Stabilité numérique et conditionnement

Jocelyne Erhel

équipe FLUMINANCE, Inria Rennes et IRMAR, France

INSA, Mai 2016

Plan

- Introduction
- Stabilité numérique
- Conditionnement
- Méthodes d'approximation

Algorithme de calcul

Evaluation de fonction

Il s'agit de calculer de $x = F(a)$, avec a donné.

Résolution d'équations

Il s'agit de résoudre $F(x, a) = 0$, avec a donné.

Algorithme

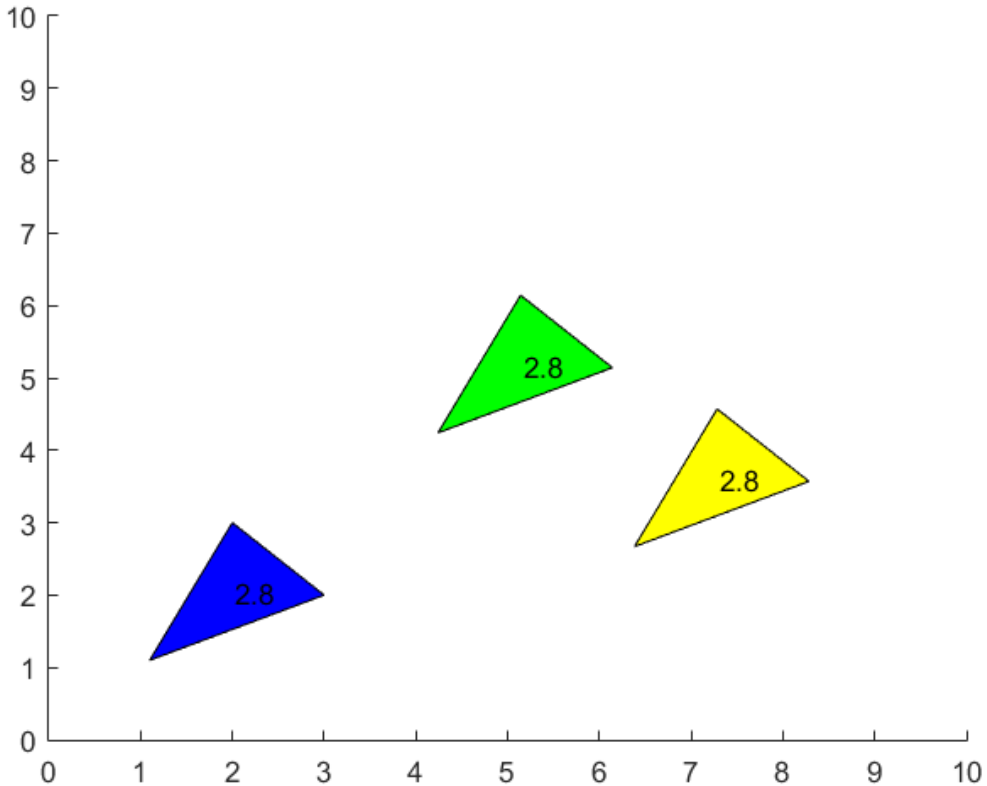
Un algorithme définit les opérations à effectuer.
Un algorithme direct calcule x en un nombre fini d'opérations.
Un algorithme itératif calcule des approximations x_n de x de façon à converger: $\lim_{n\to\infty} x_n = x$

