Fiabilité des calculs sur ordinateur

Jocelyne Erhel
INRIA-IRISA – PROJET ALADIN

Quelques exemples

Le vol manqué d'Ariane 501

voir http://www.esa.int/tidc/Press/Press96/ariane5rep.html
le premier vol du nouveau lanceur Ariane 5 a eu lieu le 4 juin 1996. Après 30 secondes de vol, le lanceur, alors à une altitude de 3700 m, a soudainement basculé, quitté sa trajectoire, s'est brisé et a explosé. L'échec était dû à la perte totale des informations de guidage et d'attitude, 37 secondes après la mise à feu du moteur Vulcain.

Le Système de Référence Inertiel a calculé une accélération horizontale beaucoup plus grande pour Ariane 5 que pour Ariane 4. Cette valeur flottante sur 64 bits n'a pu être convertie en un entier sur 16 bits, d'où la même erreur "Operand range error" dans les deux calculateurs.

Comble d'ironie, ce calcul était inutile pour Ariane 5.

Tous les logiciels ont été soigneusement vérifiés avant le lancement d'Ariane 502.

L'ordinateur et les résultats d'examen

voir http://catless.ncl.ac.uk/Risks/
Forum On Risks To The Public In Computers And Related Systems

L'histoire se passe en Australie et Nouvelle-Zélande, en 1995.
L'examen des anesthésistes a 3 parties : écrit, mémoires, oral, chacune avec plusieurs épreuves et avec divers coefficients.
Après publication des résultats, 3 candidats ont échoué.
Puis ils sont rappelés, pour leur annoncer qu'ils sont finalement reçus.

Le système informatique a changé.
Le mode d'arrondi est l'arrondi par défaut.
Le nombre de mémoires est passé de 3 à 10, avec un coefficient global inchangé de 30%.

Une erreur d'arrondi, qui ne se produisait pas avant (pas de division) a basculé la moyenne des 3 candidats du mauvais côté de la barre.

L'ordinateur et les résultats d'élections

L'histoire se passe en Allemagne, en 1996.
Le scrutin est mixte, direct et proportionnel par listes, avec une clause de 5%.
Après publication des résultats, les Verts ont un siège (5% des voix).

Le lendemain, les Verts n'ont plus de siège et le SPD a un siège supplémentaire, ce qui lui vaut la majorité à une voix au parlement.

Le résultat exact est 4,97%, donc la clause s'applique et les Verts n'ont aucun siège. Ce siège est redistribué par listes et se trouve attribué au SPD.

Arithmétique flottante

Arithmétique virgule flottante

Le format flottant est défini par

- la base b
- le nombre de chiffres de la mantisse p
- la plage d'exposants $E_{min} \ldots E_{max}$

Un nombre flottant est défini par

$$x = (-1)^s b^e a_0.a_1 a_2 \ldots a_p$$

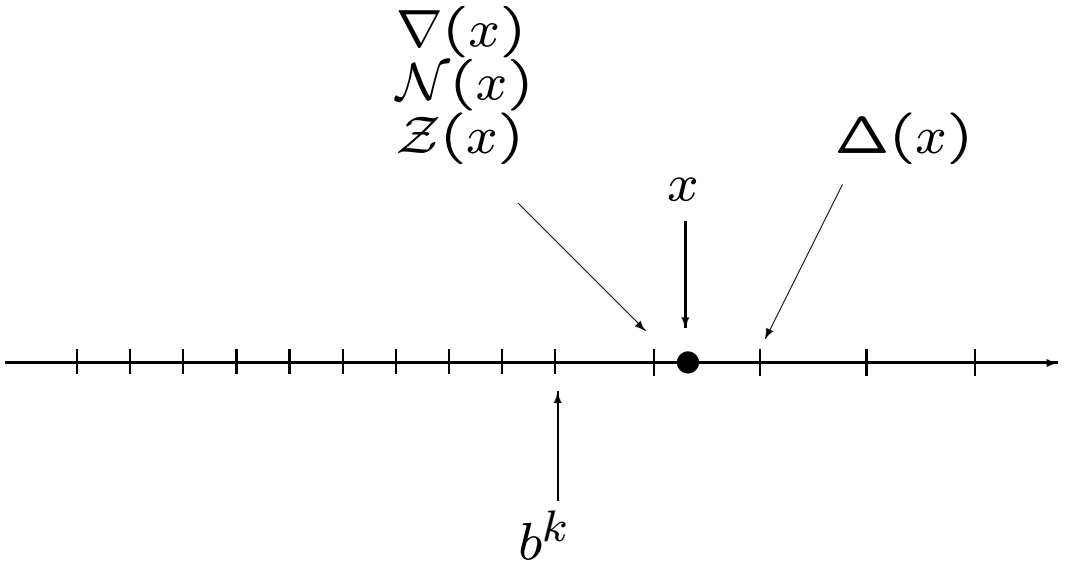
avec $a_0 \neq 0$ (écriture normalisée)

