d'ambiguité, le terme générique "arrondissage" est abrégé en "arrondi".

3.1 L'arithmétique à virgule flottante normalisée

Dans une arithmétique de base β, un nombre réel x strictement positif s'écrit $x = F\beta^E$, où F est la fraction de x, $\frac{1}{\beta} \leq F < 1$ et E est l'exposant $-\infty < E < +\infty$. En général, il faut une infinité de chiffres dans F pour représenter un réel x arbitraire.

Dans l'arithmétique de l'ordinateur, on ne conserve que les p premiers chiffres dans la fraction F, ce qui définit la mantisse m, et le domaine de variations de l'exposant E est limité. Le nombre dit "flottant" $X = m\beta^E = \mathrm{fl}(x)$ représente le réel x en machine. Une telle arithmétique s'appelle **arithmétique à virgule flottante** car la place de la virgule dépend de l'exposant E, et **normalisée** car on décide que le premier chiffre significatif non nul est après la virgule.
β et p définissent la "précision machine"[2] ou "macheps" par $\Psi = 1^+ - 1 = \beta^{-p+1}$, où 1^+ est le successeur de 1 en machine. Plus de détails sur les arithmétiques d'ordinateur sont donnés en Annexe A1.
Si l'on pouvait faire croître indéfiniment le nombre p de chiffres de la mantisse, la précision machine Ψ tendrait vers 0. On sera amené à s'intéresser aux propriétés asymptotiques des calculs lorsque le paramètre ε, représentant Ψ, tend vers 0.

3.2 Arithmétique correcte

Soit $\bullet$ l'une des 4 opérations élémentaires $\{+, -, \times, /\}$ et soit $\odot$ sa réalisation sur ordinateur. Une arithmétique est correcte si, pour tout flottant X et Y, elle satisfait

$$\mathrm{fl}(X \bullet Y) = X \odot Y.$$

On montre qu'il suffit de disposer d'**un** chiffre supplémentaire (dit **de garde**) pour stocker les résultats intermédiaires dans la division et la soustraction pour que l'arithmétique soit correcte. Une arithmétique correcte ne rajoute pas, pour les nombres représentables, d'erreur arithmétique à celle imputable à la représentation en machine du résultat.

Toutes les arithmétiques ne sont pas correctes. Par exemple, celle de CRAY sans bit[3] de garde, est incorrecte. La norme IEEE P.754 (1981) impose que l'arithmétique soit correcte pour les 4 opérations arithmétiques élémentaires, ainsi que pour la racine carrée et la congruence.

3.3 Modèle probabiliste de l'erreur d'arrondissage

Le calcul avec une mantisse à p chiffres entraîne une erreur d'arrondissage $\alpha = x - \mathrm{fl}(x)$ déterministe. Il peut être utile de modéliser cette erreur par une variable aléatoire u uniformément distribuée autour du $p$ième chiffre. C'est une idée très ancienne qui a été utilisée chez les astronomes depuis le XIX$^{\text{e}}$ siècle pour modéliser la propagation des erreurs d'arrondi à la main dans les calculs astronomiques.
Cette idée, assez naturelle, a été justifiée théoriquement (Feldstein et Goodman (1976)) en établissant la distribution du dernier chiffre après une suite "assez longue" de calcul. Cette étude a mis en évidence le fait beaucoup plus surprenant que le caractère non linéaire des calculs entraîne que la distribution du premier chiffre n'est pas uniforme mais **logarithmique** (Turner (1982)). Choisissons au hasard un nombre parmi un ensemble assez grand de résultats de calculs (fournis par un ordinateur par exemple) ou de mesures expérimentales. Il est remarquable que, pour un nombre supposé ecrit en base 10, il y ait 30% de chances pour que le premier chiffre soit 1, tandis

[2] Attention au piège du vocabulaire consacré par l'habitude. Un calcul effectué avec une *grande* précision requiert une "précision machine" ou "macheps" *petite*.

[3] bit = <u>b</u>inary dig<u>it</u> = terme américain pour désigner un chiffre significatif en base 2.

qu'il n'y en a que 4.5% pour que ce soit 9 (voir l'Annexe A1).
Ce fait, connu depuis plus d'un siècle par les astronomes (Newcomb (1881)), a été périodiquement redécouvert (Benford (1938), Hamming (1970), Knuth (1969), Stewart (1993)). Plus étonnant encore, de ce résultat peu intuitif découle ce que nous pouvons appeler **le paradoxe de Newcomb–Borel** sur le hasard des nombres (Chaitin (1991), Chatelin (1991)). En effet, en 1909, Borel démontre que presque tous les nombres réels sont normaux (Rauzy (1976, p. 75)), donc que dans n'importe quelle base, le premier chiffre en particulier est **uniformément** distribué.

Il semble donc qu'il existe une grande différence **de nature** *entre un nombre tiré au hasard parmi les* **réels mathématiques***, ou parmi ceux produits par une* **suite de calculs ou de mesures***.*

3.4 Vers une bonne arithmétique d'ordinateur

Dans les années 1970, de nombreux travaux ont permis d'établir que le choix de **la base** 2 **avec premier bit implicite** jouit de très nombreuses qualités; en particulier il permet de :

- minimiser l'erreur d'arrondissage α (Brent (1973), Cody (1973, 1984), Kuki et Cody (1973)),
- minimiser l'erreur relative moyenne dans une multiplication (Goodman et Feldstein (1975, 1977), Feldstein et Goodman (1982), Goodman, Feldstein, et Bustoz (1985)).

Ce choix a donc été recommandé par la société IEEE en 1981. Un des événements marquants pour le Calcul Scientifique dans les années 80 fut l'avènement des stations de travail disposant d'unités arithmétiques de beaucoup plus grande qualité que celles des gros ordinateurs. De plus, en se conformant toutes à la norme IEEE – ce qui est devenu un argument de vente –, elles rendent le portage des programmes beaucoup plus aisé.
Mais les performances sans cesse croissantes de ces unités arithmétiques ne risquent-elles pas de rendre obsolètes les recommandations IEEE portant sur les caractéristiques de l'arithmétique ? Il semble que oui. Une étude de Feldstein et Turner (1986) montre que la fréquence des accidents arithmétiques (overflows, underflows, élimination catastrophique) croît avec la vitesse des calculs (voir l'Annexe A1).
On peut tirer la conclusion technologique suivante :
tout accroissement significatif de la vitesse des calculs devrait être accompagné par un accroissement correspondant du domaine de variation de l'exposant et de la longueur de la mantisse.
Des études sont en cours dans la communauté pour systématiser l'utilisation de $p \geq 106$ bits (longueur totale du "mot" égale à 128 bits, soit le double des 64 bits couramment utilisés aujourd'hui).

4 Modèles probabilistes de propagation des erreurs de calcul

A partir du modèle u de l'erreur élémentaire d'arrondissage α et de l'hypothèse que *les erreurs commises à chaque pas sont indépendantes*, on peut construire un modèle probabiliste qui prédit la variable aléatoire U qui représente la solution finale x. Les principales références sont Brouwer (1937), Rademacher (1948) et surtout Henrici (1962b, 1963). Des objections ont été élevées car en pratique les erreurs sont corrélées (Huskey (1949),Marasa et Matula (1973)). Henrici s'en explique dans Henrici (1962a) à la page 55. Le modèle de Henrici pour des équations différentielles ordinaires est difficile à valider puisque en pratique les calculs sont **déterministes**. Henrici (1966) utilise une perturbation aléatoire des conditions initiales, ce qui permet d'obtenir un échantillon de solutions, mais dont rien ne permet d'affirmer qu'il est constitué de réalisations de U.

4.1 Le rôle du biais et le contrôle de l'erreur

Supposons que la loi de U admette la moyenne m. On appelle biais l'écart $m - x$ entre la moyenne et la solution exacte, c'est-à-dire l'erreur moyenne. La taille $\|m - x\|$ du biais mesure le degré de non-compensation des erreurs élémentaires. Le biais est nul pour des calculs **linéaires** et en **arrondi**. Il n'en est plus ainsi en général pour des calculs non linéaires et a fortiori en troncature. De nombreux exemples sont fournis dans Chatelin (1989b), voir aussi Becker et Becker (1977).

Lorsque $m \neq x$, la connaissance de la loi de U ne suffit pas pour le contrôle de l'erreur $U - x$, ce qui limite la portée pratique des modèles probabilistes de propagation d'erreurs pour un tel contrôle lorsque la solution exacte x n'est pas connue, comme c'est toujours le cas quand on cherche à la calculer. L'utilité de tels modèles nous semble plus devoir être trouvée en tant qu'outils d'analyse de la stabilité sous l'effet de perturbations dues à l'arithmétique. De même que toutes les analyses **directes**, les modèles probabilistes ne peuvent prétendre distinguer entre l'erreur due au problème et celle due à l'algorithme. Une analyse inverse est nécessaire pour cela. Conscientes de ces limitations, Bennani, Brunet, et Chatelin (1988) ont proposé un modèle probabiliste de propagation des erreurs dans le cadre des calculs matriciels appelé PRECIX pour résoudre $Ax = b$ et $Ax = \lambda x$. Ce modèle permet de faire une analyse directe et inverse, et permet, entre autres, de contourner la difficulté due au biais non nul.
De plus, il permet de clore une controverse scientifique qui a débuté dans les années 60 parmi les utilisateurs du calcul scientifique, quant à la portée générale du modèle de Henrici, qui établit que, pour des équations différentielles ordinaires, la loi de U est *normale* (Henrici (1962b), Hull et Swenson (1966)).
Comme on pouvait s'y attendre,
la loi normale n'est plus valable pour les problèmes non linéaires, au voisinage des points singuliers (solution non C^1).
Ce résultat est de portée générale. Il y a 35 ans déjà, van Wijngaarden (1953) proposait des exemples simples illustrant ce fait. On trouvera la démonstration (et vérification sur ordinateur) du cas particulier d'une valeur propre multiple défective dans Chatelin (1988b, 1989b), Chatelin et Brunet (1990)).

4.2 Randonisation[4] des calculs par arrondissage aléatoire

Afin de contourner, lors des simulations sur ordinateur, la difficulté posée par le caractère déterministe des calculs, Forsythe (1950) a proposé de "randoniser" les calculs par injection d'une perturbation aléatoire après chaque opération arithmétique. Cette idée peut être réalisée de la manière suivante : on simule l'effet des arrondis de calcul au $p$ième chiffre à l'aide de calculs exécutés avec $q \geq p$ chiffres et un **arrondissage aléatoire** $\tilde{\alpha}(p, q, \pi)$ au $p$ième chiffre : $\tilde{\alpha}(p, q, \pi)$ résulte de l'arrondi déterministe dû au calcul avec q chiffres, auquel s'ajoute une perturbation aléatoire π autour du $p$ième chiffre, de distribution connue. On trouve en Annexe A1 quelques exemples d'arrondissage aléatoire rencontrés dans la littérature.

Les outils de contrôle de l'erreur fondés sur des modèles probabilistes ont été progressivement abandonnés par la communauté anglo-saxonne des concepteurs de logiciels numériques. Une des principales raisons de la forte résistance de cette communauté aux méthodes probabilistes est

[4] Terme de vieux français ressuscité par Benoit Mandelbrot pour remplacer le mot franglais rando*m*isation.

subtile mais très illustrative de la richesse du domaine du Logiciel Numérique. En effet, certains algorithmes spécifiques peuvent fournir une très grande précision, quelle que soit l'instabilité du problème à résoudre, car leur structure interne est si bien ajustée qu'elle n'excite pas l'instabilité mathématique potentielle. Evidemment, l'emploi d'une arithmétique aléatoire a de fortes chances de détruire ce délicat ajustement, et en tout cas ne permet pas de l'analyser, encore moins de le prédire. C'est la raison pour laquelle la communauté se méfie des arithmétiques aléatoires, ou stochastiques.

* *

Nous continuons donc notre tour d'horizon par la présentation de l'analyse inverse des erreurs développée par Wilkinson qui a supplanté l'approche directe probabiliste. La présentation qui suit est originale (Chatelin (1988b, 1989b), Chaitin-Chatelin et Frayssé (1995), mais nous pensons que son intérêt va bien au-delà de la simple originalité. Elle propose une **théorie de la calculabilité sur ordinateur :** l'analyse théorique des erreurs, *quelle qu'en soit l'origine* (arithmétique, numérique ou physique) est centrée sur l'analyse de la stabilité mathématique du problème à résoudre.
Cette approche a le grand avantage de mettre naturellement en évidence le lien profond qui unit les stabilités numérique, algorithmique et arithmétique à la stabilité mathématique du problème posé.
Replacées dans un contexte et un vocabulaire familiers aux mathématiciens, des notions telles que l'analyse inverse des erreurs, définies au départ dans le jargon vernaculaire des concepteurs de logiciels, prennent toute leur force et leur signification. Et l'on comprend mieux pourquoi le calcul à précision finie, loin d'obéir à quelque loi magique dictée par l'arithmétique de l'ordinateur et que seuls les initiés savent déchiffrer, peut (et doit), comme tout autre processus scientifique, être interprété par l'analyse mathématique.

5 La calculabilité[5] sur ordinateur

Dans la suite, le paramètre ε représentera la précision machine Ψ et sera destiné à tendre vers zéro (c'est-à-dire que l'on fera tendre par la pensée la taille de la mantisse vers l'infini). C'est cette simple astuce qui va permettre de replacer l'étude des algorithmes implémentés sur ordinateur dans le cadre de l'analyse mathématique traditionnelle. Cette idée est aussi éloignée de l'esprit des concepteurs de logiciels informatiques que ne l'était, dans les années cinquante, l'idée de faire tendre vers zéro la taille des éléments finis dans l'esprit des ingénieurs civils qui concevaient des structures.