Calculs sur ordinateur

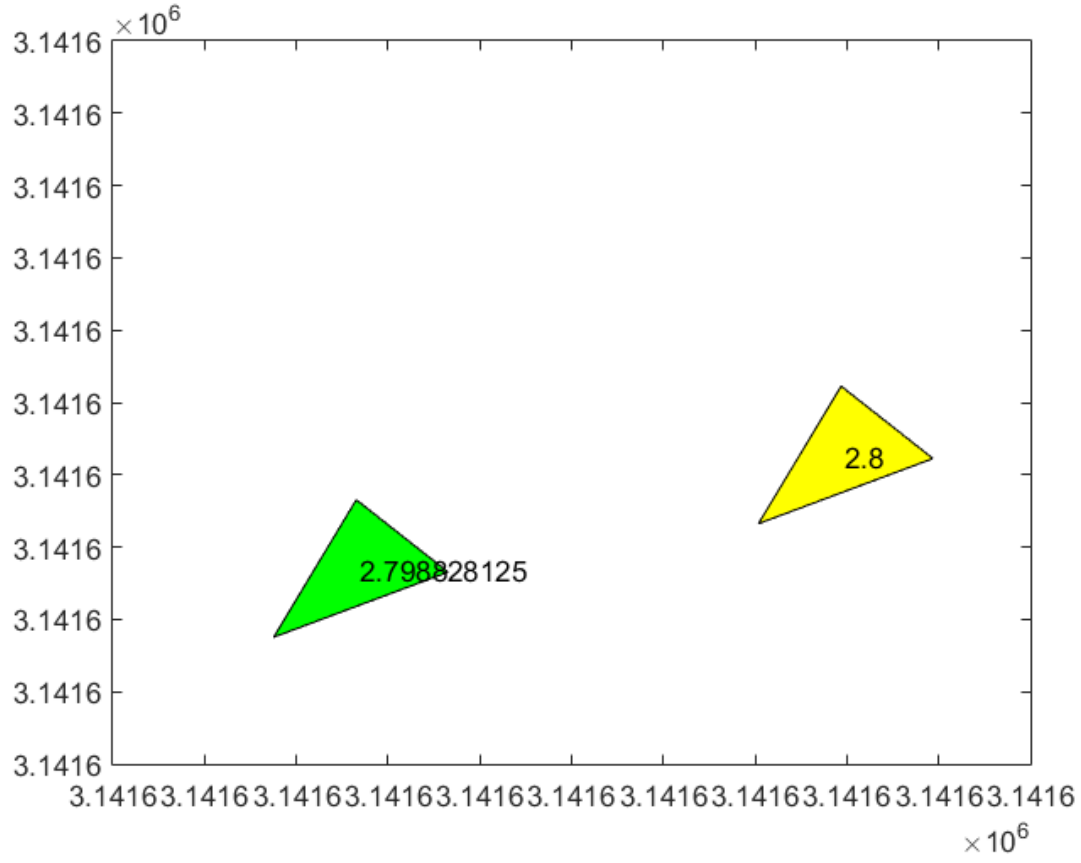
Les calculs définis par l'algorithme sont faits avec une arithmétique flottante, de précision ϵ
A cause des arrondis, le résultat du calcul est x_ϵ différent de x

Exemple d'évaluation: aire d'un triangle

Système orthonormé avec origine O et axes $0x$ et Oy
Triangle dont les sommets ont pour coordonnées $(x_1, y_1), (x_2, y_2), (x_3, y_3)$
Comment calculer l'aire du triangle ?



Exemple d'évaluation: aire d'un triangle



Aire d'un triangle (suite)

Calcul d'un déterminant 3×3

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

Algorithme 1

$$A = 1/2 \times |x_1 y_2 + x_2 y_3 + x_3 y_1 - x_1 y_3 - x_2 y_1 - x_3 y_2|$$

Somme de 6 produits avec signes différents: risque de cancellation.
Par exemple, il y a cancellation si le triangle est loin de l'origine.
L'erreur d'arrondi peut être très grande.

Aire d'un triangle (suite)