L'écart minimal entre deux mantisses est $\epsilon = b^{-p}$

Modes d'arrondi

Figure due à J-M. Muller

Tout réel est encadré par 2 nombres flottants :



4 modes d'arrondi possibles

Modes d'arrondi - propriétés

$\mathcal{N}(x)$ est l'arrondi au plus près et $\mathcal{Z}(x)$ est l'arrondi vers 0

$\nabla(x)$ est l'arrondi vers $-\infty$ et $\Delta(x)$ est l'arrondi vers $+\infty$

monotonie : x,y réels, $x \leq y \Rightarrow \mathcal{N}(x) \leq \mathcal{N}(y)$

projection : x flottant $\Rightarrow \mathcal{N}(x) = x$

précision : $|\mathcal{N}(x) - x| \leq \epsilon/2|x|$

remarque : vrai pour les autres modes d'arrondi, avec ϵ au lieu de $\epsilon/2$

Norme IEEE-754 - définition

base $b = 2$

format court : $p = 23$ et $[E_{min} : E_{max}] = [-126 : +127]$

format long : $p = 52$ et $[E_{min} : E_{max}] = [-1022 : +1023]$

1er chiffre implicite

Mode d'arrondi actif (au choix parmi 4) (en principe)

Opérations arithmétiques correctes : le résultat flottant d'une opération entre flottants est l'arrondi du résultat exact.

Exceptions gérées par des nombres spéciaux : $\pm\infty$ et NaN

Norme IEEE-754 - caractéristiques

format court : $\epsilon \simeq 10^{-6}$. Environ 7 chiffres significatifs

format long : $\epsilon \simeq 10^{-15}$. Environ 16 chiffres significatifs

format court : nombres flottants entre environ 10^{-38} et 10^{+38}

format long : nombres flottants entre environ 10^{-308} et 10^{+308}

plus petit : underflow et mise à ± 0

plus grand : overflow et mise à $\pm\infty$

impossible : inexact et mise à NaN

Intérêt de la norme IEEE-754

- Même résultat d'une machine à l'autre (en principe)
- Preuve de stabilité numérique des algorithmes
- Construction d'algorithmes précis
- arithmétique d'intervalles

Mais pas de spécification pour les fonctions élémentaires

Calcul de la limite d'une suite

Exemple dû à J-M. Muller

On calcule la limite de la suite suivante

$$\begin{cases} x_0 & = & 1{,}510005072136258 \\ x_{n+1} & = & \dfrac{3x_n^4 - 20x_n^3 + 35x_n^2 - 24}{4x_n^3 - 30x_n^2 + 70x_n - 50} \end{cases}$$

Le résultat obtenu sera, suivant le nombre de chiffres sur lesquels sont effectués les calculs intermédiaires :

Nombre de chiffres	Résultat obtenu
10	4,000000033
12	1,00000000000
14	3,0000000000000
16	4,000000000000033
18	1,0000000000000000

Et le résultat correct est 1.

Calcul formel - précision variable

Le système de calcul formel Maple ne respecte pas la norme IEEE.

Calculer avec une plus grande précision ne garantit pas le résultat.

Absorption

Sur ordinateur, en simple ou double précision, $\lim_{n\to+\infty} \sum_{i=1}^{n} \frac{1}{i} = l < \infty$

Pourtant, on apprend en maths que cette série est infinie.

Exemple en base 10, avec 3 chiffres de mantisse après la virgule :

$$\mathcal{N}(10^5 + 10) = \mathcal{N}(10^5(1 + 10^{-4})) = 10^5$$

Dès que y est beaucoup plus petit que x,
alors $x + y$ est arrondi à x.

Un petit nombre est absorbé par un grand nombre.

Cancellation catastrophique

Le problème survient lorsque le résultat d'une soustraction est petit relativement aux opérandes.

L'ordre de grandeur du résultat est le même que l'ordre de grandeur des erreurs sur les opérandes.

La soustraction est exacte mais fait ressortir les erreurs précédentes.

La cancellation amplifie les erreurs de calcul.

L'ordinateur et les factures d'électricité

L'histoire se passe aux Etats-Unis, en 1996.
La facture d'électricité mensuelle comporte 2 valeurs :
la consommation moyenne d'énergie par jour (en KW-h),
l'écart relatif par rapport au même mois de l'année précédente.
Un client voit sur la facture de février un écart de 9 %.
Il refait le calcul et trouve en fait un écart de 4%.

