Ecole Centrale de Nantes
Dépt. Info/Math
Année universitaire 2011-2012
EI 1

ANALYSE NUMERIQUE

Mazen SAAD
Mazen.Saad@ec-nantes.fr

TABLE DES MATIÈRES

INTRODUCTION

Les mathématiques appliquées et le calcul scientifique jouent un rôle croissant dans la conception de produits industriels ; ce n'est cependant qu'un maillon d'une longue chaîne qui mobilise des ressources intellectuelles nombreuses et variées pour arriver à concevoir, au mieux dans des délais impartis le produit désiré. On peut représenter très schématiquement un processus d'étude et de conception par le diagramme suivant :

- Physique mécanique, modélisation mécanique (aérodynamique, thermique, structure, ...)
- Modélisation mathématique (E.D.P.)
- Approximation : Eléments finis, volumes finis...
- Algorithme numérique, méthodes numériques pour la résolution de systèmes linéaires et non linéaires, optimisation
- Calcul informatique ...
- Expérimentation
- Exploitation des produits

La modélisation et l'approximation numérique voient leurs applications dans différents domaines, à titre d'exemples :

- Conception d'avions (aérodynamique, matériaux composites ...)
- Conception de voitures (aérodynamique, écoulement dans les moteurs, crache tests, commande optimale, structure (pneus, carrosserie,) ....
- Ingénierie pétrolière : comprendre la migration des hydrocarbures, améliorer la production des gisements pétroliers, ...
- Biologie mathématiques : propagation d'épidémie, modèle mathématique en cardiologie, cancer, tissus dentaire, pneumologie, ...
- Gestion des stocks, finance, trafic routier
- Environnement : pollution air, eau, sol
- Météo : modéliser le monde
- Et bien d'autres applications ...

Dans ce cours, nous nous intéressons à l'analyse numérique ; cette discipline elle-même peut être considérée comme partagée en deux grands thèmes :

- Approximation numérique des EDP (Eléments finis, volumes finis, méthodes spectrales, ...)
- Algorithmes numériques : résolution de grands systèmes linéaires creux, intégration numérique, résolution numérique des EDO, optimisation

L'objet de ce cours est de déterminer des méthodes pour calculer la valeur numérique (exacte ou approchée) de la solution d'une équation ou d'un système d'équations ; en particulier à l'aide d'un ordinateur.

CHAPITRE 1

ALGÈBRE LINÉAIRE

1.1. Arithmétique flottante

Il est important de se préoccuper de la manière dont sont représentés et manipulés les nombres dans une machine. Un nombre est représenté par un nombre finis de caractères, fixé à l'avance, qui dépend de l'architecture de la machine. Ainsi tous les nombres entiers ou réels ne peuvent pas être représentés. Les conséquences en sont très importantes, en particulier dans la précision des résultats lors de calculs.

Comment sont représentés et manipulés les nombres sur un ordinateur ?

La mémoire centrale est un ensemble de 'positions binaires' nommées bits. Les bits sont généralement regroupés en octets (8 bits) et chaque octet est repéré par son adresse. Chaque information devra être codée sous cette forme binaire.

En informatique,

le **kilo** vaut $1\text{K} = 2^{10} = 1024$
le **méga** vaut $1\text{M} = 2^{20} = 1048576$
le **giga** vaut $1\text{G} = 2^{30} = 1073741824$

On distingue :

– *Les nombres entiers* dont la représentation et la manipulation sont celles de l'arithmétique usuel. Il existe un plus grand entier représenté en machine.

Les entiers relatifs codés sur n chiffres binaires ont pour valeur dans $[-2^{n-1}, 2^{n-1} - 1]$. Ainsi les entiers codés sur

16 bits (=2 octets) correspond à des entiers en **simple précision** ont pour valeur dans $[-2^{15}, 2^{15} - 1] = [-32K, 32K - 1]$

32 bits (=4 octets) correspond à des entiers en **double précision** ont pour valeur dans $[-2^{31}, 2^{31} - 1] = [-2G, 2G - 1]$.

– *Les nombres flottants* qui représentent les nombres réels ou les nombres décimaux. Les nombres réels sont représentés de façon approximative en mémoire (représentation en virgule flottante), avec la convention standardisée de la forme $m \times 2^e$, où m est la mantisse $1 \leq m \leq 2$ et e l'exposant.

On utilise p chiffres binaires pour les décimaux binaires de m et q chiffres binaires pour l'exposant.

Représentation en simple précision. Sur 32 bits (4 octets), on a $p = 23$, $q = 8$ (1 bit pour le signe) ce qui permet de représenter des nombres compris, en valeur absolue, entre $2^{-128} \approx 10^{-38}$ et $2^{128} \approx 10^{38}$ car $128 = 2^q = 2^8$. La précision machine est de 7 chiffres décimaux significatifs car $2^{23} = 10^7$.

Représentation en double précision. Sur 64 bits (8 octets), on a $p = 52$, $q = 11$ et les réels en valeur absolue appartiennent $[2^{-1024}, 2^{1024}] \approx [10^{-308}, 10^{308}]$ avec 15 chiffres décimaux significatifs (car $2^{52} \approx 10^{15}$).

La représentation exacte en machine est sous forme binaire (comme on a vu), pour l'analyse que nous voulons faire ici une représentation décimale est suffisante et plus intuitive. On considère un nombre flottant de la forme $\pm a 10^q$ avec

$$a \text{ est la mantisse de la forme } 0.d_1 d_2 \cdots d_t, \qquad d_1 \neq 0$$

$$q \text{ est l'exposant (entier relatif)}$$

Bien sûr, l'entier q est soumis à la restriction :

$$-M \leq q \leq M \text{ (où M dépend de la machine).}$$

Cette représentation des nombres réels entraîne les conséquences suivantes :

- Il existe un plus petit nombre flottant ($\neq$ zéro). Le zéro machine en valeur absolue vaut $= 0.10 \cdots 10^{-M}$.
- Il existe un plus grand nombre flottant, l'infinie machine vaut $= 0.99 \cdots 910^M$.
- Tous les nombres réels n'admettent de représentation exacte :

$$\sqrt{2} \text{ est représenté par } 0.14142143 \times 10^{+1}$$

$$\pi \text{ est représenté par } 0.314... \times 10^{+1}$$

- Toute opération élémentaire (+,$*$, /) est en général entachée d'une erreur.
- Une opération peut avoir un résultat non représentable :
  Si pour le résultat $q > M$ (OVERFLOW ou dépassement de capacité.)
  Si pour le résultat $q < -M$ (UNDERFLOW).
- La représentation flottante d'un nombre peut être obtenue à partir de sa représentation décimale par
  – la troncature (on garde les t premiers décimaux)
  – l'arrondi : le tième chiffre de la mantisse est choisi au plus près.

Regardons maintenant l'erreur due à la représentation machine.

Proposition 1.1. — *La représentation flottante $fl(r)$ avec une mantisse à t chiffres d'un nombre réel r donne lieu à une erreur relative majorée par :*

$$\frac{|r - fl(r)|}{|r|} \leq 10^{1-t}.$$

Démonstration. — La représentation exacte d'un réel r s'écrit :

$$r = \pm 0.d_1 d_2 \cdots d_t d_{t+1} d_{t+2} \cdots 10^q,$$

et on a $fl(r) = \pm 0.d_1 d_2 \cdots d_t 10^q$. Ainsi $r - fl(r) = \pm 0.d_{t+1} d_{t+2} \cdots 10^{q-t}$ et on a

$$\frac{|r - fl(r)|}{|r|} = \frac{0.d_{t+1} d_{t+2} \cdots}{0.d_1 d_2 \cdots d_t d_{t+1} d_{t+2} \cdots} 10^{-t}.$$

Par ailleurs, $0.d_{t+1} d_{t+2} \cdots \leq 1$ et $0.d_1 d_2 \cdots d_t d_{t+1} d_{t+2} \cdots \geq 0.1$, d'où

$$\frac{|r - fl(r)|}{|r|} \leq 10^{1-t} (\text{par troncature}).$$

□

Quelques conséquences de cette représentation :

- $a + b = a$ si b est plus petit que le zéro machine. Par exemple, soit une machine avec $t = 2$ et $a = 0.63 \times 10^1$ et $b = 0.82 \times 10^{-4}$. Pour faire l'opération, on (la machine) réduit au même exposant, soit

  $$a + b = 0.63 \times 10^1 + 0.0000082 \times 10^1 = 0.6300082 \times 10^1,$$

  et ce dernier nombre est représenté par $fl(a + b) = 0.63 \times 10^1$ car $t = 2$.
  Conclusion : $a + b = a$ et $b \neq 0$.
- L'addition des nombres flottants n'est pas associative. Soit une machine avec $t = 4$ et $a = 0.6724 \times 10^3$, $b = 0.7215 \times 10^{-1}$ et $c = 0.5345 \times 10^1$, on a

  $$fl((a + b) + c) = 0.6777 \times 10^3 \text{ car } fl(a + b) = fl(a)$$
  $$fl(a + (b + c)) = 0.6778 \times 10^3 \text{ car } fl(b + c) \neq fl(c)$$
- Même phénomène pour la soustraction, division, multiplication ...
  Soient $y = a + b$ et $z = \frac{a}{y-b}$ alors $z = 1$. Mais par contre si $fl(y) = fl(b)$, alors on ne peut pas calculer z et un message d'erreur apparaît OVERFLOW.

1.2. Un peu de calcul matriciel

On note $M_{n,m}(K)$ l'ensemble des matrices de type (n, m) n-lignes et m-colonnes dont les coefficients appartiennent à $K = \mathbb{R}$ ou $\mathbb{C}$. On note $M_n(K)$ l'ensemble des matrices carrées d'ordre n.

Une matrice $M \in M_{n,m}(K)$ est associée à une application linéaire l de $E = K^m$ dans $G = K^n$. Soient $\{e_j\}_{j=1,m}$ base de K^m et $\{g_i\}_{i=1,n}$ une base de K^n ; la jième colonne de la matrice M est constituée des coordonnées de $l(e_j)$ dans la base $\{g_i\}_{i=1,n}$.

Produit scalaire. Soit $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$(x, y) = \sum_{i=1}^{n} x_i y_i = {}^t xy = {}^t yx.$$

Produit hermitien. Soit $(x, y) \in \mathbb{C}^n \times \mathbb{C}^n$,

$$(x, y) = \sum_{i=1}^{n} x_i \overline{y_i} = \overline{{}^t y} x = y^* x.$$

Avec $y^* = \overline{{}^t y}$ l'adjoint de y.

Définition 1.1. — *Soit $A \in M_{n,m}(K)$, on dit que A est*

- *hermitienne si $A = A^*$ $\quad (A^* = {}^t(\overline{A}) = \overline{{}^t A})$.*
- *symétrique si $A = {}^t A$*
- *unitaire si $AA^* = A^*A = I$*
- *orthogonale si A est réelle et ${}^t AA = A^t A = I$ soit encore $A^{-1} = {}^t A$*
- *normale si $AA^* = A^*A$.*

1.2.1. Valeurs et vecteurs propres. —

Définition 1.2. — *On appelle*

- *$(\lambda, u) \in \mathbb{C} \times \mathbb{C}^N$ élément propre de A si $Au = \lambda u$ et λ valeur propre de A, u vecteur propre associé à λ.*
- *$Sp(A) = \{\lambda_i; \lambda_i$ valeur propre de $A\}$= spectre de A.*
- *$\rho(A) = \max_{i=1,N} |\lambda_i|$= rayon spectral de A.*
- *$Tr(A) = \sum_{i=1}^{N} a_{ii}$= trace de A, avec $A = (a_{ij})$.*
- *Les valeurs propres de A sont les racines du polynôme :*

$$P_A(\lambda) = det(1 - \lambda I) = (-1)^N \lambda^N + (-1)^{N-1} \lambda^{N-1} + \cdots + det(A).$$

- *Les vecteurs propres de A sont les vecteurs tels que $Av = \lambda v$ et ils forment un sous espace vectoriel $E_\lambda = \{v \in K^N; Av = \lambda v\}$.*
- *$Tr(A) = \sum_{i=1}^{N} \lambda_i$, $\quad det(A) = \Pi_{i=1}^{N} \lambda_i$ (propriétés).*
- *A est semblable à B s'il existe une matrice inversible $S \in M_n(K)$ telle que $A = SBS^{-1}$.*

- *A est diagonalisable ssi $A = SDS^{-1}$ avec*
  D la matrice diagonale formée des valeurs propres,
  la ième colonne de S est un vecteur propre (à droite) associé à λ_i,
  la jième colonne de $(S^{-1})^\star$ est un vecteur propre à gauche v_j associé à λ_j. En fait les colonnes de S sont les u_j et les lignes de S^{-1} sont les v_i^*.

Théorème 1.1. — *(Factorisation unitaire d'une matrice– Théorème de Schur) Toute matrice carrée peut s'écrire*

$$A = UTU^*$$

avec U une matrice unitaire $U^{-1} = U^*$,
T une matrice triangulaire supérieure.

Conséquence sur les matrices normales :

Théorème 1.2. — *Une matrice A est normale (i.e. $AA^* = A^*A$) si et seulement si il existe U une matrice unitaire telle que*

$$A = UDU^*$$

avec D la matrice diagonale formée des valeurs propres.

Autrement dit,
une matrice normale est diagonalisable et ses vecteurs propres sont orthonormés.

Démonstration. — D'après le théorème de Shur, la matrice A s'écrit $A = UTU^*$. Or A est normale c'est à dire $AA^* = A^*A$ soit encore

$$UT^*U^*UTU^* = UTU^*UT^*U^*$$

et donc $UT^*TU^* = UTT^*U^*$, ce qui montre que $T^*T = TT^*$ et T est normale.

On va montrer que si T est une matrice triangulaire supérieure et une matrice normale alors T est diagonale.

En effet, pour tous $i, j = 1 \cdots N$, on a

$$(T^*T)_{ij} = (TT^*)_{ij}$$

ce qui équivalent à

$$\sum_{k=1}^{N} t_{ik}^* t_{kj} = \sum_{k=1}^{N} t_{ik} t_{kj}^*$$

soit encore

$$\sum_{k=1}^{N} \overline{t_{ki}} t_{kj} = \sum_{k=1}^{N} t_{ik} \overline{t_{jk}}.$$

Lorsque $i = j$, on a

$$\sum_{k=1}^{N} |t_{ki}|^2 = \sum_{k=1}^{N} |t_{ik}|^2, \tag{1.1}$$

or $t_{ki} = 0$ pour $k > i$ et $t_{ik} = 0$ pour $i > k$, l'égalité (1.1) se réduit à

$$\sum_{k=1}^{i} |t_{ki}|^2 = \sum_{k=i}^{N} |t_{ik}|^2. \tag{1.2}$$

Pour $i = 1$, on a $|t_{11}|^2 = |t_{11}|^2 + \sum_{k=2}^{N} |t_{1k}|^2$, soit $t_{1k} = 0$ pour $k \geq 2$; c'est à dire que la première ligne de la matrice T est nulle sauf le terme diagonale. Par récurrence, supposons que $t_{ij} = 0$, $i \neq j$ jusqu'à la ligne $m - 1$. Alors pour $i = m$,

$$\sum_{k=1}^{m} |t_{km}|^2 = \sum_{k=m}^{N} |t_{mk}|^2,$$

soit encore

$$|t_{mm}|^2 + \sum_{k=1}^{m-1} |t_{km}|^2 = |t_{mm}|^2 + \sum_{k=m+1}^{N} |t_{mk}|^2,$$

et par hypothèse de récurrence, $t_{km} = 0$ pour $k \leq m-1$ et o déduit $t_{mk} = 0$ pour $k \geq m+1$. Toute la ligne m est nulle sauf l'élément diagonale. Ainsi, la matrice T est diagonale.

Inversement, si $A = UDU^*$ alors A est normale car $A^*A = UD^*U^*UDU^* = UD^*DU^*$ et $AA^* = UDU^*UD^*U^* = UDD^*U^*$; or D est diagonale donc $D^*D = DD^*$, ce qui termine la preuve du résultat. □

On aboutit alors au résultat important suivant

Corollaire 1.1. — *Toute matrice symétrique réelle est diagonalisable et la base des vecteurs propres est orthonormée.*

Car si A une matrice symétrique réelle alors A est normale. De même, si A est une matrice hermitienne alors A est normale.

CHAPITRE 2

RÉSOLUTION DES GRANDS SYSTÈMES LINÉAIRES CREUX

De très nombreux phénomènes physiques sont régis par un loi de diffusion : répartition de température, concentration de produits chimiques, potentiel électrique, ... Dans tous les cas, on cherche à discrétiser les équations et à résoudre numériquement les équations mises en jeu. pour des soucis de précision, de stabilité, de pertinence des résultats, on est amené à résoudre des systèmes linéaires ou non linéaires de grandes tailles.

Voici deux exemples.

2.1. Exemple 1. Equation de la chaleur

La distribution de la température $u(x, y)$ au point (x, y) d'une plaque dont les côtés ont une température imposée $u = 0$ sur le bord et qui reçoit un apport calorifique extérieur de densité f est modélisée par une équation aux dérivées partielles. Soit $\Omega = [0, a] \times [0, b]$ désignant la plaque, la température vérifie

$$\begin{cases} -\Delta u(x,y) = -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y) \text{ dans } \Omega \\ u = 0 \text{ sur } \partial\Omega \end{cases} \tag{2.3}$$

Même si on sait qu'il existe une unique solution de ce problème, la solution de ce problème n'est pas connue analytiquement en général. On procède alors à une approximation pour se ramener à un problème à un nombre fini d'inconnus (processus de discrétisation). On introduit donc un maillage de pas h_1 dans la direction x et h_2 dans la direction y. Pour fixer les idées, on prend ici $h_1 = h_2 = h$ (voir figure 1).
Les noeuds du maillage sont les points $P_{i,j} = (x_i, y_j)$ là où la solution est approchée. On note

$$x_i = ih,\ 0 \leq i \leq N+1 \text{ les sommets du maillage dans la direction x}$$
$$y_j = jh,\ 0 \leq j \leq M+1 \text{ les sommets du maillage dans la direction y}$$

On cherche une approximation de l'équation aux noeuds du maillage (P_{ij}, $1 \leq i \leq$

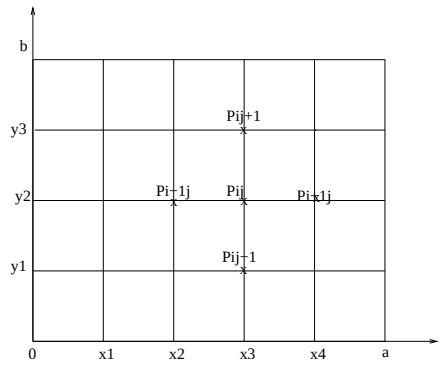


FIGURE 1. Exemple de maillage pour $N = 4$, $M = 3$

N, $1 \leq j \leq M$. Le principe de la méthode des différences finies consiste à approcher les dérivées d'une fonction par des combinaisons linéaires des valeurs de cette fonction aux points du maillage. On va décrire tout d'abord ce principe en dimension un d'espace.
Dimension 1. On s'intéresse à l'approximation de l'équation

$$\begin{cases} -u''(x) = f(x), \ \ 0 < x < a \\ u(0) = \alpha, \ u(a) = \beta \end{cases} \tag{2.4}$$

par un schéma aux différences finies sur un maillage à pas fixe $h = \frac{a}{N+1}$

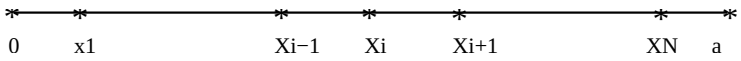


FIGURE 2. Exemple de maillage 1D.

On écrit la formule de Taylor sur un point générique x_i, on a

$$-u''(x_i) = \frac{1}{h^2}\Big(- u(x_{i-1}) + 2u(x_i) - u(x_{i+1})\Big) + 0(h^2) = f(x_i),\ i = 1...N.$$

On note par u_i une approximation de la solution exacte au point $u(x_i)$, et la méthode aux différences finies s'écrit alors

$$\begin{cases} \frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) = f_i, \ \ i = 1, N. \\ u_0 = u(0) = \alpha \\ u_{N+1} = u(a) = \beta \end{cases} \tag{2.5}$$

Le système linéaire (2.5) s'écrit $A_1 U = F$, où $A_1 \in M_{N\times N}(\mathbb{R})$ et U, $F \in \mathbb{R}^N$:

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0... & 0 \\ -1 & 2 & -1... & 0 \\ & & & \\ 0... & -1 & 2 & -1 \\ 0... & 0 & -1 & 2 \end{pmatrix}, \qquad \begin{pmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ \cdots \\ f(x_{N-1}) \\ f(x_N) + \frac{\beta}{h^2} \end{pmatrix}.$$

En dimension 2. On discrétise chaque dérivée selon sa propre direction, ainsi en appliquant la formule de Taylor dans les directions x et y, on a

$$-\frac{\partial^2 u}{\partial x^2}(P_{i,j}) = \frac{-u(P_{i-1,j}) + 2u(P_{i,j}) - u(P_{i+1,j})}{h^2} + 0(h^2)$$
$$-\frac{\partial^2 u}{\partial y^2}(P_{i,j}) = \frac{-u(P_{i,j-1}) + 2u(P_{i,j}) - u(P_{i,j+1})}{h^2} + 0(h^2).$$

En résumé, on notant $u_{i,j}$ une approximation de la solution de (2.3) au point $P_{i,j}$, la discrétisation par différences finies se ramène à la résolution du système linéaire

$$\frac{1}{h^2}(-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}) + \frac{1}{h^2}(-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}) = f_{i,j};\ 1 \leq i \leq N,\ 1 \leq j \leq M \tag{2.6}$$

$$u_{i,0} = u_{i,M+1} = u_{0,j} = u_{N+1,j} = 0,\ 1 \leq i \leq N,\ 1 \leq j \leq M. \tag{2.7}$$

C'est un système linéaire et on aimerait pour le résoudre pouvoir l'écrire sous la forme matricielle

$$AU = F$$

avec $A \in M_{r\times r}(\mathbb{R})$, U, $F \in \mathbb{R}^r$ et $r = N \times M$. Cela veut dire que nous devons ranger les inconnus $u_{i,j}$, les points intérieurs au domaine, dans un vecteur U de dimension $r = N \times M$, ceci conduit à numéroter les points du maillage.

Numérotation des points du maillage. Il y plusieurs façons pour numéroter les sommets du maillage, par exemple on peut numéroter le sommets de gauche à droite et de bas en haut (voir figure 3), ou considérer une numérotation selon les diagonales, numérotation zèbre, numérotation échiquier ...(voir TD) Dans l'exemple de la figure 3, on a $N = 4$,

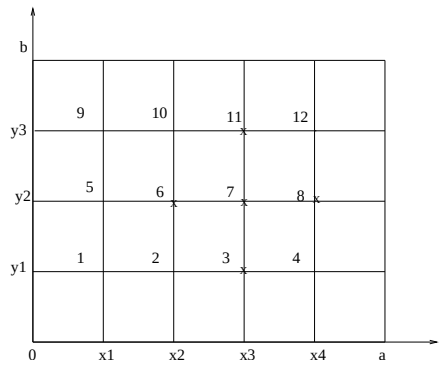


FIGURE 3. Numérotation des sommets de gauche à droite et de bas en haut.

$M = 3$ et $r = 12$, le sommet $m = 7$ correspond au point $P_{3,2} = (x_3, y_2)$ et le sommet

$m = 11$ correspond au point $P_{3,3} = (x_3, y_3)$. On vérifie alors que la numérotation globale pour $m = 1, r$ correspond alors aux points $P_{i,j}$ avec

$$m = i + (j-1)N, \text{ pour } i = 1, N, \ j = 1, M.$$

On peut écrire dans ce cas les équations du système linéaire

$$\begin{aligned}
&\text{Eq. 1} : 4u_1 - u_2 - u_5 = h^2 f_1 \\
&\text{Eq. 2} : 4u_2 - u_1 - u_3 - u_6 = h^2 f_2 \\
&... \\
&\text{Eq. 7} : 4u_7 - u_3 - u_6 - u_8 - u_{11} = h^2 f_7 \\
&...
\end{aligned}$$

ce qui correspond à l'écriture matricielle suivante

$$\frac{1}{h^2}\begin{vmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\\
4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4
\end{vmatrix}$$

cas général : .....

On peut également écrire les matrices associées à différentes numérotations. En tout cas, il est clair que la numérotation influence considérable la structure de la matrice A. La matrice A est creuse dans le sens où elle contient beaucoup de zéro. En effet, pour tous N et M, chaque ligne de la matrice A contient au plus cinq éléments non nuls.

2.2. Exemple 2. Problèmes de réseaux.

Soit un réseau fermé de noeuds P_i et d'arêtes $E_{i,j}$ reliant les noeuds P_i et P_j. C'est le cas de canalisations d'eau, de lignes électriques ...

A chaque noeud est associé un potentiel u_i. Dans chaque arête circule un fluide dont l'intensité (ou le débit) est proportionnel à la différence des potentiels $q_{i,j} = k_{i,j}(u_i - u_j)$. Le

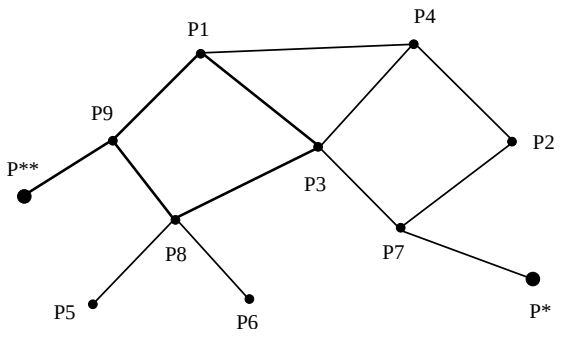


FIGURE 4. Réseau de canalisation

réseau est alimenté à partir de noeuds sources. La loi de conservation impose que le débit total est nul si le noeud est isolé (à l'intérieur du réseau), ce qui se traduit par l'équation suivante

$$\sum_{j\in V(i)} q_{ij} = \sum_{j\in V(i)} k_{i,j}(u_i - u_j) = 0 \tag{2.8}$$

$$u = u^\star \text{ donné sur les noeuds sources } P^\star, P^{\star\star} \tag{2.9}$$

avec

$$V(i) = \text{ ensemble des voisins du noeud } P_i.$$

A titre d'exemple, $V(4) = \{1, 3, 2\}$ et $V(7) = \{3, 2, P^\star\}$. Le système (2.8) est linéaire. Pour alléger les notations, on prend ici $k_{i,j} = 1$, les équations du système linéaire sont :

$$\text{Eq. } 1 : (u_1 - u_9) + (u_1 - u_4) + (u_1 - u_3) = 3u_1 - u_9 - u_3 - u_4 = 0$$

$$\text{Eq. } 2 : 2u_2 - u_4 - u_7 = 0$$

...

$$\text{Eq. } 7 : u_7 - u_3 + u_7 - u^\star = 0 \implies 2u_7 - u_3 = u^\star$$

...

ensuite, il est aisé d'écrire ce système sous la forme $AU = F$. Noter que ce système est creux parce que un noeud dépend uniquement de ses voisins.

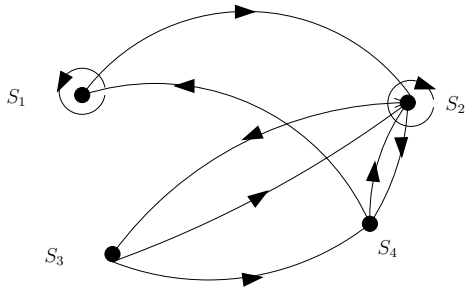
2.3. Graphe associé à une matrice et inversement

La numérotation des sommets d'un maillage ou d'un réseau modifie considérablement la structure creuse de la matrice associée. Il y a des techniques de renumérotation permettant de réduire la largeur de la bande ou du profil de la matrice. Ces techniques sont basées sur la notion de graphe.

Soit A une matrice carrée, $A = (a_{ij})$ d'ordre N. A chaque colonne de la matrice on fait correspondre un sommet S_i, $i = 1, N$.

Un **arc** relie S_i à S_j si $a_{ij} \neq 0$.
Un **graphe** est formé de l'ensemble des sommets et de arcs.
Exemple.

$$A = \begin{pmatrix} 4 & 3 & 0 & 0 \\ 0 & 2 & 1 & 2 \\ 0 & 1 & 0 & 3 \\ 6 & 5 & 0 & 0 \end{pmatrix}$$

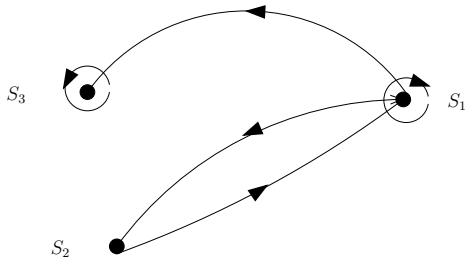


A chaque sommet, on peut associer l'ensemble de voisins :

$$V(S_i) = \{S_j,\ j \neq i, S_iS_j \text{ est un arc}\}$$

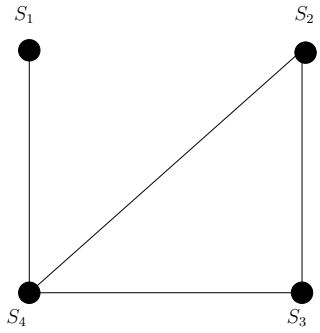
Un **chemin** allant de S_i à S_j est une suite d'arcs, si elle existe, telle que (S_i, S_{i_1}), $(S_{i_1}, S_{i_1})...(S_{i_p}, S_j)$ soient des arcs du graphe.
Un graphe est dit **fortement connexe** s'il existe au moins un chemin allant de tout sommet S_i à tout sommet S_j. Ainsi le graphe précédent est fortement connexe. Par contre, pour la matrice suivante

$$A = \begin{pmatrix} 3 & 2 & 5 \\ 4 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



le graphe n'est pas fortement connexe car il n'y a pas de chemin allant de S_3 à S_1.
Lorsque la matrice est symétrique ($a_{ij} \neq 0$ ssi $a_{ji} \neq 0$) on définit **l'arête** comme étant l'ensemble des deux arcs.

$$A = \begin{pmatrix} 3 & 0 & 0 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 1 & 4 & 3 \\ 1 & 2 & 3 & 0 \end{pmatrix}$$



2.4. Les matrices irréductibles

Soit A une matrice d'ordre N définie par blocs sous la forme suivante :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

avec A_{11} une matrice carrée d'ordre P et donc A_{22} une matrice d'ordre $N-P$. La résolution du système linéaire $Ax = b$ est équivalent à

$$\begin{cases} A_{22}x_2 = b_2 \\ A_{11}x_1 = b_1 - A_{12}x_2 \end{cases}$$

avec $x = (x_1, x_2)$, $x_1 \in \mathbb{R}^P$ et $x_2 \in \mathbb{R}^{N-P}$; de même $b = (b_1, b_2)$. Autrement dit, la résolution de ce système linéaire de taille N est **réduite** à la résolution de deux systèmes linéaires de tailles plus petites.

Définition 2.1. — *Une matrice d'ordre N est réductible ssi*
– il existe une matrice de permutation P telle que

$$B = P^t AP = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}$$

– soit encore, Il existe une partition de $\{1, 2, ..., N\}$ en deux ensembles d'indices I et J telle que $a_{i,j} = 0$ pour $i \in I$ et $j \in J$.
– soit encore, Il existe σ une permutation : $\{1, 2, ..., N\} \mapsto \{I, J\}$.

Exemple. Clairement la matrice

$$A = \begin{pmatrix} -1 & 2 & 3 & 4 \\ 5 & 6 & 10 & 11 \\ & & & \\ 0 & 0 & 11 & 13 \\ 0 & 0 & 20 & 30 \end{pmatrix}$$

est réductible. Soit maintenant la matrice

$$B = \begin{pmatrix} 30 & 0 & 20 & 0 \\ 11 & 10 & 6 & 5 \\ 13 & 0 & 11 & 0 \\ 4 & 2 & 3 & -1 \end{pmatrix}.$$

La matrice B est réductible. En effet, soit la permutation d'indice suivante et on note

i	1	2	3	4
σ	4	2	3	1

TABLE 1. Exemple de permutation.

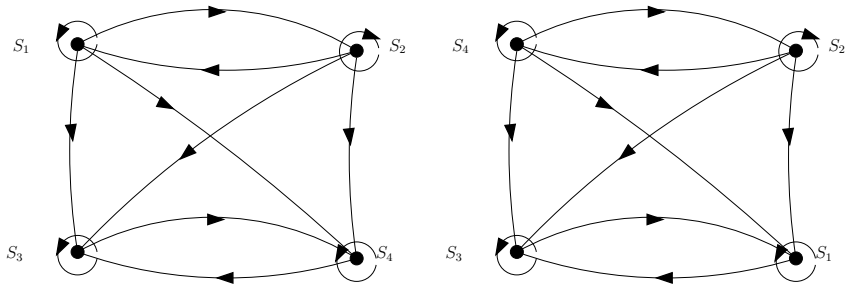


FIGURE 5. Graphe de A (à gauche) et graphe de B (à droite)

$\tilde{B} = B_\sigma$, ceci signifie que les coefficients $\tilde{b}_{i,j}$ de la matrice $\tilde{B}$ s'écrivent

$$\tilde{b}_{i,j} = b_{\sigma(i),\sigma(j)}.$$

On alors $\tilde{b}_{1,1} = b_{\sigma(1),\sigma(1)} = b_{4,4} = -1$; $\tilde{b}_{1,2} = b_{\sigma(1),\sigma(2)} = b_{4,2} = 2$; $\tilde{b}_{1,3} = b_{\sigma(1),\sigma(3)} = b_{4,3} = 3$, ainsi de suite ... on trouve finalement que $\tilde{B} = A$ qui est réductible.
Il est clair qu'il est difficile de trouver une permutation pour savoir si la matrice est réductible ou irréductible. Un moyen simple de le savoir est de regarder le graphe associé à la matrice. Les graphes associés à chaque matrice de l'exemple précédent sont représentés sur la figure 2.4.

Le graphe de A n'est pas fortement connexe car les sommets $\{S_3, S_4\}$ ne sont pas reliés aux sommets $\{S_3, S_4\}$. De même, le graphe de B n'est pas fortement connexe car les sommets $\{S_1, S_3\}$ sont indépendants des sommets $\{S_2, S_4\}$. Noter que le graphe de A est identique à celui de B quitte à renuméroter les sommets. Autrement dit, en renumérotant les sommets de la matrice B selon la permutation σ définie par le tableau 2.4 on obtient que la matrice B est réductible. On conclut que

Proposition 2.1. — *Une matrice est irréductible ssi son graphe est fortement connexe.*

2.5. Localisation des valeurs propres

Théorème 2.1. — *(Gerschgörin–Hadamard 1)*
Soit λ une valeur propre de A. Alors

$$\lambda \in \cup_{k=1}^{N} \overline{D}_k \text{ avec } \overline{D}_k = \{z \in \mathbb{C}^2, |z - a_{kk}| \leq \sum_{j=1, j\neq k}^{N} |a_{kj}|\}.$$

Les valeurs propres de A appartiennent à l'union des N disques $\overline{D}_k$ de Gerschgörin.

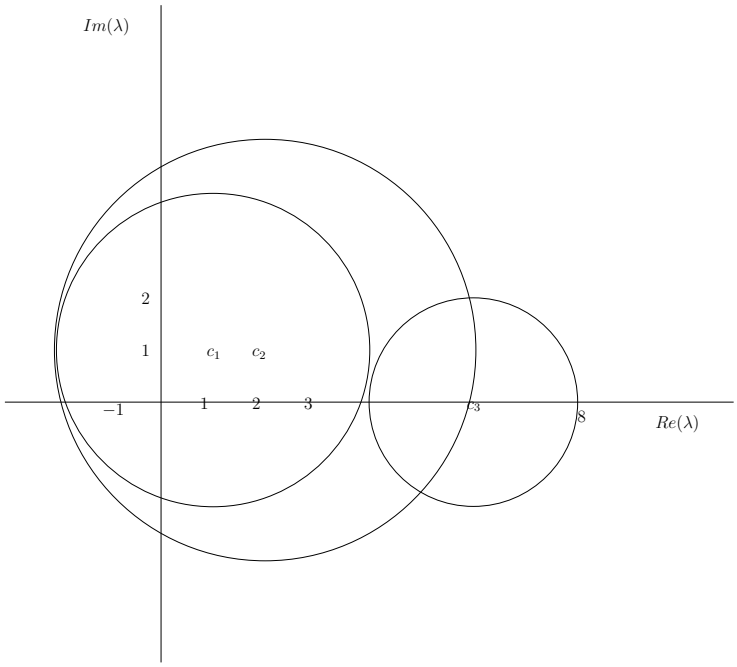


FIGURE 6. Disques de Gerschgörin.

Démonstration. — Soit u un vecteur propre associée à λ tel que $\max_i \|u_i\| = |u_k| = 1$. On a

$$Au = \lambda u \iff \sum_{i=1}^{N} a_{ij}u_j = \lambda u_j, \qquad \forall i.$$

En particulier,

$$(\lambda - a_{kk})u_k = \sum_{j\neq k} a_{kj}u_j,$$

on déduit,

$$|\lambda - a_{kk}||u_k| = |\lambda - a_{kk}| \leq \sum_{j\neq k} |a_{kj}||u_j| \leq \sum_{j\neq k} |a_{kj}| \text{ car } |u_j| \leq 1.$$

On ne connait pas k, mais on déduit $\lambda \in \cup_{k=1}^{N} \overline{D}_k$. □

Exemple. Soit

$$A = \begin{pmatrix} 1+i & i & 2 \\ -3 & 2+i & 1 \\ 1 & i & 6 \end{pmatrix}$$

Les disques de Gerschgörin sont $\lambda \in \mathbb{C}^2$ vérifiant : $|\lambda-(1+i)| \leq |i|+2 = 3$, $|\lambda-(2+i)| \leq 4$ et $|\lambda - 6| \leq 2$. On dessine dans le plan complexe ces trois disques (Fig. 2.5) et les valeurs propres sont situées dans l'union des trois disques.

