ANALYSE NUMÉRIQUE

Prof. HAMRI Nasr-eddine
Département de Mathématiques
Université Abdelhafid Boussouf - Mila

TABLE DES MATIÈRES

Chapitre

1

NOTIONS D'ERREURS

Sommaire

1 PRÉLIMINAIRES

L'outil fondamental en analyse numérique (qui nous fournit des méthodes de calcul pour l'étude et la solution approchée de problè mes mathématiques dont la résolution est généralement impossible ou impraticable), demeure la formule de Taylor. Ces solutions approchées sont le plus souvent calculées sur ordinateur au moyen d'algorithmes convenables. Dans ce qui suit nous rappelons quelques théorèmes dont la connaissance est impérative pour une meilleure compréhension de la suite.

Théorème 1 (de la valeur intermédiaire). Soit f une fonction définie sur un intervalle $[a,b]$, on définit $m = \inf_{x\in[a,b]} f(x)$ et $M = \sup_{x\in[a,b]} f(x)$. Alors pour tout y dans $[m,M]$, il existe au moins un point x dans $[a,b]$ pour lequel

$$f(x) = y.$$

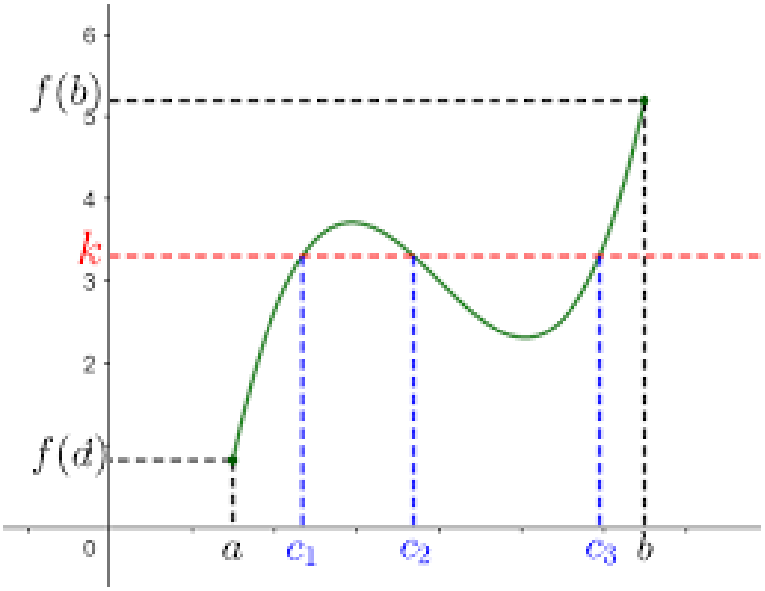

FIGURE 1.1 – *Illustration du théorème de la valeur intermédiaire.*

Théorème 2 (des accroissements finis). Soit f une fonction définie et continue sur un intervalle $[a,b]$, différentiable sur $]a,b[$. Alors il existe au moins un point c dans $[a,b]$ pour lequel

$$f(b) - f(a) = f'(c)(b-a).$$

Théorème 3 (de la moyenne). Soit $w(x)$ une fonction non négative définie et intégrable sur un intervalle $[a,b]$, et soit $f(x)$ une fonction continue sur $]a,b[$. Alors

$$\int_a^b w(x)f(x)dx = f(\xi)\int_a^b w(x)dx$$

pour $\xi \in [a,b]$.

Théorème 4 (de Taylor). Soit f une fonction définie et continue sur un intervalle $[a,b]$, $(n+1)$ fois dérivable sur $[a,b]$ pour $n \geq 0$, et soit $x, x_0 \in [a,b]$. Alors

$$f(x) = p_n(x) + R_{n+1}(x). \tag{1.1}$$

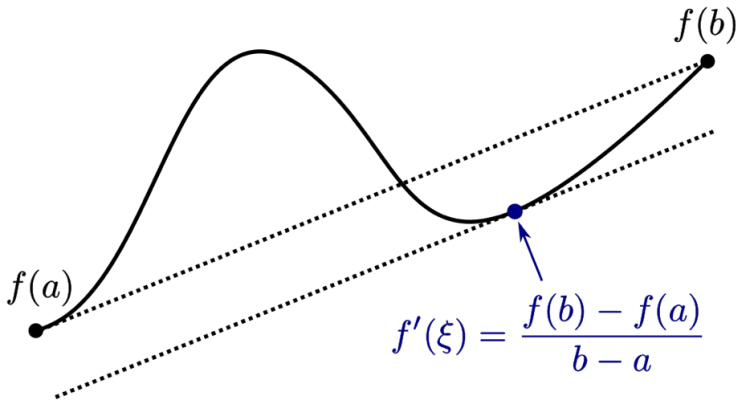

FIGURE 1.2 – *Illustration du théorème des accroissements finis.*

Où

$$\begin{aligned} p_n(x) &= f(x_0) + \frac{(x-x_0)}{1!} f'(x_0) \\ &\quad + \ldots + \frac{(x-x_0)^n}{n!} f^n(x_0) \end{aligned}$$

Et

$$\begin{aligned} R_{n+1}(x) &= \frac{1}{n!} \int_{x_0}^{x} (x-t)^n f^{(n+1)}(t) dt \\ &= \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi) \end{aligned} \tag{1.2}$$

pour $\xi \in]x_0, x[$

En utilisant la formule de Taylor on obtient par exemple les formules suivantes :

$$e^x = 1 + x + \frac{x^2}{2} + \ldots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} e^{\xi_x} \tag{1.3}$$

$$\begin{aligned} \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots + (-1)^n \frac{x^{2n}}{2n!} + \\ &\quad + (-1)^{n+1} \frac{x^{2n+2}}{(2n+2)!} \cos(\xi_x) \end{aligned} \tag{1.4}$$

$$(1-x)^{-1} = 1 + x + x^2 + \ldots + x^n + \frac{x^{n+1}}{1-x}, \quad x \neq 0 \tag{1.5}$$

De cette dernière formule nous pouvons déduire :

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \quad |x| < 1 \tag{1.6}$$

On peut calculer les séries de Taylor de n'importe quelle fonction suffisamment dérivable avec autant de termes que l'on veut. Cependant à cause de la complexité de la différentiation de plusieurs fonctions, il est souvent préférable d'obtenir indirectement leur polynôme d'approximation

de Taylor $p_n(x)$ ou leur séries de Taylor, en utilisant l'un des développement limités connus. Les trois exemples qui suivent montrent que les erreurs sont plus simples que lorsque l'on utilise la formule de l'erreur (1.2).

1.1 Exemples

1. $f(x) = e^{-x^2}$, En remplaçant x par $-x^2$ dans (1.3), on obtient :

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2} - \ldots + (-1)^n \frac{x^2 n}{n!} + (-1)^{n+1} \frac{x^{2n+2}}{(n+1)!} e^{\xi_x}$$

avec $\xi_x \in [-x^2, 0]$.

2. $f(x) = \frac{1}{\tan(x)}$, En posant $x = -u^2$ dans le développement de $\frac{1}{1-x}$ on a :

$$\frac{1}{1+u^2} = 1 - u^2 + u^4 - \ldots + (-1)^n u^{2n} + (-1)^{n+1} \frac{u^{2n+2}}{1+u^2}$$

en intégrant sur $[0, x]$ on aboutit à :

$$\frac{1}{\tan(x)} = x - \frac{x^3}{3} + \frac{x^5}{5} - \ldots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^n \int_0^x \frac{u^{2n+2}}{1+u^2} du$$

En appliquant le théorème de la moyenne, on obtient :

$$\int_0^x \frac{u^{2n+2}}{1+u^2} du = \frac{x^{2n+3}}{2n+3} \cdot \frac{1}{1+\xi_x^2}$$

avec $\xi_x \in [0, x]$.

3. $f(x) = \int_0^1 \sin(xt) dt$, Utilisant le développement de $\sin x$, et intégrant, on écrit :

$$f(x) = \sum_{j=1}^{n} (-1)^{j-1} \frac{x^{2j-1}}{2j!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \int_0^1 t^{2n+1} \cos(\xi_{xt}) dt$$

avec $\xi_{xt} \in [0, xt]$. L'intégrale dont le reste est bornée par $\frac{1}{2n+2}$, mais on peut aussi la mettre sous une forme simplifiée, et en appliquant le théorème de la moyenne on a :

$$\int_0^1 \sin(xt) dt = \sum_{j=1}^{n} (-1)^{j-1} \frac{x^{2j-1}}{2j!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \cos(\xi_x)$$

avec $\xi_x \in [0, x]$.

2 Erreurs absolues et Erreurs relatives

Un nombre approché x est légèrement différent du nombre exact X, et qui dans les calculs remplace X.

- Si $x < X$, x est dit valeur par *défaut* .
- Si $x > X$, x est dit valeur par *excès* .

On note généralement $x \approx X$.

2.1 Exemple

$$1,41 < \sqrt{2} < 1,42$$

Définition 5. On appelle erreur Δx d'un nombre approché, la valeur :

$$\Delta x = X - x,$$

C'est-à-dire

$$X = x + \Delta x$$

Définition 6. On appelle erreur absolue Δ d'un nombre x la valeur

$$\Delta = \mid X - x \mid \tag{2.1}$$

Remarque 7.
1. Si X est connu, l'erreur absolue est déterminée par (2.1).
2. Si X est inconnu, l'erreur absolue Δ est impossible à déterminer. Dans ce cas on introduit la limite supérieure de l'erreur absolue.

Définition 8. On appelle borne supérieure de l'erreur absolue tout nombre supérieur ou égal à l'erreur absolue de ce nombre. C'est-à-dire :

$$\Delta = \mid X - x \mid \leq \Delta_x$$

si Δ_x désigne la borne supérieure, donc

$$x - \Delta_x \leq X \leq x + \Delta_x$$

On note

$$X = x \pm \Delta_x.$$

2.2 Exemples

Trouver la borne supérieure d'erreur absolue de $\pi = 3,14$.

$$3,14 \leq \pi \leq 3,15$$

Donc $\mid x - \pi \mid \leq 0,01$, on peut poser $\Delta_x = 0,01$, comme $3,140 \leq \pi \leq 3,142$, une meilleure estimation de la borne d'erreur absolue est $\Delta = 0,002$.

Définition 9. On appelle erreur relative notée δ d'un nombre x, le rapport suivant :

$$\delta = \frac{\Delta}{\mid X \mid}, \quad (X \neq 0)$$

C'est-à-dire

$$\Delta = \mid X \mid \delta.$$

Définition 10. La borne supérieure de l'erreur relative δ_x est un nombre supérieur ou égal à l'erreur relative de ce nombre. C'est-à-dire :

$$\delta \leq \delta_x$$

c'est-à-dire

$$\frac{\Delta}{\mid X \mid} \leq \delta_x,$$

donc

$$\Delta \leq \mid X \mid \delta_x$$

On peut aussi utiliser

$$\Delta_x = \mid x \mid \delta_x$$

car $X \approx x$.

Remarque 11. Si l'on connait une borne d'erreur relative δ_x on a :

$$X = x(1 \pm \delta_x)$$

C'est-à-dire

$$\delta_x = \frac{\Delta_x}{x - \Delta_x}.$$

De même on obtient :

$$\Delta_x = \frac{x\delta_x}{1 - \delta_x}.$$

Exemple

En cherchant la constante de gaz, on a obtenu $R = 29,25$, l'erreur relative étant $1^0/_{00}$ trouver un encadrement de R. On a $\delta_x = 0,001$, donc $\Delta_x = R\delta_x = 0,03$, c'est-à-dire :

$$29,22 \leq R \leq 29,28.$$

3 PRINCIPALES SOURCES D'ERREURS

Les erreurs commises dans les problèmes peuvent ê tre des :

Erreurs inhérentes au problème : Erreurs dûes à la position même du problème. Le modèle théorique est trés rarement fidèle au modèle réel. Lors de l'étude d'un phénomène de la nature on est souvent contraint d'admettre certaines conditions.

Erreurs de la méthode : Il arrive qu'il soit difficile ou mê me impossible de résoudre un problème enoncé en termes exacts. On le remplace par un problème approché.

Erreurs de troncature : Associées aux processus infinis. Les fonctions données dans les formules le sont sous forme de suites infinies ou de séries, on est donc obligé de mettre fin à un certain terme de la suite. Par exemple, l'approximation d'une somme infinie par une somme finie, l'approximation de la limite d'une suite par un terme de "grand indice" ou encore l'approximation d'une inté grale par une somme finie.

Erreurs initiales : Dûes à la présence dans les formules de paramètres dont les valeurs sont approchées.

Erreurs d'arrondi : Dûes au système de numérisation.

Erreurs propagées : Les erreurs des données de départ se repercutent sur le résultat des calculs.

4 PRECISION, CHIFFRES SIGNIFICATIFS

4.1 Chiffres significatifs

Définition 12. On appelle chiffre significatif d'un nombre tous les chiffres de son écriture à partir du premier chiffre différent de zéro à gauche.

Exemple Les chifres significatifs des nombres $x = 0,0\underline{3045}$ et $x = 0,0\underline{3045000}$ sont ceux soulignés. Ils sont 4 chiffres dans le premier cas et 7 dans le deuxième.

Définition 13. Un chiffre significatif est dit *exact* si l'erreur absolue sur le nombre ne dépasse pas l'unité de l'ordre correspondant.

Exemple $x = 0,0\underline{3045}$, $\Delta(x) = 0,000003$; $x = 0,0\underline{3045}00$, $\Delta(x) = 0,0000007$; les chiffres soulignés sont exacts.

Définition 14. Si tous les chiffres significatifs sont *exacts*, on dit que le nombre est écrit avec tous les chiffres exacts.

Exemple $x = 0,03045$, $\Delta(x) = 0,000003$, x, est écrit avec 4 chiffres exacts. On parle souvent de *décimales exactes*. Dans le dernier exemple le nombre x est écrit avec 5 décimales exactes.

4.1.1 Règle pour arrondir les nombres

Soit x un nombre approché sous forme décimale. Pour l'arrondir jusqu'à n chiffres significatifs, c'est-à-dire le remplacer par un nombre x_1 [1] avec n chiffres significatifs, on rejette tous les chiffres à droite du $n^{ième}$ chiffre significatif ou s'il faut conserver les rangs, on les remplace par des zéros. Dans ces cas :

— Si le premier des chiffres rejetés est inférieur à 5, les chiffres restent inchangés.
— Si le premier des chiffres rejetés est supérieur à 5, on ajoute une unité au dernier chiffre restant.
— Si le premier des chiffres rejetés est égal à 5, et si parmi les autres chiffres rejetés il y en a des non nuls, le dernier chiffre restant est augmenté de l'unité.
• Mais si le premier des chiffres rejetés est égal à 5 alors que les autres chiffres rejetés sont nuls, le dernier chiffre conservé reste inchangé s'il est pair ou on lui ajoute une unité s'il est impair.

Exemple En arrondissant le nombre

$$x = 3,045166382535$$

jusqu'à 5 ; 4 et 3 chiffres significatifs, on obtient les nombres approchés $3,0452$, $3,045$ et $3,05$

5 Cumulation des erreurs d'arrondi

5.1 Erreurs d'arrondi sur une somme

Soient X, Y des nombres réels supposés représentés sans erreur avec N chiffres significatifs :

$$\begin{array}{ll} X = 0,a_1a_2\dots a_N.b^p, & b^{-1+p} \leq X < b^p \\ Y = 0,a'_1a'_2\dots a'_N.b^q, & b^{-1+q} \leq Y < b^q \end{array}$$

Et $\Delta(X+Y)$ l'erreur d'arrondi commise sur le calcul de $X+Y$. Supposons $p \geq q$.

— Si $X+Y < b^p$, le calcul de $X+Y$ s'accompagne de la perte des $p-q$ derniers chiffres de Y correspondants aux puissances $b^{-k+q} < b^{-N+p}$; donc $\Delta(X+Y) \leq b^{-N+p}$, alors que $X+Y \geq X \geq b^{-1+p}$.
— Si $X+Y \geq b^p$, la décimale correspondant à la puissance b^{-N+p} est elle aussi perdue, d'où $\Delta(X+Y) \leq b^{1-N+p}$.

Dans les deux cas :

$$\Delta(X+Y) \leq \varepsilon(|\,X\,|+|\,Y\,|),$$

Où $\varepsilon = b^{1-N}$ est la précision relative. Ceci est vrai quel que soit le signe de X et de Y. En général, les réels X, Y ne sont connus que par des valeurs approchées x, y avec des erreurs respectives $\Delta_x = |\,X-x\,|, \Delta_y = |\,Y-y\,|$. A ces erreurs s'ajoutent l'erreur d'arrondi :

$$\Delta(x+y) \leq \varepsilon(|\,x\,|+|\,y\,|+\Delta_x+\Delta_y).$$

Les erreurs Δ_x, Δ_y sont elles mêmes le plus souvent d'ordre ε par rapport à $|\,x\,|$ et $|\,y\,|$, de sorte que l'on pourra *"négliger"* les termes $\varepsilon\Delta_x$ et $\varepsilon\Delta_y$. On aura :

$$\Delta(x+y) \leq \Delta_x+\Delta_y+\varepsilon(|\,x\,|+|\,y\,|).$$

Remarque 15. Pour calculer une somme de réels positifs $\sum_{k=1}^{n} u_k$, on calcule les sommes partielles $s_k = u_1+u_2+\dots+u_k$ de proche en proche par les formules de récurrence :

$$\left\{ \begin{array}{lcl} s_0 & = & 0 \\ s_k & = & s_{k-1}+u_k, \quad k \geq 1 \end{array} \right.$$

1. le nombre x_1 est choisi de façon à minimiser l'erreur d'arrondi $|\,x_1-x\,|$.

Si les u_k sont connus exactement, on aura sur s_k des erreurs Δ_{s_k}telles que $\Delta_{s_1} = 0$ et

$$\Delta_{s_k} \leq \Delta_{s_{k-1}} + \varepsilon(s_{k-1} + u_k) = \Delta_{s_{k-1}} + \varepsilon s_k.$$

L'erreur globale sur s_n vérifie

$$\Delta_{s_n} \leq \varepsilon(s_2 + s_3 + \ldots + s_n).$$

5.2 Erreurs d'arrondi sur un produit

Le produit de deux mantisses de N chiffres donne une mantisse de $2N$ ou $2N-1$ chiffres dont les N ou $N-1$ derniers vont être perdus. Dans le calcul d'un produit XY il y aura donc une erreur d'arrondi

$$\Delta(xy) \leq \varepsilon \mid XY \mid, \quad \text{où} \quad \varepsilon = b^{1-N}.$$

Si X et Y ne sont connus que par des valeurs approchées x, y et si $\Delta_x = \mid X - x \mid$, $\Delta_y = \mid Y - y \mid$, on a une erreur initiale :

$$\mid XY - xy \mid = \mid X(y - Y) + (x - X)y \mid \leq \mid X \mid \Delta_y + \Delta_x \mid y \mid$$

A cette erreur s'ajoute une erreur d'arrondi :

$$\Delta(xy) \leq \varepsilon \mid xy \mid \leq \varepsilon(\mid x \mid + \Delta_x)(\mid y \mid + \Delta_y).$$

Ce qui donne la formule approximative :

$$\Delta(xy) \leq \mid x \mid \Delta_y + \Delta_x) \mid y \mid + \varepsilon \mid xy \mid .$$

Cette dernière formule nous permet d'obtenir par récurrence :

$$\Delta(x_1 x_2 \ldots x_k) \leq (k-1)\varepsilon \mid x_1 x_2 \ldots x_{k-1}.x_k \mid .$$

Remarque 16. La majoration de l'erreur d'un produit ne dépend pas de l'ordre des facteurs.

6 Représentation approchée des nombres réels

L'objet de cette section est de mettre en évidence les principales difficultés liées à la pratique des calculs numériques sur ordinateur. La capacité mémoire d'un ordinateur est par construction finie. Si X est un nombre réel, il est donc nécessaire de représenter X sous forme approchée. La notation la plus utilisée est la repré sentation avec *virgule flottante* :

$$X \simeq \pm m.b^p$$

Où b désigne la base de numération, m la mantisse, et p l'exposant. Les calculs internes sont généralement effectués en base $b = 2$, même si les résultats affichés sont finalement traduits en base 10. La mantisse m est un nombre écrit avec virgule fixe et possèdant un nombre maximum N de chiffres significatifs (impos é par la mémoire de l'ordinateur) : suivant les machines, m s'é crira

$$m = 0, a_1 a_2 \cdots a_N = \sum_{k=1}^{N} a_k b^{-k}, \qquad b^{-1} \leq m < 1.$$

Ceci entraine que la précision dans l'approximation d'un nombre réel est toujours une précision relative :

$$\frac{\Delta_x}{X} = \frac{\Delta_m}{m} \leq \frac{b^{-N}}{b^{-1}} = b^{1-N}.$$

On note $\varepsilon = b^{1-N}$ cette précision relative. **Exemple.** La même écriture peut représenter des nombres différents dans des bases différentes : $123,45$ en base 10 repr ésente le nombre $x = 1.10^2 + 2.10 + 3 + 4.10^{-1} + 5.10^{-2}$, en base 6 il représente le nombre $y = 1.6^2 + 2.6 + 3 + 4.6^{-1} + 5.6^{-2}$. D'autre part, le même nombre peut avoir un nombre fini de chiffres dans une base, et un nombre infini dans une autre base : $x = 1/3$ donne $x_3 = 0,1$ en base 3 et $x_{10} = 0,3$ en base 10.

6.1 Nombres en virgule flottante

De nombreuses manières ont été proposées pour repré senter les nombres par un ordinateur mais la plus utilisée aujourd'hui est donc la représentation dite en *"virgule flottante"* , et en base 2 (On utilise aussi la base 8 et la base 16 (numérotation hexadécimale à seize chiffres, de 0 à 9, auxquels on rajoute les lettres A,B,C,D,E et F)). La manière courante d'écrire les nombres aujourd'hui est la notation de position en base dix. On se donne donc dix symboles $0;1;2 = 1+1;3 = 2+1;...;9 = 8+1$. La représentation d'un nombre entier est simplement une suite finie de tels symboles. Par exemple 27821 n'est que le symbole 2 suivit du symbole 7,. . . Pour mettre en évidence l'aspect suite de symboles , nous écrirons :

2	7	8	2	1

Pour savoir à quel nombre correspond cette suite de symboles, on les interprète selon leur position. Ici la base est $b = 10 = 9+1$. Cela signifie que chaque fois qu'on se déplace d'un chiffre vers la gauche, la puissance de dix augmente de 1 :

$2[10^4]$	$7[10^3]$	$8[10^2]$	$2[10^1]$	$1[10^0]$

Ainsi, le nombre représenté est $2b^4 + 7b^3 + 8b^2 + 2b^1 + 1b^0 = 2.10^4 + 7.10^3 + 8.10^2 + 2.10^1 + 1.10^0$. Ceci c'était pour les nombres entiers. Qu'en est-il des écritures avec virgule ? Par exemple, que

représente $0,2$? Simplement, c'est $2/10 = 2.10^{-1}$. Et $1,74$ représente $1.10^0 + 7.10^{-1} + 4.10^{-2}$. De manière générale,

$\pm$	a_n	...	a_1	a_0	$,a_{-1}$	a_{-2}	...

représente le nombre $\pm(a_n 10^n + \cdots + a_1 10^1 + a_0 10^0 + a_{-1} 10^{-1} + a_{-2} 10^{-2} + \cdots)$, c'est-à -dire

$$\pm \sum_{i=-\infty}^{n} a_i 10^i.$$

Notons que la suite des décimales a_{-1}, a_{-2},...peut être finie ou infinie et que dans ce dernier cas la série converge. Représenter les nombres en base deux se fait exactement de la même manière

qu'en base dix excepté qu'on remplace dix par deux ! Ainsi, on se donne deux symboles 0 et 1. A une suite de tels symboles

$\pm$	a_n	...	a_1	a_0	$,a_{-1}$	a_{-2}	...

$a_i \in \{0,1\}$,

on fait correspondre le nombre

$$\pm \sum_{i=-\infty}^{n} a_i 2^i = \pm(a_n 2^n + \cdots + a_1 2 + a_0 + a_{-1} 2^{-1} + \cdots$$

Nous noterons ce nombre $(\pm a_n ... a_1 a_0, a_{-1} a_{-2} ...)_2$. On a donc $(110)_2 = 2^2 + 2^1 = 5$ et $(1,11)_2 = 2^0 + 2^{-1} + 2^{-2} = 1,75$.

6.2 Non-associativité des opérations arithmétiques.

Supposons par exemple que les réels soient calculés avec $N = 3$ chiffres significatifs et arrondis à la décimale la plus proche. Soient

$$x = 8,22 = 0,822.10, \qquad y = 0,00317 = 0,317.10^{-2}, \qquad z = 0,00432 = 0,432.10^{-2}.$$

On veut calculer la somme $x+y+z$. $(x+y)+z$ donne :

$$x+y = 8,22317 \simeq 0,822.10$$

$$(x+y)+z \simeq 8,22432 \simeq 0,822.10$$

$x+(y+z)$ donne :

$$y+z = 0,00749 \simeq 0,749.10^{-2}$$

$$x+(y+z) = 8,22749 \simeq 0,823.10$$

L'addition est donc non associative par suite des erreurs d'arrondi.

Remarque 17. En générale, dans une sommation de réels, l'erreur a tendance à être minimisée lorsqu'on somme en premier les termes ayant la plus petite valeur absolue.

6.3 Phénomènes de compensation.

Lorsqu'on tente d'effectuer des soustractions de valeurs très voisines, on peut avoir des pertes importantes de précision. **Exemple.** On veut résoudre l'équation $x^2 - 1634x + 2 = 0$ en effectuant les calculs avec $N = 10$ chiffres significatifs. On obtient

$$\Delta' = 667487, \qquad \sqrt{\Delta'} = 816,9987760,$$

$$x_1 = 817 + \sqrt{\Delta'} \simeq 1633,998776,$$

$$x_2 = 817 - \sqrt{\Delta'} \simeq 0,0012240.$$

On voit qu'on a une perte de 5 chiffres significatifs sur x_2. Ici le rem ède est simple : il suffit d'observer que $x_1.x_2 = 2$, d'où

$$x_2 = \frac{2}{x_1} = 1,223991125.10^{-3}.$$

C'est donc l'algorithme numérique utilisé qui doit être modifié.

7 SERIE D'EXERCICES

Exercice 18. Trouver une borne de l'erreur absolue du nombre $x = 3.14$ qui remplace $\pi(\pi = 3.1415926...)$ dans les deux cas suivants :

1. $3.14 < \pi < 3.142$.
2. $3.14 < \pi < 3.15$.

Exercice 19. Supposons que $x_1, x_2, x_3, ...x_n$ approchent respectivement $X_1, X_2, X_3, ..., X_n$ et que dans chaque cas la borne supérieure de l'erreur absolue est ε. Montrer que la borne supérieure de l'erreur de la somme des $x_i (i = 1, 2, ...n)$ est égale à $n\varepsilon$.

Exercice 20. Donner les bornes des erreurs relatives de $a = 1.414$ et $b = 1.41$ qui approchent $\sqrt{2} = 1.414214.....$

Exercice 21. En recherchant la constante des gaz de l'air on a obtenu $R \simeq 29.25$. La borne de l'erreur relative de cette valeur étant $1^0/_{00}$, trouver les limites entre lesquelles est comprise R.

Chapitre

2

APPROXIMATION

1 GÉNÉRALITÉS

Soit $X = \{x_i,\ \ i = 1, 2, \dots n\}$ un ensemble de n points appartenant à l'intervalle $[a, b] \subset \mathbb{R}$.

Supposons qu'à chaque x_i de X, on sache associer un $y_i \in \mathbb{R}$, résultat d'une expérience ou valeur donnée par une table, mais que cette opération soit impossible avec $x \in [a, b] \setminus X$.

Notons

$$\mathcal{D} = \{(x_i, y_i),\ \text{pour}\ \ x_i \in X\}.$$

On veut déterminer une fonction, *facilement calculable*, qui permette d'obtenir une estimation *"raisonnable"* de la réponse du phénomène pour la valeur x.

Notons g cette fonction et $\mathcal{G} = \{(x_i, g(x_i)), \text{pour}\ \ x_i \in X\}$. Si on définit une *"distance"* entre $\mathcal{D}$ et $\mathcal{G}$, nous pouvons chercher une fonction g, d'un type préalablement choisi (un polynôme par exemple), telle que cette distance soit minimale.

C'est ce qu'on appelle une **approximation**.

Un procédé général consiste à limiter la représentation de g aux combinaisons linéaires des $n+1$ fonctions d'une certaine base, choisies à priori.

Si on désigne par

$$u_1(x), u_2(x), u_3(x), \dots, u_{n+1}(x)$$

les fonctions de la base, les représentations utilisées seront les combinaisons

$$a_1 u_1(x) + a_2 u_2(x) + a_3 u_3(x) + \dots + a_{n+1} u_{n+1}(x) \tag{1.1}$$

qui dépendent des $n+1$ coefficients a_i que nous devrions calculer pour définir chaque approximation.

Remarque 22. On utilise un procédé analogue lorsque on étudie la représentation de fonctions par des séries de fonctions.

On prend une base infinie $u_1(x), u_2(x), \dots, u_{n+1}(x), \dots$ et on considère les séries

$$a_1 u_1(x) + a_2 u_2(x) + a_3 u_3(x) + \dots + a_{n+1} u_{n+1}(x) + \dots \tag{1.2}$$

L'equation (1.1) apparaît alors comme la somme partielle des $n+1$ premiers termes de (3.1).

2 APPROXIMATION

2.1 Meilleure approximation

Soit $(\mathbf{E}, d)$ un espace métrique et $\mathbf{F} \subset \mathbf{E}$. Chercher à approcher un élément f de $\mathbf{E}$ par un élément de $\mathbf{F}$, c'est donc déterminer g de $\mathbf{F}$ tel que :

$$d(f, g) = \min_{h \in \mathbf{F}} d(f, h) \tag{2.1}$$

S'il existe, cet élément sera appelé *meilleure approximation* de f dans $\mathbf{F}$ au sens de la distance d.

Théorème 23 (d'existence). Si **F** est une partie compacte de **E**, alors il existe au moins un élément g de **F** tel que :

$$d(f,g) = \min_{h\in\mathbf{F}} d(f,h).$$

Corollaire 24. Soit **E** un espace normé. Si **F** est un sous espace vectoriel de **E** de dimension finie, alors il existe au moins un élément g de **F** tel que :

$$\| f-g \| = \min_{h\in\mathbf{F}} \| f-h \|.$$

Remarque 25.

1. Nous supposerons dorénavant que **E** est l'ensemble des fonctions continues sur un intervalle $[a,b]$ de $\mathbb{R}$.
2. g est le plus souvent cherchée sous la forme suivante :

$$g(x) = a_1u_1(x) + a_2u_2(x) + a_3u_3(x) + \ldots + a_{n+1}u_{n+1}(x)$$

où les $u_i(x)$ sont des fonctions choisies dans :

— la classe des monômes :

$$1, x, x^2, \ldots, x^n.$$

— la classe des fonctions trigonométriques :

$$1, \sin x, \sin 2x, \ldots, \sin nx, \cos x, \cos 2x, \ldots, \cos nx.$$

— la classe des fonctions exponentielles :

$$1, \exp x, \exp 2x, \ldots, \exp nx$$

C'est-à-dire que :

—

$$g(x) = a_1x^n + a_2x^{n-1} + a_3x^{n-2} + \ldots + a_nx + a_{n+1},$$

ou ;

$$\begin{aligned} g(x) \quad = \quad & c + a_1\cos x + a_2\cos 2x + a_3\cos 3x + \ldots + a_n\cos nx + \\ & + b_1\sin x + b_2\sin 2x + b_3\sin 3x + \ldots + b_n\sin nx, \end{aligned}$$

ou ;

$$g(x) = a_1\exp b_1x + a_2\exp b_2x + a_3\exp b_3x + \ldots + a_n\exp b_nx,$$

ou aussi ;

$$g(x) = \frac{a_1x^n + a_2x^{n-1} + a_3x^{n-2} + \ldots + a_nx + a_{n+1}}{b_1x^n + b_2x^{n-1} + b_3x^{n-2} + \ldots + b_nx + b_{n+1}},$$

3. L'approximation de f par des polynômes, dite **approximation polynomiale**, est la plus fréquente.

Théorème 26. Si f est une fonction continue sur $[a,b]$, pour tout $\varepsilon > 0$, il existe un polynôme P_n de degré inférieur ou égal à n tel que :

$$\max_{x\in[a,b]} | f(x) - P_n(x) | < \varepsilon.$$

3 APPROXIMATION AU SENS DES MOINDRES CARRÉS

Définition 27. Soit **E** un espace vectoriel. Pour tout couple (f,g) de $\mathbf{E}\times\mathbf{E}$, on définit le produit scalaire noté $\langle f,g\rangle$ et la norme associée par :

$$\| f \| = \sqrt{\langle f,f\rangle}.$$

Pour tout $f\in\mathbf{E}$, il existe $g\in\mathbf{F}$ (**F** est un sous espace vectoriel de **E**, de dimension finie) tel que :

$$\| f-g \| = \min_{h\in\mathbf{E}} \| f-h \|.$$

Théorème 28. Une condition nécessaire et suffisante pour que g soit une meilleure approximation de f est que :

$$\langle f-g,h\rangle = 0 \qquad \text{pour tout} \quad h\in\mathbf{F} \tag{3.1}$$

g est appelée meilleure approximation de f au sens des moindres carrés.

Démonstration. La Condition est nécessaire : Soit g la meilleure approximation de f. Faisons un raisonnement par l'absurde. Supposons qu'il existe $h_1\in\mathbf{F}$ tel que :

$$\langle f-g,h_1\rangle = c\neq 0.$$

Soit $h_2\in\mathbf{F}$ défini par :

$$h_2 = g+\frac{c}{\| h_1 \|^2}h_1.$$

on a

$$\begin{aligned}\| f-h_2 \| &= \sqrt{\langle f-g-\frac{c}{\| h_1 \|^2}h_1, f-g-\frac{c}{\| h_1 \|^2}h_1\rangle}\\ &= \sqrt{\| f-g \|^2-\frac{c^2}{\| h_1 \|^2}} <\| f-g \|\end{aligned}$$

Ce qui est absurde.

La Condition est suffisante : Soit $g_1\in\mathbf{F}$ tel que : $\langle f-g_1,h\rangle = 0$ pour tout $h\in\mathbf{F}$ on a alors :

$$\begin{aligned}\| f-h \| &= \sqrt{\langle f-h,f-h\rangle}\\ &= \sqrt{\langle f-g_1-h+g_1, f-g_1-h+g_1\rangle}\\ &= \sqrt{\| f-g_1 \|^2+\| h-g_1 \|^2}.\end{aligned}$$

D'où $\| f-g_1 \|<\| f-h \|$ pour tout $h\in\mathbf{F}$ et donc $g_1=g$. cqfd

Théorème 29. La meilleure approximation au sens des moindres carrés est unique.

Démonstration. Soient g_1 et g_2 deux meilleures approximations de f on a donc :

$$\langle f-g_1,h\rangle = 0 = \langle f-g_2,h\rangle \quad \text{pour tout} \quad h\in\mathbf{F}$$

en particulier pour $h=g_1-g_2$, d'où

$$\begin{aligned}| \quad & | g_1-g_2 \| = \sqrt{\langle g_1-g_2+f-f, g_1-g_2\rangle} =\\ = \quad & \sqrt{\langle f-g_1,g_1-g_2\rangle+\langle f-g_2,g_1-g_2\rangle} = 0\end{aligned}$$

donc $g_1=g_2$. cqfd

4 CARACTÉRISATION

Il s'agit maintenant de construire le polynôme d'approximation qu'on note $Q_n(x)$ défini par :

$$Q_n(x) = a_1 + a_2 x + a_3 x^2 + \ldots + a_{n+1} x^n.$$

tel que la différence

$$\varepsilon(x) = f(x) - Q_n(x)$$

ait une norme aussi petite que possible.

Soit $\mathbf{P}_n$ l'ensemble des polynômes de degré inférieur ou égal à n. Et soit f une fonction continue sur un intervalle $[a,b]$. La meilleure approximation de f par un polynôme Q_n de $\mathbf{P}_n$ s'écrit :

$$Q_n(x) = b_1 u_n(x) + b_2 u_2(x) + \ldots + b_{n+1} u_0(x) = \sum_{i=1}^{n+1} b_i u_{n+1-i}(x).$$

(Où les u_i forment une base de $\mathbf{P}_n$, c'est-à-dire : $u_i(x) = x^i \quad i = 0, 1, \ldots, n$), et un polynôme quelconque P_n de $\mathbf{P}_n$ s'écrit :

$$P_n(x) = a_1 u_n(x) + a_2 u_2(x) + \ldots + a_{n+1} u_0(x) = \sum_{i=1}^{n+1} a_i u_{n+1-i}(x).$$

La condition nécessaire et suffisante du théorème s'écrit :

$$\langle f - \sum_{i=1}^{n+1} b_i u_{n+1-i}, \sum_{i=1}^{n+1} a_i u_{n+1-i} \rangle = 0$$

pour tout élément $(a_1, a_2, \ldots, a_n + 1)$ de $\mathbb{R}^{n+1}$. Ce qui donne le système d'équations :

$$\sum_{i=1}^{n+1} b_i \langle u_{n+1-i}, u_{n+1-j} \rangle = \langle f, u_{n+1-j} \rangle \quad j = 1, 2, \ldots, n+1 \tag{4.1}$$

qui admet une solution unique.

4.1 Norme

Remarque 30. Sur l'ensemble des fonctions continues $\mathcal{C}([a,b])$, on utilise le produit scalaire suivant : Soit ω une fonction positive n'ayant qu'un nombre fini de racines sur $[a,b]$ et telle que : $\int_a^b \omega(x) h(x) dx$ existe pour tout $h \in \mathcal{C}([a,b])$. On suppose ω continue par morceaux et on pose :

$$\langle f, g \rangle = \int_a^b \omega(x) f(x) g(x) dx.$$

Remarque 31. Le plus souvent dans le cas de l'approximation polynomiale on prend $\omega(x) = 1$.

La meilleure approximation de $f \in \mathcal{C}([a,b])$ par un polynôme de $\mathbf{P}_n$est donc le polynôme Q_n défini par :

$$\int_a^b \omega(x)[f(x) - Q_n(x)]^2 dx = \min_{P_n \in \mathbf{P}_n} \int_a^b \omega(x)[f(x) - P_n(x)]^2 dx.$$

Ce polynôme existe et vérifie :

$$\int_a^b \omega(x)[f(x) - Q_n(x)] P_n(x) dx = O \quad \text{pour tout} \quad P_n \in \mathbf{P}_n$$

Ses coefficients b_i sont donnés par le système d'équations linéaires :

$$\sum_{i=1}^{n+1} b_i \int_a^b \omega(x) x^{2n+2-i-j} dx = \int_a^b \omega(x) f(x) x^{n+1-j} dx \quad j = 1, \ldots, n+1 \tag{4.2}$$

Exemple 32. Le polynôme Q_2 qui réalise la meilleure approximation au sens des moindres carrés de la fonction $f(x) = x^3 - x^2 - \frac{1}{4}x + \frac{1}{4}$ sur l'intervalle $[-1, 1]$ avec $\omega(x) = 1$ est donné par la condition :

$$\sum_{i=1}^{3} b_i \int_{a=-1}^{b=1} 1.x^{2.2+2-i-j} dx = \int_{a=-1}^{b=1} 1.(x^3 - x^2 - \frac{1}{4}x + \frac{1}{4}) x^{2+1-j} dx \quad j = 1, 2, 3$$

c'est-à-dire le système :

$$\left\{ \begin{array}{rcrcr} \frac{2}{5}b_1 + \frac{2}{3}b_3 & = & -\frac{2}{5} + \frac{1}{6} & = & -\frac{7}{30} \\ \frac{2}{3}b_2 & = & \frac{2}{5} - \frac{1}{6} & = & \frac{7}{30} \\ \frac{2}{3}b_1 + 2b_3 & = & -\frac{2}{3} + \frac{1}{2} & = & -\frac{1}{6} \end{array} \right.$$

qui donne comme solution :

$$b_1 = -1; b_2 = \frac{7}{20}; b_3 = \frac{1}{4}.$$

le polynôme cherché est donc :

$$Q_2(x) = -x^2 + \frac{7}{20}x + \frac{1}{4}.$$

L'erreur commise pour cette approximation s'évalue comme suit :

$$\varepsilon_2 = f(x) - Q_2(x) = (x^3 - x^2 - \frac{1}{4}x + \frac{1}{4}) - \left(-x^2 + \frac{7}{20}x + \frac{1}{4}\right) = x^3 - \frac{3}{5}x.$$

Exemple 33. Le polynôme Q_2 qui réalise la meilleure approximation au sens des moindres carrés de la fonction $f(x) = |x|$ sur l'intervalle $[-1, 1]$ avec $\omega(x) = 1$ est donné par la condition :

$$\sum_{i=1}^{3} b_i \int_{-1}^{1} x^{6-i-j} dx = \int_{-1}^{1} |x| x^{3-j} dx \quad j = 1, 2, 3$$

c'est-à-dire le système :

$$\left\{ \begin{array}{rcl} \frac{2}{5}b_1 + \frac{2}{3}b_3 & = & \frac{1}{2} \\ \frac{2}{3}b_2 & = & 0 \\ \frac{2}{3}b_1 + 2b_3 & = & 1 \end{array} \right.$$

qui donne comme solution :

$$b_1 = \frac{15}{16}; b_2 = 0; b_3 = \frac{3}{16}.$$

le polynôme cherché est donc :

$$Q_2(x) = \frac{15}{16}x^2 + \frac{3}{16}.$$

L'erreur commise pour cette approximation s'évalue comme suit :

$$\varepsilon_2 = f(x) - Q_2(x) = |x| - \left(\frac{15}{16}x^2 + \frac{3}{16}\right) = \left\{ \begin{array}{lll} \frac{15}{16}x^2 - x - \frac{3}{16} & \text{si} & x < 0 \\ -\frac{15}{16}x^2 + x - \frac{3}{16} & \text{si} & x > 0 \end{array} \right. .$$

5 SERIE D'EXERCICES

Exercice 34. Trouver le polynôme P_2 qui réalise la meilleure approximation au sens des moindres carrés de la fonction $f(x) = |x-1|$ sur l'intervalle $[0,2]$ en prenant $\omega(x) = 1$.

Exercice 35. Trouver le polynôme P_2 qui réalise la meilleure approximation au sens des moindres carrés de la fonction $f(x) = |3x-5|$ sur l'intervalle $[-1,2]$ en prenant $\omega(x) = 1$.

Exercice 36. Trouver le polynôme P_2 qui réalise la meilleure approximation au sens des moindres carrés de la fonction donnée par le tableau suivant :

x	-1	$-0,5$	0	$0,5$	1
$f(x)$	$-0,75$	0	$0,25$	0	0

Chapitre

3 INTERPOLATION POLYNOMIALE

Sommaire

1 GÉNERALITÉS

Le problème de l'interpolation peut se poser dans plusieurs situations. Par exemple :

1- Une fonction f n'est pas complètement définie. On connait simplement un nombre fini de ses valeurs sur un intervalle $[a,b]$, (par exemple des données expérimentales) $x_1, x_2, \ldots, x_{n+1}$ supposés tels que :

$$a < x_1 \leq x_2 \leq \ldots \leq x_n \leq x_{n+1} < b$$

et les valeurs de f en ces points

$$f(x_1) = y_1, f(x_2) = y_2, \ldots, f(x_n) = y_n, f(x_{n+1}) = y_{n+1}$$

On peut alors vouloir déterminer une fonction φ, définie sur tout l'intervalle $[a,b]$ et constituant une certaine *"approximation de f"* (dans un sens à préciser). Il est alors souhaitable que $\varphi(x)$ soit numériquement *facile à évaluer*

2- $f(x)$ est connu pour tout x, mais son évaluation numérique est complexe. On peut alors vouloir déterminer une fois pour toutes un nombre fini de valeurs

$$f(x_1), f(x_2), \ldots, f(x_n), f(x_{n+1})$$

et pour tout

$$x \neq x_1, x_2, \ldots, x_n, x_{n+1}$$

approcher $f(x)$ en fonction de ces valeurs.

Cadre général de l'interpolation : *Chercher une fonction φ(appelé l'interpolant), d'un type préalablement choisi qui interpole f sur $[a,b]$, c'est déterminer cette fonction φ telle que*

$$\varphi(x_i) = f(x_i), \quad i = 1, 2, \ldots, n+1 \tag{1.1}$$

Cela signifie d'un point de vue géométrique, qu'il faut trouver une courbe d'équation $y = \varphi(x)$ et d'un type donné passant par le système de points

$$M_i = (x_i, y_i) \quad i = 1, 2, \ldots, n+1.$$

Le problème ainsi posé peut avoir une infinité de solutions ou ne pas avoir du tout de solution. Cependant, il admet une solution et une seule si l'on cherche non pas une fonction quelconque $\varphi(x)$ mais un polynôme de degré inférieur ou égal à n qui vé rifie (1.1) et est tel que :

$$P_n(x_1) = y_1, \quad P_n(x_2) = y_2, \ldots\ldots, P_n(x_{n+1}) = y_{n+1}.$$

Nous obtenons la formule d'interpolation :

$$y = \varphi(x).$$

Que nous utiliserons donc pour le calcul approché des valeurs de la fonction donnée $f(x)$ pour des valeurs de x différentes des points d'interpolation x_i, $i = 1, \ldots n+1$.

Remarque 37. Si $x \notin [x_1; x_{n+1}]$ on parle d'une extrapolation.

Proposition 38. Si $x_1, x_2, \ldots, x_{n+1}$ sont $(n+1)$ points distincts de $\mathbb{R}$ et si $y_1, y_2, \ldots, y_{n+1}$ sont $(n+1)$ réels, alors il existe un et un seul polynôme réel P de degré inférieur ou égal à n tel que

$$P_n(x_1) = y_1, \quad P_n(x_2) = y_2, \ldots\ldots, P_n(x_{n+1}) = y_{n+1}.$$

Démonstration. Tout polynôme réel de degré inférieur ou égal à n s'écrit de manière unique sous la forme

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

l'ensemble $\mathbb{R}[X]$ de ces polynômes est donc un espace vectoriel réel de dimension $(n+1)$. L'application

$$\begin{array}{ccc} \mathbb{R}[X] & \to & \mathbb{R}^{n+1} \\ P & \mapsto & \begin{pmatrix} P(x_1) \\ \vdots \\ P(x_{n+1}) \end{pmatrix} \end{array}$$

est linéaire et injective car un polynôme de degré inférieur ou égal à n qui a $(n+1)$ zéros distincts est nul. Il s'ensuit que cette application est aussi surjective, d'où la conclusion. cqfd

Remarque 39. Pour trouver les coéfficients du polynôme $P_n(x)$ c'est-à -dire déterminer $a_0, a_1, ..., a_n$, il suffit de résoudre le système linéaire (de Cramer) suivant :

$$\begin{pmatrix} 1 & x_1 & x_1^n \\ \vdots & \vdots & \vdots \\ 1 & x_{n+1} & x_{n+1}^n \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}$$

La matrice de ce système est une matrice de Vandermonde. Cependant, cette méthode de détermination des coefficients est peu efficace. On préfère des méthodes basées sur des formules explicites pour le polynôme $P_n(x)$ telles que les formules dites de Lagrange et de Newton établies ci-dessous.

