PETITES PERTURBATIONS ET GRANDES ERREURS

Samir AKESBI et Jean LEFORT [1]

INTRODUCTION

Calculer est une part non négligeable de l'activité mathématique. L'usage systématique de puissantes machines a rejeté dans l'ombre les problèmes de fiabilité des calculs. Ne suffit-il pas de travailler avec beaucoup de chiffres significatifs pour être sûr du résultat sans avoir à se préoccuper de la précision ou à se lancer dans un calcul d'erreurs toujours assez fastidieux ?

En fait les logiciels actuels ne font que reculer les difficultés et s'ils permettent de résoudre sans fautes un système de quelques équations à quelques inconnues (en général), les problèmes resurgissent devant un système de quelques milliers d'équations avec autant d'inconnues (et ces problèmes ne sont pas rares dans l'industrie).

Par ailleurs, des problèmes courants sont fournis avec des données mesurées, c'est-à-dire entachées d'une erreur de mesure. Comment se propage jusqu'au résultat une légère perturbation des données ? Le résultat reste-t-il fiable et avec quelle précision ?

Enfin de nombreuses méthodes mathématiques ne sont elles-mêmes que des méthodes approchées : calcul d'intégrales définies, tabulation de fonctions, résolution d'équations différentielles... On y discrétise un problème continu et l'on se rend vite compte qu'il ne sert à rien de prendre le pas de discrétisation trop petit et qu'il y a un optimum fonction de la méthode utilisée et des erreurs d'arrondis dont on ne peut s'affranchir.

QUELQUES EXEMPLES

Illustrons un peu notre propos :

1) Considérons le système linéaire $AX = B$: $\begin{pmatrix} 10 & 9 \\ 9 & 8 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 19 \\ 17 \end{pmatrix}$.

On a $\det A = -1$, $A^{-1} = \begin{pmatrix} -8 & 9 \\ 9 & -10 \end{pmatrix}$ et la solution exacte est : $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Modifions très légèrement le vecteur colonne B en $B' = \begin{pmatrix} 19{,}2 \\ 16{,}9 \end{pmatrix}$; alors la nouvelle solution est : $X' = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -8 & 9 \\ 9 & -10 \end{pmatrix}\begin{pmatrix} 19{,}2 \\ 16{,}9 \end{pmatrix} = \begin{pmatrix} -1{,}5 \\ 3{,}8 \end{pmatrix}$. On peut comparer la perturbation initiale à la perturbation finale en utilisant la norme euclidienne des vecteurs colonnes et en mettant en regard les erreurs relatives : $\dfrac{\|B'-B\|}{\|B\|} = 0{,}0088$ tandis que $\dfrac{\|X'-X\|}{\|X\|} = 2{,}6542$. L'erreur relative est ainsi multipliée par 300 et quelques !

2) Reprenons le même système linéaire et perturbons cette fois ci la matrice A en écrivant $A'Y = B$ avec $A' = \begin{pmatrix} 10 & 8{,}9 \\ 9{,}1 & 8{,}1 \end{pmatrix}$ ce qui donne : $A'^{-1} = \begin{pmatrix} 810 & -890 \\ -910 & 1000 \end{pmatrix}$, $\det A' = 0{,}01$ et la

[1] D'après la conférence prononcée par Samir Akesbi le 23 février 1994 à l'U.H.A à Mulhouse et qui avait pour titre : « *Grandes erreurs dues à des petites perturbations en calcul sur machine* ». Le texte de la conférence a été remanié et augmenté pour le présent article.

© L'OUVERT 77 (1994)

solution est $Y = \begin{pmatrix} 260 \\ -290 \end{pmatrix}$. On remarque que ni dét A' , ni A'^{-1} , ni Y n'ont de point commun du point de vue ordre de grandeur avec dét A , A^{-1} et X . Une variation d'environ 1% en valeur relative sur chacun des coefficients de A a entraîné des variations non maîtrisables (de l'ordre de 300 % en valeur relative) sur les divers résultats.

3) Prenons maintenant le système linéaire : $\begin{cases} (1+\sqrt{3})x + \sqrt{3}y &= 0 \\ 2x + (3-\sqrt{3})y &= 0 \end{cases}$ dont le déterminant est nul. Remplaçons ce système par le système approché : $\begin{cases} 2{,}732x + 1{,}732y &= 0 \\ 2x + 1{,}268y &= 0 \end{cases}$ dont le déterminant vaut alors $2.10^{-4} \neq 0$ et le système n'admet plus que la solution (0 ; 0) ce qui est fort ennuyeux si on cherche, par exemple, les vecteurs propres d'une matrice dont on n'a les valeurs propres que de façon approchée.