Algorithme 2: calcul d'un déterminant 2×2

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

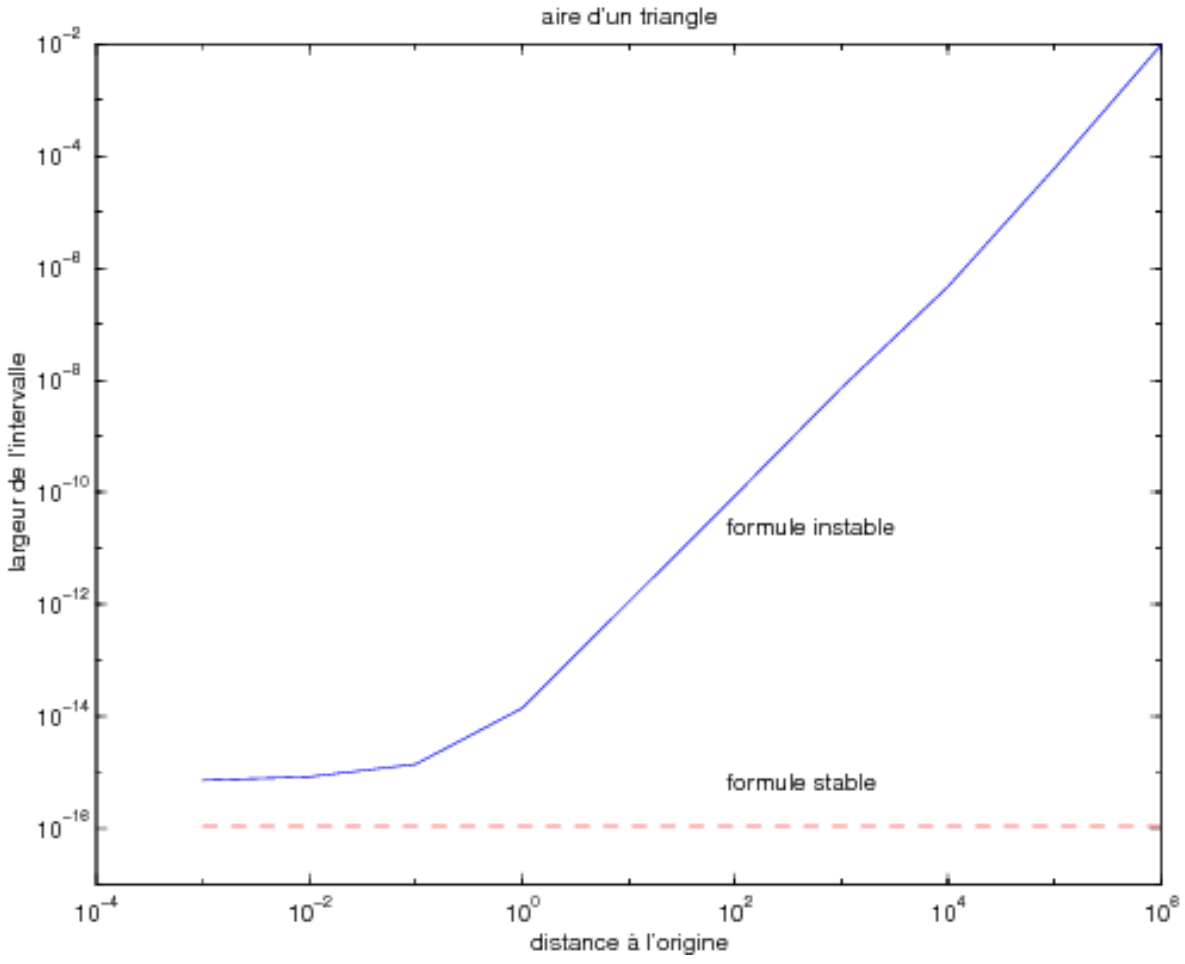
Le calcul est indépendant de la position du triangle par rapport à l'origine.
Mais il y a risque de cancellation, l'erreur d'arrondi peut être grande.

Algorithme 3: calcul précis d'un déterminant 2×2

Le calcul d'un déterminant 2×2 est identique au calcul de la partie réelle de deux nombres complexes.
On peut utiliser l'algorithme précis avec FMA et avec compensation d'erreur.
L'algorithme 3 est précis avec une erreur en $O(\epsilon)$

Calcul de l'aire d'un triangle

Calcul en arithmétique d'intervalles



Exemple de résolution: équation du second degré

Problème à résoudre

Equation $x^2 - ax + b = 0$ avec $a^2 - 4b > 0$ et $a > 0$ et $b \neq 0$
Deux racines distinctes x_1 et x_2

Algorithme 1

$D = a^2 - 4b$
$x_1 = a + \sqrt{D}$
$x_2 = a - \sqrt{D}$
Il y a risque de cancellation dans le calcul de x_2 si $D \simeq a^2$.

Algorithme 2

$x_2 = b/x_1$
L'algorithme 2 est plus précis.

Il y a dans les deux cas un risque de cancellation si D est proche de 0.
On peut calculer D avec un FMA pour éviter la cancellation.

Un algorithme est stable si l'erreur relative est bornée

$$\|x_\epsilon - x\| / \|x\| = O(\epsilon)$$

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

$$x_\epsilon = F(a_\epsilon) \text{ et } \|a_\epsilon - a\| / \|a\| = O(\epsilon)$$

ou

$$F(x_\epsilon, a_\epsilon) = 0 \text{ et } \|a_\epsilon - a\|/\|a\| = O(\epsilon)$$

Série harmonique (saison 3)

Algorithme 1

$$\begin{cases} S_1 = u_1 \\ S_i = S_{i-1} + u_i, \ i = 2, \ldots, n \\ S = S_n \end{cases}$$

Analyse des erreurs d'arrondi

$fl(S_i) = (fl(S_{i-1} + u_i)(1 + \alpha_i)$

$fl(S_i) = \sum_{i=1}^{n} u_i(1 + \eta_i)$ avec $\eta_i \leq c(n)\epsilon$ (au premier ordre)

L'algorithme 1 est inversement stable.

Série harmonique (saison 3- suite)

Algorithme 3

$$\begin{cases} S_1 = x_1 \\ r = 0 \\ \text{For } i = 2, n \\ (S_i, r_i) = Fast2Sum(S_{i-1}, x_i) \\ r = r + r_i \\ \text{EndFor} \\ S = S_n + r \end{cases}$$

$$|fl(S) - S| \leq \epsilon |S|/2$$

(au premier ordre)
L'algorithme 3 est stable.

