Arithmétique flottante

Conférence INSA - Décembre 2001

Jocelyne Erhel - Projet ALADIN

INRIA/IRISA - Rennes

- Quelques exemples
- Arithmétique flottante
- Stabilité numérique des algorithmes
- Conditionnement des problèmes
- Méthodes d'approximation
- Quelques outils de validation numérique

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

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

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

Nombres flottants

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 $\epsilon = b^{-p}$

Quelques propriétés du système flottant

L'ensemble des flottants est symétrique par rapport à 0
L'ensemble des flottants est fini

Il faut arrondir les nombres réels

Le plus grand nombre flottant est $x_{max} = b^{E_{max}} m_{max}$
Le plus petit nombre flottant > 0 est $u = b^{E_{min}}$

Il faut prévoir les dépassements vers $\pm\infty$ et vers 0

Arrondis - encadrement d'un réel

Exemples

format : $b = 10, p = 3$
$x = 1.23456$ est encadré par $x_1 = 1.234$ et $x_2 = 1.235$
et $x_2 - x_1 = 10^{-3} = \epsilon \leq |x|\epsilon$
le nombre flottant le plus proche est x_2
$x = -765.4321$ est encadré par $x_1 = -10^2 \times 7.655$ et $x_2 = -10^2 \times 7.654$
et $|x_2 - x_1| = 10^2 \times 10^{-3} \leq |x|\epsilon$
le nombre flottant le plus proche est x_2

Cas général

Tout réel (ni trop petit ni trop grand)
x est encadré par 2 nombres flottants x_1 et x_2
avec $|x_2 - x_1| \leq \epsilon |x|$

Modes d'arrondi - propriétés

Propriétés d'un arrondi

fonction des réels (ni trop petits ni trop grands) vers les flottants
monotonie : x,y réels, $x \leq y \Rightarrow \ fl(x) \leq fl(y)$
projection : x flottant $\Rightarrow fl(x) = x$

$fl(x)$ est l'un des flottants qui encadrent x
précision : $\frac{|fl(x)-x|}{|x|} \leq \epsilon$
L'erreur relative d'un arrondi est la précision machine

4 modes d'arrondi

Au plus près, vers 0, vers $-\infty$, vers $+\infty$
Arithmétique d'intervalles avec arrondis vers $\pm\infty$
Précision avec $\epsilon/2$ pour l'arrondi au plus près

Opérations arithmétiques

Exemples

Format : $b = 10$, $p = 3$
$x = 1.234$ et $y = 8.103$ alors $x \times y = 10.528488$
$a = 10 \times 8.532$ et $b = 5.276$ alors $a - b = 80.044$

Il faut arrondir le résultat d'une opération

$fl(x \times y) = 10 \times 1.053$
$fl(a - b) = 10 \times 8.004$

Opération correcte

Le résultat de l'opération entre deux nombres flottants
est l'arrondi du résultat exact

Exceptions

Exception overflow

Le résultat d'une opération est plus grand que x_{max}
exemple : x_{max}^2

Exception underflow

Le résultat d'une opération est plus petit que u
exemple : $u/2$

Exception invalide

Le résultat n'est pas prévisible
exemple : $0/0$

Propriétés des opérations arithmétiques

Propriétés conservées

0 est élément neutre de l'addition
1 est élément neutre de la multiplication
x nombre flottant, $x - x = 0$
l'addition et la multiplication sont commutatives

Propriétés NON conservées

L'addition n'est pas associative
La multiplication n'est pas associative
La multiplication n'est pas distributive par rapport à l'addition
Il existe x nombre flottant, $x \times (1/x) \neq 1$

Dans une boulangerie française le 1 janvier 2002

Sans un seul Euro en poche, mais le porte monnaie plein de FF, un client entre chez sa boulangère préférée lui présenter ses meilleurs voeux et acheter une baguette bien croustillante.