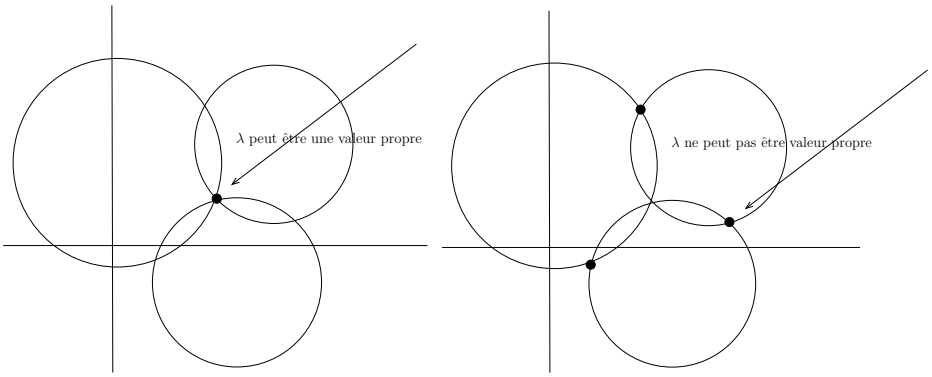


FIGURE 7. Localisation des valeurs propres sur la frontière

Théorème 2.2. — *(Gerschgörin–Hadamard 2)*
Soit A une matrice carrée et de graphe fortement connexe.
Si une valeur propre λ est située sur la frontière de la réunion des disques, alors tous les cercles de Gerschgörin passent par λ.

$$\lambda \in \partial\Big(\cup_{k=1}^{N} \overline{D}_k \Big) \implies \lambda \in \cap_{k=1}^{N} \Big(\partial \overline{D}_k \Big), \tag{2.10}$$

avec $\overline{D}_k = \{z \in \mathbb{C}^2, |z - a_{kk}| \leq \sum_{j=1, j\neq k}^{N} |a_{kj}|\}$.

Ce théorème sert essentiellement à montrer que λ n'est pas une valeur propre (voir figure 2.5).

Démonstration. — Soit λ une valeur propre et on suppose qu'elle n'est pas située à l'intérieur de la réunion des disques, alors il n'existe aucun k tel que

$$|\lambda - a_{kk}| \geq \Lambda_k, \tag{2.11}$$

où $\Lambda_k = \sum_{j\neq k} |a_{kj}|$. On pose $\max_k \|u_k\| = |u_i| = 1$, alors $|\lambda - a_{ii}| \leq \Lambda_i$ (car $\lambda \in Sp(A)$) et d'après (2.11), on a

$$|\lambda - a_{ii}| = \Lambda_i = \sum_{j\neq i} |a_{ij}|. \tag{2.12}$$

Soit $\mathcal{I} = \{i, u_i = 1, i = 1, N\}$. On a $I = \emptyset$, il contient au moins un indice. Donc, $\forall i \in I$, on a

$$\sum_{j\neq i} |a_{ij}||u_j| \geq |\sum_{j\neq i} a_{ij}u_j| = |(\lambda - a_{ii})u_i| = |\lambda - a_{ii}| = \Lambda_i = \sum_{j\neq i} |a_{ij}|,$$

on en déduit que

$$\sum_{j\neq i} |a_{ij}|(1 - |u_j|) \geq 0,$$

or $|u_j| \leq 1$, ce qui implique que

$$|a_{ij}|(1 - |u_j|) = 0, \forall j = 1, N.$$

Si $j \notin \mathcal{I}$, alors $|u_j| < 1$ et $a_{ij} = 0$ (avec $i \in \mathcal{I}$ et $j \notin \mathcal{I}$). On note $\mathcal{J}$ le complémentaire par rapport à $\{1, 2, \cdots, N\}$ de $\mathcal{I}$. Si $\mathcal{J} \neq \emptyset$, alors la partition $\mathcal{I}$, $\mathcal{J}$ serait telle que

$$\forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \text{ on a } a_{ij} = 0,$$

or la matrice est irréductible, donc cette partition est impossible et donc $J = \emptyset$. Autrement dot, $\forall i = 1, N$, on a $|u_i = 1|$ et $I = \{1, 2, \cdots N\}$, ce qui se traduit par (2.12) pour tout i, c'est à dire

$$|\lambda - a_{ii}| = \Lambda_i = \sum_{j \neq i} |a_{ij}|, \text{ pour tout } i = 1, N.$$

□

Définition 2.2. — *on dit que*
– *A est à diagonale strictement dominante si*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \ \forall i = 1, N.$$

– *A est à diagonale fortement dominante si*

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \ \forall i,$$

et il existe au moins un indice k tel que

$$|a_{kk}| > \sum_{j \neq k} |a_{kj}|.$$

Exercice. Montrer que :
Une matrice à diagonale strictement dominante est inversible.
Une matrice à diagonale fortement dominante et irréductible est inversible.

2.6. Méthodes directes pour la résolution de systèmes linéaires

On a vu que la discrétisation de systèmes modélisant des phénomènes physiques donne lieu à un système linéaire de grande taille et creux (la matrice contient beaucoup de zéro). Par exemple, pour une matrice tridiagonale de taille N, le nombre de coefficients non nuls est $3N$ à comparer avec N^2 le nombre des coefficients d'une matrice pleine ; pour $N = 3000$, il est facile de voir que 0.1% des coefficients sont utiles pour résoudre le système linéaire. Il est alors judicieux que le stockage de la matrice soit bien adapté à la structure de la matrice. Le seul critère pour résoudre un système linéaire de grande taille est la rapidité et la précision de la méthode utilisée. Ceci passe par une méthode de résolution adaptée à la nature et la structure de la matrice.

2.6.1. Méthode de Cramer. — La méthode de Gramer consiste à résoudre le système linéaire $Ax = b$ par la formule $x_i = detA_i/detA$, avec x_i est la ième composante du vecteur x et A_i est la matrice A où la ième colonne est remplacée par b.

La complexité de cette méthode c'est à dire le nombre d'opérations nécessaire pour calculer la solution est :

- N divisions
- $(N + 1)$ déterminants à calculer
- N $N!$ opérations pour calculer un déterminant

ainsi, la complexité vaut $(N + 1)N! + N$. A titre d'exemple, pour $N = 25$ la complexité est de l'ordre de 4×10^{26} opérations. Considérons maintenant un ordinateur avec 1G de ram c'est à dire qu'il effectue 10^9 opérations par seconde. Ensuite, calculons le nombre d'opérations que cet ordinateur effectue par milliard d'années :

$$10^9(\text{opérations/secondes}) \times 3600(\text{secondes/heure}) \times 24(\text{heures/jour}) \times 352(\text{jours/an}) \times 10^9(\text{an/milliard}) = 3 \times 10^{25}$$

et enfin le temps qu'il met cet ordinateur pour résoudre un système linéaire de taille 25×25 par la méthode de Cramer est

$$\frac{4 \times 10^{16}}{3 \times 10^{25}} > 10 \text{ milliards d'années}$$

ce qui est plus grand que l'âge de la terre !.

Inversion d'une matrice. Résoudre $Ax = b$ est équivaut à $x = A^{-1}b$, mais le coût pour calculer l'inverse d'une matrice par la formule $A^{-1} =^t Co(A)/detA$ est aussi de l'ordre de $(N+1)N!$. Ce que nous voulons est la résolution du système linéaire sans calculer l'inverse de cette matrice parce que dans tous les cas le calcul de A^{-1} est très coûteux.
Une méthode simple pour calculer l'inverse d'une matrice est de résoudre N systèmes linéaires. On note c_j les vecteurs colonnes de la matrice A^{-1} et A^{-1} est solution de $AA^{-1} = I$ ceci est équivalent à $Ac_j = e_j$ avec $(e_j)_j$ est la base canonique de $\mathbb{R}^N$.

2.6.2. Méthode de Gauss. — La méthode de base la plus utilisée pour la résolution des systèmes linéaires et la méthode d'élimination de Gauss. La méthode de Gauss consiste à déterminer une matrice P telle que le système équivalent $PAx = Pb$ soit triangulaire supérieure et donc simple à résoudre.

• La résolution d'un système triangulaire supérieure est aisée par la *procédure de remontée*. Soit le système $Uy = b$ avec

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & & u_{1n} \\ & u_{22} & u_{23} & \cdots & u_{2n} \\ & & & \cdots & \\ & & u_{ii} & u_{i,i+1} & u_{i,n} \\ & & & \cdots & \\ & & & u_{n-1,n-1} & u_{n-1,n} \\ & & & & u_{nn} \end{pmatrix}$$

l'algorithme de la remontée consiste à résoudre le système en commençant par la dernière équation :

$$u_{n,n}y_n = f_n \implies y_n = \frac{f_n}{u_{nn}}$$

$$u_{n-1,n-1}y_{n-1} + u_{n-1,n}y_n = f_{n-1} \implies y_{n-1} = \frac{f_{n-1} - u_{n-1,n}y_n}{u_{n-1,n-1}}$$

$$u_{ii}y_i + \sum_{j=i+1}^{N} u_{ij}y_j = f_i \implies y_i = \frac{f_i - \sum_{j=i+1}^{N} u_{ij}y_j}{u_{ii}}, \text{ pour } i = N, 1$$

```
Algorithme.
   Pour i=n à 1 par pas de -1 faire
       s=f(i)
       pour j=i+1 à n faire
            s=s-u(i,j)*y(j)
       finj
       y(i)=s/u(i,i)
   fini
```

Calculer la complexité de cet algorithme.

• Exemple d'élimination de la méthode de Gauss.

$$\begin{pmatrix} 4 & 2 & 3 & 0 \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ 2 & \frac{3}{2} & 4 & 1 \\ 0 & 2 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 9 \\ 2 \\ \frac{17}{2} \\ 8 \end{pmatrix}$$

ce qui est équivalent à résoudre :

$$\begin{cases} 4x_1+ & 2x_2+ & 3x_3+ & 0= & 9 \\ x_1+ & \frac{1}{2}x_2+ & 0+ & \frac{1}{2}x_4= & 2 \\ 2x_1+ & \frac{3}{2}x_2+ & 4x_3+ & x_4= & \frac{17}{2} \\ 0+ & 2x_2+ & x_3+ & 5x_4= & 8 \end{cases}$$

La méthode d'élimination de Gauss se fait en 3 étapes sur cet exemple.
Etape 1. Elimination de x_1 dans les les trois dernières lignes $L_2 \leftarrow L_2 - \frac{1}{4}L_1$ (i.e. L_2 est remplacé par $L_2 - \frac{1}{4}L_1$) et $L_3 \leftarrow L_3 - \frac{2}{4}L_1$. Le coefficient $a_{11} = 4$ est appelé le pivot ; ce qui donne

$$\begin{cases} 4x_1+ & 2x_2+ & 3x_3+ & 0= & 9 \\ 0+ & 0+ & -\frac{3}{4}x_3+ & \frac{1}{2}x_4= & -\frac{1}{4} \\ 0+ & \frac{1}{2}x_2+ & \frac{5}{2}x_3+ & x_4= & 4 \\ 0+ & 2x_2+ & x_3+ & 5x_4= & 8 \end{cases}$$

Etape 2. Elimination de x_2. On décrit ici la méthode sans faire de combinaison maligne des équations. Ici, il y a un problème car le pivot a_{22} est nul. On échange alors les lignes L_4 et L_2, ainsi :

$$\begin{cases} 4x_1+ & 2x_2+ & 3x_3+ & 0= & 9 \\ 0+ & 2x_2+ & x_3+ & 5x_4= & 8 \\ 0+ & \frac{1}{2}x_2+ & \frac{5}{2}x_3+ & x_4= & 4 \\ 0+ & 0+ & -\frac{3}{4}x_3+ & \frac{1}{2}x_4= & -\frac{1}{4} \end{cases}$$

et maintenant, on effectue l'opération : $L_3 \leftarrow L_3 - \frac{1}{4}L_2$

$$\begin{cases} 4x_1+ & 2x_2+ & 3x_3+ & 0= & 9 \\ 0+ & 2x_2+ & x_3+ & 5x_4= & 8 \\ 0+ & 0+ & \frac{9}{4}x_3- & \frac{1}{4}x_4= & 2 \\ 0+ & 0+ & -\frac{3}{4}x_3+ & \frac{1}{2}x_4= & -\frac{1}{4} \end{cases}$$

Etape 3. $L_4 \leftarrow L_4 + \frac{1}{3}L_3$

$$\begin{cases} 4x_1+ & 2x_2+ & 3x_3+ & 0= & 9 \\ 0+ & 2x_2+ & x_3+ & 5x_4= & 8 \\ 0+ & 0+ & \frac{9}{4}x_3- & \frac{1}{4}x_4= & 2 \\ 0+ & 0+ & 0+ & \frac{5}{12}x_4= & \frac{5}{12} \end{cases}$$

et une résolution directe de ce système par l'algorithme de remontée donne $x_1 = x_2 = x_3 = x_4 = 1$.

Algorithme général.

2.6.3. Méthode de Cholesky. — voir TD.

CHAPITRE 3

MÉTHODES ITÉRATIVES

Le principe de base de de telles méthodes est d'engendrer une suite de vecteurs x_k (les itérés) convergente vers la solution x du système linéaire $Ax = b$.

La plupart des méthodes itératives sont de la forme suivante :

Partant d'un vecteur arbitraire x_0, on engendre une suite $(x_k)_k$ définie par

$$x_{k+1} = Bx_k + c \tag{3.1}$$

avec B une matrice $M_{n\times n}(K)$, $c \in K^n$; avec $K = \mathbb{R}$ ou $\mathbb{C}$.

Définition 3.1. — *Une méthode itérative de la forme* (3.1) *est dite convergente si pour tout x_0, on a*

$$x_k \longrightarrow x, \text{ quand } k \to \infty$$

et la limite vérifie $Ax = b$. ($Ax = b$ est équivalent alors à $x = Bx + c$)

Définition 3.2. — *L'erreur d'approximation à la kième étape s'écrit*

$$e_k = x_k - x = Bx_{k-1} + c - Bx - c = B(x_{k-1} - x) = Be_{k-1}.$$

Et aussi

$$e_k = B^k e_0, \forall k \in \mathbb{N}.$$

Définition 3.3. — *Une méthode itérative est convergente si pour tout x_0, on a*

$$\lim_{k\to\infty} e_k = 0.$$

Ceci est équivalent à :

$\lim_{k\to\infty} B^k = 0 \iff \forall x \in K^n, \lim_{k\to\infty} B^k x = 0 \iff \lim_{k\to\infty} \|B\|^k = 0$ *pour toute norme matricielle.*

Théorème 3.1. — *(Convergence des méthodes itératives)*
On a

$$\lim_{k\to\infty} B^k = 0 \iff \rho(B) < 1.$$

Pour qu'une méthode itérative de la forme (3.1) *soit convergente il faut et il suffit que* $\rho(B) < 1$. *on rappelle que* $\rho(B)$ *désigne le rayon spectrale de la matrice* B ($\rho(B) = \max_i |\lambda_i(B)|$).

Démonstration. — ($\Longrightarrow$) si $\rho(B) \geq 1$, il existe λ, valeur propre de B , telle que $|\lambda| \geq 1$. Soit $x \neq 0$ vecteur propre associé à λ, alors $B^k x = \lambda^k x$ et comme $|\lambda^k| \to 1$ ou ∞ alors $B^k x$ ne converge pas vers zéro, ainsi $\lim_{k\to\infty} B^k \neq 0$.
($\Longleftarrow$) On suppose $\rho(B) < 1$, c'est à dire que $|\lambda_l| < 1$, $l = 1, r$ avec λ_l valeur propre de B. Toute matrice est semblable à une matrice de Jordan et on conclut. □

Corollaire 3.1. — *Si* $\|B\| < 1$ *alors* $\rho(B) < 1$ *et la méthode* (3.1) *converge.*

La preuve est immédiate car

$$\rho(B) \leq \|B\|.$$

En effet, soit $Bx = \lambda x$ alors $|\lambda|\|x\| = \|Bx\| \leq \|B\|\|x\|$, soit encore $|\lambda| \leq \|B\|$ pour tout $\lambda \in Sp(B)$.

Définition 3.4. — *On appelle taux asymptotique de convergence d'une méthode itérative le nombre* $R_\infty = -\log \rho(B)$. *Ce nombre est positif car* $\rho(B) < 1$.

Ce taux de convergence permet de mesurer le nombre d'itération nécessaire pour réduire l'erreur d'un certain facteur.

Proposition 3.1. — *Etant donné* $0 < \eta < 1$, *si le nombre d'itération* $k \geq \frac{-\log \eta}{R_\infty(B)}$ *alors*

$$\|e_k\| \leq \eta \|e_0\|.$$

Démonstration. — On a $e_k = B^k e_0$ et $\|e_k\| \leq \|B^k\|\|e_0\|$, $\forall k$. Pour avoir $\frac{\|e_k\|}{\|e_0\|} \leq \eta$, on impose $\|B^k\| \leq \eta$, soit $\|B^k\|^{\frac{1}{k}} \leq \eta^{\frac{1}{k}}$ alors $\log \|B^k\|^{\frac{1}{k}} \leq \frac{1}{k} \log \eta$, or $\|B^k\| < 1$ alors $k \geq \frac{\log \eta}{\log \|B^k\|^{\frac{1}{k}}} = \frac{-\log \eta}{-\log \|B^k\|^{\frac{1}{k}}}$. D'autre part,

$$\rho^k(B) = \rho(B^k) \leq \|B^k\|, \text{ alors } \rho(B) \leq \|B^k\|^{\frac{1}{k}},$$

ainsi $-\log \rho(B) = R_\infty(B) \geq -\log \|B^k\|^{\frac{1}{k}}$ et donc

$$\frac{1}{R_\infty(B)} \leq \frac{1}{-\log \|B^k\|^{\frac{1}{k}}}.$$

Enfin, on choisit alors k le nombre d'itérations :

$$k \geq \frac{-\log \eta}{R_\infty(B)})$$

pour réduire l'erreur initiale de η.

□

3.1. Méthodes itératives classiques

3.1.1. Méthode de Jacobi, Gauss-Seidel, relaxation. — Soit A une matrice d'ordre n telle que $a_{ii} \neq 0$, $i = 1, n$. On décompose A sous la forme

$$A = D - E - F$$

avec
D la diagonale de A
$-E$ la partie inférieure stricte
$-F$ la partie supérieure stricte.

• **Méthode de Jacobi.** Résoudre $Ax = b$ est équivalent à

$$Dx = (E + F)x + b.$$

La méthode de Jacobi est basée sur la décomposition précédente et elle s'écrit

$$\left\| \begin{array}{c} x_0 \text{ arbitraire} \\ Dx_{k+1} = (E + F)x_k + b \end{array} \right. \tag{3.2}$$

Il est facile à chaque itération de calculer x_{k+1} en fonction de x_k car la matrice diagonale D est inversible. Les composantes du vecteur x_{k+1} vérifient

$$a_{ii}(x_{k+1})_i = - \sum_{j=1, j\neq i}^{n} a_{ij}(x_k)_j + b_i,$$

soit encore

$$(x_{k+1})_i = \Big(- \sum_{j=1, j\neq i}^{n} a_{ij}(x_k)_j + b_i \Big) / a_{ii}.$$

Sous forme matricielle x_{k+1} s'écrit

$$x_{k+1} = D^{-1}(E + F)x_k + D^{-1}b = Jx_k + c, \tag{3.3}$$

la matrice $J = D^{-1}(E + F)$ est la matrice d'itération de Jacobi. La méthode de Jacobi converge ssi $\rho(J) < 1$.
Pour programmer cette méthode, on a besoin de stocker les vecteurs x_k et x_{k+1}.

• **Méthode de Gauss-Seidel.** Elle est basée sur cette décomposition :

$$Ax = b \iff (D - E)x = Fx + b$$

la méthode de Gauss-Seidel s'écrit

$$\left\| \begin{array}{c} x_0 \text{ arbitraire} \\ (D - E)x_{k+1} = Fx_k + b \end{array} \right. \tag{3.4}$$

$D - E$ est une matrice triangulaire inférieure et pour calculer x_{k+1} en fonction de x_k, il suffit d'appliquer l'algorithme de descente suivant :
pour $i = 1, n$

$$a_{ii}(x_{k+1})_i = -\sum_{j=1}^{i-1} a_{ij}(x_{k+1})_j - \sum_{j=i+1}^{n} a_{ij}(x_k)_j + b_i,$$

soit encore

$$(x_{k+1})_i = \Big(-\sum_{j=1}^{i-1} a_{ij}(x_{k+1})_j - \sum_{j=i+1}^{n} a_{ij}(x_k)_j + b_i \Big)/a_{ii}.$$

Noter que dans la boucle de calcul, les composantes du vecteur $(x_{k+1})_j$ pour $j = 1, i-1$ sont déjà calculés et on peut utiliser un seul vecteur pour programmer cette méthode. Sous forme matricielle x_{k+1} s'écrit

$$x_{k+1} = (D-E)^{-1}Fx_k + (D-E)^{-1}b = Gx_k + c, \tag{3.5}$$

la matrice $G = (D-E)^{-1}F$ est la matrice d'itération de Gauss-Seidel. La méthode de Gauss-Seidel converge ssi $\rho(G) < 1$.

• **Méthode de relaxation.** Elle est basée sur cette décomposition : Soit $\omega \neq 0$, la matrice A s'écrit $A = (\frac{D}{\omega} - E) + (D - \frac{D}{\omega} - F)$. Le système $Ax = b$ s'écrit $(\frac{D}{\omega} - E)x = (\frac{1-\omega}{\omega}D + F)x + b$. La méthode de relaxation s'écrit :

$$\left\| \begin{array}{c} x_0 \text{ arbitraire} \\ (\frac{D}{\omega} - E)x_{k+1} = (\frac{1-\omega}{\omega}D + F)x_k + b \end{array} \right. \tag{3.6}$$

soit encore

$$(D - \omega E)x_{k+1} = ((1-\omega)D + \omega F)x_k + \omega b,$$

La matrice d'itération s'écrit alors

$$\mathcal{L}_w = (D - \omega E)^{-}1((1\omega)D + \omega F).$$

Les composantes du vecteur x_{k+1} sont solutions de

$$\left\| \begin{array}{c} a_{ii}(x_{k+\frac{1}{2}})_i = -\sum_{j=1}^{i-1} a_{ij}(x_{k+1})_j - \sum_{j=i+1}^{n} a_{ij}(x_k)_j + b_i \\ (x_{k+1})_i = (x_k)_i + \omega((x_{k+\frac{1}{2}})_i - (x_k)_i). \end{array} \right. \tag{3.7}$$

Pour $\omega = 1$, c'est la méthode de Gauss-Sidel.

Théorème 3.2. —

1. *Soit A une matrice symétrique définie positive (ou hermitienne définie positive), alors la méthode de relaxation converge si $0 < \omega < 2$.*
2. *Soit A une matrice à diagonale strictement dominante ou à diagonale fortement dominante et irréductible) alors la méthode de Jacobi converge et la méthode de relaxation converge pour $0 < \omega \leq 1$.*

La démonstration de ce théorème est illustrée par des exemples.

3.2. Méthodes de gradients

Principe de la méthode. Soit A une matrice symétrique définie positive. Résoudre $A\overline{x} = b$ est équivalent à minimiser la fonctionnelle

$$J(x) =< Ax, x > -2 < b, x > .$$

La fonctionnelle J est quadratique et définie positive, elle admet un minimum global $\overline{x}$ solution de $J'(\overline{x}) = 2(A\overline{x} - b) = 0$. On note $e(x) = (\overline{x} - x)$ et $r(x) = b - Ax = A(\overline{x} - x)$ le résidu du système $Ax = b$. On définit l'énergie

$$E(x) =< A(\overline{x} - x), (\overline{x} - x) >=< Ae(x), e(x) >=< r(x, A^{-1}r(x) > .$$

Minimiser J est équivalent à minimiser E car

$$\begin{aligned} E(x) =& < Ax, x > -2 < A\overline{x}, x > + < A\overline{x}, \overline{x} > \\ =& < Ax, x > -2 < b, x > + < A\overline{x}, \overline{x} >= J(x) + < A\overline{x}, \overline{x} >, \end{aligned} \quad (3.8)$$

comme $< A\overline{x}, \overline{x} >$ est une constante, E et J atteignent leur minimum au point point.
Une méthode de descente est une méthode itérative sous la forme suivante

$$x_{k+1} = x_k + \alpha_k p_k, \qquad p_k \neq 0,$$

avec α_k le pas de descente,
p_k la direction de descente.
On choisit α_k et p_k de telle sorte que

$$E(x_{k+1}) < E(x_k).$$

Exemples de méthodes de descente :

•**Méthode de Richardson.** C'est la méthode de gradients à pas constant.

$$\alpha_k = \alpha > 0; \qquad p_k = r_k = b - Ax_k.$$

Les itérés vérifient

$$x_{k+1} = x_k + \alpha(b - Ax_k) = (I - \alpha A)x_k + \alpha b.$$

Théorème 3.3. — *Soit A une matrice symétrique définie positive de valeurs propres*

$$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N.$$

Alors la méthode de Richardson converge pour $0 < \alpha < \frac{2}{N}$. Le meilleur choix de α, notée α_{opt}, celui qui minimise $\rho(I - \alpha A)$ est donnée par $\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_N}$.

Démonstration. — Voir TD. □

•**Méthode de gradients à pas optimal.** On considère la direction de descente $p_k = r_k = b - Ax_k$. La méthode s'écrit

$$x_{k+1} = x_k + \alpha_k r_k = x_k + \alpha_k(b - Ax_k). \tag{3.9}$$

Le choix optimal de α_k consiste, à chaque itération, à choisir α_k pour minimiser l'énergie $E(x_{k+1})$ dans la direction r_k. Le paramètre α_k est choisi tel que

$$E(x_k + \alpha_k r_k) = E(x_{k+1}) = \min_{\alpha \in \mathbb{R}} E(x_k + \alpha r_k).$$

Calcul de α_k. On a $E(x) =< A(\overline{x} - x), \overline{x} - x >$, ainsi

$$E(x_k + \alpha r_k) =< A(x_k + \alpha r_k - \overline{x}), x_k + \alpha r_k - \overline{x} >= E(x_k) - 2\alpha < r_k, r_k > +\alpha^2 < Ar_k, r_k > .$$

C'est une équation de second degré en α et $< Ar_k, r_k >> 0$ car A est une matrice symétrique définie positive, donc le minimum est atteint par

$$\alpha_k = \frac{< r_k, r_k >}{< Ar_k, r_k >}.$$

Alors la méthode s'écrit

$$x_{k+1} = x_k + \frac{< r_k, r_k >}{< Ar_k, r_k >} r_k.$$

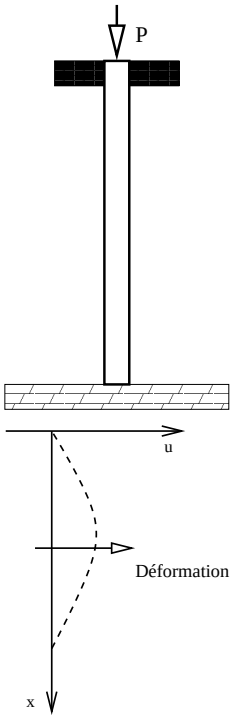
La méthode est équivalente à $x_{k+1} = (I - \alpha_k A)x_k + \alpha_k b$, pour montrer la convergence de la méthode, on ne peut pas appliquer le théorème 3.1 parce que la matrice d'itération $(I - \alpha_k A)$ dépend de k. On montre directement que $x_k \longrightarrow x = A^{-1}b$ (voir TA-2007 exo1).

•**Méthode des gradients conjugués.** On choisit α_k et la direction de descente p_k.

3.3. Calcul de valeurs propres et de vecteurs propres

La recherche de valeurs propres et de vecteurs propres est un problème qui intervient naturellement dans l'étude de la dynamique des structures. Par exemple
Exemple. Flambage d'une barre. Soit une barre de longueur l, fixée à une extrémité H. On

applique une force P vers le bas dans la direction de l'axe. Quand P est faible, pas de déformation de la barre. Quand P augmente, une valeur critique $\overline{P}$ est atteinte à partir de laquelle la barre se déforme.



Soit une barre de longueur l, fixée à une extrémité H. On applique une force P vers le bas dans la direction de l'axe. Quand P est faible, pas de déformation de la barre. Quand P augmente, une valeur critique $\overline{P}$ est atteinte à partir de laquelle la barre se déforme.

On note $u(x)$ le déplacement du point situé à l'abscisse perpendiculaire à l'axe de la barre.

Pour des petits déplacements, u vérifie

$$\begin{cases} \partial_x(a(x)\frac{du}{dx}) + Pu = 0 \\ u(0) = u(H) = 0 \end{cases} \tag{3.10}$$

$a(x)$ dépend des caractéristiques de la barre (section, module d'élasticité). Si $a = constante$, alors

$$\begin{cases} -u''(x) = Pu(x) \\ u(0) = u(H) = 0, \end{cases} \tag{3.11}$$

P est une valeur propre. Les solutions (3.12) sont $u_k(x) = \sin(\frac{k\pi x}{H}$, et $P_k = \frac{\pi^2 k^2}{H^2}$, $k \in \mathbb{N}^\star$. Pratiquement ici, on est intéressé par la plus petite valeur propre $P_1 = \pi^2/H^2$.
Si $a = a(x)$, on n'a pas de solution analytique, en général, de l'équation (11.9). On peut alors chercher une solution approchée en discrétisant le problème par exemple par la méthode des différences finies :

$$\begin{cases} \frac{1}{h^2}\Big(a_{i+\frac{1}{2}}(u_{i+1} - u_i) - a_{i-\frac{1}{2}}(u_i - u_i - 1)\Big) + Pu_i = 0, \qquad i = 1, N. \\ u_0 = u_{N+1} = 0. \end{cases} \tag{3.12}$$

Ce système est équivalent à $AU = \lambda U$ avec $\lambda = P$ la plus petite valeur propre et elle désigne la charge critique.

3.3.1. La méthode de la puissance itérée. — Elle permet le calcul d'une approximation de la valeur propre de plus grand module ainsi celle d'un vecteur propre associé.

Exemple. Soit $A = \begin{pmatrix} 10 & 0 \\ -9 & 1 \end{pmatrix}$ de valeurs propres 10 et 1 et de vecteurs propres $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ et $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Si on prend $x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \neq 0$ et que l'on calcule successivement $x_1 = Ax_0$, $x_2 = Ax_1$,...$x_{k+1} = Ax_k$ on obtient $x_1 = \begin{pmatrix} 10 \\ -8 \end{pmatrix}$; $x_2 = \begin{pmatrix} 100 \\ -98 \end{pmatrix}$; $x_3 = \begin{pmatrix} 1000 \\ -998 \end{pmatrix}$; .... Ce qui montre très vite que $x_{k+1} \approx 10x_k$ et x_{k+1} est colinéaire au vecteur propre $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Ce qui donne rapidement l'algorithme de la puissance itérée suivant

$$\begin{cases} q_0 \in \mathbb{C}^N \text{ tel que } \|q_0\| = 1 \\ \quad \text{pour } k = 1, 2, ... \\ \quad\quad x_k = Aq_{k-1} \\ \quad\quad q_k = x_k/\|x_k\|. \end{cases} \tag{3.13}$$

Moralement, ce qui se passe $Aq_{k-1} = \|x_k\|q_k$ on s'attend alors que $\|x_k\| \longrightarrow |\lambda_1|$ la plus grande valeur propre en module. Et si $\lambda_1 > 0$, alors $Aq_{k-1} \approx q_k$ et $q_k \longrightarrow u_1$ un vecteur propre associé à λ_1.

Théorème 3.4. — *Soit A une matrice d'ordre N, diagonalisable dont la valeur propre du plus grand module λ_1 est unique et simple*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|.$$

Soit $q_0 \in \mathbb{C}^N$ dont la composante selon le vecteur propre associé à λ_1 est non nulle. Alors la suite définie par (3.13) *vérifie*

i) $\lim\limits_{k\to\infty} \left(\dfrac{\overline{\lambda_1}}{|\lambda_1|}\right)^k q_k = q$ *est un vecteur propre à droite de norme* 1.

ii) $\lim\limits_{k\to\infty} \|Aq_k\| = \lim\limits_{k\to\infty} \|x_k\| = |\lambda_1|$.

iii) $\lim\limits_{k\to\infty} \dfrac{x_{k+1}(j)}{q_k(j)} = \lambda_1$, *pour* $1 \leq j \leq N$ *si* $q_k(j) \neq 0$.

Le facteur de convergence de toutes ces suites est $\dfrac{|\lambda_2|}{|\lambda_1|}$.

Démonstration. — A est diagonalisable, on note $(u_1, u_2, ..., u_N)$ une base de vecteurs propres de A associée aux λ_i, $i = 1, N$.

$$q_0 \in \mathbb{C}^N \implies q_0 = \sum_{i=1}^{N} \alpha_i u_i,$$

le choix convenable de q_0 est de choisir $\alpha_1 \neq 0$, et on écrite

$$q_0 = \alpha_1 u_1 + \sum_{i=2}^{N} \alpha_i u_i, \text{ et } \alpha_1 \neq 0,$$

ceci signifie que q_0 n'est pas orthogonal au sous espace propre à gauche associé à λ_1.

i) On calcule tout d'abord q_k en fonction de q_0. On a

$$q_1 = \frac{x_1}{\|x_1\|} = \frac{Aq_0}{\|Aq_0\|},$$

$$q_2 = \frac{Aq_1}{\|Aq_1\|} = A\Big(\frac{Aq_0}{\|Aq_0\|}\Big)\Big/\Big\|A\frac{Aq_0}{\|Aq_0\|}\Big\| = \frac{A^2q_0}{\|A^2q_0\|}.$$

Par récurrence

$$q_k = \frac{A^kq_0}{\|A^kq_0\|}. \tag{3.14}$$

D'autre part

$$A^kq_0 = A^k\Big(\sum_{i=1}^N \alpha_iu_i\Big) = \sum_{i=1}^N \alpha_i\lambda_i^ku_i = \alpha_1\lambda_1^k\Big[u_1 + \sum_{i=2}^N \frac{\alpha_i}{\alpha_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_i\Big] = \alpha_1\lambda_1^k(u_1+e_k), \tag{3.15}$$

avec $e_k = \displaystyle\sum_{i=2}^N \frac{\alpha_i}{\alpha_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_i$; or $\dfrac{|\lambda_i|}{|\lambda_1|} < 1$ pour $i = 2, N$ donc $\Big(\frac{\lambda_i}{\lambda_1}\Big)^k \longrightarrow 0$ quand $k \to \infty$ et pour tout $i = 2, N$, alors

$$e_k \longrightarrow 0 \text{ quand } k \to \infty, \tag{3.16}$$

de plus $e_k \approx \Big(\frac{\lambda_2}{\lambda_1}\Big)^k$ pour k grand. D'après la formule de récurrence (3.14) on a

$$q_k = \frac{A^kq_0}{\|A^kq_0\|} = \frac{\alpha_1\lambda_1^k(u_1+e_k)}{|\alpha_1||\lambda_1|^k\|u_1+e_k\|}, \tag{3.17}$$

ainsi

$$\frac{|\lambda_1|^k}{\lambda_1^k}q_k = \frac{\alpha_1(u_1+e_k)}{|\alpha_1|\|u_1+e_k\|} \longrightarrow \frac{\alpha_1u_1}{|\alpha_1|\|u_1\|} = q \text{ proportionnel à } u_1. \tag{3.18}$$

ii) on a

$$Aq_k = \frac{\alpha_1\lambda_1^k(Au_1+Ae_k)}{|\alpha_1||\lambda_1|^k\|u_1+e_k\|}, \tag{3.19}$$

et

$$\|Aq_k\| = \frac{\|Au_1+Ae_k\|}{\|u_1+e_k\|} \longrightarrow \frac{\|Au_1\|}{\|u_1\|} = |\lambda_1|. \tag{3.20}$$

iii)

$$\frac{x_{k+1}(j)}{q_k(j)} = \frac{Aq_k(j)}{q_k(j)} = \Big(A\frac{A^kq_0}{\|A^kq_0\|}\Big)(j)\Big/\Big(\frac{A^kq_0}{\|A^kq_0\|}\Big)(j).$$

D'après (3.15), on a

$$\frac{x_{k+1}(j)}{q_k(j)} = \frac{\alpha_1\lambda_1^{k+1}(u_1(j)+e_{k+1}(j))}{\alpha_1\lambda_1^k(u_1(j)+e_k(j))} = \lambda_1\frac{(u_1(j)+e_{k+1}(j))}{(u_1(j)+e_k(j))} \longrightarrow \lambda_1.$$

□

Proposition 3.2. —
– La méthode de la puissance itérée converge si A est diagonalisable et la valeur propre de plus grand module est unique et de multiplicité $p > 1$.
– Si la matrice n'est pas diagonalisable mais la valeur propre dominante est unique alors la méthode converge.