Conditionnement d'un problème

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

Théorie des perturbations
Perturbation des données Δa
Erreur sur le résultat Δx
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

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

L'erreur relative sur le résultat est bornée

$$\frac{\|\Delta x\|}{\|x\|} \leq C \frac{\|\Delta a\|}{\|a\|}$$

Conditionnement d'une fonction réelle

Problème d'évaluation

$x = f(a)$ avec f fonction régulière d'un intervalle de $\mathbb{R}$ dans $\mathbb{R}$
Soit $x + \Delta x = f(a + \Delta a)$ alors
$f(a + \Delta a) = f(a) + f'(a)\Delta a + 1/2 f''(b)\Delta a^2$
$\Delta x = f'(a)\Delta a + O(\Delta a^2)$

Conditionnement au point a

Le conditionnement relatif $C(a)$ vaut

$$C(a) = \frac{|a||f'a)|}{|f(a)|}$$

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

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

Exemple

En simple précision, $\epsilon \simeq 10^{-7}$
Si un problème a un conditionnement égal à 10^4
et si on le traite avec un algorithme inversement stable
alors l'erreur relative sur le résultat est d'environ 10^{-3}

Série harmonique (saison 3- suite)

Conditionnement du calcul de S

$S + \Delta S = \sum_{i=1}^{n}(u_i + \Delta u_i)$ avec $|\Delta u_i| \leq |u_i||\Delta|$
$\Delta S = \sum_{i=1}^{n} \Delta u_i$
$|\Delta S| \leq (\sum_{i=1}^{n} |u_i|)|\Delta|$
Bleu

$$C = \frac{\sum_{i=1}^{n} |u_i|}{|S|}$$

Le conditionnement C est grand si S est petit par rapport à la somme des valeurs absolues.
Il vaut 1 lorsque tous les u_i sont positifs.

Erreurs d'arrondi

L'erreur d'arrondi de l'algorithme 1 est bornée par $C\, c(n)\epsilon$.
Il a risque de cancellation et d'imprécision dans l'algorithme 1 quand C est grand.
L'erreur de l'algorithme 3 ne dépend pas du conditionnement.

Retour vers l'équation du second degré

Equation $x^2 - ax + b = 0$ avec $a^2 - 4b > 0$ et $a > 0$ et $b \neq 0$
Deux racines distinctes x_1 et x_2 non nulles et différentes de $a/2$
Perturbations Δa et Δb : nouvelle racine $x_i + \Delta x$

Calcul de l'erreur relative

$(x_i + \Delta x)^2 - (a + \Delta a)(x_i + \Delta x) + (b + \Delta b) = 0$
Au premier ordre
$(x_i^2 - ax_i + b) + (2x_i - a)\Delta x - x_i \Delta a + \Delta b = 0$
$\Delta x = \frac{x_i \Delta a - \Delta b}{2x_i - a} = \pm \frac{x_i \Delta a - \Delta b}{\sqrt{a^2 - 4b}}$

$\frac{|\Delta x|}{|x_i|} \leq \frac{|a| + |b/x_i|}{\sqrt{a^2 - 4b}} \max\left(|\Delta a/a|, |\Delta b/b|\right)$