C'est combien la baguette maintenant ?
4,30F comme dab !
Oui mais à partir d'aujourd'hui c'est en Euro.
Ah ! J'oubliais: un petit coup d'EuroCalculette..: et voila ça fait 0,66 Euro.
0,66 Euro (un petit coup d'EuroCalculette) mais ça fait 4,33 FF, ma baguette a augmenté de 3 centimes.
Je vous donne une pièce de 5 F et vous me rendez la monnaie en Euros.
Bien, 5 F ça fait (un petit coup d'EuroCalculette) 0,76 Euro moins 0,66 Euro la baguette, je vous rends 10 centimes d'Euro.
Chouette, ma première pièce en Euro. Bon, (un petit coup d'EuroCalculette) 0,10 Euros, cela fait 0,66F, tiens c'est comme le prix de la baguette en Euro. Mais attendez : 5F moins 0,66F ça met la baguette à 4,34 F.
Tout compte fait, je me demande si je ne devrais pas prendre 2 baguettes. Voyons 4,30 x 2 = 8,60F, soit (un petit coup d'EuroCalculette) 1,31 Euro. Tiens, 2 baguettes coûtent moins cher que 2 fois une baguette, c'est transcendant ce truc.
Et si je paie avec une pièce de 10F : 1,52 Euro moins 1,31 Euro les 2 baguettes = 0,21 Euro donc 1,38 F ce qui me met la baguette à (10-1,38)/2 = 8,62F/2 soit 4,31F.

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

$\epsilon \simeq 10^{-7}$. Environ 7 chiffres significatifs
nombres flottants positifs entre environ 10^{-38} et 10^{+38}

format long

$\epsilon \simeq 10^{-16}$. Environ 16 chiffres significatifs
nombres flottants positifs entre environ 10^{-308} et 10^{+308}

Exceptions

Type	Résultat flottant
underflow	± 0
overflow	$\pm\infty$
invalid	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

Des étudiants doivent calculer la limite de la suite suivante

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

Ils utilisent le logiciel de calcul formel Maple
Un premier étudiant fait l'évaluation avec 14 chiffres, il trouve 3,00000000000
Il annonce que la suite converge vers 3
Un deuxième étudiant fait l'évaluation avec 10 chiffres, il trouve 4,000000033
Il vérifie avec 16 chiffres et trouve 4,0000000000000033
Il annonce assez sûr de lui que la suite converge vers 4
Un troisième étudiant fait l'évaluation avec 12 chiffres, il trouve 1,000000000
Il vérifie avec 18 chiffres et trouve 1,00000000000000000
Il annonce assez sûr de lui que la suite converge vers 1

Quel est le résultat correct?

Calcul formel et précision variable

Deux observations importantes

Le système de calcul formel Maple (version V, jusqu'en 1999)
ne respecte pas la norme IEEE.

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

Une histoire de séries

Théorème vérifié sur ordinateur en calcul flottant

Si $u_n \geq 0$ et si $\lim_{n\to+\infty} u_n = 0$, alors la série de terme général u_n est convergente, ie $\lim_{n\to+\infty} \sum_{i=1}^{n} u_i < \infty$

Théorème appris en mathématiques

$\lim_{n\to+\infty} \sum_{i=1}^{n} \frac{1}{i} = \infty$

Quel est le théorème le plus sûr ?

Phénomène d'absorption

Exemple

en base 10, avec 3 chiffres de mantisse après la virgule
$fl(10^5 + 10) = fl(10^5(1 + 10^{-4})) = 10^5$

Absorption

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 (1/2)

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

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%.

L'ordinateur et les factures d'électricité (2/2)

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

Stabilité numérique des algorithmes

Algorithme de calcul

Un algorithme définit l'ordre des opérations
pour le calcul de $x = F(a)$, avec a donné

Les calculs sont faits avec une arithmétique flottante, précision ϵ

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

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

Une histoire d'extra-terrestre

Un être extra-terrestre décide de créer un clone
qui voyagera dans l'espace intersidéral

Lui-même et le clone disposent d'un calculateur de bord
qui émet un code confidentiel
afin de communiquer en toute sécurité

Tout se passe bien au début du voyage
Puis la communication se perd aux confins de l'univers

Voir démo sous Matlab

Une histoire d'extra-terrestre (suite)

L'extra-terrestre va voir un ami informaticien et mathématicien
Celui-ci lui bricole un nouveau système de code