4) Un troisième exemple de système linéaire est donné par le suivant que l'on cherche à résoudre par la méthode du pivot de Gauss : $\begin{cases} 37639840x - 46099201y &= 0 \\ 29180479x - 35738642y &= -1 \end{cases}$. Avec une calculatrice travaillant avec dix chiffres on trouve (x = 78,9898 et y = 64,4949) ; avec une calculatrice travaillant avec seize chiffres on obtient (x = 42407323,20 et y = 34625434,40) . Cependant la valeur exacte de la solution est (x = 46099201 et y = 37639840).

5) Considérons le polynôme $P(x)$ = 12192 x^3 – 32257 x^2 – 85344 x + 225799 et calculons P(2,645752) . La plupart des calculatrices donnent 0 (que l'on utilise la méthode de Horner ou une méthode directe) alors que la valeur exacte est 3,0563905536 10^{-8} . Mais même une calculatrice travaillant avec 16 chiffres significatifs donne 3,05590220 10^{-8} par la méthode de Horner et 3,06172296 10^{-8} par un calcul direct ce qui correspond tout de même à une erreur relative de 1,7 ‰ , alors que x est connu à moins de 10^{-6} près.

6) On se propose de calculer une valeur approchée de la dérivée seconde au point 1 de la fraction rationnelle : $f(x) = \dfrac{4970x - 4923}{4970x^2 - 9799x + 4830}$. Pour cela, on utilise la formule : $f''(1) = \lim\limits_{h \to 0} \dfrac{f(1-h) - 2f(1) + f(1+h)}{h^2}$ en prenant h petit. La valeur exacte de $f''(1)$ est 94. En prenant $h = 10^{-4}$ on devrait trouver la valeur approchée 90,79 et pour $h = 10^{-5}$, 93,76. Or un calcul sur ordinateur conduit respectivement à 113,59 et à 4952,31 .

7) N'oublions pas l'exemple classique depuis l'article de Vignes en 1984 [2] : Calculer l'expression $9x^4 - y^4 + 2y^2$ pour x = 10864 et y = 18817 , calcul dont le résultat dépend essentiellement de la marque et du type de machine utilisée, ce qui a un effet spectaculaire sur des élèves de lycée, surtout quand ils apprennent que le résultat exact est 1 !

8) Dans l'article de Vignes déjà cité, on trouve le problème suivant : calculer exp(–30) à partir de la série entière de la fonction exponentielle. Le calcul itératif fait avec 14 chiffres significatifs sur un ordinateur CDC 7600 conduit à 6,25 10^{-4} au lieu de 9,36 10^{-14} .

Nous pourrions multiplier les exemples et nous en rencontrerons d'autres dans cet article. Cherchons à comprendre ce qui se passe dans les différents cas pour savoir s'il est possible d'éviter ce genre de phénomènes et si oui, comment ?

---

[2] *Peut-on faire confiance aux ordinateurs pour faire du calcul scientifique ?* in « Gazette des mathématiciens » n°24, avril 1984 .

UNE TENTATIVE DE CLASSIFICATION

Traditionnellement, on parle de problèmes de conditionnement dès qu'apparaissent les difficultés qu'ont illustrées les exemples précédents. Mais il nous faut distinguer au moins deux types différents de conditionnement : ceux inhérents au problème proposé, et ceux inhérents à la machine utilisée qui cumule les erreurs d'arrondi. Cette classification est assez sommaire. Un même problème peut être résolu par des algorithmes différents qui sur une même machine conduiront à des résultats plus ou moins précis. Nous ne ferons dans cet article qu'une brève allusion aux questions d'algorithme pour insister essentiellement sur les problèmes mal conditionnés.

Il ne faudra cependant pas oublier que les trois types de difficultés : conditionnement, arrondi et algorithme, sont souvent mêlés. Dans le paragraphe précédent, il est possible de dire que les exemples 1 et 2 sont clairement des problèmes de conditionnement, l'exemple 3 est clairement un problème d'arrondis mais les exemples 4 et 5 mêlent les trois difficultés. Quant aux exemples de calcul de valeurs de fonction, il est encore plus difficile de séparer le conditionnement intrinsèque du problème d'avec les erreurs d'arrondis.

§1 – CONDITIONNEMENT DES SYSTÈMES LINÉAIRES

NORME D'UNE MATRICE

Nous supposons que dans l'espace $\mathbf{R}^n$ c'est la norme euclidienne qui est utilisée et nous définissons la norme euclidienne induite d'une application linéaire représentée par sa matrice A par : $\|A\| = \sup_{X\neq 0} \frac{\|AX\|}{\|X\|}$ (définition clairement indépendante de la base orthonormée choisie).

Pour démontrer quelques propriétés élémentaires de cette norme sur les applications linéaires (et donc sur les matrices) nous définirons le rayon spectral $\rho(A)$ d'une matrice (ou de l'application linéaire correspondante) comme étant $\max\{|\lambda_i|,\ \text{dét}(A-\lambda_i I) = 0\}$ c'est-à-dire que le rayon spectral est la plus grande des valeurs absolues des valeurs propres calculées sur $\mathbf{C}$ (et non sur $\mathbf{R}$).

