Analyse Numérique

Salem MATHLOUTHI

Université Virtuelle de Tunis

2007

Chapitre 1

Introduction

1.1 Rappels sur les matrices

Soit V un espace vectoriel de dimension finie n sur un corps $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$). Soit $\mathcal{B} = \{e_1, e_2, \cdots, e_n\}$ une base de V, un vecteur $v \in V$ admet une représentation unique dans la base $\mathcal{B}$:

$$v = \sum_{i=1}^{n} v_i e_i,$$

En notation matricielle, le vecteur v s'identifie à un vecteur colonne de $\mathbb{K}^n$:

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

on notera par v^T et v^* les vecteurs lignes :

$$v^T = (v_1 \ v_2 \ \cdots \ v_n), \qquad v^* = (\bar{v}_1 \ \bar{v}_2 \ \cdots \ \bar{v}_n).$$

On définit le produit scalaire dans le cas $\mathbb{K} = \mathbb{R}$ (resp. le produit hermitien dans le cas $\mathbb{K} = \mathbb{C}$) par :

$$(u, v) = u^T v = \sum_{i=1}^{n} u_i v_i,$$

(resp.

$$(u, v) = u^* v = \sum_{i=1}^{n} \bar{u}_i v_i).$$

On dit qu'une base $\mathcal{B} = \{e_1, e_2, \cdots, e_n\}$ est orthonormée si :

$$e_i^* e_j = \delta_{i,j}$$

où $\delta_{i,j}$ est le *symbole de Kronecker* : $\delta_{i,j} = 1$ *si* $i = j$, $\delta_{i,j} = 0$ *si* $i \neq j$.

Soient V et W deux espaces vectoriels sur le même corps $\mathbb{K}$, munis de bases $\{e_1, e_2, \cdots, e_n\}$ et $\{f_1, f_2, \cdots, f_m\}$ respectivement. On rappelle qu'une application linéaire $\mathcal{L}$ de l'espace vectoriel V dans l'espace vectoriel W est définie d'une manière unique par l'image de la base de V, c'est-à-dire

$$\mathcal{L}e_j = \sum_{i=1}^{m} a_{ij} f_i, \qquad 1 \leq j \leq n.$$

Si on note A la matrice $m \times n$ (m lignes et n colonnes) définie par :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

on voit que la $j$ème colonne de la matrice A représente le vecteur $\mathcal{L}e_j$ et on vérifie facilement que pour tout $v \in V$ le vecteur $\mathcal{L}v$ de W est représenté par le vecteur colonne noté Av suivant :

$$(Av)_i = (\text{ ième ligne de } A)\ v, \quad 1 \leq i \leq m.$$

On définit la matrice transposée de A, notée A^T, (resp. la matrice adjointe de A, notée A^*) :

$$(A^T)_{ij} = (A)_{ji} = a_{ji}$$

(resp.

$$(A^*)_{ij} = \overline{(A)_{ji}} = \overline{a_{ji}}).$$

Dans toute la suite, sauf mention du contraire, les matrices considérées seront supposées carrées d'ordre n, c'est-à-dire le nombre de lignes est égal au nombre de colonnes qui est égal à n.

Une matrice A carrée est par définition dite inversible s'il existe une matrice notée A^{-1}, telle que :

$$AA^{-1} = A^{-1}A = (\delta_{ij}) = I = \text{ matrice identité.}$$

On vérifie facilement que si A et B sont deux matrices inversibles, alors :

$$(BA)^{-1} = A^{-1}B^{-1}, \quad (A^T)^{-1} = (A^{-1})^T, \quad (A^*)^{-1} = (A^{-1})^*.$$

Une matrice carrée $A = (a_{ij})$ est par définition :

- *symétrique* si A est réelle et $A^T = A$,
- *hermitienne* si $A^* = A$,
- *orthogonale* si A est réelle et $A^T A = AA^T = I$,
- *unitaire* si $A^* A = AA^* = I$,
- *normale* si $A^* A = AA^*$,
- *diagonale* si $a_{ij} = 0$ pour $i \neq j$, on note $A = diag(a_{ii})$,
- *triangulaire supérieure (resp. inférieure)* si $a_{ij} = 0$ pour $i > j$ (resp. $a_{ij} = 0$ pour $i < j$),
- *semblable* à une matrice B s'il existe une matrice P inversible, dite matrice de passage, telle que $P^{-1}AP = B$,

- *diagonalisable* si elle est semblable à une matrice diagonale.

La trace d'une matrice $A = (a_{ij})$ carrée d'ordre n est définie par

$$tr(A) = \sum_{i=1}^{n} a_{ii},$$

et son déterminant est défini par

$$det(A) = \sum_{\sigma \in \mathcal{G}_n} \varepsilon_\sigma a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}$$

où $\mathcal{G}_n$ est l'ensemble de toutes les permutations de l'ensemble $\{1, 2, \cdots, n\}$ et ε_σ désigne la signature de σ.

Les valeurs propres $\lambda_i(A)$, $1 \leq i \leq n$, d'une matrice A carrée d'ordre n sont les n racines de son polynôme caractéristique :

$$P_A(x) = det(xI - A).$$

Le spectre de la matrice A est défini par

$$sp(A) = \{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\}.$$

Le rayon spectral d'une matrice A carrée d'ordre n est défini par

$$\rho(A) = \max_i \{\mid \lambda_i(A) \mid\}.$$

On rappelle que pour toute valeur propre $\lambda \in sp(A)$, il existe au moins un vecteur $v \neq 0$ tel que $Av = \lambda v$ appelé vecteur propre.

THEOREME 1.1.1 *Soit A une matrice carrée d'ordre n, il existe une matrice unitaire U telle que U^*AU soit triangulaire. Si de plus les coefficients de la matrice A sont réels et ses valeurs propres sont réelles, alors, il existe une matrice orthogonale O telle que O^TAO soit triangulaire.*

DEMONSTRATION. Soient E un espace vectoriel de dimension finie n sur le corps $\mathbb{C}$ et $\{e_1, e_2, \cdots, e_n\}$ une base de E orthonormée au sens du produit hermitien :

$$e_i^* e_j = \delta_{ij}, \quad 1 \leq i, j \leq n$$

Soit $A = (a_{ij})$ une matrice carrée d'ordre n à coefficients complexes, on peut toujours l'associer à une application linéaire $\mathcal{A} : E \to E$ relativement à la base $\{e_1, e_2, \cdots, e_n\}$ par l'image de la base :

$$\mathcal{A}(e_j) = \sum_{i=1}^{n} a_{ij} e_i, \quad 1 \leq j \leq n$$

et si λ est une valeur propre de la matrice A, alors, il existe un vecteur non nul $v \in E$ tel que :

$$\mathcal{A}(v) = \lambda v$$

En utilisant l'application linéaire $\mathcal{A}$, la première partie du théorème consiste à prouver l'existence d'une nouvelle base orthonormée (au sens du produit hermitien) $\{f_1, f_2, \cdots, f_n\}$ telle que :

$$\forall 1 \leq j \leq n, \quad \mathcal{A}f_j \in < f_1, f_2, \cdots, f_j >$$

où, $< f_1, f_2, \cdots, f_j >$ représente le sous-espace de E engendré par les vecteurs $f_1, f_2, \cdots, f_j$ sur le corps $\mathbb{C}$. Ce qui prouve l'existence d'une matrice T triangulaire supérieure associée à l'application linéaire $\mathcal{A}$ relativement à la base $\{f_1, f_2, \cdots, f_n\}$. D'autre part, la nouvelle base $\{f_1, f_2, \cdots, f_n\}$ est orthonormée au sens du produit hermitien, ce qui prouve que la matrice de passage, qu'on note U, est unitaire ($U^*U = I$) et on a : $T = U^*AU$.

Montrons par récurrence l'existence d'une base orthonormée $\{f_1, f_2, \cdots, f_n\}$ telle que :

$$\forall 1 \leq j \leq n, \quad \mathcal{A}f_j \in < f_1, f_2, \cdots, f_j >$$

Pour $n = 1$ le résultat est évident. Supposons qu'il est vrai jusqu'à l'odre m. Soit $\mathcal{A}$ est une application linéaire de $E \to E$ avec E espace vectoriel de dimension $n = m + 1$ sur le corps $\mathbb{C}$. Soit $\lambda \in \mathbb{C}$ une valeur propre de la matrice associée à l'application linéaire $\mathcal{A}$ et $v_1 \in E$ un vecteur propre associé à λ, donc :

$$\mathcal{A}v_1 = \lambda v_1$$

Quitte à diviser par $v_1^* v_1$, on peut supposer que $v_1^* v_1 = 1$. Soit $v_2, v_3, \cdots, v_n$, $n-1$ vecteurs de E tels que $\{v_1, v_2, \cdots, v_n\}$ soit une base orthonormée de E. Donc

$$\begin{array}{rcl} \mathcal{A}v_1 & = & \lambda v_1 \\ \mathcal{A}v_2 & = & \sum_{k=1}^{n} \alpha_{k,2} v_k \\ & \vdots & \\ \mathcal{A}v_n & = & \sum_{k=1}^{n} \alpha_{k,n} v_k \end{array}$$

Soit $\mathcal{B}$ l'application linéaire de l'espace vectoriel $F =< v_2, v_3, \cdots, v_n >$, sur le corps $\mathbb{C}$, dans lui même définie par :

$$\forall\ 2 \leq j \leq n, \quad \mathcal{B}v_j = \sum_{k=2}^{n} \alpha_{k,j} v_k$$

d'après l'hypothèse de récurrence, il existe une base orthonormée $\{f_2, f_3, \cdots, f_n\}$ de F telle que :

$$\forall\ 2 \leq j \leq n, \left\{ \begin{array}{rcl} f_j & = & \sum_{i=2}^{n} \gamma_{i,j} v_i \\ \mathcal{B}f_j & \in & < f_2, f_3, \cdots, f_j > \end{array} \right.$$

On pose $f_1 = v_1$. Il est clair que $\{f_1, f_2, \cdots, f_n\}$ est une base orthonormée et on a :

$$\begin{array}{rcl} \mathcal{A}f_1 & = & \lambda f_1 \\ \mathcal{A}f_j & = & \sum_{i=2}^{n} \gamma_{i,j} \mathcal{A}v_i \\ & = & \sum_{i=2}^{n} \gamma_{i,j} (\alpha_{1,i} v_1 + \mathcal{B}v_i) \\ & = & (\sum_{i=2}^{n} \gamma_{i,j} \alpha_{1,i}) f_1 + \mathcal{B}f_j \\ & \in & < f_1, f_2, \cdots, f_j > \end{array}$$

ce qui termine la démonstration par récurrence. La même démotranstion reste valable si on remplace le corps $\mathbb{C}$ par le corps $\mathbb{R}$ et on suppose que les valeurs propres de $\mathcal{A}$ sont réelles ; dans ce cas tous les coefficients seront dans $\mathbb{R}$ ce qui démontre la deuxième partie du théorème.

COROLLAIRE 1.1.1 *1) Toute matrice normale est diagonalisable et admet une base orthonormée (pour le produit hermitien) de vecteurs propres. En particulier, les matrices unitaires, orthogonales, hermitiennes et symétriques sont diagonalisables.*

2) Les valeurs propres d'une matrice hermitienne ou symétrique sont réelles.

3) Toute matrice symétrique admet une base réelle orthonormée de vecteurs propres, c'est-à-dire il existe une matrice orthogonale O telle que $O^T AO$ soit diagonale.

DEMONSTRATION.

1/ D'après le théorème 1.1.1, il existe une matrice U unitaire telle que : $U^*AU = T = (t_{ij}) =$ une matrice triangulaire. Si on note par $p_1,\ p_2, \cdots, p_n$, les colonnes de U, alors, $p_1^*,\ p_2^*, \cdots, p_n^*$ sont les lignes de U^*. Par conséquent, U est une matrice unitaire ($U^*U = I$) est équivalent à

$$\forall i = 1, n,\ \forall j = 1, n,\ (p_i)^*(p_j) = \delta_{ij}$$

On va démontrer que T est une matrice diagonale lorsque la matrice A est normale. Montrons d'abord que T est normale :

$$T^* = (U^*AU)^* = U^*A^*(U^*)^* = U^*A^*U$$

$$T^*T = U^*A^*UU^*AU = U^*A^*AU = U^*AA^*U = TT^*$$

ce qui prouve que T est normale. Pour démontrer que T est diagonale, on va comparer les coefficients de T^*T et TT^* d'indice 11 :

$$(T^*T)_{11} = \sum_{i=1}^{n} \overline{t_{i1}} t_{i1} = |t_{11}|^2$$

$$(TT^*)_{11} = \sum_{i=1}^{n} \overline{t_{1i}} t_{1i} = |t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2$$

ce qui donne : $t_{12} = t_{13} = \cdots = t_{1n} = 0$. On refait la même chose avec les coefficients d'indice 22 pour démontrer que les coefficients de la deuxième ligne de T en dehors de la diagonale sont nuls, puis avec les coefficients d'indice 33, etc ...

2/ On a :

$$U^*AU = D = diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$$

comme A est hermitienne, alors :

$$\begin{aligned} D^* &= diag(\overline{\lambda_1}, \overline{\lambda_2}, \cdots, \overline{\lambda_n}) \\ &= (U^*AU)^* = U^*A^*U \\ &= U^*AU = D \\ &= diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \end{aligned}$$

c.q.f.d.

3/ On utilise la deuxième partie du théorème 1.1.1.

Exercice d'application :

a) Montrer que si une matrice carrée A est triangulaire et normale, alors, elle est diagonale.

b) Montrer les relations

(i) det(matrice triangulaire) $= a_{11}a_{22}\cdots a_{nn}$,

(ii) sp(matrice triangulaire) $= \{a_{ii},\ i = 1, n\}$,

(iii) $tr(A + B) = tr(A) + tr(B)$, $tr(AB) = tr(BA)$, $tr(A) = \sum_{i=1}^{n} \lambda_i(A)$,

(iv) $det(A^T) = det(A) = \lambda_1(A)\lambda_2(A)\cdots\lambda_n(A)$, $det(AB) = det(BA)$,

(v) $sp(A^T) = sp(A)$, $sp(AB) = sp(BA)$,

(vi) $k \in \mathbb{N}$, $sp(A^k) = \{(\lambda_i(A))^k,\ i = 1, n\}$, (utiliser le théorème 1.1.1).

Réponse :

a) Voir la démonstration du corollaire 1.1.1.

b)(i) Supposons que $A = (a_{ij})$ est une matrice triangulaire supérieure, c'est à dire : $a_{ij} = 0$ pour $i > j$, (même raisonnement dans le cas triangulaire inférieure, en utilisant (iv)). Soit σ une permutation. Si le produit $a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}$ est non nul, alors, nécéssairement $\sigma(i) \leq i$ pour tout $i = 1, n$. D'où :

$$\begin{array}{ll} \sigma(1) \leq 1 & \Rightarrow \sigma(1) = 1 \\ \sigma(2) \neq \sigma(1) = 1;\ \sigma(2) \leq 2 & \Rightarrow \sigma(2) = 2 \\ & \vdots \\ \sigma(n) \neq \sigma(i) = i, 1 \leq i \leq n-1;\ \sigma(n) \leq n & \Rightarrow \sigma(n) = n \end{array}$$

donc, toutes les permutations donnent un produit nul, sauf peut-être l'identité. Ce qui donne :

$$det(A) = a_{11}a_{22}\cdots a_{nn}$$

(ii) Si A est une matrice triangulaire, alors, $A - xI$ est aussi triangulaire et les coefficients de la diagonale sont : $a_{ii} - x$, $i = 1, n$. D'après (i), le polynôme caractéristique de A est égal à :

$$P_A(x) = (a_{11} - x)(a_{22} - x)\cdots(a_{nn} - x)$$

c.q.f.d.