2 POLYNOME DE LAGRANGE

Soit $\{x_1, x_2, ...x_{n+1}\}$, $(n+1)$ valeurs distinctes de $[a, b]$ données. On suppose que que l'on connaisse les valeurs correspondantes de $y = f(x)$ c'est-à-dire on a :

$$f(x_1) = y_1, \;\; f(x_2) = y_2, ...f(x_{n+1}) = y_{n+1}.$$

On veut construire un polynôme $L_n(x) = \sum_{i=1}^{n+1} p_i(x) f(x_i)$, de degré inférieur ou égal à n, vérifiant :

$$L_n(x_i) = y_i, \;\; i = 1, 2, ...n+1.$$

où les fonctions $p_i(x)$ sont telles que :

$$p_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{si} \quad j = i \\ 0 & \text{si} \quad j \neq i \end{cases}$$

Où δ_{ij} est le symbole de Kronecker. **Résolution du problème :** Le polynôme à obtenir s'annulant en $(n+1)$ points $x_1, x_2 ..., x_{i-1}, x_{i+1}, ..., x_{n+1}$, il s'ecrit donc :

$$p_i(x) = C_i(x - x_1)(x - x_2)...(x - x_{i-1})(x - x_{i+1})...(x - x_n), \tag{2.1}$$

où C_i est une constante. Posant dans (3.1) $x = x_i$ et comme $p_i(x_i) = 1$, on a donc :

$$C_i(x_i - x_1)(x_i - x_2)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n) = 1$$

c'est-à-dire :

$$C_i = \frac{1}{(x_i - x_1)(x_i - x_2)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}$$

en portant cette valeur dans (3.1) on obtient :

$$p_i(x) = \frac{(x - x_1)(x - x_2)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_1)(x_i - x_2)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)} = \prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(x - x_j)}{(x_i - x_j)}. \tag{2.2}$$

Le polynôme $L_n(x)$ qui vérifie les conditions $L_n(x_i) = y_i$, est alors de la forme :

$$L_n(x) = \sum_{i=1}^{n+1} p_i(x) y_i \tag{2.3}$$

En effet, d'une part le polynôme $L_n(x)$ ainsi construit est un polyn ôme de degré inférieur ou égal à n. Et d'autre part comme $p_i(x_j) = 1$ si $j = i$ et 0 sinon, on a :

$$L_n(x_j) = \sum_{i=1}^{n+1} p_i(x_j) y_i = p_j(x_j) y_i = y_i \qquad j = 1, 2, ..., n+1$$

En portant la valeur de $p_i(x)$ dans (2.3) tirée de (2.2) on obtient :

$$\begin{aligned} L_n(x) &= \sum_{i=1}^{n+1} \frac{(x-x_1)(x-x_2)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_1)(x_i-x_2)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_n)} y_i = \\ &= \sum_{i=1}^{n+1} \left(\prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(x-x_j)}{(x_i-x_j)} \right) y_i. \end{aligned} \tag{2.4}$$

Le polynôme donné par (2.4) est appelé le *polynôme d'interpolation de Lagrange*. et

$$p_i(x_j) = \left(\prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(x-x_j)}{(x_i-x_j)} \right) \tag{2.5}$$

sont appelés les *coeficients de Lagrange*.

Proposition 40. Le polynôme d'interpolation de Lagrange est unique.

Démonstration. Faisons un raisonnement par l'absurde. Soit $\check{L}_n(x)$ un polynôme de degré inférieur ou é gal à n, distinct de $L_n(x)$ et est tel que :

$$\check{L}_n(x) = y_i \qquad i = 1, 2, ..., n+1$$

Le polynôme

$$Q_n(x) = \check{L}_n(x) - L_n(x)$$

dont le degré est aussi inférieur ou égal à n, s'annule en $n+1$ points $\{x_1, x_2...., x_{n+1}\}$, c'est-à-dire :

$$Q_n(x) = 0$$

donc

$$\check{L}_n(x) = L_n(x)$$

cqfd

Exemple 41. Construire les coefficients de Lagrange pour $n+1 = 2$ et $n+1 = 3$.

Solution 42. - Cas $n+1 = 2$: En appliquant la formule (2.5) on obtient

$$p_1(x) = \frac{x-x_2}{x_1-x_2}; \;\; p_2(x) = \frac{x-x_1}{x_2-x_1}$$

et alors

$$L(x) = \frac{x-x_2}{x_1-x_2} f(x_1) + \frac{x-x_1}{x_2-x_1} f(x_2)$$

- Cas $n+1=3$: En appliquant la formule (2.5) on obtient

$$p_1(x)=\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)};\ \ p_2(x)=\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)};\ \ p_3(x)=\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)};$$

et alors

$$L_2(x)=\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}f(x_1)+\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}f(x_2)+\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}f(x_3).$$

Exemple 43. Construire le polynôme de Lagrange de la fonction donnée par $y=f(x)=\sin\pi x$ pour les points $x_1=0$, $x_2=\frac{1}{6}$, $x_3=\frac{1}{2}$.

Solution 44. On calcule d'abord les valeurs correspondantes aux points d'interpolation de la fonction $y=f(x)$:

$$y_1=0,\quad y_2=\sin\frac{\pi}{6}=\frac{1}{2},\quad y_3=\sin\frac{\pi}{2}=1$$

On applique ensuite les formules (2.4), on obtient le polynô me de Lagrange de degré inférieur ou égal à 2 suivant :

$$L_2(x)=\frac{(x-\frac{1}{6})(x-\frac{1}{2})}{(0-\frac{1}{6})(0-\frac{1}{2})}.0+\frac{(x)(x-\frac{1}{2})}{\frac{1}{6}(\frac{1}{6}-\frac{1}{2})}.\frac{1}{2}+\frac{(x)(x-\frac{1}{6})}{\frac{1}{2}(\frac{1}{2}-\frac{1}{6})}.1$$

Ou

$$L_2(x)=-3x^2+\frac{7}{2}x$$

Exemple 45. Soit la fonction $y=f(x)$ donnée par le tableau suivant :

x	y
321,0	2,50651
322,8	2,50893
324,2	2,51081
325,0	2,51188

On demande le calcul de la valeur de f en 323,5.

Solution 46. On pose $x=323,5$; le polynôme de Lagrange sera un polynôme de degr é inférieur ou égal à $n=3$. D'après les formules (2.4), on a :

$$\begin{aligned}f(323,5) &= \frac{(323,5-322,8)(323,5-324,2)(323,5-325,0)}{(321,0-322,8)(321,0-324,2)(321,0-325,0)}.2,50651+\\ &\quad+\frac{(323,5-321,0)(323,5-324,2)(323,5-325,0)}{(322,8-321,0)(322,8-324,2)(322,8-325,0)}.2,50893+\\ &\quad+\frac{(323,5-321,0)(323,5-322,8)(323,5-325,0)}{(324,2-321,0)(324,2-322,8)(324,2-325,0)}.2,51081+\\ &\quad\frac{(323,5-321,0)(323,5-322,8)(323,5-324,2)}{(325,0-321,0)(325,0-322,8)(325,0-324,2)}.2,51188\\ &= -0,07996+1,18794+1,83897-0,43708=2,50987\end{aligned}$$

2.1 Cas où les points sont equidistants

Soit $\{x_1,x_2....,x_{n+1}\}$, $(n+1)$ points d'interpolation de $[a,b]$, on suppose que :

$$x_2=x_1+h,\ \ x_3=x_2+h=x_1+2h,\ ...x_j=x_{j-1}+h=x_1+(j-1)h,\ \ ...x_{n+1}=x_n+h=x_1+nh$$

Où $h = \frac{b-a}{n}$ représente le pas de la subdivision. Les points d'interpolation x_i sont alors dits *équidistants*. Dans ce cas, les coefficients de Lagrange peuvent être simplifiés, en posant :

$$x = x_1 + th$$

on aura :

$$t_1 = 0, \quad t_2 = 1, \quad ...t_n = n.$$

D'où

$$\begin{aligned} l_n(t) &= \sum_{i=1}^{n+1} \left(\prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(x_1 + th - x_1 - (j-1)h)}{((x_1 + (i-1)h) - x_1 + (j-1)h)} \right) y_i \\ &= \sum_{i=1}^{n+1} \left(\prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(t-j+1)}{(i-j)} \right) y_i \end{aligned} \tag{2.6}$$

Les coefficients de $l_n(t)$ sont donc indépendants du pas h et des points x_i ; ils peuvent être représentés dans une table.

Exemple 47. Soit la fonction $y = \cos x$ donnée par le tableau suivant :

x	5,0	5,1	5,2	5,3	5,4	5,5	5,6	5,7
y	0,283662	0,377977	0,468516	0,554374	0,634692	0,708669	0,775565	0,834712
t	1	2	3	4	5	6	7	8

Calculer $\cos 5,34$.

Solution 48. Posons

$$x = 0,1t + 5$$

Les valeurs de la nouvelle variable t associées aux points d'interpolation seront alors :

$$t = 0, 1, 2, 3, 4, 5, 6, 7$$

Il faut donc trouver la valeur de y pour $x = 5,34$, c'est-à-dire pour $t = 3,47$, les points étant équidistants, on a donc :

$$\begin{aligned} l_n(3,47) &= \sum_{i=1}^{8} \left(\prod_{\substack{j=1 \\ j \neq i}}^{8} \frac{(3,47-j+1)}{(i-j)} \right) y_i \\ &= 0,592864 \end{aligned}$$

Donc $\cos 5,34 = 0,592864$.

3 Estimation de l'erreur dans l'interpolation de Lagrange

Soit $L_n(x)$ le polynôme de Lagrange qui interpole la fonction $y = f(x)$ aux points $\{x_1, x_2...., x_{n+1}\}$ c'est-à -dire qui vérifie :

$$L_n(x_1) = y_1, \quad L_n(x_2) = y_2, \quad ..., L_n(x_{n+1}) = y_{n+1}.$$

La question que l'on se pose maintenant est : quelle est l'approximation du polynôme construit par rapport à la fonction $f(x)$? ou en d'autres termes quelle est la grandeur du reste :

$$R_n(x) = f(x) - L_n(x)$$

Pour évaluer cette erreur commise dans l'interpolation par le polynô me de Lagrange, nous supposons que la fonction f est continue, dé rivable et à dérivées continues jusqu'à l'ordre $(n+1)$ dans le domaine $[a,b]$ contenant x et les points d'interpolation x_i. Soit

$$g(x) = f(x) - L_n(x) - k(x-x_1)(x-x_2).....(x-x_n)$$

Où k est une constante. La fonction g possède $n+1$ racines aux points

$$x_1, x_2, ...x_n$$

Choisissons k de sorte que $g(x)$ ait une $(n+2)^{ième}$ racine en un point quelconque fixé $\tilde{x}$ de $[a,b]$, autre que les points d'interpolation. Il suffit pour cela de poser

$$f(\tilde{x}) - L_n(\tilde{x}) - k(\tilde{x}-x_1)(\tilde{x}-x_2).....(\tilde{x}-x_n)$$

D'où

$$k = \frac{f(\tilde{x}) - L_n(\tilde{x})}{(\tilde{x}-x_1)(\tilde{x}-x_2).....(\tilde{x}-x_n)} \tag{3.1}$$

car $\tilde{x}$ n'est pas un point d'interpolation. Pour cette valeur de k, la fonction $g(x)$ admet $(n+2)$ racines sur $[a,b]$ et elle s'annule aux extrémités de chacun des intervalles suivants :

$$[x_1,x_2],\ [x_2,x_3],\ \ldots\ [x_i,\tilde{x}],\ [\tilde{x},x_{i+1}],\ \ldots [x_n,x_{n+1}],$$

En appliquant le théorème de Rolle à chacun de ces intervalles, on voit que la dérivée $g'(x)$ admet au moins $(n+1)$ racines sur $[a,b]$. De même, la dérivée seconde $g''(x)$ s'annule au moins n fois sur $[a,b]$. En opérant de même pour les dérivées successives de la fonction g, on aboutit à la conclusion que dans $[a,b]$, la dérivée $g^{(n+1)}(x)$ possède au moins une racine. Notons par ξ cette racine : on a donc $g^{(n+1)}(\xi) = 0$. Comme

$$g^{(n+1)}(x) = f^{(n+1)}(x) - k(n+1)!$$

Pour $x = \xi$ on obtient :

$$0 = f^{(n+1)}(\xi) - k(n+1)!$$

Donc

$$k = \frac{f^{(n+1)}(\xi)}{(n+1)!} \tag{3.2}$$

En identifiant les équations (3.1) et (3.2) on obtient :

$$\frac{f(\tilde{x}) - L_n(\tilde{x})}{(\tilde{x}-x_1)(\tilde{x}-x_2).....(\tilde{x}-x_n)} = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

C'est-à-dire

$$f(\tilde{x}) - L_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(\tilde{x}-x_1)(\tilde{x}-x_2).....(\tilde{x}-x_{n+1}) \tag{3.3}$$

Comme $\tilde{x}$ est complètement arbitraire, l'équation (3.3) s'écrit :

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_1)(x-x_2).....(x-x_{n+1}) \tag{3.4}$$

Où $\xi \in [a,b]$ dépend de x.

Remarque 49. L'équation (3.4) est vraie pour tout point de $[a,b]$, y compris les points d'interpolation. En posant

$$M_{n+1} = \max_{x\in[a,b]} \left| f^{(n+1)}(x) \right|$$

Nous obtenons l'estimation de l'erreur absolue dans l'interpolation par le polynôme de Lagrange sous la forme :

$$|R_n(x)| = |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |(x-x_1)(x-x_2).....(x-x_n)|$$

Exemple 50. Avec quelle précision peut-on calculer $\sqrt{115}$ à l'aide d'une interpolation par le polynôme de Lagrange de la fonction $y = \sqrt{x}$si l'on prend les points d'interpolation

$$x_1 = 100, \quad x_2 = 121, \quad x_3 = 144.$$

Solution 51. Comme on a

$$y' = \frac{1}{2}x^{-\frac{1}{2}}, \quad y'' = -\frac{1}{4}x^{-\frac{3}{2}}, \quad y''' = \frac{3}{8}x^{-\frac{5}{2}}.$$

Il s'ensuit

$$M_3 = \max\left|y'''\right| = \frac{3}{8}(100)^{-\frac{5}{2}} = \frac{3}{8}10^{-5} \quad \text{pour } 100 \leq x \leq 144$$

Donc

$$\begin{aligned} |R_2(x)| &\leq \frac{3}{8}10^{-5}\frac{1}{3!}|(115-100)(115-121)(115-144)| = \\ &= \frac{1}{16}10^{-5}.15.6.29 \approx 1,6.10^{-3} \end{aligned}$$

4 POLYNOME DE NEWTON

4.1 Différences finies

Définition 52. Soit $y = f(x)$ une fonction donnée. On pose $\Delta x = x_{i+1} - x_i = h$ une valeur fixée de l'accroissement de x. On appelle *différence d'ordre un* de la fonction y l'expression :

$$\Delta y = \Delta f(x) = f(x + \Delta x) - f(x)$$

On définit de façon analogue les différences d'ordres supé rieurs

$$\Delta^n y = \Delta(\Delta^{n-1} y) \quad (n = 2, 3, ...)$$

Exemple 53. $\Delta^2 y = \Delta[f(x+\Delta x) - f(x)] = [f(x+2\Delta x) - f(x+\Delta x)] - [f(x+\Delta x) - f(x)] = f(x+2\Delta x) - 2f(x+\Delta x) + f(x)$.

Exemple 54. Construire les différences de $f(x) = x^3$, en prenant le pas $\Delta x = 1$.

Solution 55. On a $\Delta f(x) = (x+1)^3 - x^3 = 3x^2+3x+1$, $\Delta^2 f(x) = \left[3(x+1)^2 + 3(x+1) + 1\right] - (3x^2+3x+1) = 6x + 6$, $\Delta^3 f(x) = [6(x+1)+6] - (6x+6) = 6$, $\Delta^n f(x) = 0$, pour $n > 3$.

Proposition 56. Si $f(x) = P_n(x) = a_0 x^n + a_1 x^{n-1} + ... + a_n$ est un polynôme de degré n, alors $\Delta^n f(x) = \Delta^n P_n(x) = n! a_0 h^n = C$ où $\Delta x = h$ et C est une constante.

Démonstration. En effet, on a :

$$\begin{aligned} \Delta f(x) &= \Delta P_n(x) = P_n(x+h) - P_n(x) = \\ &= a_0[(x+h)^n - x^n] + a_1\left[(x+h)^{n-1} - x^{n-1}\right] + ... \\ &\quad ... + a_{n-1}[(x+h) - h] \end{aligned}$$

En utilisant la formule du binôme de Newton, on voit que $\Delta f(x) = \Delta P_n(x)$ est un polynôme de degré $(n-1)$:

$$\Delta f(x) = \Delta P_n(x) = b_0 x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-1}$$

Où

$$b_0 = nha_0$$

Suivant le même raisonnement la différence seconde $\Delta^2 f(x) = \Delta^2 P_n(x)$ est un polynôme de degré $(n-2)$:

$$\Delta^2 f(x) = \Delta^2 P_n(x) = c_0 x^{n-2} + c_1 x^{n-3} + \cdots + c_{n-2}$$

et

$$c_0 = (n-1)hb_0 = n(n-1)h^2a_0$$

En raisonnant ainsi on établit de proche en proche que

$$\Delta^n f(x) = \Delta^n P_n(x) = n!h^n a_0$$

D'où l'on conclut

$$\Delta^m f(x) = \Delta^m P_n(x) = 0 \quad \text{pour} \quad m > n$$

cqfd

Remarque 57. Le symbole Δ (delta) peut être considéré comme un *opérateur* qui associe à la fonction $y = f(x)$ la fonction $\Delta y = f(x + \Delta x) - f(x)$ (Δx étant une constante).

Remarque 58. L'opérateur Δ a les propriétés suivantes :

1. $\Delta(u + v) = \Delta u + \Delta v$;
2. $\Delta(Ku) = K\Delta u$ (K est une constante) ;
3. $\Delta^m(\Delta^n u) = \Delta^{m+n} u$ (m et n entiers non négatifs) ;
4. On pose par définition $\Delta^0 u = u$.

4.2 Différences divisées

Définition 59. Soit $\{x_1, x_2...., x_{n+1}\}$, $(n+1)$ points d'interpolation de $[a, b]$, et :

$$f(x_1) = y_1, \quad f(x_2) = y_2, \quad ..., f(x_{n+1}) = y_{n+1}.$$

on pose :

$$\Delta x_i = x_{i+1} - x_i \neq 0 \qquad (i = 1, 2, ...)$$

On appelle *différences divisées d'ordre* 1, les relations données par :

$$f[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \quad (i = 1, 2, ...)$$

Exemple 60. $f[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1}$, $f[x_2, x_3] = \frac{y_3 - y_2}{x_3 - x_2}$, etc...

D'une façon analogue on définit les *différences divisées d'ordre 2*

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}, \quad (i = 1, 2, ...)$$

Exemple 61. $f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{f[x_3, x_1]}$.

D'une façon générale, les *différences divisées d'ordre n* s'obtiennent à partir des différences divisées d'ordre $(n-1)$ à l'aide de la relation de récurrence suivante :

$$f[x_i, x_{i+1}, ...x_{i+n}] = \frac{f[x_{i+1}, ...x_{i+n}] - f[x_i, x_{i+n-1}]}{x_{i+n} - x_i}, \quad (n = 1, 2, ...; i = 1, 2, ...)$$

Remarque 62. Les différences divisées sont symétriques de leurs arguments, c'est-à-dire que les différences divisées ne changent pas avec la permutation des éléments. Plus précisément, pour toute permutation σ de $\{1, 2...., k\}$ on a

$$f\left[x_{\sigma(1)}, x_{\sigma(2)}, ...x_{\sigma(k)}\right] = f[x_1, x_2, ...x_k]$$

Exemple 63. $f[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_1 - y_2}{x_1 - x_2} = f[x_2, x_1]$

Théorème 64. Soit p_n le polynôme d'interpolation de f aux points $\{x_1, x_2...., x_{n+1}\}$, alors pour tout k de $\{1, 2...., n\}$ on a :

$$f[x_1, x_2, ...x_k] = \frac{f[x_2, ...x_{k+1}] - f[x_1, x_k]}{x_{k+1} - x_1},$$

Démonstration. L'idée est de construire une fonction polynôme qui interpole f sur $\{x_1, x_2...., x_{k+1}\}$; puis on identifie les deux expressions disponibles de son coefficient dominant. Soit p défini par

$$p(x) = \frac{x - x_1}{x_{k+1} - x_1} q_k(x) + \frac{x_{k+1} - x}{x_{k+1} - x_1} p_k(x)$$

Où q_k (respectivement p_k) désigne le polynôme d'interpolation de f sur $\{x_2, x_3...., x_{k+1}\}$ (respectivement $\{x_1, x_2...., x_k\}$). Par définition, p est un polynôme de degré inférieur ou égal à k. Montrons que p interpole f sur $\{x_1, x_2...., x_k\}$. On évalue d'abord $p(x_1)$: la contribution du premier terme de la somme est nulle ; il vient $p(x_1) = p_k(x_1)$ soit $p(x_1) = f(x_1)$ par d éfinition de p_k. On évalue alors $p(x_k)$ par le même type de raisonnement, on montre que $p(x_k) = f(x_k)$. Puis on considè re x_j pour tout j dans $\{2, 3...., k\}$. On sait que $p_k(x_j) = q_k(x_j) = f(x_j)$ grâce aux définitions de p_k et de q_k. Un calcul simple montre que : $p(x_j) = f(x_j)$; en effet

$$p(x_j) = \frac{1}{x_{k+1} - x_1} \left[\left(x_j - x_1 \right) f(x_j) + \left(x_{k+1} - x_j \right) f(x_j) \right] \tag{4.1}$$

Soit

$$p(x_j) = f(x_j)$$

p est le polynôme d'interpolation de f sur $\{x_1, x_2...., x_{k+1}\}$. Soit α le coefficient dominant de p ; d'après ce qui préc ède, $\alpha = f[x_1, ..., x_{k+1}]$. Sur l'expression (4.1), on voit que le coefficient dominant de p est donné par

$$\alpha = \frac{1}{x_{k+1} - x_1} [\text{coef dominant } (q_k) - \text{coef dominant } (p_k)]$$

comme (q_k) et (p_k) sont des polynômes d'interpolation sur $\{x_2, x_3...., x_{k+1}\}$ et $\{x_1, x_2...., x_k\}$ respectivement, l'égalité cherché e en découle. cqfd

Remarque 65. Les différences divisées forment généralement un tableau du type suivant :

x	$f(x)$	Différences divisées			
		Ordre 1	Ordre 2	Ordre 3	Ordre 4
x_1	$f[x_1]$				
		$f[x_1, x_2]$			
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$		
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$		$f[x_1, x_2, x_3, x_4, x_5]$
		$f[x_3, x_4]$		$f[x_2, x_3, x_4, x_5]$	
x_4	$f[x_4]$		$f[x_3, x_4, x_5]$		
		$f[x_4, x_5]$			
x_5	$f[x_5]$				

Exemple 66. Donner les différences divisées de la fonction donnée par le tableau suivant :

x	0	0,2	0,3	0,4	0,7	0,9
y	132,651	148,877	157,464	166,375	195,112	216,000

Solution 67. Les résultats sont portés sur le tableau suivant :

x	$f(x)$	Différences divisées			
		Ordre 1	Ordre 2	Ordre 3	Ordre 4
0	132,651				
		81,13			
0,2	148,877		15,8		
		85,87		1	
0,3	157,464		16,2		0
		89,11		1	
0,4	166,375		16,7		0
		95,79		1	
0,7	195,112		17,3		
		104,44			
0,9	216,000				

4.3 Polynôme d'interpolation de Newton :

On a par définition

$$f[x,x_1] = \frac{y_1 - y}{x_1 - x} = \frac{f(x) - f(x_1)}{x - x_1},$$

donc

$$f(x) = f(x_1) + (x - x_1) f[x,x_1]$$

et

$$f[x,x_1,x_2] = \frac{f[x,x_1] - f[x_1,x_2]}{x - x_2}$$

donc

$$f(x) = f(x_1) + (x - x_1) f[x_1,x_2] + (x - x_1)(x - x_2) f[x,x_1,x_2]$$

en réitérant le procédé, on obtient :

$$\begin{aligned} f(x) &= f(x_1) + (x - x_1) f[x_1,x_2] + (x - x_1)(x - x_2) f[x_1,x_2,x_3] + \cdots + \\ &\quad + (x - x_1)(x - x_2) \cdots\cdots (x - x_n) f[x_1,x_2,...,x_{n+1}] \\ &\quad + (x - x_1)(x - x_2) \cdots\cdots (x - x_n)(x - x_{n+1}) f[x,x_1,x_2,...,x_{n+1}] \end{aligned} \tag{4.2}$$

Remarque 68. Si f est un polynôme de degré inférieur ou égal à n on a :

$$f[x,x_1,x_2,...,x_{n+1}] = 0$$

Remarque 69. Le polynôme défini par

$$\begin{aligned} P_n(x) &= f(x_1) + (x - x_1) f[x_1,x_2] + (x - x_1)(x - x_2) f[x_1,x_2,x_3] + \cdots + \\ &\quad + (x - x_1)(x - x_2) \cdots\cdots (x - x_n) f[x_1,x_2,...,x_{n+1}] \end{aligned}$$

est le polynôme d'interpolation de degré inférieur ou égal à n de f pour les points $\{x_1, x_2 ...., x_{n+1}\}$ appelé *polynôme d'interpolation de Newton*. C'est-à-dire qu'il vérifie :

$$P_n(x_i) = f(x_i) \quad i = 1, 2, ...n + 1.$$

4.4 Erreur d'interpolation

En comparant $f(x)$ et $P_n(x)$, nous obtenons la formule générale de l'erreur :

$$\varepsilon(x) = f(x) - P_n(x) = (x - x_1)(x - x_2) \cdots\cdots (x - x_{n+1}) f[x,x_1,x_2,...,x_{n+1}]$$

que l'on peut ecrire sous la forme :

$$\varepsilon(x) = \left(\prod_{i=1}^{n+1} (x - x_i) \right) f[x,x_1,x_2,...,x_{n+1}]$$

4.5 Autre écriture du polynôme d'interpolation de Newton

4.5.1 Cas des points équidistants

En posant le pas du tableau $h = \Delta x$ ($\Delta x_i = x_{i+1} - x_i$, $i = 1, 2, ..., n+1$) et $y_1 = f(x_1), y_2 = f(x_2), \ldots, y_n = f(x_n), y_{n+1} = f(x_{n+1})$; alors nous obtenons :

$$f(x) = y_1 + k\Delta y_1 + \frac{k(k-1)}{2!}\Delta^2 y_1 + \cdots + \frac{k(k-1)...(k-n+1)}{n!}\Delta^n y_1 + R_n(x)$$

Où

$$k = \frac{x - x_1}{h} \quad \text{et} \quad \Delta y_1 = y_2 - y_{1;}\ \Delta^2 y_1 = \Delta y_2 - \Delta y_{1,}$$

et

$$R_n(x) = \frac{h^{n+1}k(k-1)...(k-n)}{(n+1)!}f^{(n+1)}(\xi_x) \simeq \frac{k(k-1)...(k-n)}{(n+1)!}\Delta^{n+1}y_1; \qquad \xi_x \in [x, x_i]$$

Exemple 70. Sachant que $\sin 26° = 0,43837$; $\sin 27° = 0,45399$ et $\sin 28° = 0,46947$, calculer $\sin 26°15'$.

Solution 71. Le tableau des valeurs se présente comme suit :

i	x_i	y_i	Δy_i	$\Delta^2 y_i$
1	26	0,43837		
			0,01562	
2	27	0,45399		−0,00014
			0,01548	
3	28	0,46947		

comme $h = 60'$ et $k = \frac{1575' - 1560'}{60} = \frac{1}{4}$ donc

$$\sin 26°15' = 0,43837 + \frac{1}{4}.0,01562 + \frac{\frac{1}{4}(\frac{1}{4}-1)}{2}(-0,00014) = 0,44229.$$

L'erreur

$$|R_2(x)| \leq \frac{\frac{1}{4}(\frac{1}{4}-1)(\frac{1}{4}-2)}{3!}(\frac{\pi}{180}) \simeq 0,2510^{-6}$$

car comme $y = \sin x$ alors $\left|y^{(n)}\right| \leq 1$.

Remarque 72. Si l'on fixe n et que l'on pose

$$h = \max_{h\in\{1,...,n\}} |x_{k+1} - x_k|$$

alors pour tout $x \in [x_1, x_2]$, on a

$$|(x - x_1)\cdots(x_1 - x_{n+1})| \leq h(h)(2h)\cdots(nh) \leq h^{n+1}n!.$$

Si f est de classe C^{n+1} sur $[a, b]$ et si on a $x_1 = a$ et $h < \frac{b-a}{n}$, alors

$$\sup_{x\in[x_1,x_2]} |f(x) - P_n(x)| \leq \frac{\sup_{x\in[a,b]}\left|f^{(n+1)}(x)\right|}{n+1}h^{n+1}.$$

Comme le second membre tend vers zéro avec h, on peut donc rendre

$$\sup_{x\in[x_1,x_2]} |f(x) - P_n(x)|$$

aussi petit que l'on veut (avec h suffisamment petit). Ceci montre que l'interpolation polynomiale peut être utilisée pour approcher les valeurs de $f(x)$.

Remarque 73. La remarque précédente n'est en général pas vraie, même si f est très régulière sur $[a,b]$. Dans le cas o ù $x_{k+1}-x_k$ ne dépend pas de k, c'est là ce qu'on appelle le phénomène de Runge. Par exmple si

$$f(x)=\frac{1}{1+25x^2}$$

et si $[a,b]=[-1,1]$, on peut montrer que le polyn ôme $P_n(x)$ interpolant $f(x)$ aux points

$$x_k=-1+\frac{2k}{n},\quad (k=1,...,n+1),$$

ne tend pas vers f lorsque $n\to\infty$.

5 INTERPOLATION CUBIQUE DE HERMITE

On cherche le polynôme P_3 de $\mathbb{R}[X]$ qui prend deux valeurs imposées y_1 et y_2 en deux points donnés x_1 et x_2, et deux valeurs imposées de la dérivée y_3 et y_4 aux deux mêmes points, c'est-à -dire :

$$P_3(x_1)=y_1,\quad P_3(x_2)=y_2,\quad P_3'(x_1)=y_3,\quad P_3'(x_2)=y_4.$$

Comme il y a quatre inconnues à déterminer dans $P_3(x)=ax^3+bx^2+cx+d$, on écrit les équations

$$\begin{cases} ax_1^3+bx_1^2+cx_1+d &= y_1\\ ax_2^3+bx_2^2+cx_2+d &= y_2\\ 3ax_1^2+2bx_1+c &= y_3\\ 3ax_2^2+2bx_2+c &= y_4 \end{cases}$$

le déterminant de ce système se factorise en $-(x_2-x_1)^4$ et on obtient :

$$\begin{aligned} a &= \frac{2(y_1-y_2)}{(x_2-x_1)^3}+\frac{y_3+y_4}{(x_1-x_2)^2}\\ b &= \frac{3(x_1+x_2)(y_1-y_2)}{(x_1-x_2)^3}-\frac{(x_1+2x_2)y_3+(2x_1+x_2)y_4}{(x_1-x_2)^2}\\ c &= \frac{6x_1x_2(y_1-y_2)}{(x_2-x_1)^3}+\frac{x_2(2x_1+x_2)y_3+x_1(2x_2+x_1)y_4}{(x_1-x_2)^2}\\ d &= \frac{x_1^2(3x_1-x_2)y_1-x_2^2(3x_2-x_1)y_2}{(x_1-x_2)^3}-\frac{x_1x_2(x_2y_3+x_1y_4)}{(x_1-x_2)^2}. \end{aligned}$$

6 SERIE D'EXERCICES

Exercice 74. On considère la fonction $f(x)$ donnée par le tableau suivant :

x	1	4	6
$f(x)$	1.5709	1.5727	1.5751

Trouver une approximation de $f(3.5)$ en utilisant le polynôme d'interpolation de Lagrange du second degré .

Exercice 75. Ecrire les différences divisées de $f(x)=x^2$ et de $g(x)=x^3$.

Exercice 76.

1. Construire le polynôme d'interpolation de Newton de la fonction $y=f(x)$ donnée par le tableau suivant :

x	0	2.5069	5.0154	7.52270
$f(x)$	0.3989423	0.3988169	0.3984408	0.3978138

2. Trouver à l'aide de ce polynôme $f(3.7608)$.

Exercice 77. Soit le tableau des valeurs $y = \log x$, trouver $\log 1005$.

x	1000	1010	1020	1030	1040	1050
y	3.0000000	3.0043214	3.0086002	3.0128372	3.0170333	3.0211893

Exercice 78. Trouver $\sin 14°$ à partir du tableau des valeurs données ci-dessous de la fonction $y = \sin x$, le pas étant $h = 5°$.

x	15°	20°	25°	30°	35°	40°	45°	50°	55°
y	0.2588	0.3420	0.4226	0.5000	0.5736	0.6428	0.7071	0.7660	0.8192

Exercice 79.

1. En interpolant par un polynôme de degré 3 et en utilisant la formule appropriée calculer pour $x = 1.05$, à l'aide de la table donnée ci-dessous, les valeurs de la fonction $y = \sin x$.

x	1.0	1.1	1.2	1.3
y	0.841471	0.891207	0.932039	0.963558

2. Donner une majoration de l'erreur.

Chapitre

4 INTEGRATION ET DÉRIVATION NUMÉRIQUE

Sommaire

1 INTÉGRATION NUMÉRIQUE

1.1 Méthode Générale

Lorsque l'intégrale définie d'une fonction continue sur un intervalle $[a,b]$ ne peut pas être évaluée analytiquement ou lorsque l'intégrale n'est pas donnée sous forme analytique mais numériquement en un certain nombre de valeurs discrètes, l'intégration numérique peut être utilisée.

Il existe plusieurs méthodes permettant d'évaluer les intégrales de fonctions bornées sur un intervalle $[a,b]$. La présence de singularité dans les fonctions (ou dans certaines fonctions) rend les calculs parfois difficiles.

Le problème de l'intégration numériqued'une fonction consiste à chercher la valeur de l'intégrale définie à partir de plusieurs valeurs de la fonction sous le signe somme. L'intégrale à évaluer étant :

$$I = \int_a^b f(x)dx. \tag{1.1}$$

L'intégration numérique consiste à remplacer l'intégrale (2.1) par une somme discrète sur un nombre fini de points :

$$I_N = \sum_{i=1}^{N} A_i f(x_i)$$

où a_i et x_i sont des variables à préciser.

Pour que l'avaluation numérique soit correcte, il est nécessaire d'imposer que toute méthode d'intégration vérifie :

$$\lim_{N\to\infty} I_N = I \tag{1.2}$$

Au delà de la vérification de ce critère, la qualité d'une méthode sera évaluée par la manière dont la convergence vers le résultat exact s'effectue.

Nous supposerons dorénavant que les fonctions sont de classe C^1(continues et à dérivées continues) sur $[a,b]$, mais aussi pour toute fonction f :

$$f'(x) < K \quad \forall x \in [a,b],$$

où K est une constante finie. Ce qui veut dire que la dérivée de f n'est pas singulière.sur $[a,b]$.

On se propose alors de chercher une approximation de I. On remplace f sur $[a,b]$ par une fonction d'interpolation φ (un polynôme par exemple) pour considérer :

$$\int_a^b f(x)dx = \int_a^b \varphi(x)dx$$

1.2 Approximation d'une intégrale

Soit ω une fonction positive définie sur $[c,d]$, on veut approcher

$$I = \int_c^d \omega(x)f(x)dx.$$

on suppose que f est connue en $(n+1)$ points $x_1, x_2, ..., x_{n+1}$. On écrit

$$I = \sum_{i=1}^{n+1} \alpha_i f(x_i) + R$$

où α_i seront choisi de telle sorte que R soit nul lorsque f est d'un type déterminé.

Lorsque f est quelconque, on suppose que

$$A = \sum_{i=1}^{n+1} \alpha_i f(x_i)$$

est une valeur approchée de I et R est suffisamment petit.

Soient $v_1, v_2, ..., v_{n+1}$ $(n+1)$ fonctions linéairement indépendantes continues sur $[a,b]$. On note $\mathcal{F}_n$ le sous espace engendré par ces fonctions. Pour que R soit nul, si $f \in \mathcal{F}_n$ on obtient :

$$I_k = \int_c^d \omega(x) v_k(x) dx = \sum_{i=1}^{n+1} \alpha_i v_k(x_i); \quad k = 1, ..., n+1.$$

On suppose que les fonctions[1] $v_1, v_2, ..., v_{n+1}$ sont telles que le système admette une soultion unique.

Soit A_n la fonction d'interpolation de f dans $\mathcal{F}_n$ vérifiant :

$$A_n(x) = \sum_{k=1}^{n+1} a_k v_k(x)$$

et

$$A_n(x_i) = f(x_i) \quad \text{pour} \quad i = 1, ..., n+1.$$

nous avons

$$\int_c^d \omega(x) A_n(x) dx = \sum_{i=1}^{n+1} \alpha_i A_n(x_i)$$

si

$$\varepsilon(x) = f(x) - A_n(x)$$

nous obtenons :

$$\int_c^d \omega(x) f(x) dx = \sum_{i=1}^{n+1} \alpha_i f(x_i) + \int_c^d \omega(x) \varepsilon(x) dx$$

Si la fonction d'interpolation $A_n(x)$ est le polynôme d'interpolation de f, c'est-à-dire que :

$$\{v_1(x), v_2(x), ..., v_{n+1}(x)\} = \{1, x, ..., x^n\}$$

On choisira $\{\alpha_1, \alpha_2, ..., \alpha_{n+1}\}$ de telle sorte que R soit nul quand f est un polynôme quelconque de degré inférieur ou égal à n.

1.3 Utilisation de l'interpolation polynomiale

Soit

$$A_n(x) = P_n(x) = \sum_{i=1}^{n+1} L_i(x) f(x_i)$$

avec

$$L_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{x - x_j}{x_i - x_j}$$

d'où

$$P = \int_c^d \omega(x) P_n(x) dx = \sum_{i=1}^{n+1} \left(\int_c^d \omega(x) L_i(x) dx \right) f(x_i)$$

Remarque 80. Sous certaines conditions sur ω, $\{\alpha_i\}$ et $\{x_i\}$ P est une approximation de I.

Remarque 81. Les α_i sont indépendants de f, ils peuvent être calculés une bonne fois pour toute.

1. vérifient les conditions de Haar.

1.4 Etude de l'erreur d'intégration

Si f est $(n+1)$ fois continument dérivable sur $[a,b]$, on sait qu'il existe $\xi_x \in [a,b]$ tel que :

$$\varepsilon(x) = f(x) - A_n(x) = L(x)\frac{f^{(n+1)}(\xi_x)}{(n+1)!}$$

avec

$$L(x) = \prod_{i=1}^{n+1}(x - x_j)$$

en intégrant on obtient :

$$R = I - P = \int_c^d \omega(x)L(x)\frac{f^{(n+1)}(\xi_x)}{(n+1)!}dx$$

si $\omega(x)L(x)$ est de signe constant dans $[c,d]$ (en particulier si $[c,d]$ ne contient aucun des points d'interpolation, avec ω de signe constant sur $[c,d]$) le théorème de la moyenne donne :

$$R = \frac{f^{(n+1)}(\eta)}{(n+1)!}\int_c^d \omega(x)L(x)dx, \qquad \eta \in [c,d]$$

si on connait une borne supérieure de $f^{(n+1)}$

$$M_{n+1} = \sup_{x\in[a,b]} \left|f^{(n+1)}(x)\right|$$

on a

$$|R| \leq \frac{M_{n+1}}{(n+1)!}\int_c^d |\omega(x)L(x)|\,dx$$

1.5 Convergence des méthodes d'intégration

Soit f une fonction continue sur $[a,b]$ et ω une fonction (poids) définie sur $[a,b]$ telle que

$$\int_c^d |\omega(x)|\,dx \leq M$$

on approche

$$I = \int_c^d \omega(x)f(x)dx$$

par

$$A = \sum_{i=1}^{n+1}\alpha_i f(x_i)$$

où α_i et x_i sont à déterminer.

Problème 82. Savoir si en augmentant le nombre de points x_i, on obtiendrait une valeur de A de plus en plus proche de I. Plus précisemment a-t-on :

$$\lim_{n\to\infty}\sum_{i=1}^{n+1}\alpha_i f(x_i) = \int_c^d \omega(x)f(x)dx?$$

La réponse n'est positive que sous certaines conditions sur $\{\alpha_i\}$ et $\{x_i\}$.

Théorème 83. Lorsque f est une fonction continue sur $[a,b]$ nous avons

$$\lim_{n\to\infty}\sum_{i=1}^{n+1}\alpha_i f(x_i)=\int_c^d \omega(x)f(x)dx \tag{1.3}$$

si les conditions suivantes sont vérifiées :

1. l'équation (1.3) est vrai pour un polynôme P quelconque
2. $\sum_{i=1}^{n+1}|\alpha_i|$ est borné pour tout n.

Démonstration. Soit P un polynôme quelconque. Posons

$$\varepsilon_f(x)=\int_c^d \omega(x)f(x)dx-\sum_{i=1}^{n+1}\alpha_i f(x_i)$$

et

$$\varepsilon_P(x)=\int_c^d \omega(x)P(x)dx-\sum_{i=1}^{n+1}\alpha_i P(x_i)$$

on donc

$$\varepsilon_f(x)=\int_c^d \omega(x)P(x)dx+\int_c^d \omega(x)(f(x)-P(x))dx-\sum_{i=1}^{n+1}\alpha_i P(x_i)+\sum_{i=1}^{n+1}\alpha_i(f(x_i)-P(x_i))$$

ou

$$\varepsilon_f(x)=\varepsilon_P(x)+\int_c^d \omega(x)(f(x)-P(x))dx-\sum_{i=1}^{n+1}\alpha_i(f(x_i)-P(x_i))$$

soit $\varepsilon>0$ un nombre déstiné à tendre vers 0. D'après le théorème de Weirstrass [2] nous pouvons prendre un polynôme P tel que

$$\max_{x\in[a,b]}|f(x)-P(x)|\leq\varepsilon$$

nous avons alors

$$\left|\varepsilon_f(x)\right|\leq|\varepsilon_P(x)|+\varepsilon\int_c^d|\omega(x)|\,dx-\varepsilon\sum_{i=1}^{n+1}|\alpha_i|$$

c'est-à-dire

$$\left|\varepsilon_f(x)\right|\leq|\varepsilon_P(x)|+\varepsilon(M-\sum_{i=1}^{n+1}|\alpha_i|)$$

si $\lim_{n\to\infty}|\varepsilon_P(x)|=0$ pour tout polynôme P et si $\sum_{i=1}^{n+1}|\alpha_i|\leq N$ pour tout n on a

$$\lim_{n\to\infty}\left|\varepsilon_f(x)\right|\leq\varepsilon(M+N)$$

cqfd

2. Si f est continue sur [a,b], il existe un polynôme de degré inférieur ou égal à n qui approche f.

1.6 Formules de Newton Cotes

On suppose que f est connue en $(n+1)$ points $x_1, x_2, ..., x_{n+1}$ équidistants, et tels que :

$$\begin{aligned} x_1 &= a,\ x_2 = x_1 + h, \\ &\cdots\cdots\cdots\cdots \\ x_i &= x_{i-1} + h = a + (i-1)h \\ &\cdots\cdots\cdots\cdots \\ x_{n+1} &= x_n + h = a + nh \end{aligned}$$

on prend $\omega(x) = 1$, $\forall x \in [a, b]$.

On pose

$$\int_{x_{1-k}}^{x_{n+1+k}} f(x)dx = \sum_{i=1}^{n+1} \alpha_i f(x_i) + R \tag{1.4}$$

— Si $k = 0$, on dit que la formule (1.4) est de type fermé.
— Si $k = 1$, on dit que la formule (1.4) est de type ouvert.
Les coefficients α_i sont donnés par

$$\alpha_i = \int_{x_{1-k}}^{x_{n+1+k}} L_i(x)dx$$

comme les x_i sont équidistants on peut poser $x = a + th$ et on a :

$$\alpha_i = \int_{x_{1-k}}^{x_{n+1+k}} l_i(t)dt$$

avec

$$l_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{t - j + 1}{i - j}$$

Remarque 84.

$$\alpha_1 = \alpha_{n+1}; \alpha_2 = \alpha_n; \alpha_3 = \alpha_{n-1}; .....\text{et} \quad \sum_{i=1}^{n+1} \alpha_i = b - a + 2kn \tag{1.5}$$

Remarque 85. Les méthodes de Newton Cotes sont convergentes pour les polynômes mais ne le sont pas pour une fonction continue quelconque.

En effet si d'après l'équation (1.5) $\sum_{i=1}^{n+1} \alpha_i$ est bien bornée, il n'en est pas de même de $\sum_{i=1}^{n+1} |\alpha_i|$ car les α_i ne sont pas toujours de même.

En intégrant les formules d'interpolation on a :

1.7 Formule de type fermé : des trapèzes et de Simpson

1. $\int_{x_1}^{x_2} f(x)dx = \frac{h}{2}(f(x_1) + f(x_2)) - \frac{h^3}{12} f''(\xi), \quad \xi \in [x_1, x_2]$ (*formule des trapèzes*)
2. $\int_{x_1}^{x_3} f(x)dx = \frac{h}{3}(f(x_1) + 4f(x_2) + f(x_3)) - \frac{h^5}{90} f''(\xi), \quad \xi \in [x_1, x_3]$ (*formule de Simpson*)
3. $\int_{x_1}^{x_4} f(x)dx = \frac{3h}{8}(f(x_1) + 3f(x_2) + 3f(x_3) + f(x_4)) - \frac{3h^5}{80} f''(\xi), \quad \xi \in [x_1, x_4]$ (*formule de Newton*)

1.8 Formule de type ouvert :

1. $\int_{x_1}^{x_3} f(x)dx = 2hf(x_2) + \frac{h^3}{3}f''(\xi), \quad \xi \in [x_1, x_3]$ (*formule de Poncelet*)
2. $\int_{x_1}^{x_4} f(x)dx = \frac{3h}{2}(f(x_2) + f(x_3)) + \frac{3h^3}{4}f''(\xi), \quad \xi \in [x_1, x_4]$
3. $\int_{x_1}^{x_5} f(x)dx = \frac{4h}{3}(2f(x_2) - f(x_3) + 2f(x_4)) + \frac{14h^5}{45}f^{(4)}(\xi), \quad \xi \in [x_1, x_5]$
4. $\int_{x_1}^{x_6} f(x)dx = \frac{5h}{24}(11f(x_2) + f(x_3) + f(x_4) + 11f(x_5)) + \frac{95h^5}{144}f^{(4)}(\xi), \quad \xi \in [x_1, x_6]$

Remarque 86. Les formules d'intégration données entre x_1 et x_2, x_3 ....peuvent être modifiées pour d'autres points d'interpolation.Exemple : $\int_{x_2}^{x_3} f(x)dx = \frac{h}{2}(f(x_2) + f(x_3)) - \frac{h^3}{12}f''(\xi), \quad \xi \in [x_2, x_3]$

$\int_{x_2}^{x_4} f(x)dx = 2hf(x_3) + \frac{h^3}{3}f''(\xi), \quad \xi \in [x_2, x_4]$

1.9 Intégration par la méthode de Gauss

1.9.1 Polynôme de Legendre

Définition 87. On appelle polynômes de Legendre, les polynômes qui s'écrivent sous la forme :

$$P_n(x) = \frac{1}{2^n n!} \cdot \frac{\partial^n}{\partial x^n}\left[(x^2-1)^n\right], \quad (n = 0, 1, ...)$$

Propriétés : Les polynômes de Legendre vérifient les propriétés suivantes :

1. $P_n(1) = 1$ et $P_n(-1) = (-1)^n \quad (n = 0, 1, 2, ...)$.
2. $\int_{-1}^{1} P_n(x)Q_k(x)dx = 0, \quad (k < n)$, où $Q_k(x)$ est un polynôme de degré $k < n$.
3. Le polynôme de Legendre $P_n(x)$ possède n racines réelles distinctes dans $[-1, 1]$.