L'extra-terrestre renouvelle l'expérience avec deux clones

Il découvre les limites de l'univers

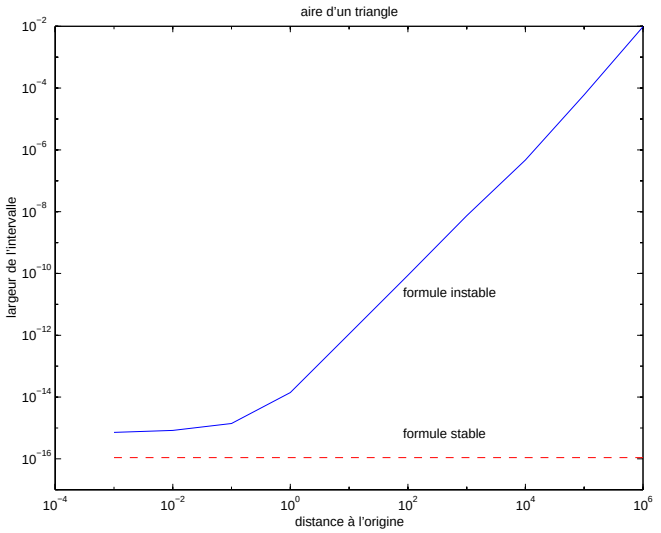
Voir démo sous Matlab

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

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

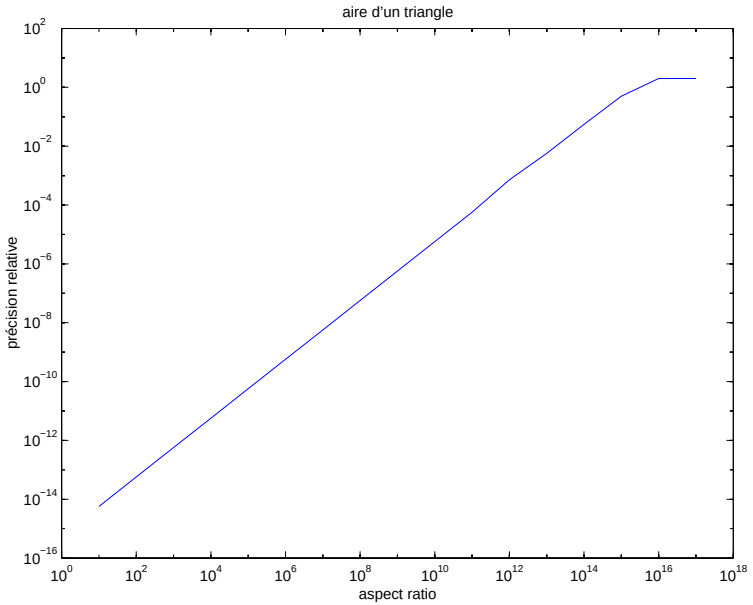
Formule stable : erreur en $O(\epsilon)$

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

Conditionnement des problèmes

Aire d'un triangle aplati

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

Voir démo sous Matlab

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

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

Stabilité numérique et conditionnement

Si l'algorithme est inversement stable
et si le problème a un conditionnement C, alors

$$\|x_\epsilon - x\| = O(C \times \epsilon)$$

Exemple

En simple précision (7 chiffres)
problème avec un conditionnement égal à 10^4
algorithme inversement stable
résultat avec environ 3 chiffres exacts

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

Voir démo sous Matlab

Une équation différentielle

Problème à résoudre

$$\begin{cases} y''(t) = \frac{2}{1-t}y'(t) - \frac{a^2}{(1-t)^4}y(t) & t \in [0,T] \\ y(0) = \sin a \\ y'(0) = a\cos a \end{cases}$$

avec $T = 0.995$, $a = \pi/3$

Résultats avec Matlab

méthode	t=0.1	t=0.8	t=0.9
ode23	0.91822917399541	-0.86589790979113	-0.86724737254210
ode45	0.91821635284222	-0.86616935489750	-0.86738032689247

Comment interpréter ces résultats?

Une équation différentielle (suite)

Résultats avec Matlab et résultat exact

méthode	t=0.1	t=0.8	t=0.9
ode23	0.91822917399541	-0.86589790979113	-0.86724737254210
ode45	0.91821635284222	-0.86616935489750	-0.86738032689247
exact	0.91821610688027	-0.86602540378444	-0.86602540378444

Conditionnement du problème

$$C_a(t) = \frac{a}{1-t}, \quad \lim_{t \to 1} C_a(t) = \infty$$

Conclusions

La méthode ode45 est plus précise que ode23, (voir doc Matlab)
La précision de la solution se dégrade lorsque t tend vers 1
parce que le problème est de plus en plus mal conditionné