(iii)(iv)(v) $A = (a_{ij})$, $B = (b_{ij})$, $AB = (c_{ij})$, $BA = (d_{ij})$, par définition du produit de deux matrices, on a :

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}, \quad d_{ij} = \sum_{k=1}^{n} b_{ik}a_{kj}$$

d'où

$$\begin{array}{rl} tr(A+B) & = \sum_{k=1}^{n}(a_{ii} + b_{ii}) \\ & = \sum_{k=1}^{n} a_{ii} + \sum_{k=1}^{n} b_{ii} \\ & = tr(A) + tr(B) \end{array}$$

$$\begin{array}{rl} tr(AB) & = \sum_{i=1}^{n} c_{ii} \\ & = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki} \\ & = \sum_{k=1}^{n}\sum_{i=1}^{n} b_{ki}a_{ik} \\ & = \sum_{k=1}^{n} d_{kk} \\ & = tr(BA) \end{array}$$

D'après le théorème 1.1.1, toute matrice est semblable à une matrice triangulaire T. D'après (ii), les valeurs propres d'une matrice triangulaire T sont les coefficients de la diagonale de T. Montrons que deux matrices semblables ont les mêmes valeurs propres : soit $\lambda \in sp(P^{-1}AP)$, par définition, il existe $v \neq 0$ tel que $P^{-1}APv = \lambda v$, d'où $APv = \lambda Pv$ et comme P est inversible, alors $Pv \neq 0$, par suite Pv est un vecteur propre de A associé à λ, ce qui montre que $\lambda \in sp(A)$. De la même manière, on montre que les valeurs propres de A sont des valeurs propres de $P^{-1}AP$. Donc, la matrice A et la matrice triangulaire associée à A par le théorème 1.1.1 ont le même polynôme caractéristique puisqu'elles ont les mêmes valeurs propres. D'autre part, un calcul simple nous montre que :

$$P_B(x) = det(B - xI) = (-1)^n[x^n - tr(B)x^{n-1} + \cdots + (-1)^n det(B)]$$

ce qui montre que $tr(A) = tr(T)$ et $det(A) = det(T)$, soit encore :

$$\begin{array}{ll} tr(A) & = \sum_{i=1}^{n} \lambda_i(A) \\ det(A) & = \prod_{i=1}^{n} \lambda_i(A) \end{array}$$

Montrons que : $det(A) = det(A^T)$. Il suffit de remarquer que si une matrice B est semblable à une matrice C , alors, B^T est semblable à C^T . Donc, A^T est semblable à T^T qui est une matrice triangulaire inférieure (ses valeurs propres sont sur sa diagonale, même raisonnement que le cas d'une matrice triangulaire supérieure). Or, la diagonale de T est la même que la diagonale de T^T , donc, T et T^T ont les mêmes valeurs propres. Par conséquent, A^T et A ont les mêmes valeurs propres. Ce qui prouve que : $det(A) = det(A^T)$ et $sp(A) = sp(A^T)$. Montrons que : $sp(AB) = sp(BA)$ (ce qui prouve en particulier que $det(AB) = det(BA)$). Il suffit de prouver que $sp(AB) \subset sp(BA)$ et par symétrie on a l'autre inclusion. Soit $\lambda \in sp(AB)$, alors il existe $v \neq 0$ tel que $ABv = \lambda v$:

1. Si $Bv \neq 0$, alors, $BA(Bv) = \lambda Bv$ (on applique B de deux côtés) d'où Bv est un vecteur propre de BA associé à la valeur propre λ , ce qui montre que $\lambda \in sp(BA)$.
2. Si $Bv = 0$, alors nécéssairement $\lambda = 0$. D'autre part, si A est inversible, il existe $w \neq 0$ tel que $Aw = v$, sinon, il existe $w \neq 0$ tel que $Aw = 0$. Dans les deux cas $BAw = 0$, ce qui montre que w est un vecteur propre de BA associé à la valeur propre $\lambda = 0$, d'où $\lambda = 0 \in sp(BA)$.

(vi) On vérifie facilement que si B est semblable à C alors B^k est semblable à C^k, $k \in \mathbb{N}$. D'où, A^k est semblable à T^k (T est la matrice triangulaire donnée par le théorème 1.1.1). Un calcul simple nous montre que T^k, $k \in \mathbb{N}$, est aussi triangulaire supérieure et les coefficients de sa diagonale sont les t_{ii}^k, $i = 1, n$.

REMARQUE 1.1.1 *D'après le corollaire 1.1.1, si A est une matrice hermitienne, il existe une matrice unitaire $P = U^* = U^{-1}$ telle que :*

$$A = P^*\ diag(\lambda_i)\ P$$

d'où, pour tout $v \in V$:

$$v^*Av = v^*P^*\ diag(\lambda_i)\ Pv = (Pv)^*\ diag(\lambda_i)\ (Pv) = \sum_{i=1}^{n} \lambda_i |\tilde{v}_i|^2$$

avec $\tilde{v}_i$ la ième composante du vecteur Pv, d'où :

$$\lambda_{max}||Pv||_2^2 = \lambda_{max}\sum_{i=1}^{n}|\tilde{v}_i|^2 \geq v^*Av \geq \lambda_{min}||Pv||_2^2 = \lambda_{min}\sum_{i=1}^{n}|\tilde{v}_i|^2$$

avec λ_{max} la plus grande valeur propre de A et λ_{min} la plus petite valeur propre de A. D'autre part, la matrice P est unitaire ce qui donne $||Pv||_2 = ||v||_2$, d'où :

$$\lambda_{max}||v||_2^2 = \lambda_{max}\sum_{i=1}^{n}|v_i|^2 \geq v^*Av \geq \lambda_{min}||v||_2^2 = \lambda_{min}\sum_{i=1}^{n}|v_i|^2$$

ce qui prouve en particulier que :

$$\lambda_{max} = \max_{||u||_2=1} u^*Au;\ \lambda_{min} = \min_{||u||_2=1} u^*Au$$

Le maximum est atteint pour un vecteur propre de A associé à λ_{max} de norme 2 égale à 1 et le minimum est atteint pour un vecteur propre de A associé à λ_{min} de norme 2 égale à 1.

DEFINITION et REMARQUE

On dit qu'une matrice hermitienne est définie positive (resp. positive) si :

$$v^*Av > 0, \forall v \in V - \{0\}, \qquad (resp. v^*Av \geq 0, \forall v \in V)\ ;$$

d'après ce qui précède, on a

$$v^*Av \geq \lambda_{min}v^*v, \text{ pour tout } v \in V\ ;$$

d'où, une matrice hermitienne est définie positive (resp. positive) si et seulement si $\lambda_{min} > 0$ (resp.$\lambda_{min} \geq 0$).

1.2 Normes et suites de matrices

Soit V un espace vectoriel de dimension finie. On rappelle les trois normes vectorielles suivantes :

$$\begin{aligned} \| v \|_1 &= \sum_i | v_i | \\ \| v \|_2 &= (\sum_i | v_i |^2)^{1/2} \\ \| v \|_\infty &= \max_i | v_i | \end{aligned}$$

DEFINITION 1.2.1 *Une norme matricielle, notée comme la norme vectorielle par $\| . \|$, est par définition une norme vectorielle :*

(i) $\| A \| \geq 0$ *et* $\| A \| = 0 \Longleftrightarrow A = 0$,

(ii) $\| \gamma A \| = | \gamma | \| A \|$ *pour tout scalaire* γ,

(iii) $\| A + B \| \leq \| A \| + \| B \|$,

de plus elle vérifie :

(iv) $\| AB \| \leq \| A \| \| B \|$.

Etant donné une norme vectorielle $\| \ . \ \|$, on lui associe une norme matricielle, appelée norme matricielle subordonnée, de la manière suivante :

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

(noter que les normes à droite sont vectorielles).

Exercice d'application : Vérifier que
(i) la norme subordonnée est bien une norme matricielle,
(ii) $\| \ A \ \| = \max_{x \neq 0} \frac{\| \ Ax \ \|}{\| \ x \ \|} = \max_{\|u\|=1} \| \ Au \ \|$

Réponse :

1. (a)

$$\begin{array}{lll} \|A\| = 0 & \Longleftrightarrow \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = 0 & \Longleftrightarrow \forall x \neq 0, \frac{\|Ax\|}{\|x\|} = 0 \\ & \Longleftrightarrow \forall x, \|Ax\| = 0 & \Longleftrightarrow A = 0 \end{array}$$

   (b) $\|\lambda A\| = \sup_{x \neq 0} \frac{\|\lambda Ax\|}{\|x\|} = \sup_{x \neq 0} \frac{|\lambda| \|Ax\|}{\|x\|} = |\lambda| \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = |\lambda| \|A\|$

   (c) A et B deux matrices de même ordre :

$$\begin{array}{lll} \forall x \neq 0, \frac{\|(A+B)x\|}{\|x\|} & = \frac{\|Ax+Bx\|}{\|x\|} & \leq \frac{\|Ax\|+\|Bx\|}{\|x\|} \\ & \leq \frac{\|Ax\|}{\|x\|} + \frac{\|Bx\|}{\|x\|} & \leq \|A\| + \|B\| \end{array}$$

   d'où : $\|A + B\| = \sup_{x \neq 0} \frac{\|(A+B)x\|}{\|x\|} \leq \|A\| + \|B\|$.

   (d) De la définition de $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$, on déduit que : $\|Ax\| \leq \|A\|\|x\|$ pour tout x. Soient A et B deux matrices de même ordre, on a :

$$\forall x \neq 0, \frac{\|ABx\|}{\|x\|} = \frac{\|A(Bx)\|}{\|x\|} \leq \frac{\|A\|\|Bx\|}{\|x\|} \leq \frac{\|A\|\|B\|\|x\|}{\|x\|} = \|A\|\|B\|$$

   ce qui prouve que : $\|AB\| \leq \|A\|\|B\|$.

2.

$$\begin{array}{ll} \|A\| & = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \neq 0} \|A(\frac{x}{\|x\|})\| \\ & \leq \sup_{\|u\|=1} \|Au\| = \sup_{\|u\|=1} \frac{\|Au\|}{\|u\|} \\ & \leq \sup_{u \neq 0} \frac{\|Au\|}{\|u\|} = \|A\| \end{array}$$

Les normes matricielles subordonnées aux normes $\| \ . \ \|_1, \| \ . \ \|_2$ et $\| \ . \ \|_\infty$ sont données par le théorème suivant :

THEOREME 1.2.1 **(1)** $\| \ A \ \|_1 = \max_j \| \ a_j \ \|_1$, *(a_j : jème colonne de A),*
(2) $\| \ A \ \|_2 = \sqrt{\rho(A^*A)}$,
(3) $\| \ A \ \|_\infty = \max_i \| \ a'_i \ \|_1$, *(a'_i : ième ligne de A),*

DEMONSTRATION.

(1) $u \in \mathbb{R}^n,\ \| u \|_1 = 1$, alors :

$$\begin{aligned} \| Au \|_1 &= \sum_{k=1,n} | \sum_{j=1,n} a_{k,j} u_j | \\ &\leq \sum_{k=1,n} \sum_{j=1,n} | a_{k,j} u_j | \\ &\leq \sum_{j=1,n} (\sum_{k=1,n} | a_{k,j} |) | u_j | \\ &\leq \max_j (\sum_{k=1,n} | a_{k,j} |) \sum_{j=1,n} | u_j | \end{aligned}$$

ce qui prouve que

$$\| A \|_1 \leq \max_j \| a_j \|_1$$

D'autre part, soit j_0 tel que :

$$\max_j \| a_j \|_1 = \| a_{j_0} \|_1$$

on pose

$$v_{j_0} = 1, \quad v_j = 0, \ \ \forall j \neq j_0$$

on obtient que $\| v \|_1 = 1$ et $\| Av \|_1 = \max_j \| a_j \|_1 \leq \| A \|_1$, d'où l'autre inégalité.

(2)

$$\begin{aligned} \| A \|_2 &= \max_{\|u\|_2 = 1} \| Au \|_2 \\ &= \max_{\|u\|_2 = 1} \sqrt{u^* A^* A u} \\ &= \sqrt{\rho(A^* A)} \end{aligned}$$

car la matrice A^*A est hermitienne positive (voir remarque 1.1.1).

(3) $u \in \mathbb{R}^n,\ \| u \|_\infty = 1$, alors :

$$\begin{aligned} \| Au \|_\infty &= \max_k | \sum_{j=1,n} a_{k,j} u_j | \\ &\leq \max_k \sum_{j=1,n} | a_{k,j} u_j | \\ &\leq \| u \|_\infty \max_k (\sum_{j=1,n} | a_{k,j} |) \end{aligned}$$

ce qui prouve que

$$\| A \|_\infty \leq \max_k \| a'_k \|_1$$

D'autre part, soit k_0 tel que :

$$\max_k \| a'_k \|_1 = \| a'_{k_0} \|_1$$

on pose

$$\forall j, \quad v_j = \begin{cases} signe(a_{k_0,j}) & , \quad \text{si } a_{k_0,j} \neq 0, \\ 0 & , \quad \text{sinon} \end{cases}$$

on obtient que $\| v \|_\infty = 1$ et $\| Av \|_\infty = \max_k \| a'_k \|_1 \leq \| A \|_\infty$, d'où l'autre inégalité.

On vérifie sans peine que la norme matricielle subordonnée à la norme 2 d'une matrice normale est égale à son rayon spectral (Indication : utiliser le corollaire 1.1.1 et le fait que la norme 2 est invariante par transformation unitaire). Une question s'impose, le rayon spectral définit-il une norme matricielle ? La réponse est donnée par l'exemple suivant :

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

on a que A est une matrice non nulle et son rayon spectral est nul.

Le théorème suivant compare le rayon spectral et une norme matricielle quelconque,

THEOREME 1.2.2 *(1) Soit A une matrice quelconque et $\| \ . \ \|$ une norme matricielle quelconque. Alors*

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

(2) Etant donné une matrice A et $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée telle que

$$\| A \| \leq \rho(A) + \varepsilon.$$

DEMONSTRATION.

(1) Soit λ une valeur propre de A telle que $\rho(A) = | \lambda |$. Soit $v \neq 0$ un vecteur propre de A associé à λ. Soit B la matrice carrée d'ordre n définie par :

$$Bx = (v^* x)v, \quad \forall x \in \mathbb{C}^n$$

On a :

$$\| AB \| \leq \| A \| \| B \|, \text{ et } \| B \| \neq 0$$

D'autre part

$$ABx = A(v^* x)v = (v^* x)\lambda v, \quad \forall x \in \mathbb{C}^n$$

par conséquent $AB = \lambda B$, d'où $\| AB \| = | \lambda | \| B \|$. Comme $\| B \| \neq 0$, alors

$$\rho(A) = | \lambda | \leq \| A \|$$

(2) D'après le théorème 1.1.1, il existe une matrice unitaire U telle que :

$$T = U^{-1} A U$$

est une matrice triangulaire supérieure :

$$T = \begin{pmatrix} t_{11} & \cdots & \cdots & t_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix}$$

avec $sp(A) = \{t_{11}, t_{22}, t_{33}, \cdots, t_{nn}\}$. Soit $\delta > 0$, on pose :

$$P_\delta = diag\{1, \delta, \delta^2, \cdots, \delta^{n-1}\}$$

On vérifie facilement que

$$\lim_{\delta \to 0} P_\delta^{-1} T P_\delta = diag\{t_{11}, t_{22}, \cdots, t_{nn}\}$$

d'où

$$\lim_{\delta \to 0} \parallel P_\delta^{-1} U^{-1} A U P_\delta \parallel_2 = \parallel diag\{t_{11}, t_{22}, \cdots, t_{nn}\} \parallel_2 = \rho(A)$$

Soit $\varepsilon > 0$, on choisit $\delta > 0$ tel que :

$$\parallel P_\delta^{-1} U^{-1} A U P_\delta \parallel_2 \leq \rho(A) + \varepsilon$$

et on définit l'application de l'espace des matrices carrées d'ordre n dans $\mathbb{R}^+$

$$B \rightarrow \parallel B \parallel = \parallel P_\delta^{-1} U^{-1} B U P_\delta \parallel_2$$

on vérifie facilement que $\parallel . \parallel$ est une norme matricielle subordonnée à la norme vectorielle suivante :

$$v \rightarrow \parallel v \parallel = \parallel P_\delta^{-1} U^{-1} v \parallel_2$$

La notion de convergence d'une suite de matrices n'est autre que la notion classique de convergence dans les espaces vectoriels normés.

THEOREME 1.2.3 *Soit B une matrice carrée. Les conditions suivantes sont équivalentes :*

(1) $\lim_{k \to \infty} B^k = 0$,

(2) $\lim_{k \to \infty} B^k v = 0$, *pour tout vecteur* v,

(3) $\rho(B) < 1$,

(4) $\parallel B \parallel < 1$ *pour au moins une norme matricielle subordonnée.*

DEMONSTRATION.

(1) $\Longrightarrow$ (2) évidente.

(2) $\Longrightarrow$ (3) Soit λ une valeur propre de B telle que $\rho(B) = \mid \lambda \mid$. Soit $v \neq 0$ un vecteur propre de B associé à λ. On a :

$$B^k v = \lambda^k v$$

d'où

$$\parallel B^k v \parallel = (\rho(B))^k \parallel v \parallel$$

ce qui prouve que $\lim_{k \to \infty} \rho(B)^k = 0$ car $\parallel v \parallel \neq 0$, d'où, $\rho(B) < 1$.

(3) $\Longrightarrow$ (4) on applique (2) du théorème 1.2.2 avec $\varepsilon = \frac{1-\rho(B)}{2}$.