Démonstration. — On a $\lambda_1 = \lambda_2 = ... = \lambda_p$ et $|\lambda_1| > \lambda_{p+1}$. On reprend la même preuve que précédemment, il vient

$$A^k q_0 = \lambda_1^k \Big[(\sum_{i=1}^{p} \alpha_i u_i) + \sum_{i=p+1}^{N} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_i \Big].$$

On pose $u = \sum_{i=1}^{p} \alpha_i u_i \neq 0$, donc

$$A^k q_0 = \lambda_1^k (u + e_k) \approx \lambda_1^k u,$$

ainsi

$$\frac{|\lambda_1|^k}{\lambda_1^k} q_k \longrightarrow q \text{ vecteur propre associé à } \lambda_1.$$

□

3.3.2. Méthode de la puissance inverse. — Elle permet la recherche de la plus petite valeur propre en module. Il suffit d'appliquer la méthode de la puissance itérée à la matrice A^{-1}. Les valeurs propres de A^{-1} sont $\mu_i = \frac{1}{\lambda_i}$ où λ_i est valeur propre de A. On suppose

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_{N-1}| > |\lambda_{N-1}|.$$

On a

$$\max_i |\mu_i| = \frac{1}{\min_i |\lambda_i|} = \frac{1}{|\lambda_N|} = |\mu_N|.$$

La méthode de la puissance inverse s'écrit :

$$\begin{cases} q_0 \in \mathbb{C}^N \text{ tel que } \|q_0\| = 1, \quad q_0 = \sum_{i=1}^{N} \alpha_i u_i, \quad \alpha_N \neq 0 \\ \text{pour } k = 1, 2, ... \\ x_k = A^{-1} q_{k-1} \\ q_k = x_k / \|x_k\|. \end{cases} \tag{3.21}$$

Le vecteur x_k est déterminé par la résolution du système linéaire

$$A x_k = q_{k-1}.$$

On peut alors décomposer la matrice A sur la forme $A = LU$ et résoudre à chaque itération le système linéaire de façon simple. On a évidemment $\lim_{k\to\infty} \|A^{-1} q_k\| = \lim_{k\to\infty} \|x_k\| = |\mu_N|$.

Application : Recherche de la valeur propre la plus proche d'un nombre donné. Soit $\tilde{\lambda}$ donné. Soit λ une valeur propre de A telle que $\tilde{\lambda}$ soit la plus proche de λ, autrement dit $\tilde{\lambda} - \lambda$ représente la valeur propre de plus petit module de $A - \tilde{\lambda}$:

$$\tilde{\lambda} \neq \lambda, \qquad |\tilde{\lambda} - \lambda| < |\tilde{\lambda} - \mu|, \qquad \forall \mu \in Sp(A)\{\lambda\}.$$

On applique alors l'algorithme de la puissance inverse à $A - \tilde{\lambda}$:

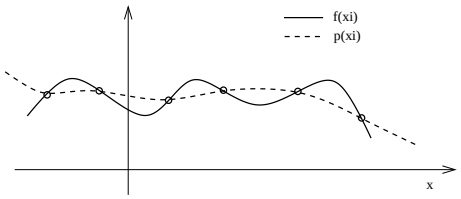
$$\left\{ \begin{array}{c} q_0 \in \mathbb{C}^N \text{ tel que } \|q_0\| = 1 \\ (A - \tilde{\lambda})x_k = q_{k-1} \\ q_k = x_k / \|x_k\|. \end{array} \right. \tag{3.22}$$

CHAPITRE 4

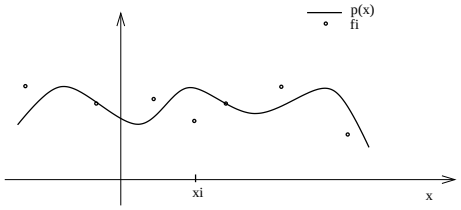
INTERPOLATION ET APPROXIMATION

4.1. Introduction

• On se donne un ensemble de points (x_i, f_i) obtenus suite à une mesure expérimentale (f_i représente la température, pression, débit, ...) pour connaître la valeur de la fonction mesurée en d'autres points dans le domaine, on peut alors représenter la fonction f par un polynôme.



On cherche P un polynôme tel que $P(x_i) = f(x_i)$. Un tel polynôme interpole la fonction mesurée aux points des mesures x_i



On cherche P un polynôme le plus proche des valeurs mesurées. L'approximation au sens des moindre carré consiste à p tel que

$$\sum_i |p(x_i) - f_i|^2 \text{ soit minimal .}$$

• On cherche à calculer une intégrale dont on ne connaît pas explicitement sa valeur. Par exemple, on approche cette comme suit

$$f(x) \approx p(x) et \int f(x)\,dx \approx \int p(x)\,dx \text{ (facile à calculer)}$$

ce qui conduit à l'intégration numérique.

• f solution d'une e.d.o

• f solution d'une équation non linéaire de la forme $f = G(f)$

• Soit f une fonction inconnue solution d'un problème aux limites (équation de la chaleur par exemple), on cherche à approcher au mieux les valeurs de f en certains points du domaine.

4.2. Interpolation de Lagrange

Soient $(n+1)$ couples : (x_0, f_0), (x_1, f_1) ... (x_n, f_n), tels que les x_i sont distincts. On cherche un polynôme P tel que

$$P(x_i) = f_i \text{ pour } i = 0, 1, ..., n.$$

Le polynôme passe par les points de mesure.

Théorème 4.1. — *Il existe un unique* $P \in \mathbb{P}_n = \{$ *polynômes de degrés* $n\}$ *tel que*

$$P(x_i) = f_i \text{ pour } i = 0, 1, ..., n.$$

Démonstration. —
Unicité. Soient $p, q \in \mathbb{P}_n$ tels que $P(x_i) = Q(x_i) = f_i$ pour $i = 0, ...n$, alors $p - q \in \mathbb{P}_n$ et il s'annule en $(n+1)$ points distincts alors $p - q \equiv 0$.
Existence. Base de polynômes de Lagrange.
Soit

$$L_i \in \mathbb{P}_n \text{ tel que } L_i(x_j) = \delta_{ij},\ i = 0, ..., n,$$

alors $\{L_i\}_{i=0,n}$ est une base de $\mathbb{P}_n$ (famille libre).
Construction de $L_i(x)$. On a $L_i(x_j) = 0$ pour $j \neq i$ donc $x - x_j$ divise le polynôme

$$L_i(x) = \lambda \prod_{j=0, j\neq i}^{n} (x - x_j) \in \mathbb{P}_n \implies \lambda \in \mathbb{R},$$

λ est calculé par $L_i(x_i) = 1$ ce qui donne $\lambda = \dfrac{1}{\prod_{j=0,j\neq i}^{n}(x_i - x_j)}$. Les polynômes de Lagrange sont

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j},\ i = 0, n. \tag{4.1}$$

□

Théorème 4.2. — *(Erreur d'interpolation de Lagrange)*
Soit $f \in \mathcal{C}^{n+1}([a, b])$ *et* $a \leq x_0 < x_1 < x_2 < ... < x_n \leq b$. *Soit* p *le polynôme de Lagrange définit par*

$$p(x_i) = f(x_i) \text{ pour } i = 0, 1, ..., n.$$

Alors

$$f(x) - p(x) = \frac{L(x)}{(n+1)!} f^{(n+1)}(\xi), \tag{4.2}$$

avec $L(x) = \prod_{j=0}^{n}(x - x_j)$, $a \leq \min(x_0, x) < \xi < \max(x, x_n) \leq b$.

Démonstration. — Si $x = x_i$ alors $f(x_i) = p(x_i)$ et $L(x) = 0$ ce qui établit (4.2).
soit $x \neq x_i$, $i = 0, n$. Considérons la fonction w définie par :

$$w(t) = f(t) - p(t) - L(t)k(x)$$

avec la fonction $k(x)$ est donnée tel que $w(x) = 0$, soit encore $k(x) = \frac{f(x)-p(x)}{L(x)}$. On a

$$w(x) = 0, w(x_i) = 0,\ i = 0, n$$

w s'annule en $(n + 2)$ points distincts et d'après le théorème de Rolle
w' s'annule en $(n + 1)$ points distincts, et donc
w'' s'annule en n points distincts, ...
$w^{(n+1)}$ s'annule en 1 point ; Il existe $\xi \in]a, b[$ tel que $w^{(n+1)}(\xi) = 0$.

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

or $p^{(n+1)}(\xi) = 0$ car $p \in \mathbb{P}_n$, ce qui donne

$$k(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!} = \frac{f(x) - p(x)}{L(x)}$$

ce qui établit (4.2). □

En majorant l'erreur d'interpolation, on a

$$|f(x) - p(x)| \leq \frac{1}{(n + 1)!} \max_{x\in[a,b]} |f^{(n+1)}(\xi)| \max_{x\in[a,b]} |L(x)|. \tag{4.3}$$

L'erreur d'interpolation résulte de deux termes : le premier terme $\max\limits_{x\in[a,b]} |f^{(n+1)}(x)|$ dépend de f et on ne peut pas l'améliorer car f est donnée, par contre le deuxième terme $\max\limits_{x\in[a,b]} |L(x)|$ dépend de la distribution des points x_i. On peut choisir l'ensemble des points $\{x_i\}_{i=0,n}$ pour que l'erreur $\max\limits_{x\in[a,b]} |L(x)|$ soit minimal.

4.2.1. Meilleur choix des points x_i. — Le polynôme $L(x) = (x - x_0)(x - x_1)...(x - x_{n+1})$ est un polynôme de degré $(n + 1)$ dont le coefficient de x^{n+1} est 1. Le meilleur choix de $\{x_i\}_{i=0,n}$ est alors les racines du polynôme $L(x)$ vérifiant

$$\max_{x\in[a,b]} |L(x)| \leq \max_{x\in[a,b]} |q(x)|, \quad \forall q \in \mathbb{P}_{n+1} \text{ et } q(x) = x^{n+1} + a_n x^n + ... \tag{4.4}$$

Nous allons voir que les polynômes de Tchebychev répondent à cette question. Les polynômes de Tchebychev sont définis par

$$T_n(x) = \cos(n \arccos(x)), \qquad x \in [-1, 1], n \geq 0.$$

Vérifions d'abord que $T_n(x) \in \mathbb{P}_n$. On a $T_0 = \cos(0) = 1$ et $T_1(x) = x$ et on a la relation de récurrence

$$T_{n+1} = 2xT_n(x) - T_{n-1} = 2^n x^{n+1} + ...$$

car en posant $\theta = Arccos(x)$

$$T_{n+1} = \cos((n+1)\theta) = \cos(n\theta)\cos(\theta) - \sin(n\theta)\sin(\theta)$$
$$= \cos(n\theta)\cos(\theta) - \frac{1}{2}(\cos((n-1)\theta) - cos((n+1)\theta) = xT_n(x) - \frac{1}{2}(T_{n-1}(x) - T_{n+1}).$$

Les racines de T_n sont : $T_n(x) = cos(n \arccos(x)) = 0$, alors $n \arccos(x) = \frac{\pi}{2} + k\pi$, ainsi $\arccos(x) = \frac{\pi}{2n} + \frac{k\pi}{n}$ pour $k = 0 \cdots n-1$, ce qui donne que les racines de T_n sont :

$$x_k = \cos(\frac{2k+1}{2n}\pi);\ k = 0 \cdots n-1.$$

Les extremas de T_n sont : $T_n(x) = \cos(n \arccos(x)) = \pm 1$, alors $\arccos(x) = \frac{k\pi}{n}$ ce qui donne les extremas :

$$y_k = \cos\frac{k\pi}{n}; \qquad T_n(y_k) = (-1)^k,\ k = 0 \cdots n.$$

Théorème 4.3. — *Les racines des polynômes de Tchebychev satisfont*

$$\max_{x\in[-1,1]} |T_n(x)| \leq \max_{x\in[-1,1]} |q(x)|, \quad \forall q \in \mathbb{P}_n \text{ et } q(x) = 2^{n-1}x^n + a_{n-1}x^{n-1} + ... \tag{4.5}$$

Démonstration. — On va montrer ce résultat par l'absurde. Soit $q(x) = 2^{n-1}x^n + a_{n-1}x^{n-1} + ... \neq T_n(x)$ et on suppose

$$\max_{x\in[-1,1]} |q(x)| \leq \max_{x\in[-1,1]} |T_n(x)|. \tag{4.6}$$

Sur chaque intervalle $[\cos(\frac{k\pi}{n}, \cos(\frac{(k+1)\pi}{n}]$, $k = 0 \cdots n-1$, T_n passe du maximum au minimum ou inversement. On pose $d(x) = q(x) - T_n(x) \neq 0$, et donc $d \in \mathbb{P}_{n-1}$. De plus q est continue et vérifie la relation 4.6, alors sur chaque intervalle $[\cos(\frac{k\pi}{n}), \cos(\frac{(k+1)\pi}{n}])$, le graphe de q intersecte au moins une fois le graphe de T_n, c'est à dire $d(x)$ s'annule au moins une fois dans cet intervalle. Alors le polynôme d s'annule n fois et comme d est un polynôme de degré $(n-1)$ alors $d = 0$, ce qui contredit que $d \neq 0$. □

Ainsi, on a montré que parmi tous les polynômes de degré n s'écrivant sous la forme $q(x) = 2^{n-1}x^n + a_{n-1}x^{n-1} + ...$, le polynôme de Tchenychev est celui qui réalise le minimum pour la norme infinie, c'est à dire

$$\|T_n\|_{\mathcal{C}^0[-1,1]} < \|q\|_{\mathcal{C}^0[-1,1]}, \forall q \in \mathbb{P}_n, q(x) = 2^{n-1}x^n + \cdots .$$

Autrement dit,

$$\max_{x\in[-1,1]} |(x-x_0)(x-x_1)\cdots(x-x_n)| \text{ est minimal}$$

si et seulement si

$$T_{n+1}(x) = 2^n(x-x_0)(x-x_1)\cdots(x-x_n) \text{ avec } x_k = (\cos\frac{2k+1}{2(n+1)}\pi);\ k = 0 \cdots n.$$

Ou encore,
pour toute distribution de points d'interpolation $(z_0, z_1, \cdots, z_n)$, alors

$$\max_{x\in[-1,1]} |(x-x_0)(x-x_1)\cdots(x-x_n)| \leq \max_{x\in[-1,1]} |(x-z_0)(x-z_1)\cdots(x-z_n)|,$$

où les x_k sont les racines de T_{n+1}.

Enfin, sur un intervalle quelconque $[a,b]$, les racines du polynôme de Tchebychev sont définies comme suit

$$\hat{T}_n : [a,b] \xrightarrow{\phi} [-1,1] \xrightarrow{T_n} \mathbb{R}$$

où $\phi(x) = \frac{2x}{b-a} - \frac{b+a}{b-a}$.
Les polynômes de Tchebychev sur un intervalle quelconque $[a,b]$ s'écrivent :

$$\hat{T}_n = (T_n \circ \phi)(x) = T_n(\phi(x)) = \cos(n \arccos \phi(x))$$

et leurs racines sont : $\phi(x_k) = (\cos\frac{(2k+1)\pi}{2n})$; $k = 0\cdots n-1$ et donc $\phi(x_k) = \frac{2x_k}{b-a} - \frac{b+a}{b-a} = \cos(\frac{(2k+1)\pi}{2n})$, ainsi $x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos(\frac{(2k+1)\pi}{2n})$, pour $k = 0\cdots n-1$.

Remarque 4.1. — *i)On a* $T_{n+1}(x) = 2^n x^{n+1} + \cdots = 2^n(x-x_0)(x-x_1)\cdots(x-x_n)$ *et comme* $\max_{x\in[-1,1]} |T_{n+1}(x)| = 1$ *et donc* $\max_{x\in[-1,1]} |(x-x_0)(x-x_1)\cdots(x-x_n)| = \frac{1}{2^n}$, *ainsi l'erreur d'interpolation s'écrit :*

$$|f(x) - p(x)| \leq \frac{1}{(n+1)!}\frac{1}{2^n} \max_{x\in[-1,1]} |f^{(n+1)}(x)|.$$

ii) Les racines de Tchebychev, elles sont plus denses aux extrémités. Cette distribution a pour effet de réduire le phénomène de Runge (les effets sur le bord).
Exemple (TP) : comparer pour $f(x) = e^{-x^2}$ *sur* $[-5,5]$ *en prenant* 10 *points équidistants et les 10 racines de* $T_{10}(x)$.

4.3. Polynôme d'interpolation de Newton

Une autre façon de construire $p \in \mathbb{P}_n$ tel que $p(x_i) = f_i$ est d'utiliser la formule de Taylor et d'introduire les différences divisées.
En effet, par la formule de Taylor, on écrit

$$p(x) = p(x_0) + (x-x_0)Q_0(x) \text{ avec } Q_0 \in \mathbb{P}_{n-1},$$

or $p(x_0) = f_0$ ainsi

$$p(x) = f_0 + (x-x_0)Q_0(x),$$

Ensuite pour que $p(x_1) = f_1$, alors $Q_0(x_1) = \frac{f_1-f_0}{x_1-x_0}$ est connu. On applique à nouveau La formule de Taylor à $Q_0(x)$

$$Q_0(x) = Q_0(x_1) + (x-x_1)Q_1(x) \text{ avec } Q_1 \in \mathbb{P}_{n-2},$$

soit encore

$$p(x) = f_0 + (x-x_0)Q_0(x_1) + (x-x_0)(x-x_1)Q_1(x),$$

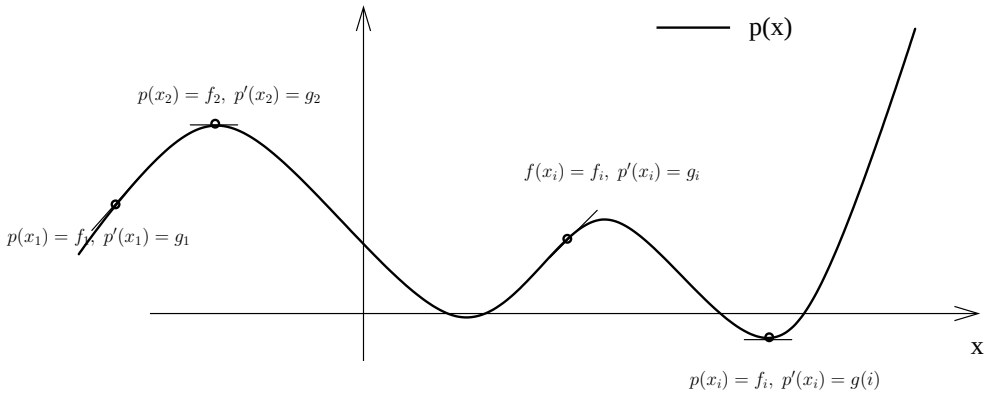


FIGURE 1. Interpolation de Hermite

Pour assurer $p(x_2) = f_2$, on impose alors

$$Q_1(x_2) = \frac{f_2 - f_0 - (x_2 - x_0)Q_0(x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

on continue le procédé en faisant le développement de Taylor de $Q_1(x)$ au point x_2. Les $Q_i(x_{i+1})$ sont appelés les différences divisées.

4.4. Interpolation de Hermite

On cherche un polynôme qui interpole la fonction f ainsi que sa dérivée aux points donnés. précisément, soient les $(n+1)$ triplet (x_i, f_i, g_i) pour $i = 0, n$. On chercher un polynôme p tel que

$$\begin{cases} p(x_i) = f_i, & i = 0, n \\ p'(x_i) = g_i, & i = 0, n \end{cases} \tag{4.7}$$

Théorème 4.4. — *Il existe un unique* $P \in \mathbb{P}_{2n+1} = \{$ *polynômes de degrés 2n+1*$\}$ *satisfaisant* (4.7).

Démonstration. —
Unicité. Soient $p, q \in \mathbb{P}_{2n+1}$ tels que $p(x_i) = q(x_i) = f_i$ et $p'(x_i) = q'(x_i) = g_i$ pour $i = 0, ...n$, alors $r = p - q \in \mathbb{P}_{2n+1}$ et $r(x_i) = r'(x_i) = 0$, alors $(x - x_i)^2$ divise le polynôme r, ainsi $r = c(x - x_0)^2...(x - x_n)^2 \in \mathbb{P}_{2(n+1)}$, or $r \in \mathbb{P}_{2n+1}$ alors $c = 0$ et $r \equiv 0$.
Existence. Base de polynômes de Hermite.
On cherche une base de polynômes de $\mathbb{P}_{2n+1}$ telle que

$$p(x) = \sum_{i=0}^{n} f_i A_i(x) + \sum_{i=0}^{n} g_i B_i(x)$$

Les conditions sur les fonctions de bases sont alors les suivantes :

$$A_i(x_j) = \delta_{ij},\ B_i(x_j) = 0 \text{ pour } i = 0, n$$

$$A_i'(x_j) = 0,\ B_i'(x_j) = \delta_{ij} \text{ pour } i = 0, n$$

les premières conditions permettent d'imposer $p(x_i) = f_i$ et les secondes $p'(x_i) = g_i$. Ces conditions permettent de construire les fonctions de bases. En effet,
Construction des polynômes A_i. On a $A_i(x_j) = A_i'(x_j) = 0$ pour $j \neq i$ alors $(x - x_j)^2$ divise A_i pour $j \neq i$, alors $A_i(x) = r(x) \prod_{j=0,\, j\neq i}^{n} (x - x_j)^2$ où $r(x) \in \mathcal{P}_{N1}$. On peut exprimer ce polynôme en fonction du polynôme de Lagrange. En effet, $L_i(x) = \prod_{j=0,\, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$, ainsi $A_i(x) = q(x)L_i^2(x)$, où $q(x) = ax + b \in P_1$. Les coefficients a et b sont tels que

$$A_i(x_i) = 1 \text{ et } A_i'(x_i) = 0.$$

On a

$$A_i(x_i) = (ax_i + b)L_i^2(x_i) = 1 = ax_i + b \text{ car } L_i(x_i) = 1.$$

$$A_i(x_i) = aL_i^2(x_i) + 2L_i(x_i)L_i'(x_i)(ax_i + b) = a + 2L_i'(x_i)(ax_i + b) = a + 2L_i'(x_i) = 0,$$

ainsi $a = -2L_i'(x_i)$ et $b = 1 - ax_i$, enfin

$$A_i(x) = (1 - 2(x - x_i)L_i'(x_i))L_i^2(x_i).$$

Calcul de B_i. Pour $j \neq i$, on a $B_i(x_j) = B_i'(x_j) = 0$, alors L_i^2 divise B_i, d'autre part $B_i(x_i) = 0$, alors $(x - x_i)$ divise aussi B_i. On déduit que $B_i(x) = c(x - x_i)L_i^2(x)$ et $c \in \mathbb{R}$. On détermine la constante par la relation $B_i'(x_i) = 1$; on a

$$B_i'(x_i) = cL_i^2(x_i) + 2c(x_i - x_i)L_i(x_i)L_i'(x_i) = c = 1,$$

ce qui donne

$$B_i(x) = (x - x_i)L_i^2(x).$$

□

Théorème 4.5. — *(Erreur d'interpolation de Hermite)*
Soit $f \in \mathcal{C}^{2n+2}([a, b])$ et $a \leq x_0 < x_1 < x_2 < ... < x_n \leq b$. Soit $p \in \mathbb{P}_{2n+1}$ le polynôme d'interpolation de Hermite définit par

$$\begin{cases} p(x_i) = f(x_i), & i = 0, n \\ p'(x_i) = f'(x_i), & i = 0, n. \end{cases} \tag{4.8}$$

Alors

$$f(x) - p(x) = \frac{L(x)}{(2n + 2)!} f^{(2n+2)}(\xi), \tag{4.9}$$

avec $L(x) = \prod_{j=0}^{n} (x - x_j)^2$, $a \leq \min(x_0, x) < \xi < \max(x, x_n) \leq b$.

La preuve est semblable à celle proposée pour le théorème 4.2.

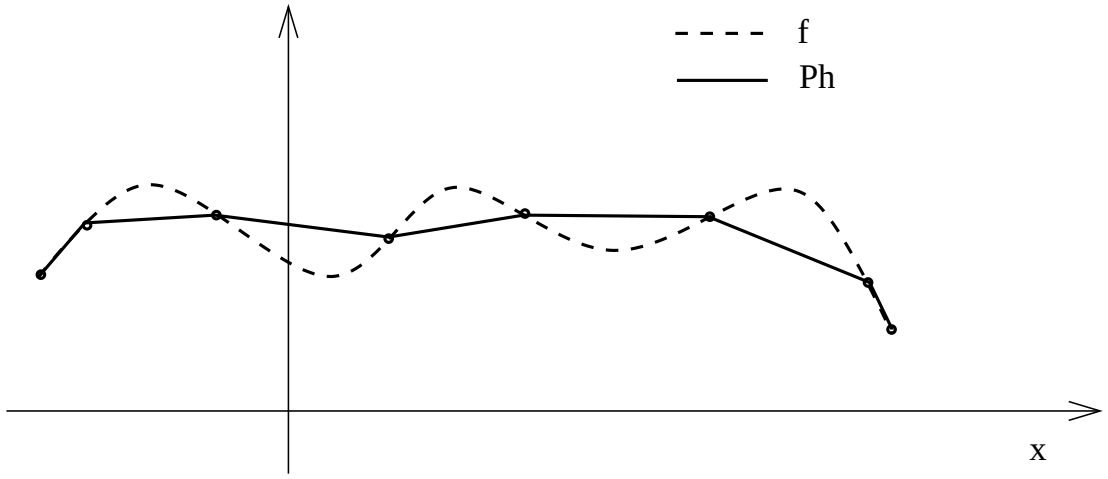


FIGURE 2. Interplation locale par des $\mathbb{P}_1$ par morceaux(gauche),par des $\mathbb{P}_2$ par morceaux(droite)

4.5. Interpolation locale

Le théorème d'interpolation 4.2 montre que l'approximation d'une fonction par des polynômes nécessite que la fonction soit régulière et le degré du polynôme soit élevé pour avoir la convergence lorsque le degré du polynôme tend vers l'infinie. L'interpolation locale consiste à interpoler la fonction sur des intervalles de petite taille par des polynômes de faible degré. On va décrire cette méthode.

Soit $[a, b]$ un compact de $\mathbb{R}$. On divise l'intervalle $[a, b]$ en M-intervalles de pas $h = (b - a)/M$. On pose $a_i = a + ih$, $i = 0, M$. Ensuite sur chaque intervalle $[a_i, a_{i+1}]$ on construit un polynôme d'interpolation $p_i(x)$ de degré m **fixe** aux points d'interpolation $(x_i)_k = a_i + \frac{kh}{m}$, $k = 0, m$. Les points $(x_i)_k$ sont situés dans l'intervalle $[a_i, a_{i+1}]$.

On considère le polynôme d'interpolation définit par morceaux comme suit

$$q_h(x) = p_i(x), \text{ pour } a_i \leq x \leq a_{i+1}. \tag{4.10}$$

Le polynôme q_h est continue mais non dérivable aux points a_i.

Théorème 4.6. — *(Erreur d'interpolation locale et convergence)*
Soit $f \in \mathcal{C}^{m+1}([a, b])$, alors

$$\forall x \in [a, b], |f(x) - q_h(x)| \leq \frac{1}{m!} \sup_{x \in [a,b]} |f^{(m+1)}(x)| h^{m+1}. \tag{4.11}$$

Démonstration. — Soit $x \in [a, b]$, il existe $i \in \{0, 1, ..., M - 1\}$ tel que $a_i \leq x \leq a_{i+1}$. D'après le théorème d'interpolation de Lagrange

$$|f(x) - p_i(x)| = |f(x) - q_h(x)| \leq \frac{\Pi_i(x)}{(m + 1)!} |f^{(m+1)}(\xi)|, \xi \in]a_i, a_{i+1}[$$

avec

$$\Pi_i(x) = |\prod_{k=0}^{m}(x-(x_i)_k)| \leq (m+1)(a_{i+1}-a_i)^{m+1} \leq (m+1)h^{m+1}$$

ce qui établit (4.11). □

Dans le théorème 4.6, l'entier m est fixe (en général $m = 1, 2$ ou 3) et on regarde la convergence par rapport à h quand $h \to 0$, on a directement

$$|f(x) - q_h(x)| \longrightarrow 0 \text{ quand } h \to 0.$$

4.6. Meilleure approximation (projection orthogonale)

Soit V un espace vectoriel muni d'un produit scalaire noté $((\cdot,\cdot))$ et $\|\cdot\|$ la norme associée. Soit V_N un sous espace de V de dimension finie. On note $\{q_1, q_2, ..., q_N\}$ une base de V_N. On dit que $u_N \in V_N$ réalise la meilleure approximation de $f \in V$ au sens suivant

$$\|u_N - f\| = \min_{v_N \in V_N} \|v_N - f\|. \tag{4.12}$$

On a $u_N = \sum_{j=0}^{N} \lambda_j^\star q_j$, avec $\lambda_j^\star \in \mathbb{R}$. Le problème (4.12) est alors équivalent à

$$\text{chercher } \{\lambda_j^\star\}_{j=1,N} \text{ réalisant le } \min_{\lambda \in \mathbb{R}^N} \|\sum_{j=0}^{N} \lambda_j q_j - f\|^2, \tag{4.13}$$

avec $\lambda = (\lambda_1, ..., \lambda_N)$. On note

$$J(\lambda) = \|\sum_{j=0}^{N} \lambda_j q_j - f\|^2.$$

Le minimum de cette fonction est caractérisé par :

$$\frac{\partial J}{\partial \lambda_k}(\lambda^\star) = 0, \text{ pour } k = 1, n,$$

car J est quadratique en λ et $J(\lambda) \longrightarrow +\infty$ quand $|\lambda| \to +\infty$.
En développant J, on a

$$J(\lambda) = \sum_{i,j=1}^{N} \lambda_i \lambda_j ((q_j, q_j)) - 2\sum_{i=0}^{N} \lambda_i ((q_i, f)) + \|f\|^2.$$

On désigne par A la matrice de coefficients $a_{i,j}$ avec $a_{i,j} = ((q_j, q_j)) \in \mathbb{R}$. La fonctionnelle J s'écrit

$$J(\lambda) = < A\lambda, \lambda > -2 < b, \lambda > + \|f\|^2,$$

avec $< \cdot, \cdot >$ désigne le produit scalaire dans $\mathbb{R}^N$.
Ainsi $\lambda^\star$ est caractérisé par

$$J'(\lambda^\star) = 0 \iff 2(A\lambda^\star - b) = 0,$$

$\lambda^\star$ est solution du système linéaire

$$A\lambda^\star = b.$$

La matrice A est symétrique définie positive.
En effet, soit $x \in \mathbb{R}^N$

$$< Ax, x >= \sum_{i=1}^{N}(Ax)_i x_i = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}x_j x_i = \sum_{i=1}^{N}\sum_{j=1}^{N}((q_j, q_i))x_j x_i$$
$$= ((\sum_{i=1}^{N} q_i x_i, \sum_{j=1}^{N} q_j x_j)) = \|X\|^2 \geq 0 \quad (4.14)$$

avec $X = \sum_{i=1}^{N} q_i x_i$. De plus pour $x \neq 0$, alors $X \neq 0$ et donc $< Ax, x >$ est strictement positive.

Nous allons voir que a structure de la matrice A dépend fortement de la base q_i, $i = 1, n$ sur plusieurs exemples.

Exemple 1. Soit $V = L^2(0,1)$ muni du produit scalaire $((f,g)) = \int_0^1 f(x)g(x)\,dx$. On divise l'intervalle $[0,1]$ en N-intervalles de pas $h = 1/N$. On pose $M_i =]ih, (i+1)h[$ pour $i = 1, ..., N-1$. Soit $V_N = \{v \in V;\ v|M_i = constante\}$. On a dimension $V_N = N =$ nombre d'intervalles. On propose la base suivante

$$q_j(x) = \begin{cases} 1 & \text{si } x \in M_i \\ 0 & \text{sinon} \end{cases}$$

Les coefficients de la matrice A sont

$$a_{i,j} = ((q_i, q_j)) = \int_0^1 q_i(x)q_j(x)\,dx = \begin{cases} 0 & \text{si } i \neq j \\ \int_{ih}^{(i+1)h} |q_i(x)|^2\,dx = \int_{ih}^{(i+1)h} dx = h \end{cases}$$

ainsi $A = h\,I$ (I la matrice identité) une matrice diagonale et facile à inverser.

Exemple 2. $V = H^1(0,1)$ base éléments finis $\mathbb{P}_1$.

Exemple 3. Soit $V = \mathcal{C}^0([-1,1])$ ou $V = L^2(-1,1)$ muni du produit scalaire $((f,g)) = \int_{-1}^1 f(x)g(x)\,dx$.
On considère $V_N = \mathbb{P}_N$ ensemble de polynômes de degré $\leq N$, et la base canonique de $\mathbb{P}_N =< 1, x, x^2, ..., x^N >$. Les coefficients de la matrice de projection orthogonale

$$a_{i,j} = \int_{-1}^1 q_i(x)q_j(x)\,dx = \int_{-1}^1 x^{i+j}\,dx \begin{cases} 0 & \text{si } i+j \text{ est impair} \\ \frac{2}{i+j+1} & \text{si } i+j \text{ est pair} \end{cases}$$

La matrice A est difficile à inverser et cette matrice n'est pas creuse.

D'après l'exemple précédent, il est alors intéressant de construire des polynômes orthogonaux associés au produit scalaire désiré.

4.7. Polynômes orthogonaux

Soit $]a,b[\in \mathbb{R}$ borné ou non. Soit un poids $\omega :]a,b[\longrightarrow \mathbb{R}^+$continue. On suppose $\forall n \in \mathbb{N}$, $\int_a^b |x|^2 w(x)\,dx$ est convergente. Soit $E = \mathcal{C}^0(]a,b[)$ muni du produit scalaire $< f,g >_\omega = \int_a^b f(x)g(x)\omega(x)\,dx$ et $\|\cdot\|_\omega$ la norme associée.

Définition 4.1. — *On appelle polynôme unitaire un polynôme dont le coefficient du plus haut degré est 1, i.e.* $p_n(x) = x^n + a_{n-1}x^{n-1} + ... + a_0$.

Théorème 4.7. — *Il existe une suite de polynômes unitaires* $(p_n)_n \in \mathbb{N}$, $deg(p_n) = n$, *orthogonaux 2 à 2 pour le produit scalaire de* E *associé au poids* ω. *Cette suite est unique.*

Démonstration. — Par récurrence selon le procédé d'orthogonalisation de Schmidt.
On a $p_0(x) = 1$ car p_0 est unitaire.
Supposons $p_0, p_1, ...p_{n-1}$ déjà construits, alors $< p_0, p_1, ...p_{n-1} >$ forme une base de $\mathbb{P}_{n-1}$.
Soit $p_n \in \mathbb{P}_n$ unitaire, alors

$$p_n(x) = x^n + \sum_{i=0}^{n-1} \lambda_j p_j(x).$$

On a $< p_n, p_k >_\omega = 0$, pour tout $k = 0, ..., n$, donc $< x^n, p_k >_\omega + \lambda_k \|p_k\|_\omega = 0$ et donc $\lambda_k = -\frac{<x^n, p_k>_\omega}{\|p_k\|_\omega}$. On a alors déterminé p_n de façon unique car le choix des λ_k est unique. □

Théorème 4.8. — *Formule de récurrence pour construire les polynômes orthogonaux) Les polynômes* p_n *du théorème 4.7 vérifient la relation de récurrence :*

$$p_n(x) = (x - \lambda_n)p_{n-1}(x) - \mu_n p_{n-2}(x),$$

avec $\lambda_n = \frac{<xp_{n-1}, p_{n-1}>_\omega}{\|p_{n-1}\|_\omega}$; $\mu_n = \frac{\|p_{n-1}\|_\omega^2}{\|p_{n-2}\|_\omega^2}$.

Exemples de polynômes orthogonaux.
Polynômes de Tchebychev. $]a,b[=]-1,1[$, $\omega(x) = \frac{1}{\sqrt{1-x^2}}$, $T_n(x) = \cos(n \arccos(x)) = cos(n\theta)$, $\theta \in [0,\pi]$. on vérifie que

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{1}{\sqrt{1-x^2}}\,dx = \int_0^\pi T_n(\cos\theta)T_m(\cos\theta)\,d\theta = \begin{cases} 0 \text{ si } n \neq m \\ \pi/2 \text{ si } n = m \neq 0 \\ \pi \text{ si } n = m = 0. \end{cases}$$

Polynômes de Legendre. $]a,b[=]-1,1[$, $\omega(x)=1$,

$$p_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n.$$

4.8. Approximation au sens des moindres carrés discrets

On dispose d'une suite de données (ou mesures) expérimentales (x_i,y_i), $i=1,n$. On cherche un polynôme $p\in\mathbb{P}_m$ avec $n\geq(m+1)$ qui réalise la meilleure approximation au sens suivant :

$$\sum_{i=1}^{n}|q(x_i)-y_i|^2 \text{ soit minimal .} \tag{4.15}$$

Autrement dit le polynôme recherché p vérifie

$$\sum_{i=1}^{n}|p(x_i)-y_i|^2=\min_{q\in\mathbb{P}_m}\sum_{i=1}^{n}|q(x_i)-y_i|^2 \text{ soit minimal .} \tag{4.16}$$

L'avantage de cette méthode est de pouvoir prendre des polynômes de petit degré, $n=2,\,3,\cdot$ Soit $q(x)=a_0+a_1x^1+a_2x^2+...+a_mx^m$, on cherche alors à minimiser la fonctionnelle $J:\mathbb{R}^{m+1}\mapsto\mathbb{R}^+$ définit par

$$J(a)=\sum_{i=1}^{n}|a_0+a_1x_i^1+a_2x_i^2+...+a_mx_i^m-y_i|^2$$

avec $a=(a_0,a_1,...,a_m)$. Le minimum est réalisé lorsque $J'(a)=0 \iff \forall k=1,m,\quad \frac{\partial J}{\partial a_k}(a)=0$. On a, pour $k=1,m$

$$\frac{\partial J}{\partial a_k}(a)=2\sum_{i=1}^{n}(a_0+a_1x_i^1+a_2x_i^2+...+a_mx_i^m-y_i)x_i^k=0. \tag{4.17}$$

On note $S_p=\sum_{i=1}^{n}x_i^p$ et $v_k=\sum_{i=1}^{n}y_ix_i^k$, donc le système (4.17) est équivalent à

$$a_0S_k+a_1S_{k+1}+a_2S_{k+2}+...+a_mS_{k+m}=v_k;\quad \forall k=1,m.$$

ce système est équivalent à

$$\begin{bmatrix} S_0 & S_1 & ... & S_m \\ S_1 & S_2 & ... & S_{m+1} \\ ... & ... & ... & ... \\ S_m & S_{m+1} & ... & S_{2m} \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{bmatrix}=\begin{bmatrix} v_0 \\ v_1 \\ ... \\ v_m \end{bmatrix}, \tag{4.18}$$

soit encore

$$Sa=v.$$

La fonctionnelle J peut s'écrire également sous la forme

$$J(a) = \|Aa - y\|^2 = \sum_{i=1}^{n} |(Aa)_i - y_i|^2,$$

la norme ici est celle associée au produit scalaire dans $\mathbb{R}^{m+1}$, et la matrice A est donnée comme suit

$$\begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^m \\ 1 & x_2 & x_2^2 & ... & x_2^m \\ ... & ... & ... & ... & ... \\ 1 & x_n & x_n^2 & ... & x_n^m \end{bmatrix} \in \mathcal{M}_{n\times(m+1)}, \tag{4.19}$$

$a \in \mathbb{R}^{m+1}$ et $y = (y_0, y_1, ..., y_n) \in \mathbb{R}^n$.
On va établir une relation entre la matrice S et la matrice A. En effet,

$$J(a) =< Aa - y, Aa - y >,$$

soit $h \in \mathbb{R}^{m+1}$, alors la différentielle de J s'écrit

$$J'(a)h =< Ah, Aa - y > + < Aa - y, Ah >= 2 <^t A(Aa - y), h >,$$

ainsi $J'(a) =^t A(Aa - y)$. Le minimum de J vérifie alors

$$^tAAa =^t Ay \tag{4.20}$$

ce qui équivalent à $Sa = v$ et alors

$$S =^t AA.$$

La matrice S est inversible car S est symétrique définie positive. En effet,

$$< Su, u >= \|Au\|^2 \geq 0,$$

si $\|Au\| = 0$ alors $u_0 + u_1x_j + u_2x_j^2 + ... + u_mx_j^m = 0$ pour $j = 1, n$; et en considérant le polynôme $p(x) = u_0 + u_1x + u_2x^2 + ... + u_mx^m$, les relations précédentes montrent que $p(x_j) = 0$ et donc le polynôme p admet n-racines et comme $n \geq (m + 1)$ et p de degré m alors $p \equiv 0 \iff u = 0$.