Si A est symétrique, alors $\|A\| = \rho(A)$. En effet, en notant λ la valeur propre de plus grand module, c'est-à-dire que l'on a $\rho(A) = |\lambda|$, et en prenant pour X, dans la définition de $\|A\|$, un vecteur propre associé à λ, on voit que $\|A\| \geq \rho(A)$. Réciproquement, on sait qu'une matrice symétrique admet une base de vecteurs propres orthonormés. Dans une telle base (e_i) on a : $X = \Sigma a_i e_i$ et $AX = \Sigma a_i \lambda_i e_i$ mais aussi $\|X\|^2 = \Sigma a_i^2$ et $\|AX\|^2 = \Sigma a_i^2 \lambda_i^2 \leq \lambda^2 \Sigma a_i^2$ ce qui prouve que $\|A\| \leq |\lambda|$ et assure la conclusion annoncée.

La première partie de la démonstration montre que l'on a toujours $\|A\| \geq \rho(A)$.

La définition de la norme d'une matrice implique que $\|AX\| \leq \|A\| \times \|X\|$. Mais on a d'une façon plus générale : $\|AB\| \leq \|A\| \times \|B\|$. Ceci se démontre grâce aux égalités et inégalités suivantes :

$$\|AB\| = \sup_{X\neq 0} \frac{\|ABX\|}{\|X\|} = \sup_{X\neq 0} \frac{\|ABX\|}{\|BX\|} \frac{\|BX\|}{\|X\|} \leq \sup_{X\neq 0} \frac{\|ABX\|}{\|BX\|} \times \sup_{X\neq 0} \frac{\|BX\|}{\|X\|} \leq \|A\| \times \|B\| .$$

On trouvera en annexe le calcul de la norme d'une matrice dans le cas général.

RETOUR SUR LES DEUX PREMIERS EXEMPLES

Le premier exemple consiste, à partir d'un système $AX = Z$, à perturber Z et à chercher la perturbation qui en résulte sur X, c'est-à-dire à transformer le système proposé en

$A(X + \Delta X) = Z + \Delta Z$ puis à calculer ΔX en fonction de ΔZ. Nous avons clairement, comme $AX = Z$, $A\Delta X = \Delta Z$ soit $\Delta X = A^{-1}\Delta Z$ et en passant aux normes :

$\|\Delta X\| = \|A^{-1}\Delta Z\| \leq \|A^{-1}\| \times \|\Delta Z\|$. Par ailleurs nous avons $\|Z\| \leq \|AX\| \leq \|A\| \times \|X\|$ et en divisant membre à membre (dans le bon sens !) les deux dernières inégalités sur les normes, il vient : $\dfrac{\|\Delta X\|}{\|X\|} \leq \|A\| \times \|A^{-1}\| \times \dfrac{\|\Delta Z\|}{\|Z\|}$.

Le deuxième exemple consiste à perturber la matrice A et à chercher la perturbation qui en résulte sur le résultat X. Nous avons donc transformé le système en $(A + \Delta A)(X + \Delta X) = Z$. Nous trouvons donc $A\Delta X + \Delta A(X + \Delta X) = 0$ soit $\Delta X = -A^{-1}\Delta A(X + \Delta X)$ et en passant aux normes : $\|\Delta X\| \leq \|A^{-1}\| \times \|\Delta A\| \times \|X + \Delta X\|$ ou encore: $\dfrac{\|\Delta X\|}{\|X + \Delta X\|} \leq \|A\| \times \|A^{-1}\| \times \dfrac{\|\Delta A\|}{\|A\|}$.

Dans les deux cas nous voyons apparaître la quantité $\|A\| \times \|A^{-1}\|$ à laquelle nous donnerons le nom de conditionnement de la matrice A et que nous noterons $\mathrm{Cond}(A)$.

Dans les exemples proposés, A est une matrice symétrique de valeurs propres $9 \pm \sqrt{(82)}$ ce qui conduit à : $\|A\| = 9 + \sqrt{(82)}$ et $\|A^{-1}\| = |9 - \sqrt{(82)}|^{-1} = 9 + \sqrt{(82)}$ et par suite $\mathrm{cond}(A) \approx 326$. Ce résultat explique bien la multiplication par 300 des erreurs relatives. Il faut toutefois noter que $\Delta A = \begin{pmatrix} 0 & -0{,}1 \\ 0{,}1 & 0{,}1 \end{pmatrix}$ et que $\|\Delta A\|$ ne peut se calculer qu'en revenant à la définition ; on trouve $10^{-1}\phi$, où ϕ est le nombre d'or, soit environ 0,16.