(4) $\Longrightarrow$ (1) évidente en utilisant la propriété de la norme matricielle suivante :

$$\parallel B^k \parallel \leq \parallel B \parallel^k$$

THEOREME 1.2.4 *Soit B une matrice carrée, et* $\parallel . \parallel$ *une norme matricielle quelconque. Alors*

$$\lim_{k \to \infty} \parallel B^k \parallel^{1/k} = \rho(B).$$

DEMONSTRATION.

En utilisant le théorème 1.1.1, on vérifie facilement que :

$$sp(B^k) = \{\lambda^k \ ; \ \lambda \in sp(B)\}$$

d'où

$$\rho(B^k) = (\rho(B))^k$$

d'après le théorème 1.2.2, on a :

$$\rho(B^k) \leq \| \ B^k \ \|$$

d'où

$$\rho(B) \leq \| \ B^k \ \|^{1/k}$$

Soit $\varepsilon > 0$, d'après (2) du théorème 1.2.2, il existe une norme matricielle $\| \ . \ \|_\varepsilon$ telle que :

$$\| \ B \ \|_\varepsilon \leq \rho(B) + \varepsilon$$

D'autre part, toutes les normes sont équivalentes car l'espace vectoriel des matrices est un espace vectoriel de dimension finie et la norme matricielle et en particulier une norme vectorielle, donc :

$$\exists c_\varepsilon > 0 \ \ ; \ \ \| \ A \ \| \leq c_\varepsilon \ \| \ A \ \|_\varepsilon, \ \forall \ A$$

Par conséquent :

$$\rho(B) \leq \| \ B^k \ \|^{1/k} \leq c_\varepsilon^{1/k} \ \| \ B^k \ \|_\varepsilon^{1/k} \leq c_\varepsilon^{1/k}(\rho(B) + \varepsilon)$$

Par passage à la limite sur k puis sur ε, on obtient ce qu'il faut.

Une importante norme matricielle non subordonnée à une norme vectorielle est la norme de *Frobenius* définie pour toute matrice $A \in \mathcal{M}_{mn}(\mathbb{R})$ par :

$$\| \ A \ \|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} \mid a_{ij} \mid^2)^{1/2},$$

on vérifie que

$$\| \ A \ \|_F^2 = trace(A^* A)$$

ce qui prouve que la norme de *Frobenius* est invariante par transformation unitaire. La norme de *Frobenius* est essentiellement la norme Euclidienne appliquée à un vecteur de mn composantes. Il est facile de voir que la norme de l'identité I est toujours égale à 1 pour une norme subordonnée. D'autre part, il est clair que $\| \ I \ \|_F = \sqrt{n}$. D'où, pour $n \geq 2$, la norme de *Frobenius* ne peut pas être subordonnée à une norme vectorielle.

1.3 Méthode de Gauss pour la résolution des systèmes linéaires

L'idée de base derrière les méthodes de résolution de $Ax = b$ est la transformation de ce problème en un problème facile à résoudre. Considérons

l'exemple d'un système en trois dimensions :

$$\begin{array}{rcrcrcl} x_1 & + & x_2 & + & 2x_3 & = & 3 \\ 2x_1 & + & 3x_2 & + & x_3 & = & 2 \\ 3x_1 & - & x_2 & - & x_3 & = & 6 \end{array},$$

ce qui corresponds à

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix}, \qquad b = \begin{pmatrix} 3 \\ 2 \\ 6 \end{pmatrix}.$$

L'inconnu x_1 peut-être éliminer de la deuxième équation et de la troisième équation en retranchant de la deuxième équation deux fois la première équation et de la troisième équation trois fois la première équation, on obtient alors, le système suivant

$$\begin{array}{rcrcrcl} x_1 & + & x_2 & + & 2x_3 & = & 3 \\ & & x_2 & - & 3x_3 & = & -4 \\ & - & 4x_2 & - & 7x_3 & = & -3 \end{array}.$$

(Il est clair que ces transformations ne changent pas la solution.)

L'inconnu x_2 peut-être éliminer de la troisième équation du système en ajoutant à la dernière équation quatre fois la deuxième équation, on obtient, le système triangulaire suivant :

$$\begin{array}{rcrcrcl} x_1 & + & x_2 & + & 2x_3 & = & 3 \\ & & x_2 & - & 3x_3 & = & -4 \\ & & & - & 19x_3 & = & -19 \end{array}.$$

La solution de notre système de départ peut-être déterminée directement à partir des équations du dernier système. Dans la dernière équation il y a seulement x_3, et on a $x_3 = 1$. En remplaçant x_3 par sa valeur dans la deuxième équation, on obtient

$$x_2 = -1.$$

Enfin, en remplaçant x_3 et x_2 dans la première équation, on obtient

$$x_1 = 2$$

soit

$$x = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}.$$

On a donc transformé notre système en un système qui a la même solution et qui est facile à résoudre, c'est le système triangulaire.

1.3.1 Système linéaire triangulaire. Méthode de remontée.

Supposons que nous voulons résoudre

$$Ux = b,$$

où U est une matrice triangulaire supérieure $n \times n$ inversible. On a alors, n équations sous la forme :

$$\begin{array}{ccccccccc} u_{11}x_1 & + & u_{12}x_2 & + & \cdots & + & u_{1n}x_n & = & b_1 \\ & & u_{22}x_2 & + & \cdots & + & u_{2n}x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & u_{nn}x_n & = & b_n \end{array} .$$

(La matrice U est inversible si et seulement si les éléments de la diagonale sont non nuls.)

La n^{eme} équation dépend uniquement de l'inconnu x_n, on a

$$u_{nn}x_n = b_n, \text{ soit } x_n = \frac{b_n}{u_{nn}}.$$

La $(n-1)^{eme}$ équation

$$u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n = b_{n-1}$$

dépend uniquement de x_n et x_{n-1}, or, x_n est connu, d'où

$$x_{n-1} = \frac{1}{u_{n-1,n-1}}(b_{n-1} - u_{n-1,n}x_n).$$

Pour $k > 0$, x_k est déterminé de la même manière que les inconnus

$$x_n, x_{n-1}, \cdots, x_{k+1}$$

par

$$x_k = \frac{1}{u_{kk}}(b_k - u_{k,k+1}x_{k+1} - u_{k,k+2}x_{k+2} - \cdots - u_{k,n}x_n).$$

Pour tout $k = 1, \cdots, n$, le calcul de x_k nécessite une division, $n-k$ additions et $n-k$ multiplications. D'où, le nombre nécessaire d'opérations élémentaires pour résoudre par la méthode de remontée un système triangulaire est :

$$\begin{array}{l} n \text{ divisions} \\ (n-1) + (n-2) + \cdots + 2 + 1 = \frac{n(n-1)}{2} \approx \frac{n^2}{2} \text{ additions} \\ (n-1) + (n-2) + \cdots + 2 + 1 = \frac{n(n-1)}{2} \approx \frac{n^2}{2} \text{ multiplications} \end{array} .$$

Dans le cas d'un système linéaire triangulaire inférieure

$$Lx = b,$$

on utilise les mêmes techniques de la méthode de remontée, au lieu de commencer par l'inconnu x_n et on monte à x_1, on commence par x_1 puis on descend à x_n. On appelle cette procédure la méthode de descente.

Exercice d'application

Montrer en utilisant la méthode de remontée que l'inverse d'une matrice triangulaire $T = (t_{ij})$ inversible est une matrice triangulaire de même nature. De plus, les éléments de la diagonale de la matrice inverse sont les inverses des éléments de la diagonale de la matrice T. (Indication : la jième colonne de la matrice T^{-1} est égale à la solution du système triangulaire $Tx = e_j$ avec e_j est le jième vecteur de la base canonique.)

1.3.2 La méthode d'élimination de Gauss

La résolution d'un système linéaire triangulaire est facile. Exactement comme le travail à la main de l'exemple, on va transformer le système linéaire $Ax = b$ en un système linéaire triangulaire ayant la même solution.

Soit A une matrice carrée d'ordre n

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} a'_1 \\ a'_2 \\ \vdots \\ a'_n \end{pmatrix}$$

où $a_{ij} \in \mathbb{R}$ pour $1 \leq i, j \leq n$ et a'_i représente la i^{eme} ligne de la matrice A, $1 \leq i \leq n$.

La première étape de la résolution de $Ax = b$ consiste à éliminer x_1 de toutes les équations sauf la première. Dans le cas où $a_{11} \neq 0$, on applique la technique de l'exemple, c'est à dire :

$$\begin{array}{c} \textit{la ligne } a'_2 \textit{ est remplacée par } a'_2 - \frac{a_{21}}{a_{11}} a'_1 \\ \textit{la ligne } a'_3 \textit{ est remplacée par } a'_3 - \frac{a_{31}}{a_{11}} a'_1 \\ \vdots \\ \textit{la ligne } a'_n \textit{ est remplacée par } a'_n - \frac{a_{n1}}{a_{11}} a'_1 \end{array},$$

de même pour le vecteur b :

$$\begin{array}{c} \textit{la composante } b_2 \textit{ est remplacée par } b_2 - \frac{a_{21}}{a_{11}} b_1 \\ \textit{la composante } b_3 \textit{ est remplacée par } b_3 - \frac{a_{31}}{a_{11}} b_1 \\ \vdots \\ \textit{la composante } b_n \textit{ est remplacée par } b_n - \frac{a_{n1}}{a_{11}} b_1 \end{array}.$$

L'élément a_{11} s'appelle le pivot. Si $a_{11} = 0$, on cherche un coefficient non nul a_{i1}, $i = 2, \cdots, n$, (un tel coefficient existe, sinon la matrice est non inversible) et on permute la ligne i avec la première ligne pour que le nouveau pivot qui est le coefficient à la position $1, 1$ soit non nul.

A l'étape k, la matrice $A^{(k)}$ a la forme suivante :

$$A^{(k)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} = \begin{pmatrix} a'_1 \\ a_2^{(2)\prime} \\ \vdots \\ a_k^{(k)\prime} \\ \vdots \\ a_n^{(k)\prime} \end{pmatrix}$$

L'étape k de la résolution consiste à éliminer l'inconnu x_k de toutes les équations sauf les k-premières. De la même manière que la première étape, si le pivot

$a_{kk}^{(k)} \neq 0$, on fait les transformations suivantes :

$$\begin{array}{c} \textit{la ligne } {a_{k+1}^{(k)}}' \textit{ est remplacée par } {a_{k+1}^{(k)}}' - \frac{a_{k+1k}^{(k)}}{a_{kk}^{(k)}} {a_k^{(k)}}' \\ \textit{la ligne } {a_{k+2}^{(k)}}' \textit{ est remplacée par } {a_{k+2}^{(k)}}' - \frac{a_{k+2k}^{(k)}}{a_{kk}^{(k)}} {a_k^{(k)}}' \\ \vdots \\ \textit{la ligne } {a_n^{(k)}}' \textit{ est remplacée par } {a_n^{(k)}}' - \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} {a_k^{(k)}}' \end{array}$$

de même pour le vecteur $b^{(k)}$

$$\begin{array}{c} \textit{la composante } b_{k+1}^{(k)} \textit{ est remplacée par } b_{k+1}^{(k)} - \frac{a_{k+1k}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} \\ \textit{la composante } b_{k+2}^{(k)} \textit{ est remplacée par } b_{k+2}^{(k)} - \frac{a_{k+2k}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} \\ \vdots \\ \textit{la composante } b_n^{(k)} \textit{ est remplacée par } b_n^{(k)} - \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} \end{array}$$

Si le coefficient $a_{kk}^{(k)} = 0$, on cherche un coefficient $a_{ik}^{(k)} \neq 0$, $i = k+1, \cdots, n$, (un tel coefficient existe sinon la matrice est non inversible) et on permute la ligne i avec la ligne k pour que le nouveau pivot qui est le coefficient à la position kk soit non nul.

Au bout de $n-1$ étapes, on obtient un système triangulaire.

1.4 Calcul de l'inverse d'une matrice

Dans la pratique, on évite le calcul de l'inverse A^{-1} d'une matrice inversible A. Dans le cas particulier où on a vraiment besoin de l'expression de la matrice A^{-1}, on utilise l'algorithme de Gauss Jordan. Le principe de la méthode de Gauss-Jordan est le même que celui de la méthode de Gauss. On initialise une matrice B à l'identité, cette matrice va jouer le rôle du second membre de la méthode de Gauss.

La première étape de la méthode de Gauss-Jordan est la même que celle de la méthode de Gauss, de plus, on applique les mêmes transformations à la matrice B. Supposons que le résultat de la $(k-1)$-ème étape est :

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & & & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ & \ddots & & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k)} & \vdots & & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}, \; B^{(k)} = \begin{pmatrix} b_{11}^{(k)} & b_{12}^{(k)} & \cdots & b_{1n}^{(k)} \\ b_{21}^{(k)} & b_{22}^{(k)} & \cdots & b_{2n}^{(k)} \\ \vdots & \vdots & & \vdots \\ b_{n1}^{(k)} & b_{n2}^{(k)} & \cdots & b_{nn}^{(k)} \end{pmatrix}$$

On cherche un pivot non nul (par exemple par la stratégie du pivot partiel) :

$$a_{lk}^{(k)} \quad / \quad | \, a_{lk}^{(k)} \, | = \sup_{k \leq i \leq n} | \, a_{ik}^{(k)} \, | \, .$$

Si $k \neq l$, on échange la ligne k avec la ligne l dans les deux matrices $A^{(k)}$ et $B^{(k)}$. On notera encore par $A^{(k)}$ la matrice obtenue après permutation de la matrice $A^{(k)}$. Puis, on fait l'élimination dans les deux matrices $A^{(k)}$ et $B^{(k)}$ de la manière suivante :

$$\begin{aligned}&(\textit{la ligne } i) \ \textit{ est remplacée par}\\ &\qquad \{(\textit{la ligne } i) - \tfrac{x_i}{pivot}(\textit{la ligne de pivot})\}\\ &x_i = \text{ le coefficient à la position } (i,k) \text{ de la matrice } A^{(k)}\end{aligned}$$

avec i allant de 1 à $k-1$ et de $k+1$ à n. Au bout de n étapes, on obtient :

$$A^{(n)} = diag(a_{kk}^{(n)}) \qquad B^{(n)}$$

Pour finir, on divise les lignes de $k = 1, \cdots, n$, par $a_{kk}^{(n)}$, on obtient alors : à gauche la matrice identité et à droite la matrice A^{-1}.

1.5 Conditionnement d'un système linéaire

Soit A une matrice carrée d'ordre n inversible et b un vecteur de $\mathbb{R}^n$. On considère le système linéaire

$$Au = b,$$

de solution exacte et unique $u = A^{-1}b$.

Dans un premier cas, supposons que le second membre du système $Au = b$ est perturbé en $b + \delta b$ et A restant inchangée. Soit $u + \delta u$ la solution exacte du système perturbé

$$A(u + \delta u) = b + \delta b.$$

Comme $Au = b$, la relation précédente implique que $A\delta u = \delta b$, d'où

$$\delta u = A^{-1}\delta b.$$

Utilisant une norme subordonnée, nous obtenons une majoration de $\| \delta u \|$:

$$\| \delta u \| \leq \| A^{-1} \| \| \delta b \| .$$

La relation précédente montre que la norme de la perturbation de la solution exacte du système $Au = b$ due à une perturbation du second membre δb est au plus égale à $\| A^{-1} \|$ multipliée par $\| \delta b \|$. D'autre part, la relation $Au = b$ implique que

$$\| b \| \leq \| A \| \| u \| .$$

Combinons les deux inégalités précédentes, nous obtenons l'inégalité importante suivante :

$$\frac{\| \delta u \|}{\| u \|} \leq \| A \| \| A^{-1} \| \frac{\| \delta b \|}{\| b \|}.$$

Par conséquent, le rapport d'amplification des erreurs relatives est au plus égal à $\| A \| \| A^{-1} \|$.

EXEMPLE

Soit A la matrice

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix},$$

elle est symétrique, son déterminant vaut 1 et sa matrice inverse

$$A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}.$$

On considère le système linéaire

$$Au = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

et on considère le système perturbé, où le second membre est légèrement modifié, la matrice A restant inchangée :

$$A(u + \delta u) = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$$

ce qui donne un rapport d'amplification des erreurs relatives de l'ordre de 2000 (avec $\| \, . \, \|_1$). Théoriquement, on obtient avec la même norme un maximum pour le rapport d'amplification égale à

$$\| \, A \, \|_1 \| \, A^{-1} \, \|_1 = 136 \times 33 = 4488.$$

Maintenant, nous perturbons la matrice A et nous laissons b fixe. Soit $u + \Delta u$ la solution exacte du système perturbé :

$$(A + \Delta A)(u + \Delta u) = b.$$

On suppose que ΔA est assez petit pour que la matrice perturbée reste inversible. De la même manière que le premier cas on montre l'inégalité suivante :

$$\frac{\| \, \Delta u \, \|}{\| \, u + \Delta u \, \|} \leq \| \, A \, \| \| \, A^{-1} \, \| \, \frac{\| \, \Delta A \, \|}{\| \, A \, \|}.$$

La quantité $\| \, A \, \| \| \, A^{-1} \, \|$ apparait à nouveau pour contrôler l'amplification des erreurs relatives.