5.1 Une hiérarchie d'approximations

Pour fixer les idées, nous considérons que nous devons résoudre l'équation non linéaire

$$(P^*) \quad F^*(x^*) = y^* \tag{1}$$

dont on suppose qu'elle admet une solution $x^* = F^{*-1}(y^*)$ localement unique. F^* est une application non linéaire de X dans Y, qui sont (par exemple) des espaces normés. F^* est supposée **bicontinue** (F^* et F^{*-1} sont continues). La continuité de F^{*-1} est aussi connue comme une propriété

[5] Le terme de "calculabilité" est employé ici dans une acception différente de celle qui est utilisée en Informatique Théorique.

de **stabilité** de (P^*). La résolution approchée de (P^*) sur ordinateur se situe en général à deux niveaux au moins : d'abord numérique (par discrétisation et/ou arrêt des itérations), puis arithmétique (par passage à la précision finie de l'ordinateur). La famille d'approximations numériques de (P) est désignée par

$$(P_h) \quad F_h(x_h) = y_h \tag{2}$$

où h est un paramètre destiné à tendre vers zéro, lié au pas de discrétisation et/ou au nombre fini d'itérations. La résolution algorithmique de (P_h), **pour h fixé**, conduit à résoudre exactement sur $\mathbb{R}$ un problème de dimension finie à l'aide d'un algorithme fini, soit

$$(P) \quad F(x) = y \quad \Longleftrightarrow x = G(y) \tag{3}$$

où G représente l'algorithme fini qui, s'il était effectué en arithmétique exacte, donnerait la solution exacte x. Or, le calcul est effectué en pratique sur un ordinateur utilisant une arithmétique à précision finie ε, dans l'ensemble $\mathbb{F}_\epsilon$ des nombres flottants de cette arithmétique. Il fournit la solution calculée X_ε du problème discret dans $\mathbb{F}_\epsilon$

$$(\mathcal{P}_\varepsilon) \quad \mathcal{F}_\varepsilon(X_\varepsilon) = Y_\varepsilon \quad \Longleftrightarrow X_\varepsilon = \mathcal{G}_\varepsilon(Y_\varepsilon). \tag{4}$$

5.2 Plongement de l'ensemble des flottants dans l'ensemble des réels

Afin de comparer plus aisément $(\mathcal{P}_\varepsilon)$ et (P), *on va plonger* $(\mathcal{P}_\varepsilon)$ *dans* $\mathbb{R}$ pour obtenir la solution calculée x_ε puis considérer que x_ε est la **solution exacte** dans $\mathbb{R}$ d'un problème perturbé de même type que (P), soit

$$(P_\varepsilon) \quad F_\varepsilon(x_\varepsilon) = y_\varepsilon \quad \Longleftrightarrow x_\varepsilon = G_\varepsilon(y_\varepsilon). \tag{5}$$

Cette démarche permet d'analyser l'erreur $x - x_\varepsilon$ due à l'arithmétique finie par comparaison dans $\mathbb{R}$ des applications F et F_ε : c'est le **principe de l'analyse inverse** (ou rétrograde, “backward”) des erreurs de Givens et Wilkinson. Elle permet de contourner la difficulté posée par l'absence de structure riche pour l'arithmétique finie (absence d'associativité, commutativité, et distributivité) et de se replacer dans le contexte arithmétique puissant de $\mathbb{R}$ où l'on peut appliquer les théorèmes de l'analyse classique.

Remarques

1. Dans certains cas, l'analyse de $x - x_\varepsilon$ se fait plus aisément ou plus finement par comparaison de G et G_ε. Cette démarche s'appelle traditionnellement **l'analyse directe** (ou progressive, “forward”) des erreurs (Bauer (1974), Voïévodine (1980), Stummel (1980) Higham (1987, 1989)).

2. Le plongement de $(\mathcal{P}_\varepsilon)$ dans $\mathbb{R}$ utilisé en Logiciel Numérique est similaire, dans son principe, à celui réalisé en Analyse Numérique Théorique pour étudier la convergence de méthodes de discrétisation pour les équations aux dérivées partielles, par exemple (Aubin (1972), Chatelin (1983)). Mais les difficultés respectives sont de nature différente. En Logiciel Numérique, le prolongement de $\mathbb{F}_\epsilon$ dans $\mathbb{R}$ est trivial, et toute la difficulté consiste à comparer F et F_ε lorsque ε tend vers zéro, alors que F_ε *n'est pas connue explicitement.* C'est un problème technique *ardu* qui n'a été résolu systématiquement que pour l'Algèbre Numérique (Anderson, Bai, Bischof et al. (1992), Chaitin-Chatelin et Frayssé (1995)).

5.3 Principe de Lax et calculabilité

On dit que x est **calculable arithmétiquement** si x_ε tend vers x quand ε tend vers zéro. Cette convergence arithmétique peut s'établir par appel au **principe de Lax** (Aubin (1972)) :

Consistance $F_\varepsilon(z) \to F(z)$ au voisinage de x, et $y_\varepsilon \to y$

$+$

Stabilité uniforme, c'est à dire équicontinuité en ε de $F_\varepsilon^{-1} = G_\varepsilon$

$\Longrightarrow$ Convergence : $x_\varepsilon \to x$ quand $\varepsilon \to 0$

Ce principe est connu de tous les numériciens théoriciens sous la forme célèbre : "un schéma stable et consistent est convergent", lorsque le paramètre est h qui représente le pas de discrétisation en temps et/ou en espace, et/ou le nombre d'itérations.
Dans le cas qui nous occupe, le calcul sur ordinateur de x_h fournit en fait $x_{h\varepsilon}$. On dit que x^* est **calculable sur ordinateur** si $x_{h\varepsilon}$ tend vers x^* quand h et ε tendent vers zéro. On voit donc apparaître comme condition suffisante de convergence la stabilité uniforme de $(P_{h\varepsilon})$ ou encore l'équicontinuité uniforme en h et ε de $F_{h\varepsilon}^{-1}$.

Techniquement, le passage sur ordinateur revient à rajouter le paramètre ε de la précision machine à la liste des paramètres par rapport auxquels il suffit d'établir la stabilité uniforme.

5.4 Fiabilité arithmétique des algorithmes

On appelle **algorithme** une procédure de calcul, qui, si elle est exécutée en arithmétique exacte, fournit la solution exacte $x = G(y)$ de (P) en un nombre *fini* de pas.

Le principe de Lax permet d'établir la calculabilité $x_\varepsilon \to x$ sous la double condition suffisante de consistance ($F_\varepsilon \to F$ et $y_\varepsilon \to y$) et de stabilité uniforme (équicontinuité en ε de F_ε^{-1}).

Regardons de plus près la condition de consistance. Pour cela, on introduit l'erreur de consistance sur G, soit $EC(z) = \|G_\varepsilon(z) - G(z)\|$ pour z dans un voisinage de y. A cause de l'erreur inhérente à la représentation des données numériques en machine, le mieux que l'on puisse établir est une borne du type

$$EC(y_\varepsilon) \leq C(y, G)\varepsilon \tag{6}$$

qui exprime que l'erreur de consistance pour G en y_ε est au mieux d'ordre 1 en ε.
Si cette condition est satisfaite lorsque G est de classe C^1, on dit que l'algorithme est **arithmétiquement fiable**.

Remarque : Cette propriété de *fiabilité arithmétique* est connue sous le nom de "**stabilité inverse**" ("backward stability") depuis Wilkinson (1961). On trouve aussi le terme "stabilité numérique vis à vis de l'arrondissage" ("numerical stability with respect to round off") dans les textes d'Analyse Numérique et même "stabilité" tout court (Dahlquist et Björck, Bulirsch et Stoer par exemple). Ce n'est pas une propriété de stabilité au sens mathématique ou physique du terme (c'est à dire équicontinuité en ε de F_ε^{-1}).
Ce choix linguistique est malheureux car il rend possible la confusion entre cette notion, liée à l'arithmétique des ordinateurs, et la notion mathématique de stabilité numérique en arithmétique exacte (voir le paragraphe suivant). C'est la raison pour laquelle nous avons proposé le terme plus approprié de **fiabilité arithmétique** (Chatelin (1989b), Chatelin et Frayssé (1991, 1995)) pour un algorithme.

5.5 Limites de la stabilité numérique

Nous considérons plus généralement une méthode numérique qui ne peut donner, en arithmétique exacte, la solution exacte x qu'asymptotiquement, à la limite des paramètres de convergence. Cette convergence est classiquement réglée par le principe de Lax.

La stabilité ne peut pas toujours être établie dans tout le domaine de variation des paramètres. Il faut parfois se contenter d'une **stabilité conditionnelle**. Cela restreint le domaine de variation des autres paramètres en fonction de la précision machine utilisée. Les ingénieurs disposent jusqu'à présent du seul outil de la stabilité uniforme en h des schémas numériques, établi par l'Analyse Numérique Théorique, c'est-à-dire ce qu'on appelle classiquement la *stabilité numérique* en Analyse Fonctionnelle (équicontinuité uniforme en h de F_h^{-1}).

Cet outil est suffisant pour entraîner la convergence si les calculs sont exacts ($\varepsilon = 0$), mais il ne l'est pas toujours si les calculs sont faits à la précision $\varepsilon > 0$. Cela dépend d'abord de la stabilité du problème à résoudre. Il est clair que des difficultés peuvent surgir lorsque la méthode d'approximation n'admet qu'une stabilité *conditionnelle à l'arithmétique* qui impose un seuil de stabilité du type $h \geq h_0(\varepsilon)$, ce qui fait jouer à la précision finie de l'arithmétique un rôle crucial. Ce fait, très bien analysé pour la dérivation numérique (voir l'Annexe 2), ne semble pas toujours être interprété correctement lorsqu'il advient dans d'autres contextes (Trefethen et Trummer (1987)). Nous y reviendrons au paragraphe 9.

C'est parce que le calcul scientifique joue un rôle primordial dans le développement des technologies de pointe que le problème de la propagation des erreurs se pose avec force. La prise en compte de phénomènes de plus en plus fins entraîne une modification profonde des techniques de modélisation. Cette modification se caractérise, entre autres, par la prise en compte d'un nombre beaucoup plus grand de degrés de liberté au niveau le plus fin de la description et par un accroissement de sa complexité (donc du nombre de niveaux considérés) afin d'augmenter la fidélité à la réalité physique ou technologique.

Le rôle de l'arithmétique, en particulier la nécessité de disposer d'une précision suffisante, va devenir essentiel dans les simulations de modèles représentant un **couplage fort** de deux ou plusieurs phénomènes physiques (Lions (1992)). Dans la zone d'interaction de deux modèles couplés apparaissent le plus souvent de nouveaux phénomènes qui requièrent une analyse morphologique (qualitative) avant toute tentative de calcul plus poussé (Aubin (1994)).
De plus, le couplage de modèles à des niveaux de description différents pose le problème de la cohérence de l'ensemble de la description. Il faut de nouveaux outils pour gérer cette cohérence. L'évolution de la modélisation qui permet la prise en compte d'un dispositif et de son environnement (voir aussi Chatelin et Trouvé (1993) où des exemples industriels Thomson–CSF sont décrits) est indissociable de l'évolution des ordinateurs mis à la disposition des ingénieurs. Le prix très bas de la mémoire et l'accroissement des performances technologiques et algorithmiques permettent d'aborder des problèmes de très grande taille. Les diverses architectures parallèles proposées permettent aussi d'espérer reproduire certaines spécificités du parallélisme naturel.

Remarque : La convergence arithmétique par exemple peut s'établir aussi bien sur l'erreur $x_\varepsilon - x$ que sur l'erreur résiduelle (ou résidu) $F(x_\varepsilon) - y$ puisque les conditions de continuité uniforme en ε de G_ε (stabilité) et de $F \circ G_\varepsilon$ (stabilité résiduelle) sont équivalentes mathématiquement. Mais en pratique, pour une étude à ε fixé, ces conditions ne sont pas équivalentes car, à cause de la composition par F, $F \circ G_\varepsilon$ ne contient plus le risque d'instabilité dû à F^{-1} qui figure dans $G_\varepsilon = F^{-1} \circ (F \circ G_\varepsilon)$. On retrouve là une différence subtile pour les mathématiques mais fondamentale sur ordinateur entre l'Analyse Numérique Théorique et le Logiciel Numérique. Une autre différence importante apparaît entre le paramètre ε de l'arithmétique, et les autres paramètres de discrétisation (temps ou espace) : on a vu que sur ordinateur, l'erreur de consistance arithmétique est au mieux d'ordre 1 en ε. Alors que si h est le pas de discrétisation numérique, on cherche à obtenir une erreur de consistance numérique $F(z) - F_h(z)$ qui soit d'un ordre en h aussi élevé que possible.

6 Mesures de stabilité

Le concept de **stabilité** joue un rôle central dans l'analyse des méthodes numériques, rôle à la fois théorique (convergence ou calculabilité) et pratique (qualité de la solution, en particulier précision obtenue ou justesse au sens de "accuracy"[6]). Regardons plus en détail la stabilité en un point régulier et les mesures associées (Rice (1966), Geurts (1982), Chatelin et Frayssé (1992, 1993)). La notion de stabilité est relative au choix des données à perturber, et les mesures de stabilité dependent des métriques utilisées. Depuis Turing (1948), il est traditionnel en Analyse Matricielle d'appeler **conditionnement** (linéaire) une telle mesure qui, en un point régulier où $x = F^{-1}(y)$ est C^1, n'est pas autre chose qu'une mesure de la différentielle de Fréchet de F^{-1} en y.
En Analyse Matricielle, deux types de métriques se sont révélées particulièrement utiles pour l'analyse; elles reviennent à définir pour une matrice A des perturbations admissibles ΔA de niveau ω par

i) $\mathcal{E}^{\mathcal{N}} = \{\Delta A;\ ||\Delta A|| \leq \omega ||A||\}$ pour l'analyse en norme (dite *"normwise"*) de type Turing (1948), ou analyse du pire des cas,

ii) $\mathcal{E}^{\mathcal{C}} = \{\Delta A;\ |\Delta a_{ij}| \leq \omega |a_{ij}|\}$ pour l'analyse élémentaire (dite *"componentwise"*) de type Bauer (1966), ou analyse du meilleur des cas.