Sur la facture de février 1995, la consommation moyenne est 11KW-h.
Sur la facture de février 1996, la consommation moyenne est 12KW-h.
Ce qui fait un écart de 9%.

Mais, avant arrondi, les nombres sont
11,21KW-h en 1995,
11,68KW-h en 1996.
Ce qui fait un écart de 4%.

Il s'est produit un phénomène de cancellation.

Un peu de formalisation

Conditionnement d'un problème

problème d'évaluation $x = F(a)$ ou problème de résolution $F(x,a) = 0$

étude de la sensibilité aux données: théorie de la perturbation :

étude de Δx en fonction de Δa

Le problème est bien posé si x est unique et est une fonction continue de a

Le problème est stable si $\|\Delta x\|/\|\Delta a\|$ est borné

Le conditionnement C est défini par

$$C = \limsup_{\|\Delta a\| \to 0} \|\Delta x\|/\|\Delta a\|$$

Attracteur de Lorenz

En 1961, Lorenz découvre l'effet papillon
sur un modèle simplifié de météorologie

La solution est très sensible aux conditions initiales :
le problème est instable

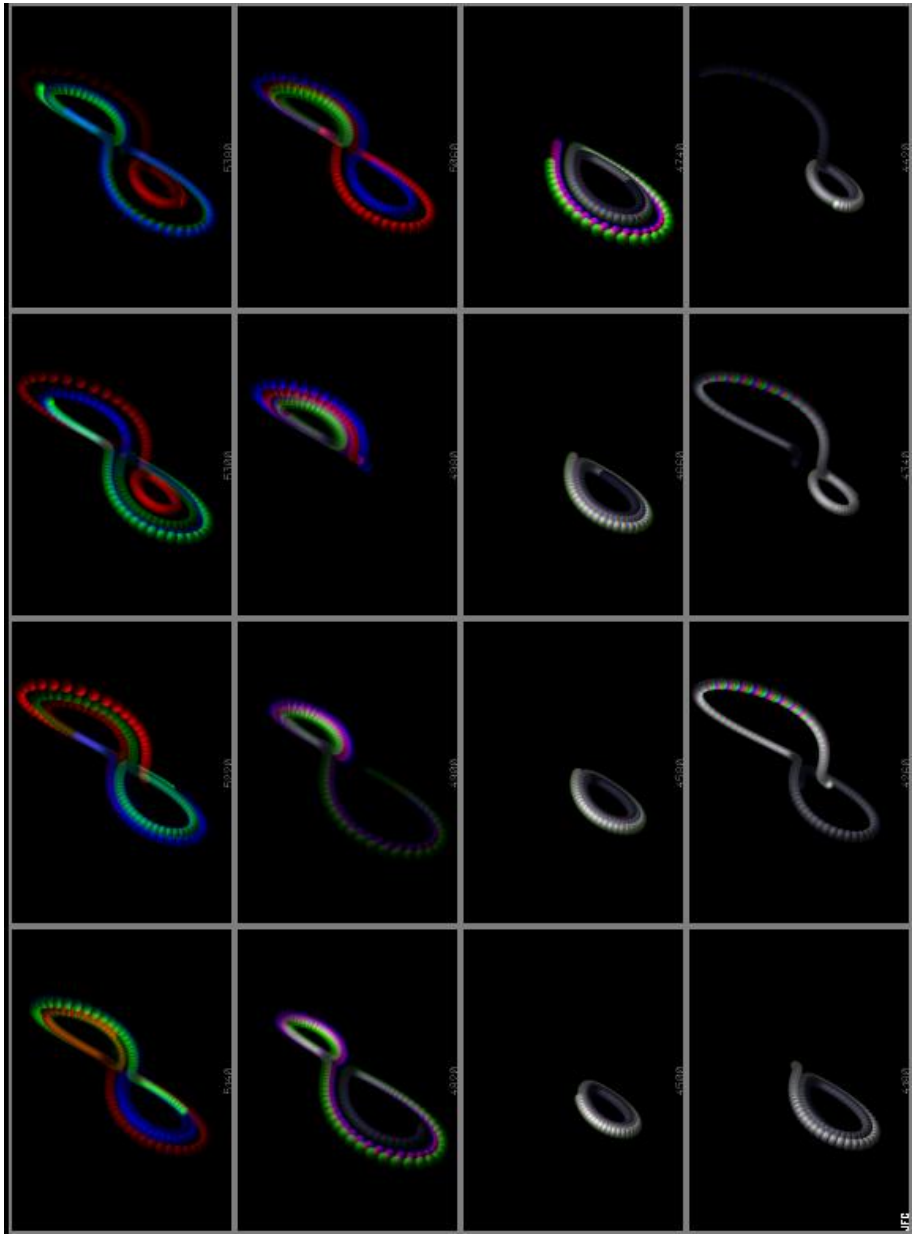
Cela se traduit par une très forte sensibilité aux erreurs d'arrondi