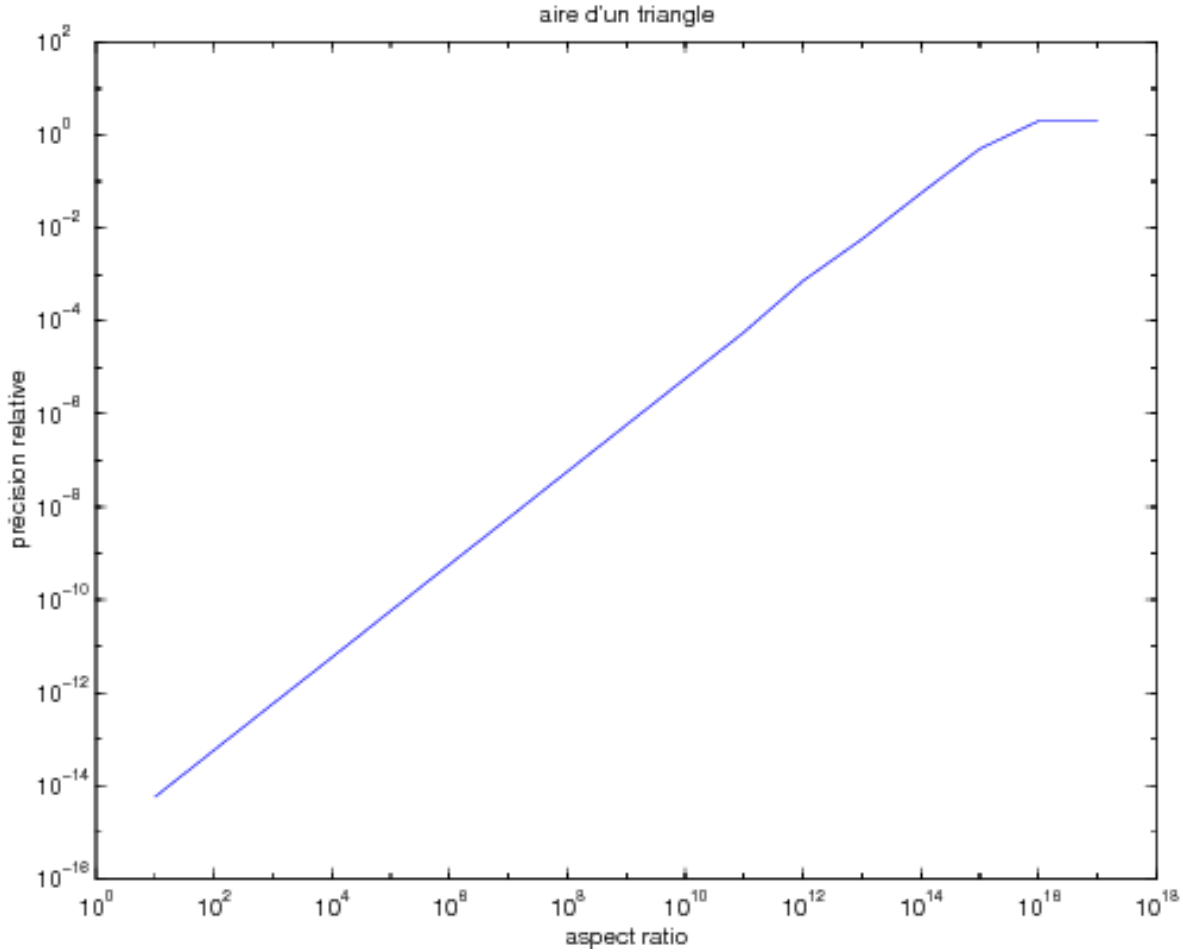
Conditionnement relatif

$$C_i = \frac{|a| + |b/x_i|}{\sqrt{a^2 - 4b}}$$

L'équation est mal conditionnée si $a^2 - 4b$ est proche de 0.
Il y a risque de cancellation et d'imprécision dans les deux algorithmes.

Aire d'un triangle aplati

L'aire d'un triangle de base a et de hauteur h est $A = \frac{1}{2} \times a \times h$
Le conditionnement est lié à l'aspect ratio du triangle défini par $\max a / \min h$
Calcul en arithmétique d'intervalles



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

Système linéaire

Problème

Résoudre $Ax = b$, avec A matrice carrée d'ordre n inversible et b vecteur.
La matrice A et le vecteur b sont les données du problème.
Le système a une solution unique $x = A^{-1}b$, le problème est bien posé.

Conditionnement du système

$$c(A) = \|A\|\|A^{-1}\|$$

Système linéaire: analyse des perturbations

Système

$Ax = b$ avec A inversible

Système perturbé

$(A + E)y = b + e$ avec $\|E\| \leq \alpha \|A\|$ et $\|e\| \leq \beta \|b\|$
Hypothèse $c(A)\alpha < 1$

Erreur relative

$$\frac{\|y - x\|}{\|x\|} \leq \frac{c(A)}{1 - c(A)\alpha}(\alpha + \beta)$$

Système linéaire: algorithme direct de résolution

Algorithme de Gauss avec Pivot Partiel

$PA = LU$
$Ly = Pb$
$Ux = y$

Algorithme inversement stable

La solution calculée x_ϵ vérifie
$(A + E)x_\epsilon = b$ avec $\|E\|_\infty \leq c(n)\rho \|A\|_\infty$

Erreur de calcul

$\frac{\|x_\epsilon - x\|}{\|x\|} = O(c(A)\epsilon)$
Sans la stratégie de pivot partiel, les erreurs d'arrondi peuvent exploser.