Nous reviendrons plus loin sur le rôle respectif de ces deux types de perturbations pour l'analyse des algorithmes. L'ensemble des conditionnements normwise et componentwise pour les problèmes $A \longmapsto x = A^{-1}b,\ A^{-1}$ ou λ, x tels que $Ax = \lambda x$, ainsi que pour les racines de polynômes, est donné dans Chaitin-Chatelin et Frayssé (1995).

Il importe de réaliser que les conditionnements des algorithmes/méthodes qui jouent un rôle théorique dans la démonstration de la convergence n'ont pas besoin d'être estimés très finement. Il suffit en effet de démontrer qu'ils sont *uniformément bornés* en les paramètres de convergence. Par contre, à cause de l'inégalité (7) donnée à la page 16 et qui est à la base de l'analyse d'erreur, on voit que l'estimation de l'erreur dépend d'une estimation fidèle du conditionnement du problème. C'est pourquoi la bibliothèque logicielle LAPACK fournit des routines de calcul de tels estimateurs.

7 Qualité arithmétique des algorithmes fiables

Peut-on mesurer la qualité de l'algorithme employé pour résoudre (P) afin d'en valider le choix ? Cela peut se faire grâce au concept essentiel **d'erreur inverse** associée à x_ε de Givens et Wilkinson, qui consiste, après avoir spécifiés les ensembles de perturbations admissibles, donc la métrique, à mesurer la distance minimale de (P) à un problème perturbé de même type résolu exactement par x_ε, soit $F_\varepsilon(x_\varepsilon) = y_\varepsilon$. C'est le **principe de Wilkinson** (1961).

[6] L'anglais dispose de deux notions non équivalentes, "precision" et "accuracy" qui se traduisent toutes deux par précision en français. Pour les distinguer, on se propose de traduire "accuracy" par "justesse".

7.1 Le calcul de l'erreur inverse sur ordinateur

L'erreur inverse en x_ε peut se calculer à l'aide du résidu $r = F(x_\varepsilon) - y$. Les formules explicites pour l'Algèbre Linéaire, avec les métriques “normwise” et “componentwise”, sont données dans Chaitin-Chatelin et Frayssé (1995).
Idéalement, les quantités $F(x_\varepsilon)$ et y sont proches. Calculer le résidu r revient donc à effectuer des soustractions de nombres voisins issus eux-mêmes de calculs, donc entachés d'erreur. Ces soustractions sont connues pour fournir une grande erreur relative (élimination catastrophique). Mais pour r voisin de 0, l'erreur absolue ne dépasse pas la précision machine Ψ. L'évaluation du résidu est donc suffisamment précise pour que sa valeur calculée puisse être valablement comparée à Ψ, ce qui suffit pour établir la fiabilité (voir ci-dessous, voir aussi l'Annexe A3).

Remarque : Les praticiens sont parfois méfiants quand il s'agit de fonder une analyse d'erreur sur le résidu calculé à la précision courante. Nous venons de voir que leurs craintes sont exagérées. On peut même utiliser le raffinement itératif sur le résidu ainsi calculé pour accroître la fiabilité d'algorithmes (voir Jankowski et Woźniakowski (1977) pour les solveurs linéaires, ainsi que Skeel (1980) pour la factorisation LU).
Par contre, il est vrai qu'une précision étendue est nécessaire pour le calcul du résidu si l'on veut utiliser le raffinement itératif pour augmenter la précision de la solution calculée. C'est le fondement des méthodes d'arithmétique d'intervalle (Kulisch et Miranker (1981)).

7.2 Indice de qualité arithmétique

Lorsqu'on fait tourner un algorithme sur ordinateur, le mieux que l'on puisse espérer est de ne pas introduire plus d'incertitudes que celles qui résultent des données (influence de l'erreur d'arrondissage de l'ordre de Ψ ou des erreurs expérimentales). On peut donc introduire **l'indice de qualité arithmétique** J associé à x_ε comme le rapport de l'erreur inverse en x_ε à la précision machine Ψ , lorsqu'il n'y a pas d'erreurs expérimentales.
Cet indice est supérieur ou égal à 1, la meilleure qualité arithmétique correspondant à $J = 1$. Et un algorithme fiable est dit **optimal** (par rapport à l'arithmétique) si son indice de qualité reste proche de 1, quel que soit le problème à résoudre. On trouvera en Annexe A3 une présentation succinte de l'analyse inverse des erreurs appliquée à l'exemple de base, l'algorithme de Gauss pour résoudre $Ax = b$.

7.3 Erreur de calcul

Soit $\tilde{x}$ la solution calculée (représentant x_ε ou plus généralement $x_{h\varepsilon}$). La majoration de l'erreur de calcul $x - \tilde{x}$ par l'analyse inverse des erreurs repose sur l'estimation des deux quantités fondamentales que sont

- K le conditionnement du problème, et
- $EI(\tilde{x})$ l'erreur inverse en $\tilde{x}$.

L'erreur de calcul $x - \tilde{x}$ entre solution exacte et calculée peut alors être estimée, pour un problème régulier, à l'aide de l'inégalité (au 1er ordre)

$$\begin{aligned} \|x - \tilde{x}\| &\leq \text{Conditionnement} \times \text{erreur inverse} \\ &\leq K \times EI(\tilde{x}) \end{aligned} \tag{7}$$

Il existe de nombreuses variantes de cette inégalité selon le type de métriques choisies sur l'espace des données et celui des solutions. En particulier, en Analyse Numérique, les mesures sont le plus souvent **relatives**. L'inégalité (7) met bien en évidence la distinction dans l'erreur de calcul entre la contribution du problème (conditionnement K) et celle de l'algorithme (erreur inverse EI). L'erreur inverse est **amplifiée** par l'instabilité mathématique du problème par un facteur ne pouvant dépasser le conditionnement. On voit qu'il n'existe pas **d'effet "boule de neige" automatique**. Cet effet n'est déclenché qu'en présence d'une trop grande instabilité du problème ($K \geq 10^{15}$ en double précision).

7.4 Pertinence du modèle de perturbation

L'analyse inverse des erreurs dépend de l'ensemble $\mathcal{E}$ des perturbations admissibles choisi, puisque les quantités K et $EI(\tilde{x})$ dans (7) dépendent de $\mathcal{E}$. De la pertinence de $\mathcal{E}$ à représenter les perturbations engendrées par le passage de l'algorithme sur ordinateur dépend la qualité de (7), comme borne supérieure de l'erreur $||x - \tilde{x}||$. Mais K mesure une instabilité **potentielle** liée à $\mathcal{E}$, qui peut ne pas être excitée par les perturbations effectivement engendrées par l'algorithme implémenté, entraînant une erreur effective **beaucoup plus petite** que celle que peut prédire la théorie. C'est une question difficile –et qui reste principalement du domaine de l'art – que de caractériser l'ensemble $\mathcal{E}$ pour un algorithme donné (voir Chaitin-Chatelin et Frayssé (1995) où sont traités de nombreux exemples, voir aussi Higham (1987, 1989) et l'Annexe A3).

C'est parce que l'algorithme de Gauss pour $Ax = b$ semble engendrer des perturbations "componentwise" que les années 80 ont vu l'essor de l'analyse componentwise (Bauer (1966), Skeel (1979)), maintenant disponible dans LAPACK. Par contre, la factorisation QR pour $Ax = b$ semble, elle, être mieux analysée par des perturbations "normwise" (Higham (1991)). Ce simple fait d'expérience illustre bien la *nécessité d'un large ensemble d'outils d'analyse. Il n'existe pas d'outil universel ayant réponse à tout.*

8 Le calcul au voisinage des singularités

La question du calcul à précision finie ne devient cruciale que si ce calcul se fait **au voisinage d'un point singulier**, lorsque la solution exacte x n'est plus C^1 et que le conditionnement linéaire classique n'existe pas ou devient infini. C'est le cas pour les problèmes bien posés lorsqu'on cherche à calculer une racine multiple. La solution n'est plus C^1 mais reste Hölder-continue d'ordre $\gamma < 1$.

De tels cas ne sont pas rares en Analyse Numérique (penser au calcul des valeurs propres en particulier). De plus, puisque les singularités ne sont pas génériques, on résout toujours sur ordinateur un problème régulier qui est très mal conditionné, et non analysable par les moyens classiques de la théorie linéaire. En fait, les propriétés du calcul au voisinage d'une singularité sont réglées par les caractéristiques de la singularité sous-jacente (Chatelin, Frayssé, et Braconnier (1995)).
Cette idée n'est pas nouvelle : elle a été présentée par Kahan (1972). Mais Kahan propose des conditionnements pour les problèmes singuliers qui imposent des contraintes sur les perturbations admissibles qui ne sont pas vérifiées par le calcul sur ordinateur. Ces conditionnements présentent donc un intérêt principalement théorique.

8.1 Conditionnement hölderien

Au voisinage d'un point singulier, le conditionnement linéaire d'un problème bien posé devient très grand : le problème est mal conditionné au sens classique[7]. Pour mesurer la difficulté de calcul d'une racine multiple, la notion de conditionnement hölderien joue le rôle tenu précédemment par le conditionnement linéaire pour une racine simple (voir Wilkinson (1963) pour les racines multiples de polynômes et Chatelin (1986) pour les valeurs propres multiples défectives).
Pour les valeurs propres multiples défectives, on ne dispose pas de formule explicite pour le conditionnement hölderien. Des estimations à l'aide de PRECISE sont données dans Chaitin-Chatelin et Frayssé (1995), qui conduisent à une très bonne estimation de l'erreur de calcul, grâce à une inégalité analogue à (7), adaptée au cas de la singularité traitée.

8.2 Distances à la singularité et sûreté de calcul

La **distance à la singularité** est un paramètre important de **l'analyse du risque** que le calcul soit non signifiant. Cette distance est, bien entendu, relative au choix de l'ensemble $\mathcal{E}$ des perturbations admissibles sur les données. Donc il en existe plusieurs. Par exemple, pour $Ax = b$, on obtient une distance "normwise" ou "componentwise" selon que l'on considère les ensembles $\mathcal{E}^{\mathcal{N}}$ ou $\mathcal{E}^{\mathcal{C}}$ définis au paragraphe 6.
Il est naturel de penser qu'il existe un lien entre le conditionnement K du problème et la distance à la singularité δ. Toujours pour $Ax = b$, un théorème attribué à Turing établit que $\delta = 1/K$, en analyse "normwise". Ce n'est plus vrai en componentwise (Rohn (1990)).
Comment l'égalité de Turing peut-elle s'étendre au cas de problèmes non linéaires ? Dans le cas du calcul des racines de systèmes d'équations algébriques, Shub et Smale (1992) ont proposé une généralisation sans modification. Mais comme dans Kahan (1972), cela amène à imposer des contraintes sur les perturbations admissibles qui ne sont pas satisfaites par le calcul sur ordinateur (Chatelin, Frayssé, et Braconnier (1995)).
Chatelin et Frayssé (1993) ont conjecturé une extension où δ est proportionnelle à $K^{\frac{1}{\gamma-1}}$, et où $\gamma < 1$ est l'ordre (de Hölder-continuité) de F en la singularité au voisinage. Cette conjecture a été vérifiée sur ordinateur à l'aide de PRECISE (voir aussi Chatelin, Frayssé, et Braconnier (1995)).
En particulier, on estime la **frontière dangereuse**, c'est-à-dire la valeur maximale de la précision machine ϵ autorisée pour que le calcul soit régulier donc **sûr**. Des calculs effectués avec une mantisse plus courte seraient vus comme singuliers par l'ordinateur. La frontière dangereuse représente la **distance à la singularité vue par l'ordinateur**. C'est aussi la taille maximale des perturbations que peut subir le problème sans violer la *limite de sécurité.*

Puisque la distance à la singularité mesure la distance de sécurité d'un calcul, on peut se poser la question : *existe-t-il un paramètre numérique qui a pour effet de "rapprocher" la singularité lors d'un calcul à précision finie ?* La réponse est oui, il existe au moins un tel paramètre. C'est la **non-normalité** dont il va être question maintenant.

9 Le rôle de la non-normalité

9.1 Bref rappel théorique

Les matrices normales A satisfont $A^*A = AA^*$ où A^* représente la transposée-conjuguée de A. Elles représentent la classe la plus générale de matrices diagonalisables admettant une base **or-**

[7] C'est la raison pour laquelle certains auteurs appellent "mal posés" de tels problèmes (Demmel (1987, 1990), Shub et Smale (1992)). Mais comme les problèmes de racines multiples sont *bien posés* au sens de Hadamard, nous préférons le terme de "singuliers".

thonormale de vecteurs propres : $A = QDQ^*$ où Q est unitaire et D est diagonale.
Une matrice non-normale est telle que $A^*A \neq AA^*$. Elle admet une décomposition de Schur : $A = QSQ^*$, où $S = D + N$ est triangulaire. Cette forme de Schur est reliée à la forme de Jordan $A = XJX^{-1}$ via la factorisation $X = QR$. En effet, $A = Q(RJR^{-1})Q^* = QSQ^*$. Donc l'instabilité de la décomposition spectrale dépend des indices[8] des valeurs propres et du conditionnement $\text{cond}(X)$ (ou $\text{cond}(R)$) de la base de Jordan. Une bonne mesure de non normalité devrait refléter le montant de l'instabilité spectrale. C'est une question difficile qui n'a pas encore reçu de solution entièrement satisfaisante. On peut citer (voir Chaitin-Chatelin et Frayssé (1995)) :

i) $\nu(A) = \|AA^* - A^*A\|$, calculable directement à partir de A,

ii) $\Delta(A) = \|N\|$, qui nécessite de connaître S,

iii) le nombre de Henrici $\text{He}(A) = \nu(A)/\|A^2\|$ qui est homogène en A. $\text{He}(A)$ grand est une condition suffisante seulement de grande instabilité spectrale.