Finalement, on vérifie que a la solution de (4.20) est bien un minimum de J. La fonctionnelle J est quadratique et d'après la formule de Taylor, on a

$$J(a + h) = J(a) + J'(a)h + \frac{1}{2} < J''(a)h, h >,$$

de plus $J''(a) = 2^tAA$ une matrice symétrique définie positive et comme $J'(a) = 0$, on déduit que

$$J(a + h) > J(a), \text{ pour tout } h \neq 0,$$

ce qui prouve que a est un (le) minimum global de J.

CHAPITRE 5

INTÉGRATION NUMÉRIQUE

Il s'agit d'approcher $I = \int_a^b f(x)\,dx$ dans le cas où on ne connaît pas une primitive de f.
Il y a deux façons de présenter le calcul.
– On approche I par une formule de quadrature globale, c'est à dire on remplace f par un polynôme d'interpolation : $\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx$ avec $p \in \mathbb{P}_n$. On a déjà vu que le polynôme d'interpolation génère des effets indésirables sur le bord : effet de Runge. Intuitivement, cette méthode ne converge pas en général et n'est pas une une bonne approximation.
– On approche f par des polynômes par morceaux (interpolation locale). On a déjà vu que l'interpolation converge et nécessite peu de régularité sur la fonction f. Cette méthode est appelé méthode composite.

5.1. Méthode composite

On découpe l'intervalle $[a, b]$ en M mailles de longueur $h = \frac{b-a}{M}$ et on pose $x_i = a + ih$, pour $i = 0, 1, 2, \cdots, M$. On a $x_0 = a$ et $x_M = b$. On écrit alors :

$$\int_a^b f(x)\,dx = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} f(x)\,dx. \tag{5.1}$$

Ensuite, il suffit de construire une méthode d'intégration sur chaque intervalle $[x_i, x_{i+1}]$. On peut se ramener à une intégrale sur $[0, 1]$ ou $[-1, 1]$ (appelé intervalle de référence). Par exemple, soit

$$x = x_i + \frac{1+t}{2}h \in [x_i, x_{i+1}] \text{ et } t \in [-1, 1],$$

ainsi

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{h}{2}\int_{-1}^{1} f(x_i + \frac{1+t}{2}h)\,dt = \frac{h}{2}\int_{-1}^{1} g(t)\,dt \tag{5.2}$$

avec $g(t) = f(x_i + \frac{1+t}{2}h)$.

5.2. Formulation de quadrature de type interpolation

Soit $I(g) = \int_\alpha^\beta g(t)\,dt$.
On appelle **formule de quadrature** à $(n+1)$ points une formule du type suivant

$$I_n(g) = \sum_{j=0}^{n} c_j g(t_j), \tag{5.3}$$

avec le points t_j sont données ou à calculer dans $[\alpha, \beta]$ et les coefficients c_j sont indépendants de la fonction g.
On définit l'erreur d'intégration de g

$$R_n(g) = I(g) - I_n(g). \tag{5.4}$$

On dit qu'une **formule de quadrature a un degré de précision** k si

$$\left\| \begin{array}{ll} \forall p \in \mathbb{P}_k, & R_n(p) = 0 \text{ (exacte pour les } \mathbb{P}_k) \\ \exists q \in \mathbb{P}_{k+1}, & R_n(q) \neq 0 \text{ (inexacte pour les } \mathbb{P}_{k+1} \end{array} \right. \tag{5.5}$$

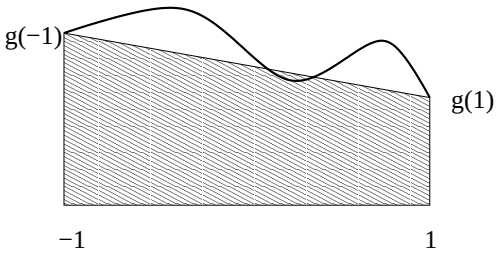
C'est équivalent à

$$\left\| \begin{array}{c} R_n(x^k) = 0 \text{ pour } i = 0, 1, ..., k \\ R_n(x^{k+1}) \neq 0 \end{array} \right. \tag{5.6}$$

Dans la formule de quadrature (5.3), il faut déterminer c_j et éventuellement t_j pour que la méthode soit exacte pour les polynômes de plus haut degré.

5.3. Formule d'intégration classique

5.3.1. La formule des trapèzes. —



L'intégrale est remplacée par l'aire du trapèze.

$$\int_{-1}^{1} g(t)\,dt \approx \alpha g(-1) + \beta g(1). \tag{5.7}$$

On cherche α, β pour que la méthode soit exacte sur les polynômes de plus haut degré.

On considère alors g un polynôme et la formule (5.7) est exacte. Il vient
pour $g(t) = 1$, $\int_{-1}^{1} dt = 2 = \alpha + \beta$,
pour $g(t) = t$, $\int_{-1}^{1} t\,dt = 0 = -\alpha + \beta$,
ainsi $\alpha = \beta = 1$ et la méthode (5.7) est complètement déterminée comme suit

$$\int_{-1}^{1} g(t)\,dt \approx g(-1) + g(1). \tag{5.8}$$

et de degré de précision au moins 1. D'autre part, pour $g(t) = t^2$, on a $\int_{-1}^{1} t^2\,dt = 2/3$ et $g(-1) + g(1) = 2$ et donc la méthode n'est pas exacte pour $\mathbb{P}_2$ et le degré de précision est exactement 1.

Erreur d'intégration. Soit $R(g) = \int_{-1}^{1} g(t)\,dt - (g(-1) + g(1))$. Soit p un polynôme d'interpolation de degré 1 tel que $p(-1) = g(-1)$ et $p(1) = g(1)$. D'après le théorème d'interpolation, on a

$$g(t) = p(t) + \frac{(t+1)(t-1)}{2} g''(\xi(t)), \qquad -1 < \xi(t) < 1,$$

et par intégration

$$\int_{-1}^{1} g(t)\,dt = \int_{-1}^{1} p(t)\,dt + \int_{-1}^{1} \frac{(t+1)(t-1)}{2} g''(\xi(t))\,dt,$$

or $p \in \mathbb{P}_1$, alors $\int_{-1}^{1} p(t)\,dt = p(1) + p(1) = g(-1) + g(1)$ et $R(g) = \int_{-1}^{1} \frac{(t+1)(t-1)}{2} g''(\xi(t))\,dt$. Ainsi

$$|R(g)| \leq \frac{1}{2}\Big(\int_{-1}^{1} (1 - t^2)\,dt \Big) \sup_{s\in[-1,1]} |g''(s)| = \frac{2}{3} \sup_{s\in[-1,1]} |g''(s)|.$$

Formule composite par la méthode des trapèzes. Revenons à (5.2) et en utilisant la formule des trapèzes, on a

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{h}{2} \int_{-1}^{1} f(x_i + \frac{1+t}{2}h)\,dt = \frac{h}{2} \int_{-1}^{1} g(t)\,dt \approx \frac{h}{2}(g(-1)+g(1)) = \frac{h}{2}(f(x_i)+f(x_{i+1}). \tag{5.9}$$

L'erreur d'intégration sur $[x_i, x_{i+1}]$ s'écrit :

$$R_i(f) = \int_{x_i}^{x_{i+1}} f(x)\,dx - \frac{h}{2}(f(x_i) + f(x_{i+1}),$$

soit encore

$$R_i(f) = \frac{h}{2}\Big(\int_{-1}^{1} g(t)\,dt - (g(-1) + g(1)) \Big) = \frac{h}{2} R(g)$$

On a $g(t) = f(x_i + \frac{1+t}{2}h)$, $g'(t) = \frac{h}{2} f'(x_i + \frac{1+t}{2}h)$ et $g''(t) = \frac{h^2}{4} f''(x_i + \frac{1+t}{2}h)$. Par conséquent

$$|R_i(f)| = \frac{h}{2}|R(g)| \leq \frac{h}{3} \sup_{s\in[-1,1]} |g''(s)| \leq \frac{h^3}{12} \sup_{\xi\in[x_i,x_{i+1}]} |f''(\xi)|. \tag{5.10}$$

La formule composite sur l'intervalle $[a, b]$ s'écrit

$$\int_a^b f(x)\,dx = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} f(x)\,dx \approx \frac{h}{2} \sum_{i=0}^{M-1} (f(x_i) + f(x_{i+1})). \tag{5.11}$$

L'erreur d'intégration composite sur $[a, b]$ s'écrit

$$R_h(f) = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} f(x)\,dx - \frac{h}{2} \sum_{i=0}^{M-1} (f(x_i) + f(x_{i+1})) = \sum_{i=0}^{M-1} R_i(f).$$

Ainsi, de (5.10), on déduit

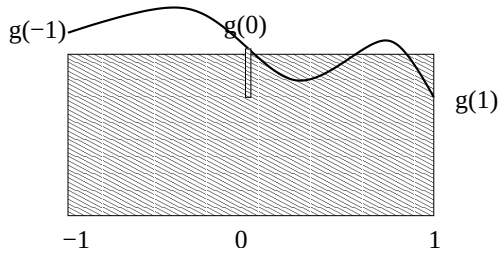
$$|R_h(f)| \leq \sum_{i=0}^{M-1} |R_i(f)| \leq \frac{h^3}{12} \sum_{i=0}^{M-1} \sup_{\xi \in [x_i, x_{i+1}]} |f''(\xi)| \leq \frac{h^3}{12} \sup_{x \in [a,b]} |f''(x)| M,$$

M dépend de h et il s'écrit $M = \frac{b-a}{h}$, ce qui donne que l'erreur d'intégration est en h^2 comme suit

$$|R_h(f)| \leq \Big(\frac{b-a}{12} \sup_{x \in [a,b]} |f''(x)| \Big) h^2.$$

On a démontré alors que si $f \in \mathcal{C}^2([a,b])$, alors la méthode d'intégration composite converge car $R_h(f) \longrightarrow 0$ quand $h \to 0$.

5.3.2. Méthode du point milieu. —



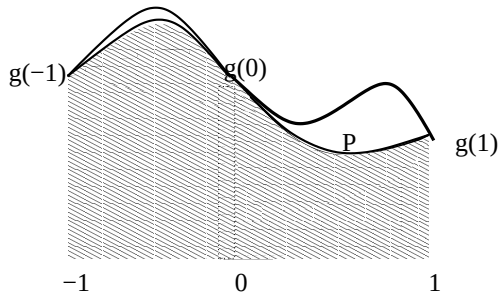
L'intégrale est remplacée par l'aire du rectangle de hauteur $g(0)$.

$$\int_{-1}^{1} g(t)\, dt \approx 2g(0) \tag{5.12}$$

Le degré de précision de la méthode est 1. Sur un intervalle de longueur h la méthode s'écrit

$$\int_{x_i}^{x_{i+1}} f(x)\, dx \approx hf(\frac{x_i + x_{i+1}}{2}) \tag{5.13}$$

5.3.3. Méthode de Simpson. —



L'intégrale est remplacée par l'aire du polynôme d'interpolation de degré deux passant par $g(-1)$, $g(0)$ et $g(1)$,

$$\int_{-1}^{1} g(t)\, dt \approx \alpha g(-1) + \beta g(0) + \gamma g(1) \tag{5.14}$$

les coefficients α, β et γ sont calculés pour que la méthode soit exacte pour $\mathbb{P}_2$, ce qui donne

$$\int_{-1}^{1} g(t)\, dt \approx \frac{1}{3} g(-1) + \frac{4}{3} g(0) + \frac{1}{3} g(1). \tag{5.15}$$

le degré de précision est 3.

5.4. Les formule de Gauss

Dans la formule de quadrature suivante :

$$\int_a^b \mu(x)f(x)\,dx \approx \sum_{i=0}^{n} c_i f(x_i), \text{ avec } \mu(x) \text{ est un poids,}$$

on désire améliorer les résultats en déterminant au 'mieux' les points $\{x_i\}_{i=0,n}$. On cherche alors les coefficients c_i et les points x_i pour que la méthode soit exacte pour des polynômes de plus haut degré.

Théorème 5.1. — *La formule de quadrature à* $(n+1)$ *points est exacte sur l'espace* $\mathbb{P}_{2n+1}$ *(des polynômes de degré* $2n+1$*) si et seulement si*

(a) elle est de type interpolation à $n+1$ *points,*

(b) les abscisses d'interpolation sont telles que

$$v(x) = \Pi_{j=0}^{n}(x - x_j) \text{ vérifie } \int_a^b x^q v(x)\mu(x)\,dx = 0, \qquad \forall q, 0 \le q \le n.$$

Démonstration. — remarquons d'abord qu'il y a $2n+2$ inconnus (c_i, x_i) pour $i = 1, n$ et $\mathbb{P}_{2n+1}$ est de dimension $2n+2$.

($\Longrightarrow$) Si la formule est exacte sur $\mathbb{P}_{2n+1}$, alors elle est exacte sur $\mathbb{P}_n$ d'où (a). D'autre part, $\forall 0 \le q \le n$, $x^q v(x) \in \mathbb{P}_{2n+1}$ car $v(x) \in \mathbb{P}_{n+1}$ et comme la formule est exacte sur $\mathbb{P}_{2n+1}$, alors

$$\int_a^b x^q v(x)\mu(x)\,dx = \sum_{i=0}^{n} c_i x_i^q v(x_i) = 0, \text{ car } v(x_i) = 0.$$

($\Longleftarrow$) Soit $p \in \mathbb{P}_{2n+1}$ que l'on divise par $v \in \mathbb{P}_{n+1}$, alors $p = vq + r$ avec $q, r \in \mathbb{P}_n$, donc

$$\int_a^b p(x)\mu(x)\,dx = \int_a^b q(x)v(x)\mu(x)\,dx + \int_a^b r(x)\mu(x)\,dx,$$

on a d'après (b)

$$\int_a^b q(x)v(x)\mu(x)\,dx = 0$$

et d'après (a)

$$\int_a^b r(x)\mu(x)\,dx = \sum_{i=0}^{n} c_i r(x_i).$$

On a $p(x_i) = v(x_i)q(x_i) + r(x_i) = r(x_i)$ car $v(x_i) = 0$, ainsi

$$\int_a^b p(x)\mu(x)\,dx = \sum_{i=0}^{n} c_i p(x_i).$$

□

5.4.1. Application. — Les polynômes orthogonaux déjà construits associés au poids $\mu(x)$ vérifient la partie (b) du théorème précédent. En effet, soit $(h_n)_n$ une famille de polynômes orthogonaux associés au produit scalaire avec poids $\mu(x)$. Alors $v(x) = C_{n+1}h_{n+1}(x)$ et $x^q \in \mathbb{P}_q$, alors $x^q = \sum_{i=0}^{q} \beta_i h_i(x)$, ainsi

$$\int_a^b x^q v(x)\mu(x)\,dx = C_{n+1}\sum_{i=0}^{q}\int_a^b \beta_i h_i(x)h_{n+1}(x)\mu(x)\,dx = 0,$$

car $0 \leq i \leq q \leq n$.

La méthode d'intégration de Gauss-Legendre. Les polynômes de Legendre pour $\mu(x) = 1$, $a = -1$ et $b = 1$ sont donnés par la formule suivante :

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n,$$

et ils vérifient :

$$(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0.$$

On a

$$\begin{aligned}
&L_0(x) = 1\\
&L_1(x) = x,\\
&L_2(x) = \frac{3}{2}x^2 - \frac{1}{2},\ \text{ses racines sont } x = \pm\frac{1}{\sqrt{3}},\\
&L_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x,\ \text{ses racines sont } x = 0,\ x = \pm\sqrt{\frac{3}{5}}.
\end{aligned}$$

Dans la méthode d'intégration suivante

$$\int_{-1}^{1} f(x)\,dx \approx \alpha f(-\frac{1}{\sqrt{3}}) + \beta f(\frac{1}{\sqrt{3}}),$$

les coefficients α et β sont calculés pour que la méthode soit exacte pour $\mathbb{P}_1$, par contre le degré de précision est forcement 3, la méthode est exacte pour $\mathbb{P}_3$. La méthode suivante

$$\int_{-1}^{1} f(x)\,dx \approx \alpha f(-\sqrt{\frac{3}{5}}) + \beta f(0) + \gamma f(\sqrt{\frac{3}{5}})$$

est exacte pour $\mathbb{P}_5$.

Gauss-Tchebychev. $\mu(x) = \frac{1}{\sqrt{1-x^2}}$ sur $]-1,1[$.

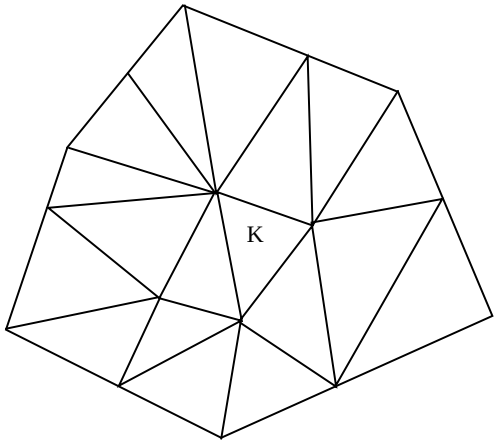
$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}\,dx \approx \sum_{i=0}^{n} c_i f(x_i),$$

avec

$$x_i = \cos\frac{2i+1}{2n+2}\pi,\ \text{les racines des polynômes de Tchebychev.}$$

5.5. Intégration numérique d'une fonction en 2D

Soit Ω un domaine polygonale. On recouvre exactement Ω par des domaines élémentaires du type triangle (ou du type rectangle).



Le domaine Ω est partitionné en N triangles K_i :

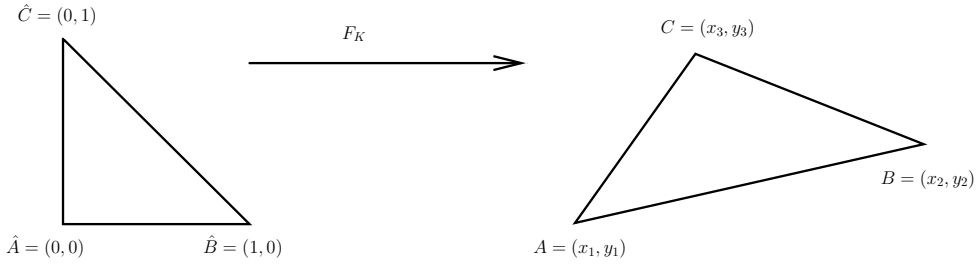
$$\Omega = \cup_{i=1}^{N} K_i$$

La méthode composite s'écrit alors

$$\int_\Omega f(x,y)\,dxdy = \sum_{i=1}^{N} \int_{K_i} f(x,y)\,dxdy. \tag{5.16}$$

Il suffit alors de déterminer une approximation de $\int_{K_i} f$ pour obtenir celle sur Ω. Comme en dimension 1, on va construire des méthodes d'intégrations sur un triangle de référence fixe $\hat{K}$ et puis en déduire celle sur un triangle quelconque.

Transformation affine de $\hat{K}$ dans K. Soit $\hat{K}$ le triangle de référence de sommets $\hat{A} = (0,0)$, $\hat{B} = (1,0)$ et $\hat{C} = (0,1)$. On désigne par K un triangle quelconque de sommets $A = (x_1, y_1)$, $B = (x_2, y_2)$ et $C = (x_3, y_3)$



On cherche la transformation affine inversible

$$F_K : \hat{K} \longmapsto K$$

$$F\begin{pmatrix}\hat{x}\\ \hat{y}\end{pmatrix} = \begin{pmatrix}x\\ y\end{pmatrix} = \begin{pmatrix}a_{11}\hat{x} + a_{12}\hat{y} + b_1\\ a_{21}\hat{x} + a_{22}\hat{y} + b_2\end{pmatrix}.$$

L'application F_K est déterminée de façon unique par

$$F_K(\hat{A}) = A,\ F_K(\hat{B}) = B,\ F_K(\hat{C}) = C.$$

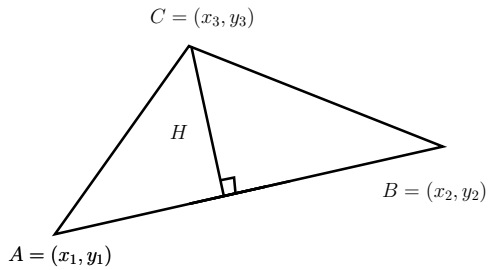
en effet

$$F\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}x_1\\y_1\end{pmatrix} \implies b_1 = x_1,\ b_2 = x_2$$
$$F\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}x_2\\y_2\end{pmatrix} \implies a_{11} = x_2 - x_1,\ a_{21} = y_2 - y_1,$$
$$F\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}x_3\\y_3\end{pmatrix} \implies a_{12} = x_3 - x_1,\ a_{22} = y_3 - y_1.$$

L'application F_K s'écrit

$$F_K(\hat{X}) = X = J_K\hat{X} + b = \begin{pmatrix}x_2 - x_1 & x_3 - x_1\\ y_2 - y_1 & y_3 - y_1\end{pmatrix}\begin{pmatrix}\hat{x}\\\hat{y}\end{pmatrix} + \begin{pmatrix}b_1\\b_2\end{pmatrix}$$

avec J_K dépend de K et F_K est inversible $\iff$ $det\, J_K \neq 0$. On a $det\, J_K = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)$. On montre que $|det\, J_K| = 2\, aire(K)$. En effet



On a $aire(K) = \frac{H|\vec{AB}|}{2}$.
On considère les deux vecteurs $\vec{AB}$ et $\vec{AC}$, on a $|\vec{AB} \times \vec{AC}| = |\vec{AB}||\vec{AC}| \sin\theta$, or $\sin\theta = \frac{H}{|\vec{AC}|}$, aussi $|\vec{AB} \times \vec{AC}| = |\vec{AB}|H = 2\, aire(K)$. D'autre part,

$$\vec{AB} \times \vec{AC} = \begin{pmatrix}x_2 - x_1\\y_2 - y_1\end{pmatrix} \times \begin{pmatrix}x_3 - x_1\\y_3 - y_1\end{pmatrix}$$
$$= (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1).$$

Changement de variable. On pose $X = F_K(\hat{X})$, on a

$$\int_K f(x,y)\, dxdy = \int_{\hat{K}} (f \circ F_K)(\hat{X}) |det\, \nabla F_K|\, d\hat{x}d\hat{y},$$

soit encore

$$\int_K f(x,y)\, dxdy = 2\, aire(K) \int_{\hat{K}} (f \circ F_K)(\hat{X})\, d\hat{x}d\hat{y}, \tag{5.17}$$

cette formule est très intéressante car on connaît $aire(K)$ en fonction des coordonnées des sommets et en posant $g = f \circ F_K$, il suffit alors de construire des formules d'intégrations sur le triangle de référence $\hat{K}$.

Exemple 1 : Intégration par $\mathbb{P}_1$-sommets. On cherche α, β et γ pour que la méthode

$$\int_{\hat{K}} g(\hat{x},\hat{y})\, d\hat{x}d\hat{y} \approx \alpha g(\hat{A}) + \beta g(\hat{B}) + \gamma g(\hat{C}) \tag{5.18}$$

soit exacte pour les polynômes de degré 1. En dimension 2 d'espace, l'espace de polynômes de degré 1 est de dimension 3, $\mathbb{P}_1 =< 1, \hat{x}, \hat{y} >$.

Pour $g = 1$; $\int_{\hat{K}} d\hat{x}d\hat{y} = \frac{1}{2} = \alpha + \beta + \gamma$
Pour $g = \hat{x}$;

$$\int_{\hat{K}} \hat{x}\, d\hat{x}d\hat{y} = \int_0^1 \Big(\int_0^{1-\hat{x}} \hat{x}\, d\hat{y}\Big) d\hat{x} = \int_0^1 \hat{x}(1-\hat{x})\, d\hat{x} = \frac{1}{6} = \beta.$$

Pour $g = \hat{y}$;

$$\int_{\hat{K}} \hat{y}\, d\hat{x}d\hat{y} = \int_0^1 \Big(\int_0^{1-\hat{x}} \hat{y}\, d\hat{y}\Big) d\hat{x} = \frac{1}{2}\int_0^1 (1-\hat{x})^2\, d\hat{x} = \frac{1}{6} = \gamma.$$

La méthode est complètement déterminée par

$$\int_{\hat{K}} g(\hat{x},\hat{y})\, d\hat{x}d\hat{y} \approx \frac{1}{6}(g(\hat{A}) + g(\hat{B}) + g(\hat{C})). \tag{5.19}$$

0n vérifie que le degré de précision est 1 car pour $g = \hat{x}^2$, $\int_{\hat{K}} \hat{x}^2\, d\hat{x}d\hat{y} = \int_0^1 \hat{x}^2(1-\hat{x})^2\, d\hat{x} = \frac{1}{12}$ et $\frac{1}{6}(g(\hat{A}) + g(\hat{B}) + g(\hat{C})) = \frac{1}{6}$.

Sur un triangle quelconque K, et on utilisant la formule (5.17), il vient

$$\int_K f(x,y)\, dxdy = 2\, aire(K) \int_{\hat{K}} (f \circ F_K)(\hat{X})\, d\hat{x}d\hat{y}$$
$$\approx \frac{1}{3} aire(K)\Big((f \circ F_K)(\hat{A}) + (f \circ F_K)(\hat{B}) + (f \circ F_K)(\hat{C})\Big) = \frac{1}{3} aire(K)(f(A) + f(B) + f(C)).$$

Enfin, la formule composite sur Ω par intégration $\mathbb{P}_1$-sommets s'écrit :

$$\int_\Omega f(x,y)\, dxdy \approx \sum_{i=1}^{N} \frac{aire(K_i)}{3}\big(f(A_{K_i}) + f(B_{K_i}) + f(C_{K_i})\big), \tag{5.20}$$

avec A_{K_i}, B_{K_i}, C_{K_i} sont les trois sommets du triangle K_i.
Erreur d'intégration (voir TA 2007).

Exemple 2 :Intégration par $\mathbb{P}_1$-centre de gravité. La méthode suivante

$$\int_{\hat{K}} g(\hat{x},\hat{y})\, d\hat{x}d\hat{y} \approx \frac{1}{2} g(\frac{1}{3},\frac{1}{3}), \tag{5.21}$$

est exacte pour $\mathbb{P}_1$.
Sur un élément K quelconque,

$$\int_K f(x,y)\, dxdy \approx aire(K) f(x_G, y_G), \tag{5.22}$$

avec (x_G, y_G) est le barycentre du triangle K.

Exemple 3 : Intégration par $\mathbb{P}_1$-milieu. Soient $\hat{M}_1 = (\frac{1}{2}, 0)$, $\hat{M}_2 = (0, \frac{1}{2})$ et $\hat{M}_3 = (\frac{1}{2}, \frac{1}{2})$ les milieux des trois côtes du triangle $\hat{K}$. La méthode d'intégration suivante

$$\int_{\hat{K}} g(\hat{x}, \hat{y})\, d\hat{x}d\hat{y} \approx \alpha g(\hat{M}_1) + \beta g(\hat{M}_2) + \gamma g(\hat{M}_3), \tag{5.23}$$

est déterminée pour $\mathbb{P}_1$ comme précédemment et s'écrit

$$\int_{\hat{K}} g(\hat{x}, \hat{y})\, d\hat{x}d\hat{y} \approx \frac{1}{6}\big(g(\hat{M}_1) + g(\hat{M}_2) + g(\hat{M}_3)\big). \tag{5.24}$$

Par contre le degré de précision est 2, la méthode est alors exacte pour $\mathbb{P}_2 =< 1, \hat{x}, \hat{y}, \hat{x}^2, \hat{y}^2, \hat{x}\hat{y} >$. Cette méthode est intéressante car elle est plus précise que la méthode $\mathbb{P}_1$-sommets et $\mathbb{P}_1$-barycentre et utilise uniquement trois points.

CHAPITRE 6

RÉSOLUTION NUMÉRIQUES DES EDO

6.1. Le problème de Cauchy

6.1.1. Les équations du premier ordre. — Soit f une fonction définie de $[t_0, T]\times\mathbb{R} \mapsto \mathbb{R}$. Le problème de Cauchy consiste à trouver une fonction $y : [t_0, T] \mapsto \mathbb{R}$ solution de

$$\begin{cases} y'(t) = f(t, y(t)), & t \in [t_0, T] \\ y(t_0) = y_0 \end{cases} \tag{6.1}$$

La condition $y(t_0) = y_0$ est une condition initiale ou la condition de Cauchy.
Si on suppose que la fonction f est continue par rapport aux deux variables t, y et que f est uniformément Lipschitzienne par rapport à y c'est à dire que

$$\exists L > 0,\ \forall t \in [t_0, T],\ \forall y_1, y_2 \in \mathbb{R}, |f(t, y_1) - f(t, y_2| \leq L|y_1 - y_2|,$$

alors le problème de Cauchy admet une unique solution $y \in \mathcal{C}^1([0, T])$ (C'est le Théorème de Cauchy).
Le problème de Cauchy est un **problème d'évolution**, c'est à dire à partir de la condition initiale, on peut calculer la solution à l'instant t comme

$$y(t) = y(t_0) + \int_{t_0}^{t} f(s, y(s))\, ds, \tag{6.2}$$

et donc la solution à l'instant t dépend uniquement de la solution aux instants $t_0 \leq s \leq t$. La solution (6.2) ne donne pas y de façon explicite sauf dans des cas simples.

6.1.2. Les systèmes du 1er ordre. — Ils s'écrivent sous la forme suivante

$$\begin{cases} Y'(t) = F(t, Y(t)), & t \in [0, T] \\ Y(t_0) = Y_0 \end{cases} \tag{6.3}$$

avec $Y = (y_1, y_2, ..., y_N)$, $F : [t_0, T] \times \mathbb{R}^N \mapsto \mathbb{R}^N$ définie par

$$F(t, Y) = (f_1(t, Y), f_2(t, Y), ..., f_N(t, Y)).$$

6.1.3. Les équations différentielles d'ordre >1. — Exemple. L'équation du pendule

$$\begin{cases} \theta''(t) = -\frac{g}{L}\sin\theta(t) \\ \theta(0) = \theta_0 \\ \theta'(t) = \theta_1 \end{cases} \tag{6.4}$$

Ce système se ramène à un système d'ordre 1 en posant

$$y_1(t) = \theta(t),\ y_2(t) = \theta'(t)$$

ainsi

$$y_1'(t) = \theta'(t),\ y_2'(t) = \theta''(t) = -\frac{g}{L}\sin\theta(t) = -\frac{g}{L}\sin y_1(t).$$

On note $Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}$ et $F(t, Y(t)) = \begin{pmatrix} y_2(t) \\ -\frac{g}{L}\sin y_1(t) \end{pmatrix}$ alors l'équation (6.6) est équivalente à

$$Y'(t) = F(t, Y(t))$$

avec $Y(0) = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$.

Dans le cas général d'une équation différentielle de la forme :

$$\begin{cases} u^{(p)}(t) = h(t, u, u'(t), \cdots, u^{(p-1)}(t)) \\ u^{(j)}(0) = u_j^0,\ j = 0, p-1 \end{cases} \tag{6.5}$$

On se ramène à un système d'ordre 1 en posant : $y_1 = u$, $y_2 = u', \cdots, y_p = u^{(p-1)}$, soit encore

$$\begin{cases} y_1 = u \implies y_1' = y_2 \\ y_2 = u' \implies y_2' = u'' = y_3 \\ y_p = u^{(p-1)} \implies y_p' = u^{(p)} = h(t, y_1, y_2, \cdots, y_p) \end{cases} \tag{6.6}$$

Remarque 6.1. — *Problème de Cauchy ≠ problème aux limites*

6.2. Approximation numérique des équations différentielles d'ordre 1

Le but est d'étudier des méthodes numériques consistantes et stables permettant le calcul de bonnes approximations de la solution exacte de l'edo (6.1).

La méthode la plus célèbre est celle de Euler.

Méthode de Euler. On se donne une subdivision de $I = [t_0, T]$ en N intervalles de pas h

$$dessin\ h = t_{n+1} - t_n$$

La méthode d'Euler consiste à approcher $y'(t_n)$ par la formule de Taylor comme suit :

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \mathcal{O}(h^2)$$

soit

$$y'(t_n) = \frac{y(t_{n+1}) - y(t_n)}{h} + \mathcal{O}(h) = f(t_n, y(t_n))$$

Soit y_n une approximation de $y(t_n)$ ($y_n \approx y(t_n)$), le schéma d'Euler s'écrit

$$y_{n+1} = y_n + hf(t_n, y_n),\ n = 0, N-1, \tag{6.7}$$

$$y(0) = y_0.$$

Interprétation de la méthode d'Euler

Définition 6.1. —
Un **Schéma à un pas** *si y_{n+1} est une fonction de t_n et y_n uniquement. Le schéma de Euler est un schéma à un pas.*
Un **Schéma à deux pas** *si y_{n+1} est une fonction de (t_n y_n) et de (t_{n-1}, y_{n-1}) uniquement.*
Un **Schéma à k pas** *si y_{n+1} est une fonction de (t_n y_n), (t_{n-1}, y_{n-1})... $(t_{n-(k-1)}, y_{n-(k-1)})$.*
Un schéma à k pas est **explicite** *si on peut exprimer y_{n+1} sous la forme*

$$y_{n+1} = E(t_n, t_{n-1}, .., t_{n-(k-1)}, y_n, y_{n-1}, ..., y_{n-(k-1)}).$$

Le schéma d'Euler est explicite. Un schéma à k pas est **implicite** *si y_{n+1} est solution d'une équation non linéaire de la forme*

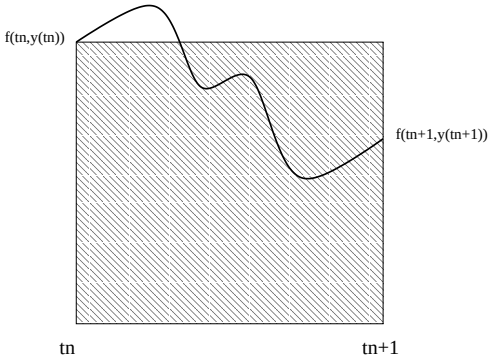
$$y_{n+1} = I(t_{n+1}, t_n, t_{n-1}, .., t_{n-(k-1)}, y_{n+1}, y_{n-1}, ..., y_{n-(k-1)}).$$

6.3. Schémas classiques

Une façon d'obtenir une multitude de schémas est d'intégrer l'edo sur $[t_n, t_{n+1}]$:

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t)\, dt$$

et ensuite d'approcher l'intégrale. Par exemple



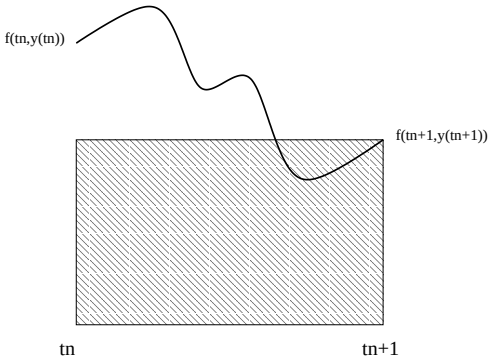
Euler explicite

Intégration par la méthode des rectangles à gauche,

$$\int_{t_n}^{t_{n+1}} f(t, y(t)\, dt \approx h f(t_n, y(t_n))$$

ce qui donne le schéma d'Euler explicite

$$y_{n+1} = y_n + h f(t_n, y_n).$$



Euler implicite

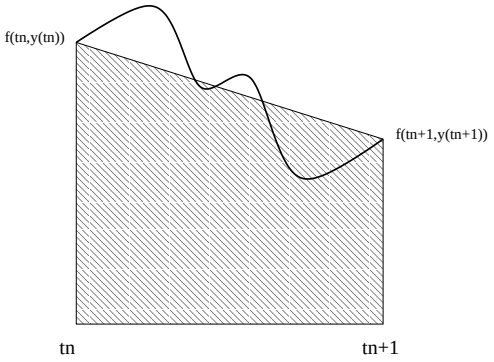
Intégration par la méthode des rectangles à droite,

$$\int_{t_n}^{t_{n+1}} f(t, y(t)\, dt \approx h f(t_{n+1}, y(t_{n+1}))$$

ce qui donne le schéma d'Euler implicite

$$y_{n+1} = y_n + h f(t_{n+1}, y_{n+1}),$$

y_{n+1} est solution d'une équation non linéaire.