Exemple 88.

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) \end{aligned}$$

1.9.2 Formule de quadrature de Gauss

Soit $y = f(t)$ une fonction définie sur $[-1, 1]$.

Problème 89. Comment choisir $t_1, t_2, ..., t_n$ et $A_1, A_2, ...A_n$ pour que la formule de quadrature

$$\int_{-1}^{1} f(t)dt = \sum_{i=1}^{n} A_i f(t_i) \tag{1.6}$$

soit exacte pour tout polynôme $f(t)$ de degré N le plus grand.

Solution 90. Comme on a $2n$ constantes t_i et A_i $(i = 1, 2, ..., n)$ alors que le polynôme de degré $2n-1$ est défini par $2n$ coefficients, ce degré maximal dans le cas général est $N = 2n - 1$.

Posant $\int_{-1}^{1} t^k dt = \sum_{i=1}^{n} A_i t_i^k \quad (k = 0, 1, ..., 2n-1)$ et $f(t) = \sum_{k=0}^{2n-1} C_k t^k$ alors

$$\int_{-1}^{1} f(t)dt = \sum_{k=0}^{2n-1} C_k \int_{-1}^{1} t^k dt = \sum_{k=0}^{2n-1} C_k \sum_{i=1}^{n} A_i t_i^k = \sum_{i=1}^{n} A_i f(t_i)$$

comme

$$\int_{-1}^{1} t^k dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \frac{2}{k+1} & \text{si } k \text{ est pair} \\ 0 & \text{si } k \text{ est impair} \end{cases}$$

pour résoudre le problème il suffit de déterminer t_i et A_i à partir du systéme non linéaire de $2n$ équations

$$\begin{cases} \sum_{i=1}^{n} A_i & = & 2 \\ \sum_{i=1}^{n} A_i t_i & = & 0 \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} A_i t_i^{2n-2} & = & \frac{2}{2n-1} \\ \sum_{i=1}^{n} A_i t_i^{2n-1} & = & 0 \end{cases} \tag{1.7}$$

Pour résoudre le système (1.7), on considère

$$f(t) = t^k P_n(t) \quad (k = 0, 1, ...n-1)$$

où $P_n(t)$ est le polynôme de Legendre.

Les degrés de ce polynôme ne dépassant pas $2n-1$, ces polynômes doivent vérifier :

$$\int_{-1}^{1} f(t)dt = \sum_{i=1}^{n} A_i f(t_i)$$

et

$$\int_{-1}^{1} t^k P_n(t)dt = \sum_{i=1}^{n} A_i t_i^k P_n(t_i) \quad (k = 0, 1, ...n-1)$$

comme

$$\int_{-1}^{1} P_n(x)Q_k(x)dx = 0, \text{ pour } (k < n)$$

on

$$\int_{-1}^{1} t^k P_n(x)dx = 0, \text{ pour } (k < n)$$

et donc

$$\sum_{i=1}^{n} A_i t_i^k P_n(t_i) = 0 \quad (k = 0, 1, ...n-1) \tag{1.8}$$

si l'on pose $P_n(t_i) = 0$ pour $(i = 1, 2, ..., n)$ les égalités (1.8) sont vérifiées pour tout A_i.Si on connait t_i, on trouve à partir du système linéaire des n premières équations du système, les constantes A_i pour $(i = 1, 2, ..., n)$.

Remarque 91. La formule (1.8) où les t_i sont les racines du polynôme de Legendre $P_n(t)$ et où les A_i pour $(i = 1, 2, ..., n)$ sont définis à partir du système, s'appelle *formule de quadrature de Gauss*.

Exemple 92. Trouver la formule de quadrature de Gauss, dans le cas de trois ordonnées ($n = 3$).

Solution 93. Comme le polynôme de Legendre de degré 3 est le polynôme :

$$P_3(t) = \frac{1}{2}(5t^3 - 3t)$$

en l'annulant on obtient ses racines qui sont données par :

$$\begin{aligned} t_1 &= -\sqrt{\frac{3}{5}} \simeq -0,7745 \\ t_2 &= 0 \\ t_3 &= \sqrt{\frac{3}{5}} \simeq 0,7745 \end{aligned}$$

pour la détermination des coefficients A_i pour $(i = 1, 2, 3)$, on obtient le système :

$$\begin{cases} \sum_{i=1}^{3} A_i & = & 2 \\ \sum_{i=1}^{3} A_i t_i & = & 0 \\ \sum_{i=1}^{3} A_i t_i^2 & = & \frac{2}{3} \end{cases}$$

c'est-à-dire

$$\begin{cases} A_1 + A_2 + A_3 & = & 2 \\ -\sqrt{\frac{3}{5}} A_1 + \sqrt{\frac{3}{5}} A_3 & = & 0 \\ \frac{3}{5} A_1 + \frac{3}{5} A_3 & = & \frac{2}{3} \end{cases}$$

dont la solution est : $A_1 = A_3 = \frac{5}{9}$, $A_2 = \frac{8}{9}$.D'où :

$$\int_{-1}^{1} f(t)dt = \sum_{i=1}^{3} A_i f(t_i) = \frac{1}{9}\left[5f(-\sqrt{\frac{3}{5}}) + 8f(0) + 5f(\sqrt{\frac{3}{5}})\right]$$

1.10 Calcul de $\int_a^b f(x)dx$

Pour calculer $\int_a^b f(x)dx$, on fait le changement de variable suivant :

$$x = \frac{b+a}{2} + \frac{b-a}{2}t$$

on obtient :

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f(\frac{b+a}{2} + \frac{b-a}{2}t)dt$$

en appliquant la formule de quadrature de Gauss on a :

$$\int_a^b f(x)dx = \frac{b-a}{2}\sum_{i=1}^{n} A_i f(x_i)$$

où

$$x_i = \frac{b+a}{2} + \frac{b-a}{2}t_i \quad (i = 1, 2, ..., n)$$

et où t_i sont les racines du polynôme de Legendre $P_n(t)$, c'est-à-dire $P_n(t_i) = 0$.

1.11 Erreur de l'intégration par la méthode de Gauss

Le reste de la formule de Lagrange à n points est donné par

$$R_n = \frac{(b-a)^{2n-1}(n!)^4 f^{(4)}(\xi)}{(2n!)^3(2n+1)}$$

d'où l'on tire

$$\begin{aligned} R_2 &= \frac{1}{135}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi) \\ R_3 &= \frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi) \end{aligned}$$

Exemple 94. Calculer par la méthode de quadrature de Gauss à trois ordonnées, l'intégrale suivante :

$$\int_0^1 \sqrt{1+2x}dx$$

Solution 95. Comme $a = 0$ et $b = 1$, d'après le changement de variable :

$$x_i = \frac{b+a}{2} + \frac{b-a}{2}t_i \quad (i = 1, 2, 3)$$

on obtient :

$$\begin{aligned} x_1 &= \frac{1}{2} + \frac{1}{2}t_1 = \frac{1}{2} + \frac{1}{2}.(-\sqrt{\frac{3}{5}}) \simeq 0,112 \\ x_2 &= \frac{1}{2} + \frac{1}{2}t_2 = \frac{1}{2} + \frac{1}{2}.0 \simeq 0,500 \\ x_3 &= \frac{1}{2} + \frac{1}{2}t_3 = \frac{1}{2} + \frac{1}{2}.(\sqrt{\frac{3}{5}}) \simeq 0,887 \end{aligned}$$

donc les coefficients C_i sont :

$$\begin{aligned} C_1 &= \frac{b-a}{2}A_1 = \frac{1}{2}.\frac{5}{9} \simeq 0,277 \\ C_2 &= \frac{b-a}{2}A_2 = \frac{1}{2}.\frac{8}{9} \simeq 0,444 \\ C_3 &= \frac{b-a}{2}A_3 = \frac{1}{2}.\frac{5}{9} \simeq 0,277 \end{aligned}$$

donc

$$\int_0^1 f(x)dx = \int_0^1 \sqrt{1+2x}dx = \sum_{i=1}^{3} C_i f(x_i) = 1,398$$

l'erreur commise dans le calcul de cette intégrale s'évalue de la manière suivante :

$$R_3 = \frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi) \text{ où } \xi \in [0,1]$$

comme $f^{(6)}(x) = -945(1+2x)^{\frac{-11}{2}}$ alors $\max_{x\in[0,1]}\left|f^{(6)}(x)\right| = 945$ donc

$$R_3 \leq \frac{945}{15750}(\frac{1}{2})^7 = 0,5.10^{-3}$$

2 SERIE D'EXERCICES

Exercice 96.

1. Soit f une fonction possédant 4 dérivées continues dans l'intervalle $[0,5]$, donner une évaluation de $\int_0^5 f(x)dx$ sachant que $x_1 = 1$; $x_2 = 2$ et $x_3 = 4$.
2. Soit g une fonction possédant 2 dérivées continues dans l'intervalle $[0,2]$, donner une évaluation de $\int_0^2 (x-1)g(x)dx$ sachant que $x_1 = 0$; $x_2 = 2$.

Exercice 97.

1. Dans l'intervalle $[a,b]$ on prend $x_1 = a$, $x_2 = a + h = b$, $f \in C^2[a,b]$, évaluer la formule des trapèzes $\int_a^b f(x)dx$.

2. On prend $x_1 = a$, $x_2 = a+h$, $x_3 = a+2h, ...x_{n+1} = a+nh = b$, retrouver la formule des trapèzes généralisée.
3. Déduire la valeur approximative de l'intégrale

$$\int_0^1 \frac{dx}{1+x}$$

pour $n = 6$.

4. Calculer la valeur exacte de cette intégrale et déterminer les erreurs absolues et relatives.

Exercice 98. On suppose que f est une fonction 3 fois continument dérivable dans $[-h,h]$ et $f^{(4)}(x)$ continue dans cet intervalle, f est donnée aux points : $x_1 = -h$, $x_2 = 0$, $x_3 = h$.

1. Etablir la formule suivante :

$$\begin{aligned}\int_{-h}^{x} f(t)dt &= \frac{2x^3 - 3hx^2 + 5h^3}{12h^2} f(-h) - \frac{x^3 - 3h^2x - 2h^3}{3h^2} f(0) + \\ &\quad + \frac{2x^3 + 3hx^2 - h^3}{12h^2} f(h) + \varepsilon(x)\end{aligned}$$

On donnera une expression de $\varepsilon(x)$.

2. On suppose que $x \in [-h,x]$:

a) Montrer que $\varepsilon(x)$ peut s'écrire sous la forme :

$$\varepsilon(x) = \frac{1}{24}(x^2 - h^2)^2 f^{(3)}(\zeta),$$

avec $\zeta \in [-h,h]$.

b) Utiliser le résultat précédent pour donner une valeur approchée de :

$$\int_{-\frac{1}{4}}^{0} \frac{dt}{1+t}$$

Quelle est la précision obtenue ?

Comparer ce résultat à celui que donnerait la formule des trapèzes utilisant les points $x_1 = -\frac{1}{4}$ et $x_2 = 0$.

Exercice 99. Déduire la formule de Gauss de la fonction f sur l'intervalle $[-1,1]$ pour le cas de trois ordonnées, on prendra pour la fonction poids $\omega(x) = 1$.

Exercice 100. 1. En utilisant la formule de Gauss à trois ordonnées, calculer l'intégrale :

$$\int_a^b f(x)dx$$

2. En déduire $\int_0^1 \sqrt{1+2x}\, dx$.

3 DÉRIVATION NUMÉRIQUE

3.1 Généralités :

Pour résoudre un certain nombre de problèmes pratiques (étudier la vitesse d'un changement à l'intérieur d'un système par exemple), il est nécessaire parfois de calculer les dérivées d'une fonction $y = f(x)$ supposée dérivable mais connue de façon discrète sur un intervalle $[a, b]$, ou que l'expression analytique compliquée de cette fonction rende difficile sa dérivation.

Comment fournir une valeur approchée de la dérivée, d'ordre un ou supérieur, de $f(x)$ en un point de $[a, b]$.

Le principe est d'approcher la fonction à dériver par un polynôme d'interpolation $P_n(x)$ dont on calcule la dérivée ensuite, c'est-à-dire en posant :

$$f'(x) = P'_n(x)$$

Les dérivées d'ordre supérieur de $f(x)$ s'obtiennent de la même façon.

Si l'on connait l'erreur d'interpolation :

$$\varepsilon(x) = f(x) - P_n(x)$$

l'erreur de la dérivée $P'_n(x)$ est donnée par :

$$r(x) = f'(x) - P'_n(x) = \varepsilon'(x)$$

Soit f une fonction numérique $y = f(x)$ dérivable (respectivement p fois dérivable), donnée aux points équidistants $\{x_1, x_2, ..., x_{n+1}\} \subset [a, b]$ par

$$y_i = f(x_i)$$

Pour chercher sur $[a, b]$ la dérivée $y' = f'(x)$, (respectivement la dérivée d'ordre p c'est-à-dire $y^{(p)} = f^{(p)}(x)$)

Nous allons chercher la valeur de la dérivée sous la forme :

$$y^{(p)}(x) = f^{(p)}(x) = \sum_{i=1}^{n+1} \alpha_i(x) f(x_i) + r(x)$$

les $\alpha_i(x)$ seront choisis de telle sorte que la fonction reste $r(x)$ soit nulle si f est d'un type déterminé (un polymôme).

Si f est quelconque et $r(x)$ suffisamment petit nous considérons que :

$$D(x) = \sum_{i=1}^{n+1} \alpha_i(x) f(x_i)$$

est une approximation de $f^{(p)}(x)$.

Soient $v_1, v_2, ..., v_{n+1}$ $(n+1)$ fonctions linéairement indépendantes p fois continument dérivables sur $[a, b]$. On note $\mathcal{F}_n$ le sous espace engendré par ces fonctions. Pour que $r(x)$ soit nul, il faut que les fonctions $v(x)$ vérifient

$$v^{(p)}(x) = \sum_{i=1}^{n+1} \alpha_i(x) v_k(x_i); \quad k = 1, ..., n+1.$$

On suppose que les fonctions [3] $v_1, v_2, ..., v_{n+1}$ sont telles que le système admette une soultion unique.

3. vérifient les conditions de Haar.

Soit P_n la fonction d'interpolation de f dans $\mathcal{F}_n$ vérifiant :

$$P_n(x) = \sum_{k=1}^{n+1} a_k v_k(x)$$

et

$$P_n(x_i) = f(x_i) \quad \text{pour} \quad i = 1, ..., n+1.$$

nous avons

$$P_n^{(p)}(x) = \sum_{i=1}^{n+1} \alpha_i(x) P_n(x_i)$$

si

$$\varepsilon(x) = f(x) - P_n(x)$$

nous obtenons :

$$f_n^{(p)}(x) = \sum_{i=1}^{n+1} \alpha_i(x) f_n(x_i) + \varepsilon^{(p)}(x)$$

Remarque 101. Si la fonction d'interpolation $P_n(x)$ est le polynôme d'interpolation de f, c'est-à-dire que :

$$\{v_1(x), v_2(x), ..., v_{n+1}(x)\} = \{1, x, ..., x^n\}$$

On choisira $\{\alpha_1(x), \alpha_2(x), ..., \alpha_{n+1}(x)\}$ de telle sorte que $r(x)$ soit nul quand f est un polynôme quelconque de degré inférieur ou égal à n.

3.2 Utilisation de l'interpolation polynomiale

Nous pouvons remplacer la fonction f par son polynôme d'interpolation de Newton (par exemple)

$$\begin{aligned} y &= f(x) = y_1 + k\Delta y_1 + \frac{k(k-1)}{2!}\Delta^2 y_1 + \frac{k(k-1)(k-2)}{3!}\Delta^3 y_1 + \cdots \\ &= y_1 + k\Delta y_1 + \frac{k^2-k}{2}\Delta^2 y_1 + \frac{k^3-3k^2+2k}{6}\Delta^3 y_1 + \cdots \end{aligned}$$

Où

$$k = \frac{x - x_1}{h} \quad \text{et} \quad h = x_{i+1} - x_i \quad (i = 1, 2, ...)$$

Comme

$$y'(x) = \frac{dy}{dx} = \frac{dy}{dk}\frac{dk}{dx} = \frac{1}{h}\frac{dy}{dk}$$

On obtient

$$y'(x) = \frac{1}{h}\left[\Delta y_1 + \frac{2k-1}{2}\Delta^2 y_1 + \frac{3k^2-6k+2}{6}\Delta^3 y_1 + \cdots\right]$$

D'une façon analogue, comme

$$y''(x) = \frac{d(y')}{dx} = \frac{d(y')}{dk}\frac{dk}{dx}$$

on a

$$y''(x) = \frac{1}{h^2}\left[\Delta^2 y_1 + (k-1)\Delta^3 y_1 + \cdots\right]$$

Remarque 102. On procède de la même manière pour chercher les dérivées d'un ordre quelconque.

Remarque 103. Pour chercher les dérivées $y'(x), y''(x)$ en un point fixé x, il faut prendre comme x_1 la valeur tabulée de l'argument la plus proche de ce point x.

Remarque 104. Lorsqu'on cherche la valeur de la dérivée ou des dérivées de y aux points d'interpolation x_i, on peut considérer toute valeur tabulée comme étant une valeur initiale ; on a alors

$$y'(x_1) = \frac{1}{h}\left[\Delta y_1 - \frac{\Delta^2 y_1}{2} + \frac{\Delta^3 y_1}{3} - \frac{\Delta^4 y_1}{4} + \frac{\Delta^5 y_1}{5} - \cdots\right]$$

et

$$y''(x_1) = \frac{1}{h^2}\left[\Delta^2 y_1 - \frac{\Delta^3 y_1}{3} + \frac{11}{12}\Delta^4 y_1 - \frac{5}{6}\Delta^5 y_1 + \cdots\right]$$

3.3 Erreur de dérivation

3.3.1 cas de $\varepsilon' = f' - P_n'$

Soit $P_n(x)$ le polynôme d'interpolation de f. Nous avons

$$P_n(x) = \sum_{i=1}^{n+1} a_i v_i(x) = \sum_{i=1}^{n+1} L_i(x) f(x_i)$$

On sait que pour tout $x \in [a,b]$, il existe $\xi_x \in [a,b]$ tel que

$$\varepsilon(x) = f(x) - P_n(x) = \frac{1}{(n+1)!}\left(\prod_{i=1}^{n+1}(x - x_i)\right) f^{(n+1)}(\xi_x).$$

Par ailleurs $\varepsilon(x) = L(x)g(x)$ avec $L(x) = \prod_{i=1}^{n+1}(x - x_i)$ et $g(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x)$, en dérivant nous obtenons

$$\varepsilon'(x) = f'(x) - P_n'(x) = L'(x)g(x) + L(x)g'(x). \tag{3.1}$$

Le point x pour lequel nous cherchons une approximation peut être :
- soit l'un des points d'interpolation x_i.
- soit un point de $[a,b]$ différent de x_i pour tout i.

L'erreur commise ne sera pas la même dans les deux cas.

a) si $x = x_i$ dans ce cas $L(x_i) = 0$ et

$$L'(x_i) = \lim_{x \to x_i} \frac{\prod_{j=1}^{n+1}(x - x_j)}{x - x_i} = \lim_{x \to x_i} \prod_{\substack{j=1 \\ j \neq i}}^{n+1}(x - x_j) = \prod_{\substack{j=1 \\ j \neq i}}^{n+1}(x_i - x_j)$$

d'où

$$\varepsilon'(x) = \frac{1}{(n+1)!}\left(\prod_{\substack{j=1 \\ j \neq i}}^{n+1}(x_i - x_j)\right) f^{(n+1)}(\xi_{x_i})$$

b) si $x \neq x_i$ pour tout i, dans ce cas nous devons connaitre une estimation de $g\prime(x)$. Si f est $(n+2)$ fois continument dérivable, pour tout $x \in [a,b]$, il existe un élément η_x tel que

$$g\prime(x) = \frac{1}{(n+2)!} f^{(n+2)}(\eta_x)$$

dans ce cas on a :

$$\varepsilon'(x) = \frac{1}{(n+1)!} L'(x) f^{(n+1)}(\xi_{x_i}) + \frac{1}{(n+2)!} L(x) f^{(n+2)}(\eta_x)$$

Si on connait des bornes de $f^{(n+1)}$ et $f^{(n+2)}$, c'est-à-dire si on note

$$M_p = \max_{x\in[a,b]} \left|f^{(p)}(x)\right|$$

nous avons alors

$$\left|\varepsilon'(x_i)\right| \leq \frac{M_{n+1}}{(n+1)!} \prod_{\substack{j=1 \\ j\neq i}}^{n+1} (x_i - x_j)$$

et

$$\left|\varepsilon'(x)\right| \leq \frac{M_{n+1}}{(n+1)!} \left|L'(x)\right| + \frac{M_{n+2}}{(n+2)!} |L(x)| \quad \text{pour } x \in [a,b]; x \neq x_i; i = 1,2,...,n$$

Si $P_m(x)$ est un polynôme de Newton contenant les différences $\Delta y_1, \Delta^2 y_1, ..., \Delta^m y_1$ et si l'erreur correspondante est donnée par :

$$\varepsilon_m(x) = f(x) - P_m(x)$$

l'erreur de la dérivée s'écrit :

$$r(x) = \varepsilon'_m(x) = f'(x) - P'_m(x)$$

comme

$$\varepsilon_m(x) = \frac{(x-x_1)(x-x_2)\cdots(x-x_m)}{(m+1)!} f^{(m+1)}(\xi) = h^{m+1} \frac{k(k-1)\cdots(k-m)}{(m+1)!} f^{(m+1)}(\xi)$$

où $\xi \in [x, x_m]$. En supposant que $f \in C^{(m+2)}$ on obtient :

$$r(x) = \frac{d\varepsilon_m(x)}{dk}\frac{dk}{dx} = \frac{h^m}{(m+1)!}\left[f^{(m+1)}(\xi)\frac{d}{dk}\left[k(k-1)\cdots(k-m)\right] + k(k-1)\cdots(k-m)\frac{d}{dk}f^{(m+1)}(\xi)\right] \tag{3.2}$$

en supposant $\frac{d}{dk}f^{(m+1)}(\xi)$ bornée et tenant compte du fait que $\frac{d}{dk}\left[k(k-1)\cdots(k-m)\right]_{k=0} = (-1)^m m!$, on en tire avec $x = x_1$ et par suite avec $k = 0$,

$$r(x) = (-1)^m \frac{h^m}{(m+1)} f^{(m+1)}(\xi) \tag{3.3}$$

Exemple 105. Calculer $y\prime(50)$ pour la fonction $y = f(x)$ donnée par le tableau suivant :

x	50	55	60	65
y	1,6990	1,7404	1,7782	1,8129

Solution 106. Formons le tableau des différences finies comme suit :

x	y	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
50	1,6990			
		0,0414		
55	1,7404		−0,0036	
		0,0378		0,0005
60	1,7782		−0,0031	
		0,0347		
65	1,8129			

3.3.2 cas de $\varepsilon^{(p)} = f^{(p)} - P_n^{(p)}$

En généralisant l'équation (3.1) on obtient :

$$\varepsilon^{(p)}(x) = \sum_{j=0}^{p} \mathrm{C}_p^j L^{(p-j)}(x) g^{(j)}(x) \tag{3.4}$$

en supposant que f soit $(n+1+p)$ continument dérivable, cette equation (3.4) s'écrit

$$\varepsilon^{(p)}(x) = \sum_{j=0}^{p} \mathbb{C}_p^j L^{(p-j)}(x) \frac{f^{(n+1+j)}(\xi_j)}{(n+1+j)!}, \quad \xi_j \in [a,b]$$

en réecrivant autrement l'équation (3.2) on obtient :

$$\frac{\partial^p}{\partial x^p} f[x, x_1, ..., x_{n+1}] = f\left[\underbrace{x, x, ...x}_{p+1\ fois}, x_1, ..., x_{n+1}\right]$$

on a :

$$\varepsilon^{(p)}(x) = \sum_{j=0}^{p} \mathbb{C}_p^j L^{(p-j)}(x) f\left[\underbrace{x, x, ...x}_{j+1\ fois}, x_1, ..., x_{n+1}\right]$$

lorsque les points sont équidisants c'est-à-dire si : $x_i = x_{i-1} + h = x_1 + (i-1)h$, $i \geq 1$ et si nous posons : $x = x_1 + th$, nous avons :

$$L_i(x) = l_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{(t-j+1)}{i-j}$$

et

$$P_n(x) = p_n(t) = \sum_{i=1}^{n+1} l_i(t) f(x_i)$$

d'où

$$P_n'(x) = p_n'(t) = \sum_{i=1}^{n+1} l_i'(t) f(x_i)$$

et

$$P_n^{(p)}(x) = p_n^{(p)}(t) = \sum_{i=1}^{n+1} l_i^{(p)}(t) f(x_i)$$

Remarque 107. Les coefficients $l_i(t)$, $l_i'(t)$, $l_i^{(p)}(t)$ ne dépendant ni de h ni de x_1, peuvent être tabulés.

Exemple 108. Trouver l'extremum de la fonction donnée par le tableau suivant :

x	1,80	1,82	1,84	1,86	1,88	1,90
y	0,5815170	0,5817731	0,5818649	0,5817926	0,5815566	0,5811571

Solution 109. Le tableau des différences est donnée par le tableau suivant :

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1,80	0,5815170			
		0,002561		
1,82	0,5817731		−0,001643	
		0,000918		0,000002
1,84	0,5818649		−0,001641	
		−0,000723		0,000004
1,86	0,5817926		−0,001637	
		−0,002360		0,000002
1,88	0,5815566		−0,001635	
		−0,003995		
1,90	0,5811571			

l'extremum est atteint pour $f\prime(x) = 0$ c'est-à-dire :

$$0 = \frac{0,000918 - 0,000723}{2} + k(-0,001641) + \frac{3k^2-1}{6}\frac{0,000002 + 0,000004}{2}$$

ou

$$0 = \frac{3}{2}k^2 - 1641k + 97$$

ou aussi

$$k = \frac{97}{1641} + \frac{1}{1094}k^2$$

ce qui donne $k = 0,05911$ d'où $x = x_1 + kh = 1,84 + 0,05911.0,02 = 1,8411822$.

3.4 Algorithmes de dérivation

Les formules de dérivation numériques déduites au paragraphe précédent pour la fonction $y = f(x)$ au point $x = x_1$ ont l'inconvenient de n'utiliser que des valeurs de la fonction pour $x > x_1$.

Les formules de dérivation qui tiennent compte des valeurs de $y = f(x)$ aussi bien pour $x > x_1$ que pour $x < x_1$ sont relativement plus exactes. Ces formules s'appellent *formules de dérivation par différences centrales*.

Soient ..., $x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3$, ...un système de points équidistants et numérotés symétriquement par rapport à x_0, à pas $x_{i+1} - x_i = h$ et $y_i = f(x_i)$ les valeurs correspondantes de $y = f(x)$. Si on pose

$$k = \frac{x - x_1}{h}$$

alors si le le polynôme d'interpolation est le polynôme de Stirling, on aura :

$$\begin{aligned} y &= f(x) = y_0 + k\Delta y_{-\frac{1}{2}} + \frac{k^2}{2}\Delta^2 y_{-1} + \frac{k^2(k^2-1)}{3!}\Delta^3 y_{-\frac{3}{2}} + \frac{k^2(k^2-1)}{4!}\Delta^4 y_{-2} + \\ &+ \frac{k^2(k^2-1)(k^2-2^2)}{5!}\Delta^5 y_{-\frac{5}{2}} + \frac{k^2(k^2-1)(k^2-2^2)}{6!}\Delta^6 y_{-3} + \cdots \end{aligned} \tag{3.5}$$

où $\Delta y_i = y_{i+1} - y_i$; $\Delta y_{-\frac{1}{2}} = \frac{\Delta y_{-1} + \Delta y_0}{2}$, $\Delta^3 y_{-\frac{3}{2}} = \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2}$, $\Delta^5 y_{-\frac{5}{2}} = \frac{\Delta^5 y_{-3} + \Delta y_{-2}}{2}$, etc...

En tenant compte de :

$$\frac{dk}{dx} = \frac{1}{h}$$

on obtient de la formule (3.5) :

$$\begin{aligned} y' &= f'(x) = \frac{1}{h}(\Delta y_{-\frac{1}{2}} + k\Delta^2 y_{-1} + \frac{3k^2-1}{6}\Delta^3 y_{-\frac{3}{2}} + \frac{2k^2-k}{12}\Delta^4 y_{-2} + \\ &+ \frac{5k^4 - 15k^2 + 4}{120}\Delta^5 y_{-\frac{5}{2}} + \frac{3k^5 - 10k^3 + 4k}{360}\Delta^6 y_{-3} + \cdots) \end{aligned}$$

et

$$\begin{aligned} y'' &= f''(x) = \frac{1}{h^2}(\Delta^2 y_{-1} + k\Delta^3 y_{-\frac{3}{2}} + \frac{6k^2-1}{12}\Delta^4 y_{-2} + \\ &+ \frac{2k^3 - 3k}{12}\Delta^5 y_{-\frac{5}{2}} + \frac{15k^4 - 30k^2 + 4}{360}\Delta^6 y_{-3} + \cdots) \end{aligned}$$

en particulier si k=0, on a :

$$y'(x_0) = \frac{1}{h}(\Delta y_{-\frac{1}{2}} - \frac{1}{6}\Delta^3 y_{-\frac{3}{2}} + \frac{1}{30}\Delta^5 y_{-\frac{5}{2}} + \cdots) \tag{3.6}$$

et

$$y''(x_0) = \frac{1}{h^2}(\Delta^2 y_{-1} - \frac{1}{12}\Delta^4 y_{-2} + \frac{1}{90}\Delta^6 y_{-3} + \cdots) \tag{3.7}$$

Exemple 110. Calculer la dérivée $y'(1)$ et la dérivée seconde $y''(1)$ de la fonction donnée par le tableau suivant :

x	0,96	0,98	1,00	1,02	1,04
y	0,7825361	0,7739332	0,7651977	0,7563321	0,7473390

Solution 111. En composant les différences de la fonction $y = f(x)$ on obtient :

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0,96	0,7825361				
		−0,086029			
0,98	0,7739332		−0,001326		
		−0,87355		0,000025	
1,00	0,7651977		−0,001301		0,000001
		−0,88656		0,000026	
1,02	0,7563321		−0,001275		
		−0,89931			
1,04	0,7473390				

et en appliquant (3.6) on a :

$$\begin{aligned} y'(1) &= \frac{1}{0,02}(-\frac{87355+88656}{2}.10^{-7} - \frac{1}{6}.\frac{25+26}{2}.10^{-7} + \frac{1}{30}.1.10^{-7}) = \\ &= -50(88005,5 + 4,2 + 0).10^{-7} = -0,4400485. \end{aligned}$$

et

$$\begin{aligned} y''(1) &= \frac{1}{0,02^2}(-1301.10^{-7} - \frac{1}{12}.1.10^{-7}) = \\ &= -2500.1301.10^{-7} = -0,325250. \end{aligned}$$

Remarque 112. Quand les points d'interpolation sont équidistants, les différences divisées sont remplacées par différences finies.

- On appelle *différences finies* d'ordre 1 *progressives*, notées $\nabla_h f$, la fonction définie par :

$$\nabla_h f(x) = \frac{1}{h}(f(x+h) - f(x))$$

-On appelle *différences finies* d'ordre 1 *régressives*, notées $\bar{\nabla}_h f$, la fonction définie par :

$$\bar{\nabla}_h f(x) = \frac{1}{h}(f(x) - f(x-h))$$

-On appelle *différences finies* d'ordre 1 *centrales*, notées $\delta_h f$, la fonction définie par :

$$\delta_h f(x) = \frac{1}{h}(f(x+\frac{h}{2}) - f(x-\frac{h}{2}))$$

- On définit les *différences finies* d'ordre k *progressives*, notées $\nabla_h^k f$, la fonction définie par :

$$\nabla_h^k f(x) = \nabla_h(\nabla_h^{k-1} f(x))$$

-On appelle *différences finies* d'ordre 1 *régressives*, notées $\bar{\nabla}_h^k f$, la fonction définie par :

$$\bar{\nabla}_h^k f(x) = \bar{\nabla}_h(\bar{\nabla}_h^{k-1} f(x))$$

-On appelle *différences finies* d'ordre 1 *centrales*, notées $\delta_h^k f$, la fonction définie par :

$$\delta_h^k f(x) = \delta_h(\delta_h^{k-1} f(x))$$

Remarque 113. On appelle différences non divisées le produit

$$\nabla^k = \nabla_h^k h^k$$

pour les différences non divisées progressives,

$$\bar{\nabla}^k = \bar{\nabla}_h^k h^k$$

pour les différences non divisées régressives, et

$$\delta^k = \delta_h^k h^k$$

pour les différences non divisées centrales.

3.5 Formules centrales de dérivation

$$f'(x_0) = \frac{1}{2h}(f(x_1) - f(x_{-1})) - \frac{h^2}{6}f^{(3)}(\xi), \quad \xi \in [x_{-1}, x_1]$$

$$f'(x_0) = \frac{1}{12h}(f(x_{-2}) - 8f(x_{-1}) + 8f(x_1) - f(x_2)) + \frac{h^4}{30}f^{(5)}(\xi), \quad \xi \in [x_{-2}, x_2]$$

$$f''(x_0) = \frac{1}{h^2}(f(x_1) - 2f(x_0) + f(x_{-1})) + \frac{h^2}{12}f^{(4)}(\xi), \quad \xi \in [x_{-1}, x_1]$$

$$f''(x_0) = \frac{1}{24h^2}(-2f(x_{-2}) + 32f(x_{-1}) - 60f(x_0) + 32f(x_1) - 2f(x_2)) + \frac{h^4}{90}f^{(6)}(\xi), \quad \xi \in [x_{-2}, x_2]$$

3.6 Formules non centrales de dérivation

$$f'(x_{-1}) = \frac{1}{2h}(-3f(x_{-1}) + 4f(x_0) - f(x_1)) - \frac{h^2}{3}f^{(3)}(\xi), \quad \xi \in [x_{-1}, x_1]$$

$$f'(x_1) = \frac{1}{2h}(f(x_{-1}) - 4f(x_0) + 3f(x_1)) + \frac{h^2}{3}f^{(3)}(\xi), \quad \xi \in [x_{-1}, x_1]$$

$$f'(x_{-1}) = \frac{1}{12h}(-25f(x_2) + 48f(x_{-1}) - 36f(x_0) + 16f(x_1) - 3f(x_2)) + \frac{h^4}{5}f^{(5)}(\xi), \quad \xi \in [x_{-2}, x_2]$$

$$f'(x_1) = \frac{1}{12h}(-f(x_{-2}) + 6f(x_{-1}) - 18f(x_0) + 10f(x_1) + 3f(x_2)) - \frac{h^4}{20}f^{(5)}(\xi), \quad \xi \in [x_{-2}, x_2]$$

$$f'(x_2) = \frac{1}{12h}(3f(x_{-2}) - 16f(x_{-1}) + 36f(x_0) - 48f(x_1) + 25f(x_2)) - \frac{h^4}{5}f^{(5)}(\xi), \quad \xi \in [x_{-2}, x_2]$$

4 SERIE D'EXERCICES

Exercice 114. Soit f une fonction possédant $(n+2)$ dérivées continues dans l'intervalle $[a,b]$, l'erreur d'interpolation polynomiale en $(n+1)$ points est donnée par :

$\epsilon_n(x) = f(x) - p_n(x) = \frac{(x-x_1)(x-x_2)....(x-x_{n+1})f^{(n+1)}(\xi_x)}{(n+1)!}$

1. Calculer $\epsilon_n'(x)$;
2. Donner une évaluation de $\epsilon_n'(x)$ en un point d'interpolation x_k;
3. Pour $n=1, x_1=a, x_2=a+h$, calculer $p_1'(a)$ et $\epsilon_1'(a)$;
4. Pour $n=2, x_1=a, x_2=a+h, x_3=a+2h$, calculer $p_2'(a)$ et $\epsilon_2'(a)$.

Exercice 115. Calculer $y'(0,97)$ de la fonction $y=f(x)$ donnée par le tableau suivant :

x	0,96	0,98	1,00	1,02	1,04
y	0,7825361	0,7739332	0,7651977	0,7563321	0,7473390

Exercice 116. Calculer $y'(50)$ et $y''(50)$ de la fonction $y=\log(x)$ donnée par le tableau suivant :

x	50	55	60	65
y	1,6990	1,7404	1,7782	1,8129

Exercice 117.

1. Soit $f(x)=e^x$, on donne le tableau suivant :

x	0,4	0,6	0,7	1,0
y	1,491825	1,822119	2,013753	2,718282

2. Calculer $f'(0,8)$ et donner une majoration de l'erreur.
3. Calculer $f''(0,8)$ et donner une majoration de l'erreur.

Chapitre

5

RÉSOLUTION DES ÉQUATIONS NON-LINÉAIRES

1 RÉSOLUTION DES ÉQUATIONS NON-LINÉAIRES

Ce chapitre est consacré à quelques méthodes numériques de résolution des équations du type :

$$f(x) = 0 \tag{1.1}$$

où l'application : $f : [a,b] \subset \mathbb{R} \longmapsto \mathbb{R}$ est supposée suffisamment régulière (continue et dérivable) sur l'intervalle $[a,b]$.

C'est-à-dire que nous allons approcher les racines de cette équation (1.1) sur $[a,b]$.

L'équation (1.1) represente une multitude de problèmes (équations algèbriques (où f est un polynôme), trigonométriques, exponentielles...).

Le problème revient donc à trouver x vérifiant $f(x) = 0$ sans qu'on puisse déterminer x explicitement.

Une équation du type (1.1) recouvre beaucoup d'applications.

Comme exemples :

1. On veut déterminer le volume V d'un gaz à une temperature T et une pression P. L'équation d'état qui lie V,T et P est la suivante :

   $$(P + \alpha(\frac{n}{V})^2)(V - n\beta) = knT$$

   où α et β sont des coefficients qui dépendent de la nature du gaz, n le nombre de molécules contenues dans le volume V et k représente la constante de Boltzman. Il est nécessaire donc de résoudre une équation non linéaire où l'inconnue est V. Ce qui revient à trouver les racines de la fonction

   $$f(V) = (P + \alpha(\frac{n}{V})^2)(V - n\beta) - knT$$

   Il s'agit donc de résoudre une équation non linéaire dont on n'est pas capable de trouver une solution exacte.

2. Le lancement d'un projectile. Sa trajectoire est décrite (par la loi de Newton), par une fonction $t \mapsto x(t) = (x_1(t), x_2(t))$ qui doit satisfaire une équation du type

   $$\ddot{x} = F(t, \dot{x}, x).$$

   Où $\ddot{x} = \frac{d^2x}{dt^2}$ est l'acceleration et $\dot{x} = \frac{dx}{dt}$ la vitesse du projectile. Chercher par exemple à savoir à quel moment le projectile retombe sur le sol revient à résoudre :

   $$x_2(t) = 0.$$

   Ainsi, si on dispose d'une méthode numérique pour estimer $x(t)$, on pourra utiliser les méthodes de ce chapitre pour résoudre

   $$x_2(t) = 0.$$

Puisqu'en général la solution d'une équation $f(x) = 0$ ne s'exprime pas par une formule, on ne peut espèrer trouver une solution exacte en un nombre fini d'étapes. Nous allons donc approcher les solutions avec une précision aussi bonne qu'on le souhaite.

Mathématiquement, cela signifie qu'on a une suite $(x_n)_{n\in\mathbb{N}}$ de solutions approchées , c'est-à-dire telle que $x_n \to x^*$ où x^* est une racine de : $f(x^*) = 0$.

Avoir des méthodes pour obtenir des solutions de $f(x) = 0$ de manière approchée est intéressant mais, si on veut les appliquer à des problèmes réels, le temps qu'il faudra attendre pour obtenir la réponse est important.

Par exemple, en ce qui concerne le projectile heurtant le sol, le coût de calcul de $x(t)$ peut être relativement élevé et on voudrait donc que la méthode de résolution de $x_2(t) = 0$ converge en aussi peu d'étapes que possible. En effet, le résultat de ce calcul est peut-être utilisé pour prendre des décisions quant à la trajectoire ultérieure du projectile.

Cette vitesse de convergence s'exprime ici par le gain de précision qu'on gagne en passant de x_n à x_{n+1}.

On s'interesse d'abord aux méthodes de séparation des racines. Il s'agit de determiner des intervalles $[a_i, b_i]$ à l'interieur desquels $f(x)$ admet une racine et une seule.

Cette séparation des racines s'effectue en général :

1. Soit sur le graphe de la fonction $y = f(x)$.
2. Soit les graphes de $y_1 = f_1(x)$ et $y_2 = f_2(x)$, si on peut mettre $f(x) = 0$ sous la forme $f_1(x) - f_2(x) = 0$.
3. Soit en se basant sur le théoreme suivant :

Théorème 118. Pour a et b donnés :

— Si $f(a).f(b) < 0$ alors f admet au moins une racine dans $[a, b]$ si de plus $f'(x) \neq 0$ quelque soit $x \in [a, b]$, la racine est unique.

— Si $f(a).f(b) > 0$ alors f n'admet pas de racine dans $[a, b]$ ou $f(x)$ admet un nombre pair de racines dans $[a, b]$.

Apres avoir isolé une racine dans $[a, b]$, on peut en obtenir une approximation à l'aide de plusieurs méthodes numériques.

Nous allons décrire quelques unes de ces méthodes ci-dessous.

2 MÉTHODE DE BISSECTION OU DE DICHOTOMIE

Cette méthode permet à la fois de montrer l'existence d'une racine d'une fonction $f : [a, b] \longmapsto \mathbb{R}$ et de l'estimer numériquement.

L'idée est : si f est continue et change de signe sur $[a, b]$, f s'annule en un certain point de $[a, b]$.

Définition 119. Choisissons un point quelconque $x_0 \in]a, b[$.

1- Si $f(x_0) = 0$, x_0 est la racine et on a fini.

Sinon, supposons par exemple que $f(a) < 0$ et $f(b) > 0$.

Soit x_0 milieu de $[a, b]$, la racine x^* supposée existante se trouve dans l'un des deux intervalles $[a, x_0]$, $[x_0, b]$, pour savoir lequel, on regarde les conditions du théorème ci-dessus.

2- Si $f(x_0) < 0$, alors il y a une racine dans $]x_0, b[$. On pose dans ce cas $a_1 = x_0$, $b_1 = b$

3- Sinon, $f(x_0) > 0$ et il doit y avoir une racine dans $]a, x_0[$. On pose $a_1 = a$, $b_1 = x_0$. On recommence la procédure en choisissant x_1 dans $[a_1, b_1]$ et ainsi de suite, ce qui donne une suite

décroissante d'intervalles $[a_n, b_n]$ avec $x_0 = \frac{a+b}{2}$, $x_1 = \frac{a_1+b_1}{2}$, ..., $x_n = \frac{a_n+b_n}{2}$ contenant chacun une racine. Et qui verifient :

$$|a - x_n| \leqslant \left(\frac{b-a}{2^{n+1}}\right) \tag{2.1}$$

Remarque 120. L'équation (2.3) permet d'estimer le nombre d'itérations nécessaires pour approcher x^* avec une précision donnée ε .

En effet, si on veut savoir à partir de quel n on a $|x_n - x^*| \leq \varepsilon$, il suffit de chercher n tel que $(1/2^n)|b-a| \leq \varepsilon$. C'est-à-dire $n \geq log_2(|b-a|/\xi)$ où ξ dénote le plus petit entier supérieur ou égal à ε .

Pour que a_n et b_n soient de bonnes approximations d'une racine x^* : $a_n \leqslant x^* \leqslant b_n$ et $a_n \to x^*, b_n \to x^*$. Il faut que la longueur de l'intervalle $[a_n, b_n]$ tende vers 0.

Le théorème donnant le résultat s'énonce comme suit :

Théorème 121. Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. Si $f(a).f(b) < 0$, la fonction f possède au moins une racine dans $]a,b[$. De plus, si on définit par récurrence $[a_0, b_0] = [a,b]$,

$$x_n = \frac{1}{2}(a_n + b_n) \text{ et } [a_{n+1}, b_{n+1}] = \begin{cases} [a_n, x_n] & \text{si} \quad f(a_n)f(x_n) < 0, \\ [x_n, x_n] = \{x_n\} & \text{si} \quad f(x_n) = 0, \\ [x_n, b_n] & \text{si} \quad f(x_n)f(b_n) < 0, \end{cases} \tag{2.2}$$

les trois suites (a_n), (b_n) et (x_n) convergent linéairement vers la même limite x^* avec $f(x^*) = 0$.

Démonstration. On suppose $f(a) < 0$ et $f(b) > 0$. Sinon on remplace f par $-f$.

Montrons par récurrence que $[a_n, b_n]$ est bien défini et que $f\ (a_n)\ f\ (b_n) < 0$ sauf si $a_n = b_n$ dans ce cas $f(a_n) = f(b_n) = 0$.

- Le cas $n = 0$ est trivial.

Supposons que la formule soit vraie pour n et montrons la pour $n+1$.

Soit $a_n = b_n$ sont racines et alors $a_{n+1} = b_{n+1} = a_n$ sont aussi racines. Soit $a_n \neq b_n$ et $f(a_n).f(b_n) < 0$, ce qui implique que

- si $f(a_n).f(x_n) < 0$ ou $f(x_n).f(b_n) < 0$, $a_{n+1} \neq b_{n+1}$ et $f(a_{n+1})f(b_{n+1}) < 0$;

- sinon, $f(a_n).f(x_n) \geq 0$ et $f(x_n).f(b_n) \geq 0$ d'où on déduit que $f(x_n) = 0$ et que $a_{n+1} = b_{n+1} = x_n$ sont des racines de f .

il est facile de constater que :

$$\forall n \in \mathbb{N}, [a_{n+1}, b_{n+1}] \subseteq [a_n, b_n] \quad \text{et} \quad |b_{n+1} - a_{n+1}| \leq \frac{1}{2}|b_n - a_n|.$$

Cela implique que la suite (x_n) est de Cauchy.

Soit $\varepsilon > 0$. Comme $\frac{1}{2n} \to 0$, il existe alors $n_0 \in \mathbb{N}$ tel que $n \geq n_0 \Longrightarrow (1/2n)|b_0 - a_0| \leq \varepsilon$.

