Chapitre 11

Instabilité numérique de suites définies par une relation de récurrence

Il s'agit d'étudier les erreurs dans le calcul numérique itératif d'une suite définie par une relation de récurrence.

Exemple : Calcul de $I_n = \displaystyle\int_0^1 \frac{x^n}{10+x}dx$, $n \in \mathbb{N}$.

On a :
$I_0 = \int_0^1 \frac{1}{10+x}dx = \ln\dfrac{11}{10}$
Pour $n > 0$
$I_n = \int_0^1 \frac{x^{n-1}(10+x)-10x^{n-1}}{10+x}dx = \int_0^1 x^{n-1}dx - 10\int_0^1 \frac{x^{n-1}}{10+x}dx = \dfrac{1}{n} - 10I_{n-1}$

L'erreur absolue sur I_n est approximativement égale à 10 fois l'erreur absolue sur I_{n-1} (sans compter l'erreur sur $\dfrac{1}{n}$) d'où on a un facteur d'amplification de l'erreur de 10^n en partant de $n = 0$. Donc il faut trouver une autre formulation algébrique.

On peut utiliser la décroissance de la suite x^n sur $[0,1]$ (d'où la décroissance de la suite I_n) et un encadrement de I_n en fonction de n : par intégration en utilisant l'encadrement $\frac{x^n}{11} \leq \frac{x^n}{10+x} \leq \frac{x^n}{10}$ on obtient $\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{10(n+1)}$ qui donne une approximation de I_n par $\frac{1}{11(n+1)}$ avec une erreur absolue $\Delta I_n \leq \frac{1}{10(n+1)} - \frac{1}{11(n+1)}$ (longueur de l'encadrement de I_n). On a une erreur relative sur I_n majorée par $\frac{1}{10}$. Ce résultat n'est pas très précis.

On peut étudier la récurrence inverse en partant de I_{n_0} :

$$I_{n-1} = \frac{1}{10}\left(\frac{1}{n} - I_n\right)$$

L'erreur absolue sur I_{n-1} est approximativement égale à $1/10$ de l'erreur absolue sur I_n (sans compter l'erreur sur $\dfrac{1}{n}$) d'où on a un facteur d'atténuation de l'erreur de $10^{-(n_0-n)}$ en partant de n_0). Donc ce choix de formulation algébrique est intéressant.

1 Effets des erreurs dues à l'arithmétique machine sur la méthode des approximations successives :

$x_{n+1} = F(x_n)$, x_0 donné
On suppose que l'on est dans les conditions de convergence de Lipschitz :
$\exists\ K < 1$, positif et un réel $a > 0$ tels que
$\forall\ y, z \in [x_0 - a, x_0 + a]$ on ait $|F(y) - F(z)| \leq K|y - z|$
et qu'on a $|F(x_0) - x_0| \leq a(1 - K)$
alors F a un point fixe unique x dans l'intervalle $[x_0 - a, x_0 + a]$
$x_n \in [x_0 - a, x_0 + a]$ et $|x_n - x| \leq \frac{K^n}{1-K}|x_1 - x_0|$
Remarque :

— le point initial x_0 de la suite peut avoir une erreur d'affectation.

— le calcul de F en un point peut être entaché d'erreur d'arrondi.

On a sur machine :

$$X_{n+1} = F(X_n) + e_n \quad \text{n = 0, 1, ...}$$

On suppose les erreurs e_n bornées par e (indépendante de n).
S'il existe $K < 1$, positif et un réel $\alpha > 0$ tel que F ait un point fixe $x \in [x - \alpha, x + \alpha]$ et $\forall y \in [x - \alpha, x + \alpha]$ on ait $|F(y) - F(x)| \leq K|y - x|$
Soit $X_0 \in [x - \alpha', x + \alpha']$ avec $0 < \alpha' \leq \alpha - \frac{e}{1-K}$
alors
$X_n \in [x - \alpha, x + \alpha]$ et $|X_n - x| \leq \dfrac{e}{1-K} + K^n(\alpha' - \dfrac{e}{1-K}) \quad (1)$
démonstration (par récurrence) :
$|X_0 - x| \leq \alpha' \leq \alpha' + \frac{e}{1-K} \leq \alpha$
On suppose que $|X_{n-1} - x| \leq \alpha$ on a :
$|X_n - x| \leq |F(X_{n-1}) - F(x) + e_{n-1}| \leq |F(X_{n-1}) - F(x)| + e \leq K|X_{n-1} - x| + e$
$\leq K^2|X_{n-2} - x| + Ke + e \leq \ldots \leq K^n|X_0 - x| + K^{n-1}e + \ldots + Ke + e$
$\leq K^n|X_0 - x| + \frac{1-K^n}{1-K}e \leq K^n|X_0 - x| + \frac{e}{1-K} - \frac{K^n}{1-K}e \leq K^n(\alpha' - \frac{e}{1-K}) + \frac{e}{1-K}$
$\leq \alpha' + \frac{e}{1-K} \leq \alpha$
remarque :
Le second membre de l'inégalité (1) tend vers $\dfrac{e}{1-K}$ quand n tend vers ∞ ($K < 1$). Ainsi la suite X_n peut osciller au voisinage de x (d'amplitude bornée par $\dfrac{e}{1-K}$) sans s'y rapprocher : les erreurs machine deviennent plus grandes que la précision donnée a priori.
exemple :
$x_{n+1} = (x_n^2 - 1/x_n^2)/(2x_n - 2/x_n)$ $n = 0, 1, ..., x_0 = 1{,}05$, de point fixe 1, on a les résultats suivants :