Attracteur de Lorenz

voir http://blanche.polytechnique.fr/lactamme/,
J-F. Colonna, Ecole Polytechnique

$$\left\{ \begin{array}{rcl} x'(t) & = & -10x + 10y \\ y'(t) & = & 28x - y - xz \\ z'(t) & = & -\frac{8}{3}z + xy \end{array} \right. \qquad (1)$$

Calcul de l'attracteur de Lorenz sur 3 ordinateurs différents
par une méthode numérique de Runge-Kutta d'ordre 4



Calcul de trajectoire

Exemple dû à J-M. Muller

On calcule une trajectoire $x(t)$ avec une vitesse $x'(t)$ qui vérifie

$$\begin{cases} x'(t) &= 10(x(t) - t^2) \\ x(0) &= 0{,}02 \end{cases} \qquad (2)$$

On résout par une méthode numérique de Runge-Kutta d'ordre 4 :

valeur	calcul sur 7 chiffres	calcul sur 15 chiffres	résultat exact
$x(1)$	$-8{,}175$	$1{,}219999906$	$1{,}22$
$x(2)$	-206935	$4{,}417920697$	$4{,}42$
$x(3)$	$-4{,}558 \times 10^9$	$-36{,}1797$	$9{,}62$
$x(4)$	$-1{,}003 \times 10^{14}$	$-1{,}008 \times 10^7$	$16{,}82$
$x(5)$	$-2{,}211 \times 10^{18}$	$-2{,}222 \times 10^{11}$	$26{,}02$

Calcul de trajectoire - conditionnement

Exemple dû à J-M. Muller

La solution exacte est $x(t) = t^2 + \frac{1}{5}t + \frac{1}{50} + (x_0 - \frac{1}{50})e^{10t}$

d'où

$$|\Delta x(t)|/|x(t)| = |x_0|(e^{10t}/|x(t)|)\ |\Delta x_0|/|x_0|$$

Le problème du calcul de $x(t)$ est très mal conditionné pour $x_0 = \frac{1}{50}$:

$$C = \frac{1}{50}e^{10t}/|t^2 + \frac{1}{5}t + \frac{1}{50}|$$

Cela se traduit par des erreurs qui augmentent avec t

Méthode d'approximation

pas de calcul ou de résolution direct $\Rightarrow$ approximation

$x_h = F_h(a)$ tel que $\lim_{h \to 0} x_h = x$

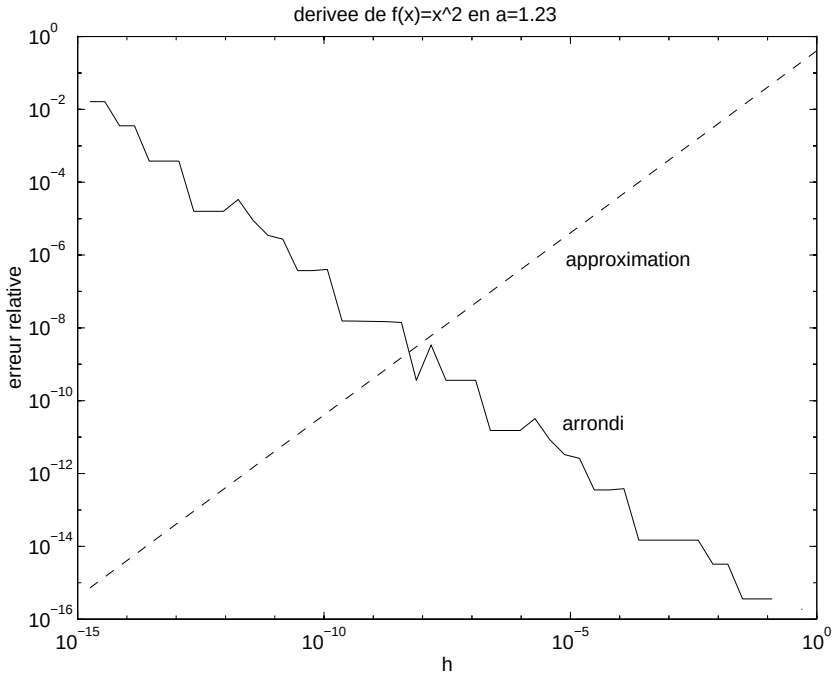
On définit l'ordre de convergence α par

$$\|x_h - x\| = O(h^\alpha)$$

Calcul de dérivée

f fonction dérivable au point a

Approximation par la formule $f'(a) \simeq f_h(a) = \frac{f(a+h)-f(a)}{h}$



Calcul de dérivée

Approximation par la formule décentrée $f'(a) \simeq f_h(a) = \frac{f(a+h)-f(a)}{h}$