Méthodes d'approximation

Convergence et ordre des méthodes 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

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

Approximation de dérivée (1/3)

f fonction dérivable au point a

Approximation par la formule

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

Voir démo sous Matlab

Nouveau théorème déduit de calculs flottants

$$\forall f, \forall a, f'(a) = 0$$

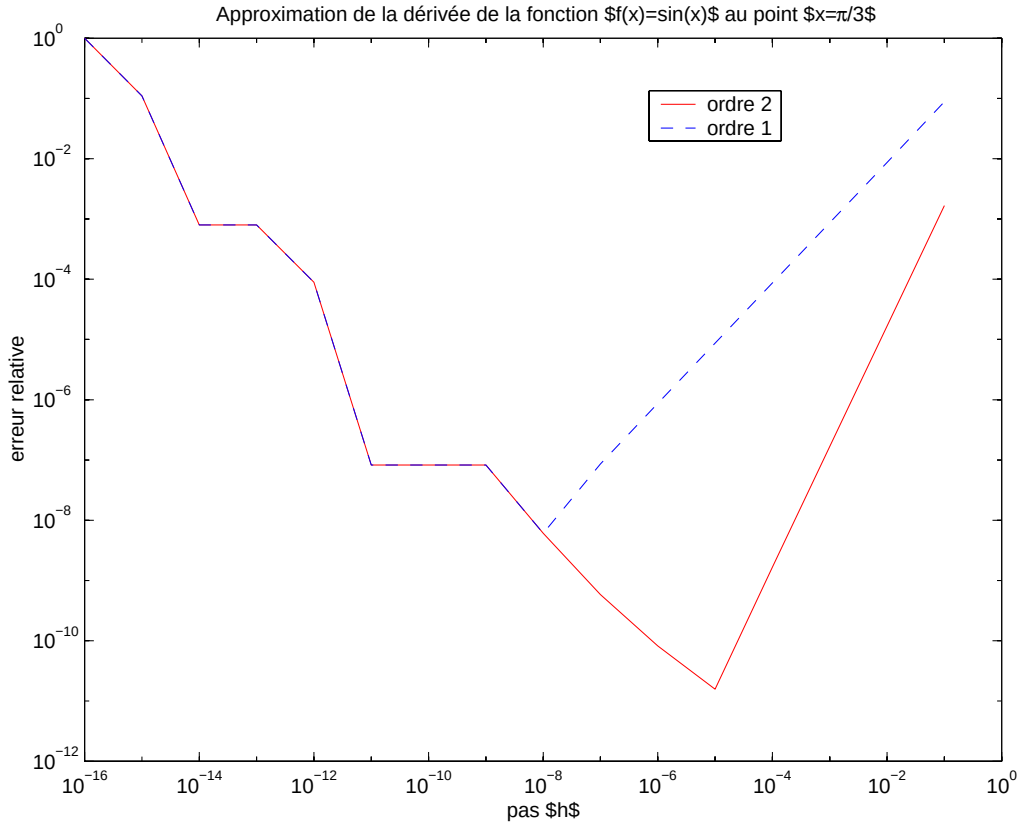
Approximation de dérivée (2/3)

formule décentrée d'ordre 1

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

formule centrée d'ordre 2

$$f'(a) \simeq (f(a+h) - f(a-h))/2h$$



Approximation de dérivée (3/3)

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 $O(\epsilon^{1/2})$

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 $O(\epsilon^{2/3})$

Un récapitulatif

Les calculs en arithmétique flottante induisent des erreurs d'arrondi
Des algorithmes de calcul mathématiquement équivalents
ne sont pas numériquement équivalents

Un algorithme inversement stable résout un problème
avec des données perturbées
Le conditionnement d'un problème mesure sa
sensibilité aux perturbations des données

Erreur en $O(C \times \epsilon)$

Une méthode d'approximation est définie par un paramètre h
La précision est mesurée par son ordre
Pour h très petit, les erreurs d'arrondi sont plus grandes
que l'erreur d'approximation

paramètre optimal avant explosion des erreurs d'arrondi

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

Bibliothèques spécialisées

* en équations différentielles, etc
contrôle d'erreur - ordre variable et pas variable
contrôle d'invariants

Système numérique Matlab

* utilise Lapack
utilise des bibliothèques spécialisées

Utilisation de résidus

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

Exemple

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

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

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 dans LAPACK

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