EXEMPLE

Considérons le même système que l'exemple précédent où, cette fois on perturbe la matrice A

$$\begin{pmatrix} 10 & 7 & 8,1 & 7,2 \\ 7,08 & 5,04 & 6 & 5 \\ 8 & 5,98 & 9,89 & 9 \\ 6,99 & 4,99 & 9 & 9,98 \end{pmatrix} (u + \Delta u) = b, \text{ de solution } \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

Nous définissons le nombre conditionnement de la matrice inversible A par :

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

Pour n'importe quelle norme matricielle subordonnée, la norme de la matrice identité I est égale à un. D'autre part, $I = AA^{-1}$ et $\| AA^{-1} \| \leq \| A \| \| A^{-1} \|$, ce qui montre que $cond(A) \geq 1$. On dit qu'un système est bien conditionné si le conditionnement de sa matrice est de l'ordre de 1 et il est mal conditionné si le conditionnement de sa matrice est très grand par rapport à 1.

1.6 Exercices : chapitre 1

Exercice 1

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Montrer que si $rang(A) = 1$ alors il existe u, $v \in \mathbb{R}^n$ tels que $A = uv^T$.
2. On considère la matrice élémentaire

$$E = I_n - \alpha\, uv^T$$

   (a) Montrer que E est inversible si et seulement si $\alpha\, u^T v \neq 1$.

   (b) On suppose que $\alpha\, u^T v \neq 1$, montrer que

$$E^{-1} = I_n - \beta\, uv^T \qquad \text{où} \quad \beta = \frac{\alpha}{\alpha\, u^T v - 1}.$$

   (c) Déterminer les valeurs propres de E.

   (d) En déduire que si $M \in \mathcal{M}_n(\mathbb{R})$ est inversible et $u \in \mathbb{R}^n$, alors

$$\det(M + uu^T) = (1 + u^T M^{-1} u)\ \det M.$$

Exercice 2

Dans $\mathcal{M}_n(\mathbb{R})$, on considère la matrice tridiagonale suivante :

$$A(a,b) = \begin{pmatrix} a & b & 0 & \cdots & \cdots & 0 \\ b & \ddots & \ddots & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & \ddots & b \\ 0 & \cdots & \cdots & 0 & b & a \end{pmatrix}, \quad a, b \in \mathbb{R}$$

Jusqu'à la troisième question, on supposera que $a = 0$ et $b = 1$.

1. Vérifier que : $u_k^T = (\sin(\frac{k\pi}{n+1}), \sin(\frac{2k\pi}{n+1}), \cdots, \sin(\frac{nk\pi}{n+1})), k = 1, \cdots, n$ sont des vecteurs propres de $A(0,1)$. Déduire le $sp(A(0,1))$.
2. En déduire le spectre de $A(a,b)$ et des vecteurs propres associés.

Exercice 3

Etant donné une norme vectorielle $\|.\|$, la norme matricielle subordonnée associée est définie par :

$$\|A\| = \sup_{u \neq 0} \frac{\|Au\|}{\|u\|}$$

(Les normes à droite sont vectorielles).

Vérifier que

1. $\|A\| = \sup_{\|u\|\leq 1} \|Au\| = \sup_{\|u\|=1} \|Au\|$
2. $\|A\| = \inf\{k > 0; \|Au\| \leq k\|u\|, \forall u\}$

Exercice 4

Soit $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$. Montrer les propriétés suivantes

1. La norme $\|.\|_2$ est invariante par transformation unitaire.
2. Si A est une matrice normale alors $\|A\|_2 = \rho(A)$.

Exercice 5

On munit $\mathbb{R}^n$ de la norme euclidienne notée $\|.\|_2$. On munit $\mathcal{M}_n(\mathbb{R})$ de la norme induite (notée aussi $\|.\|_2$) et le conditionnement associé est noté $\text{cond}_2(A)$.

Soit A une matrice inversible de $\mathcal{M}_n(\mathbb{R})$.

1. Montrer que si A est normale alors

$$cond_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}$$

où les nombres $\lambda_i(A)$ désignent les valeurs propres de A.

2. Montrer que $\text{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$, $\alpha \in \mathbb{R}^*$ et Q est une matrice orthogonale.

Exercice 6

Soit $A \in \mathcal{M}_n(\mathbb{R})$ la matrice définie par :

$$A = \begin{pmatrix} 1 & 2 & & & \\ & 1 & 2 & & \\ & & \ddots & \ddots & \\ & & & 1 & 2 \\ & & & & 1 \end{pmatrix}$$

1. Soit $x = (x_1 \cdots x_n)^T \in \mathbb{R}^n$ la solution du système linéaire $Ax = b$ où $b = (b_1 \cdots b_n)^T$ est un vecteur donné de $\mathbb{R}^n$.
   (a) Montrer que $x_k = \sum_{i=0}^{n-k} (-2)^i b_{k+i}, \quad k = 1, \ldots, n.$
   (b) En déduire A^{-1} et un vecteur $v \neq 0$ tel que $\|A^{-1}\|_\infty = \frac{\|A^{-1}v\|_\infty}{\|v\|_\infty}$.
   (c) Calculer $\text{cond}_\infty(A)$ et $\text{cond}_1(A)$.
2. Soit $(e_1, \cdots, e_n)$ la base canonique de $\mathbb{R}^n$.
   (a) Calculer $\|A^{-1}e_n\|_2$. En déduire un minorant de $\|A^{-1}\|_2$.
   (b) Calculer $\|Ae_2\|_2$. En déduire que $\text{cond}_2(A) > 2^n$.

Exercice 7

Résoudre, par la méthode de Gauss, le système linéaire $AX = b$ avec :

$$A = \begin{pmatrix} 2 & 1 & 0 & 4 \\ -4 & -2 & 3 & -7 \\ 4 & 1 & -2 & 8 \\ 0 & -3 & -12 & -1 \end{pmatrix}; \quad b = \begin{pmatrix} 2 \\ -9 \\ 2 \\ 2 \end{pmatrix}$$

Exercice 8 (Algorithme de la méthode de Gauss)

1. Ecrire un algorithme d'échange de deux vecteurs de $\mathbb{R}^n$.
2. Ecrire un algorithme de résolution de $Ax = b$ dans le cas où la matrice A est triangulaire.

3. Ecrire l'algorithme de la méthode de Gauss pour la résolution de $Ax = b$ (sans stratégie de pivot) ; puis avec pivot partiel.

Exercice 9 (Méthode de Gauss-Jordan)
Appliquer la méthode de Gauss-Jordan pour calculer l'inverse de la matrice

$$A = \begin{pmatrix} 10 & 1 & 4 & 0 \\ 1 & 10 & 5 & -1 \\ 4 & 5 & 10 & 7 \\ 0 & -1 & 7 & 9 \end{pmatrix}$$

1.7 Corrigé des exercices : chapitre 1

Réponse 1

1. Si $rang(A) = 1$, alors, il existe $u \in \mathbb{R}^n$, $u \neq 0$, tel que :

$$\forall x \in \mathbb{R}^n, \exists \alpha_x \in \mathbb{R} / Ax = \alpha_x u$$

en particulier :

$$\forall i \in \{1, \cdots, n\}, \exists v_i \in \mathbb{R} / Ae_i = v_i u$$

où $(e_i)_{i=1,n}$ est la base canonique de $\mathbb{R}^n$. D'où, si on écrit $x = \sum_{i=0}^{n} x_i e_i$, on obtient :

$$\begin{array}{l} Ax = \sum_{i=0}^{n} x_i Ae_i \\ = (\sum_{i=0}^{n} v_i x_i) u \\ = (v^T x) u \\ = (uv^T) x \end{array}$$

ce qui prouve que $A = uv^T$.

2. On remarque que pour $z \in \mathbb{R}^n$:

$$Ez = 0 \Longrightarrow z - \alpha(v^T z)u = 0 \Longrightarrow z \in < u >$$

d'où : $KerE \subset < u >=$ le sous-espace de $\mathbb{R}^n$ engendré par u. On supposera que $u \neq 0$ et $v \neq 0$, sinon $E = I_n$.

   (a) E inveresible $\Longleftrightarrow ker E = \{0\} \Longleftrightarrow Eu \neq 0$ (car $ker E \subset < u >$)
   $\Longleftrightarrow Eu = (1 - \alpha(v^T u))u \neq 0 \Longleftrightarrow 1 - \alpha(v^T u) \neq 0$.

   (b)

$$\begin{array}{rl} \forall x \in \mathbb{R}^n, EE^{-1}x & = E(x - \beta(v^T x)u) \\ & = Ex - \beta(v^T x)Eu \\ & = x - \alpha(v^T x)u - \beta(v^T x)(u - \alpha(v^T u)u) \\ & = x - \alpha(v^T x)u - (v^T x)\beta(1 - \alpha(v^T u))u \\ & = x - \alpha(v^T x)u + \alpha(v^T x)u \\ & = x \end{array}$$

   (c) Soit $V =< v >^{\perp}$=le sous-espace $\mathbb{R}^n$ orthogonal au sous-espace engendré par v. On a :

$$\forall x \in V, Ex = x$$

ce qui prouve que 1 est une valeur propre de multiplicité au moins $n - 1 = \dim V$. Si $\alpha(v^T u) \neq 0$, alors, la nième valeur propre est égale à $\lambda_n = 1 - \alpha(v^T u) \neq 1$, car $Eu = (1 - \alpha(v^T u))u$. Si $\alpha(v^T u) = 0$, alors :

$$Ex = \lambda_n x \Longrightarrow x - \alpha(v^T x)u = \lambda_n x \Longrightarrow (1 - \lambda_n)x = \alpha(v^T x)u$$

donc $\lambda_n = 1$ ou x est colinéaire à u. Comme $Eu = u$, alors $\lambda_n = 1$. Dans les deux cas $\lambda_n = 1 - \alpha(v^T u)$. D'où :

$$sp(E) = \{1, 1 - \alpha(v^T u)\}$$

(d) On écrit $M + uu^T = M(I_n - (-1)(M^{-1}u)u^T) = ME$, avec $\alpha = -1$ et $v = M^{-1}u$. On utilise les propriétés du déterminant :

$$\det(ME) = \det(M)\det(E),$$
$$\det(E) = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_n$$

où λ_i , $i = 1, n$, sont les valeurs propres de E. On obtient ce qu'il faut.

Réponse 2

1. Pour tout $k = 1, \cdots, n$, il faut vérifier que

$$u_k = (\sin(\frac{k\pi}{n+1}), \sin(\frac{2k\pi}{n+1}), \cdots, \sin(\frac{nk\pi}{n+1}))^T$$

est un vecteur propre de $A(0,1)$. Donc, pour tout $j = 1, \cdots, n$, il faut calculer :

$$x_{j+1} + x_{j-1}$$

avec $x_j = \sin(\frac{jk\pi}{n+1})$.
En utilisant la relation trigonométrique :

$$\sin(a) + \sin(b) = 2\cos(\frac{a-b}{2})\sin(\frac{a+b}{2})$$

on obtient pour tout $j = 1, \cdots, n$:

$$\begin{aligned} x_{j+1} + x_{j-1} &= \sin(\tfrac{(j+1)k\pi}{n+1}) + \sin(\tfrac{(j-1)k\pi}{n+1}) \\ &= \sin(\tfrac{(j+1)k\pi}{n+1}) + \sin(\tfrac{(j-1)k\pi}{n+1}) \\ &= 2\cos(\tfrac{k\pi}{n+1})\sin(\tfrac{jk\pi}{n+1}) \\ &= 2\cos(\tfrac{k\pi}{n+1})x_j \end{aligned}$$

d'où, u_k est un vecteur propre associé à la valeur propre $\lambda_k = 2\cos(\frac{k\pi}{n+1})$.

2. On a $A(a,b) = aI_n + bA(0,1)$. Si $b = 0$, $A(a,0) = aI_n$ ce qui donne $sp(A(a,0)) = \{a\}$. Supposons que $b \neq 0$, alors :
$\star$ $\lambda \in sp(A(a,b)) \Longrightarrow \exists v \neq 0 / A(a,b)v = \lambda v$
$\Longrightarrow A(0,1)v = \frac{\lambda - a}{b} v \Longrightarrow \frac{\lambda - a}{b} \in sp(A(0,1))$
$\star$ $\lambda \in sp(A(0,1)) \Longrightarrow \exists v \neq 0 / A(0,1)v = \lambda v$
$\Longrightarrow A(a,b)v = aI_n v + bA(0,1)v = (a + b\lambda)v \Longrightarrow a + b\lambda \in sp(A(a,b))$
D'où, $sp(A(a,b)) = \{a + 2b\cos(\frac{k\pi}{n+1}), k = 1, \cdots, n\}$.

Réponse 3 Pour tout x , on a :$\|Ax\| \leq \|A\|\|x\|$, d'où :

1.

$$\begin{aligned} \|A\| &= \sup_{x \neq 0} \tfrac{\|Ax\|}{\|x\|} = \sup_{x \neq 0} \|A(\tfrac{x}{\|x\|})\| \\ &\leq \sup_{\|u\|=1} \|Au\| \leq \sup_{\|u\| \leq 1} \|Au\| \\ &\leq \sup_{\|u\| \leq 1} \|A\|\|u\| = \|A\| \end{aligned}$$

2.

$$\inf\{k > 0 / \forall x, \|Ax\| \leq k\|x\|\} \leq \|A\|$$

Soit $k \in]0, +\infty[$ tel que pour tout x, on a : $\|Ax\| \leq k\|x\|$ alors :

$$\forall x \neq 0, \frac{\|Ax\|}{\|x\|} \leq k$$

ce qui donne que : $\|A\| \leq k$. Donc :

$$\|A\| \leq \inf\{k > 0 / \forall x, \|Ax\| \leq k\|x\|\}$$

Réponse 4

1. Soient U et V deux matrices unitaires ($U^*U = UU^* = V^*V = VV^* = I_n$), alors :
$$\begin{aligned}\|U^*AV\|_2 &= \sqrt{\rho((U^*AV)^*(U^*AV)}\\ &= \sqrt{\rho(V^*A^*UU^*AV)}\\ &= \sqrt{\rho(V^*(A^*A)V)}\\ &= \sqrt{\rho(A^*A)} = \|A\|_2\end{aligned}$$
car $V^*(A^*A)V$ et (A^*A) sont semblables et par conséquent elles ont les mêmes valeurs propres.
2. Si A est une matrice normale, alors A est diagonalisable : $D = diag(\lambda_i) = U^*AU$ avec U une matrice unitaire. D'où $D^*D = diag(|\lambda_i|^2) = U^*A^*AU$, ce qui prouve que :
$$\begin{aligned}\|A\|_2 &= \sqrt{\rho(A^*A)} = \sqrt{\rho(U^*A^*AU)}\\ &= \sqrt{\max_i |\lambda_i|^2} = \sqrt{(\max_i |\lambda_i|)^2}\\ &= \max_i |\lambda_i| = \rho(A)\end{aligned}$$

Réponse 5

1. Si A est une matrice normale, alors A est diagonalisable :
$$D = diag(\lambda_i) = U^*AU \text{ et } \|A\|_2 = \max_i |\lambda_i|$$
D'où
$$D^{-1} = diag(\frac{1}{\lambda_i}) = U^{-1}A^{-1}(U^*)^{-1} = U^*A^{-1}U$$
($U^{-1} = U^*$ et $(U^*)^{-1} = U$ car $U^*U = UU^* = I$), ce qui prouve :
$$\|A^{-1}\|_2 = \rho(A^{-1}) = \max_i \frac{1}{|\lambda_i|} = \frac{1}{\min_i |\lambda_i|}$$
car A^{-1} est aussi normale (pour le prouver, on utilise :
$(A^*A)^{-1} = A^{-1}(A^*)^{-1} = A^{-1}(A^{-1})^*$).
2. On a : $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}$ (car $sp(AB) = sp(BA)$). D'où :
$$\begin{aligned}cond_2(A) &= \|A\|_2\|A^{-1}\|_2\\ &= \sqrt{\rho(A^*A)\rho((A^{-1})^*A^{-1})}\\ &= \sqrt{\rho(A^*A)\rho((A^*)^{-1}A^{-1})}\\ &= \sqrt{\rho(A^*A)\rho(A^{-1}(A^*)^{-1})}\\ &= \sqrt{\rho(A^*A)\rho((A^*A)^{-1})}\end{aligned}$$
or $\rho(B) = \max_i |\lambda_i(B)|$ et $\rho(B^{-1}) = \max_i |\lambda_i(B^{-1})| = \max_i \frac{1}{|\lambda_i(B)|} = \frac{1}{\min_i |\lambda_i(B)|}$, d'où :
$$1 = cond_2(A) = \sqrt{\frac{\max_i |\lambda_i(A^*A)|}{\min_i |\lambda_i(A^*A)|}}$$
ce qui donne $\max_i |\lambda_i(A^*A)|$=$\min_i |\lambda_i(A^*A)|$, or, la matrice A^*A est hermitienne définie positive ($x^*A^*Ax = (Ax)^*(Ax) = \|Ax\|_2^2 > 0$, pour $x \neq 0$), donc ses valeurs propres sont strictement positives, ce qui donne :
$$\lambda_1(A^*A) = \lambda_2(A^*A) = \cdots = \lambda_n(A^*A) = r > 0$$
Par conséquent : $A^*A = U^*(rI_n)U = rI_n$, d'où : $(\frac{1}{\sqrt{r}}A)^*(\frac{1}{\sqrt{r}}A) = I_n$. Ce qui prouve que $\frac{1}{\sqrt{r}}A = Q$ est une matrice unitaire. Comme A est réelle, alors, Q est réelle, d'où, Q est orthogonale.