1.00119 ; 0.999863 ; 0.998482 ; 0.999853 ; 0.99919 ; 0.999706 ; 0.999595 ...

Remarque :
Si F est différentiable dans $]x - \tau, x + \tau[$ avec $\tau > 0$ et $|F'(x)| < 1$ alors il existe un intervalle $]x - \xi, x + \xi[$ avec $\xi > 0$ tel que $\forall x_0 \in\ ]x - \xi, x + \xi[\quad x_n$ converge vers x.

2 Applications :

2.1 Méthode de Bernoulli

On considère
— l'équation algébrique $P_n(x) = a_0 x^n + \cdots + a_n = 0$
— l'équation récurrente $a_0 y_k + \cdots + a_n y_{k-n} = 0$
Les solutions particulières de l'équation récurrente sont de la forme $y_k = \beta^k$, les constantes β sont racines de l'équation caractéristique $a_0\beta^k + \cdots + a_n\beta^{k-n} = 0$ qui est l'équation algébrique de racines distinctes x_i telles que :

$$|x_1| > |x_2| > \cdots > |x_n|$$

$$\text{on a } y_k = \sum_{i=1}^{n} c_i x_i^k \text{ soit } y_k = c_1 x_1^k (1 + \frac{c_1}{c_2}(\frac{x_2}{x_1})^k + \cdots)$$

$$\text{soit } \frac{y_k}{y_{k-1}} = x_1 \frac{(1 + \frac{c_1}{c_2}(\frac{x_2}{x_1})^k + \cdots)}{(1 + \frac{c_1}{c_2}(\frac{x_2}{x_1})^{k-1} + \cdots)} \text{ d'où } \lim_{k\to\infty} \frac{y_k}{y_{k-1}} = x_1$$

Cette méthode donne la plus grande (en module) racine x_1 (ou la plus petite (en module) en prenant le polynôme réciproque de $P_n = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$, noté $P_n^*(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$). Les valeurs de y_k sont obtenues par la relation récurrente avec les conditions initiales par exemple $y_0 = y_1 = \cdots = y_{n-2} = 0$, $y_{n-1} = 1$ pour avoir $c_1 \neq 0$.

3 Vitesse de convergence et accélération de convergence d'une suite :

On suppose x_n converge vers x.
Définition : on dit que la suite (x_n) a une convergence d'ordre $r \geq 1$, s'il existe une constante C finie, non nulle telle que :

$$\lim_{n\to\infty} \frac{|x_{n+1} - x|}{|x_n - x|^r} = C$$

C s'appelle la constante asymptotique d'erreur ou facteur de convergence.
Si $r > 1$ la suite est à convergence exponentielle
Si $r = 1$ et $C \neq 1$ la suite est à convergence linéaire
Si $r = 1$ et $C = 1$ la suite est à convergence logarithmique
La suite (x_n) définie par $x_{n+1} = F(x_n)$, x_0 donné est d'ordre entier :
c'est le nombre r tel que $F'(x) = \cdots = F^{(r-1)}(x) = 0$ et $F^{(r)}(x) \neq 0$ où x est le point fixe de F.
On a alors $C = \dfrac{|F^{(r)}(x)|}{r!}$

Evolution de la précision au fur et à mesure des itérations
on a $10^{-d_n} = |x_n - x|$, soit $d_n = -\log_{10}|x_n - x|$
d_n représente à une constante additive près indépendante de n, le nombre de chiffres décimaux exacts de x_n.
Par définition d'une suite, pour $n > N$ on a

$$|x_{n+1} - x| \approx C|x_n - x|^r$$

d'où en prenant le logarithme on a :

$$d_{n+1} \approx r d_n + R \text{ (avec } R = \text{-}\log_{10} C).$$

donc à chaque itération (pour x_n suffisamment près de x) on multiplie le nombre de chiffres exacts par r et on rajoute R.