9.2 Le rôle de la non-normalité dans l'instabilité spectrale

Reconnue depuis longtemps (Henrici (1962a)), l'instablité spectrale des problèmes non-normaux a, jusqu'à présent, été considérée par les ingénieurs comme une curiosité mathématique sans équivalent dans l'univers physique. Cette attitude fut raisonnable jusqu'à maintenant parce que la taille et la difficulté des problèmes abordés (déterminées en grande partie par la puissance des ordinateurs) l'ont permis. Mais la physique et la technologie (Braconnier, Chatelin, et Dunyach (1995), Trefethen, Trefethen, Reddy, et Driscoll (1993)) commencent à requérir la résolution de problèmes où la nonnormalité croît exponentiellement avec la taille du problème à résoudre, ou avec une constante physique comme le nombre de Reynolds en mécanique des fluides (Reddy, Schmid, et Henningson (1993)) ou l'inverse de la résistivité en magnéto-hydro-dynamique (Kerner (1989)).

Or la non-normalité a pour effet de "rapprocher" les singularités, dans les calculs à précision finie. Le spectre exact, ensemble des valeurs propres[9] de A, c'est-à-dire l'ensemble des points singuliers de $(A - zI)^{-1}$, est remplacé en calcul à précision ϵ par le **pseudospectre**, c'est-à-dire l'ensemble des points qui satisfont $Ax = \lambda x$ approximativement. Plus le problème spectral est instable, plus le pseudospectre est étendu dans $\mathbb{C}$, et plus des points réguliers, très éloignés de λ en arithmétique exacte, sont vus par le calcul à précision finie comme singuliers (Trefethen (1990), Chatelin (1993b), Chatelin et Frayssé (1993)).
On trouve dans Chatelin (1993a, p. 161–165) l'exemple très illustratif d'une matrice, dite matrice de Schur, dépendant d'un paramètre ν ($\text{He} \sim \nu$) de taille 20, dont le spectre, de forme parabolique en arithmétique exacte, devient circulaire en arithmétique à précision finie (15 chiffres décimaux, calculs par QR), pour des valeurs de $\nu \geq 10^2$. Toute l'information spectrale contenue dans A est alors perdue par passage sur l'ordinateur. Et la matrice, qui est diagonalisable en arithmétique exacte, se comporte vis à vis de l'ordinateur comme un bloc de Jordan de taille 18. On voit que l'influence de la non-normalité, qui est d'abord locale (pour $\nu < 10^2$) devient, à partir d'un certain niveau, **globale** : *un problème bien posé peut être vu comme un problème mal posé par un ordinateur travaillant en double précision (15 chiffres décimaux).*

D'où provient la non-normalité des matrices qui commencent à émerger en Physique et en Technologie ? Ces matrices proviennent le plus souvent de la discrétisation d'équations aux dérivées partielles et la nonnormalité est héritée de celle de l'opérateur différentiel. La forte instabilité spectrale qui en découle reflète souvent le fait que le spectre d'une famille de tels opérateurs, dépendant

[8] L'indice d'une valeur propre est la taille du plus grand bloc de Jordan associé.
[9] λ est valeur propre de $A \iff \exists x \neq 0$ tel que $Ax = \lambda x$.

d'un paramètre physique (nombre de Reynolds ou de Péclet par exemple) ou numérique (taille de A) présente une forte discontinuité quand le paramètre varie. Cela peut se produire quand le modèle décrit un couplage fort entre deux phénomènes naturels, provoquant une zone d'instabilités physiques (Chaitin-Chatelin et Frayssé (1995)).

Une des conséquences les plus immédiates de la non-normalité concerne la convergence des méthodes d'approximations successives pour résoudre $x = Ax + b$. Mathématiquement, l'itération de point fixe converge dès que $\rho(A) < 1$, mais il n'en est pas toujours ainsi sur ordinateur. Cela dépend du rayon spectral vu par l'ordinateur. Ce fait d'expérience connu de tous les numériciens (Hammarling et Wilkinson (1976)), n'a été relié à l'influence de la non-normalité de A que très récemment (Arioli et Romani (1992), Trefethen (1990), Chatelin (1993b)).

Il importe de réaliser que l'instabilité spectrale du modèle mathématique (éventuellement discrétisé) ne fait que refléter *l'instabilité physique* du phénomène analysé par le physicien ou l'ingénieur, instabilité inhérente à la dynamique du phénomène.
Le calcul à précision finie ne fait que révéler cette instabilité, il ne la crée pas, même si parfois il peut l'amplifier sérieusement comme nous pouvons le constater avec des méthodes itératives. La non-normalité peut diminuer sensiblement la qualité de la convergence de telles méthodes.
Lorsque la non-normalité met en cause la robustesse des méthodes itératives, il faut des outils adaptés pour analyser la fiabilité des résultats qu'elles fournissent.

9.3 La non-normalité de A : un défi pour les méthodes itératives

Les expérimentations faites jusqu'à présent semblent indiquer que la non-normalité n'a pas d'influence négative sur la fiabilité des algorithmes finis (par exemple LU pour $Ax = b$) et certains algorithmes itératifs (QR pour $Ax = \lambda x$) : l'erreur inverse relative[10] reste de l'ordre de la précision machine. Voyons ce qu'il en est pour les méthodes itératives.
Pour les phénomènes d'évolution, il est, par exemple, bien connu que la **stabilité** en arithmétique exacte des systèmes linéaires d'équations différentielles (temps continu) ou d'équations de récurrence (temps discrétisé) de matrice A, dépend de la localisation dans $\mathbb{C}$ du spectre de A (c'est un exemple de stabilité conditionnelle). En précision finie, c'est la localisation du *pseudospectre* qui signale le risque d'une instabilité effective du calcul.
Si la matrice A est non-normale, il est possible que le pseudospectre déborde du domaine de stabilité déterminé dans $\mathbb{C}$ par le spectre exact. Ce débordement est nécessaire mais non suffisant pour entraîner la divergence. Mais il entraîne au moins une diminution de la qualité de la convergence (Trefethen (1990), Chatelin (1993a), Bennani et Braconnier (1994a, 1994b)) et peut mettre en défaut des algorithmes non suffisamment robustes.
A l'heure actuelle, tous les problèmes de très grande taille ($> 10^5$), issus de la discrétisation d'équations aux dérivées partielles sont traités par des méthodes itératives. On comprend donc que la non-normalité puisse présenter un **sérieux défi** aux méthodes de calcul à précision finie lorsque le défaut de normalité est potentiellement non borné, ce qui est le cas lorsque l'opérateur différentiel de la physique a lui-même une non-normalité qui dépend **exponentiellement** d'un paramètre physique qui tend vers l'infini (Gustavsson (1991), Schmid et Henningson (1992), Reddy, Trefethen, et Pathria (1993), Trefethen, Trefethen, Reddy, et Driscoll (1993)).

En pratique, le défi se traduit par un **effet d'échelle** : certains calculs qui donnent toute satisfaction dans une gamme éprouvée de tailles de modélisation deviennent non significatifs lorsque

[10] Néanmoins, puisque $||AA^* - A^*A|| \leq 2||A||^2$, la présence du coefficient $||A||$, pouvant être très grand, au dénominateur de la formule de l'erreur inverse relative normwise, incite à rester prudent sur les conséquences à en tirer.

la taille est sensiblement augmentée (Braconnier, Chatelin, et Dunyach (1995)). Il importe que l'ingénieur dispose des outils nécessaires pour décider si la cause des ennuis numériques est que le problème est trop instable ou bien que l'algorithme, ou la méthode numérique, est trop peu fiable.

Car si la méthode de calcul s'avère fiable, les difficultés de calcul sont inhérentes au problème (donc à la physique si la modélisation est correcte) et elles ne peuvent pas disparaître. Quelle est alors la valeur de l'information contenue dans un calcul à précision finie qui fournit des résultats faux si le calcul est trop instable ? C'est à cette difficile question des "calculs impossibles" que le Calcul Qualitatif tente, à partir d'expérimentations, d'apporter quelques réponses encore bien modestes par rapport à l'importance des enjeux.

10 Le Calcul Qualitatif

Le contrôle de l'erreur de calcul dont nous avons parlé jusqu'ici n'a de sens que si cette erreur tend vers 0, donc si le problème est bien posé. Mais, vue de l'ordinateur, la différence mathématique entre problèmes bien et mal posés peut devenir floue, comme le suggère l'exemple de la matrice de Schur cité au paragraphe précédent.

Plus généralement, comment peut-on espérer analyser les résultats d'un problème mal posé (la solution ne dépend pas continuement des données) dont tous les chiffres peuvent être faux ? La quantification (calculer juste) n'est plus possible (Chatelin (1988d, 1989a,b)). Il faut se contenter d'approches plus *qualitatives*. On ne peut pas calculer exactement la trajectoire d'un ouragan mais on peut répondre à la question qualitative suivante : doit-on évacuer Miami ?

Il convient, dans l'écheveau des résultats obtenus, de démêler ce qui reste calculable de ce qui ne présente plus aucune signification, c'est-à-dire de découvrir des formes plus faibles de **stabilité qualitative**. En l'absence de théorie, c'est une approche expérimentale de la stabilité que propose le Calcul Qualitatif à partir, entre autres, d'expérimentations sur ordinateurs menées, à l'aide de PRECISE, sur des problèmes dont les données sont aléatoirement perturbées.

10.1 Retrouver les caractéristiques des singularités algébriques

Le Calcul Qualitatif peut permettre de déterminer l'ordre d'une singularité algébrique (Chatelin, Frayssé, et Braconnier (1995)) et plus généralement de calculer la forme de Jordan la plus probable (Chatelin et Braconnier (1994)). A partir de la forme de Jordan la plus problable, on peut analyser qualitativement le comportement dynamique du problème et en tirer des informations utiles. Cette démarche est très voisine de celle utilisée en Physique Expérimentale pour déterminer les exposants critiques des transitions de phase, par exemple, lorsque la divergence est en loi de puissance (Gautier (1980)).

10.2 Au-delà de la théorie linéaire des perturbations

La théorie des perturbations est asymptotique, donc le domaine d'application de l'approximation linéaire peut être trop petit pour être appréciable par le calcul en double précision. On peut alors dans certains cas (racines de polynômes, spectres de matrices par exemple) confier à l'ordinateur le soin de visualiser les effets de perturbations des données d'amplitude relative t variant entre 0 et 1. La figure 1 (Chatelin et Frayssé (1993)) illustre le cas des zéros du polynôme $\Pi(x) = (x-1)^3(x-2)^3(x-3)^2(x-4)$ (calculés comme valeurs propres de la matrice compagne associée) en montrant la superposition des racines des polynômes perturbés pour t variant de 10^{-15} à 10^{-10}. Pour chaque valeur de t, la taille de l'échantillon est de 10.

Pour $t \sim 10^{-15}$, les effets observés sont conformes à ceux prédits par la théorie classique (développement en séries de Puiseux pour racines multiples) (voir le zoom autour de la racine 3 en Figure 2).

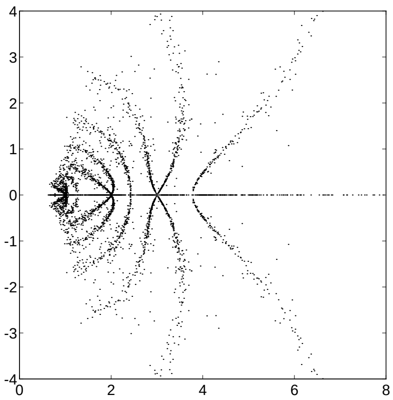


Figure 1 : *Racines de* $\Pi(x)$

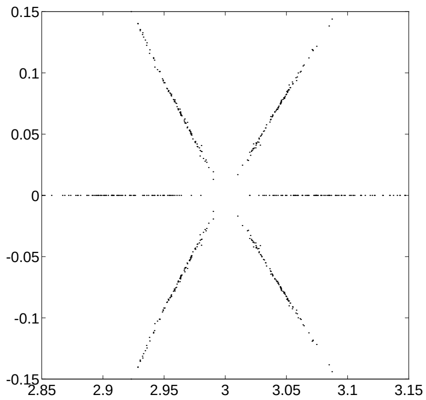


Figure 2 : *Zoom autour de* 3

Pour $10^{-14} \leq t \leq 10^{-10}$, on constate comme prévu des **effets non linéaires**. Les trois singularités 1, 2 et 3 ont une force de diffusion croissante.
De plus, pour $10^{-12} \leq t \leq 10^{-10}$, on voit apparaître des racines "parasites" autour des valeurs 1.5, 2.5 et 3.8 obtenues par couplage des racines voisines. C'est un phénomène **collectif** émergeant qui ne doit pas nécessairement être considéré comme faux. On constate donc que, pour le calcul de racines de polynômes, on obtient grâce à l'ordinateur des solutions (ici les racines parasites) qui n'ont pas nécessairement d'équivalent dans la réalité mathématique, c'est à dire dans le calcul exact, mais qui peuvent jouer un rôle important dans la réalité physique.

10.3 La logistique calculée

Le calcul sur ordinateur de la récurrence quadratique

$$x_0, \quad x_{k+1} = rx_k(1 - x_k) = f_r(x_k), \ k \geq 0 \tag{8}$$

peut, pour certaines valeurs de r, conduire avant la 100ième itération, à une valeur dont les 15 chiffres significatifs sont faux en double précision. Pourtant à chaque pas, l'évaluation arithmétique est correcte. Le phénomène est un reflet de l'instabilité mathématique (variable en fonction de r) de l'itération de point fixe $x_{k+1} = f_r(x_k)$ utilisée pour résoudre l'équation $x = f_r(x)$. Cette équation admet deux racines simples $x^* = 0$ et $x^* = 1 - 1/r$ pour $r \neq 0$, sauf en $r = 1$ où la racine unique est double.
Néanmoins, les visualisations du même calcul (8) obtenues par des ordinateurs très divers, possèdent toutes une remarquable similitude de forme (voir aussi les visualisations de Colonna (1992) concernant l'attracteur de Lorentz). D'ailleurs, la fiabilité arithmétique a pu être démontrée par une analyse inverse des erreurs (Hammel, Yorke, et Grebogi (1987)). C'est grâce à une telle fiabilité du calcul que les simulations sur ordinateur de phénomènes turbulents gardent une signification qualitative. Même si l'écoulement calculé n'est pas identique à celui observé en laboratoire, les comportements sont le plus souvent similaires dans leurs tendances..