Il y a cancellation puis absorption pour h très petit

L'erreur globale est en $O(h) + O(\epsilon/h)$

Il faut choisir $h = O(\epsilon^{1/2})$ et l'erreur est en $O(\epsilon^{1/2})$

Pour la formule centrée $f'(a) \simeq (f(a+h) - f(a-h))/2h$,
l'erreur globale est en $O(h^2) + O(\epsilon/h)$

Il faut choisir $h = O(\epsilon^{1/3})$ et l'erreur est en $O(\epsilon^{2/3})$

Algorithme de résolution

L'algorithme définit l'ordre des opérations pour le calcul de x (ou x_h)

A cause des arrondis, le résultat du calcul est x_ϵ différent de x

L'algorithme est inversement stable si x_ϵ est solution d'un problème perturbé

$$x_\epsilon = F(a_\epsilon) \ et \ \|a_\epsilon - a\| = O(\epsilon)$$

Si le problème F a un conditionnement C alors

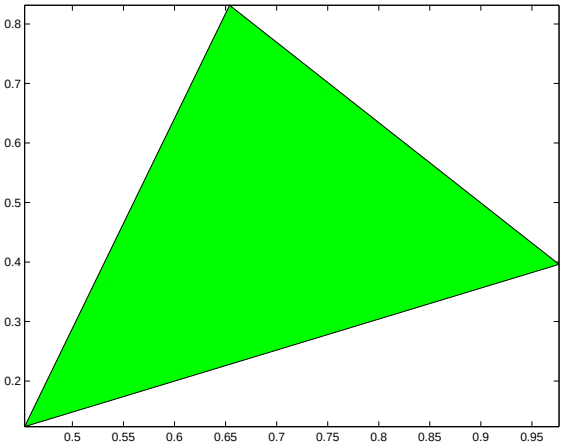
$$\|x_\epsilon - x\| \simeq C\epsilon$$

Calcul de l'aire d'un triangle

On considère un triangle défini par les 3 sommets de coordonnées $(x_1,y_1),(x_2,y_2),(x_3,y_3)$

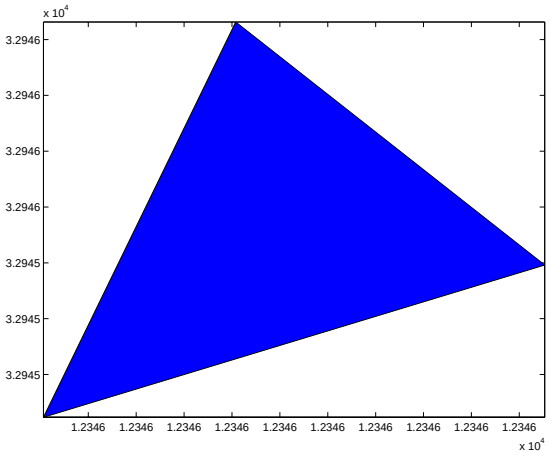
On calcule l'aire avec la formule du déterminant, en double précision

$$A = \pm\frac{1}{2} \times \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix}$$



aire calculée = 0.157876**34025000**

translation d'environ $(10^4,10^4)$



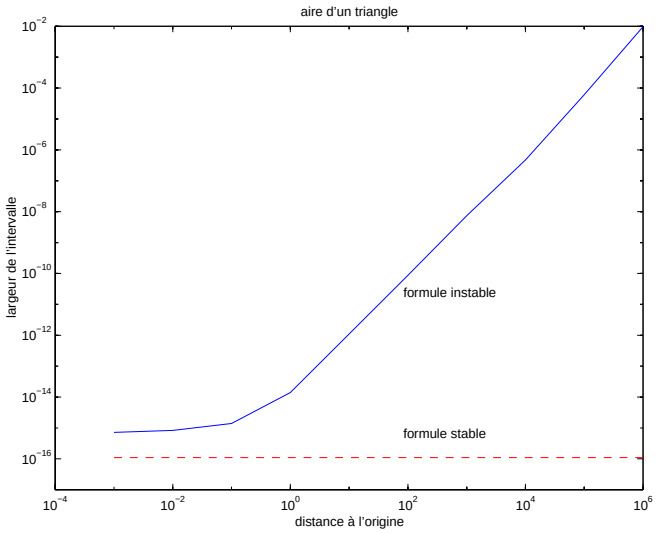
aire calculée = 0.157876**28293037**

Aire d'un triangle - calcul

$$A = \pm\frac{1}{2} \times \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix}$$