Comparaison des vitesses de convergence de deux suites :
$x_n \to x$; $y_n \to x$
y_n converge plus vite que x_n si $\lim_{n\to\infty} \frac{y_n - x}{x_n - x} = 0$
Si y_n est d'ordre r et x_n d'ordre p avec $r > p$ alors y_n converge plus vite que x_n.
Accélération de convergence de suite :
Si (x_n) converge lentement, (peut être impossible de converger en pratique) par exemple pour $r = 1$, on peut accélérer sa convergence : transformer en une autre suite qui converge plus vite vers la même limite.

3.1 Procédé Δ^2 d'Aitken

S'il existe un nombre réel a non nul et différent de 1 tel que $\lim_{n\to\infty} \frac{x_{n+1} - x}{x_n - x} = a$, on peut accélérer la suite (x_n) par la suite (1) $y_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$, on calcule en parallèle les deux suites.
(y_n) converge plus vite que (x_n)

Choix de la formulation algébrique du Δ^2 d'Aitken :
(1) est plus stable numériquement que la formulation (2) définie par

$$\frac{x_n x_{n+2} - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n}$$

(2) est numériquement instable car les erreurs de cancellation (différence de deux nombres voisins) dit aussi phénomène de compensation sont grandes.
Dans (1) on retranche à x_n un terme correctif (ces erreurs sont du deuxième ordre).
Dans (2) il n'y a pas de terme correctif, l'erreur de cancellation est du premier ordre. On peut aussi remarquer qu'au voisinage de la solution, des oscillations peuvent se produire.

3.2 ϵ-algorithme de Wynn

Dans cet algorithme on calcule des quantités à 2 indices $\epsilon_k^{(n)}$:
$\epsilon_{-1}^{(n)} = 0$, $\epsilon_0^{(n)} = x_n \quad n = 0, 1, ...$
$\epsilon_{k+1}^{(n)} = \epsilon_{k-1}^{(n+1)} + \frac{1}{\epsilon_k^{(n+1)} - \epsilon_k^{(n)}} \quad n, k = 0, 1, ...$

On place ces quantités dans un tableau à double entrée noté ϵ :

$$
\begin{array}{llllllllll}
0 = \epsilon_{-1}^{(0)} & & & & & & & & & \\
\quad \vdots & \searrow & & & & & & & & \\
\quad \vdots & & \epsilon_0^{(0)} & & & & & & & \\
\quad \vdots & \nearrow & \vdots & \searrow & & & & & & \\
0 = \epsilon_{-1}^{(1)} & & \vdots & & \epsilon_1^{(0)} & & & & & \\
\quad \vdots & \searrow & \vdots & \nearrow & \vdots & \searrow & & & & \\
\quad \vdots & & \epsilon_0^{(1)} & & \vdots & & \epsilon_2^{(0)} & & & \\
\quad \vdots & \nearrow & & \searrow & \vdots & \nearrow & \vdots & \searrow & & \\
0 = \epsilon_{-1}^{(2)} & & & & \epsilon_1^{(1)} & & \vdots & & \epsilon_3^{(0)} & \\
\quad \vdots & & & & & & & & & \searrow
\end{array}
$$

La relation à double indice relie des quantités aux quatre sommets d'un losange :

$$
\begin{array}{ccccc}
 & & \epsilon_k^{(n)} & & \\
 & \nearrow & & \searrow & \\
\epsilon_{k-1}^{(n+1)} & & & & \epsilon_{k+1}^{(n)} \\
 & \searrow & & \nearrow & \\
 & & \epsilon_k^{(n+1)} & &
\end{array}
$$

Dans le tableau ϵ on progresse de la gauche vers la droite et de haut en bas.
Si on connaît x_0, x_1, ..., x_{2k} pour k fixé, alors le tableau ϵ est fini

$$
\begin{array}{llll}
\epsilon_0^{(0)} & & & \\
\vdots & \searrow & & \\
\vdots & & \epsilon_1^{(0)} & \\
\vdots & & & \searrow \\
\vdots & & & \epsilon_{2k}^{(0)} \\
\vdots & & & \nearrow \\
\vdots & & \epsilon_1^{(2k-1)} & \\
\vdots & \nearrow & & \\
\epsilon_0^{(2k)} & & &
\end{array}
$$

On montre sous certaines hypothèses que $\epsilon_{2k}^{(n)} \to x$ avec n fixé ou k fixé.
$\epsilon_2^{(n)}$ est le procédé d'Aitken.