10.4 Pour une réhabilitation du calcul sur ordinateur à précision finie

"Le calcul à précision finie a mauvaise presse, le calcul à précision finie fait parfois peur" disions-nous dans l'introduction. Nous espérons avoir convaincu que ces peurs sont injustifiées si l'on dispose de bons outils pour tester la fiabilité arithmétique des logiciels et la fiabilité numérique des méthodes. Les outils statistiques que propose PRECISE (voir la section 11) sont un premier pas dans la direction de l'analyse des calculs impossibles.

Nous voulons aller plus loin et argumenter que, même si c'était possible, vouloir calculer juste à tout prix n'est pas forcément une bonne idée. La divergence graduelle du calcul de l'itération (8) pourrait n'être qu'une curiosité sur ordinateur si, depuis sa mise en évidence (May (1976)), on ne l'avait rencontrée souvent dans les phénomènes naturels. Pour les lasers à gaz carbonique, le diagramme observé reproduit la logistique calculée et non pas la solution mathématique, c'est à dire l'ensemble des points fixes de $x = f_r^k(x),\ k \geq 1$ (Glorieux (1988)). Or, traiter l'itération (8) en calcul formel a justement pour effet de faire disparaître la zone de divergence graduelle qui a un sens physique.
Vouloir calculer juste est certes louable; *mais juste par rapport à quoi ?* Tout dépend du but que l'on se fixe. Il peut ne servir à rien de calculer exactement des solutions que leur instabilité rend inattaignables dans la réalité physique.

De plus, une remarquable correspondance se fait jour entre le calcul en précision finie au voisinage des singularités mathématiques et les phénomènes physiques au voisinage de points critiques (Berry (1991)). Le calcul en précision finie, donc en *information finie*, au voisinage de singularités, semble pouvoir servir de modèle simplifié à certains phénomènes physiques de transition de phases. Le calcul à précision finie, justement parce qu'il introduit nécessairement des perturbations, fait apparaître des **lois du calcul** qui peuvent éclairer certains phénomènes critiques, comme la limite de l'optique ondulatoire lorsque la longueur d'onde tend vers zéro. Que l'on songe à ces merveilleux phénomènes naturels que sont un arc-en-ciel, ou le scintillement de la mer sous le soleil, et qui s'expliquent par la perturbation de caustiques (Berry (1978)).

Pour toutes ces raisons, l'arithmétique d'ordinateur à précision finie devrait jouir d'un statut épistémologique bien supérieur à celui qui est le sien aujourd'hui, celui de suspect numéro un dès qu'un résultat calculé semble douteux. Tout au contraire, il nous semble que le caractère fini de l'information traitée par une telle arithmétique peut être un **atout** dans la simulation des phénomènes naturels à des fins d'analyse.

Chaitin-Chatelin et Frayssé (1995) présente une étude assez complète du calcul au voisinage de singularités algébriques. Mais le domaine des singularités essentielles reste à approfondir, afin d'enrichir la connaissance du mécanisme global de la propagation des erreurs dans les calculs associés à des dynamiques chaotiques.
On peut espérer déboucher sur une théorie de l'information adaptée au calcul scientifique. Cette théorie doit refléter le rôle de l'erreur dans le phénomènes d'évolution, permettant de mieux comprendre comment le calcul à précision finie au voisinage de singularités peut servir de modèle simplifié pour certains mécanismes d'émergence du nouveau qui sont à la base des lois du vivant (Langton (1990, 1991), Crutchfield et Young (1990)).

11 La boîte à outils PRECISE

La boîte à outils PRECISE (**PR**ecision **E**stimation and **C**ontrol **I**n **S**cientific and **E**ngineering computing) est un ensemble d'outils logiciels développés sous MATLAB pour aider l'utilisateur à expérimenter la stabilité de méthodes numériques et d'algorithmes, là où les calculs s'avèrent délicats, c'est à dire au voisinage de singularités.

PRECISE ne prétend pas être un outil de contrôle automatique de la précision. Il existe déjà de très nombreux outils de ce type (Erhel et Philippe (1991), Chaitin-Chatelin et Frayssé (1995)). L'outil PRECISE est présenté comme une aide à la compréhension des cas dificiles, afin de fournir à l'utilisateur les connaissances sur son problème nécessaires au choix de l'outil de contrôle le plus adapté à son cas. PRECISE permet d'explorer l'influence, sur la qualité de la convergence de méthodes numériques, de la précision finie ainsi que tout autre type de perturbations des données, à son choix.

PRECISE offre les possibilités suivantes :
1) un module pour une analyse inverse statistique des erreurs. Ce module fournit des estimateurs statistiques pour

- les conditionnements en des points réguliers ou singuliers pour l'algorithme / méthode et le problème,
- les erreurs inverses,
- les indices de fiabilité et de qualité arithmétique et numérique,
- les distances à la singularité, ou frontières dangereuses
- l'ordre des singularités de Hölder.

2) un module pour l'analyse de sensibilité. Ce module fournit des sorties graphiques de

- spectres perturbés,
- portraits spectraux et pseudospectres,
- portraits de sensibilité et pseudoracines de polynômes,
- portraits de divergence d'itérations dépendant d'un paramètre.

Annexes techniques

A1. Arithmétique des ordinateurs

1. Caractéristiques de l'arithmétique

Soit $[E_{\min}, E_{\max}]$ le domaine de variation de l'exposant E du nombre $X = m\beta^E$. On rappelle que β représente la base de l'arithmétique et que la mantisse m contient p chiffres significatifs en base β. L'ensemble des nombres X ainsi représentables en machine est $\mathbb{F}$, l'ensemble des nombres **flottants normalisés** ; il est caractérisé par 4 entiers $\{\beta, p, E_{\min}, E_{\max}\}$, dont les valeurs sont données pour différents ordinateurs dans la Table 1.

Ordinateur	β	p	$E_{\min}$	$E_{\max}$
IBM 370 et 3090				
simple précision	16	6	-64	63
double précision	16	14	-64	63
Convex DEC VAX/9000	2	55	-254	255
CRAY				
simple précision	2	48	-16384	8191
double précision	2	96	-16384	8191
Norme IEEE				
simple précision	2	24	-126	127
double précision	2	53	-1022	1023
calculettes	10	10	-99	99

Table 1 : *Caractéristiques de l'arithmétique*

Pour comparer les arithmétiques binaire et hexadécimale, on utilise le tableau de conversion suivant :

chiffres hexadécimaux	bits	exposant hexadécimal	exposant binaire
6	24	-64	-256
14	56	63	252

La perte moyenne en bits due à la normalisation hexadecimale est de 1.5 bit (resp. 2.5 bits) par rapport à la normalisation binaire (resp. avec premier bit implicite). Donc l'arithmétique hexadécimale sur un équivalent de 56 bits se comporte comme une arithmétique à 53 bits effectifs au moins.

On montre que la précision machine $\Psi = 1^+ - 1 = \min\{s;\ \mathrm{fl}(1+s) = 1+s > 1\}$, et que $\frac{1}{2}\Psi \leq \left|\frac{X^+ - X}{X}\right| \leq \Psi$. Donc $\Psi = \max_{X\neq 0}\left|\frac{X^+ - X}{X}\right|$ représente la précision relative maximale de la représentation des nombres réels par l'arithmétique de l'ordinateur.

Le plus petit nombre représentable est $\omega = \beta^{E_{\min}-1}$, le plus grand est $\Omega = \beta^{E_{\max}}(1-\beta^{-p})$. Ω (resp. ω) est la limite de dépassement (resp. sous-passement) de capacité, soit en anglais "overflow" (resp. "underflow"). L'arithmétique IEEE permet un sous-passement gradué à l'aide de nombres dénormalisés. La différence $\alpha = x - \mathrm{fl}(x)$ représente l'erreur d'arrondissage : $\mathrm{fl}(x) = x(1+\delta)$ avec $|\delta| \leq \mathbf{u}$. L'arrondissage de l'unité $\mathbf{u}$ dépend du mode d'arrondissage:

- arrondi au plus près : $\mathbf{u} = \frac{1}{2}\beta^{-p+1} = \frac{1}{2}\Psi$,
- troncature : $\mathbf{u} = \beta^{-p+1} = \Psi$.

2. Distribution des chiffres de la mantisse (Feldstein et Goodman (1976))

Soit x un réel positif de fraction F normalisée (premier chiffre non nul) illimitée dans la base β. Alors $\beta F = \beta^{\log_\beta x(\mathrm{mod}.1)}$. On dira que x admet une *distribution logarithmique* si et seulement si $\log_\beta x(\mathrm{mod}.1)$ est uniformément réparti sur $[0,1]$. Sous cette hypothèse, Feldstein et Goodman (1976) ont étudié la répartition de chaque chiffre de la fraction F d'un réel choisi au hasard selon la distribution logarithmique. Soit $p_k(a)$ la probabilité que le $k$ième chiffre en base β soit égal à a, $0 \leq a \leq \beta - 1$. Ils montrent que :

i) $p_1(a) = \log_\beta(1 + \frac{1}{a})$ pour $a \geq 1$,

ii) $p_k(a) = \log_\beta \prod_{j=\beta^{k-2}}^{\beta^{k-1}-1} \frac{j\beta+a+1}{j\beta+a}$, pour $k \geq 2$ et $\lim_{k\longrightarrow\infty} p_k(a) = 1/\beta$.

Le $k$ième chiffre admet donc une répartition qui tend vers une loi uniforme lorsque k tend vers l'infini. Cette limite est atteinte très vite en pratique. La Table 2 donne, en fonction de β, la valeur de k pour laquelle $1/\beta$ est atteinte avec au moins quatre chiffres décimaux. Le théorème de Feldstein-Goodman repose sur

β	2	3	4	8	10	16
k	13	10	8	5	5	4

Table 2 : *Répartition uniforme pour le chiffre de rang $\geq k$ en fonction de la base β*

l'hypothèse théorique d'une **distribution logarithmique** pour x supposé être à fraction illimitée. Bien que l'on ne sache pas si cette hypothèse est réaliste pour représenter le résultat d'un calcul effectué avec une mantisse finie, on peut constater le bon accord entre les distributions de chiffres théoriques prédites par Feldstein-Goodman et la distribution constatée expérimentalement et portant sur un ensemble de $2\ 10^5$ nombres (Benford (1938)). La forte présomption en faveur du résultat de Feldstein–Goodman est encore renforcée par un théorème de Turner (1982) qui établit, sans aucune hypothèse, la même distribution logarithmique du chiffre de tête du résultat d'un *produit* infini de nombres dont les chiffres sont uniformément distribués. En pratique un produit de 3 nombres est suffisant pour obtenir au moins 2 chiffres décimaux communs avec la distribution logarithmique limite. On en déduit qu'il est réaliste de supposer que la distribution du premier et du dernier chiffre d'un nombre choisi au hasard *parmi les résultats de suites "assez longues" de calculs, est donnée par le théorème de Feldstein-Goodman.* Le théorème de Turner, en s'intéressant à des multiplications successives, a l'intérêt de bien montrer le rôle de la non-linéarité des calculs dans le paradoxe de Newton–Borel sur la distribution du premier chiffre (Figures 3 et 4).

3. Fréquence des accidents arithmétiques (Feldstein et Turner (1986)

Feldstein et Turner (1986) interprètent l'hypothèse théorique de distribution logarithmique des nombres réels de la manière suivante, pour les nombres flottants $X = m\beta^E$:

i) la mantisse m est logarithmiquement distribuée dans $[1/\beta, 1[$,

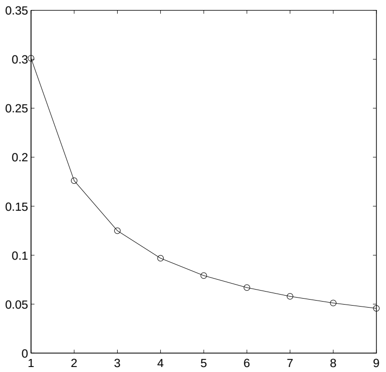


Figure 3 : *Distribution de probabilité du premier chiffre en base 10*

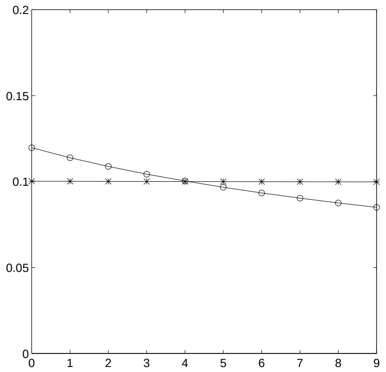


Figure 4 : *Distribution de probabilité du deuxième (o) et du quatrième (*) chiffre en base 10*

ii) l'exposant E est uniformément distribué sur les entiers de $[E_{\min}, E_{\max}]$.

Sous ces deux hypothèses, ils s'intéressent à la fréquence d'apparition d'accidents arithmétiques de type "overflow" ou "élimination catastrophique[11]". Ils établissent, entre autres, les tableaux suivants :

N	10^3	10^4	10^5	10^6
$p_N = 0.5$	35	111	351	1111
$p_N = 0.9$	90	285	901	2850
$p_N = 0.99$	292	923	2918	9228

Table 3 : $E_{\max}$ *requis pour avoir la probabilité* p_N *de non overflow dans* N *additions arithmétiques*

N	5	15	50	100
p_N	0.21	0.51	0.91	0.99

Table 4 : *Probabilité* p_N *de perte de signification pour* la moitié *des 27 bits de mantisse, au moins une fois au cours de* N *soustractions arithmétiques*

Le problème des overflows s'améliore en augmentant le domaine de variation de l'exposant. Celui de l'élimination catastrophique s'améliore en accroissant la taille de la mantisse.

4. Arithmétique aléatoire

Dans les notations du paragraphe 4.2, l'arithmétique aléatoire simule l'effet des arrondis de calcul au $p$ième chiffre à l'aide de calculs exécutés avec $q \geq p$ chiffres et un **arrondissage aléatoire** $\tilde{\alpha}(p, q, \pi)$ au $p$ième chiffre : $\tilde{\alpha}(p, q, \pi)$ résulte de l'arrondi déterministe dû au calcul avec q chiffres, auquel s'ajoute une perturbation aléatoire π autour du $p$ième chiffre, de distribution connue. Les exemples suivants d'arithmétique aléatoire ont été proposés dans la littérature :

[11] On appelle "élimination catastrophique" l'apparition brusque de chiffres non signifiants lors de la soustraction de deux nombres voisins de même signe (voir Annexe A2).