calcul en

arithmétique d'intervalles



Formule instable : cancellations

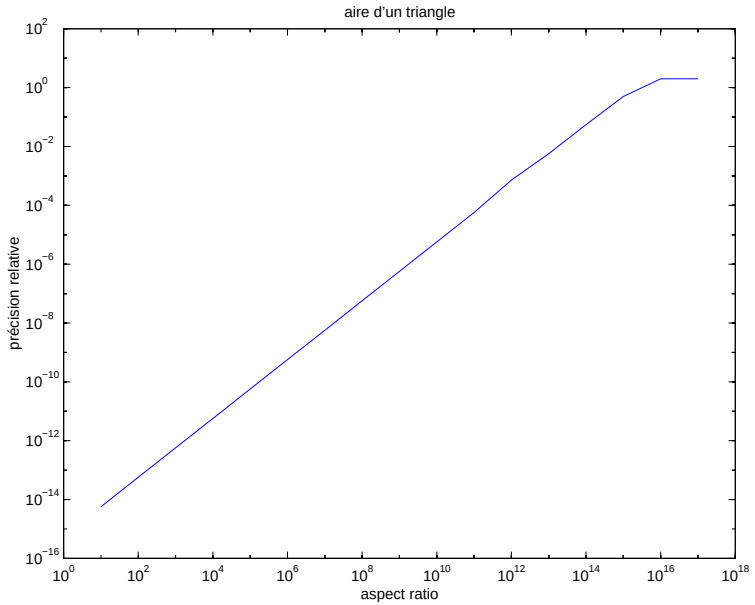
$$A = 1/2 \times |x_1y_2 + x_2y_3 + x_3y_1 - x_1y_3 - x_2y_1 - x_3y_2|$$

Formule stable : erreur en $O(\epsilon)$

$$A = \frac{1}{2} \times |(x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)|$$

Aire d'un triangle - conditionnement

L'aire d'un triangle de base a et de hauteur h est $A = \frac{1}{2} \times a \times h$
On a $|\Delta A|/A \leq (1 + a/h) \max(|\Delta a|/a, |\Delta h|/a)$
où a/h est l'aspect ratio du triangle



calcul en arithmétique d'intervalles par la formule stable

Le calcul de l'aire d'un triangle aplati est mal conditionné

Un récapitulatif

sensibilité du problème aux données	problème singulier/bien posé conditionnement du problème
méthode d'approximation	ordre de convergence singularité artificielle
algorithme de résolution	sensibilité aux arrondis stabilité numérique
arithmétique flottante	cancellation, absorption, exceptions

Quand tout va bien

- résoudre un problème bien conditionné
- choisir une méthode précise (h petit mais pas trop)
- choisir un algorithme stable
- choisir une arithmétique précise (ϵ petit)

Si tout est réuni, l'erreur finale est en $O(h^\alpha + C\epsilon/h)$

Quelques outils de validation numérique

Bibliothèques de calcul scientifique

bibliothèque Lapack "Linear Algebra package"