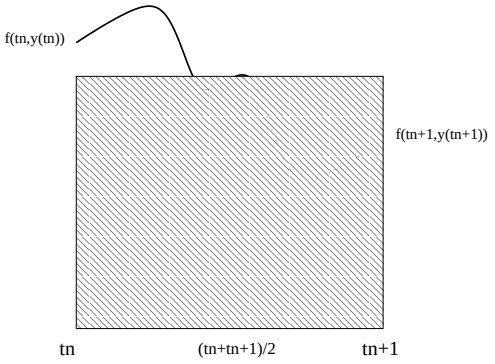
Méthode des trapèzes

Intégration par la méthode des trapèzes,

$$\int_{t_n}^{t_{n+1}} f(t, y(t)\, dt \approx \frac{h}{2}\Big(f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))\Big)$$

ce qui donne le schéma à un pas implicite

$$y_{n+1} = y_n + \frac{h}{2}\Big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\Big)$$



Point Milieu

Intégration par la méthode du point milieu,

$$\int_{t_n}^{t_{n+1}} f(t, y(t)\, dt \approx hf(t_n+\frac{h}{2}, y(t_n+\frac{h}{2}))$$

On connaît uniquement la valeur de y_n, et pour donner une approximation de la solution au point $t_n + \frac{h}{2}$ on utilise le schéma d'Euler explicite :

$$y(t_n + \frac{h}{2}) \approx y(t_n) + \frac{h}{2} f(t_n, y(t_n)),$$

le schéma d'Euler modifié s'écrit

$$y_{n+1} = y_n + h\Big(f(t_n+\frac{h}{2}, y_n+\frac{h}{2} f(t_n, y_n)\Big)$$

6.4. Etude des méthodes à un pas

Dans les méthode à un pas, le calcul de y_{n+1} fait intervenir t_n, y_n, h que l'on peut écrire sous la forme

$$\begin{cases} y_{n+1} = y_n + h\Phi(t_n, y_n, h), & n = 0, N-1 \\ y_0 \text{ donné .} \end{cases} \tag{6.8}$$

Dans le paragraphe précédent, plusieurs schémas numériques ont été présentés et afin de les comparer plusieurs critères seront étudiés :
→ Consistance, ordre d'approximation
→ Stabilité numérique
→ Convergence
→ *A-stable(comportement pour T grand)*
→ *Bien conditionnée (choix de h pour réduire le temps de calcul).*

Définition 6.2. — *(Convergence) Soit $e_n = y_n - y(t_n)$ l'erreur locale de convergence. La méthode est convergente si*

$$\max_{n=0,N} |e_n| \longrightarrow 0 \text{ quand } N \to \infty. \tag{6.9}$$

Définition 6.3. — *(Consistance) Le schéma approche-t-il l'équation ? Si on remplace la solution approchée par la solution exacte dans le schéma numérique, quelle est l'erreur commise en fonction de h ? Pour ceci, on définit l'erreur locale de consistance*

$$E_n = y(t_{n+1}) - \overline{y}_{n+1}, \tag{6.10}$$

avec $\overline{y}_{n+1}$ est la solution du schéma issue de $y(t_n)$:

$$\overline{y}_{n+1} = y(t_n) + h\Phi(t_n, y(t_n), h). \tag{6.11}$$

On dit que le schéma est consistant d'ordre $p \geq 1$ si

$$\exists K > 0 \text{ telle que } |E_n| \leq Kh^{p+1}. \tag{6.12}$$

Définition 6.4. — *(Stabilité) C'est une propriété du schéma numérique. Un schéma est stable signifie qu'une petite perturbation sur les données (y_0, Φ) n'entraîne qu'une petite perturbation sur la solution indépendamment de h. Plus précisément,*
soient $\{y_n\}_n$, $\{z_n\}_n$, $n = 0, N$, les solutions de

$$\begin{cases} y_{n+1} = y_n + h\Phi(t_n, y_n, h), & n = 0, N-1 \\ y_0 \text{ donné .} \end{cases} \tag{6.13}$$

et

$$\begin{cases} z_{n+1} = z_n + h(\Phi(t_n, z_n, h) + \varepsilon_n), & n = 0, N-1 \\ z_0 \text{ donné .} \end{cases} \tag{6.14}$$

La méthode est dite stable s'il existe deux constantes M_1 et M_2 indépendantes de h telles que

$$\max_{n=0,..,N} |y_n - z_n| \leq M_1|y_0 - z_0| + M_2 \max_{0,N-1} |\varepsilon_n|. \tag{6.15}$$

Théorème 6.1. — *(Théorème de Lax – Convergence) Soit une méthode à un pas consistante et stable, alors elle est convergente.*

Démonstration. — Le schéma est consistant, alors

$$y(t_{n+1}) = \overline{y}_{n+1} + E_n = y(t_n) + h\Phi(t_n, y(t_n), h) + E_n,$$

on pose $\varepsilon_n = E_n/h$ et donc $|\varepsilon_n| \leq Kh^p \longrightarrow 0$ quand $h \to 0$ ou $N \to \infty$.
Comme le schéma est stable et en prenant $z_n = y(t_n)$, on a

$$\max_{n=0,..,N} |y_n - y(t_n)| \leq M_1|y_0 - y(t_0)| + M_2 \max_{0,N-1} |\varepsilon_n|,$$

ce qui montre

$$\max_{n=0,..,N} |y_n - y(t_n)| \to 0,$$

et le schéma est convergent. □

6.4.1. Stabilité et consistance. —

Théorème 6.2. — *(Condition suffisante de stabilité) Si Φ est Lipschtzienne par rapport à y uniformément en h de constante M, alors la méthode à un 1 pas est stable.*

Démonstration. — Soient $(y_n)_n$ solution de (6.13) et $(z_n)_n$ solution de (6.14). En faisant la différence de deux équations, il vient

$$|y_{n+1} - z_{n+1}| \leq |y_n - z_n| + h|\Phi(t_n, y_n, h) - \Phi(t_n, z_n, h)| + h|\varepsilon_n|, \tag{6.16}$$

en tenant compte du fait que Φ est M-Lipschitz,

$$|y_{n+1} - z_{n+1}| \leq (1 + hM)|y_n - z_n| + h|\varepsilon_n|. \tag{6.17}$$

On réutilise cette estimation pour estimer $|y_n - z_n|$, on déduit

$$|y_{n+1} - z_{n+1}| \leq (1 + hM)^2 |y_{n-1} - z_{n-1}| + (1 + hM)h|\varepsilon_{n-1}| + h|\varepsilon_n|, \tag{6.18}$$

ce qui permet de conclure par récurrence que

$$|y_{n+1} - z_{n+1}| \leq (1 + hM)^{n+1} |y_0 - z_0| + h \sum_{i=0}^{n} (1 + hM)^k |\varepsilon_{n-k}|. \tag{6.19}$$

On a

$$\sum_{i=0}^{n} (1 + hM)^k = \frac{1 - (1 + hM)^{n+1}}{1 - 1 - hM} = \frac{(1 + hM)^{n+1} - 1}{hM},$$

d'autre part,

$$(1 + hM)^{n+1} \leq e^{(n+1)hM} \leq e^{(T-t_0)M}.$$

De (6.19), on déduit

$$\begin{aligned} |y_{n+1} - z_{n+1}| \quad &\leq (1 + hM)^{n+1} |y_0 - z_0| + h \sum_{i=0}^{n} (1 + hM)^k \max_{k=0,n} |\varepsilon_k| \\ &\leq (1 + hM)^{n+1} |y_0 - z_0| + \frac{(1 + hM)^{n+1} - 1}{M} \max_{k=0,n} |\varepsilon_k| \\ &\leq e^{(T-t_0)M} |y_0 - z_0| + \frac{e^{(T-t_0)M} - 1}{M} \max_{k=0,n} |\varepsilon_k| \end{aligned}$$

ce qui établit (6.15) avec $M_1 = e^{(T-t_0)M}$ et $M_2 = \frac{e^{(T-t_0)M}-1}{M}$.

6.4.2. Consistance et ordre de consistance. — On va décrire une méthode générale pour calculer l'ordre de consistance d'un schéma à un pas. On rappelle l'erreur de consistance

$$E_n = y(t_{n+1}) - y(t_n) - h\Phi(t_n, y(t_n), h).$$

On pose $t = t_n$ un point générique, et donc

$$E_n = y(t + h) - y(t) - h\Phi(t, y(t), h).$$

La formule de Taylor donne

$$y(t + h) = y(t) + hy'(t) + \frac{h^2}{2!} y''(t) + \frac{h^3}{3!} y^{(3)}(t) + ... + \frac{h^p}{p!} y^{(p)}(t) + ... \tag{6.20}$$

$$\Phi(t, y(t), h) = \Phi(t, y(t), 0) + h\frac{\partial \Phi}{\partial h}(t, y(t), 0) + \frac{h^2}{2!}\frac{\partial^2 \Phi}{\partial^2 h}(t, y(t), 0) + ... + \frac{h^{p-1}}{(p-1)!}\frac{\partial^{p-1} \Phi}{\partial^{p-1} h}(t, y(t), 0) + ... \tag{6.21}$$

alors

$$\begin{aligned} E_n \quad =& h\Big[y'(t) - \Phi(t, y(t), 0)\Big] + h^2\Big[\frac{1}{2}y''(t) - \frac{\partial \Phi}{\partial h}(t, y(t), 0)\Big] \\ &+ \frac{h^3}{2!}\Big[\frac{1}{3}y^{(3)}(t) - \frac{\partial^2 \Phi}{\partial^2 h}\Phi(t, y(t), 0)\Big] + ... + \frac{h^p}{(p-1)!}\Big[\frac{1}{p}y^{(p)}(t) - \frac{\partial^{p-1}\Phi}{\partial^{p-1}h}\Phi(t, y(t), 0)\Big]. \end{aligned}$$

On conclut que :
Le schéma est au moins d'ordre 1 (consistant) si

$$\Phi(t, y(t), 0) = y'(t) = f(t, y(t))$$

Le schéma est au moins d'ordre 2 si de plus

$$\begin{aligned} \frac{\partial \Phi}{\partial h}(t, y(t), 0) = \frac{1}{2}y''(t) = \frac{1}{2}\frac{d}{dt}f(t, y(t)) \qquad &= \frac{1}{2}\Big(\partial_t f(t, y(t)) + y'(t)\partial_y f(t, y(t))\Big) \\ &= \frac{1}{2}\Big(\partial_t f(t, y(t)) + f(t, y(t))\partial_y f(t, y(t)\Big). \end{aligned}$$

Le schéma est au moins d'ordre 3 si de plus

$$\frac{\partial^2 \Phi}{\partial^2 h}\Phi(t, y(t), 0) = \frac{1}{3}y^{(3)}(t) = \frac{1}{3}\frac{d^2}{dt^2}f(t, y(t))$$

Le schéma est au moins d'ordre p si de plus

$$\frac{\partial^{p-1}\Phi}{\partial^{p-1}h}\Phi(t, y(t), 0) = \frac{1}{p}y^{(p)}(t) = \frac{1}{p}\frac{d^{p-1}}{dt^{p-1}}f(t, y(t)).$$

On dispose alors du résultat de convergence suivant :

Théorème 6.3. — *Soit* Φ *Lipschitz par rapport à* y *(stabilité) et vérifiant* $\Phi(t, y(t), 0) = f(t, y(t))$ *pour tout* t *(consistance). Alors le schéma* (6.8) *est convergent.*

Application. Le schéma d'Euler est convergent.

6.4.3. La méthode de Runge-Kutta. — Pour calculer une approximation de la solution à l'instant t_{n+1} en fonction de celle de t_n, la méthode de Runge-Kutta utilise q solutions intermédiaires en fonction de y_n. La méthode de Runge-Kutta de rang q sous sa forme générale s'écrit :

$$\begin{cases} y_{n,i} = y_n + h\displaystyle\sum_{j=1}^{q} a_{ij} f(t_{n,j}, y_{n,j}) \;\; i = 1, ..., q \\ y_{n+1} = y_n + h\displaystyle\sum_{i=1}^{q} b_i f(t_{n,i}, y_{n,i}) \\ t_{n,i} = t_n + c_i h \;\; i = 1, ..., q \end{cases}$$

C'est une méthode à un pas et elle redonne la plupart des méthodes déjà vues. Pour q quelconque, on peut trouver des conditions sur les coefficients b_i, c_i, a_{ij} pour que la

méthode soit consistante, d'ordre 2, ...

La méthode de Runge-Kutta classique de rang 4 (ordre 4) :

$$\begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + \frac{h}{2} f(t_n, y_{n,1}), \\ y_{n,3} = y_n + \frac{h}{2} f(t_n + h/2, y_{n,2}), \\ y_{n,4} = y_n + h f(t_n + h/2, y_{n,3}), \\ \\ y_{n+1} = y_n + \frac{h}{6} f(t_n, y_{n,1}) + \frac{h}{3} f(t_n + h/2, y_{n,2}) + \frac{h}{3} f(t_n + h/2, y_{n,3}) + \frac{h}{6} f(t_n + h, y_{n,4}). \end{cases} \tag{6.22}$$

est la plus employée pour résoudre une edo.

6.5. Méthodes à pas multiple

□

CHAPITRE 7

TRAVAUX DIRIGÉS

7.1. Systèmes linéaires creux

7.1.1.

Exercice 7.1. — **(Erreurs d'arrondi)** *Soit S une sphère dans $\mathbb{R}^3$ de rayon R et tangente à un seul plan de coordonnées. Soit une autre sphère s, de rayon r et tangente au plan de coordonnées et à S.*

1. *Faire un dessin, et trouver le rapport des volumes de 2 sphères $z = v/V$.*
2. *Vérifier que $z = (\sqrt{2}-1)^6 = (3-2\sqrt{2})^3 = 99-70\sqrt{2}$. Faire un tableau de z en fonction des valeurs de $\sqrt{2}$ approchée par : 1.4, 1.414 , 1.4142136 .*

Exercice 7.2. — 1. *Les matrices suivantes sont elles symétriques, hermitiennes :*

$$A_1 = \begin{pmatrix} 1 & 2+i \\ 2+i & 2 \end{pmatrix} .$$

$$A_2 = \begin{pmatrix} 1 & 2+i \\ 2-i & 2 \end{pmatrix} .$$

2. *Soit A une matrice symétrique. Montrer que A hermitienne ssi A réelle.*
3. *Montrer que $(Ax, y) = (x, A^*y)$ dans C^N et $(Ax, y) = (x, {}^t Ay)$ dans $\mathbb{R}^N$.*
4. *Soit A une matrice définie positive alors les éléments diagonaux sont strictement positifs.*
5. *Soit A une matrice unitaire (ou orthogonale) alors $\|Ax\|_2 = \|x\|_2$ et $|detA| = 1$.*
6. *Montrer que les matrices semblables ont le même polynôme caractéristique.*
7. *Montrer que si A est hermitienne alors ses valeurs propres sont réelles.*
8. *Montrer que toute matrice symétrique réelle est définie positive ssi toutes ses valeurs propres sont strictement positives.*
9. *Soient A et B deux matrices diagonalisables. Montrer que si elles ont les mêmes vecteurs propres alors elles commutent entre elles.*

Exercice 7.3. — *Soit $A = (a_{ij})_{ij}$, $i, j = 1, N$ une matrice triangulaire inférieure, c'est à dire*

$$a_{ij} = 0 \text{ pour } j > i.$$

1. *Donner une CNS pour que A soit inversible et trouver ses valeurs propres.*
2. *Résoudre $AX = b$, avec $A = (a_{ij})_{1\leq i,j\leq N}$, $X = (x_i)_{i=1,N}$, et $b = (b_i)_{i=1,N}$. En déduire alors que si $b_i = 0$ pour $i < k$ et $b_k \neq 0$ alors $x_i = 0$ pour $i < k$ et $x_k \neq 0$*
3. *Montrer que le produit de deux matrices triangulaires inférieures est une matrice triangulaire inférieure.*
4. *Montrer que l'inverse d'une matrice triangulaire inférieure est une matrice triangulaire inférieure.*
5. *Ecrire en language libre un algorithme de résolution de $AX = b$.*

Exercice 7.4. — **Normes matricielles.** *On rappelle les normes suivantes sur $\mathbb{R}^N$: $\|x\|_1 = \sum_{i=1}^N |x_i|$, $\|x\|_\infty = \max_{i=1,N} |x_i|$ et $\|x\|_2 = (\sum_{i=1}^N |x_i|^2)^{\frac{1}{2}}$.*
Soit A une matrice carrée $A = (a_{ij})_{i,j}$, $i, j = 1, N$. On définit la norme induite ou subordonnée associée à une norme vectorielle par

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

Montrer que

1. $\|A\|_1 = \max_{j=1,N} \sum_{i=1}^N |a_{i,j}|$.
2. $\|A\|_\infty = \max_{i=1,N} \sum_{j=1}^N |a_{i,j}|$.
3. *Si B est réelle et symétrique, alors $\lambda_{min}(B) \leq \frac{(Bx,x)}{\|x\|_2^2} \leq \lambda_{max}(B)$ (quotient de Rayleigh). De plus,*

$$\max_{x\neq 0} \frac{(Bx, x)}{\|x\|_2^2} = \lambda_{max}(B),$$
$$\min_{x\neq 0} \frac{(Bx, x)}{\|x\|_2^2} = \lambda_{min}(B).$$

   En déduire que $\|A\|_2 = \sqrt{\rho({}^tAA)}$ (A est réelle).
4. *Montrer que $\rho(A) \leq \|A\|$.*
5. *Calculer les normes 1, ∞ et 2 pour la matrice*

$$\begin{pmatrix} 4 & -1 & 0 \\ -1 & -10 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Exercice 7.5. — **Conditionnement d'une matrice :** $cond_pA = \|A\|_p\|A^{-1}\|_p$**.**

1. *Vérifier les propriétés suivantes :*
   (a) *$condA \geq 1$, $cond(\alpha A) = condA$, $\alpha \in \mathbb{R}$.*
   (b) *A réelle et symétrique alors $cond_2A = \dfrac{\max|\lambda(A)|}{\min|\lambda(A)|}$.*

(c) U orthogonale alors $cond_2 U = 1$ et $cond_2(UA) = cond_2(A) = cond_2(AU)$.

2. *Soient x et $x + \delta x$ les solutions des systèmes linéaires $Ax = b$ et $A(x + \delta x) = b + \delta b$. Montrer que $\frac{\|\delta x\|}{\|x|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$. Conclure.*
3. *Soient x et $x+\delta x$ les solutions des systèmes linéaires $Ax = b$ et $(A+\delta A)(x+\delta x) = b$. Montrer que $\frac{\|\delta x\|}{\|x + \delta x\|} \leq cond(A)\frac{\|\delta A\|}{\|A\|}$.*
   En déduire que si $\|A^{-1}\delta A\| < 1$ alors $\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta A\|}{\|A\|}\frac{1}{(1 - \|A^{-1}\delta A\|)}$.

Exercice 7.6. — **Décomposition LU.** *Soit A une matrice tridiagonale ($a_{i,i-1} = a_i,\ a_{i,i} = b_i,\ a_{i,i+1} = c_i$).*

1. *Trouver L et U telle que $A = LU$, et écrire l'algorithme de décomposition de A. Donner la complexité de cet algorithme.*
2. *Résoudre le système linéaire $Ax = b$.*
3. *Application. Résoudre*

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Exercice 7.7. — **Problème de diffusion 1D.** *La discrétisation du problème de Laplace à une seule variable d'espace :*

$$\begin{cases} -u''(x) = f(x), \ \ 0 < x < L \\ u(0) = u(L) = 0 \end{cases} \tag{7.23}$$

par un schéma aux différences finies sur un maillage à pas fixe $h = \frac{L}{N+1}$ s'écrit :

$$\begin{cases} \frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) = f_i, \ \ i = 1, N \\ u_0 = u_{N+1} = 0 \end{cases} \tag{7.24}$$

1. *Ecrire le système linéaire (7.24) sous forme $AU = F$. Montrer que la matrice A est symétrique et définie positive.*
2. *Vérifier que les vecteurs propres de la matrice A sont les vecteurs V_k, avec $(V_k)_i = \sin(k\pi\frac{ih}{L})$ la ième composante, associés aux valeurs propres $\lambda_k = \frac{4}{h^2}\sin^2(k\pi\frac{h}{2L})$.*
3. *Donner $cond_2 A$. La matrice est-elle bien conditionnée.*

Exercice 7.8. — **Problème de diffusion 2D.** *Soit $D = [0, a] \times [0, b]$, on considère le problème*

$$\begin{cases} -\frac{\partial^2 u}{\partial x^2}(x, y) - \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y) \text{ dans } D \\ u = 0 \text{ sur } \partial D \end{cases} \tag{7.25}$$

La discrétisation par différences finies sur un quadrillage uniforme de pas $\delta x = \frac{a}{N+1}$ *et* $\delta y = \frac{b}{M+1}$ *est :*

$$\frac{1}{\delta x^2}(-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}) + \frac{1}{\delta y^2}(-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}) = f_{i,j};\ 1 \leq i \leq N,\ 1 \leq j \leq M \tag{7.26}$$

avec

$$u_{i,0} = u_{i,M+1} = u_{0,j} = u_{N+1,j} = 0.$$

1. *Soit le vecteur* $U = (u_m)$ *avec* $m = i + (j-1)N$. *Pour* $N = M = 5$ *écrire le système sous la forme* $AU = F$. *Montrer* A *est inversible et définie positive.*
2. *Vérifier que les vecteurs propres de la matrice sont les vecteurs* $u^{(p,q)}$ *définis par*

$$(u^{(p,q)})_{i,j} = \sin(p\pi\frac{i\delta x}{a})\sin(q\pi\frac{j\delta y}{b}).$$

   Les valeurs propres associées sont

$$\lambda_{p,q} = \frac{4}{\delta x^2}sin^2(p\pi\frac{\delta x}{2a}) + \frac{4}{\delta y^2}sin^2(q\pi\frac{\delta y}{2b})$$

3. *On suppose* $\delta x = \delta y$, *calculer* $cond_2 A$.
4. *Donner la matrice* A *associée aux graphes suivants :*

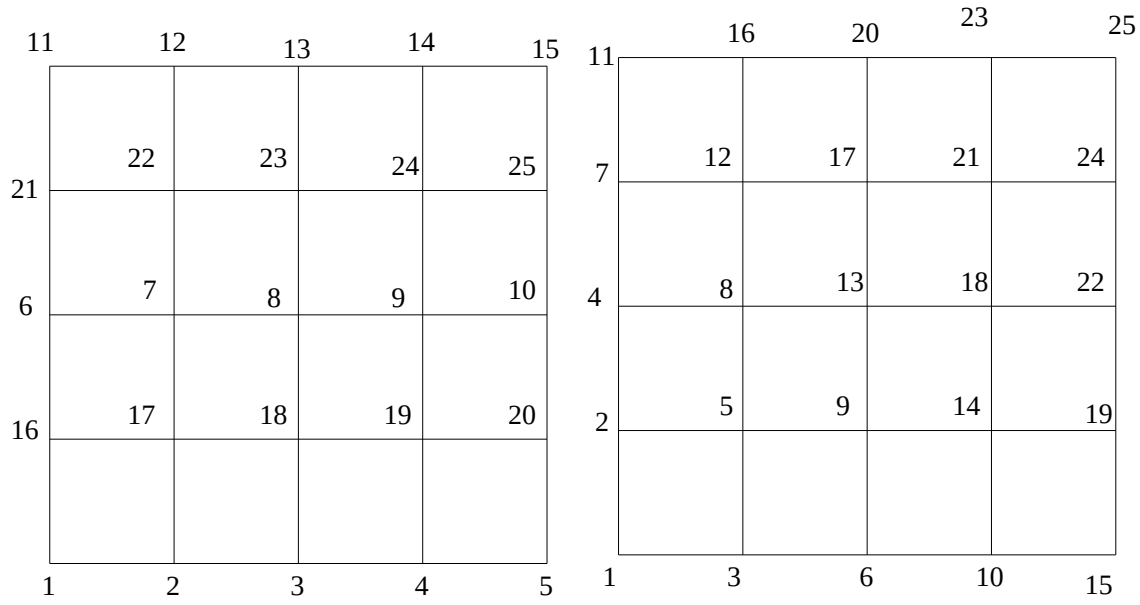


FIGURE 1. Influence de la numérotation. Maillage Zèbre à gauche, maillage diagonale à droite.

7.2. Méthodes itératives

Exercice 7.9. — *Appliquer la méthode itérative de Jacobi au système $Ax = b$ où*

$$A = \begin{pmatrix} 5 & 1 & 0 & 0 & -2 \\ -1 & 5 & 1 & 0 & 0 \\ 0 & -1 & 5 & 1 & 0 \\ 0 & 0 & -1 & 5 & 1 \\ -1 & 0 & 0 & 0 & 5 \end{pmatrix}$$

et montrer que la méthode est convergente.

Exercice 7.10. — **Méthode de Jacobi avec paramètre.**

Soit A une matrice $n \times n$ inversible avec $a_{ii} \neq 0$ et $b \in \mathbb{R}^n$. On veut résoudre le système $Ax = b$. On note D la matrice diagonale constituée de la diagonale de A. Soit $\alpha \neq 0$, on étudie la méthode itérative

$$x_{k+1} = (I - \alpha D^{-1}A)x_k + \alpha D^{-1}b$$

1. *Montrer que si x_k converge vers x alors x est solution.*
2. *Exprimer les coefficients de la matrice $D^{-1}A$ en fonction des coefficients de la matrice A.*
3. *On suppose A est à diagonale strictement dominante et $0 < \alpha \leq 1$. Montrer que la méthode est bien définie et*

$$\|I - \alpha D^{-1}A\|_\infty < 1.$$

En déduire la convergence de la méthode.

Exercice 7.11. — **Méthode de Jacobi et de Gauss-Seidel**

Soit A la matrice du système linéaire $Ax = b$, définie par : $a_{ii} = i + 1$, $i = 1, n$; $a_{i+1,i} = -1$, $i = 1, n-1$; $a_{i,i+1} = -i$, $i = 1, n-1$.

1. *Calculer la matrice d'itération $\mathcal{J}$ de Jacobi. Calculer $\|\mathcal{J}\|_\infty$, $\|\mathcal{J}\|_1$. Conclure.*
2. *Soit G la matrice d'itération de Gauss-Seidel. On pose $L = D^{-1}E$ et $U = D^{-1}F$, montrer que $G = (I - L)^{-1}U$. Montrer que le polynôme caractéristique de G s'écrit*

$$P_G(\lambda) = \lambda^N det(I - L - \frac{1}{\lambda}U),$$

et si $|\lambda| \geq 1$ alors $det(I - L - \frac{1}{\lambda}U) \neq 0$. En déduire que la méthode est convergente.
3. *Le fait d'avoir trouvé une méthode itérative (au moins) convergente prouve que la matrice A est inversible. Pourquoi ?*
4. *Décrire l'algorithme de calcul de la méthode de Gauss-Seidel appliquée à cet exemple.*

Exercice 7.12. — **Double Gauss-Seidel.** *Soit A une matrice réelle, symétrique et définie positive. on considère le splitting $A = D - E - F$ avec D la diagonale, $-E$ la partie inférieure de A et $F = E^T$. On définit la méthode itérative suivante :*

$$\begin{cases} x_0 \text{ arbitraire} \\ (D-E)y_{k+1} = Fx_k + b \\ (D-F)x_{k+1} = Ey_{k+1} + b \end{cases} \tag{7.27}$$

1. *Montrer que les itérés sont bien définis.*
2. *Montrer que la méthode est consistante (c-à-d si $x_k \longrightarrow x$ alors $Ax = b$)*
3. *Ecrire les matrices G et H telles que*

$$x_{k+1} = Gx_k + Hb.$$

4. *En posant $L = D^{-\frac{1}{2}}ED^{-\frac{1}{2}}$ et $U = D^{-\frac{1}{2}}FD^{-\frac{1}{2}}$, montrer que G est semblable à $B = (I-U)^{-1}L(I-L)^{-1}U$ et que la matrice B s'écrit également $B = (I-U)^{-1}(I-L)^{-1}LU$.*
5. *On suppose $Sp(G) \subset \mathbb{R}^+$, montrer que la méthode est convergente.*

Exercice 7.13. — **Méthode des directions alternées** *Soit A une matrice carrée telle que $A = H + V$ où H une matrice réelle symétrique définie positive (c-à-d $(Hx, x) > 0, \forall x \neq 0$) et V une matrice réelle symétrique semi-définie positive (c-à-d $(Vx, x) \geq 0, \forall x$) . Pour résoudre le système $Ax = b$, on considère la méthode itérative suivante :*

$$\begin{cases} x_0 \text{ arbitraire} \\ (I+H)y_{k+1} = (I-V)x_k + b \\ (I+V)x_{k+1} = (I-H)y_{k+1} + b \end{cases} \tag{7.28}$$

avec I la matrice identité.

1. *Montrer que les itérés sont bien définis.*
2. *Montrer que la méthode est consistante (c-à-d si $x_k \longrightarrow x$ alors $Ax = b$).*
3. *Donner la matrice d'itération T telle que*

$$x_{k+1} = Tx_k + g, \tag{7.29}$$

   Montrer que T est semblable à $\tilde{T} = (I-H)(I+H)^{-1}(I-V)(I+V)^{-1}$.
4. *Montrer que*
   (a) *la matrice $(I-H)(I+H)^{-1}$ est symétrique,*
   (b) *$\|(I-H)(I+H)^{-1}\|_2 = \max_i \left|\frac{1-\lambda_i}{1+\lambda_i}\right|$ avec λ_i une valeur propre de H,*
   (c) *$\|(I-H)(I+H)^{-1}\|_2 < 1$.*
5. *Montrer que $\|(I-V)(I+V)^{-1}\|_2 \leq 1$.*
6. *Montrer que $\rho(\tilde{T}) \leq \|\tilde{T}\|_2$. En déduire que la méthode itérative est convergente.*

Exercice 7.14. — *Soit A une matrice symétrique définie positive de valeurs propres*

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N.$$

Pour résoudre le système linéaire $Ax = b$, on considère la méthode de Richardson suivante

$$x_{k+1} = x_k + \alpha(b - Ax_k), \quad \alpha \neq 0,$$

avec α un réel non nul.

1. *Montrer que la méthode est consistante.*
2. *Ecrire la matrice d'itération et montrer que pour $0 < \alpha < 2/\lambda_N$, la méthode est convergente.*
3. *Soit $f_i(\alpha) = |1 - \lambda_i\alpha|$, $i = 1, N$. Tracer $f_1(\alpha)$,$f_N(\alpha)$ et $f_i(\alpha)$ pour $i \neq 1$ et $i \neq N$.*
4. *Trouver le meilleur choix de α, noté α_{opt}, c'est à dire celui qui minimise $\rho(I - \alpha A)$ et montrer que $\rho(I - \alpha_{opt}A) = \frac{cond_2(A)-1}{cond_2(A)+1}$*

Exercice 7.15. — **Valeur propre de plus grand module.**

Soit A une matrice réelle et symétrique. On suppose que les valeurs propres de A vérifient :

$$|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_N|.$$

On cherche à calculer numériquement la valeur propre λ_1.

1. *Soit x_0 un vecteur non nul de $\mathbb{R}^N$, on construit par récurrence à partir de x_0, une suite de réels R_k par :*

$$\begin{cases} x_{k+1} = Ax_k \\ R_k = \frac{<x_k, x_{k+1}>}{<x_k, x_k>} \end{cases} \tag{7.30}$$

   avec $< \cdot, \cdot >$ désigne le produit scalaire dans $\mathbb{R}^N$. Montrer que si x_0 n'est pas orthogonal à l'espace propre associé à λ_1, alors

$$R_k \longrightarrow \lambda_1 \text{ quand } k \to \infty,$$

   avec un facteur de convergence de l'ordre de $(\frac{|\lambda_2|}{|\lambda_1|})^2$.
2. *Proposer un algorithme pour calculer la plus petite valeur propre en module.*

7.3. Interpolation et approximation

Exercice 7.16. — **Interpolation de Lagrange**

1. *Ecrire le polynôme d'interpolation associé aux points :*

$$(x_i, f(x_i)) = (-1, -3/2), (-1/2, 0), (0, 1/4), (1/2, 0), (1, 0).$$

2. *Soient $f(x) = \cos(x)$ et $g(x) = e^{3x}$ définies sur $[0, 1]$, estimer le nombre minimum n de points pour que l'erreur entre la fonction et son polynôme d'interpolation soit inférieure à 0.1, 0.01 , 0.001.*

3. *Déterminer le polynôme q de degré minimum tel que $q(-1) = 8, q(0) = 1, q(1) = 2, q(3) = 196, q'(3) = 299$*

Exercice 7.17. — **Interpolation de Hermite**

Soient $f \in \mathcal{C}^1([a,b])$ et x_1, x_2 deux points distincts. Soit p un polynôme de degré ≤ 3 vérifiant

$p(x_i) = f(x_i)$ et $p'(x_i) = f'(x_i)$ pour $i = 1, 2$

1. *Montrer qu'un tel polynôme unique.*
2. *Existence. Trouver une base (A_1, A_2, B_1, B_2) de $\mathcal{P}_3$ telle que*

$$p(x) = f(x_1)A_1(x) + f(x_2)A_2(x) + f'(x_1)B_1(x) + f'(x_2)B_2(x),$$

et exprimer cette base en fonction des polynômes d'interpolation de Lagrange L_1 et L_2.

3. *Etablir la majoration d'interpolation suivante : si $f \in \mathcal{C}^4([a,b])$, alors il existe $\xi \in]a,b[$ tel que*

$$f(x) - p(x) = \frac{(x-x_1)^2(x-x_2)^2}{4!} f^{(4)}(\xi)$$

4. *Décrire les polynômes d'interpolation de Hermite dans le cas général.*

Exercice 7.18. — **Approximation.**

1. *Trouver a tel que $\| \sin x - a \|_{L^\infty(0,\frac{\pi}{2})} = \max_{0 \leq x \leq \frac{\pi}{2}} |\sin x - a|$ soit minimal.*
2. *Trouver b tel que $\| \sin x - b \|^2_{L^2(0,\frac{\pi}{2})} = \int_0^{\frac{\pi}{2}} |\sin x - b|^2 dx$ soit minimal.*
3. *Trouver le polynôme de degré ≤ 1 qui réalise la meilleure approximation au sens des moindres carrés de $f(x) = x^2 - \frac{x}{4} + \frac{1}{4}$ pour la norme $L^2([-1,1])$.*
4. *Trouver le polynôme de degré ≤ 2 qui réalise la meilleure approximation au sens des moindres carrés discrets associés à : $(x_i, y_i) = (-1, -3/2)$, $(-1/2, 0)$, $(0, 1/4)$, $(1/2, 0)$, $(1, 0)$.*
5. *Soient $(x_1, y_1), ..., (x_m, y_m)$ des points de $\mathbb{R}^2$, $m \geq 2$. Ecrire les équations vérifiées par*
   (a) $\min_{a,b} \sum_{i=1}^m |y_i - ax_i - b|^2$
   (b) $\min_{c,d} \sum_{i=1}^m |x_i - cy_i - d|^2$

Exercice 7.19. — 1. *Montrer que le polynôme P de degré minimum tel que $P(0) = 0$, $P(1) = 2$, $P'(1) = 0$, $P''(1) = 1$ est unique. Dire rapidement comment construire P.*

2. *Montrer que l'espace vectoriel*

$$H = \{v \in C^1[0,2], v \in \mathbb{P}_2 \text{ sur } [0,1], v \in \mathbb{P}_2 \text{ sur } [1,2]\}$$

est de dimension 4 ($\mathbb{P}_2$ est l'ensemble des polynômes de degré deux en dimension un).

3. *Vérifier que les fonctions* $v_1(x) = 1$, $v_2(x) = \sqrt{5}(x-1)^2$, $v_3(x) = \sqrt{3}(x-1)$, $v_4(x) = \sqrt{5}(x-1)|x-1|$ *forment une base de* H.
4. *Soit la fonction* f *définie et intégrable sur* $[0,2]$. *Soit* $v(x) = \sum_{i=1}^{4} \beta_i v_i(x) \in H$, *avec* $\beta_i \in \mathbb{R}$, *réalisant la meilleure approximation de* f *au sens des moindres carrées pour la norme* $L^2(0,2)$.
   Ecrire alors le système linéaire vérifié par le vecteur X *de composantes* β_i, $i = 1,4$. *Déterminer complètement la matrice du système linéaire.*

Exercice 7.20. — **Meilleure approximation au sens de Tchebychev**

Dans la majoration de l'erreur d'interpolation de Lagrange, il est intéressant de chercher les points x_i *de telle sorte que :*

$$\max_{x\in[-1,1]} |(x-x_0)...(x-x_n)| \text{ soit minimal.}$$

Soit les polynômes de Tchebychev définis par $T_n(x) = \cos(n \arccos(x))$ *pour* $x \in [-1,1]$ *et* $n \geq 0$.