Système linéaire: algorithme direct de résolution

Matrice symétrique définie positive

$x^T Ax > 0$ pour $x \neq 0$

Algorithme de Cholesky

$A = LL^T$
$Ly = b$
$L^T x = y$

Algorithme inversement stable

sans besoin de pivot et sans facteur de croissance
$(A + E)x_\epsilon = b$ avec $\|E\|_2 \leq c(n) \|A\|_2$

Erreur de calcul

$\frac{\|x_\epsilon - x\|}{\|x\|} = O(c(A)\epsilon)$

Système linéaire: algorithme itératif de résolution

Méthode itérative

Approximations x_k et résidus $r_k = b - Ax_k$
Convergence si $\lim_{k\to+\infty} r_k = 0$

Résidu et erreur relative

$e = -r_k$ et $E = 0$ d'où $\frac{\|x_k - x\|}{\|x\|} \leq c(A)\frac{\|r_k\|}{\|b\|}$

Résidu et perturbation de matrice

$$\frac{\|x_k - x\|}{\|x\|} \leq 2\frac{c(A)}{1 - c(A)\eta}\eta$$

avec $\eta = \frac{\|r_k\|}{\|A\|\|b\| + \|b\|}$ et $c(A)\eta < 1$

η est la plus petite perturbation possible.

Système linéaire: estimation du conditionnement

Valeurs singulières

$$c(A) = \frac{\sigma_1}{\sigma_n}$$

Le calcul des valeurs singulières est difficile: on fait une estimation de $c(A)$.

Sensibilité aux conditions initiales

Equation différentielle

$$\begin{cases} dy/dt = f(y) \text{ pour } t \in [0, T] \\ y(0) = y_0 \end{cases}$$

Perturbation de la condition initiale y_0

Impact sur les résultats $y(t)$?

L'équation de Lorenz

Système climatique simplifié

- σ est le nombre de Prandtl,
- ρ est le nombre de Rayleigh,
- β est un paramètre.

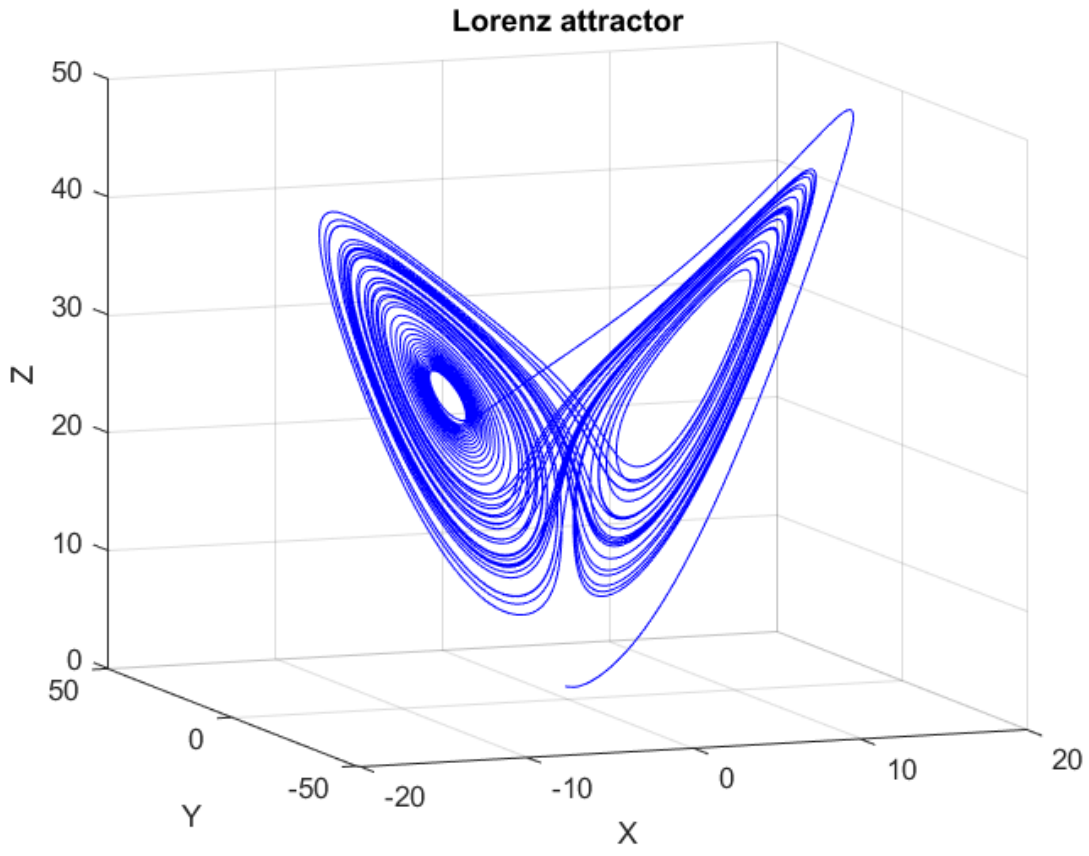
Système différentiel

$$\begin{cases} dx/dt = \sigma(y - x) \\ dy/dt = x(\rho - z) - y \\ dz/dt = xy - \beta z \end{cases}$$