* tous les algorithmes de base en algèbre linéaire

* prise en compte de l'arithmétique flottante

* algorithmes stables et fiables

* estimation de conditionnement

* optimisation de la vitesse d'exécution

bibliothèque Nag

* résolution de problèmes non linéaires, etc

* utilise Lapack

logiciels spécialisés en équations différentielles, etc

Utilisation de résidus

Le résidu du problème $F(x,a) = 0$ est $F(x_\epsilon,a)$

Exemple : problème $F(x) = a$ alors $F(x_\epsilon) = a + (F(x_\epsilon) - a)$

donc $\|x_\epsilon - x\| \leq C\|F(x_\epsilon) - a\|$

L'erreur sur la solution est le résidu multiplié par le conditionnement

Dans la mesure du possible, utiliser des résidus avec un sens physique

Estimation de conditionnement

Formule mathématique du conditionnement

exemple : aire du triangle et aspect ratio

Algorithme d'approximation du conditionnement

exemple : système linéaire

Analyse statistique de sensibilité aux données

Analyse statistique - principe

Pour un problème bien conditionné et un algorithme stable

$$\|\Delta x\| \simeq C\|\Delta a\|^q$$

La méthode consiste à perturber aléatoirement les données a

avec une amplitude croissante $\alpha = \|\Delta a\|$

et à estimer statistiquement l'erreur $\|\Delta x\|$

par $\max_{k=1,\ldots,ne} \|x_k - x_\epsilon\|$

où x_k est calculé avec a_k tel que $\|a_k - a\| \leq \alpha$

Voir articles de J. Erhel, Aladin-Irisa

Analyse statistique - méthode

3 zones d'amplitude de perturbations

- $\alpha \leq \alpha_1$: erreurs d'arrondi : $\|\Delta x\|$ est constant
- intervalle $[\alpha_1,\alpha_2]$: perturbations : $\max_{k=1,\ldots,ne} \|x_k - x_\epsilon\| \simeq C\alpha^q$
- $\alpha \geq \alpha_2$: erreur très grande ou autre régularité

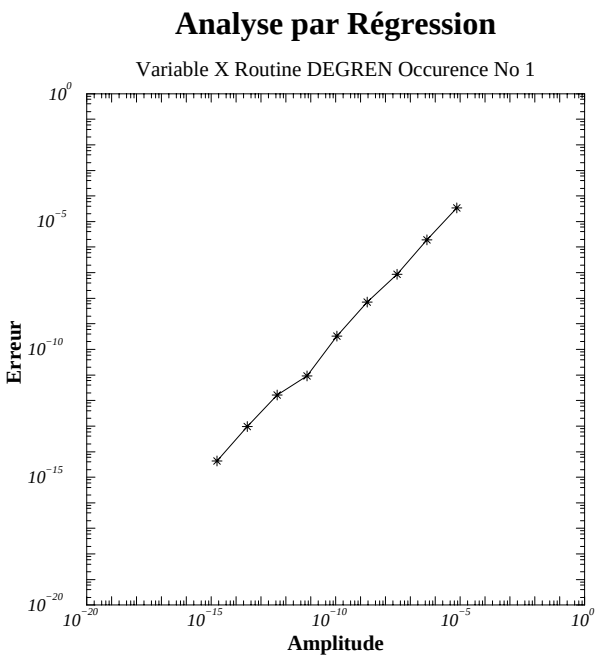
mise en oeuvre

- visualiser l'erreur en fonction de l'amplitude
- déterminer un intervalle $[\alpha_1,\alpha_2]$
- appliquer une régression log-linéaire
  pour estimer le conditionnement C et la régularité q

Exemple d'analyse statistique : racine de polynôme

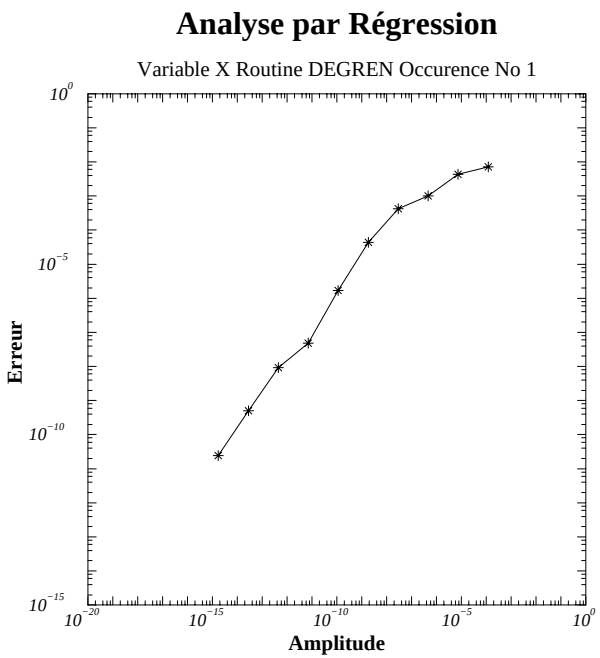
$$P(X) = (X^{n-2} + X^{n-1} + \ldots + X + 1)(X - a)(X - b)$$

$n = 10, a = 3, b = 2$



conditionnement $\simeq 3$

$n = 10, a = 2.0001, b = 2$



conditionnement $\simeq 10^4$

Encadrement de résultat

Arithmétique d'intervalles avec les arrondis dirigés

Encadrement garanti du résultat :
le résultat exact est dans l'intervalle

Mais les intervalles enflent à cause de
la décorrélation et de l'effet enveloppant