- $q = 2p$, échantillon de taille 2 pour le calcul en mode bruité (noisy mode) sur IBM 7030 (Stretch) (Bucholz (1962)).[12]
- $q = 2p$, π simule la troncature sur $r = p$ chiffres dans Hull et Swenson (1966).
- $q = p$, $\pi = (0, 1)$ avec probabilité 1/2 (troncature) (La Porte et Vignes (1974)) et $\pi = (-1, 0, 1)$ avec probabilités respectives $(1/4, 1/2, 1/4)$ (arrondi au plus près) (Vignes (1978)) dans les méthodes de permutation-perturbation et CESTAC, avec échantillon de taille 2 ou 3 (voir aussi Chesneaux et Vignes (1992), Pichat et Vignes (1993))
- $q \geq p$ dans PRECIX (Brunet (1989), Bennani, Brunet, et Chatelin (1988)) et dans SCALP (François (1989)); π est laissé au choix de l'utilisateur.

A2. Stabilité conditionnelle à l'arithmétique

Les deux exemples de base de stabilité conditionnelle à l'arithmétique sont liés à l'approximation d'une dérivée par une formule de différences finies. On sait que, mathématiquement, l'équation

$$\int_0^s x(t)dt = y(s) \Longleftrightarrow x = y'$$

est mal posée : l'application $y \longmapsto y'$ ne dépend pas continuement de y. A une variation $\Delta y = \varepsilon \sin \omega t$ correspond $\Delta x = \varepsilon\omega \cos \omega t$. Et $\|\Delta x\|_\infty = \varepsilon\omega$ peut être arbitrairement grand lorsque $\varepsilon \to 0$, si l'on choisit par exemple $\omega = 1/\varepsilon^2$.

1. Elimination catastrophique arithmétique (Chatelin (1989b))

On cherche à calculer $x = \dfrac{(1+y)-1}{y} = G(y)$ pour des valeurs de $y \neq 0$, sur ordinateur (ou calculette). Les résultats $x_\varepsilon = G_\varepsilon(y)$ en arithmétique IEEE sont donnés sur la Figure 5. Sur le graphe de gauche (resp. droite), y varie dans $[10^{-16}, 10^{-14}]$ (resp. $[5\ 10^{-17}, 5\ 10^{-16}]$). Mathématiquement, $x = G(y) = 1$ pour tout $y \neq 0$, et $\lim_{y \to 0} G(y) = 1$. Sur ordinateur, on calcule $x_\varepsilon = G_\varepsilon(y)$, et un phénomène d'élimination arithmétique apparaît pour y assez petit ($y \sim \Psi$). L'élimination se produit lorsque $1 + y$ et y ne sont plus distincts en machine. Alors $(1 + y) - 1$ est calculé à zéro.
On voit que la soustraction sur ordinateur de deux nombres voisins peut entrainer une grande erreur *relative*. $x_\varepsilon = G_\varepsilon(y)$ est discontinu pour certaines valeurs de y : il n'y a pas équicontinuité en ε pour tout y. En particulier, $\lim_{\varepsilon \to 0, y \to y} G_\varepsilon(y) = 1$ seulement si $y > \Phi(\varepsilon)$.

2. Elimination catastrophique numérique (Kulisch et Miranker (1986))

Soit

$$f(t) = \frac{4970t - 4923}{4970t^2 - 9799t + 4830}.$$

On veut calculer $x = f"(1) = 94$ par différences finies : x est approché par $x_h = \frac{f(1+h)-2f(1)+f(1-h)}{h^2} = G_h(1)$. Les valeurs $x_{h\varepsilon} = G_{h\varepsilon}$ pour h au voisinage de 0 sont données sur la Figure 6. On voit apparaître un phénomène semblable au précédent mais amplifié par la division par h^2. Il y a une perte importante de signification quand les valeurs calculées pour $f(1 + h) - f(1)$ et $f(1) - f(1 - h)$ ne diffèrent que *sur des chiffres faux*. En double précision ($\Psi \sim 10^{-15}$), $f(1) = 47$ est exact mais $f(1 + h)$ et $f(1 + h)$ n'ont que 12 chiffres corrects. L'élimination numérique a lieu pour $h \sim 10^{-5}$. Il est inutile de faire decroître h davantage. Les résultats calculés se détériorent à cause de la division par h^2. Les quatre graphes correspondent aux intervalles $[h_i, 10h_i]$, $i = 1, 2, 3, 4$ avec $h_1 = 10^{-5}$, $h_2 = 10^{-7}$, $h_3 = 10^{-9}$, $h_4 = 10^{-12}$. L'approximation de $f"(1)$ par differences finies n'est un processus stable que si $h \geq 10^{-5}$ pour $\Psi \sim 10^{-15}$.

[12] Dans ce cas précis, π n'est pas aléatoire : les calculs sont effectués avec les $r = p$ bits de poids faible tous égaux à 0 ou à 1.

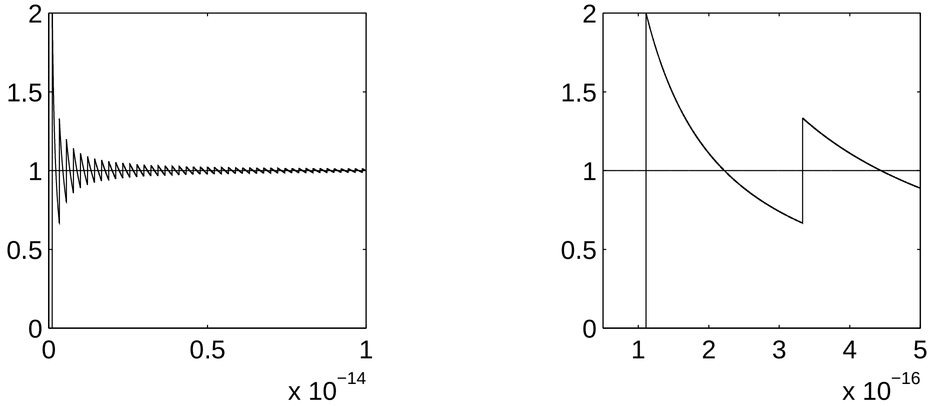


Figure 5 : *Elimination catastrophique - Calcul de* $\frac{(1+y)-1}{y}$ *quand* $y \to 0$.

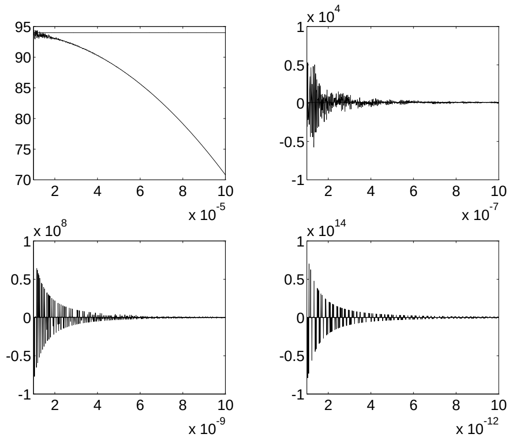


Figure 6 : *Elimination catastrophique - Calcul de* $f''(1)$ *par différences finies.*

A3. Fiabilité arithmétique

1. Calcul du résidu à la précision courante

Calculer le résidu à la précision courante est suffisant pour obtenir une information pertinente sur la fiabilité arithmétique du problème étudié. Par exemple, pour un système linéaire $Ax = b$, Higham (1991) montre que, si le résidu $r = Ax_\epsilon - b$ est calculé de manière conventionelle, le résidu calculé r_ϵ satisfait

$$|r - r_\epsilon| \leq \mathbf{u}c(|A||x_\epsilon| + |b|)$$

où c est une constante de l'ordre de n, la taille de A. Dans ce cas, l'erreur absolue commise sur le calcul de l'erreur inverse (qu'elle soit normwise ou componentwise) n'excède pas $\mathbf{u}c$, ce qui reste suffisant pour juger de la fiabilité de l'algorithme considéré.

2. Algorithme de Gauss

Voyons sur un exemple archétypique, l'algorithme de Gauss pour résoudre $Ax = b$, comment se déroule l'analyse de l'erreur. Soit $A = LU$ la factorisation exacte de Gauss. La solution calculée x_ϵ est la solution exacte dans $\mathbb{R}$ de (P_ϵ) : $A_\epsilon x_\epsilon = b_\epsilon$ avec $A_\epsilon = L_\epsilon U_\epsilon$ (Wilkinson (1961)). Soit le calcul successif de y_ϵ par $L_\epsilon y_\epsilon = b_\epsilon$ puis de x_ϵ tel que $U_\epsilon x_\epsilon = y_\epsilon$ ou encore $x_\epsilon = G_\epsilon(b_\epsilon)$, version en précision finie de $x = G(b)$.
Remarquons que G est une représentation de l'inverse différente de A^{-1}. Le résultat x est le même, mais les routes empruntées par le calcul sont différentes. En effet, $x = A^{-1}b$ se détermine par les formules de Cramer, et x existe si et seulement si A^{-1} est régulier. Tandis que, lorsque x est calculé par la méthode de Gauss, on a la condition supplémentaire d'existence que les "pivots" (mineurs) soient non nuls. Dans ce cas le problème et l'algorithme ont des stabilités **différentes**. On appelle stabilité mathématique (resp. algorithmique) la continuité de F^{-1} (resp. G).
Pour montrer la stabilité de la méthode de Gauss, Wilkinson (1961) démontre que

$$\|A - A_\epsilon\| \leq 8n^3 \rho_n \|A\|_\infty \epsilon \tag{9}$$

où n est l'ordre de A et ρ_n est le facteur de croissance des éléments de U au cours du calcul. L'inégalité (9) a trois conséquences :

i) la stabilité uniforme en ϵ de A_ϵ.
Avec $E = A - A_\epsilon$, $A_\epsilon = (I - EA^{-1})A$ et $A_\epsilon^{-1} = A^{-1}(I - EA^{-1})^{-1}$. Donc, si $\|E\| < \frac{1}{2}\|A^{-1}\|$, soit pour ϵ assez petit, $\sup_\epsilon \|A_\epsilon^{-1}\| \leq 2\|A^{-1}\|$. La convergence $x_\epsilon \longrightarrow x$ quand $\epsilon \longrightarrow 0$ en découle. On a aussi la stabilité résiduelle $\sup_\epsilon \|AA_\epsilon^{-1}\| \leq 2$.

ii) la fiabilité arithmétique de l'algorithme.
L'identité $G_\varepsilon(b_\varepsilon) - G(b_\varepsilon) = (A_\varepsilon^{-1} - A)b_\varepsilon = A_\varepsilon^{-1}(A - A_\varepsilon)A^{-1}b_\varepsilon$ permet de déduire que l'erreur de consistance en b_ε est d'ordre 1 en ε.

iii) la qualité arithmétique de l'algorithme.
L'erreur inverse est $EI(x_\varepsilon) = \|Ax_\varepsilon - b\|/(\|A\|\|x_\varepsilon\| + \|b\|)$. De l'égalité $Ax_\epsilon - b = AA_\epsilon^{-1}[(A_\epsilon - A)x + (b_\epsilon - b)]$ et des bornes sur $\|A - A_\epsilon\|$ et $\|b - b_\epsilon\|$, on déduit que EI est de l'ordre de Ψ. Les constantes obtenues dépendent de n, des détails de l'arithmétique, et surtout du facteur ρ_n qui est infini en cas de pivot nul. Même s'il n'y a pas de pivot nul, ρ_n peut croître exagérément. C'est la raison pour laquelle l'algorithme de Gauss est utilisé avec *pivotage* partiel ou complet.

3. Cas où les résultats sont plus justes que prédits

La borne supérieure de l'erreur exprimée par l'inégalité (7) à la page 16 peut ne pas être atteinte par la solution calculée lorsque la perturbation engendrée par le calcul a une structure très particulière par rapport à l'ensemble $\mathcal{E}$ des perturbations admissibles choisies pour l'analyse. Nous citons deux exemples (voir Chaitin-Chatelin et Frayssé (1995) pour plus de détails) :

i) Pour certaines matrices *triangulaires*, le conditionnement componentwise peut être borné indépendamment de A^{-1}. Donc tout algorithme qui génère des perturbations componentwise (comme l'algorithme de Gauss) donne des résultats très précis, même au voisinage de la singularité (Higham (1989)).

ii) L'algorithme de Björck-Pereyra appliqué aux matrices de Vandermonde (qui interviennent dans les problèmes d'interpolation polynômiale et qui sont connues pour leur très mauvais conditionnement) engendrent des perturbations componentwise spécifiques à ce cas particulier (Higham (1987)). Cela permet de résoudre avec une grande justesse certains systèmes de Vandermonde.

A4. Conditionnement d'une singularité

On rappelle qu'une fonction $x \longmapsto y = F(x)$ *Hölder-continue d'ordre* γ au voisinage de x satisfait

$$\|\Delta x\| \leq K\|\Delta y\|^{\gamma}.$$

En un point singulier höldérien d'ordre γ, $\|\Delta x\|$ est donc (asymptotiquement) proportionnel à $\|\Delta y\|^{\gamma}$ et le conditionnement hölderien (s'il existe) :

$$C^{\gamma} = \lim_{\|\Delta y\| \longmapsto 0} \frac{\|\Delta x\|}{\|\Delta y\|^{\gamma}},$$

joue le rôle tenu par le conditionnement linéaire $C = \lim \|\Delta x\|/\|\Delta y\|$ $(\gamma = 1)$ en un point régulier. Si l'on choisit de mesurer la régularité d'un problème par son ordre de Hölder-continuité, on peut dire que la régularité d'un problème régulier est d'ordre 1, tandis que celle d'un problème singulier Hölder-continu est d'ordre < 1.