PROPRIÉTÉS DU CONDITIONNEMENT

a) Le conditionnement d'une matrice est toujours supérieur ou égal à 1. En effet : $1 = \|I\| = \|AA^{-1}\| \leq \|A\| \times \|A^{-1}\| = \mathrm{cond}(A)$. Plus ce nombre est proche de 1, meilleur est le conditionnement.

b) On a de façon immédiate : $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$ et $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.

c) Si on note μ_n et μ_1 la plus grande et la plus petite valeur propre de la matrice symétrique ${}^tA \times A$ (ou ${}^{\bullet}A \times A$ dans le cas complexe), alors $\mathrm{cond}(A) = \sqrt{(\mu_n / \mu_1)}$. (On trouvera la démonstration en annexe).

d) Si A est une matrice orthogonale (c'est-à-dire si ${}^tA = A^{-1}$) alors $\mathrm{cond}(A) = 1$.

On voit donc que le conditionnement est une question de largeur spectrale et non pas une question de valeur du déterminant. Certes si le déterminant est petit en valeur absolue, la plus petite des valeurs propres sera faible et le conditionnement risque d'être mauvais, mais l'exemple 1 montre que cela n'est nullement une condition nécessaire.

Ce qu'il faut aussi savoir c'est que le conditionnement est en général d'autant plus mauvais que la matrice a une grande taille. Par exemple la matrice $n \times n$ suivante : $\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}$

admet les nombres $4 \sin^2(k\pi/(2n+2))$ comme valeurs propres (pour k variant de 1 à n). Cette matrice étant symétrique, son conditionnement est le rapport de la plus grande à la plus petite des valeurs propres, c'est-à-dire que pour n grand ce conditionnement est de l'ordre de $4n^2/\pi^2$.

Or cette matrice joue un rôle important dans la résolution approchée de l'équation $\frac{\partial^2 u}{\partial t^2} + f(t) = 0$ comme on peut le voir en annexe.

QUELLE SOLUTION ?

La première idée consiste à remplacer le système proposé $AX = B$ par le système $CAX = CB$ avec C inversible et CA bien conditionnée. Malheureusement le calcul du produit CA dans le cas général et pour des matrices de très grandes dimensions coûte cher en temps machine. De plus C va dépendre de A et doit être recalculé dans chaque exemple ce qui accroît le temps de calcul. C'est pourquoi on préfère la méthode suivante : on écrit la matrice A sous la forme $M - N$ où M est une matrice facilement inversible (diagonale ou triangulaire). L'équation $AX = B$ se transforme en $MX = NX + B$ ou encore $X = (M^{-1}N)X + (M^{-1}B)$. On considère alors la suite récurrente de vecteurs : $X_n = (M^{-1}N)X_{n-1} + (M^{-1}B)$ qui, moyennant certaines précautions, converge vers la solution. On prend très généralement pour M soit la diagonale de A , soit le triangle inférieur gauche de A .

CAS DES MATRICES DE DÉTERMINANT NUL

Il n'est pas question ici de chercher un conditionnement qui est évidemment infini. Mais le but est en général autre comme le montre l'exemple 3. Il s'agit de trouver un espace propre. Quand les matrices qui interviennent ne sont pas trop grossesen taille, le plus simple consiste à supprimer une équation et à résoudre le système ainsi réduit ; c'est bien ainsi que l'on opère « à la main ». Bien évidemment il vaut mieux choisir correctement la ou les équations que l'on supprime pour être sûr de trouver tout l'espace propre cherché.

Ce principe ne s'applique pas dès que la taille des matrices est importante. On utilise alors des méthodes itératives qui permettent d'atteindre en même temps et de façon indépendante les vecteurs propres et les valeurs propres correspondantes. Il n'est pas question d'aborder ici l'étude de ces méthodes et de leur précision.

§ 2 – CALCUL DES VALEURS D' UNE FONCTION

GÉNÉRALITÉS

Calculer la valeur d'une fonction pour un choix particulier de la variable ne semble pas poser de difficultés particulières surtout qu'avant les études supérieures les élèves ne rencontrent guère de fonctions non explicites, et pourtant les fonctions trigonométriques, les fonctions exponentielles, les fonctions logarithmiques sont autant de fonctions pour lesquelles on devrait se poser des problèmes de tabulation, mais peu de personnes (y compris parmi les enseignants) savent comment ces fonctions ont été tabulées ou comment elles sont implantées sur les calculatrices ou les ordinateurs.

Un simple polynôme ou une fraction rationnelle révèle cependant des surprises. Les exemples 5, 6 et 7 donnés au début de l'article montrent bien certaines des difficultés qui peuvent apparaître.