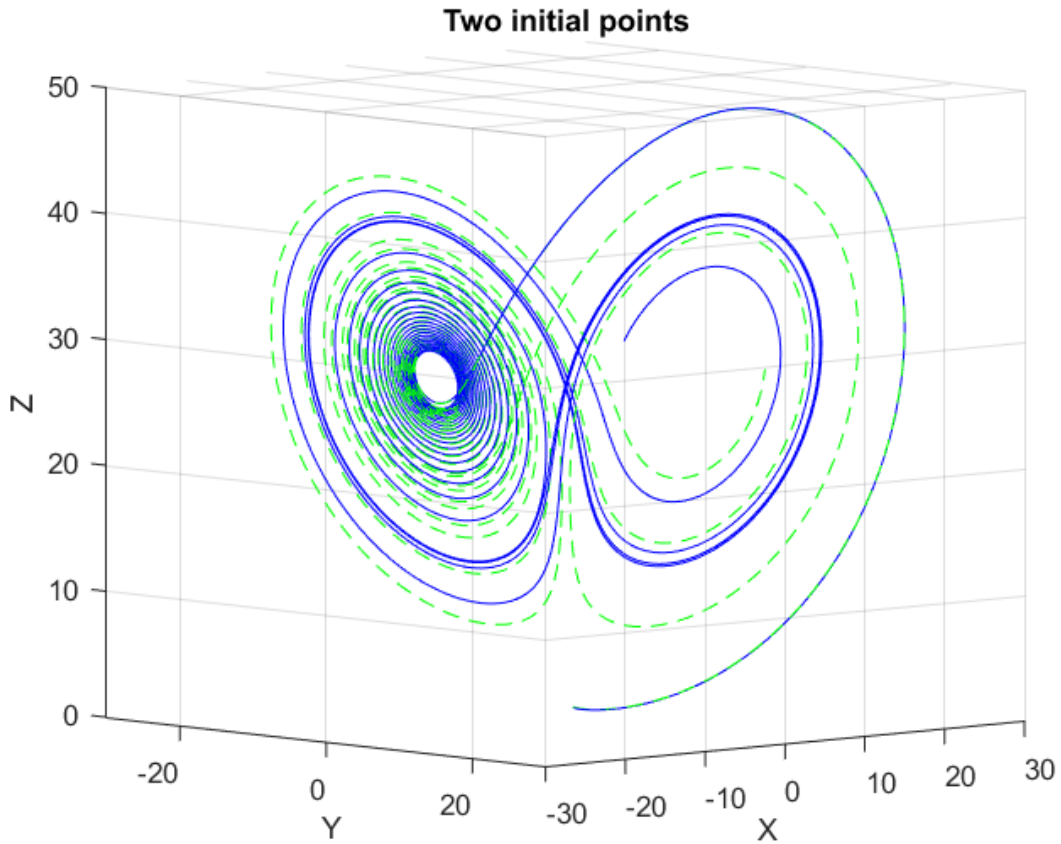
Conditions initiales x_0, y_0, z_0

Durée $[0, T]$

Attracteur de Lorenz



Effet papillon



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 d'une dérivée

f fonction réelle régulière au point a

Formule décentrée d'ordre 1

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

$$f(a+h) = f(a) + hf'(a) + h^2/2f''(a) + h^3/6f'''(a) + ...$$

Formule centrée d'ordre 2

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

$$f(a-h) = f(a) - hf'(a) + h^2/2f''(a) - h^3/6f'''(a) + ...$$

Erreurs d'approximation et d'arrondi

Cumul des erreurs

$|fl(d_h) - f'(a)|/|f'(a)| \leq |fl(d_h) - d_h|/|f'(a)| + |d_h - f'(a)|/|f'(a)|$

Erreur d'arrondi

$$\begin{aligned} fl(f(a+h)) &= f(a+h) * (1+\alpha), \\ fl(f(a)) &= f(a) * (1+\beta), \\ fl(f(a-h)) &= f(a-h) * (1+\gamma), \end{aligned}$$

avec $|\alpha| = O(\epsilon),\ |\beta| = O(\epsilon),\ |\gamma| = O(\epsilon)$

ce qui donne $\begin{array}{l} |fl(d_h) - d_h| \,/|f'(a)| = O(\epsilon/h), \\ |fl(c_h) - c_h| \,/|f'(a)| = O(\epsilon/h). \end{array}$

Erreur globale

$|fl(d_h) - f'(a)| \,/|f'(a)| = O(h) + O(\epsilon/h),$
$|fl(c_h) - f'(a)| \,/|f'(a)| = O(h^2) + O(\epsilon/h).$