La conjecture proposée par Chatelin et Frayssé (1993) peut être rapprochée des inégalités asymptotiques, valables pour $K \longrightarrow \infty$, données par Demmel (1987, 1990) :

$$\frac{a}{K^2} \leq \delta \leq \frac{b}{K},$$

qu'elle rend plus précises.

A5. Portraits de divergence pour la logistique

1. La logistique réelle

L'équation logistique (en variable réelle) s'écrit

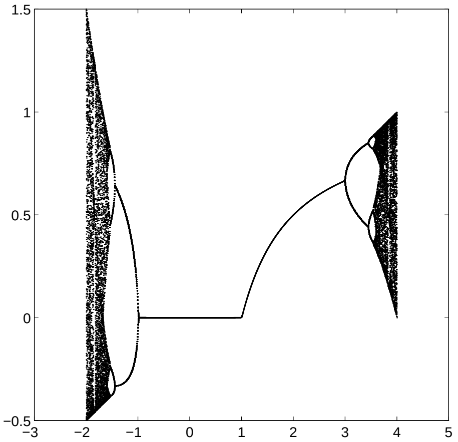
$$x = f_r(x) = rx(1-x), \tag{10}$$

où f_r est une application non linéaire du 2ième degré en x dépendant du paramètre réel r. L'équation (10) admet deux racines (simples pour $r \neq 1$) $x^* = 0$ et $x^* = 1 - 1/r$, $r \neq 0$. Lorsque $r = 1$, il n'y a qu'une racine double $x^* = 0$: l'équation admet donc une solution singulière pour $r = 1$. Que donne le calcul sur ordinateur des solutions de (10) par itération de point fixe

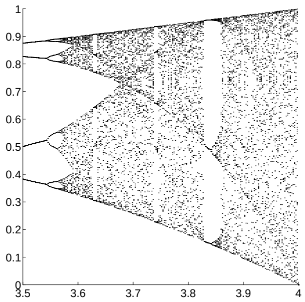
$$x_0, \; x_{k+1} = f_r(x_k), \; k \geq 0 \tag{11}$$

pour r variant de $-\infty$ à $+\infty$? Les valeurs calculées à partir de $x_0 = 1/2$, pour k variant entre 201 et 250 sont données sur la Figure 7. Le graphe de gauche (resp. droite) correspond à r dans $[-3, 5]$ (resp. $[-3.5, 4]$).

On constate que l'itération de point fixe ne converge que pour r dans $[-1, 3]$ (où $|f'(x)| \leq 1$). Pour r dans $[-2, 1]$ et $[3, 4]$, il y a divergence graduelle (les valeurs des x_k sont bornées). A l'extérieur, $r > 4$ ou $r < -2$), il y a divergence vers $\pm\infty$.



La logistique sur $[-2, 4]$



Zoom sur $[3.5, 4]$

Figure 7 : *Portrait de divergence de la logistique réelle*

1. La logistique complexe

Considérons maintenant l'analogue de (10) en variable complexe, soit

$$z = f_\lambda(z) = \lambda z(1 - z), \tag{12}$$

dont les solutions sont approchées sur ordinateur par

$$z_0 = 1/2, \ z_{k+1} = \lambda z_k(1 - z_k), \ k \geq 0 \tag{13}$$

pour $\lambda \in \mathbb{C}$. Selon les valeurs de λ, (13) peut converger, rester bornée, ou diverger. On considère l'application $\lambda \longmapsto k_\lambda$ où k_λ est l'indice de la première itération pour laquelle $|z_k(z_k - 2)| \geq 8$, où $\{z_k\}$ est calculé par (13). Cette application est le *portrait de divergence* de (13). Voir la Figure 8. La partie gauche (resp. droite) correspond à la région $[-2, 4] \times [-3, 3]$ (resp. $[-0.13, 0.13] \times [1, 1.26]$ de $\mathbb{C}$. On calcule un nombre maximum d'itérations égal à 32. Les 32 régions correspondant à $k_\lambda = 1$ à 32 sont représentées avec une échelle en teinte de gris de 9 niveaux. Le noir correspond à $k_\lambda > 32$. On reconnait le célèbre ensemble de Mandelbrot. Il est défini par $\{\lambda;\ |z_k| \not\to \infty\}$ et correspond bien à $k_\lambda > 32$.

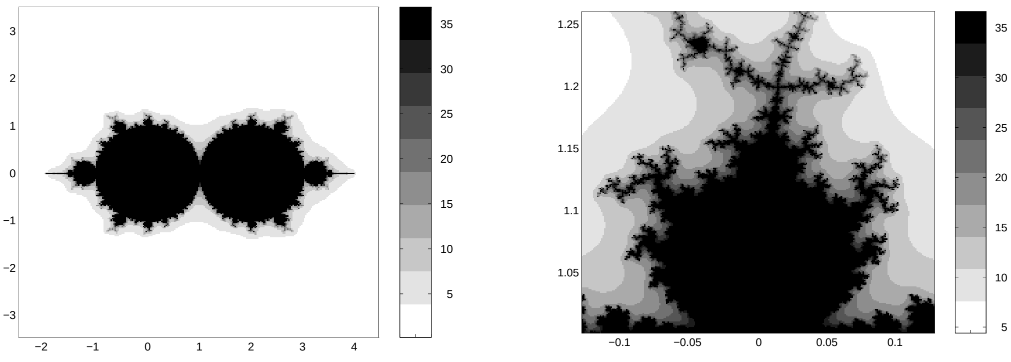


Figure 8 : *Portrait de divergence de la logistique complexe*

Bibliographie

E. ANDERSON, Z. BAI, C. BISCHOF, J. DEMMEL, J. DONGARRA, J. DU CROZ, A. GREENBAUM, S. HAMMARLING, A. MCKENNEY, S. OSTROUCHOV, ET D. SORENSEN, (1992), **LAPACK User's Guide**, SIAM, Philadelphia.

M. ARIOLI ET F. ROMANI, (1992), *Stability, convergence and conditioning of stationary iterative methods of the form* $x^{(i+1)} = Px^{(i)} + q$ *for the solution of linear systems*, IMA J. Numer. Anal., **12**, 21–30.

J. P. AUBIN, (1972), **Approximation of elliptic boundary value problems**, Wiley, New-York.

J. P. AUBIN, (1994), **Neural networks and qualitative physics**, Cambridge U. Press.

F. L. BAUER, (1966), *Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme*, ZAMM, **46**, 409–421.

F. L. BAUER, (1974), *Computational graphs and rounding error*, SIAM J. Numer. Anal., **11**, 2, 87–96.

M. BECKER ET P. BECKER, (1977), *Importance des erreurs dues à la précision limitée de l'ordinateur pour la validité des simulations*, RAIRO Inf., **11**, 389–398.

F. BENFORD, (1938), *On the law of anomalous numbers*, Proc. Amer. Philos. Soc., **78**, 551–572.

M. BENNANI, M. C. BRUNET, ET F. CHATELIN, (1988), *De l'utilisation en calcul matriciel de modèles probabilistes pour la simulation des erreurs de calcul*, C. R. Acad. Sci. Paris, **307**, 847–850. Série I.

M. BENNANI ET T. BRACONNIER, (1994a), *Comparative behaviour of eigensolvers on highly non-normal matrices.* Submitted to IMA J. Numer. Anal.

M. BENNANI ET T. BRACONNIER, (1994b), *Stopping criteria for eigensolvers.* Submitted to J. Num. Lin. Alg. Appl.

M. BENNANI, (1991), *A propos de la stabilité de la résolution d'équations sur ordinateurs*, Ph. D. dissertation, Institut National Polytechnique de Toulouse.

S. BERNSTEIN, (1927), *Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes*, Mathem. Annalen, **97**, 2–59.

M. BERRY, (1978), *Les jeux de lumière dans l'eau*, La Recherche, **72**, 760–768.

M. BERRY, (1991), *Asymptotics, singularities, and the reduction of theories*, in **Proceedings of the 9th Congress of Logic, Methodology and Philosophy of Science**, D. Prawitz, B. Skyrms, et D. Westerståhl, eds., Studies in Logic and the Foundations of Mathematics, Uppsala.

Å. BJÖRCK ET V. PEREYRA, (1970), *Solution of Vandermonde systems of equations*, Math. Comp., **24**, 112, 893–903.

B. BLISS, M. C. BRUNET, ET E. GALLOPOULOS, (1992), *Automatic parallel Program Instrumentation with Applications in Performance and Error Analysis*, in **Expert systems for Scientific Computing**, E. N. Houstis, J. R. Rice, et R. Vichnevtsky, eds., North-Holland and IMACS, 235–260.

E. Borel, (1909), *Les probabilités dénombrables et leurs applications arithmétiques*, Rend. Circ. Mat. Palermo, **27**, 247–271.

T. Braconnier, F. Chatelin, et J.-C. Dunyach, (1995), *Highly nonnormal eigenvalue problems in the aeronautical industry*, Japan J. of Industrial Appl. Math., **12**, 1.

T. Braconnier, (1994), *Sur le calcul de valeurs propres en précision finie*, Ph. D. dissertation, Université H. Poincaré, Nancy.

R. P. Brent, (1973), *On the precision attainable with various floating point number systems*, IEEE Trans. Comp. C, **22**, 6, 598–601.

D. Brouwer, (1937), *On the accumulation of errors in numerical integration*, Astron. J., **46**, 149–153.

M. C. Brunet, (1989), *Contribution à la fiabilité de logiciels numériques et à l'analyse de leur comportement : une approche statistique*, Ph. D. dissertation, Université Paris IX Dauphine.

W. Bucholz, ed., (1962), **Planning a computer : project Stretch**, Mc Graw Hill, New-York.

Carvallo, (1852), *Notes sur les conditions de stabilité de ponts suspendus*, Annales Ponts et Chaussées, **2**, 211–226.

F. Chaitin–Chatelin et V. Toumazou, (1995), *Possibilités et limites du calcul sur ordinateur.* Cours de formation continue, ENSTA, Paris, 25-29 Septembre.

F. Chaitin-Chatelin et V. Frayssé, (1995), **Lectures on Finite Precision Computations**, SIAM, Philadelphia. To appear.

F. Chaitin-Chatelin, (1994), *Is nonnormality a serious difficulty ?*, Tech. Rep. TR/PA/94/18, CERFACS. Presented at ILAS 94, Rotterdam. Submitted to Lin. Alg. Appl.

G. J. Chaitin, (1991), *Le hasard des nombres*, La Recherche, **232**, 610–515.

F. Chatelin et T. Braconnier, (1994), *About the qualitative computation of Jordan forms*, ZAMM, **74**, 105–113.

F. Chatelin et M.-C. Brunet, (1990), *A probabilistic round-off error propagation model. Application to the eigenvalue problem.*, in **Reliable Numerical Computation**, M. G. Cox et S. Hammarling, eds., Oxford, Oxford University Press, 139–160.

F. Chatelin et V. Frayssé, (1991), *Analysis of arithmetic algorithms : a statistical study*, in **10th IEEE Symposium on Computer Arithmetic**, P. Kornerup et D. Matula, eds., 10–16.

F. Chatelin et V. Frayssé, (1992), *Elements of a condition theory for the computational analysis of algorithms*, in **Iterative methods in linear algebra**, R. Beauwens et P. De Groen, eds., Amsterdam, North-Holland, 15–25.

F. Chatelin et V. Frayssé, (1993), *Qualitative Computing : elements of a theory for finite precision computation.* Lecture Notes for the Comett European Course, June 8-10, Thomson-CSF, LCR Corbeville, Orsay.

F. Chatelin et P. Trouvé, (1993), *Le calcul scientifique, pour quoi faire ?*, Bulletin d'information du Laboratoire Central de Recherches et des Laboratoires Electroniques de Rennes, Thomson–CSF, 3–10.

F. Chatelin, V. Frayssé, et T. Braconnier, (1993), *Qualitative Computing : elements of a theory for finite precision computation*, Tech. Rep. TR/PA/93/12, CERFACS. Lecture Notes for the Workshop on Reliability of Computations, March 30 - April 1, Toulouse, France.

F. Chatelin, V. Frayssé, et T. Braconnier, (1995), *Computations in the neighbourhood of algebraic singularities*, Num. Funct. Anal. Opt., **16**.

F. Chatelin, (1983), **Spectral approximation of linear operators**, Academic Press, New-York.

F. Chatelin, (1986), *Ill-conditioned eigenproblems*, in **Large scale eigenvalue problems**, J. Cullum et R. A. Willoughby, eds., North-Holland, Amsterdam, 267–282.

F. Chatelin, (1988a), *Calculer juste, est-ce bien utile ?*, Nouvelles IBM–France, 3, 18–21.

F. Chatelin, (1988b), *Etude statistique de la qualité numérique et arithmétique de la résolution approchée d'équations par calcul sur ordinateur*, Tech. Rep. F. 133, Centre Scientifique IBM–France.

F. Chatelin, (1988c), **Valeurs propres de matrices**, Masson, Paris.

F. Chatelin, (1989a), *Comment traiter les calculs impossibles ?*, La Recherche, **214**, 1268–1270.

F. Chatelin, (1989b), *Résolution approchée d'équations sur ordinateur.* Notes de DEA, Université Paris VI.

F. Chatelin, (1991), *Les ordinateurs calculent faux et c'est tant mieux.* Journées Observatoire de Paris.

F. Chatelin, (1993a), **Eigenvalues of matrices**, Wiley, Chichester. Enlarged Translation of Chatelin (1988c).

F. Chatelin, (1993b), *The influence of nonnormality on matrix computations*, in **Linear Algebra, Markov Chains and Queueing Models**, R. J. Plemmons et C. D. Meyer, eds., Springer, New-York, 13–19.

J. M. Chesneaux et J. Vignes, (1992), *Les fondements de l'arithmétique stochastique*, C. R. Acad. Sci. Paris, **315**, 1435–1440. Série I.

W. J. Cody et Al., (1984), *A proposed radix-and-word-length-independant standard for floating point arithmetic*, IEEE Micro, **4**, 86–100.

W. J. Cody, (1973), *Static and dynamic numerical characteristics of floating point arithmetic*, IEEE Trans. Comp. C, **22**, 6, 598–601.

J.-F. Colonna, (1992), *Erreurs d'arrondi*, Pour la science, **179**, 104–107.