Réponse 6

1. (a) La k-ème équation du système $Ax = b$, $k = 1, \cdots, n-1$, s'écrit :

$$x_k + 2x_{k+1} = b_k$$

et $x_n = b_n$. Par récurrence lorsque k décroît de n à 1. Pour $k = n$: $x_n = \sum_{i=0}^{0}(-2)^i b_{n+i} = b_n$. Supposons que :

$$x_{k+1} = \sum_{i=0}^{n-k-1} (-2)^i b_{k+1+i}$$

alors :

$$\begin{array}{ll} x_k = b_k - 2x_{k+1} & = b_k + \sum_{i=0}^{n-k-1}(-2)^{i+1} b_{k+i+1} \\ & = b_k + \sum_{j=1}^{n-k}(-2)^j b_{k+j} \\ & = \sum_{j=0}^{n-k}(-2)^j b_{k+j} \end{array}$$

(b) Si on note par $A^{-1} = (\beta_{ij})$, alors, la l-ième colonne $(\beta_{1l}\ \beta_{2l}\ \cdots\ \beta_{nl})^T$ de A^{-1} est égale à $A^{-1}e_l$, donc c'est la solution du système : $Ax = e_l$. D'où, en remplaçant le vecteur b par le vecteur e_l dans la question précédente, on obtient :

$$x_k = \beta_{kl} = \sum_{i=0}^{n-k}(-2)^i b_{k+i} = \left\{ \begin{array}{ll} (-2)^{l-k} & k \leq l \\ 0 & k > l \end{array} \right.$$

Ce qui prouve que :

$$\begin{array}{ll} \|A^{-1}\|_\infty & = \max_k \sum_{l=1}^{n} |\beta_{kl}| = \max_k \sum_{l=1}^{n} 2^{l-k} \\ & = (\sum_{l=1}^{n} 2^l) \max_k 2^{-k} = \frac{1}{2}\frac{2-2^{n+1}}{1-2} \\ & = 2^n - 1 \end{array}$$

On prend $v = (-1, 1, -1, \cdots, (-1)^n)$, on a $\|v\|_\infty = 1$ et :

$$\begin{array}{ll} \|A^{-1}v\|_\infty & = \max_k |\sum_{j=1}^{n}(-1)^j \beta_{kj}| \\ & = \max_k |\sum_{j=1}^{n}(-1)^j (-2)^{j-k}| \\ & = \max_k 2^{-k} |\sum_{j=1}^{n}(-1)^j (-2)^j| \\ & = \frac{1}{2}\sum_{j=1}^{n} 2^j = 2^n - 1 = \|A^{-1}\|_\infty \end{array}$$

(c) $cond_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 3(2^n - 1)$ car $\|A\|_\infty = 1 + 2 = 3$.

$$\begin{array}{ll} cond_1(A) & = \|A\|_1 \|A^{-1}\|_1 \\ & = \|A^*\|_\infty \|(A^{-1})^*\|_\infty \\ & = 3 \max_{k=1,n} \sum_{l=1}^{n} |(-2)^{k-l}| \\ & = 3(\sum_{l=1}^{n} 2^{-l}) \max_{k=1,n} 2^k \\ & = 3(\frac{2^{-1} - 2^{-n-1}}{1-2^{-1}})2^n \\ & = 3(2^n - 1) \end{array}$$

2. (a) $\|A^{-1}e_n\|_2 = \|\sum_{k=1}^{n} \beta_{kn} e_k\|_2 = \sqrt{\sum_k \beta_{kn}^2} = \sqrt{\sum_k 4^{n-k}} = 2^n \sqrt{\sum_k 4^{-k}} \geq 2^{n-1}$

(b) $\|Ae_2\|_2 = \|2e_1 + e_2\|_2 = \sqrt{5} > 2$. D'où :
$cond_2(A) = \|A\|_2 \|A^{-1}\|_2 \geq \|Ae_2\|_2 \|A^{-1}e_n\|_2 > 2.2^{n-1} = 2^n$.

Réponse 7 Utiliser l'exemple interactif du chapitre 1.

Réponse 8

1. La procédure d'échange de deux vecteurs : $u = (u_1 u_2 u_3 \cdots u_n)^T$ et $v = (v_1 v_2 v_3 \cdots v_n)^T$

Début de la procédure echang(u,v)
Pour i allant de 1 à n faire
$x = u_i$
$u_i = v_i$
$v_i = x$
Fin faire i
Fin procédure

2. La procédure de remontée pour résoudre un système linéaire triangulaire : $A = (a_{ij})$ et $b = (b_i)$

Début de la procédure remonte(A,b,x)
$x_n = b_n$
Pour k allant de n-1 à 1 faire
$x_k = b_k$
Pour i allant de k+1 à n faire
$x_k = x_k - a_{ki} \times x_i$
Fin faire i
Fin faire k
Fin procédure

3. a) Le programme de la méthode de Gauss sans stratégie de pivot (dans ce cas
on suppose qu'à chaque étape k de la méthode d'élimination de Gauss
le coefficient à la position (k,k) est non nul) :

Début du programme Gauss1
Pour i allant de 1 à n faire
lire b_i
Pour j allant de 1 à n faire
lire a_{ij}
Fin faire j
Fin faire i
Pour k allant de 1 à n-1 faire
appel procédure : elimin(A,b,k)
Fin faire k
Appel procédure : remonte(A,b,x)
Pour i allant de 1 à n faire
affiche x_i
Fin faire i
Début de la procédure remonte(A,b,x)
$x_n = b_n$
Pour k allant de n-1 à 1 faire
$x_k = b_k$
Pour i allant de k+1 à n faire
$x_k = x_k - a_{ki} \times x_i$
Fin faire i
Fin fair e k
Fin procédure
Début de la procédure elimin(A,b,k)
$pivot = a_{kk}$
Pour i allant de k+1 à n faire
Pour j allant de k+1 à n faire
$a_{ij} = a_{ij} - a_{ik} \times \frac{a_{kj}}{pivot}$
Fin faire j
Fin faire i
Fin procédure
Fin programme Gauss1

b) Le programme de la méthode de Gauss avec pivot est le même que le programme Gauss1, sauf que juste avant de faire appel à la procédure d'élimination de Gauss : elimin(A,b,k), on fait appel à la procédure : pivot(A,b,k) qui recherche le pivot de l'étape k et permute la ligne de pivot avec la ligne k et il faut ajouter la procédure pivot(A,b,k) au programme.

Début de la procédure pivot(A,b,k)
    $pivot = |a_{kk}|$
    $l = k$
    Pour i allant de k+1 à n faire
        Si $|a_{ik}| > pivot$ alors
            $pivot = |a_{ik}|$
            $l = i$
        Fin si
    Fin faire i
    Si $pivot = 0$ alors
        affiche : A non inversible
        quitter le programme
    Fin si
    Si $l \neq k$ alors
        Pour j allant de k+1 à n
            $x = a_{kj}$
            $a_{kj} = a_{lj}$
            $a_{lj} = x$
        Fin faire j
        $x = b_k$
        $b_k = b_l$
        $b_l = x$
    Fin si
Fin procédure

Réponse 9 Utiliser l'exemple interactif du chapitre 1.

Chapitre 2

Méthodes itératives pour la résolution d'un système linéaire

Les méthodes d'élimination ou de factorisation sont utilisées surtout lorsque l'ordre de la matrice est petit (matrice 100×100 par exemple) ou lorsque la matrice est pleine (i.e. peu de coefficients nuls).

Dans la pratique, beaucoup de problèmes nécessitent la résolution d'un système $Ax = b$ d'ordre assez important, avec, A une matrice creuse (i.e. beaucoup de coefficients nuls). Des systèmes de ce type sont donnés par exemple par l'application de la méthode des différences finies ou la méthode des éléments finis.

Pour ce genre de problèmes, on utilise les méthodes itératives. Etant donné un vecteur initial arbitraire x^0, on construit une suite de vecteurs

$$x^0, x^1, \cdots, x^k, \cdots$$

qui converge vers la solution x du système linéaire $Ax = b$.

2.1 Construction d'une méthode itérative

On considère le système linéaire

$$Ax = b \tag{1.1}$$

avec A une matrice carrée d'ordre n inversible et b un vecteur de $\mathbb{R}^n$. Pour toute matrice M carrée d'ordre n inversible, le système (1.1) est équivalent à

$$Mx - (M - A)x = b \tag{1.2}$$

ou encore, en posant $N = M - A$, $B = M^{-1}N$ et $c = M^{-1}b$

$$x = Bx + c. \tag{1.3}$$

Ce qui nous permet de définir la méthode itérative :

$$\begin{cases} x^0 \in \mathbb{R}^n, \ \textit{vecteur initial} \\ x^{k+1} = Bx^k + c \end{cases} . \tag{1.4}$$

Soit x la solution de (1.1), si on note $e^k = x^k - x$ le $k$ième vecteur erreur, on obtient

$$\begin{array}{ll} e^k & = x^k - x = (Bx^{k-1} + c) - (Bx + c) = B(x^{k-1} - x) \\ & = Be^{k-1} = B^k e^0 \end{array} \tag{1.5}$$

On dit que la méthode itérative (1.4) converge si la suite de vecteurs (e^k) converge vers zéro indépendamment du vecteur initial x^0, ce qui est équivalent, d'après le théorème 2.3 du chapitre précédent, à l'une des deux propositions équivalentes :

(1) $\rho(B) < 1$

(2) $\| B \| < 1$ pour au moins une norme matricielle.

REMARQUE 2.1.1 *Pour chaque choix de la matrice inversible M, on obtient une méthode itérative. Le meilleur choix de M doit satisfaire les conditions suivantes :*

(a) *la matrice M est facile à inverser,*

(b) $\rho(B) = \rho(M^{-1}N)$ *est le plus petit possible.*

Dans la condition (a), on n'a pas besoin de l'inverse de M, il suffit que le calcul de x^{k+1} en fonction de x^k, en utilisant (1.2) :

$$Mx^{k+1} = Nx^k + b,$$

soit facile. La condition (b) est une conséquence du comportement asymptotique de l'erreur e^k ; en effet,

$$\begin{array}{ll} \lim_{k \longrightarrow \infty} \left[\max_{e^0 \neq 0} \| e^k \| / \| e^0 \|\right]^{1/k} & = \lim_{k \longrightarrow \infty} \left[\max_{e^0 \neq 0} \| B^k e^0 \| / \| e^0 \|\right]^{1/k} \\ & = \lim_{k \longrightarrow \infty} \| B^k \|^{1/k} = \rho(B) \end{array}$$

la méthode est donc d'autant plus rapide que $\rho(B)$ est plus petit (voir théorème 1.6 du chapitre précédent).

On considère la décomposition suivante :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix} = D - E - F,$$

$$D = diag(a_{11}, a_{22}, \cdots, a_{nn})$$

$$E = -\begin{pmatrix} 0 & & & \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{pmatrix}$$

$$F = -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{pmatrix}.$$

On suppose dans toute la suite que D est inversible et on pose

$$L = D^{-1}E, \quad U = D^{-1}F,$$

(1) *la méthode de Jacobi*

$$M = D, \ N = E + F, \quad J = M^{-1}N = L + U$$

le calcul de x^{k+1} à partir de x^k se fait directement par le système

$$\begin{cases} a_{11}x_1^{k+1} = & -a_{12}x_2^k & -a_{13}x_3^k & \cdots & -a_{1n}x_n^k & +b_1 \\ a_{22}x_2^{k+1} = -a_{21}x_1^k & & -a_{23}x_3^k & \cdots & -a_{2n}x_n^k & +b_2 \\ & \vdots & & & & \\ a_{nn}x_n^{k+1} = -a_{n1}x_1^k & -a_{n2}x_2^k & \cdots & -a_{nn-1}x_{n-1}^k & & +b_n \end{cases}$$

On remarque que la méthode de Jacobi nécessite n mémoires pour le vecteur x^k et n mémoires pour le vecteur x^{k+1}. La matrice J s'appelle la matrice de Jacobi.

(2) *la méthode de Gauss-Seidel*

$$M = D - E, \quad N = F, \quad H = M^{-1}N = (I - L)^{-1}U$$

le calcul de x^{k+1} à partir de x^k se fait directement par le système

$$\begin{cases} a_{11}x_1^{k+1} = & -a_{12}x_2^k & -a_{13}x_3^k & \cdots & -a_{1n}x_n^k & +b_1 \\ a_{22}x_2^{k+1} = -a_{21}x_1^{k+1} & & -a_{23}x_3^k & \cdots & -a_{2n}x_n^k & +b_2 \\ & \vdots & & & & \\ a_{nn}x_n^{k+1} = -a_{n1}x_1^{k+1} & -a_{n2}x_2^{k+1} & \cdots & -a_{n,n-1}x_{n-1}^{k+1} & & +b_n \end{cases}$$

la méthode de Gauss-Seidel nécessite seulement n mémoires, la composante x_i^{k+1} prend la place de x_i^k qui ne sera pas utilisée pour le calcul de $x_{i+1}^{k+1}, x_{i+2}^{k+1}, \cdots, x_n^{k+1}$.

(3) *la méthode de relaxation*

$$\omega \neq 0, \quad \begin{cases} M & = & \dfrac{D}{\omega} - E, \quad N = \dfrac{1-\omega}{\omega}D + F, \\ H(\omega) & = & M^{-1}N = (I - \omega L)^{-1}[(1-\omega)I + \omega U], \end{cases}$$

le calcul de x^{k+1} à partir de x^k se fait directement par le système

$$\left\{\begin{array}{llllllll} a_{11}x_1^{k+1} & = & a_{11}x_1^k & -\omega\{ & a_{11}x_1^k & +a_{12}x_2^k & +\cdots & +a_{1n}x_n^k - b_1\} \\ a_{22}x_2^{k+1} & = & a_{22}x_2^k & -\omega\{ & a_{21}x_1^{k+1} & +a_{22}x_2^k & +\cdots & +a_{2n}x_n^k - b_2\} \\ & \vdots & & & & & & \\ a_{nn}x_n^{k+1} & = & a_{nn}x_n^k & -\omega\{ & a_{n1}x_1^{k+1} & +\cdots & +a_{n,n-1}x_{n-1}^{k+1} & +a_{nn}x_n^k - b_n\} \end{array}\right.$$

la méthode de relaxation nécessite seulement n mémoires. Le paramètre ω s'appelle le paramètre de relaxation et pour $\omega = 1$ on retrouve la méthode de Gauss-Seidel.