Comme dans le cas des systèmes linéaires, on peut se poser la question de savoir quelle est la variation relative de $f(x)$ pour une variation relative donnée de x . Or la formule de Taylor ne permet de répondre que pour des variations absolues : $f(x) = f(x_0) + (x - x_0) f'(x_0) + o(x - x_0)$, ce qui montre que « l'erreur » sur la fonction est égale à « l'erreur » sur la variable multipliée par $|f'(x_0)|$. Mais souvent, ce qui est intéressant, c'est la façon dont se propage « l'erreur rela-

tive » qui est donnée par $\frac{f(x)-f(x_0)}{f(x_0)} = \frac{x-x_0}{x_0}\frac{x_0 f'(x_0)}{f(x_0)}$ à un $o(x-x_0)$ près ce qui fait que c'est la quantité $\frac{x_0 f'(x_0)}{f(x_0)}$ qui va mesurer le « conditionnement » de la fonction f.

Ceci ne tient pas compte des erreurs d'arrondi qui peuvent intervenir au cours des calculs. Il est connu de tous que le calcul sur calculatrice de $(1 + 10^{-15} - 1)/10^{-15}$ fournit la réponse 0 tandis que celui de $(1 - 1 + 10^{-15})/10^{-15}$ donne la réponse 1 . Ce qui montre encore une fois que l'algorithme utilisé joue un rôle important dans la précision du résultat.

REPRÉSENTATION GRAPHIQUE

Tous les logiciels qui proposent le tracé graphique d'une fonction commence par tabuler la fonction en des points convenablement choisis puis ils relient les points obtenus par une courbe ou un segment [3] . Le plus souvent il s'agit de points régulièrement répartis à l'intérieur d'un intervalle qui est défini automatiquement ou bien que l'utilisateur peut préciser. Dans les deux cas, mais surtout dans le premier le logiciel n'aboutit à rien si la fonction présente des variations trop rapides. C'est ainsi que les exemples 5 et 6 ne donnent rien de bien probant sur la plupart des calculatrices graphiques et même avec un logiciel de calcul formel comme « Dérive ». Par contre, avec un tel logiciel, une fois étudiées les variations de la fonction, il est tout à fait possible d'obtenir un tracé acceptable en précisant une fenêtre de taille correcte. Ainsi dans l'exemple 5 il faut choisir x entre -3 à $+3$ et y entre -10^5 à $+3.\ 10^5$. Mais la fonction de l'exemple 6 ou même l'inverse de la fonction de l'exemple 5 sont plus rebelles. Cherchons à comprendre pourquoi.

Le polynôme $12192\,X^3 - 32257\,X^2 - 85344\,X + 225799$ se factorise en $(12192\,X - 32257)(X^2 - 7)$ c'est-à-dire que ses racines sont $\pm\sqrt{7}$ soit environ $\pm$ 2,645 751 311 06... et un nombre rationnel 32257 / 12192 voisin de 2,645 751 312 33... Deux des racines sont donc séparé d'un peu plus de 10^{-9} . Entre ces deux nombres, la fonction atteint un minimum relatif de l'ordre de $-2{,}613\ 10^{-14}$. Il est très difficile, voire impossible de mettre en évidence une telle situation graphiquement (c'est-à-dire même avec un fort agrandissement de la région considérée). Il est alors facile de croire à l'existence d'une racine double.

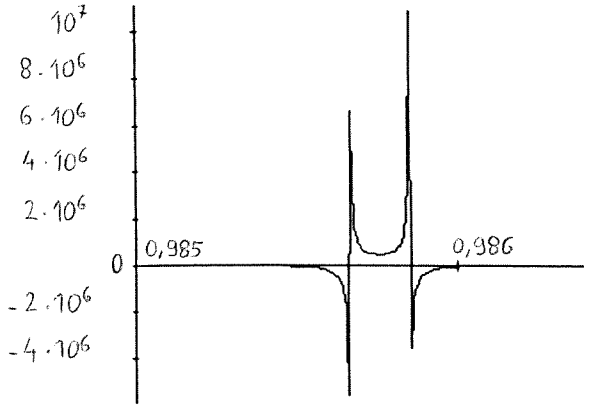
Considérons alors la fraction rationnelle correspondant à l'inverse de ce polynôme. La courbe représentative admet trois asymptotes verticales dont deux sont séparées d'environ 10^{-9} . Entre ces deux asymptotes, la courbe admet un maximum d'ordonnée $-3{,}827\ 10^{13}$, valeur pratiquement inaccessible quelque soit l'échelle adoptée sur chacun des axes.

Il est facile de multiplier de tels exemples en utilisant des réduites convenables des racines carrées d'entiers (par exemple 99/70 pour $\sqrt{2}$, etc.).