H. Cramer, (1946), **Mathematical methods of statistics**, Princeton University Press.

J. Crutchfield et K. Young, (1990), *Computation at the onset of Chaos*, in **Complexity, Entropy and the Physics of Information**, W. Zurek, ed., Addison-Wesley, 223–269.

J. W. Demmel, (1987), *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., **51**, 251–289.

J. W. DEMMEL, (1990), *Nearest defective matrices and the geometry of ill-conditioning*, in **Reliable Numerical Computation**, M. G. Cox et S. Hammarling, eds., Oxford, Clarendon Press, 35–55.

J. W. DEMMEL, (1992), *The componentwise distance to the nearest singular matrix*, SIAM J. Matrix Anal. Appl., **13**, 1, 10–19.

J. W. DEMMEL, (1993), *Geometry and High Accuracy Algorithms.* The James H. Wilkinson Prize in Numerical Analysis and Scientific Computing award and presentation, SIAM Annual Meeting.

S. DINNEMATIN, (1987), *Les ordinateurs calculent faux*, Science et Vie Micro, **41**.

I. EKELAND, (1984), **Le calcul, l'imprévu**, Seuil.

J. ERHEL ET B. PHILIPPE, (1989), *Qualité numérique des logiciels scientifiques : étude de faisabilité du projet AQUARELS*, Note Technique 8904, IRISA, Rennes, France.

J. ERHEL ET B. PHILIPPE, (1991), *Dossier de spécifications d'AQUARELS*, Rapport Convention Recherche 873, CNES - INRIA.

A. FELDSTEIN ET R. GOODMAN, (1976), *Convergence estimates for the distribution of trailing digits*, J. ACM, **23**, 287–297.

A. FELDSTEIN ET R. GOODMAN, (1982), *Loss of significance in floating point subtraction and addition*, IEEE Trans. Comp. C, **31**, 328–335.

A. FELDSTEIN ET P. TURNER, (1986), *Overflow, underflow and severe loss of significance in floating point addition and subtraction*, IMA J. Numer. Anal., **6**, 241–251.

G. E. FORSYTHE, (1950), *Round-off errors in numerical integration on automatic machinery*, Bull. AMS, **56**, 61. Abstract.

P. FRANÇOIS, (1989), *Contribution à l'étude de méthodes de contrôle automatique de l'erreur d'arrondi : la méthodologie SCALP*, Ph. D. dissertation, Institut National Polytechnique de Grenoble.

V. FRAYSSÉ, (1992), *Reliability of computer solutions (Sur la fiabilité des calculs sur ordinateurs)*, Ph. D. dissertation, Institut National Polytechnique de Toulouse.

F. GAUTIER, (1980), *Introduction to the study of phase transitions*, in **Solid state phase transformations in metal and alloys**, Aussois 78, Les Editions de Physique, Orsay, 223–236.

A. J. GEURTS, (1982), *A contribution to the theory of condition*, Numer. Math., **39**, 85–96.

W. GIVENS, (1954), *Numerical computation of the characteristic values of a real symmetric matrix*, Tech. Rep. ORNL-1574, Oak Ridge National Laboratory.

P. GLORIEUX, (1988), *Chaos et crises dans les lasers*, Images de la physique 88, Courrier du CNRS. suppl. no 71.

S. K. GODUNOV, (1992), *Spectral portraits of matrices and criteria of spectrum dichotomy*, in **Computer arithmetic and enclosure methods**, J. Herzberger et L. Atanassova, eds., North-Holland and IMACS.

D. Goldberg, (1991), *What every scientist should know about floating-point arithmetic*, ACM Computing Surveys, **23**, 1, 5–48.

H. H. Goldstine et J. von Neumann, (1951), *Numerical inverting of matrices of high order II*, Proc. Amer. Math. Soc., **2**, 188–202.

G. Golub et C. Van Loan, (1989), **Matrix computations**, John Hopkins University Press. Second edition.

G. H. Golub et J. H. Wilkinson, (1976), *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Review, **18**, 4, 578–619.

R. Goodman et A. Feldstein, (1975), *Round-off errors in products*, Computing, **15**, 263–273.

R. Goodman et A. Feldstein, (1977), *Effect of guard digits and normalizations options in floating point multiplication*, Computing, **18**, 93–106.

R. Goodman, A. Feldstein, et J. Bustoz, (1985), *Relative error in floating point multiplication*, Computing, **35**, 127–139.

L. H. Gustavsson, (1991), *Energy growth of three-dimensional disturbances in plane Poiseuille flow*, J. Fluid Mech., **224**, 241–260.

S. Hammarling et J. H. Wilkinson, (1976), *The practical behaviour of linear iterative methods with particular reference to SOR*, Tech. Rep. NAC 69, NPL, Teddington.

S. H. Hammel, J. A. Yorke, et C. Grebogi, (1987), *Do numerical orbits of chaotic dynamical processes represent true orbits ?*, J. Complexity, **3**, 136–145.

R. W. Hamming, (1970), *On the distribution of numbers*, Bell Syst. Tech. J., **49**, 1609–1625.

P. Henrici, (1962a), *Bounds for iterates, inverses, spectral variation and field of values of non-normal matrices*, Numer. Math., **4**, 24–40.

P. Henrici, (1962b), **Discrete variable methods in ordinary differential equations**, Wiley, New-York.

P. Henrici, (1963), **Error propagation for difference method**, Wiley, New-York.

P. Henrici, (1966), *Test of probabilistic models for propagation of roundoff errors*, Comm. ACM, **9**, 409–410.

N. J. Higham, (1987), *Error analysis of the Björck-Pereyra algorithm for solving Vandermonde systems*, Numer. Math., **50**, 613–632.

N. J. Higham, (1989), *The accuracy of solutions to triangular systems*, SIAM J. Numer. Anal., **26**, 5, 1252–1265.

N. J. Higham, (1991), *Iterative Refinement Enhances the Stability of QR Factorization Methods for Solving Linear Equations*, BIT, **31**, 447–468.

T. E. Hull et J. R. Swenson, (1966), *Tests of probabilistic models for propagation of round-off*, Com. of the ACM, **9**, 2, 108–113.

H. D. Huskey, (1949), *On the precision of a certain procedure of numerical integration*, J. Res. Nat. Bur. Stand., **42**, 57–62.

IEEE P.754, (1981), *A proposed standard for binary floating point arithmetic.*

M. Jankowski et H. Woźniakowski, (1977), *Iterative refinement implies numerical stability*, BIT, **17**, 303–311.

B. Kågström et A. Ruhe, (1981), *An algorithm for numerical computation of the Jordan canonical form of a complex matrix*, ACM, **6**, 3, 398–419.

W. Kahan, (1972), *Conserving confluence curbs ill-condition*, Technical Report 6, Department of Computer Science, University of California, Berkeley.

C. S. Kenney et A. J. Laub, (1994), *Small sample statistical condition estimates for general matrix functions*, SIAM J. Sc. Comp., **15**, 36–61.

W. Kerner, (1989), *Large scale complex eigenvalues problems*, J. Comp. Phys., **85**, 1–85.

D. E. Knuth, (1969), **The art of computer programming. Vol. 2 : Semi-numerical algorithms**, Addison Wesley, Reading Mass.

H. Kuki et W. J. Cody, (1973), *A statistical study of the accuracy of floating point number systems*, Com. ACM, **16**, 4, 610–620.

U. Kulisch et W. L. Miranker, (1981), **Computer arithmetic in theory and practice**, Academic Press, New-York.

U. Kulisch et W. L. Miranker, (1986), *The arithmetic of the digital computer : a new approach*, SIAM Review, **28**, 1, 1–40.

M. La Porte et J. Vignes, (1974), **Arithmétique des ordinateurs, Tome 1 : systèmes linéaires**, Technip, Paris.

C. Langton, (1990), *Computation at the edge of chaos : phase transitions and emergent computation*, Physica D, **42**, 1, 12–37.

C. Langton, (1991), *Life at the edge of chaos*, in **Artificial Life II**, C. Langton, C. Taylor, J. Farmer, et S. Rasmussen, eds., ACM Conference, Paris, 41–91.

P. S. Laplace, (1812), *Théorie analytique des probabilités*, Paris.

J. L. Larson et A. H. Sameh, (1980), *Efficient calculation of the effects of roundoff errors*, Computing, **24**, 275–297.

P. Levy, (1937), **Théorie de l'addition des variables aléatoires**, Gauthier-Villars, Paris.

J.-L. Lions, (1992), *World mathematical year 2,000 and computer sciences*, in **Future Tendencies in Computer Science, Control and Applied Mathematics**, A. Benssoussan et J.-P. Verjus, eds., no. 653 in Lecture Notes in Computer Science, Springer-Verlag, 3–16.

E. Lorentz, (1963), Trans. N.Y. Acad. Sc., **25**, 409.

J. D. Marasa et D. W. Matula, (1973), *A simulative study of correlated error propagation in various finite precision arithmetics*, IEEE Trans. Comp. C, **22**, 587–597.

E. Marshall, (1992), *Fatal error : how Patriot overlooked a Scud*, Science, 1347.

[Mat1990] THE MATH WORKS INC., (1990), **Pro - Matlab User's Guide for SUN workstations**. Matlab Version 4.0a.

R. MAY, (1976), *Simple mathematical models with very complicated dynamics*, Nature, 261.

W. MILLER, (1975), *Software for roundoff analysis*, ACM transactions on mathematical software, **1**, 108–128.

J. M. MULLER, (1989), **Arithmétique des ordinateurs**, Masson.

S. NEWCOMB, (1881), *Note on the frequency of use of different digits in natural numbers*, Amer. J. Math., **4**, 39–40.

W. OETTLI ET W. PRAGER, (1964), *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., **6**, 405–409.

M. PICHAT ET J. VIGNES, (1993), **Ingénierie du contrôle de la précision des calculs sur ordinateur**, Editions Technip, Paris.

H. A. RADEMACHER, (1948), *On the accumulation of errors in processes of integration on high-speed calculating machines*, The annals of the Comp. Lab., Harvard U., **16**, 176–185.

G. RAUZY, (1976), **Propriétés statistiques de suites arithmétiques**, Collection SUP, Presses Universitaires de France.

S. C. REDDY, P. J. SCHMID, ET D. S. HENNINGSON, (1993), *Pseudospectra of the Orr-Sommerfeld operator*, SIAM J. Appl. Math., **53**, 15–47.

S. C. REDDY, L. N. TREFETHEN, ET D. PATHRIA, (1993), *Pseudospectra of the convection-diffusion operator*, Technical Report CTC93 TR126, Cornell Theory Center.

J. R. RICE, (1966), *A theory of condition*, SIAM J. Numer. Anal., **3**, 2, 287–310.

J. RIGAL ET J. GACHES, (1967), *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., **14**, 3, 543–548.

J. ROHN, (1990), *Non singularity under data rounding*, Lin. Alg. and its Appl., **139**, 171–174.

F. SCHLESINGER, (1917), *On the errors in a sum of a number of tabular quantities*, Astron. J., **30**, 183–190.

P. J. SCHMID ET D. S. HENNINGSON, (1992), Phys. Fluids A, **4**.

M. SHUB ET S. SMALE, (1992), *Complexity of Bezout's theorem I. Geometric aspects*, in **Proceedings of World Congress of Nonlinear Analysis**, W. de Gruyter, ed., Berlin. To appear.

R. D. SKEEL, (1979), *Scaling for numerical stability in Gaussian Elimination*, J. Assoc. Comput. Mach., **26**, 3, 494–526.

R. D. SKEEL, (1980), *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., **35**, 817–832.

P. H. STERBENZ, (1974), **Floating-point computation**, Prentice Hall, Englewood Cliffs.

I. STEWART, (1993), *1 est plus probable que 9*, Pour la science, **190**, 94–98.

F. Stummel, (1980), *Rounding error analysis of elementary numerical algorithms*, Computing Suppl., **2**, 169–195.

L. N. Trefethen et M. Trummer, (1987), *An instability phenomenon in spectral methods*, SIAM J. Numer. Anal., **24**, 1008–1023.

L. N. Trefethen, A. E. Trefethen, S. C. Reddy, et T. A. Driscoll, (1993), *Hydrodynamics stability without eigenvalues*, Science, **261**, 578–584.

L. N. Trefethen, (1990), *Approximation theory and numerical linear algebra*, in **Algorithms for Approximation II**, J. C. Mason et e. M. G. Cox, eds., London, Chapman and Hall, 336–359.

L. N. Trefethen, (1992), *Pseudospectra of matrices*, in **Numerical Analysis 1991**, D. F. Griffiths et G. A. Watson, eds., Longman, Harlow, London.

L. N. Trefethen, (Book in preparation), *Spectra and pseudospectra : the behavior of nonnormal matrices and operators*, Department of Computer Science, Cornell University.

A. M. Turing, (1948), *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., **1**, 287–308.

P. R. Turner, (1982), *The distribution of leading significant digits*, IMA J. Numer. Anal., **2**, 407–412.

A. van der Sluis, (1970), *Domain of uncertainty for perturbed operator equations*, Comp., **5**, 312–323.

A. van Wijngaarden, (1953), *Erreurs d'arrondiment dans les calculs systématiques*, in **Les machines à calculer et la pensée humaine**, Colloque CNRS 37, Paris.

J. Vignes, (1978), *New methods for evaluating the validity of the results of mathematical computations*, Math. Compt. Simul. IMACS, **20**, 4, 227–249.

V. Voïévodine, (1980), **Principes numériques d'algèbre linéaire**, Editions Mir, Moscou.

R. von Mises, (1936), *Les lois de probabilités pour les fonctions statistiques*, Ann. Inst. H. Poincaré, **6**, 185–212.

J. von Neumann et H. H. Goldstine, (1947), *Numerical inverting of matrices of high order I*, Proc. Amer. Math. Soc., **53**, 1021–1099.

J. H. Wilkinson, (1961), *Error Analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., **8**, 3, 281–330.

J. H. Wilkinson, (1963), **Rounding errors in algebraic processes**, vol. 32, Her Majesty's stationery office, London.