Valeur optimale du pas d'approximation

Schéma décentré

$O(h) = O(\epsilon/h)$ donc $h = O(\sqrt{\epsilon})$

En double précision: pas $h = O(10^{-8})$ et erreur minimale en $O(10^{-8})$

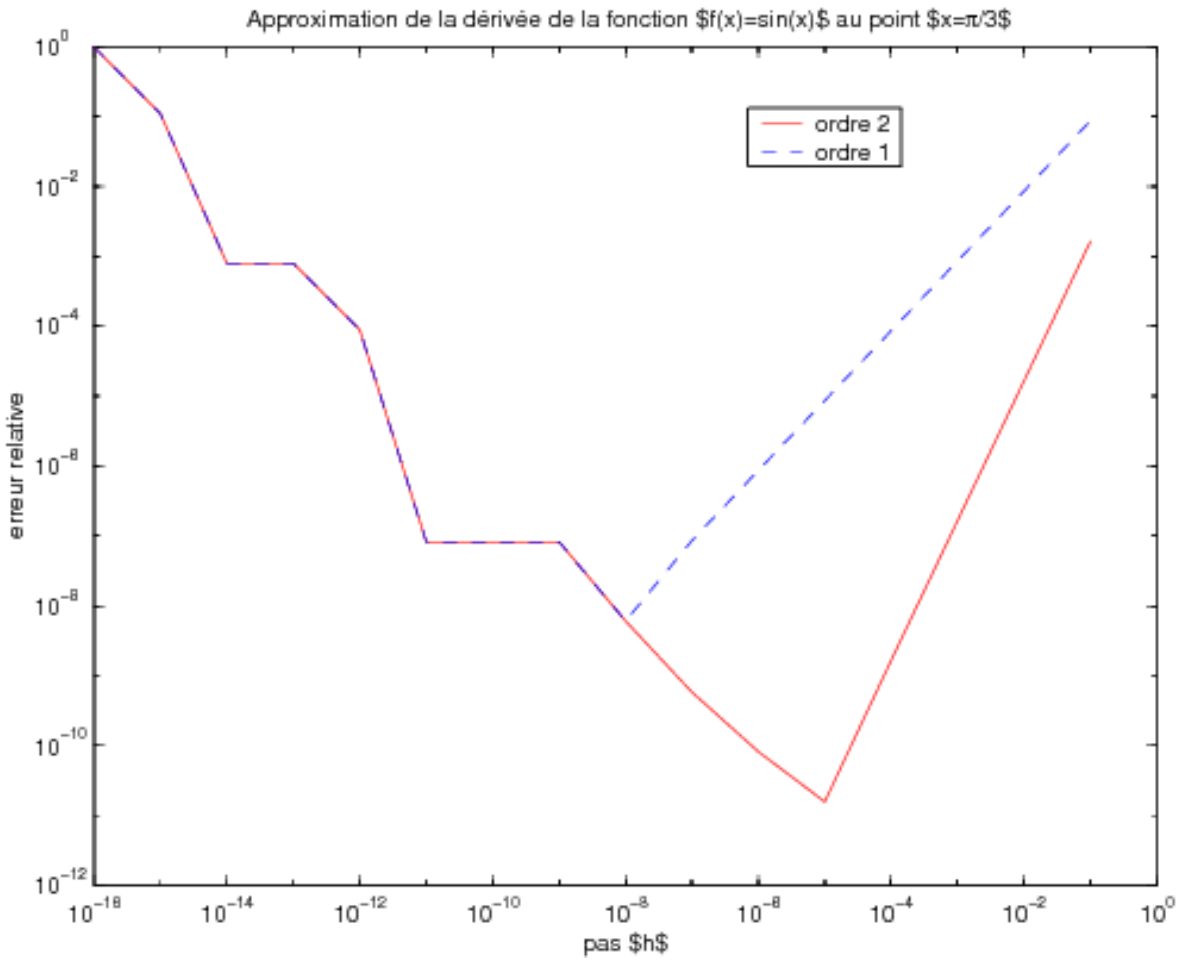
Schéma centré

$O(h^2) = O(\epsilon/h)$ donc $h = O(\epsilon^{1/3})$

En double précision: pas $h = O(10^{-5})$ et erreur minimale en $O(10^{-11})$

paramètre optimal avant explosion des erreurs d'arrondi

Approximation d'une dérivée



Conclusion

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

$$\text{Erreur en } O(C \times \epsilon)$$

Une méthode d'approximation est définie par un paramètre h
La valeur optimale de h équilibre approximation et arrondis