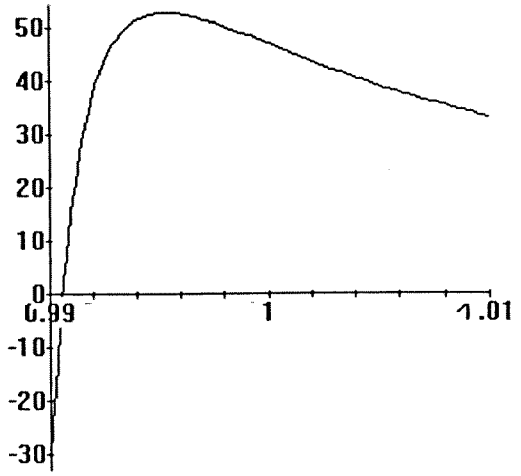
La fonction de l'exemple 6 présente les mêmes vicissitudes puisque l'on a : $f(x) = \frac{4970x-4923}{4970x^2-9799x+4830} = \frac{1680}{70x-69} - \frac{1633}{71x-70}$ ce qui montre que les deux asymptotes verticales sont situées aux environs de 0,9857 et 0,9859 et qu'entre ces deux asymptotes, la courbe présente un minimum dont les coordonnées valent environ (0,9858 ; 467127) , alors qu'il existe un maximum de coordonnées voisines de (0,9953 ; 52,85) c'est-à-dire à droite des deux asymptotes. Il est donc quasiment impossible de dessiner sur un même graphique cartésien l'ensemble de la courbe. On trouvera ci-après deux extraits, à des échelles différentes,

[3] Nous ne cherchons pas ici à analyser la gestion des asymptotes verticales. L'une des méthodes utilisées consiste à ne pas relier deux points dont la différence des ordonnées est trop grande.

obtenus à l'aide du logiciel « maple » et qui permettent de se rendre compte du phénomène ; il semble difficile de faire mieux.



x varie de 0,985 à 0,9865



TABULER UNE FONCTION

Ceux qui sont suffisamment âgés pour avoir fait du calcul numérique à l'aide de tables de logarithmes se souviennent que le calcul de $\log(\sin x)$ et $\log(\tan x)$ ne se faisait pas de la même façon selon que x était supérieur ou non à 4° . Dans le premier cas, on trouvait le nombre voulu directement dans la table (ou par interpolation linéaire) ; dans le second cas il fallait recourir aux quantités $\log\left(\frac{\sin x}{x}\right)$ et $\log\left(\frac{\tan x}{x}\right)$ où x était exprimé en minutes d'arc, quantités qui étaient tabulées à part. Cet artifice avait pour but de pallier la variation trop rapide des fonctions $\log(\sin x)$ et $\log(\tan x)$ au voisinage de 0 (présence d'une asymptote verticale).

La plupart des fonctions étudiées sont dérivables et par suite lipschitziennes sur des intervalles fermés convenables. Mais rien n'est dit sur la valeur de la constante et une fonction, même dérivable sur tout $\mathbf{R}$, peut présenter localement des variations très brutales qui remettent en cause la possibilité d'effectuer une interpolation linéaire entre deux points consécutifs.

a) **Le schéma de Horner** : Cette méthode, très en vogue dans les programmes actuels, minimise le nombre d'opérations (essentiellement des multiplications) dans le calcul des valeurs d'un polynôme. Elle n'assure pas la meilleure précision possible compte tenu du matériel utilisé. Pour s'en rendre compte étudions les résultats partiels dans l'exemple 5 :

opérations effectuées	résultats pour $x = 2{,}645\,752$
$A = 12192\,x - 32257$	A = 0,008 384
$B = A \times x$	B = 0,022 181 984 768
$C = B - 85344$	C = – 85 343,977 818 015 232
$D = C \times x$	D = – 225 798,999 999 969 436 094 464
$E = D - 225799$	E = 0,000 000 030 563 905 536

On remarque que les résultats intermédiaires sont beaucoup trop grands (en particulier C et D) devant le résultat final qui est obtenu par soustraction et qu'il y a ainsi une perte de précision tout à fait catastrophique. On peut améliorer la précision du calcul sur machine en divisant le polynôme par $12192\,x$ avant tout calcul, pour remultiplier par cette même quantité à la fin. Mais c'est un palliatif difficile à mettre en oeuvre dans le cas général comme on peut s'en rendre compte dans l'étude de l'exemple 7 en effectuant le calcul selon divers regroupements.

b) **La précision relative** : Nous avons vu dans les généralités que ce qui est important c'est souvent la précision relative du résultat. Il est en effet assez futile d'obtenir une précision absolue de 10^{-5} sur un résultat de l'ordre de 10^{-8} ! Or la précision relative est gérée par la quantité $\dfrac{x_0 f'(x_0)}{f(x_0)}$ dont on voit bien qu'elle risque d'être grande (donc mauvaise) non seulement quand $f'(x_0)$ est grand, mais aussi quand x_0 est voisin d'une racine de f. C'est exactement ce qui se passe dans le cas de l'exemple 5 où cette quantité vaut à peu près $7{,}687\ 10^6$. Un exemple encore plus simple est donné par l'étude de la variation de $x^2 - 2$ entre $x = 1{,}414$ et $x = 1{,}415$ où $f(x)$ passe de la valeur $-0{,}000\,604$ à la valeur $0{,}002\,225$. On devra donc faire un calcul permettant d'évaluer l'erreur dès que le résultat paraîtra très faible