Pour les $m \geq n \geq n_0$, on a $x_m \in [a_m, b_m] \subseteq [a_n, b_n]$ et alors

$$|x_m - x_n| \leq |b_n - a_n| \leq \frac{1}{2}|b_{n-1} - a_{n-1}| \leq ... \leq \frac{1}{2^n}|a_0 - b_0| \leq \varepsilon$$

La suite (x_n) est donc bien de Cauchy.

Par conséquent, il existe un $x^* \in [a,b]$ tel que $x_n \to x^*$.

En outre, comme $|x_n - a_n| \leq |b_n - a_n| \to_{n\to\infty} n_0$ et $|b_n - x_n| \leq |b_n - a_n| \to_{n\to\infty} n_0$, il est facile de montrer que (a_n) et (b_n) convergent aussi vers x^*. Puisque $f(a_n)f(b_n) \leq 0$ pour tout n, on en déduit en passant à la limite sur n et en utilisant la continuité de f que $f(x^*)^2 \leq 0$, c'est-à-dire $f(x^*) = 0$.

Nous avons montré que f possède une racine (x^*) et que les suites (a_n), (b_n) et (x_n) convergent toutes trois vers x^* .

Cette convergence est linéaire. Nous allons le voir pour (x_n), (il en est de même pour (a_n) et (b_n)).

Comme $(x_m)_{m\geq n} \subseteq [a_n, b_n]$ et que $x_m \to x^*$, on a $x \in [a_n, b_n]$. En conséquence

$$|x_n - x^*| \leq |b_n - a_n| \leq \frac{1}{2^n}|b_0 - a_0| \tag{2.3}$$

où $c = 1/2 \in]0,1[$. cqfd

3 MÉTHODE DES APPROXIMATIONS SUCCESSIVES (du type $x_{n+1} = F(x_n)$)

Remarque 122. L'équation (1.1) peut toujours se mettre sous la forme

$$x = F(x) \tag{3.1}$$

Il suffit par exemple de poser : $F(x) = x + f(x)$.

Définition 123. Soit $F : \mathbb{R} \to \mathbb{R}$ une fonction numérique.
Si $x \in \mathbb{R}$ est tel que $F(x) = x$, on dit que x est un point fixe de F.

Après avoir isolé une racine dans l'intervalle $[a, b]$, on peut utiliser la proposition suivante pour l'approcher :

Proposition 124. : Soit $F : [a, b] \subset \mathbb{R} \longmapsto [a, b] \subset \mathbb{R}$ une fonction Lipschitzienne de rapport k avec $0 < k < 1$ (on dit dans ce cas, strictement contractante). C'est-à-dire :

$$\forall x, y \in [a, b], |\, F(x) - F(y)\,| \leq k|x - y| \tag{3.2}$$

Alors la suite définie par :

$$x_{n+1} = F(x_n), \forall x_0 \in [a, b] \tag{3.3}$$

converge vers la racine x^* quand n tend vers l'infini.
De plus on a l'estimation de l'erreur comme suit :

$$|x_n - x^*| \leq \frac{k^n}{1-k}|x_1 - x_0| \tag{3.4}$$

Remarques :

Remarque 125. La condition strictement contractante peut être remplacée par :

$$|F'(x)| < 1, \forall x \in [a, b] \tag{3.5}$$

Remarque 126. Si $0 \leq F'(x) < 1$, la suite (x_n) converge vers x^* de façon monotone.

Remarque 127. Si $-1 < F'(x) \leq 0$, la suite (x_n) converge vers x^* alternativement par excès et par defaut.

Remarque 128. Si $|F'(x)| > 1$, la suite (x_n) diverge.

Remarque 129. L'écriture de l'équation (1.1) sous une forme (3.1) quelconque n'est pas unique et ne donne pas toujours une méthode convergente, en effet :
Si on cherche la racine de $\tan x - x = 0$ pour $\pi \leq x \leq \frac{3\pi}{2}$, on écrit soit :

1. $x = \tan x$ c'est-à-dire :

$$F_1(x) = \tan x$$

2. soit $x = \pi + \arctan x$ c'est-à-dire :

$$F_2(x) = \pi + \arctan x$$

On obtient :

$$F_2'(x) = \frac{1}{1+x^2}$$

et $|F_2'(x)| < 1$, la méthode converge. Mais $F_1'(x) = 1 + \tan^2 x$ et $|F_1'(x)| > 1$, la méthode diverge.

Remarque 130. Si on cherche les deux racines de l'équation $x^2 - 6x + 8 = 0$ qui sont $x_1 = 2$ et $x_2 = 4$, on peut écrire :

1. Soit

$$x = \frac{x^2+8}{6}$$

c'est-à-dire :

$$F_1(x) = \frac{x^2+8}{6}$$

2. Ou bien

$$x = \sqrt{6x-8}$$

c'est-à-dire :

$$F_2(x) = \sqrt{6x-8}$$

On obtient :

$$F_1'(x) = \frac{1}{3}x$$

donc

$$|F_1'(x)| < 1, \text{ seulement si } |x| < 3 \tag{3.6}$$

et

$$F_2'(x) = \frac{3}{\sqrt{6x-8}}$$

donc

$$|F_2'(x)| < 1, \text{ seulement si } x > \frac{17}{6} \tag{3.7}$$

Pour l'équation $x = F_1(x)$ on prendra l'intervalle $[0;3]$ ce qui donnera la racine $x^* = 2$, et pour l'équation $x = F_2(x)$ on prendra l'intervalle $[3;5]$ ce qui donnera la racine $x^* = 4$.

4 MÉTHODE DU TYPE $x_{n+1} = x_n - \frac{f(x_n)}{g(x_n)}$

On peut écrire ces algorithmes sous la forme (3.1) avec : $F(x) = x - \frac{f(x)}{g(x)}$

Donc si $|F'(x)| < 1, \forall x \in [a,b]$ ce qui veut dire : $|1 - \frac{g(x)f'(x) - g'(x)f(x)}{g(x)^2} F'(x)| < 1$, le schéma : $x_{n+1} = x_n - \frac{f(x_n)}{g(x_n)}$ converge vers la solution x^* de (1.1) pour x_0 convenablement choisi.

4.1 Méthode de la sécante

Soit c un point de $[a,b]$ tel que $f(c) \neq 0$. On choisit un point initial x_0 tel que $f(x_0).f(c) < 0$. La corde (ou sécante) joignant les points $M_c = (c, f(c))$ et $M_{x_0} = (x_0, f(x_0))$ coupe l'axe des x en un point dont l'abscisse est notée x_1. Et on recommence la procédure avec M_c et $M_1 = (x_1, f(x_1))$. Et ainsi de suite.

On obtient une suite (x_n) définie par :

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n) - f(c)}(x_n - c) = F(x_n).$$

La convergence vers la solution x^* de (1.1) est assurée par un choix convenable de c tel que $|F'(x)| < 1$ dans un voisinage contenant les points (x_n) d'itération.

Remarque 131. Si $f''(x) > 0$ sur $[a,b]$, alors si le point c est tel que $f(c) > 0$, la suite (x_n) est monotone convergente vers la solution x^*, par excès si $f'(x) < 0$ sur $[a,b]$, par défaut si $f'(x) > 0$ sur $[a,b]$.

Si $f''(x) > 0$ sur $[a,b]$, alors si le point c est tel que $f(c) < 0$, la suite (x_n) est monotone convergente vers la solution x^*, par excès si $f'(x) > 0$ sur $[a,b]$, par défaut si $f'(x) < 0$ sur $[a,b]$.

4.2 Méthode de la fausse position ou de Régula-falsi

On peut améliorer la convergence de la méthode de la bissection en s'inspirant de la méthode de dichotomie. L'idée est, au lieu de prendre pour x_n le point milieu de $[a_n, b_n]$, il vaudrait peut-être mieux choisir x_n comme l'intersection du segment de droite joignant $(a_n, f(a_n))$ et $(b_n, f(b_n))$ avec l'axe des $"x"$: $\mathbb{R} \times \{0\}$.

Cela donne la formule :

$$x_n = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)} f(a_n).$$

On peut espérer ainsi que x_n converge plus vite vers x^* . Le désavantage de ce choix est que nous aurons besoin de plus d'hypothèses sur f pour montrer cette convergence.

Théorème 132. Soit $f \in C([a,b];\mathbb{R}) \cap C^1(]a,b[;\mathbb{R})$ une fonction convexe ou concave et $f(a)f(b) < 0$. Définissons a_n, b_n, x_n par la récurrence suivante : $a_0 = a, b_0 = b$ et

$$x_n = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)} f(a_n), \qquad [a_{n+1}, b_{n+1}] = \begin{cases} [a_n, x_n] & \text{si} \quad f(a_n)f(x_n) < 0, \\ [x_n, b_n] & \text{si} \quad f(x_n)f(b_n) < 0, \end{cases} . \tag{4.1}$$

Alors, soit il existe un n tel que $f(x_n) = 0$, soit x_n est bien défini pour tout n et x_n converge à l'ordre 1 vers x^* où x^* est l'unique racine de f dans $[a,b]$.

Cette méthode dite de Regula -Falsi converge dans les mêmes conditions que la méthode de la sécante et en général plus vite.

4.3 Méthode de la tangente ou Méthode de Newton

Soit x_0 un point de $[a,b]$. La tangente à la courbe $y = f(x)$ au point $M_0 = (x_0, f(x_0))$ coupe l'axe des x en un point d'abscisses x_1. En itérant le procédé, on obtient une suite d'abscisses (x_n) définie par :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = F(x_n)$$

La méthode de Newton peut être vue comme un cas limite de la méthode de la sécante où les deux points x_{n-1} et x_n sont tellement proches que $(f(x_n) - f(x_{n-1}))/(x_n - x_{n-1})$ se confond avec $f'(x_n)$. Ainsi on obtiendra x_{n+1} à partir de x_n en regardant l'intersection de la tangente f au point x_n avec l'axe des x. Comme cette tangente est constituée de l'ensemble des points (x,y) tels que

$y = f(x_n) + f'(x_n)(x - x_n)$ et que le point recherché est du type $(x_{n+1}, 0)$. Au voisinage de la racine, cette méthode converge plus vite que la méthode de la sécante. et $x_n \to x^*$ à l'ordre 2.

La condition $|F'(x)| < 1$ dans un voisinage contenant les points (x_n) d'itération est en général satisfaite car :

$$F'(x^*) = \frac{f(x^*).f''(x^*)}{(f'(x^*))^2} = 0$$

Posons dans le voisinage contenant les points (x_n) d'itération, M=sup$|f''(x^*)|$ et m=inf$|f'(x^*)|$. La formule de Taylor dans ce voisinage contenant les points (x_n) d'itération donne l'estimation de l'erreur pour une itération

$$|x_n - x^*| \leq \frac{M}{m} |x_{n-1} - x^*|^2$$

Ce qui donne

$$|x_n - x^*| \leq \left(\frac{M}{m}\right)^{2^n - 1} |x_0 - x^*|^{2^n}$$

Remarque 133. Lorsque la méthode converge, suivant le signe de $f''(x)$, nous avons :

1. $f''(x) > 0$, si $f'(x) > 0$ sur $[a, b]$ (respectivement $f'(x) < 0$) , alors la suite (x_n) est monotone convergente vers la solution x^*, par excès (respectivement par défaut).
2. $f''(x) < 0$, si $f'(x) < 0$ sur $[a, b]$ (respectivement $f'(x) > 0$) , alors la suite (x_n) est monotone convergente vers la solution x^*, par excès (respectivement par défaut).

Remarque 134. 1. Quelques faiblesses de la méthode de Newton lorsqu'on travaille sur de trop grands voisinages de la racine x^* .

Soit $f : [-\pi/2, \pi/2] \to \mathbb{R} : x \mapsto \sin x$. Cette fonction possède une racine simple unique : $x^* = 0$.
La méthode de Newton s'écrit :

$$x_0 \in [-\pi/2, \pi/2], \qquad x_{n+1} = x_n - \tan x_n, \qquad n \geq 0$$

Choisissons $x_0 = \alpha$ où α est la racine strictement positive de $\tan x = 2x$.
Dans ce cas,

$$x_1 = x_0 - \tan x_0 = \alpha - 2\alpha = -\alpha$$

et

$$x_2 = x_1 - \tan x_1 = -\alpha - \tan(-\alpha) = -\alpha + \tan\alpha = -\alpha + 2\alpha = \alpha.$$

On est revenu à x_0. !
Ensuite le processus recommence :

$$x_3 = -\alpha, \quad x_4 = \alpha, \quad x_5 = -\alpha, ...$$

La suite $(x_n)_{n\in\mathbb{N}}$ alterne donc entre α et $-\alpha$.
On dit que c'est une *orbite périodique* de période 2 ou un *cycle d'ordre deux*. En conséquence (x_n) ne converge pas vers 0. Cela met en évidence qu'on doit partir suffisament près de la racine afin d'assurer la convergence de la méthode.
Ici on peut montrer que, si $|x_0| < \alpha$, alors $x_n \to 0 = x^*$.

2. La méthode de Newton n'est pas nécessairement plus performante que les autres méthodes si on est trop loin de la racine. Il est donc important de determiner un intervalle $[a, b]$, contenant la racine x^*, le plus petit possible, de manière à ce que x_0 soit le plus proche possible de x^* . Sinon l'algorithme peut converger vers une autre racine ou même diverger.

5 MÉTHODE DU POINT FIXE

Si on regarde la méthode de Newton d'un point de vue abstrait, on voit qu'on obtient x_{n+1} à partir de x_n en évaluant toujours la même expression. Plus précisément, on a $x_{n+1} = F(x_n)$ avec $F(x) = x - f(x)/f'(x)$. Si $x_n \to x^*$, on déduit immédiatement de la continuité de F que $x^* = F(x^*)$. On dit alors que x^* est un *point fixe* de F. Or, il se fait que les points fixes de F correspondent aux racines simples de f .

Nous disposons maintenant d'un cadre pour rechercher de nouveaux algorithmes.

En effet, à toute fonction F dont les points fixes correspondent aux solutions du problème, on peut associer un schéma récursif $x_{n+1} = F(x_n)$.

Quelles sont donc les propriétés que F doit posséder pour être intéressante ? C'est-à-dire :

1. On doit avoir (x_n) convergente et sa limite x^* sera alors un point fixe de F ;
2. Et (x_n) doit tendre vers x^*aussi vite que possible et l'ordre de convergence doit être aussi élevé que possible.

Théorème 135. Soit $F : [a,b] \to [a,b]$ une fonction. On suppose qu'il existe une constante $K \in [0,1[$ telle que

$$\forall x, y \in [a,b], \qquad |F(x) - F(y)| \leq K|x - y|. \tag{5.1}$$

Alors, F possède un unique point fixe $x^* \in [a,b]$ et pour tout $x_0 \in [a,b]$, la suite $(x_n)_{n\in\mathbb{N}}$ 0 définie par $x_{n+1} = F(x_n)$ converge vers x^* .

Remarque 136. Une fonction qui satisfait (5.1) pour $K \in [0,+\infty[$ est dite lipchitzienne. Lorsque $K < 1$, on dit que F est une contraction.

Remarque 137. Le plus petit K qui satisfait (5.1) (le K optimal) est appelé la constante de Lipschitz de la fonction F et se note $Lip(F)$. Ainsi

$$|(x) - (y)|Lip(F) = Lip_{[a,b]}(F) = sup_{x,y\in[a,b]x\neq y} \frac{|F(x) - F(y)|}{\left|x - y\right|}.$$

Remarque 138. Les fonctions qui satisfont (5.1) sont continues. L'inverse n'est pas vrai.

Remarque 139. Si $F \in C^1(]a,b[;\mathbb{R})$, on peut montrer grâce au théorème de la moyenne que

$$Lip(F) = sup_{x]a,b[}|F(x)|.$$

En conséquence, F sera une contraction si et seulement si $sup_{x\in]a,b[}|F'(x)| < 1$. Si de plus F est dérivable en a et b, il découle de la compacité de $[a,b]$ que F est une contraction si et seulement si, pour tout$x \in [a,b]$, $\quad |F'(x)| < 1$.

Remarque 140. Du point de vue de l'existence, l'intérêt de ce théorème est qu'il est valable en dimension supérieure à 1. En effet, en dimension 1, la continuité suffit. Notons cependant que, dans ce cas, la convergence des suites (x_n) n'est pas assurée et en fait n'a pas nécessairement lieu. Leur comportement peut d'ailleurs être fort complexe .

Le théorème (135) donne un critère pour la convergence des suites sur un intervalle $[a,b]$.Et on peut l'appliquer au voisinage d'un point fixe. On en conclut que pour tout $x_0 \in I_\varepsilon$, la suite $(x_n)_{n\in\mathbb{N}}$ définie par $x_{n+1} = F(x_n)$ converge bien vers x^* .

Remarque 141. Si $|F'(x^*)| < 1$, les suites qui entrent dans un petit voisinage de x^* convergent vers x^* . On dit que x^* est un point fixe *attractif*.

Si $|F'(x^*)| > 1$, même si une suite entre dans un petit voisinage de x^* , elle est forcée d'en ressortir. On dit de x^* que c'est un point fixe *répulsif*.

Si $|F'(x^*)| = 1$, on ne peut rien dire Les deux situations ci-dessus peuvent se produire. Ou aucune d'elles. Cependant on peut penser $|F'(x^*)| = 1$ comme une transition entre $|F'(x^*)| < 1$ et $|F'(x^*)| > 1$, c'est à dire entre un point fixe qui était attractif et devient répulsif. De telles situations sont communes et, typiquement, lorsque $|F'(x^*)| = 1$, une *bifurcation* a lieu.

Nous avons examiné la convergence – ou non – des suites vers un point fixe. Nous voudrions aussi connaitre la vitesse de convergence de x_n vers x^* . Globalement, le théorème (135) ne nous offre qu'une convergence linéaire. En effet, l'équation (5.1) implique

$$|x_{n+1} - x^*| = |F(x_n) - F(x^*)| \leq K|x_n - x^*|.$$

Lorsqu'on est suffisamment proche du point fixe x^*, la méthode de Newton est quadratique. En faisant un développement deTaylor avec reste de F au point x^*. On écrit

$$F(x) = F(x^*) + F(x^*)(x - x^*) + \cdots + \frac{F^{(k-1)}(x^*)}{(k-1)!}(x - x^*)^{k-1} + \frac{F'^{(k)}(\xi)}{k!}(x - x^*)^k$$

Et nous avons le théorème suivant :

Théorème 142. Sous les hypothèses du théorème (135), si de plus on a $F \in C^k(]a, b[; \mathbb{R})$ et $F'(x^*) = 0, ..., F^{(k-1)}(x^*) = 0$, alors (x_n) converge vers x^* à l'ordre k. Plus précisément, on a

$$|x_{n+1} - x^*| \leq c|x_n - x^*|^k$$

où $c > |kF^{(k)}(x^*)|/k!$ peut être choisi arbitrairement proche de $|F^{(k)}(x^*)|/k!$.

6 SERIE D'EXERCICES

Exercice 143. Montrer en utilisant la propriété de valeur intermédiaire que toute fonction continue $f : [a, b] \to [a, b]$ possède au moins un point fixe.

Exercice 144. Utiliser l'algorithme de dichotomie pour calculer à 0.01 près la racine de :

$$f(x) = e^x \sin x - 1$$

dans l'intervalle $[0, \pi/2]$.

Exercice 145. En utilisant une méthode de convergence de la forme $x_{n+1} = F(x_n)$, trouver la racine à 0.01 près de :

$$f(x) = xe^x - 1 = 0$$

dans l'intervalle $[1/2, 1]$.

Exercice 146. On considère la fonction f définie par $f(x) = x^3 + x - 1$, $x \in \mathbb{R}$.

1. Montrer que $f(x) = 0$ admet une racine réelle unique $x^* \in]0, 1[$.
2. Déterminer par la méthode de dichotomie, une approximation de x^* à 10^{-1} près en utilisant le test d'arrêt $|x_{n+1} - x_n| \leq \epsilon$. Comparer le nombre d'itérations effectif pour avoir cette précision avec le nombre $N = (\frac{\log(\frac{b-a}{\epsilon})}{\log 2}) = 3$.
3. Effectuer deux itérations avec la méthode de Newton en partant de $x_0 = 1$.

Exercice 147. En utilisant la méthode de Newton, chercher la racine à 0.001 près de l'équation :

$$f(x) = x^4 + x^2 + 2x - 1 = 0$$

dans l'intervalle $[0, 1]$.

Exercice 148. Trouver par la méthode de Newton, la racine positive minimale de l'équation :

$$\tan x = x$$

0.0001 près.

Exercice 149.

1. Trouver le nombre de racines réelles distinctes de :

$$P_3(x) = x^3 - 3x^2 - x + 3$$

2. Existent-t-ils des racines multiples ?

Exercice 150.

1. Trouver le nombre de racines réelles distinctes de :

$$P_3(x) = x^3 - 5x^2 + 7x - 3$$

2. Existent-t-ils des racines multiples ?

Exercice 151.

1. Trouver le nombre de racines réelles distinctes de :

$$P_3(x) = x^4 - x^3 - 3x^2 + 5x - 2$$

2. Existent-t-ils des racines multiples ?

Exercice 152. Résoudre graphiquement l'équation cubique :

$$x^3 - 1.75x + 0.75 = 0$$

Exercice 153.

1. Trouver par la méthode de Krylov, le polynôme caractéristique de la matrice suivante :

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ 3 & 1 & 2 \end{pmatrix}$$

2. Localiser les différentes valeurs propres.
3. Donner, par la méthode de Newton, une estimation de la valeur propre négative à 0.001 près.

Exercice 154. On considère la fonction $f(x) = 4 + 8x^2 - x^4$.

1. Combien f possède-t-elle de racines ? Si on décide d'utiliser la méthode de bissection, quelles sont les paires de points initiaux qu'on peut choisir pour obtenir chacune des racines ?
2. Si on opte pour la méthode de Newton, donner des intervalles autour de chacune des racines sur lesquels la méthode de Newton converge.

Exercice 155. L'équation $x^3 + 4x^2 - 10 = 0$ peut se réécrire sous la forme d'un point fixe des trois façons suivantes :

$$\begin{aligned} x &= \varphi_1(x) = \sqrt{\frac{10 - x^3}{4}} \\ x &= \varphi_2(x) = \frac{10}{x^2 + 4x} \\ x &= \varphi_3(x) = \sqrt{\frac{10}{x + 4}} \end{aligned}$$

1. Montrer que l'équation ci-dessus possède une unique racine (qui est positive) et donc que φ_i, $i = 1, 2, 3$, possèdent un seul point fixe.
2. Calculer les dix premières itérées des suites (x_n) définies par $x_{n+1} = \varphi_i(x_n)$ et $x_0 = 1$ pour $i = 1, 2, 3$. Qu'en déduire ?
3. Tracer les graphes des fonctions φ_i. Comment les comportements observés ci-dessus se voient-ils sur ces graphiques ? Observer également la vitesse de convergence.

Exercice 156. En mécanique céleste, le calcul des positions planétaires donne lieu à l'équation de Képler :

$$m = x - Esin(x)$$

où nous allons considérer les valeurs $m = 0,8$ et $E = 0,2$. Utilisez la méthode du point fixe pour résoudre cette équation en partant des valeurs initiales $x_0 = 1$, 0 et -1 respectivement.

7 RÉSOLUTION DES SYSTÈMES D'ÉQUATIONS NON-LINÉAIRES

7.1 Résolution d'une équation algèbrique

Soit

$$P_x(x) = a_1 x^n + a_2 x^{n-1} + a_3 x^{n-2} + ... + a_n x + a_{n+1} = \sum_{i=1}^{n+1} a_i x^{n+1-i}$$

un polynôme de degré inferieur ou égal à n.

On suppose que tous les coefficients a_i sont réels.

Il existe différentes méthodes pour chercher les racines de ce polynôme, c'est à dire chercher x qui vérifie

$$P(x) = 0$$

.

Nous allons donner quelques méthodes qui nous permettront de chercher les racines réels supposées existantes de ce polynôme.

7.2 Propriétés sur les racines d'un polynôme

Si $x_1, x_2, x_3, ...x_n$ sont les n racines de $P_n(x) = 0$, on a les propriétés suivantes qui sont vérifiées (d'après le théorème de d'Alermbert).

$$\begin{cases} \sum_{i=1}^{n} x_i & = & -\frac{a_2}{a_1} \\ \sum_{i=1}^{n-1} x_i (\sum_{j=i+1}^{n} x_i) & = & \sum_{1 \le i_1 < i_2 \le n} x_{i_1} x_{i_2} \quad = \quad \frac{a_3}{a_1} \\ ..... & = & ..... \\ \sum_{1 \le i_1 < i_2 < ....< i_p \le n} x_{i_1} x_{i_2} ..... x_{i_{p-1}} x_{i_p} & = & (-1)^p \frac{a_{p+1}}{a_1} \\ ..... & = & ..... \\ x_1 x_2 ..... x_{n-1} x_n & = & (-1)^n \frac{a_{n+1}}{a_1} \end{cases}$$

En posant $S_k = \sum_{i=1}^{n} x_i^k$, on obtient les relations dites de Newton :

$$\begin{cases} a_1 S_1 + a_2 & = & 0 \\ a_1 S_2 + a_2 S_1 + 2a_3 & = & 0 \\ ..... & = & ..... \\ a_1 S_n + a_2 S_{n-1} + ... + a_n S_1 + n a_{n+1} & = & 0 \end{cases}$$

7.3 Théorème de Sturm

Le théorème de Sturm permet de calculer le nombre de racines réelles distinctes d'un polynôme dans un intervalle donné.

7.3.1 Suite de Sturm

On se donne un polynôme $P = x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$. La suite de Sturm (ou chaîne de Sturm à partir du polynôme P) est une suite finie de polynômes $P_0, P_1, ..., P_m$. Elle est construite par récurrence :

$P_0 = P$;

$P_1 = P\prime$, où P'est la dérivée de P, c'est-à-dire le polynôme $P' = nx^{n-1} + ... + a_1$;

Pour $i \geq 2$, P_i est l'opposée du reste de la division de P_{i-2} par P_{i-1}.

La construction s'arrête au dernier polynôme non nul.

Pour obtenir cette suite, on calcule les restes intermédiaires que l'on obtient en appliquant l'algorithme d'Euclide à P_0 et sa dérivée P_1 :

$$\left\{\begin{array}{ccc} P_0 & = & P_1Q_1 - P_2 \\ P_1 & = & P_2Q_2 - P_3 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ P_{m-2} & = & P_{m-1}Q_{m-1} - P_m \\ P_{m-1} & = & P_mQ_m \end{array}\right.$$

Si P possède uniquement des racines distinctes, le dernier terme est une constante non nulle. Si ce terme est nul, P admet des racines multiples, et on peut dans ce cas appliquer le théorème de Sturm en utilisant la suite $T_0, T_1, ..., T_{m-1}, 1$ que l'on obtient en divisant les $P_1, P_2, ..., P_{m-1}$ par P_m.

Et le nombre de racines réelles de $P_n(x) = 0$ supposées distinctes est donné donc par le théorème suivant :

Théorème 157 (Théorème de Sturm). Le nombre de racines réelles distinctes dans un intervalle $[a, b]$ d'un polynôme à coefficients réels, dont a et b ne sont pas des racines, est égal au nombre de changements de signe de la suite de Sturm aux bornes de cet intervalle.

Plus formellement, si nous notons $N(y)$ le nombre de changements de signe (zéro n'est pas compté comme un changement de signe) observés dans la suite $P(y), P_1(y), P_2(y), \ldots, P_m(y)$ alors le nombre de racines réelles distinctes de l'équation dans l'intervalle $[a, b]$ (où a et b ne sont pas des racines de P) est donné par $N = N(a) - N(b)$.

Remarque 158. Si l'equation $P_n(x) = 0$ admet une racine multiple, soit $(j + 1)$ le premier indice tel que $P_{j+1}(x) = 0$. Les racines de $P_n(x) = 0$ seront alors les racines simples de $P_j(x) = 0$. Le nombre de racines distinctes est donné par le théorème de Sturm en arretant la suite $(p_n(y))$ au terme $p_i(y)$.

Exemple 159. Supposons que l'on souhaite connaître le nombre de racines dans un certain intervalle du polynôme $p(x) = x^4 + x^3 - x - 1$.

On commence par calculer les deux premiers termes.

$p_0(x) = p(x) = x^4 + x^3 - x - 1$

$p_1(x) = p\prime(x) = 4x^3 + 3x^2 - 1$ {\{\begin{aligned}p_{0}(x)& = p(x) = x^{4}+x^{3}-x-1\\p_{1}(x)& = p\prime(x) = 4x^{3} + 3x^{2} - 1\end{aligned}}}

En divisant p_0 par p_1 on obtient le reste $-\frac{3}{16}x^2 - \frac{3}{4}x - \frac{15}{16}$, et en le multipliant par -1 on obtient $p^2(x) = \frac{3}{16}x^2 + \frac{3}{4}x + \frac{15}{/16}$. Ensuite, on divise p_1 par p_2 et en multipliant le reste par -1, on obtient $p_3(x) = -32x - 64$. Puis on divise p_2 par p_3 et en multipliant le reste par -1, on obtient $p_4(x) = -\frac{3}{16}$.

Finalement, la suite de Sturm du polynôme P est donc :

$p_0(x) = x^4 + x^3 - x - 1$

$p_1(x) = 4x^3 + 3x^2 - 1$

$p_2(x) = \frac{3}{16}x^2 + 34x + \frac{15}{16}$

$p_3(x) = -32x - 64$ $p_4(x) = -\frac{3}{16}$

Pour trouver le nombre de racines totales, c'est à dire entre $-\infty$ et $+\infty$, on évalue p_0, p_1, p_2, p_3, et p_4 en $-\infty$ et on note la séquence de signes correspondante :$+ - + + -$. Elle contient trois changements de signe (+ à −, puis − à +, puis + à −).

On fait la même chose en $+\infty$ et obtient la séquence de signes $+ + + - -$, qui contient juste un changement de signe. D'après le théorème de Sturm, le nombre total de racines du polynôme P est $3 - 1 = 2$. Nous pouvons faire une vérification en remarquant que $p(x) = x^4 + x^3 - x - 1$ se factorise en$(x^2 - 1)(x^2 + x + 1)$, où on voit que $x^2 - 1$ a deux racines (-1 et 1) alors que $x^2 + x + 1$ n'a pas de racines réelles.

Exemple 160. Soit $P_3(x) = x^3 + 2x^2 - x - 2$. Alors

$p_1(x) = x^3 + 2x^2 - x - 2$
$p_2(x) = 3x^2 + 4x - 1$
$p_3(x) = 7x + 8$ (à un facteur multiplicatif positif près)
$p_4(x) = 81$ (à un facteur multiplicatif positif près)

Qu'on peut mettre sous forme du tableau suivant :

y	$p_1(y)$	$p_2(y)$	$p_3(y)$	$p_4(y)$	$N(y)$
-3	$-$	$+$	$-$	$+$	3
0	$-$	$-$	$+$	$+$	1
3	$+$	$+$	$+$	$+$	0

Il y a donc $3-1=2$ racines réelles distinctes dans $[-3,0]$
et $3-0=3$ racines réelles distinctes dans $[-3,3]$

Exemple 161. Soit $P_6(x) = x^6 + 4x^5 + 4x^4 - x^2 - 4x - 4$. Alors

$p_1(x) = x^6 + 4x^5 + 4x^4 - x^2 - 4x - 4$.
$p_2(x) = 6x^5 + 20x^4 + 16x^3 - 2x - 4$
$p_3(x) = 4x^4 + 8x^3 + 3x^2 + 14x + 16$
$p_4(x) = x^3 + 6x^2 + 12x + 8$
$p_5(x) = -17x^2 - 58x - 48$
$p_6(x) = -x - 2$
(Tous ces polynôme sont définis à un facteur multiplicatif positif près)
Qu'on peut mettre sous forme du tableau suivant :

y	$p_1(y)$	$p_2(y)$	$p_3(y)$	$p_4(y)$	$p_5(y)$	$p_6(y)$	$N(y)$
-3	$+$	$-$	$+$	$-$	$-$	$+$	4
-2	0	0	0	0	0	0	
-1	0	$-$	$+$	$+$	$-$	$-$	2
0	$-$	$-$	$+$	$+$	$-$	$-$	2
1	0	$+$	$+$	$+$	$-$	$-$	1
2	$+$	$+$	$+$	$+$	$-$	$-$	1
3	$+$	$+$	$+$	$+$	$-$	$-$	1

Il y a donc $3-1=2$ racines réelles négatives distinctes dans $[-3,0]$, et une racine réelle distinctes dans $[0,3]$. Comme $p_6(x) = -x-2$ la valeur -2 est une racine double.

8 RÈSOLUTION DE SYSTÈMES NON LINÉAIRES

Nous allons maintenant résoudre le système de m équations à m inconnues $x_1, x_2, ..., x_m$ de la forme :

$$\begin{cases} f_1(x_1, x_2, ..., x_m) &= 0 \\ f_2(x_1, x_2, ..., x_m) &= 0 \\ ..... &= 0 \\ f_m(x_1, x_2, ..., x_m) &= 0 \end{cases} \tag{8.1}$$

Définition 162. Le système (8.1) peut toujours se mettre sous la forme :

$$F(x) = x \ , \ x \in \mathbb{R}^m \ , F : \mathbb{R}^m \longmapsto \mathbb{R}^m$$

x est dit point fixe de $F(x)$.

Le théorème suivant assure l'existence et l'unicité d'un point fixe. Il nous permettra d'approcher les solutions de systèmes algébriques non linéaires.

Théorème 163 (du point fixe). Soit E un espace métrique complet non vide, $F : E \longmapsto E$ une contraction stricte. Alors F admet un point fixe et un seul donné par la méthode des approximations successives :

$$x_{n+1} = F(x_n)$$

pour x_0 quelconque.

Démonstration. **1- Existence :** Soit la suite $(x_n) \in E$ définie par $x_{n+1} = F(x_n)$. Nous allons montrer que la suite (x_n) est de Cauchy.

Comme F est une contraction, nous avons :

$$\begin{aligned} d(x_2, x_1) &\leq k d(x_1, x_0) \\ d(x_3, x_2) \leq k d(x_2, x_1) &\leq k^2 d(x_1, x_0) \\ \cdots\cdots \cdots &\leq \cdots\cdots\cdots \\ d(x_{n+1}, x_n) \leq k d(x_n, x_{n-1}) &\leq k^n d(x_1, x_0) \end{aligned}$$

Ce qui nous donne

$$\begin{aligned} d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n) \\ d(x_{n+p}, x_n) &\leq k^n (k^{p-1} + k^{p-2} + k^{p-3} \cdots + k + 1) d(x_1, x_0) \leq \frac{k^n}{1-k} d(x_1, x_0) \end{aligned}$$

Nous en déduisons que $d(x_{n+p}, x_n) \xrightarrow{0}$ quand $n \xrightarrow{+} \infty$. Donc la suite (x_n) est fde Cauchy et par suite admet une limite x et comme F est continue, $F(x_n)) \xrightarrow{x}$. On a donc $F(x) = x$.

2- Unicité Si x et y sont deux points fixes, on doit avoir

$$d(x, y) \leq k d(x, y) < d(x, y)$$

si $d(x, y) \neq 0$.

On a donc nécessairement $d(x, y) = 0$ et $x = y$.

cqfd

Proposition 164. Soit $F : E \longmapsto E$. Posons $F_2 = FoF$, $F_3 = FoFoF$, ...,$F_p = FoF_{p-1}$; F_p est appelée l'iterée d'ordre p de F. Nous avons alors le résultat suivant : Si l'une des iterées F_p est strictement contractante, alors F admet un point fixe unique.

Remarque 165.

1. Le procédé $x_{n+1} = F(x_n)$ est un algorithme permettant de trouver le point fixe de F. De plus la suite (x_n) converge rapidement vers x, car

$$d(x_n, x) = \lim_{p \longmapsto +\infty} d(x_n, x_{n+p}) \leq \frac{k^n}{1-k} d(x_1, x_0)$$

2. L'itérée F_p contraction stricte n'implique pas nécessairement que F soit continue ou contractante.
3. La condition $k < 1$ de contraction stricte, est indispensable. Car $k \leq 1$ ne suffit pas pour garantir ni l'existence ni l'unicité du point fixe.

8.1 Méthode des approximations successives (type Jacobi ou Gauss-Seidel)

L'équation

$$x = F(x); \quad x \in \mathbb{R}^m, \quad F : \mathbb{R}^m \longmapsto \mathbb{R}^m$$

peut s'écrire sous la forme :

$$A(x) = b; \quad x \in \mathbb{R}^m; \quad b \in \mathbb{R}^m; \quad A : \mathbb{R}^m \longmapsto \mathbb{R}^m$$

ou bien sous la forme

$$x = B(x) + c; \quad c \in \mathbb{R}^m; \quad B : \mathbb{R}^m \longmapsto \mathbb{R}^m$$

ou sous la forme d'un système de m équations :

$$x_i = B_i(x) + c_i; \quad i = 1, 2, ..., m$$

— S'il existe un domaine Ω convexe contenant x solution de $x = F(x)$ tel que

$$\forall x \in \Omega, \quad \sum_{j=1}^{m} |\frac{\partial B_i}{\partial x_i}| \leq d < 1; \quad i = 1, 2...., m$$

alors pour tout vecteur initial $x^{(0)}$ pris dans Ω, la suite de vecteurs $x^{(k)}$ définis par le schema itératif

$$x^{(k+1)} = B(x^{(k)}) + c \tag{8.2}$$

converge vers x d'après le théorème du point fixe. Le schema (8.2) s'appelle *"Methode de Jacobi non lineaire"*.

— Sous les mêmes hypothèses, la suite de vecteurs $x^{(k)}$ définis par le schema itératif :

$$x_i^{(k+1)} = B_i(x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}, x_i^{(k)}, ..., x_m^{(k)}) + c_i; \quad i = 1, 2, ..., m \tag{8.3}$$

converge vers x, d'après le théorème du point fixe.
Le schema (8.3) s'appelle *"Methode de Gauss-Seidel non lineaire"*.

Chapitre

6

RESOLUTION D'UN SYSTEME LINEAIRE

1 METHODES DIRECTES

1.1 Rappel

1.1.1 Rang d'une matrice

Définition 166. Soit $A \in M_{m,n}(\mathbb{R})$ une matrice de m lignes et de n colonnes. Le rang de A est égal au nombre maximum de colonnes de A linéairement indépendant et on le note :

$$rang\ A$$

Le rang de A est donc aussi égal à la dimension de l'image de toute application linéaire représenté par A.

Proposition 167. - $rang\ A = rang\ A^t$ - si $A \in M_{m,n}(\mathbb{R})$ alors $rang\ A \leq \inf(m,n)$ - si $A \in M_n(\mathbb{R})$ alors A inversible $\Leftrightarrow$ $rang\ A = n$.

Définition 168. Soit $A \in M_{m,n}(\mathbb{R})$, on appelle matrice extraite de A une matrice obtenue par sélection de lignes et de colonnes. Par exemple si I (respectivement J) est un sous ensemble de $\{1,2,...m\}$ (respectivement $\{1,2,...n\}$) on définit une matrice extraite B de A par :

$$B = \left(a_{ij}\right)_{i\in I, j\in J}$$

Théorème 169. Soit $A \in M_{m,n}(\mathbb{R})$ une matrice de *rang* r alors :
- le rang de toute matrice extraite de A est inférieur ou égal à r,
- toute matrice carrée inversible extraite de A est d'ordre inf érieur ou égal à r,
- il existe une matrice carrée extraite de A d'ordre r qui est inversible.

1.2 Systèmes linéaires

Définition 170. On appelle système de m équations à n inconnues $x_1, x_2, ..., x_n$ la famille d'équations :

$$\left\{\begin{array}{ccc} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n & = & b_2 \\ ..... & ... & ... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n & = & b_m \end{array}\right. \tag{1.1}$$

que l'on peut mettre sous la forme matricielle suivante :

$$Ax = b$$

où $A \in M_{m,n}(\mathbb{R})$ est une matrice donnée par ses éléments (a_{ij}), $1 \leq i \leq m, 1 \leq j \leq n$, $b \in \mathbb{R}^m$ de composantes $(b_1, b_2, ..., b_m)$ et x le vecteur inconnu de composantes $(x_1, x_2, ..., x_n)$.

Remarque 171. Si $b = 0$, on dit que le système (1.1) est homogène. Dans le cas où $m < n$, on dit que le système est sous determiné et dans le cas où $n > m$ on parle d'un système sur determiné.

Théorème 172. Une condition suffisante pour que le système (1.1) admette au moins une solution est que $rang\ A = m$.

Corollaire 173. Soit $A \in M_n(\mathbb{R})$ une matrice carrée, une condition nécessaire et suffisante pour que le système $Ax = b$ admette une solution unique est que A soit inversible. Autrement dit $\det A \neq 0$ (ou $rangA = n$). Le système (1.1) est alors dit système de Cramer.

1.3 Résolution d'un système triangulaire supérieur

Définition 174. Soit $A \in M_n(\mathbb{R})$ une matrice carrée, A est dite triangulaire supérieure si :

$$a_{ij} = 0, \quad \forall i > j$$

Dans ce cas le système $Ax = b$ est dit triangulaire supérieur.

Remarque 175. On ne traitera pas le cas des systèmes triangulaires inférieurs car la technique de résolution est identique.

Le système d'équations $Ax = b$ triangulaire supérieur a la forme suivante :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n & = & b_1 \\ a_{22}x_2 + ... + a_{2n}x_n & = & b_2 \\ ..... & \dots & \dots \\ a_{nn}x_n & = & b_n \end{cases}$$

Théorème 176. Soit le système triangulaire supérieur $Ax = b$ où $A \in M_n(\mathbb{R})$ est une matrice carrée et $b \in \mathbb{R}^n$, si :

$$a_{kk} \neq 0, \quad \forall k \in [1, n]$$

alors le système admet une solution unique et cette solution x^* est telle que :

$$x_k^* = \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j^*}{a_{kk}}, \quad k = \{n, n-1, ..., 1\} \tag{1.2}$$

Nous allons étudier les méthodes de résolution du système de n équations linéaires à n inconnues $Ax = b$, par les mé thodes directes. Nous entendons par méthodes directes des méthodes qui mènent à la solution en un nombre fini d'opérations élémentaires. Ces méthodes sont utilisées seulement si le nombre d'équations du système n'est pas trop élevé (généralement $n \leq 100$). La méthode de Cramer en est une, mais elle est numériquement inacceptable. Car sa mise en oeuvre demande le calcul de $n+1$ dé terminants et n divisions. Pour calculer chaque déterminant, nous devons effectuer $n!n$ multiplications et $n! - 1$ additions soit un total de $(n+1)^2 n! - 1$ opé rations élémentaires Par exemple, pour $n = 5$ on obtient 4319 opérations élé mentaires. Pour $n = 10$ on obtient à peu près 4.10^8 opérations él émentaires Or, dans la pratique, nous aurons à résoudre des systèmes d'ordres $n = 100$, $n = 1000$ voire même plus. Il est donc impossible de r ésoudre de tels systèmes par la méthode de Cramer. Dans ce chapitre, nous présentons essentiellement la méthode d'é liminations successives de Gauss et son interprétation matricielle, laquelle débouche sur la méthode de Cholesky pour un système à matrice définie positive. Si la matrice A n'est plus

triangulaire, nous sommes amenés à chercher une matrice M inversible telle que la matrice produit MA soit triangulaire. On résoudra alors le système :

$$MAx = Mb$$

par l'algorithme (1.2). Nous nous limitons bien entendu à des systèmes $Ax = b$ avec $detA \neq 0$.

2 Méthode de Gauss

Soit $A \in M_n(\mathbb{R})$ une matrice carrée donnée et $b \in \mathbb{R}^n$. On cherche x^* solution du système linéaire :

$$Ax = b$$

La méthode de Gauss consiste à construire un système é quivalent plus facile à résoudre (à matrice triangulaire supé rieure par exemple). Deux systèmes linéaires définis par deux matrices $A \in M_n(\mathbb{R})$ et $U \in M_n(\mathbb{R})$ sont dits équivalents si leurs solutions sont identiques.

Remarque 177. - Les transformations élémentaires suivantes appliquées à un système linéaire engendre un système linéaire équivalent : - Une équation peut être remplacée par cette même é quation à laquelle on ajoute ou on retranche un certain nombre de fois une autre ligne. - La multiplication d'une équation par une constante non nulle. - La permutation de deux lignes ou de deux colonnes.

La représentation d'un système linéaire peut se faire à travers une matrice de dimension $n.(n+1)$ appelé matrice augmentée. La matrice est noté $\tilde{A} = [A|b]$ et a pour forme gé nérale :

$$\tilde{A} = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \cdots & \cdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right)$$

La résolution du système linéaire ayant pour matrice augmenté e $\tilde{A}$ peut se faire en appliquant des transformations élé mentaires permettant d'obtenir un système équivalent. L'objectif de l'algorithme de Gauss est la construction d'un système triangulaire supérieur équivalent, en annulant au fur et à mesure les termes en dessous de la diagonale.

Définition 178. On appelle pivot de la transformation, l'élément a_{kk} de la matrice utilisée pour annuler les termes a_{jk}, $j > k$. La ligne k est alors appelée ligne pivot.

Théorème 179. Soit un système linéaire défini par une matrice A d'ordre n et $b \in \mathbb{R}^n$. Si A est non singulière alors il existe une matrice U d'ordre n triangulaire supérieure et $y \in \mathbb{R}^n$ tels que $Ux = y$ soit équivalent à $Ax = b$. La résolution du système $Ax = b$ se fait ensuite par résolution du système triangulaire supérieur.

Démonstration. Construisons la matrice augmentée $\tilde{A}^{(1)} = \left[A^{(1)}|b^{(1)}\right]$

$$\tilde{A}^{(1)} = \left(\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \cdots & \cdots & \ddots & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right)$$

l'exposant indiquant le nombre de fois qu'une valeur a été stocké e à la location i, j donnée. La première étape de l'algorithme de Gauss est d'annuler l'ensemble des coefficients de la premi ère colonne en dessous de la diagonale. Cela s'obtient si $a_{11} \neq 0$ en réalisant la transformation suivante sur la ligne $i > 1$:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - g_{i1} a_{1j}^{(1)}, \quad j \in [1, n+1]$$

où $g_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$, on obtient le système équivalent à l'étape 2 suivant donné par sa matrice augment ée :

$$\tilde{A}^{(2)} = \left(\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \cdots & \cdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right)$$

les étapes suivantes consistent à refaire le même procédé pour les colonnes suivantes. Ainsi l'étape k consiste à éliminer l'inconnu x_k dans les équations $k+1,...,n$. Ce qui donne les formules suivantes définies pour les lignes $i = k+1,...,n$ en supposant que le $k^{ième}$ pivot $a_{kk}^{(k)} \neq 0$:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - g_{ik} a_{kj}^{(k)}, \quad j \in [k, n+1] \tag{2.1}$$

avec $g_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$. A la dernière étape c'est-à-dire à $k = n$, on obtient le syst ème équivalent suivant :

$$\tilde{A}^{(n)} = \left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \cdots & \cdots & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{n-1n-1}^{(n-1)} & a_{n-1n}^{(n-1)} & b_{n-1}^{(n-1)} \\ 0 & \cdots & \cdots & 0 & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right)$$

la matrice U est donc définie comme étant la matrice $\tilde{A}^{(n)}$ et y le vecteur $b^{(n)}$. cqfd

Remarque 180. - La ligne i de la matrice $\tilde{A}^{(k)}$ n'est plus modifiée par l'algorithme dès lors que $i \leq k$. - A l'étape k, on pratique l'élimination sur une matrice de taille $n-k+1$ lignes et $n-k+2$ colonnes.