1. *Vérifier que* $T_0(x) = 1, T_1(x) = x, T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ *et en déduire que* T_n *est un polynôme de degré* n *de la forme* $T_n(x) = 2^{n-1}x^n + ...$
2. *Donner les racines et les extremum de* T_n.
3. *Soit* $q(x) = 2^{n-1}x^n + ...$ *et* $q(x) \neq T_n(x)$. *Montrer que*

$$\max_{x\in[-1,1]} |T_n(x)| < \max_{x\in[-1,1]} |q(x)|.$$

   Conclure
4. *Etendre ce résultat à un intervalle* $[a,b]$.
5. *Comparer la fonction* $f(x) = \frac{1}{1+x^2}$ *définie sur* $[-8,8]$ *avec :*
   – *le polynôme d'interpolation obtenu à partir de* $n = 10, 20, ..$ *points équidistants*
   – *le polynôme d'interpolation obtenu à partir de* $n = 10, 20, ..$ *points de Tchebychev.*

7.4. Intégration numérique

Exercice 7.21. — **Formule de moyenne.**

1. *Déterminer la formule de quadrature suivante*

$$\int_0^1 g(t)\,dt \approx \alpha g(\frac{1}{2})$$

   pour qu'elle soit exacte pour des polynômes de degré le plus haut possible.
2. *Donner une estimation de l'erreur d'intégration.*
3. *Soit* $\{x_i\}$, $i = 0, ..., M$ *une subdivision de* $[a,b]$ *de pas* h. *Utiliser la formule de moyenne composite pour approcher* $\int_a^b f(x)dx$.
4. *Soit* $\varepsilon = 10^{-1}, 10^{-2}, 10^{-8}$, *trouver* M *pour que cette formule de quadrature approche* $\int_0^3 sin(x)e^{-x^2}dx$ *avec une précision* ε

Exercice 7.22. — **Formule de Simpson.**

1. *Déterminer la formule de quadrature*
$$\int_{-1}^{1} g(x)dx \approx \alpha g(-1) + \beta g(0) + \gamma g(1)$$
et donner l'erreur l'intégration.
2. *Sur une subdivision de pas h de l'intervalle $[a, b]$, approcher $\int_a^b f(x)dx$ en utilisant la formule de quadrature ci-dessus.*

Exercice 7.23. — *Construire les formules d'intégration numérique ci-dessous, degré de précision, erreur d'intégration*
$$\int_a^{a+h} f(x)dx \approx \alpha_0 f(a) + \alpha_1 f(a + \frac{h}{3}) + \alpha_2 f(a + \frac{2h}{3}) + \alpha_3 f(a + h)$$
$$\int_0^1 g(t)\, dt \approx Ag(0) + Bg(t_0)$$
$$\int_{-1}^1 g(t)\, dt \approx \alpha g(-1) + \beta g'(0) + \gamma g(1)$$
$$\int_{-1}^1 g(t)\, dt \approx \alpha g(-1) + \beta g''(0) + \gamma g(1)$$

Exercice 7.24. — *On veut calculer numériquement l'intégrale* $J = \int_0^{+\infty} \frac{1}{1+x^6}dx$.

1. *En majorant $\frac{1}{1+x^6}$ par $\frac{1}{x^6}$, déterminer un nombre $a > 0$ tel que*
$$\int_a^{+\infty} \frac{1}{1+x^6}dx \leq \eta,$$
avec η un réel positif donné.
2. *Déterminer A, B et t_0 pour que la formule de quadrature*
$$\int_0^1 g(t)dt \approx Ag(0) + Bg(t_0) \tag{7.31}$$
soit exacte pour des polynômes de degré deux. Quel est le degré de précision de la méthode ?
3. *On note $R(g) = \int_0^1 g(t)dt - (Ag(0) + Bg(t_0))$. Donner une majoration de l'erreur l'intégration.*
4. *Soit $\{x_i\}$, $i = 0, ..., N$ une subdivision de $[0, a]$ de pas h. Utiliser la formule (7.31) composite pour approcher*
$$I_a = \int_0^a f(x)dx,$$
avec $f(x) = \frac{1}{1+x^6}$. Estimer l'erreur d'intégration en fonction de h.
5. *Dire comment approcher l'intégrale J à ε près.*

Exercice 7.25. — **Intégration 2D par $\mathbb{P}_1$-centre de gravité**

On considère le triangle $\hat{K} = \{(\hat{x}, \hat{y}) \in \mathbb{R}^2; \hat{x} \geq 0, \hat{y} \geq 0, \hat{x} + \hat{y} \leq 1\}$ *et on note* $\hat{G} = (1/3, 1/3)$ *son centre de gravité.*

1. *Déterminer le coefficient* α *pour que la formule d'intégration*
$$\int_{\hat{K}} g(\hat{x}, \hat{y}) d\hat{x} d\hat{y} \approx \alpha g(\hat{G}) \tag{7.32}$$
soit exacte pour les polynômes de plus haut degré.
2. *on note* $\tilde{K}$ *le triangle de sommets B(1,0), D=(1,1) et C=(0,1) et Q le carré* $[0, 1 \times [0, 1]$. *Montrer que*
$$\int_{\tilde{K}} g(\hat{x}, \hat{y}) d\hat{x} d\hat{y} = \int_{\hat{K}} g(1 - \eta, 1 - \xi) d\eta d\xi \tag{7.33}$$
Déduire de ce qui précède une formule d'intégration $\tilde{K}$ *et sur le carré Q.*
3. *Soit* Q_h *un carré de coté h,* $Q_h = [x_0, x_0 + h] \times [y_0, y_0 + h]$. *En déduire une formule d'intégration sur* Q_h.
4. *Soit* $h > 0$, *et* Ω *un rectangle de cotés* Nh *et* Mh, *que l'on discrétise en* NM *carrés de côté h. Expliciter la formule d'intégration composition sur* Ω.

Exercice 7.26. — **Méthode de Gauss-Legendre.** *On considère la formule de quadrature suivante*
$$\int_{-1}^{1} f(x) dx \approx \alpha f(-\frac{1}{\sqrt{3}}) + \beta f(\frac{1}{\sqrt{3}}).$$

1. *Déterminer cette formule pour que de degré de précision soit le plus haut possible et montrer que le degré de précision est* 3.
2. *On suppose* $f \in \mathcal{C}^4([-1, 1])$, *donner l'erreur d'interpolation de Hermite construit sur les abscisses de Gauss* $x_0 = \frac{-1}{\sqrt{3}}$ *et* $x_1 = \frac{1}{\sqrt{3}}$. *En déduire une majoration de l'erreur d'intégration.*
3. *Donner les racines du polynôme de Legendre de degré 2, que conclure ?*
4. *Comment choisir* $(x_0, x_1, \alpha, \beta)$ *dans la formule*
$$\int_a^b f(x) dx \approx \alpha f(x_0) + \beta f(x_1)$$
pour que le degré de précision soit maximal.
5. *Construire une méthode de degré de précision 5, 7, 2n+1.*

Exercice 7.27. — **Gauss-Tchebytchef** *On considère*
$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1 - x^2}} dx \approx \alpha f(x_0) + \beta f(x_1).$$
Comment choisir x_0, x_1, α, β *pour que le degré de précision soit maximal. Donner une majoration de l'erreur d'intégration. Donner une approximation de* $\int_{-1}^{1} \frac{Arccosx}{\sqrt{1-x^2}} dx$.

7.5. Equations différentielles

Exercice 7.28. — *Méthodes à un pas*

Soit $f : [a, b] \times \mathbb{R} \longrightarrow \mathbb{R}$ *uniformément L-Lipschitz par rapport à la deuxième variable. Soit le problème de Cauchy*

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(a) = y_a \end{cases}$$

Soit $N > 0$ *, on pose* $h = (b-a)/N$ *et* $t_i = a + ih$ *pour* $0 \leq i \leq N$.

1. *On considère la méthode d'Euler*

$$y_{n+1} = y_n + hf(t_n, y_n).$$

Montrer que la méthode est consistante et stable.

2. *On considère le schéma de Runge-Kutta 2 (Euler modifié) suivant*

$$y_{n+1} = y_n + hf(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n))$$

consistance ? ordre ? convergence ?

3. *On considère le schéma de Taylor à un pas suivant*

$$y_{n+1} = y_n + h\Phi(t_n, y_n, h)$$

avec $\Phi(t_n, y_n, h) = f(t_n, y_n) + \frac{h}{2} g_2(t_n, y_n)$ *et* $g_2(t_n, y_n) = \frac{\partial f}{\partial t}(t_n, y_n) + f(t_n, y_n)\frac{\partial f}{\partial y}(t_n, y_n)$. *Montrer que ce schéma est consistant et donner l'ordre de consistance. Montrer que si* g_2 *et* L_2*-Lipschitz le schéma est consistant et convergent. Construire des schémas d'ordres supérieurs.*

4. *Dans les méthodes à un pas, on considère*

$$\Phi(t, y, h) = a_1 f(t, y) + a_2 f(t + p_1 h, y + p_2 h f(t, y)).$$

Choisir a_1*,* a_2*,* p_1*,* p_2 *pour que la méthode soit d'ordre 2. Etudier la convergence.*

5. *Appliquer les schémas précédents à l'équation*

$$\begin{cases} y'(t) = y(t) sint \;\; t \in (0, \pi) \\ y(0) = 1 \end{cases}$$

Exercice 7.29. — **Méthodes implicites**

Soit $f : [a, b] \times \mathbb{R} \longrightarrow \mathbb{R}$ *uniformément L-Lipschitz par rapport à la deuxième variable. Soit le problème de Cauchy*

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(a) = y_a \end{cases}$$

Soit $N > 0$ *, on pose* $h = (b-a)/N$ *et* $t_i = a + ih$ *pour* $0 \leq i \leq N$. *On propose le schéma suivant pour approcher la solution* $y(t)$ *:*

$$y_{n+1} = y_n + \frac{h}{2} \left(f(t_n, y_n) + f(t_{n+1}, y_{n+1}) \right).$$

1. *Montrer que si $h < 2/L$ le schéma est bien défini. Donner l'ordre de ce schéma.*
2. *Montrer que pour tout $h < h_0 < 2/L$ assez petit, on a*

$$|y(t_{n+1}) - y_{n+1}| \leq \left(1 + \frac{hL}{1 - \frac{hL}{2}}\right) |y(t_n) - y_n| + \frac{ch^3}{1 - \frac{hL}{2}},$$

avec c une constante positive..
3. *En déduire que*

$$\max_{0 \leq n \leq N} |y(t_n) - y_n| \leq Ch^2.$$

4. *On considère à présent le θ-schéma, défini par*

$$y_{n+1} = y_n + hf\left(\theta t_{n+1} + (1-\theta)t_n, \theta y_{n+1} + (1-\theta)y_n\right).$$

Sous quelles conditions le schéma est bien défini. Consistance et ordre du schéma ?

Exercice 7.30. — *On considère l'équation différentielle*

$$y'(t) = -y(t)^2, \ t \in [0,T], \ \text{et } y(0) = 1. \tag{7.34}$$

Pour calculer une solution approchée, on propose le schéma suivant

$$y_{n+1} = y_n - hy_n y_{n+1}$$

avec $t_i = ih$ et $h = T/N$.

1. *Montrer que $0 < y_n \leq 1$ pour tout n.*
2. *Montrer que le schéma est consistant et donner l'ordre de consistance.*
3. *Calculer la solution exacte de (7.34). Calculer $y(t_1)$, $y(t_2)$ et $y(t_n)$ en fonction de h. Calculer y_1, y_2 et y_n en fonction de h. Conclure.*
4. *Montrer que le schéma est convergent.*
5. *On considère maintenant le système d'équations différentielles*

$$x'(t) = y(t)^2 - x(t), \ x(0) = x_0,$$

$$y'(t) = -y(t)^2 + x(t), \ y(0) = y_0.$$

On considère le schéma suivant

$$x_{n+1} = x_n + h(y_n y_{n+1} - x_{n+1}) \tag{7.35}$$

$$y_{n+1} = y_n + h(-y_n y_{n+1} + x_{n+1}) \tag{7.36}$$

   (a) *Montrer que $x_n + y_n = x_0 + y_0$, pour tout n.*
   (b) *Soit $Z_n = (x_n, y_n)^T$, Montrer que Z_{n+1} est défini comme solution d'un système linéaire que l'on écrira. Montrer que si Z_0 est positif alors Z_n est positif.*

Exercice 7.31. — *On considère l'équation différentielle suivante :*

$$\begin{cases} y'(t) = t - t\,y(t)\,, \quad t \in]0,1[\,, \\ y(0) = 0. \end{cases} \tag{7.37}$$

1. *Calculer $y''(t)$ en fonction de t et $y(t)$.*
2. *On note*

$$f(t,y) = t\,(1-y)\,, \qquad \textit{pour } \; t \in [0,1],\; y \in \mathbb{R}\,,$$

et

$$Df(t,y) = \frac{\partial f}{\partial t}(t,\,y) + f(t,\,y)\,\frac{\partial f}{\partial y}(t,\,y)\,.$$

On propose le schéma numérique à un pas suivant : étant donné un entier $N \geq 1$, on définit une subdivision régulière de l'intervalle $[0,1]$ par :

$$t_i = i\,h\,, \quad i = 0 \cdots N, \;\; \textit{où } h = 1/N\,.$$

Pour $i = 0 \cdots N$, on approche $y(t_i)$ par y_i défini par récurrence par

$$\begin{cases} y_0 = 0 \\ y_{i+1} = y_i + h\,f(t_i,\,y_i) + \frac{h^2}{2}\,Df(t_i,\,y_i)\,, \end{cases} \tag{7.38}$$

i) Soit une suite de réels positifs (u_i) vérifant

$$\begin{cases} u_0 = 0\,, \\ u_{i+1} \;\leq\; (1+a)\,u_i + b \quad \textit{pour } i \in \mathbb{N}\,, \end{cases} \tag{7.39}$$

où a et b sont des constantes strictement positive. Montrer que

$$u_i \;\leq\; b\,\frac{e^{a\,i} - 1}{a}\,, \quad \textit{pour } i \in \mathbb{N}\,. \tag{7.40}$$

ii) Quelle relation a-t-on entre $y''(t_i)$ et $Df(t_i, y(t_i))$.

iii) Ecrire le développement de Taylor à l'ordre 3 de $y(t_i + h)$ par rapport à h.

iv) Montrer que les erreurs d'approximation $e_i = y(t_i) - y_i$ vérifient

$$|e_i| \;\leq\; \frac{h^2}{6}\,\frac{e^{K t_i} - 1}{K}\,M_3\,, \quad \textit{pour } i = 0, \cdots, N\,.$$

où $M_3 = \sup_{t\in[0,1]} |y^{(3)}(t)|$ et K est une constante positive (indépendante de h et de N) à préciser. Quel est l'ordre de convergence de la méthode ?

v) Résoudre (7.37).

Exercice 7.32. — *Résolution d'un système Soit $f \in \mathcal{C}^2(\mathbb{R})$. On considère une équation différentielle de la forme $y' = f(y)$, $y(0) = y_0$. On introduit un pas de discrétisation h et pour tout $n \in \mathbb{N}$, on approche $y(nh)$, par y_n évalué par le schéma*

$$y_{n+1} = y_n + \alpha h f(y_n) + \beta h f(y_n + \gamma h f(y_n))$$

où α, β, γ sont des paramètres réels à déterminer.

1. *Déterminer des relations entre α, β et γ pour que la méthode soit précise à l'ordre deux.*
2. *On suppose que f vérifie une condition de Lipschitz sur $\mathbb{R}$. Montrer que la méthode converge.*
3. *On considère maintenant l'équation du second ordre $y'' = -y$, $y(0) = y_0$, $y'(0) = z_0$, avec y_0, z_0 réels donnés. Mettre cette équation sous la forme d'un système du premier ordre (on notera $Y = (y, z)$ la solution), puis discrétiser en utilisant la schéma précédent. On obtient une méthode de la forme $Y_{n+1} = A_h Y_n$.*
4. *Discuter du comportement de Y_n lorsque n tend vers l'infini, dans le cas d'une précision au second ordre.*
5. *Existe-il une méthode du premier ordre pour laquelle Y_n reste bornée ?*
6. *reprendre les questions précédentes lorsque l'équation est de la forme $y'' = -y - by'$, avec $b > 0$.*

7.6. TA – 2007 avec correction

Exercice 7.33. — **Méthodes des gradients** *Soit A une matrice symétrique définie positive, et $\bar{x}$ tel que $A\bar{x} = b$.*

1. *Montrer que $\bar{x}$ réalise le minimum de $J(x) = (Ax, x) - 2(b, x)$ et de $E(x) = (A(x - \bar{x}), x - \bar{x})$.*

   Les extremas de J sont solutions de $J'(\bar{x}) = 0 = 2(A\bar{x} - b)$, soit $A\bar{x} = b$. De plus, $J''(\bar{x}) = 2A$ une matrice sym. définie positive, alors $\bar{x}$ est un min.
   On a $E(x) = (Ax, x) - 2(A\bar{x}, x) + (A\bar{x}, \bar{x}) = (Ax, x) - 2(b, x) + (A\bar{x}, \bar{x}) = J(x) + (A\bar{x}, \bar{x})$, comme $(A\bar{x}, \bar{x})$ est une constante, E et J atteignent leur min au même point $\bar{x}$.

2. *Montrer $E(x) =< r(x), A^{-1}r(x) >$ avec $r(x) = b - Ax$.*

   Car $r(x) = A(\bar{x} - x)$.

3. *Soit $r_k = b - Ax_k \neq 0$ une direction de descente. On considère la méthode itérative suivante : $x_{k+1} = x_k + \alpha_k r_k$. Trouver α_k celui qui minimise $E(x_{k+1})$, c'est à dire : trouver α_k tel que $E(x_k + \alpha_k r_k) = \min_{\alpha \in \mathbb{R}} E(x_k + \alpha r_k)$. Cette méthode est celle du gradient à paramètre optimal.*

   On a

$$E(x_k + \alpha r_k) =< A(x_k + \alpha r_k - \bar{x}), x_k + \alpha r_k - \bar{x} >= E(x_k) - 2\alpha < r_k, r_k > +\alpha^2 < Ar_k, r_k > .$$

   C'est une équation du second degré en α et $< Ar_k, r_k >> 0$ car A est sym. définie positive, alors le min est atteint par

$$\alpha_k =< r_k, r_k > / < Ar_k, r_k > .$$

4. *Montrer que $r_{k+1} = r_k - \alpha_k Ar_k$ et $(r_k, r_{k+1}) = 0$.*

   On a $r_{k+1} = b - Ax_{k+1} = b - A(x_k + \alpha_k r_k) = r_k - \alpha_k Ar_k$.

   On a $< r_k, r_{k+1} >=< r_k, r_k - \alpha_k Ar_k >=< r_k, r_k > -\alpha_k < r_k, Ar_k >= 0$ par définition de α_k.

5. *On note $E_k =< r_k, A^{-1}r_k >$. Montrer que $E_{k+1} = E_k(1 - \frac{<r_k, r_k>^2}{<A^{-1}r_k, r_k><Ar_k, r_k>})$.*

   On a $E_{k+1} =< r_{k+1}, A^{-1}r_{k+1} >=< r_{k+1}, A^{-1}(r_k - \alpha_k Ar_k) >=< r_{k+1}, A^{-1}r_k > -\alpha_k < r_{k+1}, r_k >$ et comme $< r_{k+1}, r_k >= 0$ d'où le résultat. Ensuite,

$$E_{k+1} =< r_{k+1}, A^{-1}r_k >=< r_k - \alpha_k Ar_k, A^{-1}r_k >=$$
$$< r_k, A^{-1}r_k > -\alpha_k < Ar_k, A^{-1}r_k >= E_k - \alpha_k < r_k, r_k > \text{ car } A \text{ est symétrique.}$$

   Ainsi, de la définition de α_k on a

$$E_{k+1} = E_k - \frac{< r_k, r_k >^2}{< Ar_k, r_k >} = E_k\Big(1 - \frac{< r_k, r_k >^2}{< A^{-1}r_k, r_k >< Ar_k, r_k >}\Big).$$

6. *Montrer que $E_{k+1} \leq E_k(1 - 1/cond_2(A))$. En déduire que la méthode est convergente.*

   La matrice est sym. définie positive alors ses valeurs propres sont strictement positives et $cond_2(A) = \lambda_{max}/\lambda_{min}$. On sait (feuile TD1) que le quotient de Rayleich vérifie :

$$\lambda_{min}\|r_k\|^2 \leq< Ar_k, r_k >\leq \lambda_{max}\|r_k\|^2$$

   et

$$\frac{1}{\lambda_{max}}\|r_k\|^2 \leq< A^{-1}r_k, r_k >\leq \frac{1}{\lambda_{min}}\|r_k\|^2$$

Ainsi,

$$\frac{\lambda_{min}}{\lambda_{max}}\|r_k\|^4 \leq < Ar_k, r_k >< A^{-1}r_k, r_k > \leq \frac{\lambda_{max}}{\lambda_{min}}\|r_k\|^4$$

soit encore

$$\frac{1}{cond_2(A)} \leq \frac{< Ar_k, r_k >< A^{-1}r_k, r_k >}{\|r_k\|^4} \leq cond_2(A),$$

On déduit immédiatement ce qu'il faut.

Convergence. Par récurrence, on a

$$E_{k+1} \leq E_k(1 - \frac{1}{cond_2(A)}) \leq E_0(1 - \frac{1}{cond_2(A)})^k.$$

or $cond_2 A \geq 1$, *ainsi* $0 \leq 1 - \frac{1}{cond_2(A)} < 1$ *et alors* $E_k \longrightarrow 0$ *quand* $k \to \infty$.
D'autre part, $E_k =< A(x_k - \bar{x}), x_k - \bar{x} > \geq \lambda_{min}\|x_k - \bar{x}\|^2$ *alors* $\|x_k - \bar{x}\|^2$ *tend vers zéro quand* k *tend vers l'infinie.*

7. *Appliquer cet algorithme à la matrice*

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

La méthode convergence en deux itérations.

8. *Dire comment peut-on appliquer cet algorithme sur une matrice quelconque et inversible.*

Il suffit d'appliquer la méthode des gradients sur le système linéaire $A^TAx = A^Tb$. *La matrice* A^TA *est symétrique définie positive.*

Exercice 7.34. — *Intégration 1D Construire la formule d'intégration suivante*

$$\int_{-1}^{1} g(t)dt \approx \alpha g(-1) + \beta g''(0) + \gamma g(1),$$

degré de précision, erreur d'intégration ?. (Indication : faire le développement de Taylor pour $g(t)$, $g(-1)$, $g(1)$ *au point* 0 *à l'ordre 4.) Détailler la méthode composite pour approcher* $J(f) = \int_a^b f(x)dx$ *en utilisant la formule d'intégration précédente. Donner l'erreur d'intégration globale en fonction de* h.

pour $g(t) = 1$ *alors* $2 = \alpha + \gamma$
pour $g(t) = t$ *alors* $0 = -\alpha + \gamma$
pour $g(t) = t^2$ *alors* $2/3 = \alpha + 2\beta + \gamma$
donc pour $\alpha = \gamma = 1$ *et* $\beta = -2/3$, *la méthode est exacte sur* $\mathbb{P}_2$. *D'autre part pour* $g(t) = t^3$, *on a* $\int_{-1}^{1} g(t)dt = 0$ *et* $g(-1) - \frac{2}{3}g''(0) + g(1) = 0$ *et la méthode est exacte sur* $\mathbb{P}_3$. *Pour* $g(t) = t^4$, $\int_{-1}^{1} g(t)dt = 2/5$ *et* $g(-1) - \frac{2}{3}g''(0) + g(1) = 2$, *ainsi le degré de précision de la méthode est* 3.

Erreur d'intégration.

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2}g''(0) + \frac{t^3}{6}g^{(3)}(0) + \frac{t^4}{24}g^{(4)}(\xi_t), \;\; 0 < \xi_t < |t|$$

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(0) + \frac{1}{6}g^{(3)}(0) + \frac{1}{24}g^{(4)}(\xi_1), \;\; 0 < \xi_1 < 1$$

$$g(-1) = g(0) - g'(0) + \frac{1}{2}g''(0) - \frac{1}{6}g^{(3)}(0) + \frac{1}{24}g^{(4)}(\xi_2), \;\; -1 < \xi_2 < 0)$$

ainsi,

$$\int_{-1}^{1} g(t)dt = 2g(0) + \frac{1}{3}g''(0) + \frac{1}{24}\int_{-1}^{1} t^4 g^{(4)}(\xi_t)dt$$

$$g(1) + g(-1) = 2g(0) + g''(0) + \frac{1}{24}(g^{(4)}(\xi_1) + (g^{(4)}(\xi_2)).$$

On a

$$R(g) = \int_{-1}^{1} g(t)dt - (\frac{1}{3}g(-1) - \frac{2}{3}g''(0) + \frac{1}{3}g(1) = \frac{1}{24}\int_{-1}^{1} t^4 g^{(4)}(x_t)dt - \frac{1}{24}(g^{(4)}(\xi_1) - (g^{(4)}(\xi_2))$$

en majorant, on a

$$|R(g)| \leq (\frac{1}{60} + \frac{1}{24} + \frac{1}{24}) \max_{-1\leq t\leq 1} |g^{(4)}(t)|.$$

Formule Composite. $J(f) = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} f(x)dx$ *et sur chaque intervalle de longueur h, on applique la formule d'intégration précédente*

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{h}{2}\int_{-1}^{1} f(x_i + \frac{1+t}{2}h)dt \approx \frac{h}{2}\Big(\frac{1}{3}f(x_i) - \frac{2}{3}\frac{h^2}{4}f''(x_i + h/2) + \frac{1}{3}f(x_{i+1})\Big),$$

on a appliqué la formule avec $g(t) = f(x_i + \frac{1+t}{2}h)$.

Erreur d'intégration sur $[x_i, x_{i+1}]$ *:* $R_i(f) = \frac{h}{2}R(g)$. *on a*

$$|R_i(f)| = \frac{h}{2}|R(g)| \leq \frac{h}{20} \max_{-1\leq t\leq 1} |g^{(4)}(t)| = \frac{h}{20}\frac{h^4}{2^4} \max_{x_i\leq x\leq x_{i+1}} |f^{(4)}(x)|.$$

.

Erreur d'intégration sur $[a, b]$.

$$R_h(f) = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} f(x)\,dx - \frac{h}{2}\sum_{i=0}^{M-1} \Big(\frac{1}{3}f(x_i) - \frac{2}{3}\frac{h^2}{4}f''(x_i + h/2) + \frac{1}{3}f(x_{i+1})\Big) = \sum_{i=0}^{M-1} R_i(f)$$

par conséquent,

$$|R_h(f)| \leq \sum_{i=0}^{M-1} |R_i(f)| \leq \frac{h^5}{320}M \max_{a\leq x\leq b} |f^{(4)}(x)| = \Big(\frac{b-a}{320} \max_{a\leq x\leq b} |f^{(4)}(x)|\Big)h^4.$$

Exercice 7.35. — *Intégration 2D On considère le triangle* $\hat{K} = \{(\hat{x}, \hat{y}) \in \mathbb{R}^2; x \geq 0, y \geq 0, \hat{x} + \hat{y} \leq 1\}$ *et on note* $O = (0, 0)$, $A = (1, 0)$ *et* $C = (0, 1)$ *les sommets*

1. *Déterminer les coefficient* α, β *et* γ *tels que la formule d'intégration*

$$\int_{\hat{K}} g(\hat{x}, \hat{y})d\hat{x}d\hat{y} \approx \alpha g(O) + \beta g(A) + \gamma g(C) \tag{7.41}$$

soit exacte pour les polynômes de degré ≤ 1. *Degré de précision ?*

L'espace $\mathbb{P}_1$ *en dimension 2 est engendré par* $< 1, \hat{x}, \hat{y} >$ *et de dimension 3.*
pour $g(\hat{x}, \hat{y}) = 1$, *on a* $\int_{\hat{K}} d\hat{x}d\hat{y} = \frac{1}{2} = \alpha + \beta + \gamma$,
pour $g(\hat{x}, \hat{y}) = \hat{x}$, *on a* $\int_{\hat{K}} \hat{x}d\hat{x}d\hat{y} = \int_0^1 (\int_0^{1-\hat{x}} \hat{x}d\hat{y})d\hat{x} = \int_0^1 \hat{x}(1 - \hat{x})d\hat{x} = \frac{1}{6} = \beta$,
pour $g(\hat{x}, \hat{y}) = \hat{y}$, *on a* $\int_{\hat{K}} \hat{x}d\hat{y}d\hat{y} = \int_0^1 (\int_0^{1-\hat{x}} \hat{y}d\hat{y})d\hat{x} = \frac{1}{6} = \gamma$.
Donc $\alpha = \beta = \gamma = 1/6$ *et la méthode est exacte pour les polynômes de degré 1.*

Ensuite, pour $g(\hat{x},\hat{y})=\hat{x}^2$, *on a* $\int_{\hat{K}}\hat{x}^2d\hat{x}d\hat{y}=\int_0^1(\int_0^{1-\hat{x}}\hat{x}^2d\hat{y})d\hat{x}=\int_0^1\hat{x}^2(1-\hat{x})d\hat{x}=\frac{1}{12}\neq\frac{1}{6}$. *Ce qui prouve que la méthode n'est pas exacte pour* $\mathbb{P}_2$ *et le degré de précision est 1.*

2. *On rappelle la formule de Taylor :* $\forall a,\,b\in\mathbb{R}^2$,

$$g(b)=g(a)+g'(a)(b-a)+\int_0^1(1-t)g''(ta+(1-t)b)(b-a,b-a)dt$$

Soit $a_0=(x_0,y_0)$ *un point fixe dans* $\hat{K}$. *Appliquer la formule de Taylor précédente en considérant successivement* $b=(x,y)$, $b=O$, $b=A$ *et* $b=C$. *Donner alors une estimation de l'erreur lorsque* $g\in C^2(\mathbb{R}^2)$:

$$|R_{\hat{K}}(g)|\leq 2M_{\hat{K}}(g)$$

avec $R_{\hat{K}}(g)=\int_{\hat{K}}g(\hat{x},\hat{y})d\hat{x}d\hat{y}-(\alpha g(O)+\beta g(A)+\gamma g(C))$ *et*

$$M_{\hat{K}}(g)=\max(\sup_{(\hat{x},\hat{y})\in\hat{K}}|\partial^2_{\hat{x}\hat{x}}g|,\sup_{(\hat{x},\hat{y})\in\hat{K}}|\partial^2_{\hat{y}\hat{y}}g|,\sup_{(\hat{x},\hat{y})\in\hat{K}}|\partial^2_{\hat{x}\hat{y}}g|).$$

On applique Taylor avec $b=(\hat{x},\hat{y})$ *et* $a=(x_0,y_0)$

$$g(\hat{x},\hat{y})=g(x_0,y_0)+g'(x_0,y_0)(\hat{x}-x_0,\hat{y}-y_0)+R_0(g) \tag{7.42}$$

avec $R_0(g)=\int_0^1(1-t)g''(ta_0+(1-t)b)(b-a,b-a)dt$. *On intègre sur* $\hat{K}$, *il vient*

$$\int_{\hat{K}}g(\hat{x},\hat{y})d\hat{x}d\hat{y}=\frac{1}{2}g(x_0,y_0)+\partial_xg(x_0,y_0)\int_{\hat{K}}(\hat{x}-x_0)d\hat{x}d\hat{y}$$
$$+\partial_yg(x_0,y_0)\int_{\hat{K}}(\hat{y}-y_0)d\hat{x}d\hat{y}+\int_{\hat{K}}R_0(g)d\hat{x}d\hat{y}$$

soit encore

$$\int_{\hat{K}}g(\hat{x},\hat{y})d\hat{x}d\hat{y}=\frac{1}{2}g(x_0,y_0)+(\frac{1}{6}-\frac{x_0}{2})\partial_xg(x_0,y_0)+(\frac{1}{6}-\frac{y_0}{2})\partial_yg(x_0,y_0)+\int_{\hat{K}}R_0(g)d\hat{x}d\hat{y}$$

On applique encore Taylor successivement avec $b=O$, $b=A$, $b=C$ *et* $a=a_0$

$$g(0,0)=g(x_0,y_0)+g'(x_0,y_0)(-x_0,-y_0)+R_1(g) \tag{7.43}$$
$$g(1,0)=g(x_0,y_0)+g'(x_0,y_0)(1-x_0,-y_0)+R_2(g) \tag{7.44}$$
$$g(0,1)=g(x_0,y_0)+g'(x_0,y_0)(-x_0,1-y_0)+R_3(g) \tag{7.45}$$

avec

$$R_1(g)=\int_0^1(1-t)g''(ta_0+(1-t)O)(a_0-O,a_0-O)dt$$
$$R_1(g)=\int_0^1(1-t)g''(ta_0+(1-t)A)(a_0-A,a_0-A)dt$$
$$R_1(g)=\int_0^1(1-t)g''(ta_0+(1-t)C)(a_0-C,a_0-C)dt$$

d'où

$$\frac{1}{6}(g(0,0)+g(1,0)+g(0,1))=\frac{1}{2}g(x_0,y_0)+g'(x_0,y_0)(\frac{1}{6}-\frac{x_0}{2},\frac{1}{6}-\frac{y_0}{2})+\frac{1}{6}(R_1(g)+R_2(g)+R_3(g))$$

ainsi

$$R_{\hat{K}}(g)=\int_{\hat{K}}g(\hat{x},\hat{y})d\hat{x}d\hat{y}-\frac{1}{6}(g(O)+g(A)+g(C))=\int_{\hat{K}}R_0(g)d\hat{x}d\hat{y}-\frac{1}{6}(R_1(g)+R_2(g)+R_3(g)).$$

Pour $i = 0, 1, 2, 3$ on a

$$R_i(g) = \int_0^1 (1-t)g''(\cdot)((\xi,\eta),(\xi,\eta))dt = \int_0^1 (1-t)(\partial^2_{\hat{x}\hat{x}}g(\cdot) + \eta^2\partial^2_{\hat{y}\hat{y}}g(\cdot) + 2\xi\eta\partial^2_{\hat{x}\hat{y}}g(.))dt,$$

avec $\xi = x - x_0$ ou $1 - x_0$ ou $-x_0$ et comme x, $x_0 \in [0,1]$ alors$|\xi| \leq 1$, de même $\eta = y - y_0$ ou $1 - y_0$ ou $-y_0$ et comme y, $y_0 \in [0,1]$ alors$|\eta| \leq 1$, en majorant R_i, on a

$$|R_i(g)| \leq M_{\hat{K}}(g)\int_0^1 (1-t)(\xi^2 + \eta^2 + 2|\xi||\eta|)dt \leq 4M_{\hat{K}}(g)\int_0^1 (1-t)dt = 2M_{\hat{K}}(g)$$

enfin

$$|R_{\hat{K}}(g)| \leq \frac{1}{2}(2M_{\hat{K}}) + \frac{1}{6}3(2M_{\hat{K}}) = 2M_{\hat{K}}.$$

3. *On note $\tilde{K}$ le triangle de sommets A, B=(1,1) et C, et Q le carré $[0,1\times[0,1]$. Déduire de ce qui précède une formule d'intégration sur le carré, et une estimation de l'erreur.*

Sur $\tilde{K}$, on applique la formule précédente en utilisant les sommets du triangle, il vient

$$\int_{\tilde{K}} g(\hat{x},\hat{y})d\hat{x}d\hat{y} \approx \frac{1}{6}(g(A) + g(B) + g(C))$$

et l'erreur d'intégration est comme précédemment :

$$R_{\tilde{K}}(g) \leq 2M_{\tilde{K}}(g).$$

Le carré est la réunion de deux triangles :

$$\int_Q g(\hat{x},\hat{y})d\hat{x}d\hat{y} = \int_{\hat{K}} g(\hat{x},\hat{y})d\hat{x}d\hat{y} + \int_{\tilde{K}} g(\hat{x},\hat{y})d\hat{x}d\hat{y}$$
$$\approx \frac{1}{6}(g(O) + g(A) + g(C)) + \frac{1}{6}(g(A) + g(B) + g(C))$$

et l'erreur d'intégration est majorée comme suit

$$|R_Q| \leq 2M_{\hat{K}}(g) + 2M_{\tilde{K}}(g) \leq 4M_Q$$

avec

$$M_Q(g) = \max(\sup_{(\hat{x},\hat{y})\in Q}|\partial^2_{\hat{x}\hat{x}}g|, \sup_{(\hat{x},\hat{y})\in Q}|\partial^2_{\hat{y}\hat{y}}g|, \sup_{(\hat{x},\hat{y})\in Q}|\partial^2_{\hat{x}\hat{y}}g|).$$

4. *Soit $h > 0$, et Ω un rectangle de cotés $Nh = l$ et $Mh = L$, que l'on discrétise en NM carrés de côté h. Soit $f \in C^2(\Omega)$. Etablir une estimation $R_h(f)$ de*

$$I(f) = \int\int_\Omega f(x,y)dxdy, \tag{7.46}$$

par la méthode composite et donner une estimation d'erreur.