CALCUL APPROCHÉ DE LIMITES

C'est un exercice classique que de deviner la valeur d'une limite en utilisant la calculatrice. L'expérience montre qu'il ne faut pas prendre x trop près de x_0 dans le calcul approché de $\lim_{x \to x_0} f(x)$ faute de quoi le résultat risque de ne pas correspondre à la réalité. C'est ce qui se passe dans les exemples 6 et 8 donnés au début de l'article. Cela provient du fait que pour $h = x - x_0$ trop petit les erreurs d'arrondi finissent par être du même ordre de grandeur que le résultat. Étudions de ce point de vue les deux exemples proposés.

Dans l'exemple 6, il y a d'abord l'erreur due à la méthode qui peut être évaluée à l'aide d'une formule de Taylor. On trouve $f^{IV}(1).h^2/12$ soit ici $-2\,320\,950\,146\,h^2$. On voit donc que h doit être très petit (de l'ordre de 10^{-6}) si on veut obtenir une précision acceptable sur le calcul de $f''(1)$. Mais c'est alors que l'erreur due aux arrondis devient très grande puisque, d'une part $f'(1)$ vaut -1657 et d'autre part au cours des calculs la machine est amenée à soustraire deux grands nombres pour obtenir le résultat.

Dans l'exemple 8, le calcul de $\exp(-30)$ se fait à l'aide de la série « exponentielle » ; cette série est alternée (dans le cas présent) et les termes croissent en valeur absolue jusqu'à $(-30)^{30}/(30\ !)$ qui est de l'ordre de $7{,}762\ 10^{11}$ pour décroître ensuite. Comme le résultat final vaut environ $9{,}36\ 10^{-14}$, le résultat qui sera affiché par la machine sera essentiellement la somme des erreurs d'arrondi, aussi loin qu'on puisse aller dans le développement en série de l'exponentielle. Ici le terme d'ordre 10^{-14} est atteint vers $n = 107$ et le terme d'ordre 10^{-16}, vers $n = 111$. On voit bien ici que l'algorithme utilisé n'est pas bon. Ce n'est d'ailleurs pas celui qui est utilisé dans les calculatrices.

CONCLUSION PROVISOIRE

Si le premier vol de la navette spatiale américaine n'a eu lieu qu'au printemps 1981 avec près de deux ans de retard sur le projet initial c'est en particulier en raison des difficultés de l'installation d'une protection thermique. Il a fallu trouver un compromis entre une configuration facilitant les manœuvres lors du retour où la navette évolue comme un planeur ne disposant d'aucun moyen de propulsion et une configuration permettant de limiter le flux de température au cours du freinage aérodynamique (la vitesse passe de 28 000 à 300 km/h en une demi-heure). Lors d'un des premiers essais, la navette s'est posée avec un gros trou sur le

bord d'attaque de la dérive. Fort heureusement aucun organe vital ne se trouvait derrière ce trou.

Pourquoi cette mauvaise prévision. Tout simplement parce que le flux thermique est régi par des équations aux dérivées partielles qu'on ne sait résoudre que par des méthodes numériques. On se contente donc de telles méthodes en calculant des valeurs aux nœuds d'un maillage que l'on espère assez serré. Ce ne fut pas le cas car la température variait de façon trop rapide dans la zone où se forma le trou, atteignant près de 1300°.

Cette mésaventure montre bien l'importance de l'évaluation des erreurs dans tout calcul approché. Puisse cela inciter les professeurs à mettre en garde leurs élèves face à un recours trop systématique à la calculatrice en attendant de réhabiliter un certain type de calcul d'erreur

§ 3 – ANNEXES

ANNEXE 1 : CALCUL DE LA NORME D'UNE MATRICE

On utilise le fait que $\|X\| = {}^tX\,X$ (plus exactement dét$({}^tX\,X)$ mais on identifie les matrices 1×1 avec les scalaires). On peut alors écrire : $\|A\| = \sup\dfrac{\|AX\|}{\|A\|} = \sup\sqrt{\dfrac{{}^tX\,{}^tAAX}{{}^tXX}} = \sqrt{\sup\dfrac{{}^tX\,{}^tAAX}{{}^tXX}}$.