Algorithmes spécifiques pour éviter ces deux effets

Procédures d'encadrement

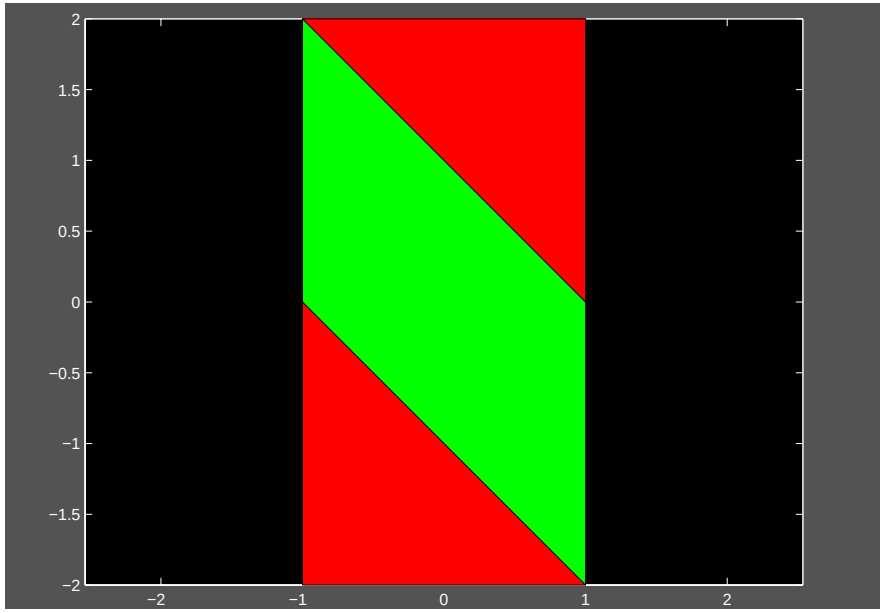
voir thèse d'Olivier Beaumont, Aladin-Irisa, 1999.

Comment grandissent les intervalles

Le découplage des données entraîne une surestimation de l'intervalle résultat

Exemple : $P(x) = x - x^2$
par intervalles : $P([0,1]) \subseteq [0,1]$
réellement : $P([0,1]) = [0,1/4]$

Le Wrapping effect enfle l'intervalle résultat par inclusion d'un connexe de $\mathbb{R}^n$ dans un produit d'intervalles



Comment réduire les intervalles : utilisation de simplexes

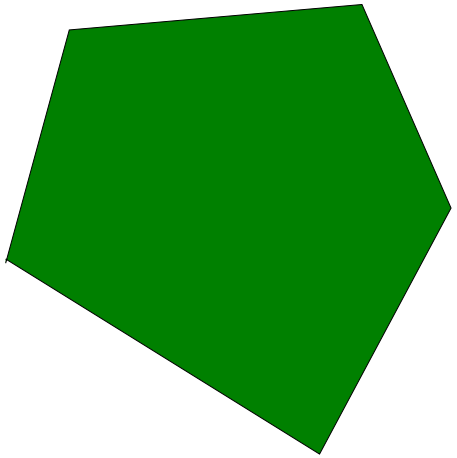
L'objectif est d'éviter le découplage

Exemple : évaluation d'un polynôme $[P](X) = \sum_{i=0}^{n} [a_i] X^i$ pour $X \in I$

On transforme le problème en l'évaluation de
$[f](x) = [f](x_1, \ldots, x_n) = \sum_{i=0}^{n} [a_i] x_i$

On corrèle les variables x_i par des contraintes linéaires à d'aide de polynômes de Chebyshev qui définissent un polyhèdre convexe S

On résout $\min_{x \in S} [f](x)$ et $\max_{x \in S} [f](x)$ par l'algorithme du simplexe pour déterminer un encadrement de $[P](I)$



Procédures d'encadrement

Le principe général est de calculer une approximation

puis de raffiner la solution en résolvant un problème de point fixe

Exemple : $f(x) = 0$

approximation x_0 et matrice $C \simeq J(x_*)^{-1}$

itérations $x_{i+1} - x_0 = x_0 - Cf(x_0) + (I - CJ(x_i))(x_i - x_0)$

opérateur de Krawczyk $K([y],x_0,C) = x_0 - Cf(x_0) + (I - CJ(x_0 + [y]))[y]$

Si $K([y],x_0,C) \subset [y]$ alors il existe une solution $x \in x_0 + [y]$

Quelques références bibliographiques

* Lecture on Finite Precision Computations

F. Chatelin et V. Frayssé, SIAM, 1995

* Qualité des calculs sur ordinateur *vers des arithmétiques plus fiables ?*

Coordonné par M. Daumas et J-M. Muller, Masson, 1997

* Accuracy and Stability of Numerical Algorithms

N. Higham, SIAM, 1995

* Arithmétique des ordinateurs

J-M. Muller, Masson, 1989

* La théorie du chaos *vers une nouvelle science*

J. Gleick, Flammarion, 1991