Exemple :

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

La matrice de Jacobi J et la matrice de Gauss-Seidel H sont données par :

$$J = \frac{1}{2}\begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ 2 & -2 & 0 \end{pmatrix}, \quad H = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ -1 & 1 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

Le système qui donne la suite de la méthode de Jacobi est donné par :

$$\left\{\begin{array}{ll} 2x_1^{k+1} & = -x_2^k - x_3^k + 1 \\ 2x_2^{k+1} & = -x_1^k - x_3^k \\ x_3^{k+1} & = x_1^k - x_2^k \end{array}\right.$$

Le système qui donne la suite de la méthode de Gauss-Seidel est donné par :

$$\left\{\begin{array}{ll} 2x_1^{k+1} & = -x_2^k - x_3^k + 1 \\ 2x_2^{k+1} & = -x_1^{k+1} - x_3^k \\ x_3^{k+1} & = x_1^{k+1} - x_2^{k+1} \end{array}\right.$$

Pour les deux méthodes, on démarre de $x^0 = (0\ 0\ 0)^T$. A chaque itération k de la méthode de Jacobi, le vecteur $x = (x_1\ x_2\ x_3)^T$ joue le rôle de x^k et le vecteur $y = (y_1\ y_2\ y_3)^T$ joue le rôle de x^{k+1}. Par exemple, si on veut calculer x^3 :
$k = 0 : x = (0\ 0\ 0)^T$ et y est donné par :

$$\left\{\begin{array}{lll} y_1 & = (-x_2 - x_3 + 1)/2 & = 1/2 \\ y_2 & = (-x_1 - x_3)/2 & = 0 \\ y_3 & = x_1 - x_2 & = 0 \end{array}\right.$$

$k = 1 : x = (1/2\ 0\ 0)^T$ et y est donné par :

$$\left\{\begin{array}{lll} y_1 & = (-x_2 - x_3 + 1)/2 & = 1/2 \\ y_2 & = (-x_1 - x_3)/2 & = -1/4 \\ y_3 & = x_1 - x_2 & = 1/2 \end{array}\right.$$

$k = 2 : x = (1/2\ \ -1/4\ 1/2)^T$ et y est donné par :

$$\left\{\begin{array}{lll} y_1 & = (-x_2 - x_3 + 1)/2 & = (1/4 - 1/2 + 1)/2 = 3/8 \\ y_2 & = (-x_1 - x_3)/2 & = (-1/2 - 1/2)/2 = -1/2 \\ y_3 & = x_1 - x_2 & = 1/2 + 1/4 = 3/4 \end{array}\right.$$

Ce qui donne :

$$x^3 = \begin{pmatrix} 3/8 \\ -1/2 \\ 3/4 \end{pmatrix}$$

A chaque itération k de la méthode de Gauss-Seidel, le vecteur $x = (x_1\ x_2\ x_3)^T$ joue le rôle de x^k et au même temps le rôle de x^{k+1}. Par exemple, si on veut calculer x^2 :
$k = 0 : x = (0\ 0\ 0)^T$ et x est donné par :

$$\begin{cases} x_1 & = (-x_2 - x_3 + 1)/2 & = 1/2 \\ x_2 & = (-x_1 - x_3)/2 & = (-1/2 - 0)/2 = -1/4 \\ x_3 & = x_1 - x_2 & = 1/2 + 1/4 = 3/4 \end{cases}$$

$k = 1 : x = (1/2\ \ -1/4\ 3/4)^T$ et x est donné par :

$$\begin{cases} x_1 & = (-x_2 - x_3 + 1)/2 & = (1/4 - 3/4 + 1)/2 = 1/4 \\ x_2 & = (-x_1 - x_3)/2 & = (-1/2 - 3/4)/2 = -5/8 \\ x_3 & = x_1 - x_2 & = 1/4 + 5/8 = 7/8 \end{cases}$$

Ce qui donne :

$$x^2 = \begin{pmatrix} 1/4 \\ -5/8 \\ 7/8 \end{pmatrix}$$

On remarque que les itérations de la méthode de Jacobi nécessitent 6 variables : le vecteur x et le vecteur y, donc 6 mémoires et les itérations de la méthode de Gauss-Seidel nécessitent seulement 3 variables : le vecteur x, donc 3 mémoires.

2.2 Convergence

Dans le cas d'un système $Ax = b$, avec A une matrice hermitienne définie positive, le théorème suivant donne une condition suffisante pour qu'une méthode itérative associée à une décomposition $A = M - N$ converge.

THÉORÈME 2.2.1 *Soit A une matrice hermitienne définie positive, décomposée sous la forme*

$$A = M - N,\ M :\ \textit{matrice inversible.}$$

Pour que la méthode itérative associée à cette décomposition converge, c'est à dire

$$\rho(M^{-1}N) < 1$$

il suffit que la matrice hermitienne $M^* + N$ soit définie positive.

DÉMONSTRATION : Vérifions d'abord que $M^* + N$ est hermitienne :

$$M^* + N = A^* + N^* + N = A + N + N^* = M + N^*.$$

Soit $\lambda \in sp(M^{-1}N)$ et soit $v \neq 0$ tel que

$$M^{-1}Nv = \lambda v$$

On pose

$$u = M^{-1}Av = (I - M^{-1}N)v = v - \lambda v$$

soit encore

$$\lambda v = v - u$$

on a alors :

$$\begin{aligned} \mid \lambda \mid^2 v^*Av &= (v-u)^*A(v-u) \\ &= v^*Av - v^*Au - u^*Av + u^*Au \\ &= v^*Av - (Av)^*u - u^*Av + u^*Au \\ &= v^*Av - (Mu)^*u - u^*Mu + u^*Au \\ &= v^*Av - u^*(M^* + M - A)u \\ &= v^*Av - u^*(M^* + N)u \end{aligned}$$

d'où

$$(1- \mid \lambda \mid^2)v^*Av = u^*(M^* + N)u$$

comme $v \neq 0$, $u \neq 0$ et A, $M^* + N$ sont définies positives, on déduit que

$$\mid \lambda \mid < 1.$$

COROLLAIRE 2.2.1 *Soit A une matrice hermitienne définie positive, alors la méthode de relaxation converge pour $0 < \omega < 2$. En particulier, la méthode de Gauss-Seidel est convergente lorsque A est hermitienne définie positive.*

DÉMONSTRATION : La matrice A est hermitienne définie positive, d'où, les coefficients de la matrice diagonale D sont strictement positifs et $E^* = F$. D'autre part

$$M = \frac{D}{\omega} - E, \quad N = \frac{1-\omega}{\omega}D + F$$

d'où

$$M^* + N = \frac{2-\omega}{\omega}D$$

donc, pour $0 < \omega < 2$, la matrice hermitienne $M^* + N$ est définie positive.

THÉORÈME *(Kahan (1958)) Le rayon spectral de la matrice de relaxation*

$$H(\omega) = (I - \omega L)^{-1}\{\omega U + (1-\omega)I\}$$

vérifie pour tout $\omega \neq 0$:

$$\rho(H(\omega)) \geq \mid \omega - 1 \mid .$$

Ce qui montre, en particulier, que la méthode de relaxation ne converge pas pour $\omega \notin]0, 2[$.

DÉMONSTRATION : On a

$$\prod_i \lambda_i(H(\omega)) = det(H(\omega)) = \frac{det(\omega U + (1-\omega)I)}{det(I - \omega L)} = (1-\omega)^n$$

or

$$\mid \prod_i \lambda_i(H(\omega)) \mid \leq \rho(H(\omega))^n$$

d'où

$$\rho(H(\omega)) \geq \mid \omega - 1 \mid .$$

DÉFINITION 2.2.1 *Soit A une matrice carrée et $J = L + U$ la matrice de Jacobi associée à A. On pose $J(\alpha) = \alpha L + \alpha^{-1}U$. On dit que la matrice A est correctement ordonnée si $sp(J(\alpha))$ est indépendant de α. Autrement dit*

$$sp(J(\alpha)) = sp(J), \quad \forall \alpha \neq 0.$$

THÉORÈME *(Young (1950)) Soit A une matrice correctement ordonnée et $\omega \neq 0$, alors*
(a) $\mu \in sp(J) \Longrightarrow -\mu \in sp(J)$
(b) $\mu \in sp(J)$ *et*

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2, \tag{2.1}$$

alors $\lambda \in sp(H(\omega))$.
(c) $0 \neq \lambda \in sp(H(\omega)))$ *et μ vérifie* (2.1)*, alors $\mu \in sp(J)$.*

DÉMONSTRATION :
(a) $J(-1) = -L - U = -J$, d'où

$$sp(J) = sp(J(-1)) = sp(-J).$$

(b)-(c) Soit $\lambda \in \mathbb{C}^*$, on a :

$$\begin{aligned} det(\lambda I - H(\omega)) &= det\{(I - \omega L)(\lambda I - H(\omega))\} \\ &= det\{(\lambda + \omega - 1)I - \lambda\omega L - \omega U\} \end{aligned}$$

soient $\alpha, \beta \in \mathbb{C}$ tels que

$$\alpha^2 = \lambda\omega^2, \quad et \quad \beta = \alpha/\omega$$

alors,

$$det(\lambda I - H(\omega)) = \alpha^n det\{\frac{(\lambda + \omega - 1)}{\alpha} I - (\beta L + \beta^{-1}U)\}.$$

D'autre part, μ est une solution de (2.1) si et seulement si

$$\mu = \pm\frac{(\lambda + \omega - 1)}{\alpha}.$$

Donc

$$\begin{cases} \lambda \in sp(H(\omega)) \\ \mu = \pm\frac{(\lambda+\omega-1)}{\alpha} \end{cases} \iff \begin{cases} det\{\mu I - (\pm\beta L + (\pm\beta)^{-1}U)\} = 0 \\ \mu = \pm\frac{(\lambda+\omega-1)}{\alpha} \end{cases}$$
$$\iff \begin{cases} \mu \in sp(J(\pm\beta)) = sp(J) \\ \mu^2 = \frac{(\lambda+\omega-1)^2}{\lambda\omega^2} \end{cases}$$

COROLLAIRE 2.2.2 *Soit A une matrice correctement ordonnée, alors*

$$\rho(H) = (\rho(J))^2.$$

En particulier, si A est correctement ordonnée la méthode de Gauss-Seidel converge si et seulement si la méthode de Jacobi converge, et si les deux méthodes convergent et $\rho(J) \neq 0$, alors :

$$0 < \rho(H) < \rho(J) < 1.$$

DÉMONSTRATION : Pour $\omega = 1$, l'équation (2.1) devient :

$$\lambda^2 = \lambda\mu^2$$

et d'après les propriétés b) et c) du théorème de Young, on a :
(a) $\mu \in sp(J) \Longrightarrow \lambda = 0 \in sp(H)$ et $\lambda = \mu^2 \in sp(H)$
(b) $\lambda \in sp(H) \Longrightarrow \pm\mu \in sp(J)$ avec $\mu^2 = \lambda$
par conséquent, $\rho(J) = 0$ si et seulement si $\rho(H) = 0$ et si $\rho(J) \neq 0$ alors $\rho(H) = (\rho(J))^2$.

2.3 Détermination théorique du paramètre de relaxation optimal

Soit $A = D - E - F$ une matrice correctement ordonnée, avec D inversible. Nous supposons que la méthode de Jacobi est convergente. D'après le corollaire précédent la méthode de relaxation converge pour $\omega = 1$, et par continuité, il y a convergence sur un intervalle contenant 1. Le problème que nous voulons résoudre est de déterminer le paramètre de relaxation optimal ω_b tel que :

$$\rho(H(\omega_b)) = \min_{\omega} \rho(H(\omega)).$$

THÉORÈME 2.3.1 *Soit A une matrice correctement ordonnée, si toutes les valeurs propres de J sont réelles et $0 \leq \rho(J) < 1$, alors, la méthode de relaxation converge si seulement si $0 < \omega < 2$; de plus, on a :*

(1) $\omega_b = \dfrac{2}{1 + \sqrt{1 - \rho(J)^2}} = 1 + \left(\dfrac{\rho(J)}{1 + \sqrt{1 - \rho(J)^2}} \right)^2$
(2) $\rho(H(\omega_b)) = \omega_b - 1$
(3) $\rho(H(\omega)) > \rho(H(\omega_b)), \quad \forall \omega \neq \omega_b.$

DÉMONSTRATION : On peut limiter l'étude à $\omega \in]0, 2[$ (voir théorème de Kahan). D'après le théorème de Young, les valeurs propres non nulles de J sont des paires $\pm\mu_i$, avec :

$$0 \leq \mu_i \leq \rho(J)$$

et à chaque $\mu \in sp(J)$ correspond une paire de valeurs propres de $H(\omega)$

$$\{\lambda_m(\omega, \mu),\ \lambda_M(\omega, \mu)\}$$

solutions de (2.1) (on suppose que $\mid \lambda_m(\omega, \mu) \mid \leq \mid \lambda_M(\omega, \mu) \mid$), et à chaque valeur propre non nulle $\lambda \in H(\omega)$ correspond une valeur propre $0 \leq \mu \in sp(J)$ telle que

$$\lambda = \lambda_m(\omega, \mu) \quad ou \quad \lambda = \lambda_M(\omega, \mu).$$

D'où

$$\rho(H(\omega)) = \max_{\mu \in sp(J)} \mid \lambda_M(\omega, \mu) \mid$$

D'après l'équation (2.1)

$$(\omega - 1)^2 = \lambda_m(\omega, \mu)\lambda_M(\omega, \mu) \leq \mid \lambda_M(\omega, \mu) \mid^2$$

d'où

$$| \lambda_M(\omega, \mu) | \geq | \omega - 1 |$$

et on a :

$$| \lambda_m(\omega, \mu) | = | \lambda_M(\omega, \mu) | = | \omega - 1 |$$

dans le cas d'une racine double ou dans le cas de deux racines complexes conjuguées. D'après (2.1), si $\lambda \in sp(H(\omega))$ est réelle, alors, nécessairement $\lambda \geq 0$ et on a, pour $\omega \neq 0$:

$$\frac{\lambda + \omega - 1}{\omega} = \pm \mu \sqrt{\lambda}. \tag{3.1}$$

On définit

$$g_\omega(\lambda) = \frac{\lambda + \omega - 1}{\omega}, \quad \omega \neq 0$$

et

$$m_\mu(\lambda) = \mu \sqrt{\lambda}, \quad 0 \leq \mu \leq \rho(J) < 1.$$

Le graphe de g_ω est la droite qui passe par $(1, 1)$ et de pente $\frac{1}{\omega}$. Géométriquement, les solutions réelles de (2.1) sont données par l'intersection des deux courbes associées à g_ω et m_μ :

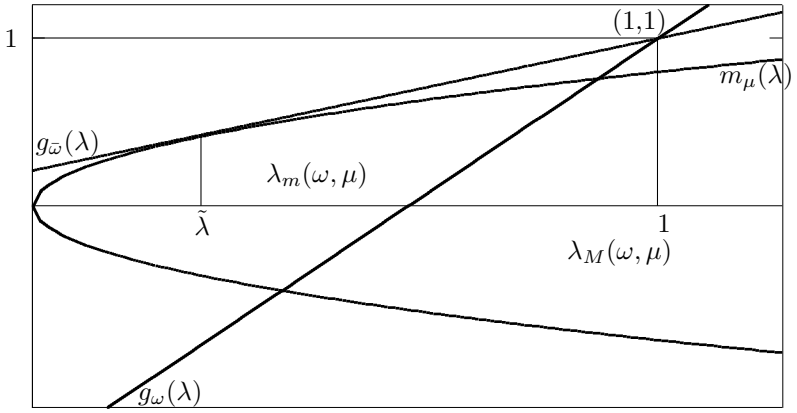


fig. 1.5

Pour $\omega > \bar{\omega}$, les racines de (2.1) sont complexes conjuguées (car la droite d'équation $g_\omega(\lambda)$ ne coupe pas les courbes $\pm m_\mu(\lambda)$). Il est clair, d'après le graphe ci-dessus, que pour $\omega \neq 0$ et tout $\forall \mu \in sp(J)$:

$$| \lambda_M(\omega, \mu) | = \begin{cases} \leq | \lambda_M(\omega, \rho(J)) | < 1, & si\ \lambda_M(\omega, \mu) \in \mathbb{R} \\ = | \omega - 1 | < 1, & sinon \end{cases}$$

or

$$| \lambda_M(\omega, \rho(J)) | \geq | \omega - 1 |$$

d'où

$$\forall \mu \in sp(J), \quad | \lambda_M(\omega, \mu) | \leq | \lambda_M(\omega, \rho(J)) | < 1$$

soit encore

$$\rho(H(\omega)) = | \lambda_M(\omega, \rho(J)) | < 1,$$

ce qui prouve que $H(\omega)$ converge pour $0 < \omega < 2$. Encore d'après le graphe ci-dessus, on a :

$$\rho(H(\omega)) = \begin{cases} \mid \lambda_M(\omega, \rho(J)) \mid > \mid \lambda_M(\bar{\omega}, \rho(J)) \mid, & \forall \omega < \bar{\omega} \\ \mid \omega - 1 \mid, & \forall \omega \geq \bar{\omega} \end{cases}$$

D'autre part $\bar{\omega}$ est la valeur de $\omega \in]0, 2[$ pour laquelle l'équation

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 (\rho(J))^2 \Longleftrightarrow \lambda^2 + \{2(\omega - 1) - \omega^2 (\rho(J))^2\}\lambda + (\omega - 1)^2 = 0$$

a une racine double, c'est à dire :

$$\{2(\omega - 1) - \omega^2 (\rho(J))^2\}^2 - 4(\omega - 1)^2 = 0$$

d'où

$$\omega^2 (\rho(J))^2 - 4\omega + 4 = 0$$

Cette équation admet une racine ≥ 2 et une autre égale à :

$$\begin{aligned} \bar{\omega} &= \frac{2 - \sqrt{4 - 4(\rho(J))^2}}{(\rho(J))^2} \\ &= \frac{2}{1 + \sqrt{1 - (\rho(J))^2}} \\ &= 1 + \left(\frac{\rho(J)}{1 + \sqrt{1 - \rho(J)^2}} \right)^2 \end{aligned}$$

et on a

$$\rho(H(\bar{\omega})) = \mid \lambda_M(\bar{\omega}, \rho(J)) \mid = \mid \bar{\omega} - 1 \mid = \bar{\omega} - 1.$$

D'où

$$\begin{aligned} \min_{\omega} \rho(H(\omega)) &= \min \left[\rho(H(\bar{\omega})); \min_{\omega \geq \bar{\omega}} \omega - 1 \right] \\ &= \bar{\omega} - 1 = \rho(H(\bar{\omega})) \end{aligned} .$$

Par conséquent

$$\omega_b = \bar{\omega}$$

vérifie ce qu'il faut pour le théorème.