3.3 Le procédé d'extrapolation de Richardson

on pose $T_0^{(n)} = x_n \quad n = 0, 1, ...$ on calcule des quantités à deux indices $T_k^{(n)}$ par

$$T_{k+1}^{(n)} = \frac{z_n T_k^{(n+1)} - z_{n+k+1} T_k^{(n)}}{z_n - z_{n+k+1}} \quad n,\, k = 0,\, 1,\, ...$$

on utilise ici une suite auxiliaire (z_n)
On place ces quantités dans un tableau T à double entrée :

$$\begin{array}{lllllllll}
x_0 = T_0^{(0)} & & & & & & & & \\
\quad \vdots & \searrow & & & & & & & \\
\quad \vdots & & T_1^{(0)} & & & & & & \\
\quad \vdots & \nearrow & \vdots & \searrow & & & & & \\
x_1 = T_0^{(1)} & & \vdots & & T_2^{(0)} & & & & \\
\quad \vdots & \searrow & \vdots & \nearrow & \vdots & \searrow & & & \\
\quad \vdots & & T_1^{(1)} & & \vdots & & T_3^{(0)} & & \\
\quad \vdots & \nearrow & & \searrow & \vdots & \nearrow & \vdots & \searrow & \\
x_2 = T_0^{(2)} & & & & T_2^{(1)} & & \vdots & & T_4^{(0)} \\
\quad \vdots & & & & & & & & \searrow
\end{array}$$

la relation à double indice relie des quantités situées aux trois sommets d'un triangle

$$\begin{array}{lll}
T_k^{(n)} & & \\
\quad \vdots & \searrow & \\
\quad \vdots & & T_{k+1}^{(n)} \\
\quad \vdots & \nearrow & \\
T_k^{(n+1)} & &
\end{array}$$

On progresse dans le tableau T comme pour l'ϵ-algorithme.
en pratique, si on connaît x_0, x_1, ..., x_k pour k fixé, le tableau infini devient fini :

$$\begin{array}{lll}
T_0^{(0)} & & \\
\quad \vdots & \searrow & \\
\quad \vdots & & T_k^{(0)} \\
\quad \vdots & \nearrow & \\
T_0^{(k)} & &
\end{array}$$

<u>Résultat</u> :
Soit (z_n) une suite strictement décroissante de réels positifs telle que $z_n \overset{n\to\infty}{\longrightarrow} 0$, s'il existe $a > 1$ tel que $\dfrac{z_n}{z_{n+1}} \geq a$ $n = 0,\, 1,\, ...$ alors $\lim\limits_{k\to\infty} T_k^{(n)} = x$ et $\lim\limits_{n\to\infty} T_k^{(n)} = x$ et réciproquement.
ex : $z_n = b^n \quad 0 < b < 1$
<u>Remarque</u> : avec ce résultat, les colonnes et les diagonales du tableau T convergent vers x

Dans les conditions du résultat, la suite $(T_k^{(n+1)})$ à k fixé converge plus vite que la suite $(T_k^{(n)})$ à k fixé si et seulement si

$$\lim_{n\to\infty} \frac{T_k^{(n+1)} - x}{T_k^{(n)} - x} = \lim_{n\to\infty} \frac{z_{n+k+1}}{z_n}$$

Remarque : Soit $x_n = f(z_n)$ avec $z_n \overset{n\to\infty}{\longrightarrow} 0$
On considère $[a,b]$ le plus petit intervalle contenant la suite z_n, $0 \in [a,b]$
Si f est suffisament différentiable sur $[a,b]$ et si les dérivées sont continues, on a :

$$T_k^{(n)} - x = (-1)^{k+1} \frac{z_n \cdots z_{n+k}}{(k+1)!} f^{(k+1)}(\xi) \text{ où } \xi \in [a,b]$$

D'où le résultat suivant :

1. Si pour k fixé, $|f^{(k+1)}(z)| \leq M \quad \forall z \in [a,b]$, M constante indépendante de z, alors $\lim_{n\to\infty} T_k^{(n)} = x$
2. Si pour tout k et pour tout $z \in [a,b]$, $|f^{(k+1)}(z)| \leq M$ (constante indépendante de k et z) alors
   — $\lim_{k\to\infty} T_k^{(n)} = x \quad n = 0, 1, ...$
   — $\lim_{n\to\infty} T_k^{(n)} = x \quad k = 0, 1, ...$

Dans ce résultat, on n'impose à la suite (z_n) que la condition de convergence vers 0.

Application :
Méthode de Romberg en quadrature numérique.