Le rectangle Ω s'écrit : $\Omega = \cup_{\{1\leq i\leq N, 1\leq j\leq M\}}Q_{ij}$ avec Q_{ij} est un rectangle de côté h, plus précisément $Q_{ij} = [x_i, x_i + h] \times [y_j, y_j + h]$.

$$I(f) = \int\int_\Omega f(x,y)dxdy = \sum_{1\leq i\leq N, 1\leq j\leq M}\int_{Q_{ij}} f(x,y)dxdy. \tag{7.47}$$

Approximation de $I_{ij} = \int_{Q_{ij}} f(x,y)dxdy = \int_{x_i}^{x_i+h}\int_{y_i}^{y_j+h} f(x,y)dxdy$. *On pose* $\xi = \frac{x-x_i}{h}$ *et* $\eta = \frac{y-y_j}{h}$, *ainsi*

$$\begin{aligned}
I_{ij}(f) &= h^2\int_0^1\int_0^1 f(x_i+h\xi, y_j+h\eta)d\xi d\eta = h^2\int_Q g(\xi,\eta)d\xi d\eta\\
&\approx \frac{1}{6}(g(O)+g(A)+g(C)) + \frac{1}{6}(g(A)+g(B)+g(C))\\
&\approx \frac{h^2}{6}(f(x_i,y_j)+f(x_i+h,y_j)+f(x_i,y_j+h))\\
&+ \frac{h^2}{6}(f(x_i+h,y_j)+f(x_i,y_j+h)+f(x_i+h,y_j+h))
\end{aligned}$$

ainsi l'erreur d'intégration sur Q_{ij} *est donnée par*

$$|R_{i,j}(f)| \le 4h^2 M_Q(g) \le 4h^4 M_{i,j}(f),$$

avec

$$M_{i,j}(f) = \max(\sup_{(x,y)\in Q_{ij}}|\partial^2_{xx}f|, \sup_{(x,y)\in Q_{ij}}|\partial^2_{yy}f|, \sup_{(x,y)\in Q_{ij}}|\partial^2_{xy}f|).$$

Enfin, l'erreur d'intégration sur Ω *par la formule composite s'écrit*

$$\begin{aligned}
R_\Omega(f) &= \sum_{1\le i\le N, 1\le j\le M}\int_{Q_{ij}} f(x,y)dxdy\\
&- h^2\sum_{1\le i\le N, 1\le j\le M}\Big(\frac{1}{6}(f(x_i,y_j)+f(x_i+h,y_j)+f(x_i,y_j+h))\\
&+ \frac{1}{6}(f(x_i+h,y_j)+f(x_i,y_j+h)+f(x_i+h,y_j+h))\Big)\\
&= \sum_{1\le i\le N, 1\le j\le M} R_{i,j}(f),
\end{aligned}$$

et alors

$$|R_\Omega(f)| \le \sum_{1\le i\le N, 1\le j\le M}|R_{i,j}(f)| \le 4h^4 NM M_\Omega(f),$$

avec

$$M_\Omega(f) = \max(\sup_{(x,y)\in\Omega}|\partial^2_{xx}f|, \sup_{(x,y)\in\Omega}|\partial^2_{yy}f|, \sup_{(x,y)\in\Omega}|\partial^2_{xy}f|),$$

or $Nh = l$ *et* $Mh = L$, *on a finalement*

$$|R_\Omega(f) \le 4lLM_\Omega(f)h^2.$$

5. *Application. Soit* $\epsilon > 0$. *Déterminer* h *maximal de telle façon que* $R_h(f)$ *approche*

$$I(f) = \int_0^2\int_0^3 e^{-x^2y^3}dxdy$$

avec une erreur majorée par ε.

On a $f(x,y) = e^{-x^2y^3}$, $l = 2$, $L = 3$. *On dérivant la fonction* f, *on peut donner une estimation des dérivées d'ordre 2, par exemple :*

$$M_\Omega(f) \le 11736,$$

ainsi

$$|R_\Omega(f)| \leq 24 \times 11736h^2 \leq \varepsilon$$

d'où h.

Exercice 7.36. — *Equations différentielles : Une méthode à un pas* *Soit $g \in C^1(\mathbb{R})$, à valeurs positives. On considère une équation différentielle du premier ordre, de la forme $y' = f(y)$, que l'on discrétise par une méthode à un pas de la forme*

$$y_{n+1} = \frac{y_n}{1 + hg(y_n)}$$

où h est le pas de discrétisation, $n \in \mathbb{N}$ et y_n approche $y(nh)$.

1. *Ecrire cette méthode sous la forme générique d'une méthode à un pas. Exprimer f en fonction de g pour que la méthode soit consistance.*

   C'est une méthode à un pas,

   $$y_{n+1} = y_n + h\Phi(t_n, y_n, h) = \frac{y_n}{1 + hg(y_n)}$$

   alors

   $$\Phi(t_n, y_n, h) = \frac{-y_n g(y_n)}{1 + hg(y_n)}.$$

   Le schéma est consistant ssi

   $$\Phi(t_n, y_n, 0) = -y_n g(y_n) = f(t_n, y_n)$$

   ainsi

   $$f(y) = -yg(y).$$

2. *On suppose que $y(0) = y_0$. Montrer que pour tout $n \in \mathbb{N}$ $y_n \in [0, y_0]$.*

   Par récurrence. $1 + hg(y_n) \geq 1$ car g est positive, et $y_{n+1} = \frac{y_n}{1+hg(y_n)} \leq y_n \leq y_0$.

3. *En déduire la stabilité, puis la convergence de la méthode.*

   On a

   $$\Phi(t, y, h) = \frac{-yg(y)}{1 + hg(y)},$$

   avec $0 \leq y \leq y_0$ et et donc les fonctions g et g' sont uniformément continues sur $[0, y_0]$. On note $M_1 = max_{0\leq y\leq y_0}|g(y)|$ et $M_2 = max_{0\leq y\leq y_0}|g'(y)|$
   Stabilité :

   $$|\Phi(t, y, h) - \Phi(t, z, h)| = \Big|\frac{yg(y)}{1 + hg(y)} - \frac{zg(z)}{1 + hg(z)}\Big|$$
   $$\leq |yg(y)(1 + hg(z)) - zg(z)(1 + hg(y))| = |(yg(y) - zg(z)) + hg(z)g(y)(y - z)|$$
   $$= |y(g(y) - g(z)) + g(z)(y - z) + hg(z)g(y)(y - z)|$$
   $$\leq |y||g(y) - g(z)| + |g(z)||y - z| + h|g(z)g(y)||y - z|,$$

   ensuite, on a $|g(y) - g(z)| \leq M_2|y - z|$(th. des accroissements finis), on déduit

   $$|\Phi(t, y, h) - \Phi(t, z, h)| \leq M_2|y_0||y - z| + M_1|y - z| + M_1^2 h|y - z|,$$

   et enfin pour h petit, c'est à dire pour $h \leq h_0$ et h_0 fixe, on a

   $$|\Phi(t, y, h) - \Phi(t, z, h)| \leq M|y - z|,$$

   avec $M = M_2|y_0| + M_1 + M_1^2 h_0$. Le schéma étant stable et consistent, alors il est convergent.

7.7. TA-2008

Exercice 7.37. — *Méthode des directions alternées Soit A une matrice réelle symétrique définie positive, on décompose A sous la forme*

$$A = rI + H + V$$

avec $r \in \mathbb{R}$, $r > 0$, I la matrice identité, H et V matrices réelles symétriques telles que $rI + H$ et $rI + V$ soient inversibles. On note par la suite $B = \frac{1}{r}H$ et $C = \frac{1}{r}V$. Pour résoudre le système $Ax = b$, on considère la méthode itérative suivante :

$$\begin{cases} x_0 \text{ arbitraire} \\ (rI + H)y_{k+1} = -Vx_k + b \\ (rI + V)x_{k+1} = -Hy_{k+1} + b \end{cases} \tag{7.48}$$

1. *Montrer que la méthode est consistante (c-à-d si $x_k \longrightarrow x$ alors $Ax = b$).*
2. *Montrer que la méthode itérative ci-dessus converge si et seulement si*
$$\rho((rI + V)^{-1}H(rI + H)^{-1}V) < 1.$$
3. *Montrer que les matrices $B(I + B)^{-1}$ et $C(I + C)^{-1}$ sont symétriques.*
4. *Montrer*
$$\rho((rI + V)^{-1}H(rI + H)^{-1}V) \leq \rho(B(I + B)^{-1})\rho(C(I + C)^{-1}).$$
(Indication : montrer que la matrice d'itération est semblable à une autre matrice et utiliser la norme $\|\cdot\|_2$).
5. *Montrer que $\rho(B(I + B)^{-1}) < 1$ si et seulement si $\frac{1}{2}I + B$ est définie positive.*
6. *En déduire que si les matrices $\frac{r}{2}I + H$ et $\frac{r}{2}I + V$ sont définies positives, la méthode itérative est convergente.*
7. *Préciser r, H et V dans le cas de la matrice suivante A*

$$\begin{vmatrix} 4 & -1 & & & -1 & & & & & & \\ -1 & 4 & -1 & & & -1 & & & & & \\ & -1 & 4 & -1 & & & -1 & & & & \\ & & -1 & 4 & -1 & & & -1 & & & \\ -1 & & & -1 & 4 & -1 & & & -1 & & \\ & -1 & & & -1 & 4 & -1 & & & -1 & \\ & & -1 & & & -1 & 4 & -1 & & & -1 \\ & & & -1 & & & -1 & 4 & -1 & & \\ & & & & -1 & & & -1 & 4 & -1 & \\ & & & & & -1 & & & -1 & 4 & -1 \\ & & & & & & -1 & & & -1 & 4 \end{vmatrix}$$

Exercice 7.38. — *Méthodes à deux pas On considère le problème de Cauchy suivant*

$$\begin{cases} y'(t) = f(t, y(t)) \;\; t \in (0, T) \\ y(0) = \eta \end{cases}$$

avec f L-lipschitz par rapport à y. Soit le schéma :

$$\begin{cases} y_{n+1} = y_{n-1} + 2hf(t_n, y_n) \ \ n \geq 1 \\ y_1 = y_0 + hf(0, y_0) \\ y_0 = \eta, \end{cases}$$

avec $h = T/N$, $t_n = nh$.

1. *Majorer les erreurs de consistance*

$$E_0 = y(t_1) - y(0) - hf(0, y(0)), \quad E_n = y(t_{n+1}) - y(t_{n-1}) - 2hf(t_n, y(t_n))$$

en fonction de $M_k = \max_{t\in(0,T)} |y^{(k)}(t)|$.

2. *Montrer que pour tout $x \geq 0$*

$$e^{-2x} + 2xe^{-x} \leq 1; \qquad 2x \leq e^x; \qquad 1 + x^2 \leq e^x.$$

3. *Soit $(\theta_n)_n$ définie par*

$$\theta_{n+1} \leq \theta_{n-1} + 2\beta\theta_n + \alpha_n, \ n \geq 1$$

avec $\theta_i, \alpha_i, \beta \in \mathbb{R}^+$. Montrer que

$$\theta_n \leq e^{(n-1)\beta}\sqrt{\theta_0^2 + \theta_1^2} + \sum_{i=1}^{n-1} \alpha_i e^{(n-1-i)\beta}, \ \ n \geq 2.$$

4. *Montrer que*

$$\sum_{i=1}^{n-1} e^{(n-1-i)hL} \leq \frac{e^{LT} - 1}{hL}.$$

Soit $e_n = y(t_n) - y_n$, $n \geq 0$. Montrer que

$$|e_{n+1}| \leq |e_{n-1}| + 2hL|e_n| + |E_n|.$$

En déduire que $\max_{0\leq n\leq N} |y(t_n) - y_n| \leq Ch^2$, où C est une constante à déterminer en fonction de M_2, M_3 et L. Conclure.

Exercice 7.39. — *Approximation Soit $x_0 \in \mathbb{R}$ fixé et $p \in \mathcal{C}^1([0, 1])$, on considère l'intégrale*

$$\int_0^1 (x - x_0)p(x)\, dx.$$

1. *Pour chacun des cas suivants, construire une formule d'intégration exacte lorsque p est un polynôme de degré inférieur ou égal à 1 :*
   (a) *en utilisant les valeurs $p(0)$ et $p(1)$.*
   (b) *en utilisant qu'une seule valeur $p(a)$ où a est à déterminer.*
2. *Soit $R > 0$, pour $f, g \in \mathcal{C}^0([0, R])$ on considère le produit scalaire*

$$(f, g) = \int_0^R xf(x)g(x)\, dx.$$

Soit $N \in \mathbb{N}$. On pose $h = \frac{R}{N}$, $x_i = ih$ pour $i = 1, \cdots, N$ et

$$V_h = \{f \in \mathcal{C}^0([0, R]); f \text{ affine sur } [x_i, x_{i+1}]\}.$$

Déterminer la dimension de V_h. Soit l'ensemble des fonctions $(\Phi_i)_{i=0}^{i=N}$ où Φ_i est définie par

$$\Phi_i(x) = \begin{cases} (x - x_{i-1})/h & si \quad x_{i-1} \leq x \leq x_i \\ (x_{i+1} - x)/h & si \quad x_i \leq x \leq x_{i+1}, \ pour\ 1 \leq i \leq N-1 \end{cases}$$

$$\Phi_0(x) = \begin{cases} (x_1 - x)/h & si\ 0 \leq x \leq x_1, \\ 0 & sinon \end{cases}$$

$$\Phi_N(x) = \begin{cases} (x - x_{N-1})/h & si\ x_{N-1} \leq x \leq R, \\ 0 & sinon \end{cases}$$

Tracer Φ_i. Montrer que $\{\Phi_0, \cdots, \Phi_N\}$ définie une base de V_h.

3. *On note M la matrice de projection orthogonale définie par $M_{ij} = (\Phi_i, \Phi_j)$. Expliciter M et montrer que M est symétrique, définie positive. Montrer qu'il existe un unique $p_\perp \in V_h$ tel que*
$$(f - p_\perp, \Phi_i) = 0 \ pour\ i = 0, \cdots N.$$
4. *En utilisant les formules d'intégration précédentes, construire deux matrices M_1, M_2 associées à des projections approchées et comparer les matrices obtenues. Pour f donnée, calculer $p_\perp$ dans le cas avec $N = 3$.*

Exercice 7.40. — *Intégration 2D On considère le triangle $\hat{K} = \{(\hat{x}, \hat{y}) \in \mathbb{R}^2; \hat{x} \geq 0, \hat{y} \geq 0, \hat{x} + \hat{y} \leq 1\}$ et on note $A = (0,0)$, $B = (1,0)$ et $C = (0,1)$ ses trois sommets et $\hat{G} = (1/3, 1/3)$ son barycentre.*

1. *Déterminer le coefficient α pour que la formule d'intégration*
$$\int_{\hat{K}} g(\hat{x}, \hat{y}) d\hat{x} d\hat{y} \approx \alpha g(\hat{G}) \tag{7.49}$$
soit exacte pour les polynômes de plus haut degré. Montrer que le degré de précision est 1.
2. *On rappelle la formule de Taylor : $g \in C^2(\mathbb{R}^2)$, $g : \mathbb{R}^2 \mapsto \mathbb{R}$, alors $\forall a, b \in \mathbb{R}^2$,*
$$g(b) = g(a) + g'(a)(b-a) + \int_0^1 (1-t) g''(tb + (1-t)a)(b-a, b-a) dt.$$
Appliquer la formule de Taylor précédente en considérant $b = (\hat{x}, \hat{y})$ et $a = \hat{G}$. Lorsque $g \in C^2(\mathbb{R}^2)$, donner alors une estimation de l'erreur
$$R_{\hat{K}}(g) = \int_{\hat{K}} g(\hat{x}, \hat{y}) d\hat{x} d\hat{y} - \alpha g(\hat{G}))$$
en fonction de
$$M_{\hat{K}}(g) = \max \Big(\sup_{(\hat{x},\hat{y}) \in \hat{K}} |\partial^2_{\hat{x}\hat{x}} g|, \sup_{(\hat{x},\hat{y}) \in \hat{K}} |\partial^2_{\hat{y}\hat{y}} g|, \sup_{(\hat{x},\hat{y}) \in \hat{K}} |\partial^2_{\hat{x}\hat{y}} g| \Big).$$
3. *On note $\tilde{K}$ le triangle de sommets $B(1,0)$, $D = (1,1)$ et $C = (0,1)$ et Q le carré $[0,1 \times [0,1]$. Montrer que*
$$\int_{\tilde{K}} g(\hat{x}, \hat{y}) d\hat{x} d\hat{y} = \int_{\hat{K}} g(1-\eta, 1-\xi) d\eta d\xi \tag{7.50}$$
Déduire de ce qui précède une formule d'intégration sur $\tilde{K}$ et sur le carré Q. Donner une estimation de l'erreur d'intégration sur $\tilde{K}$ et sur Q

4. *Soit Q_h un carré de coté h, $Q_h = [x_0, x_0 + h] \times [y_0, y_0 + h]$. En déduire une formule d'intégration sur Q_h et une estimation de l'erreur d'intégration sur Q_h*
5. *Soit $h > 0$, et Ω un rectangle de cotés $Nh = l$ et $Mh = L$, que l'on discrétise en NM carrés de côté h. Soit $f \in C^2(\Omega)$. Etablir une approximation de*

$$I(f) = \int\int_{\Omega} f(x, y)dxdy, \tag{7.51}$$

par la méthode composite et donner une estimation d'erreur $R_h(f)$.
6. *Application. Soit $\epsilon > 0$. Déterminer h maximal de telle façon que la formule de quadrature approche*

$$I(f) = \int_0^2 \int_0^3 e^{-x^2y^3} dxdy$$

avec une précision ε.

CHAPITRE 8

DEVOIR SURVEILLÉ D'ANALYSE NUMÉRIQUE (2010) ET SON CORRIGÉ

Exercice 1.

Soit (x_i, y_i) $(i = 1, \cdots, n)$ une famille de points donnés. On cherche $a = (a_1, a_2, a_3)$ réalisant le minimum de

$$J(a_1, a_2, a_3) = \sum_{i=1}^{n} x_i^2 (a_1 + a_2 x_i + a_3 x_i^2 - y_i)^2$$

1. *Calculer $\frac{\partial J}{\partial a_j}(a)$, pour $j = 1, 2, 3$.*
2. *Calculer $J"(a) = (\frac{\partial^2 J}{\partial^2 a_j a_k})_{jk}$ pour $j, k = 1, 2, 3$.*
3. *Ecrire le système linéaire caractérisant le minimum du problème*

$$\min_{a \in \mathbb{R}^3} J(a). \tag{8.1}$$

4. *Application. Soit la famille :*

$$(x_1, y_1) = (-1, \frac{3}{2}), \quad (x_2, y_2) = (-\frac{1}{2}, -\frac{3}{2}), \quad (x_3, y_3) = (\frac{1}{2}, 0), \quad (x_4, y_4) = (1, 0).$$

Donner explicitement la solution de (8.1).
Montrer qu'il s'agit bien d'un minimum.

Exercice 2.

Soit la matrice symétrique suivante

$$A = \begin{pmatrix} 3 & -1 & 0 & \cdots & 0 & 1 \\ -1 & 3 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & -1 & 3 & -1 & 0 \\ 0 & \cdots & & -1 & 3 & -1 \\ 1 & 0 & \cdots & 0 & -1 & 3 \end{pmatrix}$$

Plus précisement, $A = (a_{ij})_{ij}$

$$a_{ij} = \begin{cases} 3 & si\ j = i \\ -1 & si\ j = i+1\ ou\ j = i-1 \\ 1 & si\ (i = 1,\ j = N)\ ou\ (i = N,\ j = 1) \\ 0 & sinon \end{cases} \tag{8.2}$$

1. *Montrer que la matrice A est définie positive. Est-elle inversible ? Montrer que $1 \leq \lambda \leq 5$ où λ est une valeur propre de A.*
2. *Soient les matrices $L = (l_{ij})_{ij}$ et $U = (u_{ij})_{ij}$*

$$L = \begin{pmatrix} d_1 & & & & & \\ l_2 & d_2 & & & & \\ 0 & \ddots & \ddots & & & \\ 0 & \ddots & l_{N-2} & d_{N-2} & & \\ 0 & \cdots & & l_{N-1} & d_{N-1} & \\ a_1 & a_2 & \cdots & a_{N-2} & l_N & d_N \end{pmatrix}, \quad U = \begin{pmatrix} 1 & u_1 & 0 & \cdots & 0 & b_1 \\ & 1 & u_2 & 0 & \cdots & b_2 \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & u_{N-2} & b_{N-2} \\ & & & & 1 & u_{N-1} \\ & & & & & 1 \end{pmatrix};$$

soit encore

$$l_{ij} = \begin{cases} d_i & si\ j = i \\ l_i & si\ j = i-1 \\ a_j & si\ i = N \\ 0 & sinon \end{cases}, \qquad u_{ij} = \begin{cases} 1 & si\ j = i \\ u_i & si\ j = i+1 \\ b_i & si\ j = N \\ 0 & sinon \end{cases} \tag{8.3}$$

Donner, par identification, les relations entre les coefficients des matrices L et U pour que $A = LU$. Ecrire un algorithme permettant le calcul des coefficients des matrices L et U.

3. *Soit $A = M - N$ une décomposition de A avec $M = \frac{1}{\omega} I$, où $\omega > 0$ et I la matrice identité dans $\mathbb{R}^{n\,n}$. Les valeurs propres de A sont classées par ordre décroissant*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\ .$$

Pour résoudre le système linéaire $Ax = b$ où b un vecteur non nul de $\mathbb{R}^n$, on considère la méthode itérative :

$$\begin{cases} x_0 \in \mathbb{R}^n\ , \\ M\, x_{k+1} = N\, x_k + b\ . \end{cases} \tag{8.4}$$

Montrer que la méthode est consistante.

4. *Montrer que la méthode converge si et seulement si*

$$0 < \omega < \frac{2}{\lambda_1}\ .$$

5. *Montrer que le rayon spectral de la matrice $B = M^{-1}N$ est donné par*

$$\rho(M^{-1}N) = max\{|1 - \omega\lambda_1|,\ |1 - \omega\lambda_n|\}\ .$$

6. *Montrer que le meilleur choix de ω, celui qui rend la convergence de la méthode la plus rapide, est donné par :*
$$\omega = \frac{2}{\lambda_1 + \lambda_n} .$$
(indication : tracer $\omega \in]0, +\infty[\mapsto \ g_i(\omega) = |1 - \omega\lambda_i| \ , \quad i = 1, n$)

Exercice 3.

Soit $f : \mathbb{R} \mapsto \mathbb{R}$ telle que $f \in \mathcal{C}^3([a,b];\mathbb{R})$ et soit $a < b$ deux réels donnés. On cherche à approcher f sur l'intervalle $[a,b]$ par un polynôme p tel que
$$\begin{cases} p(a) = f(a) \\ p(b) = f(b) \\ p'(a) = f'(a) \end{cases} \tag{8.5}$$

1. *Construire les polynômes λ_1, λ_2 et λ_3 de degré 2 définis par*
$$\begin{cases} \lambda_1(a) = \lambda_1'(a) = 0, \quad \lambda_1(b) = 1 \\ \lambda_2(a) = \lambda_2(b) = 0, \quad \lambda_2'(a) = 1 \\ \lambda_3(b) = \lambda_3'(a) = 0, \quad \lambda_3(a) = 1 \end{cases}$$
2. *Montrer que (λ_1, λ_2, λ_3) forme une base de $\mathbb{P}_2$ (ensemble de polynômes de degré inférieur ou égal à 2).*
3. *Montrer que tout polynôme de $\mathbb{P}_2$ vérifiant (8.5) est unique.*
4. *Construire un polynôme d'interpolation de $\mathbb{P}_2$, vérifiant (8.5), exprimé dans la base (λ_1, λ_2, λ_3).*
5. *Montrer que l'erreur d'interpolation s'écrit :*
$$f(x) - p(x) = \frac{(x-a)^2(x-b)}{6} f^{(3)}(\xi(x)) \qquad a < \xi(x) < b. \tag{8.6}$$
(Indication : Définir pour $x \neq a$ et $x \neq b$ la fonction
$\phi(t) = f(t) - g(t) - A(x)(t-a)^2(t-b)$ et $A(x)$ est donné par $\phi(x) = 0$.)
6. *(a) Déterminer les réels α, β et γ dans la formulation suivante*
$$\int_{-1}^{1} g(t)\, dt = \alpha g(-1) + \beta g'(-1) + \gamma g(1) + R(g) \tag{8.7}$$
pour que la méthode soit exacte sur les polynômes de plus haut degré (le reste $R(g)$ est nul lorsque la méthode est exacte). Quel est le degré de précision de la méthode ?
   (b) Soit $g \in \mathcal{C}^3([-1,1];\mathbb{R})$, donner une majoration de l'erreur d'intégration en fonction des dérivées de g. (Indication : profiter de l'égalité (8.6)).
   (c) Soit $\{x_i, \ i = 0, ..., M\}$, une subdivision de $[0,5]$ de pas h. On cherche à approcher $J(u) = \int_0^5 u(x)\, dx$.
      (i) Construire une méthode d'intégration pour approcher $J_i(u) = \int_{x_i}^{x_i+h} u(x)\, dx$ en utilisant (8.7) et donner une majoration de l'erreur de l'intégration sur l'intervalle $[x_i, x_i + h]$.

(ii) Décrire la méthode composite en utilisant (8.7) *pour approcher* $J(u)$ *et majorer l'erreur d'intégration de la formule composite.*

CORRIGE

Corrigé exercice 1.

1. *On dérive par rapport aux variables* a_i. *Il vient*

$$\frac{\partial J}{\partial a_1}(a) = 2\sum_{i=1}^{n} x_i^2(a_1 + a_2x_i + a_3x_i^2 - y_i)$$

$$\frac{\partial J}{\partial a_2}(a) = 2\sum_{i=1}^{n} x_i^3(a_1 + a_2x_i + a_3x_i^2 - y_i)$$

$$\frac{\partial J}{\partial a_3}(a) = 2\sum_{i=1}^{n} x_i^4(a_1 + a_2x_i + a_3x_i^2 - y_i)$$

2. *La matrice est symétrique.*

$$\frac{\partial^2 J}{\partial^2 a_1}(a) = 2\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial^2 J}{\partial^2 a_1 a_2}(a) = 2\sum_{i=1}^{n} x_i^3$$

$$\frac{\partial^2 J}{\partial^2 a_1 a_3}(a) = 2\sum_{i=1}^{n} x_i^4$$

$$\frac{\partial^2 J}{\partial^2 a_2}(a) = 2\sum_{i=1}^{n} x_i^4$$

$$\frac{\partial^2 J}{\partial^2 a_2 a_3}(a) = 2\sum_{i=1}^{n} x_i^5$$

$$\frac{\partial^2 J}{\partial^2 a_3}(a) = 2\sum_{i=1}^{n} x_i^6.$$

3. *Le min est caractérisé par* $J'(a) = 0$, *soit encore*

$$\frac{\partial J}{\partial a_1}(a) = 0 \iff (\sum_{i=1}^{n} x_i^2)a_1 + (\sum_{i=1}^{n} x_i^3)a_2 + (\sum_{i=1}^{n} x_i^4)a_3 = \sum_{i=1}^{n} x_i^2 y_i$$

$$\frac{\partial J}{\partial a_2}(a) = 0 \iff (\sum_{i=1}^{n} x_i^3)a_1 + (\sum_{i=1}^{n} x_i^4)a_2 + (\sum_{i=1}^{n} x_i^5)a_3 = \sum_{i=1}^{n} x_i^3 y_i$$

$$\frac{\partial J}{\partial a_3}(a) = 0 \iff (\sum_{i=1}^{n} x_i^4)a_1 + (\sum_{i=1}^{n} x_i^5)a_2 + (\sum_{i=1}^{n} x_i^6)a_3 = \sum_{i=1}^{n} x_i^4 y_i$$

C'est un système linéaire :

$$Aa = b$$

avec

$$A = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 \\ \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 & \sum_{i=1}^{n} x_i^6 \end{pmatrix} \quad et\ b = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i^3 y_i \\ \sum_{i=1}^{n} x_i^4 y_i \end{pmatrix} \tag{8.8}$$

4. *On commence par calculer les coefficients de la matrice*

$$\sum_{i=1}^{n} x_i^2 = 2 + 1/2 = 5/2, \quad \sum_{i=1}^{n} x_i^3 = 0, \quad \sum_{i=1}^{n} x_i^4 = 2 + 1/8 = 17/8$$

$$\sum_{i=1}^{n} x_i^5 = 0, , \quad \sum_{i=1}^{n} x_i^6 = 2 + 1/32 = 65/32$$

$$\sum_{i=1}^{n} x_i^2 y_i = 9/8; \quad \sum_{i=1}^{n} x_i^3 y_i = -21/16; \quad \sum_{i=1}^{n} x_i^4 y_i = 45/32.$$

$$\begin{pmatrix} 5 & 0 & 17/4 \\ 0 & 17/4 & 0 \\ 17/4 & 0 & 65/16 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 9/4 \\ -21/8 \\ 45/16 \end{pmatrix}$$

Donc $a_1 = 5/4$, $a_2 = -21/34$, $a_3 = 2$. *Pour montrer que c'est un min, on va montrer que* $J"(a)$ *est définie positive.*

$$J"(a) = 2A = \begin{pmatrix} 5 & 0 & 17/4 \\ 0 & 17/4 & 0 \\ 17/4 & 0 & 65/16 \end{pmatrix}$$

La matrice $2A$ *est réelle et symétrique. Elle est définie positive ssi ses valeurs propres sont strictement positives. On a* $\lambda_3 = 17/4$ *et* $\lambda_1 + \lambda_2 = 5 + 65/16 > 0$ *et* $\lambda_1\lambda_2 = \frac{5\times 65}{16} - \frac{17\times 17}{16} = (325 - 289)/16$ *donc* $\lambda_1 > 0$ *et* $\lambda_2 > 0$.

Corrigé exercice 2.

1. *D'après le théorème de Gerschgorin, les valeurs propres sont dans l'union des disques de centre* a_{ii} *et de rayon* $\sum_{j\neq i} |a_{ij}|$. *Ici il y a un seul disque* $D(3, 2)$ *c'est à dire*

$$|\lambda - 3| \leq 2$$

Par ailleurs, A est symétrique et réelle alors ses valeurs propres sont réelles. On déduit alors

$$-2 \leq \lambda - 3 \leq 2 \iff 1 \leq \lambda \leq 5.$$

2. *On fait le produit $LU = A$, il vient*

$$A = LU = \begin{pmatrix} d_1 & d_1u_1 & 0 & \cdots & 0 & d_1b_1 \\ l_2 & l_2u_1 + d_2 & d_2u_2 & 0 & \cdots & l_2b_1 + d_2b_2 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & l_i & l_iu_{i-1} + d_i & d_iu_i & \ddots & l_ib_{i-1} + d_ib_i \\ 0 & \cdots & & & \cdots & \vdots \\ a_1 & a_1u_1 + a_2 & \cdots & a_{N-3}u_{N-3} + a_{N-2} & a_{N-2}u_{N-2} + l_N & c \end{pmatrix}$$

avec $c = \sum_{i=1}^{N-2} a_ib_i + l_Nu_{N-1} + d_N$. On déduit alors les relations suivantes :

$$\begin{cases} d_1 = 3,\ d_1u_1 = -1,\ d_1b_1 = 1 \\ l_2 = -1,\ l_2u_1 + d_2 = 3,\ d_2u_2 = -1,\ l_2b_1 + d_2b_2 = 0 \\ l_i = -1,\ l_iu_{i-1} + d_i = 3,\ d_iu_i = -1,\ l_ib_{i-1} + d_ib_i = 0,\ \textit{pour } i = 2, N-1 \\ a_1 = 1;\ a_{i-1}u_{i-1} + a_i = 0 \textit{ pour } i = 2, N-2;\ a_{N-2}u_{N-2} + l_N = -1;\ c = 3 \end{cases}$$

On va maintenant décrire l'algorithme, c'est à dire comment calculer les coefficients des matrices L et U :

$$d_1 = 3,\ u_1 = -\frac{1}{d_1},\ b_1 = \frac{1}{d_1}$$

$$l_2 = -1,\ d_2 = 3 - l_2u_1,\ u_2 = -\frac{1}{d_2},\ b_2 = -\frac{l_2b_1}{d_2}$$

• *Calcul des coefficients l_i, d_i, u_i, b_i*
pour i = 2 à N-1 faire
$l_i = -1$,
$d_i = 3 - l_iu_{i-1}$,
$u_i = -\frac{1}{d_i}$
$b_i = -\frac{l_ib_{i-1}}{d_i}$,
fini
• *Calcul des coefficients a_i*
$a_1 = 1$
Pour i = 2 à N-2 faire
$a_i = -a_{i-1}u_{i-1}$
fini
• *Calcul du coefficient l_N*
$l_N = -1 - a_{N-2}u_{N-2}$

• *Calcul du coefficient d_N. La relation $c = 3$ permet de calculer le coefficient manquant d_N comme*

$$d_N = 3 - \sum_{i=1}^{N-2} a_i b_i + l_N u_{N-1}.$$

3. *Si $x_k \longrightarrow x$, alors la limite x verifie :*

$$Mx = Nx + b \iff (M - N)x = b \iff Ax = b.$$

4. *On a*

$$x_{k+1} = M^{-1}N\, x_k + M^{-1}b.$$

La méthode converge ssi $\rho(M^{-1}N) < 1$. On a $M^{-1}N = M^{-1}(M-A) = I - M^{-1}A = I - \omega A$. Les valeurs propres de la matrice d'itération $I - \omega A$ sont $1 - \omega\lambda_i$. Ainsi,

$$\rho(M^{-1}N) = \max_{i=1,\cdots,n} |1 - \omega\lambda_i|.$$

Alors

$$\rho(M^{-1}N) < 1 \iff |1 - \omega\lambda_i| < 1 \iff -1 < 1 - \omega\lambda_i < 1 \iff 0 < \omega < \frac{1}{\lambda_i},\ \forall i = 1 \cdots n. \quad (8.9)$$

Si la méthode converge, alors on a

$$0 < \omega < \frac{1}{\lambda_1}$$

parce que la relation (8.9) est valable en particulier pour $i = 1$. D'autre part, comme $\lambda_1 \geq \lambda_i$, pour tout i , alors si $0 < \omega < \frac{1}{\lambda_1}$, on déduit

$$0 < \omega < \frac{1}{\lambda_1} \leq \frac{1}{\lambda_i},$$

ce qui montre que la relation (8.9) est vérifiée et la méthode converge.

5. *On a déjà vu que*

$$\rho(M^{-1}N) = \max_{i=1,\cdots,n} |1 - \omega\lambda_i|.$$

Les valeurs propres positives de A sont classées par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$$

donc

$$1 - \omega\lambda_1 \leq 1 - \omega\lambda_2 \leq \cdots \leq 1 - \omega\lambda_n.$$

ce qui donne le résultat.

6. *Le meilleur choix de ω est celui qui minimise la rayon spectral par rapport à ω. En effet,*

$$\rho(M^{-1}N) = \max(g_1(\omega), g_n(\omega)) = f(\omega)$$

Le meilleur choix de ω vérifie

$$f(\omega_\star) = \min_{\omega} f(\omega),$$

c'est à dire

$$f(\omega_\star) \leq f(\omega), \forall\omega.$$

On trace les deux courbes $g_1(\omega)$ et $g_n(\omega)$, ensuite on trace $f(\omega)$ et enfin on cherche (graphiquement) le min de f.

Corrigé exercice 3.

1. *Construction de λ_1. Le réel a est une racine double $\lambda_1(x) = c(x-a)^2$ et $\lambda_1(b) = c(b-a)^2 = 1$ alors $\lambda_1(x) = \frac{1}{(b-a)^2}(x-a)^2$.*
   Construction de λ_2. $\lambda_2(x) = c(x-a)(x-b)$ et $\lambda_2'(x) = c((x-b)+(x-a))$. Comme $\lambda_2'(a) = c(b-a) = 1$, d'où $\lambda_2(x) = \frac{1}{b-a}(x-a)(x-b)$.
   Construction de λ_3. $\lambda_3(x) = (\alpha x+\beta)(x-b)$. On a $\lambda_3(a) = (\alpha a+\beta)(a-b) = 1$ et $\lambda_3'(x) = (\alpha x+\beta)+\alpha(x-b) = 0$ donc $\lambda_3'(a) = (\alpha a+\beta)+\alpha(a-b) = 0$. on a alors $(\alpha a+\beta) = \frac{1}{a-b} = -\alpha(a-b)$ ce qui donne $\alpha = -\frac{1}{(a-b)^2}$ et $\beta = \frac{1}{a-b} - \alpha a = \frac{1}{a-b} + \frac{a}{(a-b)^2}$.
2. *On montre que cette famille est libre.*

$$\alpha\lambda_1(x) + \beta\lambda_2(x) + \gamma\lambda_3(x) = 0.$$

   pour $x = a$, on déduit $\gamma = 0$; pour $x = b$ on déduit $\alpha = 0$, pour $x = (a+b)/2$ on trouve $\beta = 0$ ou bien on dérivant $\alpha\lambda_2(x) = 0$ et on prend $x = a$.
3. *Soient $p, q \in \mathbb{P}_2$, alors $r = p - q \in \mathbb{P}_2$ et vérifie $r(a) = r(b) = r'(a) = 0$, ainsi $r(x) = c(x-a)^2(x-b)$; or $r \in \mathbb{P}_2$ alors $c = 0$ et $r(x) = 0$ pour tout x.*
4. *Pour $x = a$ ou $x = b$, l'égalité (8.6) est vérifiée. Pour $x \neq a$ et $x \neq b$, on définit la fonction*

$$\phi(t) = f(t) - p(t) - A(x)(t-a)^2(t-b)$$

   avec $A(x)$ est définit par $g(x) = 0$, soit encore

$$A(x) = \frac{f(x) - p(x)}{(x-a)^2(x-b)}.$$

   On a $\phi(a) = \phi(b) = \phi(x) = 0$, donc d'après le théorème de Rolle, il existe $\xi_1 \in]a, x[$ et $\xi_2 \in]x, b[$ tels que $\phi'(\xi_1) = \phi'(\xi_2) = 0$. Par ailleurs, comme $\phi'(a) = 0$ alors il existe $\xi_3 \in]a, \xi_1[$ et $\xi_4 \in]\xi_1, \xi_2[$ tels que $\phi"(\xi_3) = \phi"(\xi_4) = 0$, enfin il existe $\xi \in]\xi_3, \xi_4[$ tel que $\phi^{(3)}(\xi) = 0$.
   On a $\phi^{(3)}(t) = f^{(3)}(t) - p^{(3)}(t) - 6A(x)$. On a $p^{(3)}(t) = 0$ car $p \in \mathbb{P}_2$. On a montré alors que pour tout $x \in]a, b[$, il existe $a < \xi(x) < b$ vérifiant

$$\phi^{(3)}(\xi) = f^{(3)}(\xi) - 6A(x) = 0,$$

   soit encore

$$A(x) = \frac{f(x) - p(x)}{(x-a)^2(x-b)} = \frac{1}{6}f^{(3)}(\xi).$$

   D'où le résultat .
5. *pour $g(t) = 1$, $2 = \alpha + \gamma$*
   pour $g(t) = t$, $0 = -\alpha + \beta + \gamma$
   pour $g(t) = t^2$, $2/3 = \alpha - 2\beta + \gamma$
   on résout ce système linéaire : on obtient

$$\alpha = 4/3, \quad \beta = 2/3, \quad \gamma = 2/3.$$

Degré de précision : Pour $g(t) = t^3$, on a

$$\int_{-1}^{1} g(t)\ dt = 0$$

et

$$\alpha g(-1) + \beta g'(-1) + \gamma g(1) = -4/3 + 2 + 2/3 = -4/3$$

donc la méthode n'est pas exacte pour $\mathbb{P}_3$ et enfin le degré de précision est 2.