2.4 Exercices : chapitre 2

Exercice 1

1. Calculer dans chacun des cas suivants le rayon spectral de la matrice de la méthode de Jacobi et le rayon spectral de la matrice de la méthode de Gauss-Seidel pour la résolution du système $Ax = b$

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ et } A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

   Que peut-on déduire ?

2. Montrer que si la matrice A est à diagonale dominante stricte :

$$\forall 1 \leq i \leq n, |a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

alors la méthode itérative de Jacobi pour la résolution de $Ax = b$ est convergente (on pourra montrer que $\|J\|_\infty < 1$).

3. Etudier la convergence de la méthode de relaxation (pour la résolution du système $Ax = b$) lorsque

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

Exercice 2
Soit $A = (a_{ij})$ une matrice d'ordre n telle que $a_{ii} \neq 0$ pour tout $i = 1, \ldots, n$. On rappelle que la matrice de Jacobi $J = L + U$ et la matrice de Gauss-Seidel $H = (I - L)^{-1}U$ avec $A = D - E - F$, $L = D^{-1}E$ et $U = D^{-1}F$. On dit que la matrice A est correctement ordonnée si :

$$\forall \alpha \neq 0, sp(J(\alpha)) = sp(J)$$

avec $J(\alpha) = \alpha L + \alpha^{-1}U$.
1. Calculer la matrice $J = (c_{ij})$ et la matrice $P_\alpha J P_\alpha^{-1} = (d_{ij})$ en fonction des coefficients de A avec $P_\alpha = diag(1, \alpha, \alpha^2, \ldots, \alpha^{n-1})$.
2. Déduire qu'une matrice tridiagonale est correctement ordonnée.
3. En utilisant un théorème du cours, montrer que si A est tridiagonale alors :

$$sp(H) = \{0\} \cup \{\mu^2 / \mu \in sp(J)\}$$

Exercice 3
Le but de l'exercice est d'étudier le problème

$$\begin{cases} -u"(x) &= f(x), x \in]0,1[, \\ u(0) &= u(1) = 0 \end{cases}$$

Soit $h = \frac{1}{N+1}$ avec $N \in \mathbb{N}^*$. Le problème discrétisé correspondant est : étant donné le vecteur $F = (f_i)_{1 \leq i \leq N}$ trouver $U = (u_i)_{1 \leq i \leq N}$ tel que

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

1. Montrer que le système (1) est équivalent à la résolution de

$$AU = F$$

où on explicitera la matrice $A = (a_{ij})_{1 \leq i,j \leq N}$.

2. Calculer les valeurs propres des matrices de Jacobi et Gauss-Seidel pour la résolution de $AU = F$ en fonction de celles de A (On pourra utiliser un exercice du chapitre précédent).
3. Comparer ces deux méthodes.
4. Donner le paramètre optimal de relaxation pour la matrice $\mathcal{L}_\omega$.

Exercice 4
Partie I :
Dans cette partie, on se propose de prouver quelques propriétés utiles pour la partie II. Soient K et M deux matrices carrées d'ordre n. On suppose que K et M sont symétriques et $(I+K)$ inversible, I étant la matrice identité d'ordre n.

1. Montrer que $\rho(K) = \|K\|_2$.
2. On rappelle que $\rho(L) \leq \|L\|$ pour toute matrice L et toute norme matricielle $||.||$. En utilisant 1), montrer que :
$$\rho(KM) \leq \rho(K)\rho(M).$$
3. Montrer que
   (a) $\lambda \in sp(K) \Rightarrow \lambda \neq -1$ et $\frac{\lambda}{1+\lambda} \in sp(K(I+K)^{-1})$,
   (b) $\beta \in sp(K(I+K)^{-1}) \Rightarrow \beta \neq 1$ et $\frac{\beta}{1-\beta} \in sp(K)$.
4. Montrer que : $\rho(K(I+K)^{-1}) < 1 \Leftrightarrow \frac{1}{2}I + K$ est définie positive.

Partie II :
Soit A une matrice carrée d'ordre n symétrique. Cette partie est consacrée à l'étude d'une méthode itérative de résolution du système linéaire $Ax = b$. On introduit la décomposition
$$A = D + H + V$$
où $D = cI, c > 0$ et où H et V sont deux matrices symétriques telles que les deux matrices $D+H$ et $D+V$ soient inversibles. On considère la méthode itérative suivante :
$$\left\{ \begin{array}{lcl} (D+H)x^{(k+\frac{1}{2})} & = & -Vx^{(k)} + b \\ (D+V)x^{(k+1)} & = & -Hx^{(k+\frac{1}{2})} + b \end{array} \right. \tag{2.1}$$

1. Exprimer $x^{(k+1)}$ en fonction de $x^{(k)}$. En déduire
$$\lim_{k\to\infty} x^{(k)} = x \Leftrightarrow \rho((D+V)^{-1}H(D+H)^{-1})V) < 1$$
2. On pose $B = D^{-1}H, C = D^{-1}V$.
   (a) Montrer que $\rho((D+V)^{-1}H(D+H)^{-1}V) = \rho(B(I+B)^{-1}C(I+C)^{-1})$
   (b) Montrer que les matrices $B(I+B)^{-1}$ et $C(I+C)^{-1}$ sont des matrices symétriques. (On pourra utiliser, après justification, que $B(I+B)^{-1} = I - (I+B)^{-1}$).
   (c) En déduire que $\rho((D+V)^{-1}H(D+H)^{-1}V) \leq \rho(B(I+B)^{-1})\rho(C(I+C)^{-1})$
3. Montrer que la méthode itérative (2.1) converge dès que $\frac{1}{2}D + H$ et $\frac{1}{2}D + V$ sont définies positives.

2.5 Corrigé des exercices : chapitre 2

Réponse 1
1. a) $A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$:
Jacobi : $M = D = I$,$N = E + F = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}$ et

$J = M^{-1}N = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}$
$\det(\lambda I - J) = \lambda^3 \Longrightarrow sp(J) = \{0\} \Longrightarrow \rho(J) = 0$
Gauss-Seidel : $M = D - E = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}$, $N = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$ et
$H = M^{-1}N = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 1 \end{pmatrix}$
$sp(H) = \{0, 2, 1\} \Longrightarrow \rho(H) = 2$.
Conclusion : la méthode de Jacobi converge et la méthode de Gauss-Seidel diverge.
b)$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$:
Jacobi : $M = D = 2I$,
$N = E + F = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{pmatrix}$ et $J = M^{-1}N = \frac{1}{2}\begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{pmatrix}$
$\det(\lambda I - J) = \lambda^3 + 5\lambda \Longrightarrow sp(J) = \{0, \pm\sqrt{5}\} \Longrightarrow \rho(J) = \sqrt{5}$
Gauss-Seidel : $M = D - E = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 2 & 0 \\ -1 & -1 & 2 \end{pmatrix}$,
$N = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix}$ et $H = M^{-1}N = \begin{pmatrix} 0 & 0.5 & -0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & -0.5 \end{pmatrix}$
$sp(H) = \{0, \pm 0.5\} \Longrightarrow \rho(H) = 0.5$.
Conclusion : la méthode de Jacobi diverge et la méthode de Gauss-Seidel converge.
2. La matrice de Jacobi est égale à :

$$J = M^{-1}N = \begin{pmatrix} 0 & \frac{-a_{12}}{a_{11}} & \frac{-a_{13}}{a_{11}} & \cdots & \frac{-a_{1n}}{a_{11}} \\ \frac{-a_{21}}{a_{22}} & 0 & \frac{-a_{23}}{a_{22}} & \cdots & \frac{-a_{2n}}{a_{22}} \\ \vdots & & & & \vdots \\ \vdots & & & & \frac{-a_{n-1n}}{a_{n-1n-1}} \\ \frac{-a_{n1}}{a_{nn}} & & & \frac{-a_{nn-1}}{a_{nn}} & 0 \end{pmatrix}$$

d'où,

$$\begin{aligned} \|J\|_\infty &= \max_i \sum_{j=1, j\neq i}^{n} |\frac{-a_{ij}}{a_{ii}}| \\ &= \max_i \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| < 1 \end{aligned}$$

3. $A = I - \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = D - E - F$, d'où :

$$\begin{aligned} H(\omega) &= (\frac{D}{\omega} - E)^{-1}(\frac{1-\omega}{\omega} I + F) \\ &= \begin{pmatrix} \frac{1}{\omega} & 0 & 0 \\ 1 & \frac{1}{\omega} & 0 \\ 0 & 0 & \frac{1}{\omega} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1-\omega}{\omega} & 0 & 0 \\ 0 & \frac{1-\omega}{\omega} & -1 \\ 0 & 0 & \frac{1-\omega}{\omega} \end{pmatrix} \\ &= \begin{pmatrix} 1-\omega & 0 & 0 \\ -\omega(1-\omega) & 1-\omega & -\omega \\ 0 & 0 & 1-\omega \end{pmatrix} \end{aligned}$$

ce qui donne $sp(H(\omega)) = \{1-\omega\}$, soit $\rho(H(\omega)) = |1-\omega|$. Par conséquent, la méthode de relaxation converge pour $\omega \in]0, 2[$.

Réponse 2

1.

$$D^{-1}E = \begin{pmatrix} 0 & & \cdots & & 0 \\ -\frac{a_{21}}{a_{22}} & 0 & & & \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & & \vdots \\ \vdots & & & & \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & -\frac{a_{nn-1}}{a_{nn}} & 0 \end{pmatrix}$$

$$D^{-1}F = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ & 0 & -\frac{a_{23}}{a_{22}} & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & 0 & & \vdots \\ & & & & -\frac{a_{n-1n}}{a_{n-1n-1}} \\ 0 & & \cdots & & 0 \end{pmatrix}$$

$J = D^{-1}E + D^{-1}F = (c_{ij})$, d'où :

$$c_{ij} = \begin{cases} 0 & \text{si} \quad i = j \\ -\frac{a_{ij}}{a_{ii}} & \text{si} \quad i \neq j \end{cases}$$

Si on note par d_{ij} les coefficients de la matrice $P_\alpha J P_\alpha^{-1}$, on obtient :

$$\begin{aligned} d_{ij} &= e_i^T P_\alpha J P_\alpha^{-1} e_j \\ &= e_i^T P_\alpha J \alpha^{1-j} e_j \\ &= \alpha^{1-j} e_i^T P_\alpha \textstyle\sum_{k=1}^n c_{kj} e_k \\ &= \alpha^{1-j} e_i^T \textstyle\sum_{k=1}^n \alpha^{k-1} c_{kj} e_k \\ &= \alpha^{1-j} \alpha^{i-1} c_{ij} = \alpha^{i-j} c_{ij} \end{aligned}$$

d'où :

$$d_{ij} = \begin{cases} 0 & \text{si} \quad i = j \\ -\alpha^{i-j} \frac{a_{ij}}{a_{ii}} & \text{si} \quad i \neq j \end{cases}$$

2. Si A est tridiagonale, les coefficients non nuls de A situés sur la diagonale a_{ij} avec $i - j = 0$, la sous-diagonale a_{ij} avec $i - j = 1$ et la sur-diagonale a_{ij} avec $i - j = -1$. D'autre part, les coefficients non nuls de la matrice $-E$ sont les coefficients de la sous-diagonale et les coefficients non nuls de $-F$ sont les coefficients de la sur-diagonale. Ce qui donne que :

$$P_\alpha J P_\alpha^{-1} = \alpha L + \alpha^{-1} U = J(\alpha)$$

Donc, $J(\alpha)$ et J sont semblables et par conséquent, elles ont le même spectre.

3. D'après un théorème du cours, les valeurs propres $\{\lambda \in sp(H)\}$ et les valeurs propres $\{\mu \in sp(J)\}$ sont liées par la relation suivante :

$$\lambda^2 = \lambda \mu^2$$

car $\omega = 1$. D'où, chaque valeur propre $\mu \in sp(J)$ est associée à deux valeurs propres de H : $\lambda = 0$ et $\lambda = \mu^2$. D'autre part, $\det(H) = \det(I - L)^{-1} \det(U) = 0$, donc $\lambda = 0$ est une valeur propre de H, ce qui donne :

$$sp(H) = \{0\} \cup \{\mu^2, \mu \in sp(J)\}$$

Réponse 3 1.

$$-\frac{1}{h^2}\begin{cases} -2u_1 & +u_2 & & = f_1 \\ u_1 & -2u_2 & +u_3 & = f_2 \\ \vdots & & & \\ u_{k-1} & -2u_k & +u_{k+1} & = f_k \\ \vdots & & & \\ u_{N-1} & -2u_N & & = f_N \end{cases}$$

ce qui donne le système $AU = F$ avec :

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & & & 0 \\ -1 & 2 & -1 & & & \\ 0 & -1 & 2 & -1 & & \\ \vdots & & & & & 0 \\ \vdots & & & & & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix} \text{ et } F = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ \vdots \\ f_N \end{pmatrix}$$

2. $A = D - E - F$ avec $D = \frac{2}{h^2} I$,

$$E = \frac{1}{h^2}\begin{pmatrix} 0 & 0 & 0 & & & 0 \\ 1 & 0 & 0 & & & \\ 0 & 1 & 0 & 0 & & \\ \vdots & & & & & 0 \\ \vdots & & & & & 0 \\ 0 & \cdots & & 0 & 1 & 0 \end{pmatrix} \text{ et } F = \frac{1}{h^2}\begin{pmatrix} 0 & 1 & 0 & & & 0 \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ \vdots & & & & & 0 \\ \vdots & & & & & 1 \\ 0 & \cdots & & 0 & 0 & 0 \end{pmatrix}$$

d'où :

$$J = D^{-1}E + D^{-1}F = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & & & 0 \\ 1 & 0 & 1 & & & \\ 0 & 1 & 0 & 1 & & \\ \vdots & & & & & 0 \\ \vdots & & & & & 1 \\ 0 & \cdots & & 0 & 1 & 0 \end{pmatrix}$$

J est donc tridiagonale. D'après un exercice du chapitre précédent :

$$sp(J) = \{\cos(\frac{k\pi}{n+1}); k = 1, \ldots, n\}$$

ce qui donne $\rho(J) = \cos(\frac{\pi}{n+1})$. D'après l'exercice précédent :

$$sp(H) = \{0\} \cup \{\cos^2(\frac{k\pi}{n+1}); k = 1, \ldots, n\}$$

Ce qui donne : $\rho(H) = \cos^2(\frac{\pi}{n+1})$.
3. Les deux méthodes sont convergentes et la convergence de la méthode de Gauss-Seidel est plus rapide que la convergence de la méthode de Jacobi car $\rho(H) = \rho(J)^2 < \rho(J) < 1$.
4. D'après un théorème du cours :

$$\omega_{optim} = \frac{2}{1 + \sqrt{1 - \rho(J)^2}} = \frac{2}{1 + \sin(\frac{\pi}{n+1})}.$$