Remarque 181. Si lors de l'élimination l'élément $a_{kk}^{(k)}$ à l'é tape k est nul alors la ligne k ne peut pas être utilisée comme ligne pivot. Dans ce cas, on cherche une ligne $j > k$ telle $a_{jk}^{(k)} \neq 0$. Si une telle ligne existe, alors on permute la ligne j et la ligne k sinon le système n'admet pas de solution.

Remarque 182. Pour minimiser les erreurs d'arrondi, on choisit la valeur du pivot la plus grande en valeur absolue. Pour ce faire deux stratégie sont possibles :

1. La méthode dite à *pivot partiel* : Au $k^{ième}$ pas de l'élimination, on choisit comme ligne de pivot celle qui, parmi les $n-k+1$ restantes, a l'élément de module maximum en colonne et on permute dans $\tilde{A}^{(k)}$ la $k^{ième}$ ligne naturelle et celle qui réalise ce maximum.
2. La méthode dite à *pivot total* : Au $k^{ième}$ pas de l'élimination, on choisit comme pivot l'élément de plus grand module dans la matrice d'ordre $n-k+1$ restante. On permute donc dans $\tilde{A}^{(k)}$ la $k^{ième}$ colonne naturelle et celle du pivot, ce qui modifiera l'ordre des composantes du résultat. A la fin du processus, il ne faudra pas oublier de remettre dans l'ordre initial les composantes de la solution x.

2.1 Interprétation matricielle de la méthode de Gauss

Supposons que l'on puisse effectuer l'élimination sans permutation des lignes et des colonnes. Considérons alors les matrices

$$G^{(k)} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ \vdots & \vdots & 1 & & & \vdots \\ 0 & \cdots & -g_{k+1k} & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -g_{nk} & 0 & \cdots & 1 \end{pmatrix}, \quad k = 1, 2, ..., n-1$$

avec

$$g_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad \text{pour} \quad i = k+1, ..., n.$$

le système (2.1) peut se mettre sous la forme

$$\tilde{A}^{(k+1)} = G^{(k)}\tilde{A}^{(k)}$$

ce qui donne

$$\tilde{A}^{(n)} = G^{(n-1)}.G^{(n-2)}.G^{(n-3)}.....G^{(1)}.\tilde{A}^{(1)}$$

avec

$$\tilde{A}^{(n)} = \left[A^{(n)}|b^{(n)}\right]$$

Posons

$$\begin{aligned} U &= A^{(n)} \\ L &= \left(G^{(n-1)}.G^{(n-2)}.G^{(n-3)}.....G^{(1)}\right)^{-1} \end{aligned}$$

U (pour Upper) est une matrice triangulaire supérieure et L (pour Lower) est une matrice triangulaire inférieure à diagonale unité . Donc nous avons écrit A sous la forme : $A = LU$ où

$$L = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ g_{21} & 1 & \ddots & \cdots & \cdots & \vdots \\ g_{31} & g_{32} & 1 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ g_{n1} & \cdots & \cdots & \cdots & \cdots & 1 \end{pmatrix}, \text{ et } U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & a_{n-1n-1}^{(n-1)} & a_{n-1n}^{(n-1)} \\ 0 & \cdots & \cdots & \cdots & 0 & a_{nn}^{(n)} \end{pmatrix}$$

nous sommes donc amenés à résoudre successivement les deux syst èmes triangulaires :

$$\begin{cases} Ly = b \\ Ux = y \end{cases} \quad \text{où } y = b^{(n)}$$

3 Méthodes LU

La première phase de la méthode de Gauss consistait à transformer le système $Ax = b$ en un système triangulaire $Ux = y$ avec U une matrice triangulaire supérieure. Supposons qu'aucune permutation n'ait été effectuée, on peut alors montrer que U et y ont été obtenus à partir de A et b en les multipliant par une même matrice R triangulaire et inversible, c'est-à-dire

$$U = RA \quad \text{et} \quad y = Rb$$

on a donc $A = R^{-1}U$. Et si on pose $L = R^{-1}$et $U = R$, on peut donc dé composer A en un produit de matrice triangulaire inférieure L et une matrice triangulaire supérieure U. La méthode de Gauss appartient donc à la classe des méthodes dites méthodes LU. Elles consistent à obtenir une décomposition de la matrice A du type LU et à résoudre le système triangulaire $Ly = b$ puis ensuite le système triangulaire $Ux = y$ (L et U étant supposé es inversibles).

$$Ax = b \Longleftrightarrow LUx = b \Leftrightarrow \begin{cases} Ly = b \\ Ux = y \end{cases}$$

3.1 Décomposition LU

Définition 183. Une matrice A non singulière, admet une factorisation triangulaire si il existe une matrice L triangulaire inférieure et une matrice U triangulaire supérieure telles que :

$$A = LU$$

Théorème 184. Soit le système linéaire $Ax = b$, si au cours de l'élimination de Gauss de la matrice A, aucun pivot n'est nul alors il existe une matrice L triangulaire inférieure et une matrice U triangulaire supé rieure telles que :

$$A = LU$$

si de plus on impose $l_{kk} = 1$ alors la factorisation est unique.

La matrice U s'obtient en appliquant la méthode de Gauss tandis que la matrice L s'écrit de la manière suivante :

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ l_{21} & 1 & \cdots & 0 & 0 \\ l_{31} & l_{32} & \ddots & 0 & 0 \\ \vdots & \cdots & \ddots & 1 & 0 \\ l_{n1} & l_{n2} & \cdots & l_{nn-1} & 1 \end{pmatrix}$$

où pour $i > 1$ on a $l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$. Ainsi la matrice L est composée des facteurs multiplicatifs permettant d'annuler les éléments sous le pivot. Comme il existe des problèmes simples pour lesquelles un des pivots est nul, le théorème suivant permet d'étendre la factorisation LU à un cadre plus général.

Théorème 185. Une condition nécessaire et suffisante pour qu'une matrice A inversible puisse se factoriser sous la forme $A = LU$ est que $\det(A_k) \neq 0$, $\forall k = 1, 2, ...n-1$. Où $A_k = (a_{ij})_{\substack{i=1,2,..,k \\ j=1,2,.....,k}}$.

Démonstration. 1) Si $A = LU$, $A_k = L_k U_k$ et si A est inversible, $\det U = \prod_{i=1}^{n} u_{ii} = \prod_{i=1}^{n} a_{ii}^{(i)} \neq 0$. Donc $\det(A_k) = \det(L_k).\det(U_k) = \det(U_k) = \prod_{i=1}^{k} a_{ii}^{(i)} \neq 0$. 2) Supposons que $\det(A_k) \neq 0 \ \ \forall k = 1, 2, ..., n-1$. Celà est vrai en particulier pour $k = 1$, donc $a_{11}^{(1)} = \det(A_1)$ et la première étape de l'élimination de Gauss est possible. Par récurrence, si on a obtenu $A^{(k)}$ pour $k \leq n-1$

$$A^{(k)} = G^{(k-1)}.G^{(k-2)}.G^{(k-3)}.....G^{(1)}.A^{(1)}$$

alors $\det(A_k^{(k)}) = \det(G_k^{(k-1)})... \det(G_k^{(1)}) \det(A^{(1)}) = \det(A^{(1)}) \neq 0$ $A_k^{(k)}$ étant triangulaire on a $\prod_{i=1}^{k} a_{ii}^{(i)} \neq 0$ donc $a_{kk}^{(k)} \neq 0$, donc la $k^{ième}$ étape de l'élimination est possible. On obtiendra finalement A=LU.

cqfd

Théorème 186 (méthode à pivot partiel). Soit A une matrice carrée d'ordre n inversible, alors il existe une matrice de permutation P telle que les pivots de PA soient non nuls. Ainsi il existe deux matrices L et U telles que $PA = LU$.

Remarque 187. le système linéaire $Ax = b$ est équivalent au système $PAx = Pb$ et la résolution du système se fait selon les étapes suivantes :

1. Construire U, L et P,
1. Calculer Pb,
2. Résoudre $Ly = Pb$ (système triangulaire inférieur),
3. Résoudre $Ux = y$ (système triangulaire supérieur).

4 Méthode de Cholesky

Certains systèmes présentent des propriétés particuliè res. Les matrices associées à ces systèmes peuvent être sym ètriques, à bande, etc...La méthode de cholesky a pour but la r ésolution de systèmes linéaires pour lesquels la matrice associ ée est symétrique définie positive.

Définition 188 (Matrice symétrique). Soit A une matrice carrée d'ordre n. On dit que A est symé trique si on a

$$A = A^t.$$

Définition 189 (Matrice définie positive). Soit A une matrice carrée d'ordre n. On dit que A est définie positive si elle vérifie la condition suivante :

$$\forall x \in \mathbb{R}^n \text{ et } x \neq 0, \quad \langle Ax, x\rangle > 0$$

où $\langle .,. \rangle$ désigne le produit scalaire dans $\mathbb{R}^n$. C'est-à-dire :

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \quad \forall x, y \in \mathbb{R}^n$$

On définit la norme induite par :

$$\|x\| = \langle x, x\rangle^{\frac{1}{2}}$$

Proposition 190. Si A est une matrice symétrique définie positive alors :

1. $a_{ii} > 0$,
2. $a_{ij} < a_{ii}a_{jj} \quad \forall i \neq j$,
3. $\max_{j,k} \left|a_{jk}\right| < \max_i |a_{ii}|$.

Théorème 191. Si la matrice A est une matrice carrée définie positive alors elle est inversible.

Corollaire 192. Si la matrice A est une matrice carrée définie positive alors le système linéaire $Ax = b$ où $x, b \in \mathbb{R}^n$ admet une solution et une seule.

Théorème 193. Soit M une matrice carrée telle et non singulière alors la matrice $A = MM^t$ est symétrique définie positive.

Exemple 194. Soit

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

alors

$$A = MM^t = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

est définie positive.

4.1 Factorisation de Cholesky

Théorème 195. Soit A une matrice carrée d'ordre n symétrique définie positive, A peut alors se décomposer et de manière unique en

$$A = LL^t$$

où L est une matrice triangulaire inférieure avec des élé ments diagonaux positifs.

Ainsi ce théorème permet de déduire que la méthode de construction des matrices définies positives engendre en fait l'ensemble des matrices symétriques définies positives. Si A est une matrice symétrique définie positive alors le systè me $Ax = b$ peut être décomposé en $LL^t x = b$ et ce système peut se résoudre en résolvant les systèmes triangulaires :

$$\begin{cases} Ly &= b \\ L^t x &= y \end{cases}$$

4.2 Algorithme de décomposition de Cholesky

Soit A une matrice symétrique définie positive alors on a $A = LL^t$. Pour résoudre le système

$$Ax = b \tag{4.1}$$

le théorème précédent nous permet d'écrire (4.1) sous la forme $LL^t x = b$ avec L une matrice triangulaire inférieure inversible. On est donc amené à résoudre

$$\begin{cases} Ly &= b \\ L^t x &= y \end{cases}$$

Le problème consiste donc à construire explicitement la matrice $L = (l_{ij})$ triangulaire inférieure telle que

$$A = LL^t \quad \text{où} \quad A = (a_{ij})$$

ce qui équivaut à

$$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk}, \quad j \leq i.$$

Soit :

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ 0 & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{nn} \end{pmatrix}$$

en remarquant que a_{ij} est le produit de la ligne i de L et la colonne j de L^t alors on a :

$$a_{i1} = \sum_{k=1}^{n} l_{ik} l_{1k} = l_{i1} l_{11} + l_{i2} l_{12} + \cdots + l_{in} l_{1n} = l_{i1} l_{11}$$

en particulier pour $i = 1$, on a $l_{11} = \sqrt{a_{11}}$ (l_{11} est bien positif). la connaissance de l_{11} permet de construire la première colonne de la matrice L car :

$$l_{i1} = \frac{a_{i1}}{l_{11}}$$

En raisonnant de la même manière pour la deuxième colonne de L , on a :

$$a_{i2} = \sum_{k=1}^{n} l_{ik} l_{2k} = l_{i1} l_{21} + l_{i2} l_{22}$$

en prenant $i = 2$ alors $a_{22} = l_{21}^2 + l_{22}^2$. D'où l'on tire

$$l_{22} = \sqrt{a_{22} - l_{21}^2}$$

ensuite on a :

$$l_{i2} = \frac{a_{i2} - l_{i1} l_{21}}{l_{22}} \qquad i = 3, 4, ..., n$$

On peut généraliser la procédure au calcul de la colonne j en supposant que les $(j-1)$ colonnes ont déjà été calculé e. Ainsi :

$$a_{ij} = \sum_{k=1}^{n} l_{ij} l_{1j} = l_{ij} l_{j1} + l_{ij} l_{j2} + \cdots + l_{ik} l_{jk} + \cdots + l_{in} l_{jj}$$

et seul l_{ij} et l_{jj} ne sont pas connus. Si on pose $i = j$, on obtient :

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$$

et par conséquent

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}} \quad i > j \quad \begin{array}{rcl} j & = & 2, ..., n \\ i & = & j+1, ..., n \end{array}$$

Remarque 196. La décomposition de A symétrique définie positive, sous la forme $A = LL^t$ est unique à une matrice diagonale unité près. C'est-à-dire si $A = LL^t = MM^t$, alors $M = DL$ avec D matrice diagonale telle que $d_{ii} = \pm 1$.

Remarque 197. La méthode de Cholesky permet de calculer $\det A$ par

$$\det A = \prod_{i=1}^{n} \prod l_{ii}^2$$

5 SERIE D'EXERCICES

Exercice 198. Résoudre le système d'équations linéaires suivant :

$$\left\{ \begin{array}{rcl} -3x_1 - x_2 & = & 5 \\ -2x_1 + x_2 + x_3 & = & 0 \\ 2x_1 - x_2 + 4x_3 & = & 15 \end{array} \right.$$

1. En appliquant les formules de Cramer.
2. En triangularisant la matrice du système associée par la méthode de Gauss.

Exercice 199. Résoudre le système d'équations linéaires suivant :

$$\left\{ \begin{array}{rcl} 2x_1 + x_2 - 5x_3 + x_4 & = & 8 \\ x_1 - 3x_2 - 6x_4 & = & 9 \\ 2x_2 - x_3 + 2x_4 & = & -5 \\ x_1 + 4x_2 - 7x_3 + 6x_4 & = & 0 \end{array} \right.$$

En appliquant le principe de triangularisation de Gauss.

Exercice 200. Soit le système d'équations linéaires suivant :

$$\left\{ \begin{array}{rcl} x_1 + 2x_3 & = & 2 \\ 5x_2 + 4x_3 & = & 0 \\ 2x_1 + 4x_2 + 14x_3 & = & 5 \end{array} \right.$$

1. Montrer que la matrice associée à ce système est définie positive.
2. Résoudre ce système en utilisant la méthode de Choleski.

Exercice 201. Soit les systèmes d'équations linéaires suivants :

$$\left\{ \begin{array}{rcl} x_1 + x_2 + 2x_3 & = & 1 \\ 5x_1 + 5x_2 & = & 3 \\ 3x_1 + x_2 + x_3 & = & -2 \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{rcl} x_1 + 4x_2 + x_3 + 3x_4 & = & 2 \\ -x_2 + 3x_3 - x_4 & = & 0 \\ 3x_1 + x_2 + 2x_4 & = & 1 \\ x_1 - 2x_2 + 5x_3 + x_4 & = & -2 \end{array} \right.$$

Effectuer la résolution en mettant, si cela est possible, la matrice associée à chacun de ces deux systèmes, sous forme d'un produit de deux matrices triangulaires de structures différentes.

6 METHODES INDIRECTES

6.0.1 Introduction

Les méthodes directes de résolution de systèmes linéaires fournissent une solution x au problème $Ax = b$ en un nombre fini d'op érations. Si l'ordre n de la matrice A est élevé, le nombre d'opérations est aussi élevé et de plus, le résultat obtenu n'est pas rigoureusement exact. Par ailleurs, il existe des cas où les structures du système liné aire ne sont pas tirés à profit par les méthodes directes. C'est par exemple le cas des systèmes où la matrice A est très creuse. C'est la raison pour laquelle, dans ce cas , on préfère utiliser des méthodes itératives. L'objectif est de construire une suite de vecteurs $\left\{x^{(k)}\right\}_{k=1,2...,n}$ qui tend vers un vecteur $\bar{x}$, solution exacte du problème $Ax = b$. Souvent, on part d'une approximation $\left\{x^{(0)}\right\}$ de $\bar{x}$ obtenue en géné ral par une méthode directe.

6.1 Les méthodes itératives

L'objectif est de résoudre un système du type $Ax = b$. Pour cela, nous allons décomposer la matrice A en

$$A = M - N$$

de sorte que M soit inversible. Ainsi, le système devient :

$$Mx = Nx + b$$

et nous chercherons par récurrence une suite de vecteurs $x^{(i)}$ obtenu à partir d'un vecteur $x^{(0)}$ et de la relation

$$Mx^{(k+1)} = Nx^{(k)} + b$$

c'est-à-dire

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$$

Cette relation est une relation de récurrence du premier ordre. Nous pouvons en déduire une relation reliant l'erreur $e^{(k)} = x^{(k)} - \bar{x}$ à $e^{(k-1)} = x^{(k-1)} - \bar{x}$:

$$M(x^{(k)} - \bar{x}) = N(x^{(k-1)} - \bar{x})$$

puisque $M\bar{x} = N\bar{x} + b$ et donc $e^{(k)} = M^{-1}Ne^{(k-1)}$ pour $k = 1, 2, ...$ Si on pose $B = M^{-1}N$, nous avons alors

$$e^{(k)} = Be^{(0)}$$

La convergence de la suite $x^{(k)}$ vers la solution $\bar{x}$ est donné par le proposition suivant :

Proposition 202. Le choix de la décomposition de A devra obeir aux règles suivantes :

Remarque 203.

Proposition 204.

1. Le rayon spectral $\rho(M^{-1}N)$ doit être strictement infé rieur à 1.
2. La résiolution de $Mx^{(k)} = Nx^{(k-1)} + b$ doit être simple et n écessiter le moins d'opérations possibles
3. Pour obtenir la meilleure convergence, $\rho(M^{-1}N)$ doit être le plus petit possible.

On voit que la convergence dépend de la décomposition.

6.2 Différentes décomposition de A

On écrit la matrice A sous la forme

$$A = D + E + F$$

avec D la matrice diagonale suivante :

$$D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

E la matrice triangulaire inférieure suivante

$$E = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn-1} & 0 \end{pmatrix}$$

et F la matrice triangulaire supérieure

$$F = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & 0 & \ddots & \vdots \\ 0 & & \ddots & a_{n-1n} \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

Nous obtiendrons donc la décomposition $A = M - N$ à partir de diffé rents types de regroupement de ces matrices D, E et F.

6.3 Méthode de Jacobi

On pose

$$M = D \quad \text{et} \quad N = -(E + F)$$

ainsi, $B = M^{-1}N = D^{-1}(-E - F)$, ce qui implique :

$$x^{(k+1)} = D^{-1}(-E - F)x^{(k)} + D^{-1}b$$

si on exprime cette relation en fonction des éléments de la matrice A nous avons :

$$x_i^{(k+1)} = -\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \qquad i = 1, 2, ..., n$$

6.4 Méthode de Gauss-Seidel

Cette méthode utilise

$$M = D + E \quad \text{et} \quad N = -F$$

D'où

$$B = -(D + E)^{-1} F,$$

et alors on a :

$$x^{(k+1)} = -(D + E)^{-1} F x^{(k)} + (D + E)^{-1} b$$

le calcul de l'inverse de $(D+E)$ peut être évité. Si on écrit $(D+E)\,x^{(k+1)} = -Fx^k + b$, on obtient

$$\sum_{j=1}^{n} a_{ij} x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij} x_j^{(k)} b_i,$$

d'où

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \frac{1}{a_{ii}} \sum_{j=i+1}^{i-1} a_{ij} x_j^{(k)} + \frac{b_i}{a_{ii}}, \qquad i = 1, 2, ..., n.$$

6.5 Méthode de relaxation

On donne un paramètre $\omega \in]0, 2[$, appelé facteur de relaxation, et on pose

$$M = \frac{D}{\omega} + E \quad \text{et} \quad N = \left(\frac{1-\omega}{\omega}\right) D - F$$

et par conséquent

$$\left(\frac{D}{\omega} + E\right) x^{(k+1)} = \left(\left(\frac{1-\omega}{\omega}\right) D - F\right) x^{(k)} + b$$

d'où

$$x_i^{(k+1)} = (1-\omega)\, x_i^{(k)} + \frac{\omega}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \right) \qquad i = 1, 2, ..., n.$$

Comme on peut le constater, la méthode de Gauss-Seidel correspond à la méthode de relaxation pour $\omega = 1$.

7 Convergence des méthodes itératives

La convergence des méthodes itératives dépend fortement du rayon spectral de A, Nous étudions d'abord les propriétés de certaines matrices et la localisation de leurs valeurs propres.

Définition 205. Soit $A \in \mathcal{M}_{m,n}(\mathbb{R})$ une matrice. On définit la norme matricielle induite à partir de la norme vectorielle sur $\mathbb{R}^n$ par

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Proposition 206. Soit A et B deux matrices telles que leur multiplication soit compatible alors on a :

$$\|AB\| \leq \|A\| \|B\|$$

pour toute norme induite.

Théorème 207 (Gerschgorin-Hadamard). Les valeurs propres de la matrice A appartiennent à la reunion des n disques D_k pour $k = 1, 2, ..., n$ du plan complexe ($\lambda \in \cup_{k=1}^{n} D_k$ où D_k, appelé disque de Gerschgorin, est dé fini par :

$$|z - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq i}} \left| a_{kj} \right|$$

7.1 Cas général

On considère une méthode itérative définie comme :

$$\begin{cases} x^{(0)} & \text{donné} \\ x^{(k+1)} & = \quad Cx^{(k)} + D \end{cases}$$

Théorème 208. Soit A une matrice carré d'ordre n, pour que $\lim_{k\to\infty} A^k = 0$, il faut et il suffit que $\rho(A) < 1$.

Théorème 209. Si il existe une norme induite telle que $\|C\| < 1$ alors la méthode itérative décrite ci-dessus est convergente quelque soit $x^{(0)}$ et elle converge vers la solution de :

$$(I_d - C)x = D$$

Théorème 210. Une condition nécessaire et suffisante de convergence de la méthode ci-dessus est que :

$$\rho(C) < 1$$

Remarque 211. la condition de convergence donnée par le rayon spectral n'est pas dé pendante de la norme induite, cependant elle peut être utile car le calcul du rayon spectral peut être difficile.

7.1.1 Cas des matrices à diagonale dominante

Définition 212. Une matrice est dite à diagonale dominante si :

$$\forall i, 1 \le i \le n, \quad |a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

Théorème 213. Si A est une matrice à diagonale strictement dominante, alors A est inversible et en outre, les méthodes de Jacobi et de Gauss-Seidel convergent.

Démonstration. si A est une matrice à diagonale strictement dominante, on montre que A est inversible en démontrant que 0 n'est pas une valeur propre (c'est-à-dire $KerA = 0$). Posons $B = M^{-1}N$ est soit λ et v tels que $Bv = \lambda v$ avec $v \neq 0$. Puisque l'on s'interesse à $\rho(B) < 1$, on s'interesse en fait à la plus grande valeur propre de plus grand module de B. Ainsi, on peut supposer que $\lambda \neq 0$. L'équation $Bv = \lambda v$ devient :

$$\left(M - \frac{1}{\lambda}N\right)v = 0$$

- Pour Jacobi ; l'équation devient :

$$\left(D + \frac{1}{\lambda}E + \frac{1}{\lambda}F\right)v = 0$$

soit $C = D + \frac{1}{\lambda}E + \frac{1}{\lambda}F$. si $|\lambda| \ge 1$, on aurait :

$$|c_{ii}| = |a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \ge \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{\lambda}\right| = \sum_{\substack{j=1 \\ j\neq i}}^{n} |c_{ij}|$$

donc C serait à diagonale strictement dominante et par conséquent inversible. C inversible implique que $Cv = 0$ donc $v = 0$. Or $v \neq 0$, d'o ù la contradiction et donc on a bien $|\lambda| < 1$. - Pour Gauss-Seidel ; l'équation devient :

$$\left(D + E + \frac{1}{\lambda}F\right)v = 0$$

en posant encore $C = D + E + \frac{1}{\lambda}F$. et en supposant $|\lambda| \ge 1$, on aurait :

$$|c_{ii}| = |a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \ge \sum_{j<1}^{n} \left|\frac{a_{ij}}{\lambda}\right| + \sum_{j>1}^{n} \left|\frac{a_{ij}}{\lambda}\right| = \sum_{\substack{j=1 \\ j\neq i}}^{n} |c_{ij}|$$

et on obtient le même type de contradiction. cqfd

7.1.2 Cas des matrices symétriques définies positives

Théorème 214. Si A est une matrice symétrique définie positive, alors les mé thodes de Gauss-Seidel et de relaxation pour ($\omega \in]0,2[$) convergent.

La convergence de la méthode est d'autant plus rapide que $\rho(M^{-1}N)$ est petit. Or cette matrice $B = M^{-1}N$ dépend de ω. Une é tude théorique des valeurs propres de B montre que l'allure de la courbe $\rho(B)$ en fonction de B est décroissante entre 0 et ω_{opt} et croissante entre ω_{opt} et 2. Par ailleurs, on a toujours $1 < \omega_{opt} < 2$. On a donc intérêt à choisir ω le plus proche possible de ω_{opt}.

7.1.3 La méthode de correction

Soit le vecteur reste en x défini comme :

$$r(x) = b - Ax$$

et $\left\{r^{(k)}\right\}$ le reste en $\left\{x^{(k)}\right\}$. On appelle également l'erreur en k le vecteur

$$e^{(k)} = x^{(k)} - \bar{x}$$

où $\bar{x}$ est la solution. si on a une approximation $\left\{x^{(0)}\right\}$ de x, la relation suivante est vérifiée :

$$Ae^{(0)} = A(x^{(0)} - \bar{x}) = A(x^{(0)}) - b = -r^{(0)}$$

ce qui signifie que $e^{(0)}$ est la solution du système $Ax = -r^{(0)}$ et théoriquement, on a $\bar{x} = x^{(0)} - e^{(0)}$. Pratiquement, en appliquant au système $Ax = -r^{(0)}$ la méthode directe qui nous a fourni $x^{(0)}$, on n'obtient pas directement $e^{(0)}$, mais une approximation $y^{(0)}$ de $e^{(0)}$. Si on pose $x^{(1)} = x^{(0)} - y^{(0)}$, $x^{(1)}$ est une nouvelle approximation de $\bar{x}$, en itérant les calculs précédents, on obtient :

$$Ae^{(1)} = A(x^{(1)} - \bar{x}) = A(x^{(1)}) - b = -r^{(1)}$$

la résolution du système $Ax = -r^{(1)}$ donnera une approximation $y^{(1)}$ de $e^{(1)}$, et une nouvelle approximation $x^{(2)}$ de $\bar{x}$:

$$x^{(2)} = x^{(1)} - y^{(1)} = x^{(0)} - y^{(0)} - y^{(1)}$$

Ces calculs peuvent être itérer autant de fois que nécessaire, pour s'arrêter lorsque le reste est suffisament petit. A la $k^{ième}$ itération, les relations suivantes sont vérifiées pour $y^{(k-1)}$ approximation de $e^{(k-1)}$:

$$x^{(k)} = x^{(k-1)} - y^{(k-1)} = x^{(0)} - \sum_{i=0}^{k-1} y^{(i)}$$

avec $y^{(i)}$ une approximation de $e^{(i)}$, solution de $Ax = -r^{(i)}$ et $i = 0, 1, 2, ..., k-1$. Si nous nous arrêtons lorsque $k = N$, il est nécessaire de ré soudre $N+1$ systèmes linéaires : d'abord $Ax = b$, pour obtenir $x^{(0)}$ puis $Ax = -r^{(i)}$ et $i = 0, 1, 2, ..., N-1$ afin d'obtenir $y^{(i)}$. Une fois la matrice A décomposée (en LU ou Cholesky), il s'agit donc de résoudre les systèmes $LUx = -r^{(i)}$ où $-r^{(i)}$ a été calculé par la relation $r^{(i)} = b - Ax^{(i)}$.

8 SERIE D'EXERCICES

Exercice 215. Résoudre le système d'équations linéaires suivant :

$$\begin{cases} 10x_1 - 2x_2 - 2x_3 &= 6 \\ -x_1 + 10x_2 - 2x_3 &= 7 \\ -x_1 - x_2 + 10x_3 &= 8 \end{cases}$$

Par la méthode des approximations successives. Arrêter les calculs dès que :

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| < 10^{-2}$$

Exercice 216. Résoudre le système d'équations linéaires suivant :

$$\begin{cases} 10x_1 - 2x_2 - 2x_3 &= 6 \\ -x_1 + 10x_2 - 2x_3 &= 7 \\ -x_1 - x_2 + 10x_3 &= 8 \end{cases}$$

Par la méthode de Seidel. Arrêter les calculs dès que :

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| < 10^{-2}$$

Exercice 217. Résoudre le système d'équations linéaires suivant :

$$\begin{cases} 10x_1 - 2x_2 - 2x_3 &= 6 \\ -x_1 + 10x_2 - 2x_3 &= 7 \\ -x_1 - x_2 + 10x_3 &= 8 \end{cases}$$

Par la méthode de relaxation. Faire les calculs avec deux décimales.

Exercice 218. Résoudre le système d'équations linéaires suivant :

$$\begin{cases} 10x_1 + x_2 + x_3 &= 12 \\ 2x_1 + 10x_2 + x_3 &= 13 \\ 2x_1 + 2x_2 + 10x_3 &= 14 \end{cases}$$

Par la méthode de relaxation. Faire les calculs avec quatre décimales.

Chapitre

7

RÉSOLUTION NUMÉRIQUE des E.D.O. d'ORDRE UN

1 Introduction

Les équations différentielles ordinaires ou E.D.O. sont utilisées pour modéliser un grand nombre de phénomènes mécaniques, physiques, chimiques, biologiques etc...

Soit f une fonction continue

$$f: \quad [a,b]\times\mathbb{R}^{\ltimes} \to \mathbb{R}^{\ltimes} \tag{1.1}$$
$$(t,y) \mapsto f(t,y) \tag{1.2}$$

telle que :

$$f = (f_1, f_2, ..., f_n)$$

et

$$f_i: \quad [a,b]\times\mathbb{R}^{\ltimes} \to \mathbb{R}$$
$$(t,y) \mapsto f_i(t,y)$$

Définition 219. **1- E.D.O. du premier ordre :** On appelle équation différentielle ordinaire du premier d'ordre une équation de la forme :

$$y'(t) = f(t, y(t)) \qquad t \in [a,b] \tag{1.3}$$

2- E.D.O. d'ordre p : On appelle équation différentielle d'ordre p une équation de la forme :

$$y^{(p)}(t) = f(t, y(t), y'(t), y^{(2)}(t), ..., ..., y^{(p-1)}(t)) \qquad t \in [a,b] \tag{1.4}$$

f est une fonction continue donnée

$$f: \quad [a,b]\times(R^n)^p \to \mathbb{R}^{\ltimes}$$
$$(x,y) \mapsto f(x,y)$$

1. Une fonction y de classe C^1 vérifiant l'équation (1.3) est dite solution de l'équation différentielle du premier ordre.
2. Une fonction y de classe C^p vérifiant l'équation (1.4) est dite solution de l'équation différentielle d'ordre p.

Proposition 220. Toute équation différentielle d'ordre n sous forme canonique peut s'écrire comme un système de n équations différentielles du premier ordre.

Remarque 221. L'équation (1.3) est donc équivalente au système suivant :

$$\begin{cases} y'_1(t) &= f_1(t, y_1, ..., y_n) \\ ..... & \\ y'_n(t) &= f_n(t, y_1, ..., y_n) \end{cases} \tag{1.5}$$

Cela se fait en posant

$$\begin{cases} z_1 &= y \\ z_2 &= y' \\ ..... & ... \quad ... \\ z_p &= y^{p-1} \end{cases} \tag{1.6}$$

où $z_1, z_2, ..., z_p$ sont des fonctions de classe C^1 et l'équation différentielle d'ordre p (1.4) est est équivalente au système :

$$\left\{\begin{array}{lcl} z'_1 & = & y \\ z'_2 & = & y' \\ \ldots\ldots & \ldots & \ldots \\ z'_p & = & f(t, z_1, \ldots\ldots, z_p) \end{array}\right. \tag{1.7}$$

qui s'écrit aussi sous la forme

$$z'(t) = g(t, z(t))$$

Avec

$$\begin{aligned} g : [a,b] \times (R^n)^p &\to (R^n)^p \\ (t,y) &\mapsto f(t,y) \end{aligned}$$

Ce qui veut dire que l'étude d'une équation différentielle d'ordre p dans $\mathbb{R}^{\ltimes}$ est ramenée à une équation différentielle d'ordre 1 dans $\mathbb{R}^{\ltimes p}$.

Toute équation différentielle d'ordre p sous forme canonique peut s'écrire comme un système de p équations différentielles du premier ordre.

2 PROBLEME DE CAUCHY

Définition 222. On appelle problème de Cauchy, le problème qui consiste en la recherche d'un fonction y de classe C^1 vérifiant

$$\left\{\begin{array}{lll} y'(t) = & & f(t, y(t)) \\ y(a) = & y(0), \quad y_0 \quad \text{donné dans} & \mathbb{R}^{\ltimes} \end{array}\right. \tag{2.1}$$

Soit f une fonction continue

$$\begin{aligned} f : \quad [a,b] \times \mathbb{R}^{\ltimes} &\to \mathbb{R}^{\ltimes} \\ (t,y) &\mapsto f(t,y) \end{aligned}$$

Théorème 223. Soit le problème de Cauchy (2.1). Si f vérifie en plus de la continuité, la condition de Lipchitz, c'est-à-dire

$$\|f(t,y) - f(t,y^*)\| \leq k\|y - y^*\|; k > 0. \quad \forall t \in [a,b]; \forall y, y^* \in \mathbb{R}^{\ltimes} \tag{2.2}$$

alors le problème de Cauchy (2.1) admet une et une solution de classe C^1.

De très nombreux résultats mathématiques existent sur les problèmes de Cauchy.

Dans ce qui suit nous allons nous intéresser à certaines méthodes numériques de résolution de ce type de problème.

L'ensemble des méthodes numériques que nous allons étudier auront pour but la résolution d'un problème de Cauchy quelconque. Elles pourront donc être utilisées pour la résolution d'une très grande variété d'E.D.O.

Deux questions se posent, dans la résolution de ce type de problème (Cauchy).

1. Trouver une approximation numérique de la solution.
2. Majorer l'erreur commise à partir de cette approximation.

3 MÉTHODE de TAYLOR d'ORDRE 2

Pour résoudre numériquement le problème de Cauchy (2.1), On ecrit :

$$y(t) = y_0 + \frac{(t-a)}{1!} y'(a) + \frac{(t-a)^2}{2!} y''(a) + \ldots$$

Avec y_0 donnée

$$\begin{aligned} y\prime(a) &= f(a,y_0) = y\prime_0 \\ y\prime\prime(a) &= f(a,y_0) = y\prime_0 + \frac{\delta f}{\delta t}(a,y_0) + y\prime_0 \frac{\delta f}{\delta y}(a,y_0) \\ \cdots &\cdots \cdots \end{aligned}$$

Les formules de dérivation se compliquent très vite, et il est très souvent impossible d'avoir une idée sur le rayon de convergence de cette série (de Taylor).

Cette méthode est en général utilisée localement au voisinage du point $t_0 = a$.

4 MÉTHODES NUMERIQUES PAR PAS

Dans ce genre de méthodes, on va subdiviser l'intervalle $[a,b]$ par des points $t_1, t_2, \ldots\ldots, t_N$ équidistants : $t_{n+1} = t_n + h$. avec $h = \frac{b-a}{N}$ le pas de la subdivision.

On calcule N nombres $y_1, y_2, \ldots\ldots N$ ayant une valeur proche de celle de la solution y aux points $t_n = a_n + h$; $n = 0, \ldots\ldots, N$.

Ensuite, on fait une interpolation pour relier ces points et définir une fonction y_h sur $[a,b]$.

L'erreur de discretisation dépendante de h est estimée par la formule ;

$$e_n = y_n - y(t_n)$$

.

On distingue deux types d'algorithmes par :

1. Les algorithmes à pas séparés ou méthodes à un pas qui permettent de calculer y_{i+1} à partir de y_i.
2. Les algorithmes à pas liés ou méthodes à pas multiples qui permettent de calculer y_{i+1} à partir des $y_i, y_{i-1}, \ldots$ précédents.

5 MÉTHODE d'EULER-CAUCHY

Cette méthode étant la plus simple des méthodes numériques par pas.

En partant du développement de Taylor on a :

$$y(t_{n+1}) = y(t_n) + hy\prime(t_n) + R$$

D'où :

$$\frac{y(t_{n+1}) - y(t_n)}{h} = y\prime(t_n) + \frac{R}{h}$$

Si $\frac{R}{h}$ est suffisamment petit, on peut considérer que

$$\frac{y(t_{n+1}) - y(t_n)}{h}$$

est une bonne approximation de $y\prime(t_n)$ d'où **l'algorithme de Euler-Cauchy**.

$$\begin{cases} y_{n+1} = y_n + hf(t_n, y_n); \quad n = 0, 1....N \\ y(0) = y(a) \end{cases} \tag{5.1}$$

On peut interpréter cet algorithme comme :

connaissant y_n, on calcule y_{n+1} comme étant l'ordonnée du point d'intersection de la droite $t = t_{n+1}$ avec la droite passant par le point (t_n, y_n) ayant pour pente $f(t_n, y_n)$ c'est-à -dire la pente de la tangente en (t_n, y_n) à la courbe solution.

Théorème 224. Si la fonction f vérifie les hypothèses suivantes :

1- f est continue

2- f est lipchitzienne de rapport $K > 0$.

La methode de Euler-Cauchy (5.1) converge.

De plus on a une estimation de l'erreur sous la forme :

$$|e_n| = |y_n - y(t_n)| \leq \frac{e^{K(t_n - t_0)} - 1}{K} M(h, y\prime)$$

et

$$\max_{n=1,\ldots,N} |e_n| \to 0 \text{quand} h \to 0$$

5.1 Estimation de l'erreur dans la méthode d'Euler-Cauchy

On va chercher une majoration de l'erreur qui ne dépendra que des données.

Soit la proposition :

Proposition 225. Soit

$$c = \sup_{t \in [a,b]} |f(t, 0)|.$$

Alors

$$\|y_h\| \leq |y_0| e^{K(b-a)} + c \frac{e^{K(b-a)} - 1}{K} = D$$

et

$$\|y_h\| \leq D$$

Théorème 226. On pose

$$M_1 = \sup_{t \in [a,b]} |f(t, y)|$$

et

$$M_D(\delta, f) = \sup_{t, t' \in [a,b]} |f(t, y) - f(t', y)|$$

avec

$$\|y\| \leq D$$

et

$$t, t' \in [a, b]$$

Alors

$$|e_n| \leq (M_D(h, f) + hKM_1) \frac{e^{K(t_n - t_0)} - 1}{K}$$

La majoration de l'erreur donnée par ce théorème en fonction seulement des données est difficile a calculer numériquement.

On peut simplifier cette estimation en ajoutant une hypothèse supplémentaire.

Théorème 227. Soit Ω le domaine défini par :

$$\Omega = \{(t, y) \in \mathbb{R}^{\not\vDash} | t \in [a, b], |y| \leq D\}$$

avec

$$D = |y_0| e^{K(b-a)} + c \frac{e^{K(b-a)} - 1}{K}$$

On suppose :

1. f continue de $[a,b]\times\mathbb{R}\to\mathbb{R}$
2. f lipchitzienne en y
3. f de classe C^1 sur Ω

et on pose :

$$N(t)=\frac{1}{2}\max_{t\in[a,b]}|y''(t)|.$$

Alors

$$|e_n|\leq hN(t_n)\frac{e^{K(t_n-a)}-1}{K}$$

pour $n=0,1,.....,N$

6 MÉTHODE DE RUNGE-KUTTA

Les algorithmes de Runge-Kutta (RK) consistent à calculer à chaque pas des valeurs intermédiaires. La méthode (RK) classique est donnée par le schéma suivant :

$$\begin{cases} y_1=y_n+hF(t_n,y_n;h) \\ y_0=\eta \\ \text{avec} \quad F(t,y;h)=\frac{1}{6}(k_1+2k_2+2k_3+k_4) \\ \text{où} \quad k_1=f(t,y) \qquad k_2=f(t+\frac{h}{2},y+\frac{h}{2}k_1) \\ k_3=f(t+\frac{h}{2},y+h\frac{k_2}{2}); \qquad k_4=f(t+h,y+hk_3) \end{cases} \tag{6.1}$$

Remarque 228.
1. Les méthodes (RK) sont convergentes.
2. Elles ne nécessitent pas le calcul des dérivées successives de f.
3. Elles donnent de très bons résultats notamment pour la résolution des problèmes de Cauchy.

7 SERIE D'EXERCICES

Exercice 229. On considère l'équation différentielle

$$\begin{cases} y' & = & 2y \\ y(0) & = 5 \end{cases}$$

1. Vérifier que la solution exacte est $y(t)=5e^{2t}$.
2. Soit $h=\frac{1}{n}$, pour $i=0,...,n$, montrer que les approximations fournies par le schéma d'Euler peuvent s'écrire $y_i=5(1+2h)^i$.
3. Représenter graphiquement l'erreur

$$e(h)=\max_{0\leq i\leq n}|y(t_i)-y_i|.$$

en fonction de h, en calculant $e(0.005),e(0.01),e(0.05),e(0.1),e(0.5)$.

4. Que pensez-vous de la relation $e(h)\approx Kh$, avec K une constante ?

Exercice 230. On considere le probleme d'equation differentielle

$$\begin{cases} y' & = & -11y \\ y(0) & = 2 \end{cases}$$

Pour résoudre cette équation numériquement sur l'intervalle $[0,1]$, on se donne, pour chaque entier n, un pas $h = \frac{1}{n}$ et des noeuds $x_i = ih, i = 1,...,n$.

1. En répétant le raisonnement de l'exercice précédent, on pourrait montrer que, l'application de la méthode d'Euler conduit aux approximations :
$$y_i = 2(1-11h)^i.$$
2. Représenter les approximations obtenues pour pour $h = 0.2, 0.1, 0.09, 0.1$.
Indication : la solution de ce problème est de la forme $y = Ae^{at}$, où A et a sont faciles à calculer.

Exercice 231. On considere le problème d'équations différentielles

$$y' = 2t - 3y, \quad y(0) = 1.$$

1. Montrer que la solution exacte est donnée par :
$$y = -\frac{2}{9} + \frac{2}{3}t + \frac{11}{9}e^{-3t}.$$
2. Vérifier que les approximations obtenues en prenant $h = 0.25$ et la méthode de Taylor d'ordre 2 ou la méthode d'Euler modifiée, sont égales.
3. Ecrire la formule aux différentes $y_{i+1} = y_i + h\phi(t_i, y_i)$, obtenues par chacune des deux approches. Expliquer pourquoi, dans ce cas particulier, les deux formules coıncident.

Exercice 232. L'égalité suivante découle directement du theoreme fondamental du calcul

$$y(t+h) = y(t) + \int_t^{t+h} y'(t)dt.$$

En appliquant la formule de Simpson à l'integrale, on obtient alors l'approximation

$$y(t+h) \approx y(t) + \frac{h}{6}(y'(t) + 4y'(t+\frac{h}{2}) + y'(t+h)).$$

Supposons que y soit solution de $y' = f(t,y)$, l'approximation précédente s'écrit

$$y(t+h) \approx y(t) + \frac{h}{6}(f(t,y(t)) + 4f(t+\frac{h}{2}) + f(t+h, y(t+h))).$$

a) En remplaçant $y(t+\frac{h}{2})$ et $y(t+h)$ par les approximations données par la formule d'Euler modifiée, déduire de l'équation précédente un schéma numérique à un pas du type

$$y_{i+1} = y_i + \frac{h}{6}(f(t_i,y_i) + 2\frac{h}{3}f(t_i + \frac{h}{2}, y_i + k_1) + \frac{h}{6}f(t_i + h, y_i + k_2)))$$

pour lequel on déterminera les coefficients k_1, k_2 en fonction de h , y_i et $f(t_i, y_i)$.

b) On considère le cas particulier

$$f(t,y) = -y + t + 1$$

, pour lequel la solution exacte est

$$y(t) = t + e^{-t}.$$

En vous inspirant des exemples donnés et en choisissant $t_0 = 0, x_n = 1, y_0 = 1$, calculer

$$e(h) = \max|y_i - y(t_i)||i = 1,...npourn = 2,4,8,16$$

. Reporter cette fonction sur un graphe log – log et en déduire l'ordre de la méthode.

Exercice 233. On considère le problème

$$\begin{cases} y' & = & \frac{-3y}{t^2} \\ y(1) & = 2e^3 & \end{cases}$$

Comparer les approximations de la solution obtenues par la méthode d'Euler avec $h = 0.0016$ par la méthode du point milieu avec $h = 0.04$ et par la méthode de Runge-Kutta 4 avec $h = 0.2$. La comparaison se fera sur la base de la précision et du cout de calcul, i.e. le nombre de fois qu'il faut évaluer $f(t, y)$.

Chapitre

8

CALCUL DES VALEURS PROPRES ET VECTEURS PROPRES

Sommaire

1 Introduction

De nombreuses méthodes numériques supposent la connaissance des valeurs propres, des vecteurs propres et du rayon spectral d'une matrice. En outre de nombreux problèmes se ramènent à la recherche des valeurs propres d'une matrice. Le plus souvent, on fait appel à deux types de méthodes numé riques pour calculer les valeurs propres et les vecteurs propres (appelé s aussi éléments propres) d'une matrice. *les méthodes directes* sont celles qui permettent d'obtenir les éléments propres à partir de la connaissance explicite du polynôme caracté ristique ; les autres sont essentiellement des *méthodes itératives.* Ces dénominations présentent une certaine ambiguité. En effet, le plus souvent, la détermination du polynôme caracté ristique est obtenue par un procédé itératif et la recherche des racines de ce polynôme est presque toujours itérative. Soit $A \in \mathcal{M}_n(\mathbb{C})$. Nous allons chercher ses valeurs propres λ_i dont la multiplicité sera notée m_i.

2 RAPPELS

3 Calcul direct de $\det(A - \lambda I)$

On se donne $(n+1)$ valeurs distinctes $\lambda_1, \lambda_2, ..., \lambda_{n+1}$ quelconques ; on calcule pour chacune d'elles la valeur

$$y_i = \det(A - \lambda_i I) \qquad i = 1, 2, ..., n+1.$$

On obtient ainsi un ensemble de valeurs $\{(\lambda_i, y_i)\}_{i=1,2,...,n+1}$. On détermine alors, le polynô me d'interpolation passant par ces points. Il sera identique, à un facteur multiplicatif près, au polynôme caractéristique de A. On peut alors chercher ses racines par l'une des méthodes connues ; ce qui aboutira à une approximation des valeurs propres de A.