6. *Soit $p \in \mathbb{P}_2$ le polynôme d'interpolation vérifiant :*

$$p(-1) = g(-1), \quad p'(-1) = g'(-1), \quad p(1) = g(1)$$

d'après la question précédente avec $f = g$, $a = -1$ et $b = 1$, on déduit qu'il existe $\xi \in]-1, 1[$ tel que

$$g(t) - p(t) = \frac{(t+1)^2(t-1)}{6} g^{(3)}(\xi(t)).$$

On intégre en t, il vient

$$\int_{-1}^{1} g(t)dt - \int_{-1}^{1} p(t)dt = \int_{-1}^{1} \frac{(t+1)^2(t-1)}{6} g^{(3)}(\xi(t))dt.$$

On a $p \in \mathbb{P}_2$, ainsi la formule d'intégration est exacte

$$\int_{-1}^{1} p(t)dt = \alpha p(-1) + \beta p'(-1) + \gamma p(1) = \alpha g(-1) + \beta g'(-1) + \gamma g(1),$$

ce qui donne

$$R(g) = \int_{-1}^{1} \frac{(t+1)^2(t-1)}{6} g^{(3)}(\xi(t))dt.$$

On donne une majoration de l'erreur :

$$|R(g)| \leq \Big(\int_{-1}^{1} \frac{(t+1)^2(1-t)}{6}\, dt \Big) \max_{\xi \in [-1,1]|} g^{(3)}(\xi)| \leq \frac{8}{3} \max_{\xi \in [-1,1]|} g^{(3)}(\xi)|.$$

7. *On fait un changement de variable en posant $x = x_i + \frac{1+t}{2}h$, ainsi*

$$\int_{x_i}^{x_{i+1}} u(x)dx = \frac{h}{2} \int_{-1}^{1} u(x_i + \frac{1+t}{2}h)dt,$$

on applique la formule avec $g(t) = u(x_i + \frac{1+t}{2}h)$, et alors $u'(t) = \frac{h}{2}u'(x_i + \frac{1+t}{2}h$:

$$\int_{x_i}^{x_{i+1}} u(x)dx \approx \frac{h}{2}\Big(\alpha u(x_i) + \beta \frac{h}{2} u'(x_i) + \gamma u(x_{i+1})\Big),$$

Erreur d'intégration sur $[x_i, x_{i+1}]$: $R_i(u) = \frac{h}{2}R(g)$. on a

$$|R_i(u)| = \frac{h}{2}|R(g)| \leq \frac{4h}{3} \max_{-1 \leq t \leq 1} |g^{(3)}(t)| = \frac{4h}{3}\frac{h^3}{2^3} \max_{x_i \leq x \leq x_{i+1}} |u^{(3)}(x)|.$$

.

8. *Formule Composite.* $J(u) = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} u(x)dx$ *et sur chaque intervalle de longueur* h*, on applique la formule d'intégration précédente*

Erreur d'intégration sur $[0, 5]$.

$$R_h(u) = \sum_{i=0}^{M-1} \int_{x_i}^{x_{i+1}} u(x)\,dx - \frac{h}{2} \sum_{i=0}^{M-1} \Big(\alpha u(x_i) + \beta \frac{h}{2} u'(x_i) + \gamma u(x_{i+1})\Big) = \sum_{i=0}^{M-1} R_i(u)$$

par conséquent,

$$|R_h(u)| \leq \sum_{i=0}^{M-1} |R_i(u)| \leq \frac{h^4}{6} M \max_{0\leq x\leq 5} |u^{(3)}(x)| = \Big(\frac{5}{6} \max_{0\leq x\leq 5} |u^{(3)}(x)|\Big) h^3.$$

CHAPITRE 9

DEVOIR SURVEILLÉ D'ANALYSE NUMÉRIQUE (2011)

Exercice 1.

On munit l'espace vectoriel $\mathcal{C}^1([-1,1];\mathbb{R})$ du produit scalaire suivant

$$< f, g >= \int_{-1}^{1} f(x)g(x)\,dx + \int_{-1}^{1} f'(x)g'(x)\,dx$$

On note par $\|f\| = \sqrt{< f, f >}$.
Trouver le polynôme de degré ≤ 2 qui réalise la meilleure approximation au sens des moindres carrés de $f(x) = x^4$ pour la norme définie ci-dessus.

Exercice 2.

Soit $M_n(\mathbb{R})$ l'ensemble des matrices à n lignes et n colonnes et à éléments réels. La norme matricielle considérée est celle associée à la norme vectorielle euclidienne sur $\mathbb{R}^n$ (notée $|| \cdot ||_2$ dans le cours).

Soient A, A_1 et A_2 trois matrices carrées et réelles vérifiant :
i) $A = A_1 + A_2$,
ii) A_1 est symétrique définie positive,
iii) A_2 est symétrique semi-définie positive.

Soit r un réel strictement positif ($r > 0$).

1. Montrer que les matrices $A_1 + r\,I$, $A_2 + r\,I$ et A sont inversibles (où I désigne la matrice identité).
2. Montrer que :
$$(A_1 + r\,I)^{-1}\,(r\,I - A_1) \;=\; (r\,I - A_1)\,(A_1 + r\,I)^{-1}\,. \tag{9.1}$$
3. Exprimer les valeurs propres de la matrice produit ci-dessus en fonction de celles de A_1.
4. Soit le système linéaire suivant
$$A\,x = b\,, \tag{9.2}$$

où b est donné dans $\mathbb{R}^n$.
Montrer que l'équation (9.2) équivaut à :

$$(A_1 + r\,I)\,x \;=\; (r\,I - A_2)\,x + b\;, \tag{9.3}$$

et aussi à :

$$(A_2 + r\,I)\,x \;=\; (r\,I - A_1)\,x + b\;. \tag{9.4}$$

5. Pour résoudre numériquement (9.2), on considère les deux suites (x_m) et (y_m) de $\mathbb{R}^n$ définies par la donnée de x_0 et par les deux relations de récurrence suivantes :

$$\begin{cases} (A_1 + r\,I)\,y_m \;=\; (r\,I - A_2)\,x_m + b \\ (A_2 + r\,I)\,x_{m+1} \;=\; (r\,I - A_1)\,y_m + b\;. \end{cases} \tag{9.5}$$

Montrer que x_{m+1}, donné par (9.5), s'écrit sous la forme :

$$x_{m+1} = B\,x_m + c\;,$$

où $B \in M_n(\mathbb{R})$ et $c \in \mathbb{R}^n$.

6. En déduire que

$$x_m - \,x \;=\; B^m\,(x_0 - x)\;.$$

7. On pose

$$U = (A_2 + r\,I)\;B\;(A_2 + r\,I)^{-1}.$$

Montrer que

(a)

$$U = \Phi_1\;\Phi_2 \;\text{ où } \Phi_i = (r\,I - A_i)\,(A_i + r\,I)^{-1}\;, \quad i = 1, 2,$$

(b) $\rho(\Phi_1) < 1$ et $\rho(\Phi_2) \leq 1$ (où $\rho(M)$ d ésigne le rayon spectral d'une matrice M),

(c)

$$\rho(B) = \rho(U) \leq ||\,U\,||_2 \leq \rho(\Phi_1)\;\rho(\Phi_2)\;.$$

8. Que peut-on déduire sur les suites (x_m) et (y_m) lorsque m tend vers $+\infty$?

Exercice 3.

Soit $f : \mathbb{R} \mapsto \mathbb{R}$ telle que $f \in \mathcal{C}^5([a, b]; \mathbb{R})$ et soit $a < b$ deux réels donnés.

1. Déterminer la formule d'intégration suivante

$$\int_a^{a+h} f(x)\,dx \approx \alpha f(a) + \beta f(a + h) \tag{9.6}$$

pour qu'elle soit exacte pour des polynômes de degré le plus haut possible. Préciser le degré de précision.

2. On note
$$I = \int_a^{a+h} f(x)\,dx,\ I(h) = \alpha f(a) + \beta f(a+h) \text{ et } R(h) = I - I(h).$$
On rappelle la formule de Taylor à l'ordre p, Il existe $0 < \theta < 1$ tel que
$$R(h) = R(0) + hR'(0) + \frac{h^2}{2!}R''(0) + \cdots + \frac{h^{p-1}}{(p-1)!}R^{(p-1)}(0) + \frac{h^p}{p!}R^{(p)}(\theta h).$$
Calculer $R'(h)$, $R"(h)$, $R^{(3)}(h)$ et $R^{(4)}(h)$. En déduire que
$$R(h) = -\frac{h^3}{12}f"(a) - \frac{h^4}{24}f^{(3)}(a+\theta h) - \frac{h^5}{48}f^{(4)}(a+\theta h) \tag{9.7}$$
où $0 < \theta < 1$.
3. On utilisant la formule d'intégration (9.6), donner une approximation de
$$I_1 = \int_a^{a+h/2} f(y)\,dy \text{ et } I_2 = \int_{a+h/2}^{a+h} f(y)\,dy.$$
En écrivant que $I = I_1 + I_2$, donner une approximation de I notée $J(h)$. On définit
$$r(h) = I - J(h)$$
Montrer que
$$r(h) = -\frac{h^3}{48}f"(a) - \frac{h^4}{96}f^{(3)}(a+\tau h/2) - \frac{h^5}{96}d_\tau, \tag{9.8}$$
où $0 < \tau < 1$ et $d_\tau = \left(\frac{1}{8}f^{(4)}(a+\tau\frac{h}{2}) + f^{(4)}(a+\tau h)\right)$. (On peut par exemple faire un développement de Taylor à l'ordre 4).
4. Montrer que
$$\frac{4J(h) - I(h)}{3} - I = c_1 h^4 + c_2 h^5$$
où c_1 et c_2 sont des coefficients à déterminer.
5. Donner une approximation d'ordre 4 de I.
6. On note $I_4(h)$ une approximation de I à l'ordre 4. Construire une méthode d'intégration d'ordre 5.

CHAPITRE 10

TRAVAUX SUR ORDINATEUR INITIATION À MATLAB

Matlab est un logiciel de calcul numérique produit par MathWorks (voir le site web http :// www.mathworks.com/). Il est disponible sur plusieurs plateformes. Matlab est un langage simple et très efficace, optimisé pour le traitement des matrices et pour le calcul numérique. Matlab est l'abréviation en anglais de **"matrix laboratory"**, tous les objets en jeu sont des matrices, y compris les scalaires (matrices 1×1).
On propose une initiation non exhaustive des commandes utiles en calcul scientifique.
Première utilisation de Matlab , l'aide en ligne Une fois Matlab lancé, les instructions de Matlab doivent être tapé dans la **fenêtre de commandes** le sigle >> signifie une ligne de commande, Par exemple :

```
>>1+1
ans =
2
>>t=1+1
t =
2
>>t=1+1;
>>t
t=
2
>>u=sin(t)
u =
0.9093
>>v=exp(u)
v =
2.4826
>>format long
>>v
v =
2.48257772801500
```

```
>>format short
>>v
v =
2.4826
>>who
Your variables are:
ans t u v
>>whos
Name Size Bytes Class
ans 1x1 8 double array
t 1x1 8 double array
u 1x1 8 double array
v 1x1 8 double array
>> clear
>> v
```

On remarque que :

1. par défaut, tout calcul est affecté dans la variable **ans**
2. la commande **format** permet de modifier l'affichage du format des différentes variables
3. les commandes **who, whos** permettent de lister l'ensemble des variables utilisées
4. la commande **clear** efface le contenu de tous les variables utilisées.

10.1. La commande ;

```
>> y=sin(0.15*pi);
```

le calcul de y a été effectué mais il n'est pas affiché à l'écran. Pour l'afficher, il suffit de taper y

```
>> y
y= 0.4540
```

10.2. Variables spéciales

pi, *i*, *realmin*, *realmax*, *eps*, *ans* ... variables prédéfinies.

```
>> eps
ans =
2.2204e-16
>> 1. + eps
1.0000
>> realmax
ans =
1.7977e+308
>> realmin
ans =
```

```
2.2251e-308
>> v=1.e+400
v =
   Inf
>> pi
```

Le zéro machine *eps* est le plus grand réel positif tel que $1 + eps <= 1$!!!!!.

```
>>y = 0.444444444444444e-30
>>x = 0.222222222222222e-20
>> x+y
     0.222222222666667e-20
>>u = 0.444444444444444e+30
>>v = 0.222222222222222e+20
>> u+v
     0.444444444666666e+30
```

Le calcul se fait en réduisant à la puissance la plus élevée. Dans le premier cas, le se fait de la manière suivante

$$x + y = 10^{-20}(0.222222222222222 + 0.000000000444444)$$

et dans le second

$$u + v = 10^{30}(0.444444444444444 + 0.000000000222222).$$

10.3. Nombres complexes

```
>> c1 = 1-2i
c1 =
1.0000 - 2.0000i
>> c2 = 3*(2-sqrt(-1)*3)
c2 =
6.0000 - 9.0000i
>>c3= conj(c2)
>> real(c1)
>> imag(c2)
>> abs(c2)
>> angle(c1)
ans =
-1.1071
```

10.4. Affichage

FORMAT Set output format. All computations in MATLAB are done in double precision. FORMAT may be used to switch between different output display formats as follows : FORMAT SHORT (default) Scaled fixed point format with 5 digits. FORMAT LONG Scaled fixed point format with 15 digits.

```
>> pi
ans =
3.1416
>> format long
>> pi
ans =
3.141592653589
```

10.5. Les commentaires

Il est important de pouvoir écrire des commentaires lors de l'élaboration d'un programme. Pour cela, sur une ligne tout ce qui après le symbole % est non lu.
>> pi % pour connaitre la valeur de pi

10.6. Vecteurs - Matrices

Les tableaux (vecteurs, matrices, ...) s'utilisent dans la fenêtre de commande matlab ou dans des programmes.
Les indices des vecteurs et matrices commencent toujours à 1. L'indice 0 n'existe pas en matlab.
La ième composante d'un vecteur x est appelée $x(i)$; le coefficient de la ième ligne, jème colonne d'une matrice A est $A(i, j)$. Il n'y a pas à déclarer les tableaux, mais il est cependant recommandé de réserver une place mémoire pour les matrices en les initialisant à 0. (voir plus loin).

10.6.1. Création de vecteurs. — Un vecteur ligne est défini

(i) soit par une relation de la forme

$$x = [x1\ x2 \quad xn]$$

ou

$$x = [x1, x2, \quad , xn]$$

où les éléments x_i sont séparés par des espaces ou des virgules et mis entre [ ]. **Exemples :**

```
>> x1 = [1.1 2.3  3.5 -4. -8.]   % (vecteur de 5 composantes)
x1 =
    1.1000    2.3000    3.5000   -4.0000   -8.0000
>> x2 = [3  -5.2   2*sqrt(3)]   % (vecteurs de 3 composantes).
```

```
x2 =
    3.0000   -5.2000    3.4641
>>x2(3)
```

(ii) soit par une subdivision d'un intervalle donné (a, b) en sous-intervalles de longueur donnée h,

x=a :h :b aller de a à b par pas de h,
x=a :b (par défaut h = 1).

```
>> x = 5 :-1 : 1
x = 5 4 3 2 1
>> x=1: 1.1 : 5
x =    1.0000    2.1000    3.2000    4.3000
>> x=5:1
x =
   Empty matrix: 1-by-0
>> x = 0 : pi/2 : 2 * pi
x =
         0    1.5708    3.1416    4.7124    6.2832
```

(ii) soit par l'instruction

x=linspace(a, b, n) définit un vecteur de n composantes
$x(i) = a + (i-1)\frac{b-a}{n-1}$, pour $i = 1 \cdots n$.
Par défaut : $n = 100$.

```
>>y= linspace(0,pi,9)
y =
 0    0.3927    0.7854    1.1781    1.5708    1.9635    2.3562    2.7489    3.1416
```

(iv) soit par l'utilisation d'une fonction. Par exemple, si x est un vecteur ligne, alors sin(x) et cos(x) (par exemple) sont des vecteurs ligne de même nombre de composantes.

10.6.2. Vecteur transposé. — Un vecteur colonne peut être défini comme le transposé d'un vecteur ligne, x' **désigne le vecteur transposé du vecteur** x**.**

```
>> x = (1 : 4)' ; y = [3 -4.5 2.1]'  %  deux vecteurs colonnes x et y de composantes
```

ii) ou directement, en utilisant des ; au lieu de ,.

$$x = [x1; x2; \quad ; xn]$$

```
>> z = [3.1452 ; -3 ; 4. ; 5.256] % définit un vecteur colonne z de composantes
```

iii) en utilisant une fonction. Si x est une vecteur colonne, $sin(x)$ et $cos(x)$, par exemple, sont des vecteurs colonne de même longueur que x.

```
>> x = (0 : 0.2 : 1); % x est un vecteur colonne
>> y = exp(x); % y est un vecteur colonne
```

10.6.3. Opérations sur les vecteurs. — Quelques opérations sur les vecteurs :

$+$	:	addition de deux vecteurs u et v de même longueur.
$dot(u, v)$ ou $u' * v$	:	produit scalaire de deux vecteurs u et v.
$.*$	:	multiplie deux vecteurs composantes par composantes.
$./$	:	divise deux à deux les composantes de deux vecteurs .
$.\hat{}$	:	élève les composantes d'un vecteur à la puissance des composantes du second.
$sum(u)$	:	somme des composantes d'un vecteur u.
$mean(u)$	:	moyenne des composantes d'un vecteur u.
$length(u)$	:	donne la longueur d'un vecteur u.
$min(u)$	:	donne la plus petite composante d'un vecteur u.
$max(u)$	:	donne la plus grande composante d'un vecteur u.

```
>> u = (1 : 4) , v = (2 : 5)
u =
     1     2     3     4
v =
     2     3     4     5
>> z = u .* v        % z(i) = u(i)*v(i)
z =
     2     6    12    20
>> w = v./u          % w(i) = v(i) = u(i)
     w =
    2.0000    1.5000    1.3333    1.2500
>> t = v .^u        % t(i) = v(i) ^ u(i)
    t =
     2     9    64   625
>> x = (0 : 0.2 : 1)
     x =
         0    0.2000    0.4000    0.6000    0.8000    1.0000
>> z = exp(x).* cos(x)
         z =
    1.0000    1.1971    1.3741    1.5039    1.5505    1.4687
%z est un vecteur dont les composantes sont les éléments
%z(i) = exp(x(i))  cos(x(i)) pour i = 1;    ; 6 ,
 >> disp('mean(z) = '), mean(z)
 >>disp('length(z) = '), length(z)
 >>disp('min(z) = '), min(z)
 >>disp('max(z) = '), max(z)
 >>disp('sum(z) = '), sum(z)
 >>disp('max(abs(z)) = '), max(abs(z))
```

disp('phrase a afficher') : affiche à l'écran la phrase "phrase a afficher"

10.7. Création de matrices

i) Une matrice de n lignes et p colonnes peut être définie par une relation de la forme :

$$A = [a_{11}\ a_{12}\ \ a_{1p}; a_{21}\ a_{22}\ \ a_{2p}; \cdots a_{n1}\ a_{n2}\ \ a_{np}]$$

```
>> A = [1 2 3 4 ; 5 6 7 8 ; 9 10 11 12]   % A est une matrice de 3 lignes et 4 colonnes
A =
1 2 3 4
5 6 7 8
9 10 11 12
```

A' désigne la transposée de A si A est réelle, l'adjointe de A si A est complexe.
$A(i, j)$ désigne l'élément de la ligne i et colonne j .

ii) Une matrice peut aussi être construite à partir de vecteurs ou de matrices plus petites **concaténation** :
Exemple

```
>> A = [1 2  3; 4 5  6]
A =
1 2 3
4 5 6
>> B = [A; 7  8  9]
B =
1 2 3
4 5 6
7 8 9
```

10.7.1. Création de matrices $n \times p$. — Initialisation de matrices

A=zeros(n,p)	:	initialisation à 0 d'une matrice n lignes et p colonnes
A=eye(n)	:	matrice identité d'ordre n
A=ones(n, p)	:	matrice n lignes, p colonnes, composée de 1
A= rand(n, p)	:	matrice de nombres aléatoires $\in]0; 1[$.

Exemple.

```
>> x = zeros(1, 5)  % définit le vecteur nul de 5 composantes
x = 0 0 0 0 0
>> x = zeros(5 ,  1) %définit un vecteur colonne de 5 composantes egales a 0.
>> x = ones(1, 5)  %définit le vecteur ligne de 5 composantes
x = 1 1 1 1 1
```

10.8. Opérations sur les matrices

Quelques opérations possibles

+ *	:	addition, multiplication de deux matrices compatibles.
x = A(1, :)	:	première ligne de A.
y = A(:, 2 : 3)	:	deuxième et troisième colonnes de A.
w = A(3, i : j)	:	du i-ieme au j-ieme elements de la ligne 3.
u = A(:,2)	:	deuxième colonne de A.
z = A(:)	:	mise sous forme d'une colonne.
c *A	:	multiplie tous les éléments de A par le scalaire c.
A.m	:	élévation à la puissance m de chaque élément de la matrice A.
A^m	:	élévation à la puissance m de la matrice A.
size(A)	:	donne les dimensions de la matrice A ([m,n]=size(A))
eig(A)	:	vecteur donnant les valeurs propres de la matrice A
det(A)	:	donne le déterminant de la matrice A.
rank(A)	:	rang de la matrice A
trace(A)	:	trace de la matrice A
spy(A)	:	représentation graphique de la matrice A (dans le cas de matrices de grande taille).

Exemple :

```
>>A= rand(6,6)
>>DetA=det(A)
>>ValProp = eig(A)
>>A(1,:)=1
>>A(4,3)=-3.3333
>>rangA= rank(A)
>>B = rand(20,30);
>>B(3:6,10:18)=0.;
>>spy(B)
>>[n,m]=size(B)
>>NbLigneB = size(B,1)
>>NbColonneB= size(B,2)
```

10.9. M-Files ou scripts

Un script (ou M-file) est un fichier (premiertp.m par exemple) contenant des instructions Matlab.

Tous les fichiers matlab doivent se terminer par le suffix .m

Voici un exemple de script ou de programme matlab :

```
% mon premier programme matlab
%premiertp.m calcule det, valeurs propres d'une matrice
%
clear all
%%% Les matrices
n=input('Donner la dimension de la matrice n = ')
```

```
A= rand(n,n)
DetA=det(A)
ValProp = eig(A)
A(1,:)=1
A(4,3)=-3.3333
rangA= rank(A)

nbl=input('Donner le nombre des lignes (> 10)de la matrice B   = ')
nbc=input('Donner le nombre des colonnes(>20) de la matrice   = ')

B = rand(nbl,nbc);
B(4:8,10:18)=0.;
spy(B)

[n,m]=size(B)
NbLigneB = size(B,1)
NbColonneB= size(B,2)
```

Matlab vous offre un éditeur pour écrire et mettre au point vos M-files :

Pour **exécuter** le programme il suffit de cliquer sur **run** dans le menu de l'éditeur,
ou
dans la fenêtre de commande, on exécute un M-file en utilisant le nom du script comme commande :
>> premiertp

Les M-files (ou programme) sont **exécutés séquentiellement** dans le "workspace", c'est à dire qu'ils peuvent accéder aux variables qui s'y trouvent déjà, les modifier, en créer d'autres etc.

10.10. Fonctions

Une fonction est un script admettant des variables d'entrées et des variables de sorties, par exemple $f(x, y) = x^2 + y^2$ admet x et y comme variables d'entrées et le résultat $z = x^2 + y^2$ comme argument de sortie. Voici une fonction "fonc" définie dans un fichier *fonc.m*

```
function [z,t] = fonc(x,y,m)
z = x^2+y^2;
t=x-y+m;
end
```

Les variables x, y et m sont les variables d'entrées et ils ne doivent pas être modifiées à l'intérieur de la fonction. Les variables z et t sont les arguments de sortie, la fonction doit affecter une valeur pour z et t.

Utilisation de cette fonction. Une fonction est appelé depuis un script (un programme principal) ou dans une fenêtre commande.

>> [z,t] = fonc(1., 0., -4.)
>>[y1,y2]= fonc(-1., sin(1.), sqrt(2.))

10.11. HELP

Toutes les fonctionnalités de matlab sont illustrées à partir du menu **help**.
>>help function
>> help for
>> help switch
>> help if

10.12. Boucles et contrôle

10.12.1. Opérateurs logiques de comparaison. —

<	plus petit
>	plus grand
<=	plus petit ou égal
>=	plus grand ou égal
==	égal (compare si deux nombres sont égaux ou non)
~=	différent ou pas égal
/&	et logique
|	ou logique
~	not

10.12.2. It then else. — Faire help if.
Exemple.

```
a= log(2.)
b= sqrt(2.)/2.
if (a > b)
disp(' a est plus grand que b')
else
disp('a est plus petit que b')
end
```

10.12.3. For. — Faire help for.
Exemple.

```
A = zeros(4,5)
for i = 1: 4
     for j=5:-1:1
           A(i,j) = i+j^2 ;
     end
```

```
 end
 A
```

10.12.4. switch. — Selon la valeur d'un paramètre faire.
help switch
Exemple.

```
switch m % faire selon la valeur de m
     case 0  % si m=0
             x=0
     case {1, 2, 9} % si m=1 ou m=2 ou m=9
             x= m^2
     case {-3,-1, 99} % si m=-3 ou  m=-1 ou m= 99
             x= m+ m^2
     otherwise % sinon
             x=-9999999
end
```

10.13. Graphismes

>> help 2-D Plots
mettre dans un script *courbes.m* les instructions suivantes

```
% courbes.m
% Exemples de graphismes en 2D
% graphe 1
figure
x=0:0.05:5;
y=sin(x.^2);
plot(x,y);
xlabel('Time')
ylabel('Amplitude')
title('ma première courbe')
legend('toto')

%graphe 2
figure
z = -2.9:0.2:2.9;
bar(z,exp(-z.*z));

%graphe 3
figure
subplot(2,1, 1)
plot(x,y);
```

```
subplot(2,1,2)
bar(z,exp(-z.*z));

% graphe 4
figure
Y=[0:0.05: 1];Z1=sin(2*pi*Y);Z2=cos(2*pi*Y);
plot(Y,Z1,':b',Y,Z2,'+k');
title('Exemple de courbes');
xlabel('Y');ylabel('Z');
legend('sin','cos');
```

Commenter chaque instruction de ce programme.

Exercice 10.1. — *Définir la fonction gaussienne.m,*

```
function [g] = gaussienne(x,xc,s)
g=exp( - (x-xc)^2/s^2)
end
```

tracer la fonction pour différente valeur de xc et s sur la même courbe (faire un script tracer-gauss.m).
Dans la fonction gaussienne l'argument x est un scalaire, écrire une fonction "gaussiennev(x,xc,s)" avec x un vecteur. Tracer la fonction pour différente valeur de xc et s sur la même courbe.

10.14. tic toc

Calcul le temps CPU d'exécution d'un programme. Il suffit de faire l'instruction *tic* au début du script et ensuite de faire *toc* en fin du script pour avoir le temps d'exécution du programme.

10.15. Fonctions mathématiques

sin cos tan sinh cosh tanh ...
asin acos atan asinh acosh atanh ...
*exp log log*10 *sqrt*

fix(x)	:	donne le plus petit entier inférieur ou égal au réel x
floor(x)	:	donne la partie entière de x
ceil(x)	:	le plus proche entier plus grand ou égal à x
round(x)	:	
mod(x)	:	le reste de la division
sign	:	

autres fonctions : *factor isprime primes gcd*(*pgcd*) *lcm*(*ppcm*)

CHAPITRE 11

TRAVAUX SUR ORDINATEUR EQUATION DE LA CHALEUR EN 1D

11.1. Equation de la chaleur

On s'intéresse à l'approximation de l'équation de diffusion suivante :

$$\begin{cases} -u''(x) = f(x), \quad \text{pour } 0 < x < 1 \\ u(0) = u_g,\ u(1) = u_d \end{cases}$$

Soit $x_0, x_1, x_2, \ldots, x_{Ns}, x_{Ns+1}$, une subdivision régulière de l'intervalle $[0, 1]$ de pas h.
On note u_i une approximation de $u(x_i)$. La solution $u =^t (u_1, u_2, ..., u_{Ns})$ du système précédent est obtenue par la méthode des différences finies suivante :

$$\begin{cases} \dfrac{1}{h}\left(-\dfrac{u_{i-1} - u_i}{h} - \dfrac{u_{i+1} - u_i}{h}\right) = f_i, \quad \text{pour } i = 1, \ldots, Ns \\ u_0 = u_g,\ u_{Ns+1} = u_d \end{cases} .$$

1. Écrire ce système sous la forme $AU = b$, où A est une matrice symétrique.
2. Définir sous Matlab :
   (a) la fonction `f(x,m)` qui admet le réel x et l'entier m comme arguments et est définie par :

$$f(x, m) = \begin{cases} 0. & \text{si } m = 0 \\ -2. & \text{si } m = 1 \\ x^2 & \text{si } m = 2 \\ 4.\pi^2\, sin(2.\pi\, x) & \text{si } m = 3 \\ 10.\, e^{-x} & \text{si } m = 4 \\ & \ldots \end{cases} ;$$

```
function y = f(x,m)
% Terme source
switch m
    case 0
        y = zeros(size(x)) ;
    case 1
        y = ....
```

```
    case 2
        y = .....
    case 3
        y = .....
    case 4
        y = .....
    otherwise
        error('Argument m indéfini dans f');
end
end
```

(b) la fonction `solex(x,`u_g`,`u_d`,m)` qui est la solution exacte de l'équation associée à la fonction `f(x,m)` et aux valeurs u_g et u_g.

```
function y = solex(x, ug, ud, m)
% Solutions exactes
switch m
    case 0


 end
 end
```

Dans un script :

(c) Créer le vecteur X des abscisses des sommets du maillage ;

(d) Créer la matrice A et le vecteur b, et résoudre $AU = b$ (en utilisant la résolution de matlab).

(e) Comparer la solution exacte et la solution calculée en fonction de Ns et m. Tracer la solution exacte et la solution approchée.

```
 % Ce programme résoud l'équation de la chaleur en dimension 1
% Paramètres d'entrée
% m          : cas d'étude pour le terme de forçage f(x)
% Ns        : nombre de points de maillage (nombre de sommets)
% ug et ud : Conditions aux limites en x=0 et x=1
clear all
clf
disp('========================')
disp('==== Equation de la diffusion =======')
disp('========================')
disp(' Le second membre, choix de la fonction f ')
disp(' pour m=0 ; f=0 et u = ')
disp(' pour m=1 ; f=-2x')
disp(' pour m=2 ; f=x^2')
disp(' pour m=3 ; f=4*pi')
```

```
disp(' pour m=4 ; f=exp(x)')
% Données du problème m, Ns, ug,ud

% creation du vecteur

% Système linéaire :  Remplir la  matrice A

% Second membre b

% Résolution du système linéaire

tic;
U    = A \ b;
time_matlab = toc;

% Calcul de la solution exacte aux points du maillage
Uex = solex(X, ug, ud, m);

% Erreur relative par inversion directe
err  = abs((Uex-U));
ErreurMax= max(err)/max(abs(Uex))
Erreur2= norm(err)/norm(Uex)


% Graphismes

% Méthode de Jacobi

% Méthode de Gauss-Seidel

% Méthode LU

% Comparaison
```

(f) Résoudre le système $AU = b$ par la méthode de Jacobi en tirant profit du fait que A est matrice tridiagonale. Comparer en temps CPU la résolution de matlab et la résolution proposée.

```
function [x, k, erreur, converge] = jacobi_tri(A,b,itermax, tolerence)
%-------------------------------------------------------------------------
% résoud le système linéaire Ax=b par la méthode de Jacobi pour une matrice
% tridiagonale
% Les seuls coefficients non nuls de A sont A(i,i-1), A(i,i) et A(i,i+1)
```

```
% A= D-E-F
% D Xk1 = (E+F)Xk +b
% itermax = nombre d'iterations maximales
% tolerence = l'erreur entre deux itérés successives pour la convergence
%
% k = nombre d'itération que l'algorithme effectue si l'algoritme converge
% si la methode ne converge pas alors k = itermax
% erreur : l'erreur entre xk1 et xk ;
% si la methode converge erreur <= tolerence
%
% converge = 1 si la methode converge sinon = 0
% ------------------------------------------------------------------------
[n,n] = size(A);
xk1 = zeros(n,1);
xk = zeros(n,1);
x= zeros(n,1) ;
converge = 0 ;

for k=1:itermax
    % resolution d'une iteration de Jacobi
    xk1(1) = (b(1)- A(1,2)*xk(2) ) /A(1,1) ;
    for i=2: n-1
        xk1(i) =     A COMPLETER       ;
    end
    xk1(n) =  A COMPLETER   ;
    %
    erreur = norm (xk1-xk);
    if (erreur <= tolerence)
       x=xk1;
       converge =1
       break %%% on sort de la boucle
    else
       xk=xk1;
    end
end
end
```

(g) Résoudre le système $AU = b$ par la méthode de Gauss-Seidel en tirant profit du fait que A est matrice tridiagonale. Comparer en temps CPU la résolution de matlab et la résolution proposée.

(h) Stocker A sous forme d'une matrice tridiagonale, faire la décomposition LU et résoudre le système linéaire. Comparer en temps CPU la résolution de matlab et la résolution proposée.

11.2. Flambage d'une barre (facultatif)

Soit une barre verticale, représentée par un segment [0,1], fixée aux extrémités en 0 et 1, soumise à une force P. Quand P est faible, la barre n'est pas déformée. Lorsqu'on augmente P, on atteint une valeur critique à partir de laquelle la barre se déforme. Le problème est de déterminer cette force critique.

La déformation horizontale $u(x)$ de la barre pour $x \in [0,1]$ est solution du problème suivant :

$$\begin{cases} -(c(x)u'(x))' = Pu(x), & x \in (0,1) \\ u(0) = 0,\ u(1) = 0 \end{cases} \tag{11.9}$$

où $c(x)$ est une fonction positive qui dépend des caractéristiques de la barre (section, élasticité).

Le problème (11.9) est un problème de valeur propre. On cherche la plus petite valeur de P, c'est à dire la plus petite valeur propre de l'opérateur $-(c(x)u'(x))'$. Dans le cas simple où la fonction c est une constante, par exemple pour $c = 1$ les solutions sont les couples (u, P) avec

$$P = (k\pi)^2, \qquad u(x) = sin(k\pi x), k \geq 1.$$

Dans le cas où la fonction c dépend de x, on cherche une solution approchée par la méthode des différences finies en considérant l'approximation suivante (mêmes notations x_i et u_i que dans la partie **2**) :

$$\begin{cases} \dfrac{1}{h}\left(-c_{i-\frac{1}{2}}\dfrac{u_{i-1}-u_i}{h} - c_{i+\frac{1}{2}}\dfrac{u_{i+1}-u_i}{h}\right) = Pu_i, \quad \text{pour } i = 1,\ldots,Ns \\ u_0 = 0,\ u_{Ns+1} = 0 \end{cases} . \tag{11.10}$$

avec $c_{i+\frac{1}{2}} = c(\frac{x_i+x_{i+1}}{2})$.

1. Vérifier que le système (11.10) vérifie $AU = PU$ avec $U = (u_1, u_2, \cdots, u_{Ns})$ et

$$A = \frac{1}{h^2}\begin{pmatrix} c_{\frac{1}{2}} + c_{\frac{3}{2}} & -c_{\frac{3}{2}} & & \\ -c_{\frac{3}{2}} & c_{\frac{3}{2}} + c_{\frac{5}{2}} & -c_{\frac{5}{2}} & \\ & \ldots & & \\ & & -c_{Ns-\frac{1}{2}} & c_{Ns+\frac{1}{2}} + c_{Ns-\frac{1}{2}} \end{pmatrix}.$$

2. En utilisant la fonction $eig(A)$ et dans le cas $c = 1$, calculer les valeurs propres de A et comparer avec la solution exacte en fonction de Ns.
3. Etudier le cas où $c(x) = (x - \frac{1}{2})^2$.

BIBLIOGRAPHIE

[1] P. Lascaux, R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Tomes 1 et 2 Masson 1986.

[2] G. Allaire, S.M. Kaber, *Algèbre linéaire numérique* Ellipse, mathématiques 2e cycle édition, 2002.

[3] M. Crouzeix, AL Mignot *Analyse numérique des équations différentielles*, collec. Math. Appli. pour la maitrise. Masson, 1984.

[4] J.P. Demailly, *Analyse numérique et équations différentielles*, collection Grenoble Sciences.

[5] O.G. Ciarlet, *Introduction et analyse numérique matricielle et à l'optimisation*, Dunod.