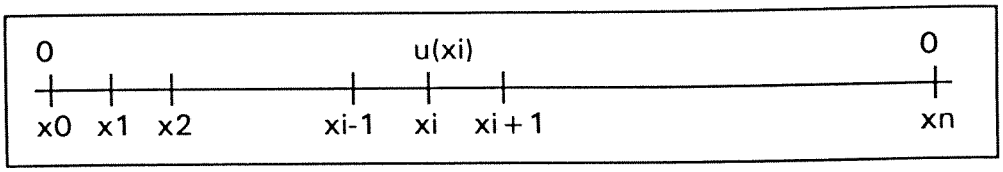
Or ${}^tA\,A$ est une matrice symétrique qui admet donc une base de vecteurs propres orthonormés (e_i) associés aux valeurs propres respectives (μ_i) avec $0 < \mu_i \le \mu_j$ si $i < j$. Les μ_i sont strictement positifs car $\|Ae_i\| = {}^te_i\,{}^tA\,A\,e_i = \mu_i$. Si on écrit $X = \Sigma a_i\,e_i$ on obtient ${}^tX\,{}^tA\,A\,X = \Sigma a_i^2\mu_i$ et ${}^tX\,X = \Sigma a_i^2$ ce qui permet d'affirmer que $\dfrac{\|AX\|^2}{\|X\|^2} = \dfrac{\sum a_i^2\,\mu_i}{\sum a_i^2} \le \mu_n$ la plus grande des valeurs propres et μ_n est manifestement atteinte(par le vecteur e_n) et par suite $\|A\| = \sqrt{\mu_n}$.

Comme la matrice A est inversible, on sait que pour tout X il existe un unique Y tel que $X = AY$ et on peut écrire $\|A^{-1}\|^2 = \sup\limits_X \dfrac{\|A^{-1}X\|}{\|X\|} = \sup\limits_Y \dfrac{\|Y\|}{\|AY\|} = \dfrac{1}{\inf\limits_Y \dfrac{\|AY\|}{\|Y\|}} = \dfrac{1}{\mu_1}$ où μ_1 est la plus petite des valeurs propres précédemment mises en évidence (le raisonnement est analogue).

En regroupant les deux résultats ci-dessus, on a le résultat $\mathbf{cond}(A) = \sqrt{(\mu_n/\mu_1)}$ où μ_n et μ_1 sont respectivement la plus grande et la plus petite valeur propre de ${}^tA\,A$.

ANNEXE 2 : RESOLUTION APPROCHEE DE $\dfrac{\partial^2 u}{\partial t^2} + f(t) = 0$

C'est le problème du mouvement d'une corde vibrante attachée à ses deux extrémités et dont on connaît l'accélération en tout point.



Découpons le segment représentant la corde en n morceaux égaux de longueur h dont les extrémités sont $x_0, x_1, \ldots x_n$. Les formules de Taylor nous permettent d'écrire :

$$u(x_{i+1}) = u(x_i) + h\,u'(x_i) + \tfrac{1}{2}h^2\,u''(x_i) + o(h^2)$$
$$u(x_{i-1}) = u(x_i) - h\,u'(x_i) + \tfrac{1}{2}h^2\,u''(x_i) + o(h^2)$$

En additionnant membre à membre et en négligeant le $o(h^2)$, il vient :

$$u(x_{i-1}) - 2\,u(x_i) + u(x_{i+1}) = h^2\,u''(x_i) = -h^2 f(x_i)$$

Comme on suppose que $u(x_0) = u(x_n) = 0$, puisque la corde est attachée à ses deux extrémités, on peut écrire les différentes égalités ci-dessus obtenues en faisant varier i de 1 à $n-1$ sous la forme : $\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_{n-1}) \end{bmatrix} = h^2 \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \end{bmatrix}$. On a bien mis en évidence cette matrice tridiagonale.

On comprend bien que pour obtenir une précision importante il faut diminuer la valeur de h mais ce faisant on augmente la taille de la matrice et donc son conditionnement. Il semble naturel de penser qu'il existe une valeur de h optimale qui dépend de la précision avec laquelle est connue la fonction f.

HISTOIRE D'ALGORITHMES

Du caillou à la puce

Jean-Luc CHABERT

Évelyne BARBIN, Michel GUILLEMOT, Anne MICHEL-PAJUS,

Jacques BOROWCZYK, Ahmed DJEBBAR, Jean-Claude MARTZLOFF

Au sommaire :

Algorithmes des opérations arithmétiques – Les carrés magiques – Autour des méthodes de fausse position – Autour de l'algorithme d'Euclide – De la mesure du cercle au calcul de π – Les méthodes de Newton – Résolutions d'équations par approximations successives – Des algorithmes de l'arithmétique – Résolution de systèmes d'équations linéaires – Tables et interpolation – Quadratures approchées – Résolutions approchées d'équations différentielles – Approximation de fonction – Accélération de convergence.

Voir un extrait p. 52.

(Dans la collection "*Regards sur la science*", aux éd. Belin (1993.)