4 Méthode de Krylov

La méthode consiste à calculer les coefficients du polynôme caractéristique dont on approche les racines à l'aide des mé thodes connues. Les vecteurs propres associés sont alors détermin és par les formules appropriées. Plus précisément, soit :

$$P(\lambda) = (-1)^n(\lambda^n - \sum_{k=1}^{n} a_k \lambda^{n-k})$$

le polynôme caractéristique de A. D'après le théorème de Cayley-Hamilton (A annule son polynôme caractéristique), on a donc $P(A) = 0$; donc :

$$A^n = \sum_{k=1}^{n} a_k A^{n-k}$$

Prenons un vecteur quelconque x_0 non nul, on a :

$$A^n x_0 = \sum_{k=1}^{n} a_k A^{n-k} x_0 \tag{4.1}$$

Notons a le vecteur de composante (a_i) ; $i = 1, 2, ..., n$

$$\begin{aligned} x_1 &= Ax_0 \\ x_2 &= A^2 x_0 \\ &..... \\ x_{n-1} &= A^{n-1} x_0 \end{aligned}$$

et B la matrice dont les n colonnes sont les vecteurs

$$x_{n-1}, x_{n-2}, ..., x_1, x_0$$

L'équation (4.1) peut s'écrire :

$$A^n x_0 = Ba \tag{4.2}$$

Pour x_0 donné, on cherche a solution de (4.2). On obtiendra ainsi, si le système est inversible, les coefficients du polyn ôme caractéristique et on pourra utiliser une des méthodes de r ésolution des équations non linéaires pour calculer ses racines qui sont les valeurs propres de A. Cette méthode permet en outre de déterminer les vecteurs propres associés aux valeurs propres calculées. Pour simplifier, nous supposerons que les valeurs propres $\lambda_1, \lambda_2, ..., \lambda_n$ sont distinctes. Si nous appelons $v_1, v_2, ..., v_n$ les vecteurs propres associés respectivement à $\lambda_1, \lambda_2, ..., \lambda_n$, nous pouvons décomposer le vecteur x_0, choisi arbitrairement dans la recherche des coefficients (a_i), suivant la base des $\{v_i\}_{i=1,2,...,n}$. On a alors

$$x_0 = \alpha_1 v_1 + \alpha_2 v_2 + ... + \alpha_n v_n$$

comme

$$\begin{aligned} Av_i &= \lambda_i v_i \\ A^2 v_i &= \lambda_i^2 v_i \\ &\cdots\cdots \end{aligned}$$

on aura

$$\begin{aligned} x_1 &= \alpha_1 \lambda_1 v_1 + \alpha_2 \lambda_2 v_2 + ... + \alpha_n \lambda_n v_n \\ \cdots &\quad \cdots \\ x_{n-1} &= \alpha_1 \lambda_1^{n-1} v_1 + \alpha_2 \lambda_2^{n-1} v_2 + ... + \alpha_n \lambda_n^{n-1} v_n \end{aligned}$$

Considérons une combinason linéaire des vecteurs $x_0, x_1, ..., x_{n-2}, x_{n-1}$. On a

$$\begin{aligned} x_{n-1} &= \beta_{i1} x_{n-2} + \cdots + \beta_{in-1} x_0 = \\ &\quad \alpha_1 \varphi_i(\lambda_1) v_1 + \alpha_2 \varphi_i(\lambda_2) v_2 + \cdots + \alpha_n \varphi_i(\lambda_n) v_n \\ \text{où } \varphi_i(\lambda) &= \lambda^{n-1} + \beta_{i1} \lambda^{n-2} + \cdots + \beta_{in-1} \end{aligned} \tag{4.3}$$

on peut choisir par exemple la formule (4.3) s'écrit alors

$$x_{n-1} + \beta_{i1} x_{n-2} + ... + \beta_{in-1} x_0 = \alpha_i \varphi_i(\lambda_i) v_i.$$

Donc, si $\alpha_i \neq 0$ la combinaison linéaire obtenue permet de d éterminer le vecteur propre v_i. Avec le choix (**??**), les coefficients β_{ij} s'obtiennent facilement par identification. Plus précisement nous avons

$$\begin{aligned} \beta_{i0} &= 1 \\ \beta_{ij} &= \lambda_i \beta_{ij-1} - a_j \end{aligned}$$

Remarque 234. Si les vecteurs $x_0, x_1, ..., x_{n-2}, x_{n-1}$ sont linéairement ind épendant, la matrice B est inversible et on obtient bien les coefficients caractéristiques dont on peut calculer les valeurs propres. Mais il arrive que ces vecteurs ne soient pas linéairement indé pendants. Par exemple x_k s'exprime comme combinaison linéaire des pr écédents et de même des suivants. On peut alors appliquer la m éthode précédente avec les vecteurs $x_0, x_1, ..., x_{k-1}$; ce qui donnera un polynôme dont les racines seront racines du polynô me caractéristique et donc valeurs propres de A. On évite en général cette complication en changeant de vecteur initial.

5 MÉTHODE DE LEVERRIER

Les coefficients du polynôme caractéristique sont déterminés par la formule (5.2) suivante. On utilise ensuite les méthodes de résolution des équations non linéaires pour calculer les racines de ce polynôme ; ce qui détermine les valeurs propres. Posons

$$P(x) = a_1 x^n + a_2 x^{n-1} + ... + a_{n+1} \quad \text{avec} \quad a_1 \neq 0.$$

Les relations de Newton entre les racines $x_1, x_2, ..., x_n$ et les coefficients de ce polynôme sont donnés par

$$\left\{ \begin{array}{rcl} a_2 + a_1 S_1 & = & 0 \\ 2a_3 + a_2 S_1 + a_1 S_2 & = & 0 \\ \cdots\cdots\cdots\cdots & \cdots & \cdots \\ ka_{k+1} + a_k S_1 + \cdots + a_1 S_k & = & 0 \\ \cdots\cdots\cdots\cdots & \cdots & \cdots \\ na_{n+1} + a_n S_1 + \cdots + a_1 S_n & = & 0 \end{array} \right. \tag{5.1}$$

avec $S_k = \sum_{i=1}^{n} x_i^k$. Donc, en considérant le polynôme caractéristique de A, dont les racines sont les valeurs propres λ_i de A, on a

$$\begin{array}{rcl} S_k & = & T_r(A^k) \\ a_1 & = & (-1)^n \end{array}$$

ce qui nous permet de calculer les coefficients a_k pour $k = 2, ..., n+1$. Plus précisément on a :

$$a_k = -\frac{1}{k-1} a_{k-1} S_1 + ... + a_2 S_{k-2} + a_1 S_{k-1} \tag{5.2}$$

Remarque 235. La méthode de Leverrier présente un grave inconvénient : elle impose le calcul des puissances souvent élevées de la matrice initiale. Par contre son algorithme est simple et il n'y a pas lieu d'envisager des cas particuliers

6 Valeurs et Vecteurs Propres

Les premiers vecteurs et valeurs propres viennent des équations diffé rentielles (Lagrange 1759, *théorie du son* ; Lagrange 1781, des matrices 6×6 dans le but de calculer les perturbations sé culaires des orbites des 6 planètes connues à l'époque, *Oeuvres V*, p. 125-490). Aujourd'hui, le calcul des valeurs et vecteurs propres est indispensable dans toutes les branches de la science, en particulier pour la solution des systèmes des équations diffé rentielles linéaires, en théorie de stabilité, pour les questions de convergence de processus itératifs, et en physique et chimie (mécanique, circuits, cinétique chimique, équation de Schr ödinger). FIG. V.1 : Une application

linéaire comme champ de vecteurs (à gauche) ; transformée sur la base des vecteurs propres (à droite). Observons en figure V.1 (à gauche) le champ de vecteurs d'une é quation différentielle

$y' = Ay$. Deux directions sautent aux yeux : ce sont les directions où le vecteur Av prend la même direction que le vecteur v , c'est-à-dire., o ù

$$Av = \lambda v \quad \text{ou} \quad (A - \lambda I)v = 0 \tag{6.1}$$

Si cette équation est vérifiée, $\lambda \in \mathbb{C}$ s'appelle *valeur propre* de la matrice A et $v \in \mathbb{C}^n (v \neq 0)$ est le *vecteur propre* correspondant. L'équation (10.1) possède une solution v non nulle si et seulement si

$$P_A(\lambda) = \det((A - \lambda I) = 0$$

Le polynôme $P_A(\lambda)$ est le *polynôme caractéristique* de la matrice A . Les valeurs propres de A sont alors les z éros du polynôme caractéristique.

7 La condition du calcul des valeurs propres

A cause des erreurs d'arrondi, les éléments d'une matrice A , pour laquelle on cherche les valeurs propres, ne sont pas exacts. Ils sont plut ôt égaux à

$$\tilde{a}_{ij} = a_{ij}(1 + \varepsilon_{ij}) \quad \text{avec} \quad \left|\varepsilon_{ij}\right| \leq eps$$

(*eps* étant la précision de l'ordinateur, est supposée être très petite). Il est alors très important d'étudier l'influence de ces perturbations sur les valeurs propres et sur les vecteurs propres de la matrice. Pour montrer ceci, considérons la famille de matrices

$$A(\varepsilon) = A + \varepsilon C \quad \text{où} \quad |\varepsilon| \leq eps \quad \text{et} \quad \left|c_{ij}\right| \leq \left|a_{ij}\right|$$

(souvent, la dernière hypothèse va être remplacée par $\|C\| \leq \|A\|$).

Théorème 236 (Gershgorin). Soit A une matrice $n \times n$ (avec des éléments dans $\mathbb{R}$ ou dans $\mathbb{C}$).

a) Si λ est une valeur propre de A, alors il existe un indice i tel que

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right|$$

c'est-à-dire. que toutes les valeurs propres de A se trouvent dans l'union des disques $D_i = \left\{ \lambda;\ |\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right| \right\}$

b) Si une composante connexe de $\bigcup_{i=1}^{n} D_i$ consiste de k disques, elle contient exactement k valeurs propres de A.

Démonstration. Soit $v \neq 0$ un vecteur propre et choisissons l'indice i tel que $|v_i| \geq |v_j|$ pour tout j . La ligne i de l'équation $Av = \lambda v$ donne

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} v_j = (\lambda - a_{ii}) v_i.$$

En divisant par v_i et en utilisant l'inégalité du triangle, on obtient U

$$|\lambda - a_{ii}| = \left| \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} \frac{v_j}{v_i} \right| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

L'affirmation (b) est vraie si A est une matrice diagonale. Le cas gén éral est obtenu par un argument de continuité en faisant tendre les éléments en dehors de la diagonale vers zéro. cqfd

Théorème 237. Soit A une matrice diagonalisable, c'est-à-dire, il existe P avec $P^{-1}AP = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ et soit $A(\varepsilon) = A + \varepsilon C$. Alors, pour chaque valeur propre $\lambda(\varepsilon)$ de $A(\varepsilon)$, il existe un λ_i avec

$$|\lambda(\varepsilon) - \lambda_i| \leq \varepsilon . \kappa_\infty(P) . \|C\|_\infty$$

Démonstration. Nous transformons la matrice$A(\varepsilon) = A + \varepsilon C$ par la mê me matrice, qui transforme A sous forme diagonale :

$$P^{-1}A(\varepsilon)P = diag(\lambda_1, \lambda_2, ..., \lambda_n) + \varepsilon P^{-1}CP$$

Si l'on dénote par e_{ij} les éléments de $P^{-1}CP$, le th éorème de Gershgorin implique l'existence d'un indice i tel que $|\lambda(\varepsilon) - (\lambda_i + \varepsilon e_{ii})| \leq \varepsilon \sum_{j\neq i} |e_{ij}|$. L'inégalité triangulaire donne alors

$$|\lambda(\varepsilon) - \lambda_i| \leq \varepsilon . \max_i (\sum_j |e_{ij}|) \leq \varepsilon . \left\|P^{-1}CP\right\|_\infty \leq \varepsilon . \left\|P^{-1}\right\|_\infty . \|C\|_\infty . \|P\|_\infty$$

ce qui démontre l'affirmation du théorème, car $\kappa_\infty(P) = \left\|P^{-1}\right\|_\infty . \|C\|_\infty$ (condition de T). cqfd

Remarque 238. La condition du calcul des valeurs propres depend de la condition de la matrice de transfor- mation P. Si la matrice A est symétrique (P est orthogonale), le problème est bien conditionné.Toutefois, observons qu'on obtient seulement une estimation pour l'erreur absolue et non pour l'erreur relative.

Théorème 239 (différentiabilité des valeurs propres). Soit λ_1 une racine simple de $P_A(\lambda) = 0$ Alors, pour $|\varepsilon|$ suffisamment petit, la matrice $A(\varepsilon) = A + \varepsilon C$ possède une valeur propre unique $\lambda_1(\varepsilon)$ proche de λ_1 . La fonction $\lambda_1(\varepsilon)$ est différentiable (même analytique) et on a

$$\lambda_1(\varepsilon) = \lambda_1 + \varepsilon \frac{u_1^* C v_1}{u_1^* v_1} + O(\varepsilon^2) \tag{7.1}$$

où v_1 est le vecteur propre à droite ($Av_1 = \lambda v_1$) et u_1 est le vecteur propre à gauche ($u_1^* A = \lambda u_1^*$). On peut supposer que $\|v_1\| = \|u_1\| = 1$

Démonstration. Soit $p(\lambda, \varepsilon) = P_{A+\varepsilon C}(\lambda) = \det(A + \varepsilon C - \lambda I)$. Comme

$$p(\lambda_1, 0) = 0 \quad \text{et} \quad \frac{\partial p(\lambda_1, 0)}{\partial \lambda} \neq 0$$

le théorème des fonctions implicites garantit l'existence d'une fonction différentiable $\lambda_1(\varepsilon)$ (même analytique), tel que $\lambda_1(0) = \lambda_1$ et $p(\lambda_1(\varepsilon), \varepsilon) = 0$. Il existe donc un vecteur $v_1(\varepsilon)$ tel que

$$(A(\varepsilon) - \lambda_1(\varepsilon) I) v_1(\varepsilon) = 0. \tag{7.2}$$

La matrice dans (11.2) étant de rang $n-1$, on peut fixer une composante à 1 et appliquer la règle de Cramer. Ceci montre que les autres composantes sont des fonctions rationnelles des éléments de la matrice $A+\varepsilon C-\lambda_1(\varepsilon)I$ et donc diff érentiables. Après la normalisation à $v_1(\lambda)^T v_1(\lambda)=1$, la fonction $v_1(\lambda)$ reste diffé rentiable. Pour calculer $\lambda_1^{'}(0)$, nous pouvons dériver l' équation (11.2) par rapport à ε et poser ensuite $\varepsilon=0$. Ceci donne

$$(A-\lambda_1 I)v_1^{'}(0)+(C-\lambda_1^{'}(0)I)v_1=0 \tag{7.3}$$

En multipliant cette relation par u_1^*, on obtient $u_1^*(C-\lambda_1^{'}(0)I)v_1=0$, ce qui permet de calculer $\lambda_1^{'}(0)$ et démontre la formule (11.1). cqfd

Conséquences. La formule (11.1) du théorème pré cédent montre que plus le vecteur propre de droite est parallèle au vecteur propre de gauche, mieux la valeur propre correspondante est bien conditionnée (par exemple, pour les matrices symétriques les deux vecteurs sont identiques) ; plus ils se rapprochent de l'orthogonalité, plus la valeur propre est mal conditionnée. Si la matrice n'est pas symétrique (ou normale), le calcul de λ_1 (valeur propre simple) peut être mal conditionné. Considé rons par exemple la matrice

$$A=\begin{pmatrix}1 & \alpha\\ 0 & 2\end{pmatrix} \quad \text{où} \quad v_1=\begin{pmatrix}1\\0\end{pmatrix}, \quad u_1=\frac{1}{\sqrt{1+\alpha^2}}\begin{pmatrix}1\\-\alpha\end{pmatrix}$$

Dans cette situation, la formule (11.1) nous donne $\lambda_1(\varepsilon)-\lambda_1=\varepsilon.(c_{11}-\alpha c_{21})+O(\varepsilon^2)$ et le calcul de$\lambda_1=1$ est mal conditionné si α est grand. Exemple 1.4 Considérons la matrice (boîte de Jordan)

$$A=\left.\begin{pmatrix}\lambda_1 & 1 & & \\ & \lambda_1 & \ddots & \\ & & \ddots & 1\\ & & & \lambda_1\end{pmatrix}\right\} n \tag{7.4}$$

Le polynôme caractéristique de $A+\varepsilon C$ satisfait

$$\det(A+\varepsilon C-\lambda I)=(\lambda_1-\lambda)^n-(-1)^n.\varepsilon.c_{n1}+O(\varepsilon^2)+O(\varepsilon.|\lambda_1-\lambda|).$$

Si $c_{n1}\neq 0$, les termes $O(\varepsilon^2)$ et $O(\varepsilon.|\lambda_1-\lambda|)$ sont négligeables par rapport à ε . Les valeurs propres de $A+\varepsilon C$ sont alors approximativement données par les racines de

$$(\lambda_1-\lambda)^n-(-1)^n.\varepsilon.c_{n1}=0 \tag{7.5}$$

c'est-à-dire $\lambda=\lambda_1+(\varepsilon.c_{n1})^{1/n}$ (observer que $(\varepsilon.c_{n1})^{1/n}$ donne n valeurs complexes distinctes - multiples des racines de l'unité). **Expérience numérique**. Prenons la matrice (11.4) avec $\lambda_1=1$ et $n=5$. Les éléments de la matrice C sont des nombres aléatoires dans l'intervalle $[-1,1]$. Le dessin 7 ci-contre montre les 5 valeurs propres de $A+\varepsilon C$ pour $\varepsilon=10^{-4},10^{-5},...,10^{-10}$. L'erreur est $\approx 10^{-1}$ pour $\varepsilon=10^{-5}$ et $\approx 10^{-2}$ pour $\varepsilon=10^{-10}$, ce qui correspond à la formule (11.5) pour $n=5$. **Conséquence**. Si la dimension n d'une boite de Jordan est plus grande que 1 , le calcul de la valeur propre de cette matrice est tr ès mal conditionné.

7.1 Condition du calcul des vecteurs propres

Considérons la situation où toutes les valeurs propres de A sont distinctes. La démonstration du théorème sur la diffé rentiabilité des valeurs propres montre (voir formule (11.2)) que les vecteurs propres normalisés $v_i(\varepsilon)$ de $A+\varepsilon C$ sont des fonctions différentiables de ε . Pour é tudier la condition du calcul des vecteurs propres, nous exprimons $v_1'(0)$ dans la base des vecteurs propres (de droite) '

$$v_1'(0)=\sum_{i=1}^{n}\alpha_i v_i. \tag{7.6}$$

La formule (11.3) donne alors

$$\sum_{j=2}^{n}(\lambda_j-\lambda_1)\alpha_j v_j+(C-\lambda_1'(0)I)\,v_1=0. \tag{7.7}$$

En multipliant (11.7) par le vecteur propre de gauche u_1^* (observer que $u_1^*v_1=0$ pour $i\neq j$), on obtient α_i (pour $i\geq 2$) de la relation $(\lambda_i-\lambda_1)\alpha_i u_i^*v_i+u_i^*Cv_1=0$. La normalisation $\|v_1(\varepsilon)\|_2^2=1$ donne (en la dérivant) $v_1^*v_1'(0)=0$ et on en déduit que $\alpha_1=-\sum_{j=2}^n\alpha_i v_1^*v_i$. Si l'on insère les formules pour α_i dans (11.6), on obtient pour $v_1(\varepsilon)=v_1+\varepsilon v_1'(0)+O(\varepsilon^2)$ la relation

$$v_1(\varepsilon)=v_1+\varepsilon\sum_{j=2}^{n}\frac{u_i^*Cv_1}{(\lambda_1-\lambda_i)u_i^*v_i}(v_i-v_1v_1^*v_i)+O(\varepsilon^2). \tag{7.8}$$

De cette formule, on voit que la condition du calcul du vecteur propre v_1 dépend de la grandeur $u_i^*v_i$ (comme c'est le cas pour la valeur propre; voir la formule (11.1)) et aussi de la distance entre λ_1 & et les autres valeurs propres de A . **Un algorithme dangereux** La première méthode (déjà utilisée par Lagrange) pour calculer les valeurs propres d'une matrice A est la suivante : *calculer d'abord les coefficients du polynôme caractéristique $P_A(\lambda)$ et déterminer ' ensuite les zéros de ce polynôme*. Si la dimension de A est très petite (disons $n\leq 3$) ou si l'on fait le calcul en arithmétique exacte, cet algorithme peut être très utile. Par contre, si l'on fait le calcul en virgule flottante, cet algorithme peut donner des mauvaises surprises. Considérons, par exemple, le problème de calculer les valeurs propres de la matrice diagonale

$$A=diag(1,2,3,...,n)$$

dont le polynôme caractéristique est

$$P_A(\lambda)=(1-\lambda)(2-\lambda)(3-\lambda)\cdots(n-\lambda)=(-1)^n\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_1\lambda+a_0 \tag{7.9}$$

Les coefficients calculés satisfont $\tilde{a}=a_i(1+\varepsilon_i)$ avec $|\varepsilon_i|\leq eps$. Cette perturbation dans les coefficients provoque une grande erreur dans les zé ros de (11.9). Les résultats numériques pour $n=9,11,13,15$ (avec $eps\approx 6.10^{-8}$, simple précision) sont dessinés dans la figure V.2. **Conclusion**. Eviter le calcul des coefficients du polynôme caract éristique. Un tel algorithme est numériquement instable.

8 La méthode de la puissance

Un algorithme simple pour calculer les valeurs propres d'une matrice A est basé sur l'itération

$$y_{k+1}=Ay_k \tag{8.1}$$

où y_0 est un vecteur arbitraire. Dans le théorème suivant, on démontre que $y_k=A^ky_0$ (*méthode de la puissance*) tend vers un vecteur propre de A et que le *quotient de Rayleigh* $y_k^*Ay_k/y_k^*y_k$ est une approximation d'une valeur propre de A .

Théorème 240. Soit A une matrice diagonalisable de valeurs propres $\lambda_1,\lambda_2,...,\lambda_n$ et de vecteurs propres $v_1,v_2,...,v_n$ (normalisés par $\|v_i\|_2=1$). Si $|\lambda_1|>|\lambda_2|\geq ...\geq|\lambda_n|$, les vecteurs y_k de l'itération (12.1) vérifient

$$y_k=\lambda_1^k(a_1v_1+O(|\lambda_2/\lambda_1|)^k \tag{8.2}$$

(le nombre a_1 est défini par $y_0=\sum_i a_iv_i$). Le quotient de Rayleigh satisfait (si $a_1\neq 0$)

$$\frac{y_k^*Ay_k}{y_k^*y_k}=\lambda_1+O(|\lambda_2/\lambda_1|^k) \tag{8.3}$$

Si A est une matrice normale (c'est-à-dire. que les vecteurs propres sont orthogonaux), l'erreur dans (12.3) est $O(|\lambda_2/\lambda_1|^{2k})$

Démonstration. Exprimons le vecteur de départ y_0 dans la base des vecteurs propres, c'est-à-dire $y_0 = \sum_{i=1}^{n} a_i v_i$. Par ré currence, on voit que

$$y_k = A^k y_0 = \sum_{i=1}^{n} a_i \lambda_i^k v_i = \lambda_1^k (a_1 v_1 + \sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i \tag{8.4}$$

ce qui démontre la formule (12.2). De cette relation, on déduit que

$$y_k^* A y_k = y_k^* y_{k+1} = \sum_{i=1}^{n} |a_i|^2 |\lambda_i|^{2k} \lambda_i + \sum_{i \neq j}^{n} \tilde{a}_i a_j \tilde{\lambda}_i^k \lambda_j^{k+1} v_i^* v_j \tag{8.5}$$

$$y_k^* y_k = \sum_{i=1}^{n} |a_i|^2 |\lambda_i|_i^{2k} + \sum_{i \neq j}^{n} \tilde{a}_i a_j \tilde{\lambda}_i^k \lambda_j^k v_i^* v_j. \tag{8.6}$$

Si $a_1 \neq 0$, la formule (12.3) est une conséquence de

$$\frac{y_k^* A y_k}{y_k^* y_k} = \frac{|a_1|^2 . |\lambda_1|^{2k} . \lambda_1 . (1 + O(|\lambda_2/\lambda_1|^k)}{|a_1|^2 . |\lambda_1|^{2k} . (1 + O(|\lambda_2/\lambda_1|^k)}. \tag{8.7}$$

Pour une matrice normale, le deuxième terme dans les formules (12.5) et (12.6) est absent et l'expression $O(|\lambda_2/\lambda_1|^k$ peut être remplacée par $O(|\lambda_2/\lambda_1|^{2k}$ dans (12.7) et dans (12.3). cqfd

Exemple 241. Considérons la matrice

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

dont la valeur propre la plus grande est $\lambda_1 = 2(1 + \cos(\pi/4)) \approx 3,41421356$. Quelques itérations de la méthode de la puissance nous donnent

$$y_0 = (1,1,1)^T \quad y_1 = (3,4,3)^T \quad y_2 = (10,14,10)^T$$

et une première approximation de λ_1 est obtenue par

$$\frac{y_1^* A y_1}{y_1^* y_1} = \frac{y_1^* y_2}{y_1^* y_1} = \frac{116}{34} \approx 3,41176$$

Remarques. Les éléments du vecteur y_k croissent exponentiellement avec k . Il est alors recommandé de normaliser y_k après chaque itération, c'est-à-dire. de remplacer y_k par $y_k / \left\| y_k \right\|$. Sinon, on risque un "overflow". Si $|\lambda_2/\lambda_1|$ est proche de 1, la convergence est très lente. Pour accélérer la convergence, on utilise la modification suivante :

9 Méthode de la puissance inverse de Wielandt

Supposons qu'on connaisse une approximation μ de la valeur propre cherchée λ_1 (il n'est pas · né cessaire de supposer que λ_1 soit la plus grande valeur propre de A). L'idée est d'appliquer l'itération (12.1) à la matrice $(A - \mu I)^{-1}$ (Les valeurs propres de cette matrice sont $(\lambda_i - \mu)^{-1}$. Si μ est proche de λ , on a ·

$$\frac{1}{\left|\lambda_1 - \mu\right|} \gg \frac{1}{\left|\lambda_i - \mu\right|} \quad \text{pour} \quad i \geq 2$$

et la convergence va être très rapide. L'itération devient alors $y_{k+1} = (A - \mu I)^{-1} y_k$ ou

$$(A - \mu I) y_{k+1} = y_k \tag{9.1}$$

Après avoir calculé la décomposition LU de la matrice $A - \mu I$, une itération de (13.1) ne coûte pas plus cher qu'une de (12.1). Pour la matrice A de l'exemple précédent, choisissons $\mu = 3,41$ et $y_0 = (1;1,4;1)^T$ Deux itérations de (13.1) nous donnent

$$y_1 = \begin{pmatrix} 236,134453781513 \\ 333,949579831933 \\ 236,134453781513 \end{pmatrix}, \quad y_2 = \begin{pmatrix} 56041,9461902408 \\ 79255,2785820210 \\ 56041,9461902408 \end{pmatrix}$$

et on obtient

$$\frac{1}{\lambda_1 - 3,41} \approx \frac{y_1^*(A-\mu I)^{-1} y_1}{y_1^* y_1} = \frac{y_1^* y_2}{y_1^* y_1} \approx 237,328870774159$$

De cette relation, on calcule λ_1 et on obtient l'approximation 3,41421356237333 . Les 13 premiers chiffres sont corrects. La méthode de la puissance (et celle de Wielandt) est importante pour la compréhension d'autres algorithmes. Si l'on veut calculer toutes les valeurs propres d'une matrice, on utilise des méthodes encore plus sophistiquées. En pratique, on procède de la manière suivante :

— on distingue les cas : A symétrique ou A quelconque.
— on cherche P telle que $P^{-1}AP$ devienne une matrice de Hessenberg (ou une matrice tridiagonale, si A est symétrique) ; voir V.3.
— on applique l'algorithme QR à la matrice H (voir V.6).
— si H est une matrice tridiagonale et symétrique, on peut é galement appliquer la méthode de bissection (voir V.4).

10 VALEURS PROPRES ET VECTEURS PROPRES

Les premiers vecteurs et valeurs propres viennent des équations différentielles (Lagrange 1759, *théorie du son* ; Lagrange 1781, des matrices 6×6 dans le but de calculer les perturbations séculaires des orbites des 6 planètes connues à l'époque, *Oeuvres V*, p. 125-490). Aujourd'hui, le calcul des valeurs et vecteurs propres est indispensable dans toutes les branches de la science, en particulier pour la solution des systèmes des équations différentielles linéaires, en théorie de stabilité, pour les questions de convergence de processus itératifs, et en physique et chimie (mécanique, circuits, cinétique chimique, équation de Schrödinger).

FIG. V.1 : Une application linéaire comme champ de vecteurs (à gauche) ; transformée sur la base des vecteurs propres (à droite).

Observons en figure V.1 (à gauche) le champ de vecteurs d'une équation différentielle $y' = Ay$.

Deux directions sautent aux yeux : ce sont les directions où le vecteur Av prend la même direction que le vecteur v , c'est-à-dire., où

$$Av = \lambda v \quad \text{ou} \quad (A - \lambda I)v = 0 \tag{10.1}$$

Si cette équation est vérifiée, $\lambda \in \mathbb{C}$ s'appelle *valeur propre* de la matrice A et $v \in \mathbb{C}^n (v \neq 0)$ est le *vecteur propre* correspondant. L'équation (10.1) possède une solution v non nulle si et seulement si

$$P_A(\lambda) = \det((A - \lambda I) = 0$$

Le polynôme $P_A(\lambda)$ est le *polynôme caractéristique* de la matrice A . Les valeurs propres de A sont alors les zéros du polynôme caractéristique.

11 LA CONDITION DU CALCUL DES VALEURS PROPRES

A cause des erreurs d'arrondi, les éléments d'une matrice A , pour laquelle on cherche les valeurs propres, ne sont pas exacts. Ils sont plutôt égaux à

$$\tilde{a}_{ij} = a_{ij}(1+\varepsilon_{ij}) \quad \text{avec} \quad \left|\varepsilon_{ij}\right| \leq eps$$

(*eps* étant la précision de l'ordinateur, est supposée être très petite). Il est alors très important d'étudier l'influence de ces perturbations sur les valeurs propres et sur les vecteurs propres de la matrice.

Pour montrer ceci, considérons la famille de matrices

$$A(\varepsilon) = A + \varepsilon C \quad \text{où} \quad |\varepsilon| \leq eps \quad \text{et} \quad \left|c_{ij}\right| \leq \left|a_{ij}\right|$$

(souvent, la dernière hypothèse va être remplacée par $\|C\| \leq \|A\|$).

Théorème 242 (Gershgorin). Soit A une matrice $n \times n$ (avec des éléments dans $\mathbb{R}$ ou dans $\mathbb{C}$).

a) Si λ est une valeur propre de A, alors il existe un indice i tel que

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|a_{ij}\right|$$

c'est-à-dire. que toutes les valeurs propres de A se trouvent dans l'union des disques $D_i = \left\{\lambda;\ |\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|a_{ij}\right|\right\}$

b) Si une composante connexe de $\bigcup_{i=1}^{n} D_i$ consiste de k disques, elle contient exactement k valeurs propres de A.

Démonstration. Soit $v \neq 0$ un vecteur propre et choisissons l'indice i tel que $|v_i| \geq \left|v_j\right|$ pour tout j . La ligne i de l'équation $Av = \lambda v$ donne

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} v_j = (\lambda - a_{ii}) v_i.$$

En divisant par v_i et en utilisant l'inégalité du triangle, on obtient U

$$|\lambda - a_{ii}| = \left|\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} \frac{v_j}{v_i}\right| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|a_{ij}\right|$$

L'affirmation (b) est vraie si A est une matrice diagonale. Le cas général est obtenu par un argument de continuité en faisant tendre les éléments en dehors de la diagonale vers zéro. cqfd

Théorème 243. Soit A une matrice diagonalisable, c'est-à-dire, il existe P avec $P^{-1}AP = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ et soit $A(\varepsilon) = A + \varepsilon C$. Alors, pour chaque valeur propre $\lambda(\varepsilon)$ de $A(\varepsilon)$, il existe un λ_i avec

$$|\lambda(\varepsilon) - \lambda_i| \leq \varepsilon . \kappa_\infty(P) . \|C\|_\infty$$

Démonstration. Nous transformons la matrice$A(\varepsilon) = A + \varepsilon C$ par la même matrice, qui transforme A sous forme diagonale :

$$P^{-1}A(\varepsilon)P = diag(\lambda_1, \lambda_2, ..., \lambda_n) + \varepsilon P^{-1}CP$$

Si l'on dénote par e_{ij} les éléments de $P^{-1}CP$, le théorème de Gershgorin implique l'existence d'un indice i tel que $|\lambda(\varepsilon) - (\lambda_i + \varepsilon e_{ii})| \leq \varepsilon \sum_{j \neq i} \left|e_{ij}\right|$. L'inégalité triangulaire donne alors

$$|\lambda(\varepsilon) - \lambda_i| \leq \varepsilon . \max_i (\sum_j \left|e_{ij}\right|) \leq \varepsilon . \left\|P^{-1}CP\right\|_\infty \leq \varepsilon . \left\|P^{-1}\right\|_\infty . \|C\|_\infty . \|P\|_\infty$$

ce qui démontre l'affirmation du théorème, car $\kappa_\infty(P) = \left\|P^{-1}\right\|_\infty . \|C\|_\infty$ (condition de T). cqfd

Remarque 244. La condition du calcul des valeurs propres depend de la condition de la matrice de transfor- mation P. Si la matrice A est symétrique (P est orthogonale), le problème est bien conditionné.Toutefois, observons qu'on obtient seulement une estimation pour l'erreur absolue et non pour l'erreur relative.

Théorème 245 (différentiabilité des valeurs propres). Soit λ_1 une racine simple de $P_A(\lambda) = 0$ Alors, pour $|\varepsilon|$ suffisamment petit, la matrice $A(\varepsilon) = A + \varepsilon C$ possède une valeur propre unique $\lambda_1(\varepsilon)$ proche de λ_1 . La fonction $\lambda_1(\varepsilon)$ est différentiable (même analytique) et on a

$$\lambda_1(\varepsilon) = \lambda_1 + \varepsilon \frac{u_1^* C v_1}{u_1^* v_1} + O(\varepsilon^2) \tag{11.1}$$

où v_1 est le vecteur propre à droite ($Av_1 = \lambda v_1$) et u_1 est le vecteur propre à gauche ($u_1^* A = \lambda u_1^*$). On peut supposer que $\|v_1\| = \|u_1\| = 1$

Démonstration. Soit $p(\lambda, \varepsilon) = P_{A+\varepsilon C}(\lambda) = \det(A + \varepsilon C - \lambda I)$. Comme

$$p(\lambda_1, 0) = 0 \quad \text{et} \quad \frac{\partial p(\lambda_1, 0)}{\partial \lambda} \neq 0$$

le théorème des fonctions implicites garantit l'existence d'une fonction différentiable $\lambda_1(\varepsilon)$ (même analytique), tel que $\lambda_1(0) = \lambda_1$ et $p(\lambda_1(\varepsilon), \varepsilon) = 0$. Il existe donc un vecteur $v_1(\varepsilon)$ tel que

$$(A(\varepsilon) - \lambda_1(\varepsilon) I) v_1(\varepsilon) = 0. \tag{11.2}$$

La matrice dans (11.2) étant de rang $n-1$, on peut fixer une composante à 1 et appliquer la règle de Cramer. Ceci montre que les autres composantes sont des fonctions rationnelles des éléments de la matrice $A + \varepsilon C - \lambda_1(\varepsilon) I$ et donc différentiables. Après la normalisation à $v_1(\lambda)^T v_1(\lambda) = 1$, la fonction $v_1(\lambda)$ reste différentiable.

Pour calculer $\lambda_1^{'}(0)$, nous pouvons dériver l'équation (11.2) par rapport à ε et poser ensuite $\varepsilon = 0$. Ceci donne

$$(A - \lambda_1 I) v_1^{'}(0) + (C - \lambda_1^{'}(0) I) v_1 = 0 \tag{11.3}$$

En multipliant cette relation par u_1^*, on obtient $u_1^*(C - \lambda_1^{'}(0) I) v_1 = 0$, ce qui permet de calculer $\lambda_1^{'}(0)$ et démontre la formule (11.1). cqfd

Conséquences. La formule (11.1) du théorème précédent montre que plus le vecteur propre de droite est parallèle au vecteur propre de gauche, mieux la valeur propre correspondante est bien conditionnée (par exemple, pour les matrices symétriques les deux vecteurs sont identiques) ; plus ils se rapprochent de l'orthogonalité, plus la valeur propre est mal conditionnée.

Si la matrice n'est pas symétrique (ou normale), le calcul de λ_1 (valeur propre simple) peut être mal conditionné. Considérons par exemple la matrice

$$A = \begin{pmatrix} 1 & \alpha \\ 0 & 2 \end{pmatrix} \quad \text{où} \quad v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad u_1 = \frac{1}{\sqrt{1+\alpha^2}} \begin{pmatrix} 1 \\ -\alpha \end{pmatrix}$$

Dans cette situation, la formule (11.1) nous donne $\lambda_1(\varepsilon) - \lambda_1 = \varepsilon . (c_{11} - \alpha c_{21}) + O(\varepsilon^2)$ et le calcul de$\lambda_1 = 1$ est mal conditionné si α est grand.

Exemple 1.4 Considérons la matrice (boîte de Jordan)

$$A = \left.\begin{pmatrix} \lambda_1 & 1 & & \\ & \lambda_1 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_1 \end{pmatrix}\right\} n \tag{11.4}$$

Le polynôme caractéristique de $A + \varepsilon C$ satisfait

$$\det(A + \varepsilon C - \lambda I) = (\lambda_1 - \lambda)^n - (-1)^n . \varepsilon . c_{n1} + O(\varepsilon^2) + O(\varepsilon . |\lambda_1 - \lambda|).$$

Si $c_{n1} \neq 0$, les termes $O(\varepsilon^2)$ et $O(\varepsilon . |\lambda_1 - \lambda|)$ sont négligeables par rapport à ε . Les valeurs propres de $A + \varepsilon C$ sont alors approximativement données par les racines de

$$(\lambda_1 - \lambda)^n - (-1)^n . \varepsilon . c_{n1} = 0 \tag{11.5}$$

c'est-à-dire $\lambda = \lambda_1 + (\varepsilon . c_{n1})^{1/n}$ (observer que $(\varepsilon . c_{n1})^{1/n}$ donne n valeurs complexes distinctes - multiples des racines de l'unité).

Expérience numérique. Prenons la matrice (11.4) avec $\lambda_1 = 1$ et $n = 5$. Les éléments de la matrice C sont des nombres aléatoires dans l'intervalle $[-1, 1]$. Le dessin 7 ci-contre montre les 5 valeurs propres de $A + \varepsilon C$ pour $\varepsilon = 10^{-4}, 10^{-5}, ..., 10^{-10}$. L'erreur est $\approx 10^{-1}$ pour $\varepsilon = 10^{-5}$ et $\approx 10^{-2}$ pour $\varepsilon = 10^{-10}$, ce qui correspond à la formule (11.5) pour $n = 5$.

Conséquence. Si la dimension n d'une boite de Jordan est plus grande que 1 , le calcul de la valeur propre de cette matrice est très mal conditionné.

11.1 Condition du calcul des vecteurs propres

Considérons la situation où toutes les valeurs propres de A sont distinctes. La démonstration du théorème sur la différentiabilité des valeurs propres montre (voir formule (11.2)) que les vecteurs propres normalisés $v_i(\varepsilon)$ de $A + \varepsilon C$ sont des fonctions différentiables de ε . Pour étudier la condition du calcul des vecteurs propres, nous exprimons $v_1'(0)$ dans la base des vecteurs propres (de droite) '

$$v_1'(0) = \sum_{i=1}^{n} \alpha_i v_i. \tag{11.6}$$

La formule (11.3) donne alors

$$\sum_{j=2}^{n} (\lambda_j - \lambda_1)\alpha_j v_j + (C - \lambda_1'(0) I)\, v_1 = 0. \tag{11.7}$$

En multipliant (11.7) par le vecteur propre de gauche u_1^* (observer que $u_1^* v_1 = 0$ pour $i \neq j$), on obtient α_i (pour $i \geq 2$) de la relation $(\lambda_i - \lambda_1)\alpha_i u_i^* v_i + u_i^* C v_1 = 0$. La normalisation $\|v_1(\varepsilon)\|_2^2 = 1$ donne (en la dérivant) $v_1^* v_1'(0) = 0$ et on en déduit que $\alpha_1 = -\sum_{j=2}^{n} \alpha_i v_1^* v_i$. Si l'on insère les formules pour α_i dans (11.6), on obtient pour $v_1(\varepsilon) = v_1 + \varepsilon v_1'(0) + O(\varepsilon^2)$ la relation

$$v_1(\varepsilon) = v_1 + \varepsilon \sum_{j=2}^{n} \frac{u_i^* C v_1}{(\lambda_1 - \lambda_i) u_i^* v_i} (v_i - v_1 v_1^* v_i) + O(\varepsilon^2). \tag{11.8}$$

De cette formule, on voit que la condition du calcul du vecteur propre v_1 dépend de la grandeur $u_i^* v_i$ (comme c'est le cas pour la valeur propre ; voir la formule (11.1)) et aussi de la distance entre λ_1 & et les autres valeurs propres de A .

Un algorithme dangereux

La première méthode (déjà utilisée par Lagrange) pour calculer les valeurs propres d'une matrice A est la suivante : *calculer d'abord les coefficients du polynôme caractéristique $P_A(\lambda)$ et déterminer ' ensuite les zéros de ce polynôme*. Si la dimension de A est très petite (disons $n \leq 3$) ou si l'on fait le calcul en arithmétique exacte, cet algorithme peut être très utile. Par contre, si l'on fait le calcul en virgule flottante, cet algorithme peut donner des mauvaises surprises.

Considérons, par exemple, le problème de calculer les valeurs propres de la matrice diagonale

$$A = diag(1, 2, 3, ..., n)$$

dont le polynôme caractéristique est

$$P_A(\lambda) = (1-\lambda)(2-\lambda)(3-\lambda)\cdots(n-\lambda) = (-1)^n \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 \tag{11.9}$$

Les coefficients calculés satisfont $\tilde{a} = a_i(1+\varepsilon_i)$ avec $|\varepsilon_i| \leq eps$. Cette perturbation dans les coefficients provoque une grande erreur dans les zéros de (11.9). Les résultats numériques pour $n = 9, 11, 13, 15$ (avec $eps \approx 6.10^{-8}$, simple précision) sont dessinés dans la figure V.2.

Conclusion. Eviter le calcul des coefficients du polynôme caractéristique. Un tel algorithme est numériquement instable.

12 LA METHODE DE LA PUISSANCE

Un algorithme simple pour calculer les valeurs propres d'une matrice A est basé sur l'itération

$$y_{k+1} = Ay_k \tag{12.1}$$

où y_0 est un vecteur arbitraire. Dans le théorème suivant, on démontre que $y_k = A^k y_0$ (*méthode de la puissance*) tend vers un vecteur propre de A et que le *quotient de Rayleigh* $y_k^* A y_k / y_k^* y_k$ est une approximation d'une valeur propre de A .

Théorème 246. Soit A une matrice diagonalisable de valeurs propres $\lambda_1, \lambda_2, ..., \lambda_n$ et de vecteurs propres $v_1, v_2, ..., v_n$ (normalisés par $\|v_i\|_2 = 1$).

Si $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_n|$, les vecteurs y_k de l'itération (12.1) vérifient

$$y_k = \lambda_1^k (a_1 v_1 + O(|\lambda_2/\lambda_1|)^k \tag{12.2}$$

(le nombre a_1 est défini par $y_0 = \sum_i a_i v_i$). Le quotient de Rayleigh satisfait (si $a_1 \neq 0$)

$$\frac{y_k^* A y_k}{y_k^* y_k} = \lambda_1 + O(|\lambda_2/\lambda_1|^k) \tag{12.3}$$

Si A est une matrice normale (c'est-à-dire. que les vecteurs propres sont orthogonaux), l'erreur dans (12.3) est $O(|\lambda_2/\lambda_1|^{2k})$

Démonstration. Exprimons le vecteur de départ y_0 dans la base des vecteurs propres, c'est-à-dire $y_0 = \sum_{i=1}^n a_i v_i$. Par récurrence, on voit que

$$y_k = A^k y_0 = \sum_{i=1}^n a_i \lambda_i^k v_i = \lambda_1^k (a_1 v_1 + \sum_{i=2}^n a_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i \tag{12.4}$$

ce qui démontre la formule (12.2). De cette relation, on déduit que

$$y_k^* A y_k = y_k^* y_{k+1} = \sum_{i=1}^n |a_i|^2 |\lambda_i|^{2k} \lambda_i + \sum_{i \neq j}^n \tilde{a}_i a_j \tilde{\lambda}_i^k \lambda_j^{k+1} v_i^* v_j \tag{12.5}$$

$$y_k^* y_k = \sum_{i=1}^n |a_i|^2 |\lambda_i|_i^{2k} + \sum_{i \neq j}^n \tilde{a}_i a_j \tilde{\lambda}_i^k \lambda_j^k v_i^* v_j. \tag{12.6}$$

Si $a_1 \neq 0$, la formule (12.3) est une conséquence de

$$\frac{y_k^* A y_k}{y_k^* y_k} = \frac{|a_1|^2 . |\lambda_1|^{2k} . \lambda_1 . (1 + O(|\lambda_2/\lambda_1|^k)}{|a_1|^2 . |\lambda_1|^{2k} . (1 + O(|\lambda_2/\lambda_1|^k)}. \tag{12.7}$$

Pour une matrice normale, le deuxième terme dans les formules (12.5) et (12.6) est absent et l'expression $O(|\lambda_2/\lambda_1|^k$ peut être remplacée par $O(|\lambda_2/\lambda_1|^{2k}$ dans (12.7) et dans (12.3). cqfd

Exemple 247. Considérons la matrice

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

dont la valeur propre la plus grande est $\lambda_1 = 2(1 + \cos(\pi/4)) \approx 3,41421356$. Quelques itérations de la méthode de la puissance nous donnent

$$y_0 = (1,1,1)^T \quad y_1 = (3,4,3)^T \quad y_2 = (10,14,10)^T$$

et une première approximation de λ_1 est obtenue par

$$\frac{y_1^* A y_1}{y_1^* y_1} = \frac{y_1^* y_2}{y_1^* y_1} = \frac{116}{34} \approx 3,41176$$

Remarques. Les éléments du vecteur y_k croissent exponentiellement avec k . Il est alors recommandé de normaliser y_k après chaque itération, c'est-à-dire. de remplacer y_k par $y_k / \left\| y_k \right\|$. Sinon, on risque un "overflow". Si $|\lambda_2/\lambda_1|$ est proche de 1, la convergence est très lente. Pour accélérer la convergence, on utilise la modification suivante :

13 METHODE DE LA PUISSANCE INVERSE DE WIELANDT

Supposons qu'on connaisse une approximation μ de la valeur propre cherchée λ_1 (il n'est pas · nécessaire de supposer que λ_1 soit la plus grande valeur propre de A). L'idée est d'appliquer l'itération (12.1) à la matrice $(A - \mu I)^{-1}$ (Les valeurs propres de cette matrice sont $(\lambda_i - \mu)^{-1}$. Si μ est proche de λ , on a ·

$$\frac{1}{\left|\lambda_1 - \mu\right|} \gg \frac{1}{\left|\lambda_i - \mu\right|} \quad \text{pour} \quad i \geq 2$$

et la convergence va être très rapide. L'itération devient alors $y_{k+1} = (A - \mu I)^{-1} y_k$ ou

$$(A - \mu I) y_{k+1} = y_k \tag{13.1}$$

Après avoir calculé la décomposition LU de la matrice $A - \mu I$, une itération de (13.1) ne coûte pas plus cher qu'une de (12.1).

Pour la matrice A de l'exemple précédent, choisissons $\mu = 3,41$ et $y_0 = (1;1,4;1)^T$ Deux itérations de (13.1) nous donnent

$$y_1 = \begin{matrix} 236,134453781513 \\ 333,949579831933, \\ 236,134453781513 \end{matrix} \quad y_2 = \begin{matrix} 56041,9461902408 \\ 79255,2785820210 \\ 56041,9461902408 \end{matrix}$$

et on obtient

$$\frac{1}{\lambda_1 - 3,41} \approx \frac{y_1^* (A - \mu I)^{-1} y_1}{y_1^* y_1} = \frac{y_1^* y_2}{y_1^* y_1} \approx 237,328870774159$$

De cette relation, on calcule λ_1 et on obtient l'approximation $3,41421356237333$. Les 13 premiers chiffres sont corrects.

La méthode de la puissance (et celle de Wielandt) est importante pour la compréhension d'autres algorithmes. Si l'on veut calculer toutes les valeurs propres d'une matrice, on utilise des méthodes encore plus sophistiquées. En pratique, on procède de la manière suivante :

— on distingue les cas : A symétrique ou A quelconque.
— on cherche P telle que $P^{-1}AP$ devienne une matrice de Hessenberg (ou une matrice tridiagonale, si A est symétrique) ; voir V.3.
— on applique l'algorithme QR à la matrice H (voir V.6).
— si H est une matrice tridiagonale et symétrique, on peut également appliquer la méthode de bissection (voir V.4).

14 Transformation sous forme tridiagonale (ou de Hessenberg)

Avec la transformation $v = Pu$ (où est P une matrice inversible) le problème

$$Av = \lambda v$$

devient

$$P^{-1}APu = \lambda u$$

Donc, les valeurs propres de A et de $P^{-1}AP$ sont les mêmes et les vecteurs propres v_i de A se transforment par $v_i = Pu_i$. Le but de ce paragraphe est de trouver une matrice P telle que $P^{-1}AP$ devienne "plus simple". La situation idéale serait trouvée si $P^{-1}AP$ devenait diagonale ou triangulaire - mais une telle transformation nécessiterait déjà la connaissance des valeurs propres. Alors, on cherche P tel que $P^{-1}AP$ soit sous forme de Hessenberg

$$P^{-1}AP = H = \begin{pmatrix} * & * & \cdots & \cdots & * \\ * & * & \ddots & & \vdots \\ & * & \ddots & \ddots & * \\ & & \ddots & \ddots & * \\ & & & * & * \end{pmatrix} \tag{14.1}$$

c'est-à-dire, $h_{ij} = 0$ pour $i > j + 1$. Pour arriver à ce but, nous considérons deux algorithmes.

14.1 a) A l'aide des transformations élémentaires

Comme pour l'élimination de Gauss, nous utilisons les transformations pour faire apparaitre les zéros - colonne par colonne - dans (17.1). Dans un premier pas, nous choisissons $k \geq 2$ tel que $|a_{k1}| \geq |a_{j1}|$ pour $j \geq 2$ et nous permutons les lignes 2 et k, c'est-à-dire, nous formons PA o ù P est une matrice de permutation convenable. Pour ne pas changer les valeurs propres, il faut également permuter les colonnes 2 et k (ceci correspond au calcul de $A' = PAP^{-1}$ car $P^2 = I$ (et donc $P = P^{-1}$). Si $a'_{21} = 0$, on a aussi $a'_{i1} = 0$ pour $i \geq 3$ et le premier pas est terminé. Sinon, nous déterminons

$$L_2 = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & -l_{n2} & \cdots & 0 & 1 \end{pmatrix} \quad \text{telle que} \quad L_2A' = \begin{pmatrix} a'_{11} & a'_{12} & \cdots & a'_{1n} \\ a'_{21} & a'_{22} & \cdots & a'_{2n} \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}$$

Pour ceci, on définit $l_{i2} = \frac{a'_{i1}}{a'_{21}}$. Une multiplication à droite avec

$$L_2^{-1} = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & l_{n2} & \cdots & 0 & 1 \end{pmatrix}$$

ne change pas la première colonne de L_2A'. On répète la même procédure avec la sous-matrice de $L_2A'L_2^{-1}$ de dimension $n-1$, et ainsi de suite. A cause des multiplications à droite avec L_i^{-1}, cet algorithme coûte deux fois plus cher que l'élimination de Gauss. Pour la matrice

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 1 \end{pmatrix}$$

on prend

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/2 & 1 \end{pmatrix}$$

et on obtient

$$L_2 A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 0 & 5/2 & -1/2 \end{pmatrix}, \quad \text{puis} \quad L_2 A L_2^{-1} = \begin{pmatrix} 3 & 5/2 & 1 \\ 2 & 5/2 & 3 \\ 0 & 9/4 & -1/2 \end{pmatrix} = H$$

Cet exemple montre un désavantage de cet algorithme : si l'on part avec une matrice symétrique A, la matrice de Hessenberg H, obtenue par cet algorithme, n'est plus symétrique en général.

14.2 b) A l'aide des transformations orthogonales

Il est souvent préférable de travailler avec des réflexions de Householder Commençons par une réflexion pour les coordonnées $2,...,n$ laissant fixe la première coordonnée :$\bar{Q}_2 = I - 2\bar{u}_2\bar{u}_2^T$ ($\|\bar{u}_2\|_2 = 1$) tel que $\bar{Q}_2\bar{A}_1 = \alpha_2 e_1$ où $\bar{A}_1 = (a_{21},...,a_{n1})$. En posant $u_2 = (0,\bar{u}_2)^T$ et $Q_2 = I - 2u_2u_2^T$, la matrice Q_2A contient des zéros dans la première colonne à partir du troisième élément. La multiplication à droite avec $Q_2^{-1} = Q_2^T = Q_2$ ne change pas cette colonne :

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \xrightarrow{Q_2A} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ \alpha_2 & * & * \\ 0 & * & * \end{pmatrix} \xrightarrow{Q_2AQ_2} \begin{pmatrix} a_{11} & * & * \\ \alpha_2 & * & * \\ 0 & * & * \end{pmatrix}$$

Dans le pas suivant, on applique la même procédure à la sous-matrice de dimension $n-1$, etc. Finalement, on arrive à la forme de Hessenberg (17.1) avec la tranformation $P^{-1} = Q_{n-1}....Q_2$ qui est une matrice orthogonale (c'est-à-dire, $P^{-1} = P^T$). Nous avons un double avantage avec cet algorithme :

— il ne faut pas faire une recherche de pivot ;
— si A est symétrique, alors $P^{-1}AP$ est aussi symétrique, et donc tridiagonale.

14.3 Méthode de bissection pour des matrices tridiagonales

Considérons une matrice symétrique tridiagonale

$$A = \begin{pmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & e_3 & & \\ & e_3 & \ddots & \ddots & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n \end{pmatrix}$$

On observe tout d'abord que si un élément e_i est nul, la matrice A est déjà décomposée en deux sous-matrices du m ême type, qui ensemble fournissent les valeurs propres de A . On peut donc supposer, sans restreindre la généralité, que

$$c_i \neq 0 \quad \text{pour} \quad i = 2,...,n. \tag{14.2}$$

Pour cette matrice, il est possible de calculer la valeur $P_A(\lambda)$ du polynôme caractéristique sans connaitre ses coefficients. En effet, si l'on pose

$$A_1 = (d_1), \qquad A_2 = \begin{pmatrix} d_1 & e_2 \\ e_2 & d_2 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} d_1 & e_2 & \\ e_2 & d_2 & e_3 \\ & e_3 & d_3 \end{pmatrix}, \qquad ...$$

et si l'on définit

$$p_i(\lambda) = \det(A_i - \lambda I),$$

on obtient

$$\begin{aligned} p_0(\lambda) &= 1 \\ p_1(\lambda) &= d_1 - \lambda \\ p_i(\lambda) &= (d_i - \lambda)p_{i-1}(\lambda) - e_i^2 p_{i-2}(\lambda), \qquad i = 2, ..., n. \end{aligned} \tag{14.3}$$

La formule de récurrence dans (17.3) est obtenue en dé veloppant le déterminant de la matrice $A_i - \lambda I$ par rapport à la dernière ligne (ou colonne). En principe, on peut maintenant calculer les valeurs propres de A (c'est- à-dire. les zéros de $p_n(\lambda)$ de la manière suivante : chercher un intervalle où $p_n(\lambda)$ change de signe et localiser une racine de $p_n(\lambda) = 0$ par bissection. Les évaluations de $p_n(\lambda)$ sont faites à l'aide de la formule (17.3). Mais il existe une astuce interessante qui permet d'améliorer cet algorithme.

Théorème 248. Si l'équation (17.2) est vérifié, les polyn ômes $p_i(\lambda)$ définis par (17.3) satisfont

a) $p_n'(\hat{\lambda})p_{n-1}(\hat{\lambda}) < 0$ si $p_n(\hat{\lambda}) = 0$ $(\hat{\lambda} \in \mathbb{R})$

b) $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) < 0$ si $p_i(\hat{\lambda}) = 0$ pour un $\{i \in 1, 2, ..., n-1\}$

c) $p_0(\lambda)$ ne change pas de signe sur $\mathbb{R}$.

Démonstration. L'affirmation (c) est triviale. Si $p_i(\hat{\lambda}) = 0$ pour un $\{i \in 1, 2, ..., n-1\}$, la formule de récurrence (17.3) donne l'inégalité $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) \leq 0$. Pour démontrer (b), il suffit d'exclure le cas $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) = 0$. Si deux valeurs consécutives de la suite $\left\{p_i(\hat{\lambda})\right\}$ sont nulles, la formule de récurrence montre que $p_i(\hat{\lambda}) = 0$ pour tout i, ce qui contredit $p_0(\lambda) = 1$. Nous démontrons par récurrence que toutes les racines de $p_i(\lambda)$ sont réelles, simples et séparées par celles de $p_{i-1}(\lambda)$. Il n'y a rien à démontrer pour $i = 1$. Supposons la propriété vraie pour i et montrons qu'elle est encore vraie pour $i+1$. Comme les z éros $\lambda_1 < \lambda_2 < ... < \lambda_i$ sont séparés par ceux de $p_{i-1}(\lambda)$ et comme $p_{i-1}(-\infty) = +\infty$, nous avons $sign\ p_{i-1}(\lambda_j) = (-1)^{j+1}$. Alors, on d éduit de (b) que $sign\ p_{i+1}(\lambda_j) = (-1)^j$. Ceci et le fait que $p_{i+1}(\lambda) = (-1)^{j+1}\lambda^{i+1} + ...$ montrent que $p_{i+1}(\lambda)$ possède un zéro ré el dans chacun des intervalles ouverts $(-\infty, \lambda_1), (\lambda_1, \lambda_2), ..., (\lambda_i, \infty)$ L'affirmation (a) est maintenant une conséquence de (b) et du fait que toutes les racines de $p_{i-1}(\lambda)$ sont réelles simples ; cqfd

Définition 249 (suite de Sturm). Une suite $\{p_0, p_1, ..., p_n\}$ de polynômes à coefficients réels s'appelle une suite de Sturm, si elle vérifie les conditions (a), (b), (c) du Théorème (275)

Considérons une suite de Sturm $\{p_0, p_1, ..., p_n\}$. Si l'on définit

$$\omega(\lambda) = \text{nombre de changements de signes de } \{p_0(\lambda), p_1(\lambda), ..., p_n(\lambda)\}$$

alors le polynôme $p_n(\lambda)$ possède exactement

$$\omega(b) - \omega(a)$$

zéros dans l'intervalle $[a, b]$ (si $p_i(\lambda) = 0$, on définit $signp_i(\lambda) = sign\ p_{i-1}(\lambda)$).

Démonstration. Par continuité, l'entier $\omega(\lambda)$ peut changer sa valeur seulement si une valeur des fonctions $p_i(\lambda)$ devient nulle. La fonction $p_0(\lambda)$ ne change pas de signe. Supposons alors que $p_i(\tilde{\lambda}) = 0$ pour un $i \in \{1, 2, ..., n-1\}$. La condition (b) et la continuité de $p_j(\lambda)$ montrent que seulement les deux situations suivantes sont possibles (ε petit) :

	$\tilde{\lambda} - \varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda} + \varepsilon$
$p_{i-1}(\lambda)$	+	+	+
$p_i(\lambda)$	±	0	±
$p_{i+1}(\lambda)$	−	−	−

	$\tilde{\lambda} - \varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda} + \varepsilon$
$p_{i-1}(\lambda)$	−	−	−
$p_i(\lambda)$	±	0	±
$p_{i+1}(\lambda)$	+	+	+

cqfd

Chaque fois, on a $\omega(\tilde{\lambda}+\varepsilon) = \omega(\tilde{\lambda}) = \omega(\tilde{\lambda}-\varepsilon)$ et la valeur de $\omega(\lambda)$ ne change pas si λ traverse un zéro de $p_i(\lambda)$ pour $i \in \{1,2,...,n-1\}$ Il reste à étudier la fonction $\omega(\lambda)$ dans un voisinage d'un zéro $\tilde{\lambda}$ de $p_n(\lambda)$. La propriété (a) implique que pour les signes de $p_j(\lambda)$ on a seulement les deux possibilités suivantes :

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{n-1}(\lambda)$	+	+	+
$p_n(\lambda)$	+	0	−

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{n-1}(\lambda)$	−	−	−
$p_n(\lambda)$	−	0	+

c'est-à-dire, $\omega(\tilde{\lambda}+\varepsilon) = \omega(\tilde{\lambda}-\varepsilon)+1$. Ceci démontre que la fonction $\omega(\lambda)$ est constante par morceaux et augmente de 1 sa valeur si λ traverse un zéro de $p_n(\lambda)$.

14.4 Méthode de bissection.

Si l'on applique ce théorème à la suite (17.3), la diff érence $\omega(b)-\omega(a)$ est égale au nombre de valeurs propres de (17.2) dans l'intervalle $[a,b]$. On obtient toutes les valeurs propres de A de la manière suivante :

— on cherche un intervalle $[a,b]$ qui contienne toutes les valeurs propres de A (par exemple, en appliquant le théorème de Gershgorin). On a donc que $\omega(a) = 0$ et $\omega(b) = n$.

— on pose $c = \frac{(a+b)}{2}$ et on calcule $\omega(c)$. Les différences $\omega(c)-\omega(a)$ et $\omega(b)-\omega(c)$ indiquent combien de valeurs propres de A sont dans $[a,c)$ et combien sont dans $[c,b)$

— on continue à diviser les intervalles qui contiennent au moins une valeur propre de A.

On peut facilement modifier cet algorithme pour calculer la valeur propre la plus petite ou la $3^{ème}$ plus grande valeur propre, etc. Pour éviter un "overflow" dans le calcul de $p_n(\lambda)$ (si n et λ sont grands), il vaut mieux travailler avec

$$f_i(\lambda) = \frac{p_i(\lambda)}{p_{i-1}(\lambda)} \qquad i = 1,2,...,n$$

et utiliser le fait que

$$\omega(\lambda) = \text{nombre d'éléments négatifs parmi } \{f_1(\lambda), f_2(\lambda), ..., f_n(\lambda)\}$$

(attention : si $p_{i-1}(\lambda)$ est zéro, on pose $f_i(\lambda) = -\infty$; cette valeur compte pour un élément négatif). Pour une programmation de l'algorithme, on utilise la récurrence

$$\begin{aligned} f_1(\lambda) &= d_1 - \lambda \\ f_i(\lambda) &= d_i - \lambda - \begin{cases} e_i^2/f_{i-1}(\lambda) & \text{si} \quad f_{i-1}(\lambda) \neq 0 \\ |e_i|/eps & \text{si} \quad f_{i-1}(\lambda) = 0 \end{cases} \end{aligned}$$

La formule pour le cas $f_{i-1}(\lambda) \neq 0$ est une conséquence de (17.3). Si $f_{i-1}(\lambda) = 0$ (c'est-à-dire $p_{i-1}(\lambda) = 0$), on remplace cette valeur par $|e_i|.eps$. Ceci correspond à ajouter la perturbation $|e_i|.eps$ à d_{i-1}

15 L'itération orthogonale

Dans ce paragraphe, nous allons généraliser la méthode de la puissance afin de pouvoir calculer les deux (trois,) valeurs propres dominantes en même temps. Cette généralisation motivera l'ité ration QR qui constitue l'algorithme le plus important pour le calcul des valeurs propres d'une matrice.

15.1 Généralisation de la méthode de la puissance (pour calculer les deux valeurs propres dominantes).

Considérons une matrice A dont les valeurs propres satisfont

$$|\lambda_1| > |\lambda_2| > ... > |\lambda_n|. \tag{15.1}$$

La méthode de la puissance est basée sur l'itération $y_{k+1} = Ay_k$ et nous permet d'obtenir une approximation de λ_1 à l'aide du quotient de Rayleigh. Pour calculer (en même temps) la deuxième valeur propre λ_2, nous prenons deux vecteurs y_0 et z_0 satisfaisant $y_0^* z_0 = 0$ et nous considérons l'itération

$$\begin{aligned} y_{k+1} &= Ay_k \\ z_{k+1} &= Az_k - \beta_{k+1} y_{k+1} \end{aligned} \tag{15.2}$$

où β_{k+1} est déterminé par la condition $y_{k+1}^* z_{k+1} = 0$. Par induction, on voit que

$$\begin{aligned} y_k &= A^k y_0 \\ z_k &= A^k z_0 - \gamma_k y_k \end{aligned}$$

où γ_k est tel que

$$y_k^* z_k = 0 \tag{15.3}$$

Ceci signifie que le calcul de $\{z_k\}$ correspond à la m éthode de la puissance appliquée à z_0, combinée avec une orthogonalisation (projection de $A^k z_0$ sur le complément orthogonal de y_k). En exprimant les vecteurs initiaux dans la base de vecteurs propres $v_1, v_2, ..., v_n$ de la matrice A (on suppose $\|v_i\|_2 = 1$),

$$y_0 = \sum_{i=1}^{n} a_i v_i, \qquad z_0 = \sum_{i=1}^{n} b_i v_i, \tag{15.4}$$

les vecteurs y_k, z_k deviennent

$$y_k = \sum_{i=1}^{n} a_i \lambda_i^k v_i, \qquad z_0 = \sum_{i=1}^{n} (b_i - \gamma_k a_i)\, \lambda_i^k v_i,$$

Comme nous l'avons constaté précédement, pour $k \to \infty$, le terme $a_1 \lambda_1^k v_1$ est dominant dans y_k (si $a_1 \neq 0$) et on obtient une approximation du premier vecteur propre v_1 . Que peut-on dire pour la suite $\{z_k\}$? La condition (18.3) d'orthogonalité implique que

$$\sum_{i=1}^{n} \sum_{i=1}^{n} a_i \left(b_j - \gamma_k a_j\right) \bar{\lambda}_i^k \lambda_j^k v_i^* v_j = 0 \tag{15.5}$$

Cette relation définit γ_k. Comme le terme avec $i = j = 1$ est dominant, on voit que $\gamma_k \approx b_1/a_1$ Par la suite, nous allons supposer que $a_1 \neq 0$ et $a_1 b_2 - a_2 b_1 \neq 0$. En divisant (18.5) par $\bar{\lambda}_1^k$ on obtient

$$\bar{a}_1(b_1 - \gamma_k a_1)\lambda_1^k(1 + O(|\lambda_2/\lambda_1|^k)) = -\bar{a}_1(b_2 - \gamma_k a_2)\lambda_2^k(v_1^* v_2 + O(|\lambda_2/\lambda_1|^k) + O(|\lambda_3/\lambda_2|^k)).$$

Maintenant, on peut insérer cette formule dans (18.4) et on en d éduit

$$z_k = \lambda_{21}^k (b_2 - \gamma_k a_2)(v_2 - v_1^* v_2 . v_1 + O(|\lambda_2/\lambda_1|^k) + O(|\lambda_3/\lambda_2|^k)) \tag{15.6}$$

Visiblement, le vecteur z_k s'approche (pour $k \to \infty$) d'un multiple de $v_2 - v_1^* v_2 . v_1$, qui est la projection orthogonale de v_2 à l'hyperplan $v_1^{\perp}$. Concernant les valeurs propres, on a le résultat suivant.

Théorème 250. Considérons les vecteurs y_k, z_k donnés par (18.2) et notons

$$U_k = \left(y_k / \left\| y_k \right\|_2 , z_k / \| z_k \|_2 \right) \tag{15.7}$$

(observer que $U_k^* U_k = 1$). Si (18.1) est vérifié , on a que

$$U_k^* A U_k \longrightarrow \begin{pmatrix} \lambda_1 & * \\ 0 & \lambda_2 \end{pmatrix} \qquad \text{pour} \qquad k \longrightarrow \infty \tag{15.8}$$

Démonstration. L'élément (1,1) de la matrice $U_k^* A U_k$ est le quotient de Rayleigh (12.3) qui converge vers λ_1 . En utilisant (18.6), on voit que l'élément (2,2) satisfait

$$\frac{z_k^* A z_k}{z_k^* z_k} \longrightarrow \frac{\left(v_2 - v_1^* v_2 . v_1 \right)^* \left(\lambda_2 v_2 - \lambda_1 v_1^* v_2 . v_1 \right)}{\left(v_2 - v_1^* v_2 . v_1 \right)^* \left(v_2 - v_1^* v_2 . v_1 \right)} = \frac{\lambda_2 (1 - \left| v_1^* v_2 \right|^2)}{1 - \left| v_1^* v_2 \right|^2} = \lambda_2$$

De façon similaire, on obtient pour l'élément (2,1)

$$\frac{z_k^* A y_k}{\| z_k \|_2 \left\| y_k \right\|_2} \longrightarrow \frac{\left(v_2 - v_1^* v_2 . v_1 \right)^* \lambda_1 v_1}{\left\| v_2 - v_1^* v_2 . v_1 \right\|_2 \| v_1 \|_2} = 0$$

Finalement, l'élément (1,2) de $U_k^* A U_k$ satisfait

$$\frac{y_k^* A z_k}{\left\| y_k \right\|_2 \| z_k \|_2} \longrightarrow \frac{v_1^* (\lambda_2 v_2 - \lambda_1 v_1^* v_2 . v_1)}{\| v_1 \|_2 \left\| v_2 - v_1^* v_2 . v_1 \right\|_2} = \frac{(\lambda_2 - \lambda_1) v_1^* v_2}{\sqrt{1 - \left| v_1^* v_2 \right|^2}}.$$

Cette expression est en général non nulle. cqfd

Remarque 251. Avec la notation (18.7), l'itération (18.2) peut être écrite sous la forme

$$A U_k = U_{k+1} R_{k+1}$$

où R_{k+1} est une matrice 2×2 qui est triangulaire supé rieure.

15.2 Méthode de la puissance (pour le calcul de toutes les valeurs propres)

ou simplement *itération orthogonale*. La généralisation de l'algorithme précédent au cas où l'on veut calculer toutes les á valeurs propres d'une matrice est évidente : on choisit une matrice orthogonale U_0, c'est-à-dire, on choisit n vecteurs orthogonaux (les colonnes de U_0) qui jouent le rôle de y_0, z_0, etc. Puis, on effectue l'itération

```
for k=1,2,..
    Z_{k}=AU_{k+1}          (d\U{e9}composition QR)
    U_{k}R_{k}=Z_{k}
end
```

Si (18.1) est vérifié et si la matrice U_0 est bien choisie ($a_1 \neq 0, a_1 b_2 - a_2 b_1 \neq 0$, , etc), une géné ralisation du théorème précédent donne la convergence

$$T_k = U_k^* A U_k \tag{15.9}$$

vers une matrice triangulaire dont les éléments de la diagonale sont les valeurs propres de A . On a donc transformé A en forme triangulaire à l'aide d'une matrice orthogonale (*décomposition de Schur*). Il y a une possibilité intéressante pour calculer T_k de (18.9) directement à partir de T_{k-1}. D'une part, on déduit de (**??**) que

$$T_{k-1} = U_k^* A U_k = \left(U_{k-1}^* U_k \right) R_k \tag{15.10}$$

D'autre part, on a

$$T_k = U_k^* A U_k = U_k^* A U_{k-1}^* U_k = R_k \left(U_{k-1}^* U_k \right).$$

On calcule la décomposition QR de la matrice T_{k-1} et on é change les deux matrices de cette décomposition pour obtenir T_k

15.3 L' algorithme QR

La méthode QR, due à J.C.F. Francis et à V.N. Kublanovskaya, est la méthode la plus couram- ment utilisée pour le calcul de l'ensemble des valeurs propres (P.G. Ciarlet 1982) the QR iteration, and it forms the backbone of the most effective algorithm for computing the Schur decomposition. (G.H. Golub & C.F. van Loan 1989) La version simple du célèbre algorithme QR n'est rien d'autre que la méthode du paragraphe précédent. En effet, si l'on pose $Q_k = U_{k-1}^* U_k$ et si l'on commence l'itération avec $U_0 = I$, les formules (18.9) et (18.10) nous permettent d' écrire l'algorithme précédent comme suit : (décomposition QR)

```
T_{0}=A
for  k=1,2,..
    Q_{k}R_{k}=T_{k-1}
    T_{k}=R_{k}Q_{k}
end
```

Les T_k qui sont les mêmes que dans le paragraphe V.5, convergent (en général) vers une matrice triangulaire. Ceci nous permet d'obtenir toutes les valeurs propres de la matrice A car les T_k ont les mêmes valeurs propres que A (voir (18.9)). Cet algorithme important a été développé indépendamment par J.G.F. Francis (1961) et par V.N. Kublanovskaya (1961). Un algorithme similaire, qui utilise la décomposition LR à la place de la dé composition QR, a été introduit par H. Rutishauser (1958).

Exemple 252. Appliquons la méthode QR à la matrice

$$A = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 3 & 6 & 8 & 4 \\ 0 & 5 & 4 & 3 \\ 0 & 0 & 4 & 3 \end{pmatrix}$$

On peut montrer que, pour une matrice de Hessenberg A , toutes les matrices T_k sont aussi sous forme de Hessenberg. Pour 'étudier la convergence vers une matrice triangulaire, il suffit alors de considérer les éléments $t_{i+1,i}^{(k)} (i = 1, 2, ..., n-1)$ de la sous-diagonale. On constate que

$$\frac{t_{i+1,i}^{(k+1)}}{t_{i+1,i}^{(k)}} \approx \frac{\lambda_{i+1}}{\lambda_i} \tag{15.11}$$

$(\lambda_1 \approx 14,3, \lambda_2 \approx 7,86, \lambda_3 \approx 2,70, \lambda_4 \approx -1,86)$. Comme, les éléments $t_{i+1,i}^{(k)}$ convergent, pour $k \longrightarrow \infty$, liné airement vers 0 (voir la figure V.4, où les valeurs sont dessinées en fonction du nombre k de l'itération).

Remarque 253. **(a)** Comme le calcul de la décomposition QR d'une matrice pleine est très coûteux ($O(n^3)$ opérations), on applique l'algorithme QR uniquement aux matrices de Hessenberg. Dans cette situation une itération nécessite seulement $O(n^3)$ opérations.

(b) La convergence est très lente en général (seulement *linéaire*). Pour rendre efficace cet algorithme, il faut absolument trouver un moyen pour accélérer la convergence.

(c) Considérons la situation où A est une matrice ré elle qui possède des valeurs propres complexes (l'hypothèse (18.1) est violée). L'algorithme QR produit une suite de matrices T_k qui sont toutes réelles. Dans cette situation, les T_k ne convergent pas vers une matrice triangulaire, mais deviennent triangulaires par blocs (sans démonstration). Comme la dimension des blocs dans la diagonale vaut en général 1 ou 2, on obtient également des approximations des valeurs propres.

15.4 Accélération de la convergence

D'après l'observation (18.11), nous savons que

$$t_{n,n-1}^{(k)} = O(|\lambda_n/\lambda_{n-1}|^k)$$

La convergence vers zéro de cet élément ne va être rapide que si $|\lambda_n| \ll |\lambda_{n-1}|$. Une idée géniale est d'appliquer l'algorithme QR à la matrice $A - pI$ où $p \approx \lambda_n$ Comme les valeurs propres de $A - pI$ sont $\lambda_i - p$, on a la propriété $|\lambda_n - p| \ll |\lambda_i - p|$ pour $i = 1, ..., n-1$ et l'élément $t_{n,n-1}^{(k)}$ va converger rapidement vers zéro. Rien ne nous empêche d'amé liorer l'approximation p après chaque itération. L'algorithme QR avec "shift" devient alors :

```
T_{0}=A
for
   k=1,2,..
determiner le parametre p_{k-1}$
Q_{k}R_{k}=T_{k-1}-p_{k-1}I    (decomposition QR)
T_{k}=R_{k}Q_{k}+p_{k-1}
end
```

Les matrices T_k de cette itération satisfont

$$Q_k^* T_{k-1} Q_k = Q_k^*(Q_k R_k + p_{k-1} I)Q_k = R_k Q_k + p_{k-1} I = T_k$$

Ceci implique que, indépendamment de la suite p_k , les matrices T_k ont toutes les mêmes valeurs propres que $T_0 = A$. Pour décrire complètement l'algorithme QR avec shift, il faut encore discuter le choix du paramètre p_k et il faut donner un critère pour arrêter l'itération.

15.4.1 Choix du "shift"-paramètre.

On a plusieurs possibilités :

— $p_k = t_{n,n}^{(k)}$: ce choix marche très bien si les valeurs propres de la matrice sont réelles.

— on considère la matrice

$$\begin{pmatrix} t_{n-1,n-1}^{(k)} & t_{n-1,n}^{(k)} \\ t_{n,n-1}^{(k)} & t_{n,n}^{(k)} \end{pmatrix} \tag{15.12}$$

Si les valeurs propres de (18.12) sont réelles, on choisit pour p_k celle qui est la plus proche de $t_{n,n}^{(k)}$. Si elles sont de la forme $\alpha \pm i\beta$ avec $\beta \neq 0$ (donc complexes), on prend d'abord $p_k = \alpha + i\beta$ et pour l'itération suivante $p_{k+1} = \alpha - i\beta$

15.5 Critère pour arrêter l'itération.

L'idée est d'itérer jusqu'à ce que $t_{n,n-1}^{(k)}$ ou $t_{n-1,n-2}^{(k)}$ soit suffisamment petit. Plus précisément, on arr ête l'itération quand

$$t_{l,l-1}^{(k)} \leq eps.(\left|t_{l-1,l-1}^{(k)}\right| + \left|t_{l,l}^{(k)}\right|) \qquad \text{pour} \quad l = n \quad \text{ou} \quad l = n-1 \tag{15.13}$$

— Si (18.13) est vérifié pour $l = n$ on accepte $t_{n,n}^{(k)}$ comme approximation de λ_n et on continue l'ité ration avec la matrice $\left(t_{i,j}^{(k)}\right)_{1 \leq i,j \leq n-1}$

— Si (18.13) est vérifié pour $l = n-1$, on accepte les deux valeurs propres de (18.12) comme approximations de λ_n et λ_{n-1} et on continue l'itération avec la matrice $\left(t_{i,j}^{(k)}\right)_{1 \leq i,j \leq n-2}$

Exemple 254. Nous avons appliqué l'algorithme QR à la matrice (**??**) avec le shift $p_k = t_{n,n}^{(k)}$. La convergence de $t_{i+1,i}^{(k)}$ vers z éro est illustrée dans la figure V.5. Une comparaison avec la figure V.4 nous montre que la convergence est beaucoup plus rapide (convergence quadratique). Après 5 itérations, on a $\left|t_{4,3}^{(k)}\right| \leq 10^{-15}$. Encore 4 itérations pour la matrice de dimension 3 donnent $\left|t_{3,2}^{(k)}\right| \leq 10^{-15}$ Il ne reste plus que 3 itérations à faire pour la matrice de dimension 2 pour avoir $\left|t_{2,1}^{(k)}\right| \leq 10^{-15}$. En tout, 12 itérations ont donné toutes les valeurs propres avec une précision de 15 chiffres.

15.6 Le "double shift" de Francis

Dans la situation où A est une matrice réelle ayant des valeurs propres complexes, il est recommandé de choisir un shift-paramètre p_k qui soit complexe. Une application directe de l'algorithme préc édent nécessite un calcul avec des matrices complexes. L'observation suivante permet d'éviter ceci.

Lemme 255. Soit T_k une matrice réelle, $p_k = \alpha + i\beta$ et $p_{k+1} = \alpha - i\beta$.Alors, on peut choisir les décompositions dans l'algorithme QR de manière à ce que T_{k+2} soit réelle.

Remarque 256. La décomposition QR d'une matrice est unique sauf qu'on peut remplacer QR par $(QD)^{-1}\left(D^{-1}R\right)$ où $D = diag(d_1, ..., d_n)$ avec $|d_i| = 1$.

Démonstration. La formule (**??**) montre que

$$T_{k+2} = (Q_{k+1}Q_{k+2})^* T_k (Q_{k+1}Q_{k+2}) \tag{15.14}$$

cqfd

Il suffit alors de démontrer que le produit $Q_{k+1}Q_{k+2}$ est ré el. Une manipulation à l'aide de formules pour T_k donne

$$\begin{aligned} Q_{k+1}Q_{k+2}R_{k+2}R_{k+1} &= Q_{k+1}(T_{k+1} - p_{k+1}I)R_{k+1} = Q_{k+1}(R_{k+1}Q_{k+1} + p_{k+1}I - p_kI)R_{k+1} = \\ &= (Q_{k+1}R_{k+1})^2 + (p_k - p_{k+1})Q_{k+1}R_{k+1} = (T_k - p_kI)^2 + (p_k - p_{k+1})(T_k - p_kI) = \\ &= T_k^2 - (p_k + p_{k+1})T_k + p_kp_{k+1}I = M \end{aligned} \tag{15.15}$$

On a donc trouvé une décomposition QR de la matrice M qui, en cons équence des hypothèses du lemme, est une matrice réelle. Si, dans l'algorithme QR, la décomposition est choisie de manière à ce que les éléments diagonaux de R_{k+1} et R_{k+2} soient r éels, alors, à cause de l'unicité de la décomposition QR, les matrices $Q_{k+1}Q_{k+2}$ et $R_{k+2}R_{k+1}$ sont réelles. Une possibilité de calculer T_{k+2} à partir de T_k est de calculer de (18.15), de faire une décomposition QR (réelle) de M et de calculer T_{k+2} à l'aide de (18.14). Cet algorithme n'est pas pratique car le calcul de T_k^2 nécessite $O(n^3)$ op érations, même si T_k est sous forme de Hessenberg. Il y a une astuce intéressante pour obtenir T_{k+2} à partir de T_k en $O(n^2)$ opérations. Elle est basée sur la propriét é suivante.

Théorème 257. Soit une matrice donnée et supposons que

$$Q^*TQ = S \tag{15.16}$$

où Q est orthogonale et S est sous forme de Hessenberg satisfaisant $s_{i,i-1} \neq 0$ pour $i = 2, ..., n$ Alors, Q et S sont déterminé es de manière "unique" par la première colonne de Q.

Remarque 258. On a "unicité" dans le sens suivant : si $\hat{Q}^*T\hat{Q}$ est de type Hessenberg avec une matrice orthogonale $\hat{Q}$ satisfaisant $\hat{Q}e_1$, alors $\hat{Q} = QD$ où $D = diag(d_1, ..., d_n)$ avec $|d_i| = 1$.

Démonstration. Notons les colonnes de Q par q_i. Alors, la relation (18.16) implique

$$Tq_i = \sum_{j=1}^{i+1} s_{ji}q_j, \qquad q_j^*Tq_i = s_{ji}. \tag{15.17}$$

Si q_1 est fixé, la valeur s_{11} est donnée par la deuxiè me formule de (18.17). Avec cette valeur, on obtient de la premiè re formule de (18.17) que q_2 est un multiple de $Tq_1 - s_{11}q_1$. Ceci détermine q_2 à une unité prè s. Maintenant, les valeurs s_{21}, s_{12}, s_{22} sont déterminées et q_3 est un multiple de $Tq_2 - s_{21}q_1 - s_{22}q_2$ etc. cqfd

Si les hypothèses du lemme précédent sont vérifiées, on peut calculer la matrice réelle T_{k+2} en $O(n^2)$ opérations de la manière suivante :

— calculer $M\varepsilon_1$, la première colonne de M (formule (18.15)) ;
— déterminer une matrice de Householder H_1 telle que $H_1(M\varepsilon_1) = \alpha e_1$
— transformer $H_1^T T_k H_1$ sous forme de Hessenberg à l'aide de matrices de Householder $H_2, ..., H_{n-1}$ (voir le paragraphe V.3) ; c'est-à-dire., calculer $H^T T_k H$ où $H = H_1 H_2 .... H_{n-1}$.

Comme $H_i e_1 = e_1$ pour $i = 2, ..., n-1$, la première colonne de H est un multiple de celle de M (observer $H_1^T = H_1$). Par la formule (18.15), la première colonne de $Q_{k+1}Q_{k+2}$ est aussi un multiple de Me_1. Par conséquent, pour un bon choix des dé compositions $Q_{k+1}R_{k+1}$ et $Q_{k+2}R_{k+2}$ on a $H = Q_{k+1}Q_{k+2}$ la matrice obtenue par cet algorithme est égale à T_{k+2} (voir (18.14)).

15.7 Etude de la convergence

Supposons d'être déjà proche de la limite et considérons, par exemple, la matrice

$$T_0 = A = \begin{pmatrix} 2 & a \\ \varepsilon & 1 \end{pmatrix}$$

où ε est un nombre petit. Avec le choix $p_0 = 1$ pour le shift-paramètre, on obtient

$$T_0 - p_0 I = \begin{pmatrix} 1 & a \\ \varepsilon & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\varepsilon^2}} & -\frac{\varepsilon}{\sqrt{1+\varepsilon^2}} \\ \frac{\varepsilon}{\sqrt{1+\varepsilon^2}} & \frac{1}{\sqrt{1+\varepsilon^2}} \end{pmatrix} \begin{pmatrix} \sqrt{1+\varepsilon^2} & \frac{a}{\sqrt{1+\varepsilon^2}} \\ 0 & -\frac{a\varepsilon}{\sqrt{1+\varepsilon^2}} \end{pmatrix} = Q_1 R_1$$

et

$$T_0 - p_0 I = R_1 Q_1 = \begin{pmatrix} * & * \\ -\frac{a\varepsilon^2}{1+\varepsilon^2} & * \end{pmatrix}$$

— si A est symétrique (c'est-à-dire, $a = \varepsilon$) on a $t^{(1)}_{n,n-1} = O(\varepsilon^2)$, donc convergence *cubique*.
— si A n'est pas symétrique (p.ex. on a donc convergence quadratique.

Ces propriétés restent vraies pour des matrices générales (sans démonstration).

16 Exercices

Exercice 259. Calculer les valeurs propres de la matrice tridiagonale (dimension $n, b.c > 0$)

$$A = \begin{pmatrix} a & c & & & \\ b & a & c & & \\ & b & a & c & \\ & & b & \ddots & \ddots \\ & & & \ddots & \end{pmatrix}$$

Indication. Les composants du vecteur propre $(v_1, v_2, ..., v_n)^T$ satisfont une équation aux diffé rences finies avec.$v_0 = v_{n+1} = 0$ Vérifier que $v_j = Const.(\alpha_1^j - \alpha_2^j)$ où

$$\alpha_1 - \alpha_2 = \frac{\lambda - a}{c}, \qquad \alpha_1.\alpha_2 = \frac{b}{c}, \qquad \left(\frac{\alpha_1}{\alpha_2}\right)^{n+1} = 1$$

Résultat : $\lambda_j = a - 2\sqrt{bc}.\cos\left(\frac{j\pi}{n+1}\right), \qquad j = 1, 2, .., n.$

Exercice 260. Considérer la matrice

$$A(\varepsilon) = \begin{pmatrix} 1 & \varepsilon & 0 \\ -1 & 0 & 1 \\ 1 & -1+\varepsilon & -\varepsilon \end{pmatrix}$$

cette matrice possède une valeur propre de la forme

$$\lambda(\varepsilon) = i + \varepsilon.d + O(\varepsilon^2)$$

Calculer d et dessiner la tangente à la courbe $\lambda(\varepsilon)$ au point $\lambda(0)$

(a) Calculer par la méthode de la puissance, la plus grande valeur propre de la matrice

$$A = \begin{pmatrix} 99 & 1 & 0 \\ 1 & 100 & 1 \\ 0 & 1 & 98 \end{pmatrix}$$

(b) Pour accélerer considérablement la vitesse de convergence, appliquer la méthode de la puissance à la matrice $A - pI$ avec un choix intelligent de p.

(c) Avec quel choix de p obtient-on la valeur propre la plus petite ?

Exercice 261. Considérons la matrice tridiagonale

$$A = \begin{pmatrix} b_1 & c_1 & \\ a_1 & b_2 & c_2 \\ & a_2 & \\ & & \ddots \end{pmatrix}$$

Montrer que, si $a_i c_i > 0$ pour $i = 1, ..., n-1$, toutes les valeurs propres de A sont réelles. *Indication*. Trouver $D = diag(d_1, ..., d_n)$ telle que DAD^{-1} soit symétrique.

Exercice 262. Soit A une matrice symétrique et B quelconque. Montrer que pour chaque valeur propre λ_B de B il existe une valeur propre λ_A de A telle que

$$|\lambda_A - \lambda_B| \leq \|A - B\|_2 .$$

Indication. Montrer l'existence d'un vecteur v tel que $v = (A - \lambda_B)^{-1}(A - B)v$. En déduire que $1 \leq \left\|(A - \lambda_B)^{-1}(A - B)\right\| \leq \left\|(A - \lambda_B)^{-1}\right\| \|(A - B)\|$.

Exercice 263. (Schur, 1909). Soit A une matrice symétrique. Montrer que pour chaque indice i il existe une valeur propre λ de A telle que

$$|\lambda - a_{ii}| \leq \sqrt{\sum_{j \neq i} \left|a_{ij}\right|^2}$$

Indication. Appliquer l'exercice 5 avec une B convenable.

Exercice 264. Soit A une matrice réelle avec pour valeur propre $\alpha + i\beta$. Montrer que l'itération

$$\begin{pmatrix} \bar{\alpha} I - A & -\bar{\beta} I \\ \bar{\beta} I & \bar{\alpha} I - A \end{pmatrix} \begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix}$$

où $\bar{\alpha} \approx \alpha$ et $\bar{\beta} \approx \beta$) permet de calculer la valeur propre $\alpha + i\beta$ et le vecteur propre correspondant. Indication. Considérer les parties réelles et complexes de l'itération de Wielandt. On obtient alors

$$\frac{u_k^T A u_k + v_k^T A v_k}{u_k^T u_k + v_k^T v_k} \longrightarrow \alpha, \qquad \frac{u_k^T A v_k + v_k^T A u_k}{u_k^T u_k + v_k^T v_k} \longrightarrow \beta$$

Exercice 265. Considérons la matrice de Hilbert,

$$A = \begin{pmatrix} 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \\ 1/4 & 1/5 & 1/6 \end{pmatrix}$$

(a) Transformer A en une matrice tridiagonale ayant les mêmes valeurs propres.

(b) En utilisant une suite de Sturm, montrer que toutes les valeurs propres sont positives et qu'une valeur propre est plus petite que 0.001

(c) Calculer approximativement la condition de A pour la norme Euclidienne.

Exercice 266. La formule de récurrence

$$(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x)$$

pour les *polynômes de Legendre* ressemble à

$$p_i(\lambda) = (d_i - \lambda)p_{i-1}(\lambda) - \varepsilon_i^2 p_{i-2}(\lambda), \qquad i = 2, \ldots, n.$$

pour les polynômes $\det(A_i - \lambda I)$. Trouver une matrice tridiagonale A de dimension n telle que les valeurs propres de A sont les racines de $P_n(x)$.

Exercice 267. Soit $p(x)$ un polynôme de degré n et supposons que toutes les racines soient simples. Démontrer que la suite définie par l'algorithme d'Euclide,

$$\begin{aligned} p_n(x) &= p(x), \qquad p_{n-1}(x) = -p'(x) \\ p_i(x) &= q_i(x)p_{i-1}(x) - \gamma_i^2 p_{i-2}(x), \qquad i = n, \ldots, 2. \end{aligned}$$

est une suite de Sturm. Pour le polynôme $p(x) = x^5 - 6x^4 + 3x^3 + 3x^2 + 2x + 8$.

(a) déterminer le nombre de racines réelles.

(b) Combien de racines sont complexes ?

(c) Combien de racines sont réelles et positives ?

Exercice 268. Pour un φ donné notons $c = \cos\varphi$ et $s = \sin\varphi$. La matrice Ω_{kl}, définie par

$$(\Omega_{kl})_{ij} = \begin{cases} 1 & \text{si } i = j, j \neq k, j \neq l \\ c & \text{si } i = j = k, \text{ ou } i = j = l \\ s & \text{si } i = k, \text{ et } j = l \\ -s & \text{si } i = l, \text{ et } j = k \\ 0 & \text{sinon} \end{cases}$$

s'appelle rotation de Givens.

(a) Montrer qu'elle est orthogonale.

(b) Soit A une matrice symétrique.Déterminer φ tel que le (k,l)-ième élément de $A' = \Omega_{kl} A \Omega_{kl}^T$ s'annule.

Resultat. $\cot 2\varphi = (a_{kk} - a_{ll})/(2a_{kl})$.

Exercice 269. La méthode de Jacobi (1846) pour le calcul des valeurs propres d'une matrice symétrique :

i) on choisit a_{kl} $(k > l)$ tel que $|a_{kl}| = \max_{i>j} |a_{ij}|$;

ii) on détermine A' comme dans l'exercice 11.

Montrer que, si on répète cette procédure, on a convergence vers une matrice diagonale, dont les éléments sont les valeurs propres de A Indication. Montrer que $\sum_{i>j} |a'_{ij}|^2 = \sum_{i>j} |a_{ij}|^2 - |a_{kl}|^2$

Exercice 270. On considère la matrice

$$A = \begin{pmatrix} 7 & 0,5 \\ 0,0001 & 8 \end{pmatrix}$$

dont on cherche à calculer les valeurs propres.

(a) Faire une itération de l'algorithme QR sans shift.

(b) Faire une itération de l'algorithme QR avec shift.

(c) Estimer la position des valeurs propres de A à l'aide du Th éorème de Gershgorin.

(d) Calculer les valeurs propres de A à l'aide du polynôme caractéristique.

Exercice 271. Montrer que si la matrice $T_0 = A$ est une matrice de Hessenberg (ou tridiagonale), alors les matrices T_k, $k \geq 1$ construites par l'algorithme QR sont également des matrices de Hessenberg (tridiagonales).

Exercice 272. Donner une estimation grossière du nombre d'opérations qui sont n écessaires pour effectuer la décomposition QR d'une matrice de Hessenberg et pour calculer ensuite le produit RQ.

Exercice 273. Soit T_0 une matrice de Hessenberg dont tous les éléments de la sous-diagonale sont non-nuls. Montrer que, si p_0 est une valeur propre de T_0, une itération de l'algorithme QR avec shift p_0 donne

$$t^{(1)}_{n,n-1} = 0.$$

Exercice 274. Expliquer, comment le calcul de T_k à partir de T_{k-1}

$$Q_k R_k = T_{k-1} - p_{k-1} I, \qquad T_k = R_k Q_k + p_{k-1} I.$$

peut être effectué sans soustraire (et additionner) explicitement la matrice $p_{k-1} I$ Indication. Laissez-vous inspirer par le "double shift" algorithme de Francis.

17 TRANSFORMATION SOUS FORME TRIDIAGONALE (ou de HESSENBERG)

Avec la transformation $v = Pu$ (où est P une matrice inversible) le problème

$$Av = \lambda v$$

devient

$$P^{-1}APu = \lambda u$$

Donc, les valeurs propres de A et de $P^{-1}AP$ sont les mêmes et les vecteurs propres v_i de A se transforment par $v_i = Pu_i$. Le but de ce paragraphe est de trouver une matrice P telle que $P^{-1}AP$ devienne "plus simple". La situation idéale serait trouvée si $P^{-1}AP$ devenait diagonale ou triangulaire - mais une telle transformation nécessiterait déjà la connaissance des valeurs propres.

Alors, on cherche P tel que $P^{-1}AP$ soit sous forme de Hessenberg

$$P^{-1}AP = H = \begin{pmatrix} * & * & \cdots & \cdots & * \\ * & * & \ddots & & \vdots \\ & * & \ddots & \ddots & * \\ & & \ddots & \ddots & * \\ & & & * & * \end{pmatrix} \tag{17.1}$$

c'est-à-dire, $h_{ij} = 0$ pour $i > j + 1$. Pour arriver à ce but, nous considérons deux algorithmes.

17.1 a) A l'aide des transformations élémentaires

Comme pour l'élimination de Gauss, nous utilisons les transformations pour faire apparaitre les zéros - colonne par colonne - dans (17.1).

Dans un premier pas, nous choisissons $k \geq 2$ tel que $|a_{k1}| \geq |a_{j1}|$ pour $j \geq 2$ et nous permutons les lignes 2 et k, c'est-à-dire, nous formons PA où P est une matrice de permutation convenable. Pour ne pas changer les valeurs propres, il faut également permuter les colonnes 2 et k (ceci correspond au calcul de $A' = PAP^{-1}$ car $P^2 = I$ (et donc $P = P^{-1}$). Si $a'_{21} = 0$, on a aussi $a'_{i1} = 0$ pour $i \geq 3$ et le premier pas est terminé. Sinon, nous déterminons

$$L_2 = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & -l_{n2} & \cdots & 0 & 1 \end{pmatrix} \quad \text{telle que} \quad L_2 A' = \begin{pmatrix} a'_{11} & a'_{12} & \cdots & a'_{1n} \\ a'_{21} & a'_{22} & \cdots & a'_{2n} \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}$$

Pour ceci, on définit $l_{i2} = \frac{a'_{i1}}{a'_{21}}$. Une multiplication à droite avec

$$L_2^{-1} = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & l_{n2} & \cdots & 0 & 1 \end{pmatrix}$$

ne change pas la première colonne de $L_2 A'$.

On répète la même procédure avec la sous-matrice de $L_2 A' L_2^{-1}$ de dimension $n-1$, et ainsi de suite. A cause des multiplications à droite avec L_i^{-1}, cet algorithme coûte deux fois plus cher que l'élimination de Gauss.

Pour la matrice

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 1 \end{pmatrix}$$

on prend

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/2 & 1 \end{pmatrix}$$

et on obtient

$$L_2 A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 0 & 5/2 & -1/2 \end{pmatrix}, \quad \text{puis} \quad L_2 A L_2^{-1} = \begin{pmatrix} 3 & 5/2 & 1 \\ 2 & 5/2 & 3 \\ 0 & 9/4 & -1/2 \end{pmatrix} = H$$

Cet exemple montre un désavantage de cet algorithme : si l'on part avec une matrice symétrique A, la matrice de Hessenberg H, obtenue par cet algorithme, n'est plus symétrique en général.

17.2 b) A l'aide des transformations orthogonales

Il est souvent préférable de travailler avec des réflexions de Householder Commençons par une réflexion pour les coordonnées $2,...,n$ laissant fixe la première coordonnée : $\bar{Q}_2 = I - 2\bar{u}_2\bar{u}_2^T$ ($\|\bar{u}_2\|_2 = 1$) tel que $\bar{Q}_2\bar{A}_1 = \alpha_2 e_1$ où $\bar{A}_1 = (a_{21},...,a_{n1})$. En posant $u_2 = (0,\bar{u}_2)^T$ et $Q_2 = I - 2u_2u_2^T$, la matrice Q_2A contient des zéros dans la première colonne à partir du troisième élément. La multiplication à droite avec $Q_2^{-1} = Q_2^T = Q_2$ ne change pas cette colonne :

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \xrightarrow{Q_2A} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ \alpha_2 & * & * \\ 0 & * & * \end{pmatrix} \xrightarrow{Q_2AQ_2} \begin{pmatrix} a_{11} & * & * \\ \alpha_2 & * & * \\ 0 & * & * \end{pmatrix}$$

Dans le pas suivant, on applique la même procédure à la sous-matrice de dimension $n-1$, etc. Finalement, on arrive à la forme de Hessenberg (17.1) avec la tranformation $P^{-1} = Q_{n-1}....Q_2$ qui est une matrice orthogonale (c'est-à-dire, $P^{-1} = P^T$).

Nous avons un double avantage avec cet algorithme :

— il ne faut pas faire une recherche de pivot ;

— si A est symétrique, alors $P^{-1}AP$ est aussi symétrique, et donc tridiagonale.

17.3 Méthode de bissection pour des matrices tridiagonales

Considérons une matrice symétrique tridiagonale

$$A = \begin{pmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & e_3 & & \\ & e_3 & \ddots & \ddots & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n \end{pmatrix}$$

On observe tout d'abord que si un élément e_i est nul, la matrice A est déjà décomposée en deux sous-matrices du même type, qui ensemble fournissent les valeurs propres de A .

On peut donc supposer, sans restreindre la généralité, que

$$c_i \neq 0 \quad \text{pour} \quad i = 2,...,n. \tag{17.2}$$

Pour cette matrice, il est possible de calculer la valeur $P_A(\lambda)$ du polynôme caractéristique sans connaitre ses coefficients. En effet, si l'on pose

$$A_1 = (d_1), \qquad A_2 = \begin{pmatrix} d_1 & e_2 \\ e_2 & d_2 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} d_1 & e_2 & \\ e_2 & d_2 & e_3 \\ & e_3 & d_3 \end{pmatrix}, \qquad ...$$

et si l'on définit

$$p_i(\lambda) = \det(A_i - \lambda I),$$

on obtient

$$\begin{aligned} p_0(\lambda) &= 1 \\ p_1(\lambda) &= d_1 - \lambda \\ p_i(\lambda) &= (d_i - \lambda)\, p_{i-1}(\lambda) - e_i^2 p_{i-2}(\lambda), \qquad i = 2,...,n. \end{aligned} \tag{17.3}$$

La formule de récurrence dans (17.3) est obtenue en développant le déterminant de la matrice $A_i - \lambda I$ par rapport à la dernière ligne (ou colonne).

En principe, on peut maintenant calculer les valeurs propres de A (c'est-à-dire. les zéros de $p_n(\lambda)$ de la manière suivante : chercher un intervalle où $p_n(\lambda)$ change de signe et localiser une racine de $p_n(\lambda) = 0$ par bissection. Les évaluations de $p_n(\lambda)$ sont faites à l'aide de la formule (17.3). Mais il existe une astuce interessante qui permet d'améliorer cet algorithme.

Théorème 275. Si l'équation (17.2) est vérifié, les polynômes $p_i(\lambda)$ définis par (17.3) satisfont

a) $p_n'(\hat{\lambda})p_{n-1}(\hat{\lambda}) < 0$ si $p_n(\hat{\lambda}) = 0$ $(\hat{\lambda} \in \mathbb{R})$

b) $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) < 0$ si $p_i(\hat{\lambda}) = 0$ pour un $\{i \in 1,2,...,n-1\}$

c) $p_0(\lambda)$ ne change pas de signe sur $\mathbb{R}$.

Démonstration. L'affirmation (c) est triviale.

Si $p_i(\hat{\lambda}) = 0$ pour un $\{i \in 1,2,...,n-1\}$, la formule de récurrence (17.3) donne l'inégalité $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) \leq 0$. Pour démontrer (b), il suffit d'exclure le cas $p_{i-1}(\hat{\lambda})p_{i+1}(\hat{\lambda}) = 0$. Si deux valeurs consécutives

de la suite $\left\{p_i(\hat{\lambda})\right\}$ sont nulles, la formule de récurrence montre que $p_i(\hat{\lambda}) = 0$ pour tout i, ce qui contredit $p_0(\lambda) = 1$.

Nous démontrons par récurrence que toutes les racines de $p_i(\lambda)$ sont réelles, simples et séparées par celles de $p_{i-1}(\lambda)$.

Il n'y a rien à démontrer pour $i = 1$. Supposons la propriété vraie pour i et montrons qu'elle est encore vraie pour $i+1$. Comme les zéros $\lambda_1 < \lambda_2 < ... < \lambda_i$ sont séparés par ceux de $p_{i-1}(\lambda)$ et comme $p_{i-1}(-\infty) = +\infty$, nous avons $sign\ p_{i-1}(\lambda_j) = (-1)^{j+1}$. Alors, on déduit de (b) que $sign\ p_{i+1}(\lambda_j) = (-1)^j$. Ceci et le fait que $p_{i+1}(\lambda) = (-1)^{j+1}\lambda^{i+1} + ...$ montrent que $p_{i+1}(\lambda)$ possède un zéro réel dans chacun des intervalles ouverts $(-\infty, \lambda_1), (\lambda_1, \lambda_2), ..., (\lambda_i, \infty)$

L'affirmation (a) est maintenant une conséquence de (b) et du fait que toutes les racines de $p_{i-1}(\lambda)$ sont réelles simples ; cqfd

Définition 276 (suite de Sturm). Une suite $\{p_0, p_1, ..., p_n\}$ de polynômes à coefficients réels s'appelle une suite de Sturm, si elle vérifie les conditions (a), (b), (c) du Théorème (275)

Considérons une suite de Sturm $\{p_0, p_1, ..., p_n\}$. Si l'on définit

$$\omega(\lambda) = \text{nombre de changements de signes de } \{p_0(\lambda), p_1(\lambda), ..., p_n(\lambda)\}$$

alors le polynôme $p_n(\lambda)$ possède exactement

$$\omega(b) - \omega(a)$$

zéros dans l'intervalle $[a, b]$ (si $p_i(\lambda) = 0$, on définit $signp_i(\lambda) = sign\ p_{i-1}(\lambda)$).

Démonstration. Par continuité, l'entier $\omega(\lambda)$ peut changer sa valeur seulement si une valeur des fonctions $p_i(\lambda)$ devient nulle. La fonction $p_0(\lambda)$ ne change pas de signe. Supposons alors que $p_i(\tilde{\lambda}) = 0$ pour un $i \in \{1, 2, ..., n-1\}$. La condition (b) et la continuité de $p_j(\lambda)$ montrent que seulement les deux situations suivantes sont possibles (ε petit) :

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{i-1}(\lambda)$	+	+	+
$p_i(\lambda)$	±	0	±
$p_{i+1}(\lambda)$	−	−	−

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{i-1}(\lambda)$	−	−	−
$p_i(\lambda)$	±	0	±
$p_{i+1}(\lambda)$	+	+	+

cqfd

Chaque fois, on a $\omega(\tilde{\lambda}+\varepsilon) = \omega(\tilde{\lambda}) = \omega(\tilde{\lambda}-\varepsilon)$ et la valeur de $\omega(\lambda)$ ne change pas si λ traverse un zéro de $p_i(\lambda)$ pour $i \in \{1, 2, ..., n-1\}$

Il reste à étudier la fonction $\omega(\lambda)$ dans un voisinage d'un zéro $\tilde{\lambda}$ de $p_n(\lambda)$. La propriété (a) implique que pour les signes de $p_j(\lambda)$ on a seulement les deux possibilités suivantes :

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{n-1}(\lambda)$	+	+	+
$p_n(\lambda)$	+	0	−

	$\tilde{\lambda}-\varepsilon$	$\tilde{\lambda}$	$\tilde{\lambda}+\varepsilon$
$p_{n-1}(\lambda)$	−	−	−
$p_n(\lambda)$	−	0	+

c'est-à-dire, $\omega(\tilde{\lambda}+\varepsilon) = \omega(\tilde{\lambda}-\varepsilon)+1$. Ceci démontre que la fonction $\omega(\lambda)$ est constante par morceaux et augmente de 1 sa valeur si λ traverse un zéro de $p_n(\lambda)$.

17.4 Méthode de bissection.

Si l'on applique ce théorème à la suite (17.3), la différence $\omega(b) - \omega(a)$ est égale au nombre de valeurs propres de (17.2) dans l'intervalle $[a, b]$. On obtient toutes les valeurs propres de A de la manière suivante :

— on cherche un intervalle $[a, b]$ qui contienne toutes les valeurs propres de A (par exemple, en appliquant le théorème de Gershgorin). On a donc que $\omega(a) = 0$ et $\omega(b) = n$.

— on pose $c = \frac{(a+b)}{2}$ et on calcule $\omega(c)$. Les différences $\omega(c) - \omega(a)$ et $\omega(b) - \omega(c)$ indiquent combien de valeurs propres de A sont dans $[a, c)$ et combien sont dans $[c, b)$

— on continue à diviser les intervalles qui contiennent au moins une valeur propre de A.

On peut facilement modifier cet algorithme pour calculer la valeur propre la plus petite ou la $3^{ème}$ plus grande valeur propre, etc.

Pour éviter un "overflow" dans le calcul de $p_n(\lambda)$ (si n et λ sont grands), il vaut mieux travailler avec

$$f_i(\lambda) = \frac{p_i(\lambda)}{p_{i-1}(\lambda)} \qquad i = 1, 2, ..., n$$

et utiliser le fait que

$$\omega(\lambda) = \text{nombre d'éléments négatifs parmi } \{f_1(\lambda), f_2(\lambda), ..., f_n(\lambda)\}$$

(attention : si $p_{i-1}(\lambda)$ est zéro, on pose $f_i(\lambda) = -\infty$; cette valeur compte pour un élément négatif).

Pour une programmation de l'algorithme, on utilise la récurrence

$$\begin{aligned} f_1(\lambda) &= d_1 - \lambda \\ f_i(\lambda) &= d_i - \lambda - \begin{cases} e_i^2/f_{i-1}(\lambda) & \text{si} \quad f_{i-1}(\lambda) \neq 0 \\ |e_i|/eps & \text{si} \quad f_{i-1}(\lambda) = 0 \end{cases} \end{aligned}$$

La formule pour le cas $f_{i-1}(\lambda) \neq 0$ est une conséquence de (17.3). Si $f_{i-1}(\lambda) = 0$ (c'est-à-dire $p_{i-1}(\lambda) = 0$), on remplace cette valeur par $|e_i|.eps$. Ceci correspond à ajouter la perturbation $|e_i|.eps$ à d_{i-1}

18 L'ITERATION ORTHOGONALE

Dans ce paragraphe, nous allons généraliser la méthode de la puissance afin de pouvoir calculer les deux (trois,) valeurs propres dominantes en même temps. Cette généralisation motivera l'itération QR qui constitue l'algorithme le plus important pour le calcul des valeurs propres d'une matrice.

18.1 Généralisation de la méthode de la puissance (pour calculer les deux valeurs propres dominantes).

Considérons une matrice A dont les valeurs propres satisfont

$$|\lambda_1| > |\lambda_2| > ... > |\lambda_n|. \tag{18.1}$$

La méthode de la puissance est basée sur l'itération $y_{k+1} = Ay_k$ et nous permet d'obtenir une approximation de λ_1 à l'aide du quotient de Rayleigh. Pour calculer (en même temps) la deuxième valeur propre λ_2, nous prenons deux vecteurs y_0 et z_0 satisfaisant $y_0^* z_0 = 0$ et nous considérons l'itération

$$\begin{aligned} y_{k+1} &= Ay_k \\ z_{k+1} &= Az_k - \beta_{k+1} y_{k+1} \end{aligned} \tag{18.2}$$

où β_{k+1} est déterminé par la condition $y_{k+1}^* z_{k+1} = 0$. Par induction, on voit que

$$\begin{aligned} y_k &= A^k y_0 \\ z_k &= A^k z_0 - \gamma_k y_k \end{aligned}$$

où γ_k est tel que

$$y_k^* z_k = 0 \tag{18.3}$$

Ceci signifie que le calcul de $\{z_k\}$ correspond à la méthode de la puissance appliquée à z_0, combinée avec une orthogonalisation (projection de $A^k z_0$ sur le complément orthogonal de y_k).

En exprimant les vecteurs initiaux dans la base de vecteurs propres $v_1, v_2, ..., v_n$ de la matrice A (on suppose $\|v_i\|_2 = 1$),

$$y_0 = \sum_{i=1}^{n} a_i v_i, \qquad z_0 = \sum_{i=1}^{n} b_i v_i, \tag{18.4}$$

les vecteurs y_k, z_k deviennent

$$y_k = \sum_{i=1}^{n} a_i \lambda_i^k v_i, \qquad z_0 = \sum_{i=1}^{n} (b_i - \gamma_k a_i)\, \lambda_i^k v_i,$$

Comme nous l'avons constaté précédement, pour $k \to \infty$, le terme $a_1 \lambda_1^k v_1$ est dominant dans y_k (si $a_1 \neq 0$) et on obtient une approximation du premier vecteur propre v_1 . Que peut-on dire pour la suite $\{z_k\}$?

La condition (18.3) d'orthogonalité implique que

$$\sum_{i=1}^{n} \sum_{i=1}^{n} a_i \left(b_j - \gamma_k a_j\right) \bar{\lambda}_i^k \lambda_j^k v_i^* v_j = 0 \tag{18.5}$$

Cette relation définit γ_k. Comme le terme avec $i = j = 1$ est dominant, on voit que $\gamma_k \approx b_1/a_1$

Par la suite, nous allons supposer que $a_1 \neq 0$ et $a_1 b_2 - a_2 b_1 \neq 0$. En divisant (18.5) par $\bar{\lambda}_1^k$ on obtient

$$\bar{a}_1(b_1 - \gamma_k a_1)\lambda_1^k(1 + O(|\lambda_2/\lambda_1|^k)) = -\bar{a}_1(b_2 - \gamma_k a_2)\lambda_2^k(v_1^* v_2 + O(|\lambda_2/\lambda_1|^k) + O(|\lambda_3/\lambda_2|^k)).$$

Maintenant, on peut insérer cette formule dans (18.4) et on en déduit

$$z_k = \lambda_{21}^k (b_2 - \gamma_k a_2)(v_2 - v_1^* v_2 . v_1 + O(|\lambda_2/\lambda_1|^k) + O(|\lambda_3/\lambda_2|^k)) \tag{18.6}$$

Visiblement, le vecteur z_k s'approche (pour $k \to \infty$) d'un multiple de $v_2 - v_1^* v_2 . v_1$, qui est la projection orthogonale de v_2 à l'hyperplan $v_1^{\perp}$. Concernant les valeurs propres, on a le résultat suivant.

Théorème 277. Considérons les vecteurs y_k, z_k donnés par (18.2) et notons

$$U_k = \left(y_k / \left\|y_k\right\|_2, z_k / \|z_k\|_2\right) \tag{18.7}$$

(observer que $U_k^* U_k = 1$). Si (18.1) est vérifié, on a que

$$U_k^* A U_k \longrightarrow \begin{pmatrix} \lambda_1 & * \\ 0 & \lambda_2 \end{pmatrix} \qquad \text{pour} \qquad k \longrightarrow \infty \tag{18.8}$$

Démonstration. L'élément (1,1) de la matrice $U_k^* A U_k$ est le quotient de Rayleigh (12.3) qui converge vers λ_1 . En utilisant (18.6), on voit que l'élément (2,2) satisfait

$$\frac{z_k^* A z_k}{z_k^* z_k} \longrightarrow \frac{\left(v_2 - v_1^* v_2 . v_1\right)^* (\lambda_2 v_2 - \lambda_1 v_1^* v_2 . v_1)}{\left(v_2 - v_1^* v_2 . v_1\right)^* (v_2 - v_1^* v_2 . v_1)} = \frac{\lambda_2 (1 - \left|v_1^* v_2\right|^2)}{1 - \left|v_1^* v_2\right|^2} = \lambda_2$$

De façon similaire, on obtient pour l'élément (2,1)

$$\frac{z_k^* A y_k}{\|z_k\|_2 \left\|y_k\right\|_2} \longrightarrow \frac{\left(v_2 - v_1^* v_2 . v_1\right)^* \lambda_1 v_1}{\left\|v_2 - v_1^* v_2 . v_1\right\|_2 \|v_1\|_2} = 0$$

Finalement, l'élément (1,2) de $U_k^* A U_k$ satisfait

$$\frac{y_k^* A z_k}{\|y_k\|_2 \|z_k\|_2} \longrightarrow \frac{v_1^*(\lambda_2 v_2 - \lambda_1 v_1^* v_2 . v_1)}{\|v_1\|_2 \|v_2 - v_1^* v_2 . v_1\|_2} = \frac{(\lambda_2 - \lambda_1) v_1^* v_2}{\sqrt{1 - |v_1^* v_2|^2}}.$$

Cette expression est en général non nulle. cqfd

Remarque 278. Avec la notation (18.7), l'itération (18.2) peut être écrite sous la forme

$$AU_k = U_{k+1} R_{k+1}$$

où R_{k+1} est une matrice 2×2 qui est triangulaire supérieure.

18.2 Méthode de la puissance (pour le calcul de toutes les valeurs propres)

ou simplement *itération orthogonale*. La généralisation de l'algorithme précédent au cas où l'on veut calculer toutes les á valeurs propres d'une matrice est évidente : on choisit une matrice orthogonale U_0, c'est-à-dire, on choisit n vecteurs orthogonaux (les colonnes de U_0) qui jouent le rôle de y_0, z_0, etc. Puis, on effectue l'itération

```
for k=1,2,..
    Z_{k}=AU_{k+1}          (d\U{e9}composition QR)
    U_{k}R_{k}=Z_{k}
end
```

Si (18.1) est vérifié et si la matrice U_0 est bien choisie ($a_1 \neq 0, a_1 b_2 - a_2 b_1 \neq 0$, , etc), une généralisation du théorème précédent donne la convergence

$$T_k = U_k^* A U_k \tag{18.9}$$

vers une matrice triangulaire dont les éléments de la diagonale sont les valeurs propres de A . On a donc transformé A en forme triangulaire à l'aide d'une matrice orthogonale (*décomposition de Schur*).

Il y a une possibilité intéressante pour calculer T_k de (18.9) directement à partir de T_{k-1}. D'une part, on déduit de (**??**) que

$$T_{k-1} = U_k^* A U_k = \left(U_{k-1}^* U_k\right) R_k \tag{18.10}$$

D'autre part, on a

$$T_k = U_k^* A U_k = U_k^* A U_{k-1}^* U_k = R_k \left(U_{k-1}^* U_k\right).$$

On calcule la décomposition QR de la matrice T_{k-1} et on échange les deux matrices de cette décomposition pour obtenir T_k

18.3 L'algorithme QR

La méthode QR, due à J.C.F. Francis et à V.N. Kublanovskaya, est la méthode la plus couram- ment utilisée pour le calcul de l'ensemble des valeurs propres (P.G. Ciarlet 1982)

the QR iteration, and it forms the backbone of the most effective algorithm for computing the Schur decomposition. (G.H. Golub & C.F. van Loan 1989)

La version simple du célèbre algorithme QR n'est rien d'autre que la méthode du paragraphe précédent. En effet, si l'on pose $Q_k = U_{k-1}^* U_k$ et si l'on commence l'itération avec $U_0 = I$, les formules (18.9) et (18.10) nous permettent d'écrire l'algorithme précédent comme suit : (décomposition QR)

```
T_{0}=A
for  k=1,2,..
    Q_{k}R_{k}=T_{k-1}
    T_{k}=R_{k}Q_{k}
end
```

Les T_k qui sont les mêmes que dans le paragraphe V.5, convergent (en général) vers une matrice triangulaire. Ceci nous permet d'obtenir toutes les valeurs propres de la matrice A car les T_k ont les mêmes valeurs propres que A (voir (18.9)).

Cet algorithme important a été développé indépendamment par J.G.F. Francis (1961) et par V.N. Kublanovskaya (1961). Un algorithme similaire, qui utilise la décomposition LR à la place de la décomposition QR, a été introduit par H. Rutishauser (1958).

Exemple 279. Appliquons la méthode QR à la matrice

$$A = \begin{pmatrix} 10 & 2 & 3 & 5 \\ 3 & 6 & 8 & 4 \\ 0 & 5 & 4 & 3 \\ 0 & 0 & 4 & 3 \end{pmatrix}$$

On peut montrer que, pour une matrice de Hessenberg A , toutes les matrices T_k sont aussi sous forme de Hessenberg. Pour 'étudier la convergence vers une matrice triangulaire, il suffit alors de considérer les éléments $t^{(k)}_{i+1,i} (i = 1, 2, ..., n-1)$ de la sous-diagonale. On constate que

$$\frac{t^{(k+1)}_{i+1,i}}{t^{(k)}_{i+1,i}} \approx \frac{\lambda_{i+1}}{\lambda_i} \tag{18.11}$$

$(\lambda_1 \approx 14,3, \lambda_2 \approx 7,86, \lambda_3 \approx 2,70, \lambda_4 \approx -1,86)$. Comme, les éléments $t^{(k)}_{i+1,i}$ convergent, pour $k \longrightarrow \infty$, linéairement vers 0 (voir la figure V.4, où les valeurs sont dessinées en fonction du nombre k de l'itération).

Remarque 280.
- **(a)** Comme le calcul de la décomposition QR d'une matrice pleine est très coûteux ($O(n^3)$ opérations), on applique l'algorithme QR uniquement aux matrices de Hessenberg. Dans cette situation une itération nécessite seulement $O(n^3)$ opérations.
- **(b)** La convergence est très lente en général (seulement *linéaire*). Pour rendre efficace cet algorithme, il faut absolument trouver un moyen pour accélérer la convergence.
- **(c)** Considérons la situation où A est une matrice réelle qui possède des valeurs propres complexes (l'hypothèse (18.1) est violée). L'algorithme QR produit une suite de matrices T_k qui sont toutes réelles. Dans cette situation, les T_k ne convergent pas vers une matrice triangulaire, mais deviennent triangulaires par blocs (sans démonstration). Comme la dimension des blocs dans la diagonale vaut en général 1 ou 2, on obtient également des approximations des valeurs propres.

18.4 Accélération de la convergence

D'après l'observation (18.11), nous savons que

$$t^{(k)}_{n,n-1} = O(|\lambda_n/\lambda_{n-1}|^k)$$

La convergence vers zéro de cet élément ne va être rapide que si $|\lambda_n| \ll |\lambda_{n-1}|$. Une idée géniale est d'appliquer l'algorithme QR à la matrice $A - pI$ où $p \approx \lambda_n$ Comme les valeurs propres de $A - pI$ sont $\lambda_i - p$, on a la propriété $|\lambda_n - p| \ll |\lambda_i - p|$ pour $i = 1, ..., n-1$ et l'élément $t^{(k)}_{n,n-1}$ va converger rapidement vers zéro. Rien ne nous empêche d'améliorer l'approximation p après chaque itération. L'algorithme QR avec "shift" devient alors :

```
T_{0}=A
for
   k=1,2,..
determiner le parametre p_{k-1}$
Q_{k}R_{k}=T_{k-1}-p_{k-1}I   (decomposition QR)
T_{k}=R_{k}Q_{k}+p_{k-1}
end
```

Les matrices T_k de cette itération satisfont

$$Q_k^* T_{k-1} Q_k = Q_k^*(Q_k R_k + p_{k-1} I)Q_k = R_k Q_k + p_{k-1} I = T_k$$

Ceci implique que, indépendamment de la suite p_k , les matrices T_k ont toutes les mêmes valeurs propres que $T_0 = A$.

Pour décrire complètement l'algorithme QR avec shift, il faut encore discuter le choix du paramètre p_k et il faut donner un critère pour arrêter l'itération.

18.4.1 Choix du "shift"-paramètre.

On a plusieurs possibilités :

— $p_k = t_{n,n}^{(k)}$: ce choix marche très bien si les valeurs propres de la matrice sont réelles.

— on considère la matrice

$$\begin{pmatrix} t_{n-1,n-1}^{(k)} & t_{n-1,n}^{(k)} \\ t_{n,n-1}^{(k)} & t_{n,n}^{(k)} \end{pmatrix} \tag{18.12}$$

Si les valeurs propres de (18.12) sont réelles, on choisit pour p_k celle qui est la plus proche de $t_{n,n}^{(k)}$.
Si elles sont de la forme $\alpha \pm i\beta$ avec $\beta \neq 0$ (donc complexes), on prend d'abord $p_k = \alpha + i\beta$ et pour l'itération suivante $p_{k+1} = \alpha - i\beta$

18.5 Critère pour arrêter l'itération.

L'idée est d'itérer jusqu'à ce que $t_{n,n-1}^{(k)}$ ou $t_{n-1,n-2}^{(k)}$ soit suffisamment petit. Plus précisément, on arrête l'itération quand

$$t_{l,l-1}^{(k)} \leq eps.(\left|t_{l-1,l-1}^{(k)}\right| + \left|t_{l,l}^{(k)}\right|) \qquad \text{pour} \quad l = n \quad \text{ou} \quad l = n-1 \tag{18.13}$$

— Si (18.13) est vérifié pour $l = n$ on accepte $t_{n,n}^{(k)}$ comme approximation de λ_n et on continue l'itération avec la matrice $\left(t_{i,j}^{(k)}\right)_{1\leq i,j\leq n-1}$

— Si (18.13) est vérifié pour $l = n-1$, on accepte les deux valeurs propres de (18.12) comme approximations de λ_n et λ_{n-1} et on continue l'itération avec la matrice $\left(t_{i,j}^{(k)}\right)_{1\leq i,j\leq n-2}$

Exemple 281. Nous avons appliqué l'algorithme QR à la matrice (**??**) avec le shift $p_k = t_{n,n}^{(k)}$. La convergence de $t_{i+1,i}^{(k)}$ vers zéro est illustrée dans la figure V.5. Une comparaison avec la figure V.4 nous montre que la convergence est beaucoup plus rapide (convergence quadratique). Après 5 itérations, on a $\left|t_{4,3}^{(k)}\right| \leq 10^{-15}$. Encore 4 itérations pour la matrice de dimension 3 donnent $\left|t_{3,2}^{(k)}\right| \leq 10^{-15}$ Il ne reste plus que 3 itérations à faire pour la matrice de dimension 2 pour avoir $\left|t_{2,1}^{(k)}\right| \leq 10^{-15}$. En tout, 12 itérations ont donné toutes les valeurs propres avec une précision de 15 chiffres.

18.6 Le "double shift" de Francis

Dans la situation où A est une matrice réelle ayant des valeurs propres complexes, il est recommandé de choisir un shift-paramètre p_k qui soit complexe. Une application directe de l'algorithme précédent nécessite un calcul avec des matrices complexes. L'observation suivante permet d'éviter ceci.

Lemme 282. Soit T_k une matrice réelle, $p_k = \alpha + i\beta$ et $p_{k+1} = \alpha - i\beta$.Alors, on peut choisir les décompositions dans l'algorithme QR de manière à ce que T_{k+2} soit réelle.

Remarque 283. La décomposition QR d'une matrice est unique sauf qu'on peut remplacer QR par $(QD)^{-1}\left(D^{-1}R\right)$ où $D = diag(d_1,...,d_n)$ avec $|d_i| = 1$.

Démonstration. La formule (**??**) montre que

$$T_{k+2} = (Q_{k+1}Q_{k+2})^* T_k (Q_{k+1}Q_{k+2}) \tag{18.14}$$

cqfd

Il suffit alors de démontrer que le produit $Q_{k+1}Q_{k+2}$ est réel. Une manipulation à l'aide de formules pour T_k donne

$$\begin{aligned} Q_{k+1}Q_{k+2}R_{k+2}R_{k+1} &= Q_{k+1}(T_{k+1} - p_{k+1}I)R_{k+1} = Q_{k+1}(R_{k+1}Q_{k+1} + p_{k+1}I - p_kI)R_{k+1} = \\ &= (Q_{k+1}R_{k+1})^2 + (p_k - p_{k+1})Q_{k+1}R_{k+1} = (T_k - p_kI)^2 + (p_k - p_{k+1})(T_k - p_kI) = \\ &= T_k^2 - (p_k + p_{k+1})T_k + p_kp_{k+1}I = M \end{aligned} \tag{18.15}$$

On a donc trouvé une décomposition QR de la matrice M qui, en conséquence des hypothèses du lemme, est une matrice réelle. Si, dans l'algorithme QR, la décomposition est choisie de manière à ce que les éléments diagonaux de R_{k+1} et R_{k+2} soient réels, alors, à cause de l'unicité de la décomposition QR, les matrices $Q_{k+1}Q_{k+2}$ et $R_{k+2}R_{k+1}$ sont réelles.

Une possibilité de calculer T_{k+2} à partir de T_k est de calculer de (18.15), de faire une décomposition QR (réelle) de M et de calculer T_{k+2} à l'aide de (18.14). Cet algorithme n'est pas pratique car le calcul de T_k^2 nécessite $O(n^3)$ opérations, même si T_k est sous forme de Hessenberg.

Il y a une astuce intéressante pour obtenir T_{k+2} à partir de T_k en $O(n^2)$ opérations. Elle est basée sur la propriété suivante.

Théorème 284. Soit une matrice donnée et supposons que

$$Q^*TQ = S \tag{18.16}$$

où Q est orthogonale et S est sous forme de Hessenberg satisfaisant $s_{i,i-1} \neq 0$ pour $i = 2,...,n$ Alors, Q et S sont déterminées de manière "unique" par la première colonne de Q.

Remarque 285. On a "unicité" dans le sens suivant : si $\hat{Q}^*T\hat{Q}$ est de type Hessenberg avec une matrice orthogonale $\hat{Q}$ satisfaisant $\hat{Q}e_1$, alors $\hat{Q} = QD$ où $D = diag(d_1,...,d_n)$ avec $|d_i| = 1$.

Démonstration. Notons les colonnes de Q par q_i. Alors, la relation (18.16) implique

$$Tq_i = \sum_{j=1}^{i+1} s_{ji}q_j, \qquad q_j^*Tq_i = s_{ji}. \tag{18.17}$$

Si q_1 est fixé, la valeur s_{11} est donnée par la deuxième formule de (18.17). Avec cette valeur, on obtient de la première formule de (18.17) que q_2 est un multiple de $Tq_1 - s_{11}q_1$. Ceci détermine q_2 à une unité près. Maintenant, les valeurs s_{21}, s_{12}, s_{22} sont déterminées et q_3 est un multiple de $Tq_2 - s_{21}q_1 - s_{22}q_2$ etc. cqfd

Si les hypothèses du lemme précédent sont vérifiées, on peut calculer la matrice réelle T_{k+2} en $O(n^2)$ opérations de la manière suivante :

— calculer $M\varepsilon_1$, la première colonne de M (formule (18.15));
— déterminer une matrice de Householder H_1 telle que $H_1(M\varepsilon_1) = \alpha e_1$
— transformer $H_1^T T_k H_1$ sous forme de Hessenberg à l'aide de matrices de Householder $H_2, ..., H_{n-1}$ (voir le paragraphe V.3); c'est-à-dire., calculer $H^T T_k H$ où $H = H_1 H_2 .... H_{n-1}$.

Comme $H_i e_1 = e_1$ pour $i = 2, ..., n-1$, la première colonne de H est un multiple de celle de M (observer $H_1^T = H_1$). Par la formule (18.15), la première colonne de $Q_{k+1}Q_{k+2}$ est aussi un multiple de Me_1. Par conséquent, pour un bon choix des décompositions $Q_{k+1}R_{k+1}$ et $Q_{k+2}R_{k+2}$ on a $H = Q_{k+1}Q_{k+2}$ la matrice obtenue par cet algorithme est égale à T_{k+2} (voir (18.14)).

18.7 Etude de la convergence

Supposons d'être déjà proche de la limite et considérons, par exemple, la matrice

$$T_0 = A = \begin{pmatrix} 2 & a \\ \varepsilon & 1 \end{pmatrix}$$

où ε est un nombre petit. Avec le choix $p_0 = 1$ pour le shift-paramètre, on obtient

$$T_0 - p_0 I = \begin{pmatrix} 1 & a \\ \varepsilon & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\varepsilon^2}} & -\frac{\varepsilon}{\sqrt{1+\varepsilon^2}} \\ \frac{\varepsilon}{\sqrt{1+\varepsilon^2}} & \frac{1}{\sqrt{1+\varepsilon^2}} \end{pmatrix} \begin{pmatrix} \sqrt{1+\varepsilon^2} & \frac{a}{\sqrt{1+\varepsilon^2}} \\ 0 & -\frac{a\varepsilon}{\sqrt{1+\varepsilon^2}} \end{pmatrix} = Q_1 R_1$$

et

$$T_0 - p_0 I = R_1 Q_1 = \begin{pmatrix} * & * \\ -\frac{a\varepsilon^2}{1+\varepsilon^2} & * \end{pmatrix}$$

— si A est symétrique (c'est-à-dire, $a = \varepsilon$) on a $t^{(1)}_{n,n-1} = O(\varepsilon^2)$, donc convergence *cubique*.
— si A n'est pas symétrique (p.ex. on a donc convergence quadratique.

Ces propriétés restent vraies pour des matrices générales (sans démonstration).

19 Exercices

Exercice 286. Calculer les valeurs propres de la matrice tridiagonale (dimension $n, b.c > 0$)

$$A = \begin{pmatrix} a & c & & & \\ b & a & c & & \\ & b & a & c & \\ & & b & \ddots & \ddots \\ & & & \ddots & \end{pmatrix}$$

Indication. Les composants du vecteur propre $(v_1, v_2, ..., v_n)^T$ satisfont une équation aux différences finies avec.$v_0 = v_{n+1} = 0$ Vérifier que $v_j = Const.(\alpha_1^j - \alpha_2^j)$ où

$$\alpha_1 - \alpha_2 = \frac{\lambda - a}{c}, \qquad \alpha_1.\alpha_2 = \frac{b}{c}, \qquad \left(\frac{\alpha_1}{\alpha_2}\right)^{n+1} = 1$$

Résultat : $\lambda_j = a - 2\sqrt{bc}.\cos\left(\frac{j\pi}{n+1}\right), \qquad j = 1, 2, .., n.$

Exercice 287. Considérer la matrice

$$A(\varepsilon) = \begin{pmatrix} 1 & \varepsilon & 0 \\ -1 & 0 & 1 \\ 1 & -1+\varepsilon & -\varepsilon \end{pmatrix}$$

cette matrice possède une valeur propre de la forme

$$\lambda(\varepsilon) = i + \varepsilon.d + O(\varepsilon^2)$$

Calculer d et dessiner la tangente à la courbe $\lambda(\varepsilon)$ au point $\lambda(0)$

(a) Calculer par la méthode de la puissance, la plus grande valeur propre de la matrice

$$A = \begin{pmatrix} 99 & 1 & 0 \\ 1 & 100 & 1 \\ 0 & 1 & 98 \end{pmatrix}$$

(b) Pour accélerer considérablement la vitesse de convergence, appliquer la méthode de la puissance à la matrice $A - pI$ avec un choix intelligent de p.

(c) Avec quel choix de p obtient-on la valeur propre la plus petite ?

Exercice 288. Considérons la matrice tridiagonale

$$A = \begin{pmatrix} b_1 & c_1 & \\ a_1 & b_2 & c_2 \\ & a_2 & \\ & & \ddots \end{pmatrix}$$

Montrer que, si $a_i c_i > 0$ pour $i = 1, ..., n-1$, toutes les valeurs propres de A sont réelles.

Indication. Trouver $D = diag(d_1, ..., d_n)$ telle que DAD^{-1} soit symétrique.

Exercice 289. Soit A une matrice symétrique et B quelconque. Montrer que pour chaque valeur propre λ_B de B il existe une valeur propre λ_A de A telle que

$$|\lambda_A - \lambda_B| \leq \|A - B\|_2 .$$

Indication. Montrer l'existence d'un vecteur v tel que $v = (A - \lambda_B)^{-1}(A - B)v$. En déduire que $1 \leq \left\|(A - \lambda_B)^{-1}(A - B)\right\| \leq \left\|(A - \lambda_B)^{-1}\right\| \|(A - B)\|$.

Exercice 290. (Schur, 1909). Soit A une matrice symétrique. Montrer que pour chaque indice i il existe une valeur propre λ de A telle que

$$|\lambda - a_{ii}| \leq \sqrt{\sum_{j \neq i} |a_{ij}|^2}$$

Indication. Appliquer l'exercice 5 avec une B convenable.

Exercice 291. Soit A une matrice réelle avec pour valeur propre $\alpha + i\beta$. Montrer que l'itération

$$\begin{pmatrix} \bar{\alpha} I - A & -\bar{\beta} I \\ \bar{\beta} I & \bar{\alpha} I - A \end{pmatrix} \begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix}$$

où $\bar{\alpha} \approx \alpha$ et $\bar{\beta} \approx \beta$) permet de calculer la valeur propre $\alpha + i\beta$ et le vecteur propre correspondant.
Indication. Considérer les parties réelles et complexes de l'itération de Wielandt. On obtient alors

$$\frac{u_k^T A u_k + v_k^T A v_k}{u_k^T u_k + v_k^T v_k} \longrightarrow \alpha, \qquad \frac{u_k^T A v_k + v_k^T A u_k}{u_k^T u_k + v_k^T v_k} \longrightarrow \beta$$

Exercice 292. Considérons la matrice de Hilbert,

$$A = \begin{pmatrix} 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \\ 1/4 & 1/5 & 1/6 \end{pmatrix}$$

(a) Transformer A en une matrice tridiagonale ayant les mêmes valeurs propres.

(b) En utilisant une suite de Sturm, montrer que toutes les valeurs propres sont positives et qu'une valeur propre est plus petite que 0.001

(c) Calculer approximativement la condition de A pour la norme Euclidienne.

Exercice 293. La formule de récurrence

$$(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x)$$

pour les *polynômes de Legendre* ressemble à

$$p_i(\lambda) = (d_i - \lambda)p_{i-1}(\lambda) - \varepsilon_i^2 p_{i-2}(\lambda), \qquad i = 2,...,n.$$

pour les polynômes $\det(A_i - \lambda I)$. Trouver une matrice tridiagonale A de dimension n telle que les valeurs propres de A sont les racines de $P_n(x)$.

Exercice 294. Soit $p(x)$ un polynôme de degré n et supposons que toutes les racines soient simples. Démontrer que la suite définie par l'algorithme d'Euclide,

$$\begin{aligned} p_n(x) &= p(x), \qquad p_{n-1}(x) = -p'(x) \\ p_i(x) &= q_i(x)p_{i-1}(x) - \gamma_i^2 p_{i-2}(x), \qquad i = n,...,2. \end{aligned}$$

est une suite de Sturm.

Pour le polynôme $p(x) = x^5 - 6x^4 + 3x^3 + 3x^2 + 2x + 8$.

(a) déterminer le nombre de racines réelles.

(b) Combien de racines sont complexes ?

(c) Combien de racines sont réelles et positives ?

Exercice 295. Pour un φ donné notons $c = \cos\varphi$ et $s = \sin\varphi$.La matrice Ω_{kl}, définie par

$$(\Omega_{kl})_{ij} = \begin{cases} 1 & \text{si } i = j, j \neq k, j \neq l \\ c & \text{si } i = j = k, \text{ ou } i = j = l \\ s & \text{si } i = k, \text{ et } j = l \\ -s & \text{si } i = l, \text{ et } j = k \\ 0 & \text{sinon} \end{cases}$$

s'appelle rotation de Givens.

(a) Montrer qu'elle est orthogonale.

(b) Soit A une matrice symétrique.Déterminer φ tel que le (k,l)-ième élément de $A' = \Omega_{kl} A \Omega_{kl}^T$ s'annule.

Resultat. $\cot 2\varphi = (a_{kk} - a_{ll})/(2a_{kl})$.

Exercice 296. La méthode de Jacobi (1846) pour le calcul des valeurs propres d'une matrice symétrique :

i) on choisit a_{kl} $(k > l)$ tel que $|a_{kl}| = \max_{i>j} |a_{ij}|$;

ii) on détermine A' comme dans l'exercice 11.

Montrer que, si on répète cette procédure, on a convergence vers une matrice diagonale, dont les éléments sont les valeurs propres de A

Indication. Montrer que $\sum_{i>j} |a'_{ij}|^2 = \sum_{i>j} |a_{ij}|^2 - |a_{kl}|^2$

Exercice 297. On considère la matrice

$$A = \begin{pmatrix} 7 & 0,5 \\ 0,0001 & 8 \end{pmatrix}$$

dont on cherche à calculer les valeurs propres.

(a) Faire une itération de l'algorithme QR sans shift.

(b) Faire une itération de l'algorithme QR avec shift.

(c) Estimer la position des valeurs propres de A à l'aide du Théorème de Gershgorin.

(d) Calculer les valeurs propres de A à l'aide du polynôme caractéristique.

Exercice 298. Montrer que si la matrice $T_0 = A$ est une matrice de Hessenberg (ou tridiagonale), alors les matrices T_k, $k \geq 1$ construites par l'algorithme QR sont également des matrices de Hessenberg (tridiagonales).

Exercice 299. Donner une estimation grossière du nombre d'opérations qui sont nécessaires pour effectuer la décomposition QR d'une matrice de Hessenberg et pour calculer ensuite le produit RQ.

Exercice 300. Soit T_0 une matrice de Hessenberg dont tous les éléments de la sous-diagonale sont non-nuls. Montrer que, si p_0 est une valeur propre de T_0, une itération de l'algorithme QR avec shift p_0 donne

$$t^{(1)}_{n,n-1} = 0.$$

Exercice 301. Expliquer, comment le calcul de T_k à partir de T_{k-1}

$$Q_k R_k = T_{k-1} - p_{k-1} I, \qquad T_k = R_k Q_k + p_{k-1} I.$$

peut être effectué sans soustraire (et additionner) explicitement la matrice $p_{k-1} I$

Indication. Laissez-vous inspirer par le "double shift" algorithme de Francis.