Réponse 4

Partie I :
1. $\|K\|_2 = \sqrt{\rho(K^*K)}$, comme K est symétrique, d'après un théorème du cours :

$$\exists U unitaire / K = U^* diag(\lambda_i) U$$

avec $sp(K) = \{\lambda_i\}$. D'où, $K^*K = U^* diag(|\lambda_i|^2)U$, ce qui prouve que :

$$\rho(K^*K) = \max_i |\lambda_i|^2 = (\max_i |\lambda_i|)^2 = \rho(K)^2$$

2. $\rho(KM) \leq \|KM\|_2 \leq \|K\|_2 \|M\|_2 = \rho(K)\rho(M)$, car K et M sont des matrices symétriques.
3.a) $\lambda \in sp(K)$ alors, il existe $u \neq 0$ vecteur propre de K associé à λ tel que :

$$Ku = \lambda u$$

Comme $I + K$ est inversible, alors $\lambda \neq -1$ (sinon $(I + K)u = 0$ et $u \neq 0$).
D'autre part :

$$\begin{aligned} Ku = \lambda u &\Longrightarrow (I+K)u = (\lambda+1)u \\ &\Longrightarrow u = (\lambda+1)(I+K)^{-1}u \\ &\Longrightarrow Ku = (\lambda+1)K(I+K)^{-1}u \\ &\Longrightarrow \lambda u = (\lambda+1)K(I+K)^{-1}u \\ &\Longrightarrow K(I+K)^{-1}u = \tfrac{\lambda}{1+\lambda}u \\ &\Longrightarrow \tfrac{\lambda}{1+\lambda} \in sp(K(I+K)^{-1}) \end{aligned}$$

b) $\beta \in sp(K(I + K)^{-1})$ alors, il existe $v \neq 0$ vecteur propre de $K(I + K)^{-1}$ associé à β tel que :

$$K(I+K)^{-1}v = \beta v$$

Comme $v \neq 0$, alors $u = (I + K)^{-1}v \neq 0$ et on a $v = (I + K)u$, d'où :

$$\begin{aligned} K(I+K)^{-1}v = \beta v &\Longrightarrow Ku = \beta(I+K)u \\ &\Longrightarrow (1-\beta)Ku = \beta u \end{aligned}$$

ce qui montre que $\beta \neq 1$ et on a :

$$Ku = \frac{\beta}{1-\beta}u$$

ce qui donne :

$$\frac{\beta}{1-\beta} \in sp(K)$$

4. $\frac{1}{2}I + K$ est symétrique et elle est définie positive si et seulement si :

$$sp(\frac{1}{2}I + K) \subset]0, +\infty[$$

Supposons que $\rho(K(I + K)^{-1}) < 1$. Soit $\gamma \in sp(\frac{1}{2}I + K)$, alors , il existe $w \neq 0$ tel que :

$$(\frac{1}{2}I + K)w = \gamma w$$

ce qui donne $K\omega = (\gamma - \frac{1}{2})w$, donc, $\lambda = \gamma - \frac{1}{2} \in sp(K)$. D'après a), $\lambda = \gamma - \frac{1}{2} \neq -1$ et on a :

$$\frac{\gamma - \frac{1}{2}}{1 + \gamma - \frac{1}{2}} = \frac{\gamma - \frac{1}{2}}{\gamma + \frac{1}{2}} \in sp(K(I+K)^{-1})$$

Or $\rho(K(I + K)^{-1}) < 1$, d'où :

$$|\frac{\gamma - \frac{1}{2}}{\gamma + \frac{1}{2}}| = |\frac{\frac{1}{2} - \gamma}{\gamma + \frac{1}{2}}| = |\frac{1}{\gamma + \frac{1}{2}} - 1| < 1$$

ce qui donne :

$$-1 < \frac{1}{\gamma + \frac{1}{2}} - 1 < 1$$

d'où : $\frac{1}{2} < \gamma + \frac{1}{2}$, ou encore : $\gamma > 0$. Supposons que $\frac{1}{2}I + K$ est définie positive. Soit $\beta \in sp(K(I+K)^{-1})$. D'après b) :

$$\beta \neq 1 et \frac{\beta}{1-\beta} \in sp(K)$$

d'où $\frac{1}{2} + \frac{\beta}{1-\beta} \in sp(\frac{1}{2}I + K)$. Par conséquent :

$$\frac{1}{2} + \frac{\beta}{1-\beta} = \frac{1+\beta}{2(1-\beta)} > 0$$

ou encore :

$$\frac{1+\beta}{1-\beta} > 0$$

ce qui donne $\{1+\beta > 0 \text{ et } 1-\beta > 0\}$ ou $\{1+\beta < 0 \text{ et } 1-\beta < 0\}$. La première possibilité donne $-1 < \beta < 1$ et la deuxième est impossible, d'où, ce qu'il faut.

Partie II.

1. De la première équation, on a :

$$x^{(k+\frac{1}{2})} = -(D+H)^{-1}Vx^{(k)} + (D+H)^{-1}b$$

La deuxième équation devient alors :

$$x^{(k+1)} = (D+V)^{-1}H(D+H)^{-1}Vx^{(k)} - (D+V)^{-1}H(D+H)^{-1}b + b$$

ce qui donne une méthode itérative de matrice $(D+V)^{-1}H(D+H)^{-1}V$, donc, la convergence de la suite (x^k) est équivalente à $\rho((D+V)^{-1}H(D+H)^{-1}V) < 1$.

2. a) On a :

$$\begin{aligned}(D+V)^{-1}H(D+H)^{-1}V &= (I+D^{-1}V)^{-1}D^{-1}H(I+D^{-1}H)^{-1}D^{-1}V \\ &= (I+C)^{-1}B(I+B)^{-1}C \\ &= (I+C)^{-1}B(I+B)^{-1}C(I+C)^{-1}(I+C)\end{aligned}$$

d'où, la matrice $(D+V)^{-1}H(D+H)^{-1}V$ est semblable à $B(I+B)^{-1}C(I+C)^{-1}$, donc, elles ont le même spectre et par conséquent le même rayon spectral.

b)

$$\begin{aligned}B(I+B)^{-1} &= (B+I-I)(I+B)^{-1} = (B+I)(I+B)^{-1} - I(I+B)^{-1} \\ &= I - (I+B)^{-1}\end{aligned}$$

et il est clair que $I-(I+B)^{-1}$ est une matrice symétrique. Même chose pour $C(I+C)^{-1}$.

c) On a :

$$\rho((D+V)^{-1}H(D+H)^{-1}V) = \rho(B(I+B)^{-1}C(I+C)^{-1})$$

Comme $B(I+B)^{-1}$ et $C(I+C)^{-1}$ sont deux matrices symétriques, d'après la partie I :

$$\rho(B(I+B)^{-1}C(I+C)^{-1}) \leq \rho(B(I+B)^{-1})\rho(C(I+C)^{-1})$$

3. D'après la partie I. question 4. $\rho(B(I+B)^{-1}) < 1$ si et seulement si $\frac{1}{2}I + B = D^{-1}(\frac{1}{2}D + H)$ est définie positive ce qui est encore équivalent à $\frac{1}{2}D + H$ est une matrice définie positive car $D = cI$ avec $c > 0$. On montre de même que : $\frac{1}{2}D + V$ est définie positive si et seulement si $\rho(C(I+C)^{-1}) < 1$. Par conséquent, $\frac{1}{2}D + H$ et $\frac{1}{2}D + V$ sont définies positives nous donne que $\rho(B(I+B)^{-1})\rho(C(I+C)^{-1}) < 1$, ce qui donne, d'après 2.a) :

$$\rho((D+V)^{-1}H(D+H)^{-1}V) < 1$$

D'où, d'après 1., la convergence de la méthode itérative (1).

Chapitre 3

Méthodes itératives pour le calcul des valeurs propres et des vecteurs propres

3.1 Introduction

Il est connu, d'après le théorème d'Abel, qu'il n'existe pas de méthode qui, au bout d'un nombre fini d'opérations, donne les racines d'un polynôme de degré $n \geq 5$. D'autre part, il est facile de vérifier que tout polynôme de degré $n \geq 2$ écrit sous la forme :

$$p(x) = x^n + a_1 x^{n-1} + \cdots + a_n$$

est égal, à un coefficient multiplicatif près, au polynôme caractéristique de la matrice

$$A = \begin{pmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_n \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{pmatrix}$$

Par conséquent, les méthodes de calcul des valeurs propres d'une matrice A sont, en général, des méthodes itératives d'approximation, qui convergent (dans un sens à préciser) vers les valeurs propres.

Une façon pour approcher les valeurs propres d'une matrice consiste à construire une suite de matrices semblables à A :

$$A_k = P_k^{-1} A P_k$$

qui convergent vers une matrice dont les valeurs propres sont faciles à calculer, diagonale ou triangulaire par exemple.

3.2 Méthode de Jacobi

La méthode de Jacobi s'applique uniquement au calcul des valeurs propres et des vecteurs propres des matrices symétriques.

On a vu, chapitre 1 corollaire 1.1, qu'une matrice symétrique est diagonalisable dans $\mathbb{R}$ et possède une base orthonormée de vecteurs propres, c'est à dire : il existe une matrice orthogonale O et n nombres réels $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ tels que

$$O^T A O = diag(\lambda_1, \cdots, \lambda_n).$$

De plus, la $i$ème colonne de O est égale à un vecteur propre associé à la valeur propre λ_i.

DÉFINITION 3.2.1 *Soient p et q deux entiers vérifiant $1 \leq p < q \leq n$, et θ un nombre réel. On définit la matrice de rotation dans le plan (p, q) et d'angle θ par :*

$$\Omega = \begin{pmatrix} 1 & & & & & & & & & & \\ & 1 & & & & & & & & & \\ & & . & & & & & & & & \\ & & & . & & & & & & & \\ & & & & cos\theta & & & & -sin\theta & & \\ & & & & & 1 & & & & & \\ & & & & & & 1 & & & & \\ & & & & & & & . & & & \\ & & & & sin\theta & & & & cos\theta & & \\ & & & & & & & & & 1 & \\ & & & & & & & & & & . \\ & & & & & & & & & & & 1 \end{pmatrix}$$

Dans toute la suite et à chaque fois qu'on utilise une rotation d'angle θ dans le plan (p, q), l'entier p est supposé strictement inférieur à l'entier q.

3.2.1 Résultats préliminaires

On remarque facilement les points suivants :

– La matrice de rotation Ω est une matrice ortogonale :

$$\Omega^T \Omega = I$$

– Dans la transformation $A = (a_{ij}) \longrightarrow B = (b_{ij}) = \Omega A \Omega^T$, avec Ω la rotation d'angle θ dans le plan (p, q), seules les $p$ième et $q$ième colonnes et lignes de la matrice A sont modifiées :

$$\begin{cases} b_{ij} = a_{ij} \ si \ i \neq p, q \ \ et \ \ j \neq p, q, \\ b_{pi} = a_{pi} cos\theta - a_{qi} sin\theta \ \ si \ \ i \neq p, q, \\ b_{qi} = a_{pi} sin\theta + a_{qi} cos\theta \ \ si \ \ i \neq p, q, \\ b_{pp} = a_{pp} cos^2\theta + a_{qq} sin^2\theta - a_{pq} sin2\theta \\ b_{qq} = a_{pp} sin^2\theta + a_{qq} cos^2\theta + a_{pq} sin2\theta \\ b_{pq} = b_{qp} = a_{pq} cos2\theta + \frac{a_{pp} - a_{qq}}{2} sin2\theta \end{cases} \tag{3.1}$$

– La norme de *Frobenius* est invariante par transformation orthogonale, d'où,

$$\sum_{i,j=1}^{n} b_{ij}^2 = \sum_{i,j=1}^{n} a_{ij}^2 \tag{3.2}$$

PROPOSITION 3.2.1 *Soient $A = (a_{ij})$ une matrice symétrique et $B = (b_{ij}) = \Omega A \Omega^T$, où Ω est la rotation d'angle θ dans le plan (p, q), alors :*

(i) $b_{pp} + b_{qq} = a_{pp} + a_{qq}$

(ii) $b_{pp} - b_{qq} = (a_{pp} - a_{qq}) \cos 2\theta - 2a_{pq} \sin 2\theta$

(iii) $b_{pp}^2 + 2b_{pq}^2 + b_{qq}^2 = a_{pp}^2 + 2a_{pq}^2 + a_{qq}^2$

(iv) $\sum_{i \neq j} b_{ij}^2 = \sum_{i \neq j} a_{ij}^2 - 2(a_{pq}^2 - b_{pq}^2)$

DÉMONSTRATION :

Les formules (i) et (ii) se déduisent facilement des formules (3.1). On vérifie facilement par le calcul que :

$$\begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

d'où, si on applique la norme de *Frobenius* à cette égalité et en utilisant encore une fois que la norme de *Frobenius* est invariante par transformation ortogonale, on obtient (iii). D'autre part, la formule (3.2) nous donne :

$$\begin{aligned} \sum_{i,j=1}^{n} b_{ij}^2 &= (\sum_{i \neq j} b_{ij}^2) + (\sum_{i \neq p,q} b_{ii}^2) + b_{pp}^2 + b_{qq}^2 \\ &= \sum_{i,j=1}^{n} a_{ij}^2 \\ &= (\sum_{i \neq j} a_{ij}^2) + (\sum_{i \neq p,q} a_{ii}^2) + a_{pp}^2 + a_{qq}^2 \end{aligned}$$

et la première formule de (3.1) nous donne :

$$\sum_{i \neq p,q} b_{ii}^2 = \sum_{i \neq p,q} a_{ii}^2$$

il suffit donc d'utiliser (iii) pour obtenir (iv).

3.2.2 Principe de la méthode de Jacobi

La méthode de Jacobi consiste à construire une suite de matrices semblables à la matrice A, de la manière suivante :

$$\begin{aligned} A_1 &= A \\ A_{k+1} &= \Omega_k A_k \Omega_k^T \end{aligned}$$

où Ω_k est une matrice de rotation dans un plan $(\bar{p}, \bar{q})$ et d'angle $\bar{\theta}$ bien choisis. Il est clair que (A_k) est une suite de matrices semblables à A et on a :

$$A_{k+1} = (\Omega_k \Omega_{k-1} \cdots \Omega_1) A (\Omega_k \Omega_{k-1} \cdots \Omega_1)^T$$

On veut que A_k tende vers une matrice diagonale (qui est nécessairement semblable à A) quand $k \to \infty$. Par conséquent, si on note par $A_{k+1} = (b_{ij}(\theta, p, q))$ et $A_k = (a_{ij})$, on a intérêt à choisir $\bar{\theta}$ et $(\bar{p}, \bar{q})$ qui minimisent la somme des carrés des éléments hors diagonaux de la matrice A_{k+1}, c'est à dire :

$$\sum_{i \neq j} b_{ij}^2(\bar{\theta}, \bar{p}, \bar{q}) = \min_{\theta, p \neq q} \sum_{i \neq j} b_{ij}^2(\theta, p, q) \leq \sum_{i \neq j} a_{ij}^2$$

D'après la formule (iv) de la proposition 3.2.1, il faut et il suffit de choisir $\bar{p}, \bar{q}$ et $\bar{\theta}$ tels que :

$$a^2_{\bar{p}\bar{q}} - b^2_{\bar{p}\bar{q}}(\bar{\theta}, \bar{p}, \bar{q}) = \max_{\theta, p \neq q}(a^2_{pq} - b^2_{pq}(\theta, p, q))$$

donc, $\bar{p}, \bar{q}$ sont choisis tels que :

$$\mid a_{\bar{p}\bar{q}} \mid = \max_{p \neq q} \mid a_{pq} \mid$$

et $\bar{\theta}$ tel que :

$$b_{\bar{p}\bar{q}}(\bar{\theta}, \bar{p}, \bar{q}) = 0$$

PROPOSITION 3.2.2 *Pour tout (p, q) tel que $a_{pq} \neq 0$, il existe une unique valeur non nulle de θ, $-\frac{\pi}{4} < \theta \leq \frac{\pi}{4}$, telle que :*

$$b_{pq}(\theta, p, q) = 0$$

DÉMONSTRATION : D'après les formules (3.1), on a :

$$b_{pq} = b_{qp} = 0 \Longleftrightarrow cotg(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}}$$

or, la fonction $cotg(2\theta)$ est bijective de $]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}]$ dans $\mathbb{R}$. D'où, l'existence et l'unicité de $\theta \in]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}]$ tel que $b_{pq}(\theta, p, q) = 0$.

3.2.3 Convergence de la méthode de Jacobi

Soit $A_1 = A = (a_{ij}) = (a^1_{ij})$ et $A_k = (a^k_{ij})$ la matrice de l'itération k de la méthode de Jacobi. Soit p_k, q_k tels que

$$\mid a^k_{p_k q_k} \mid = \max_{p \neq q} \mid a^k_{pq} \mid$$

Soit $\theta_k \in]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}]$ l'unique solution de

$$cotg(2\theta) = \frac{a^k_{q_k q_k} - a^k_{p_k p_k}}{2a^k_{p_k q_k}} \tag{3.3}$$

et Ω_k la matrice de rotation dans le plan (p_k, q_k) d'angle θ_k. On définit la matrice A_{k+1} de l'itération $k+1$ par :

$$A_{k+1} = \Omega_k A_k \Omega_k^T$$

THÉORÈME 3.2.1 *La suite (A_k) de matrices obtenues par la méthode de Jacobi est convergente, et*

$$\lim_{k \longrightarrow \infty} A_k = D$$

où D est une matrice diagonale qui a le même spectre que A.

DÉMONSTRATION : On note par

$$A_k = D_k + E_k, \quad \text{avec } D_k = diag\{a_{ii}^k\}$$

$$\varepsilon_k = \| E_k \|_F^2 = \sum_{i \neq j} | a_{ij}^k |^2$$

Il est clair que :

$$| a_{p_k q_k}^k |^2 \leq \varepsilon_k \leq n(n-1) | a_{p_k q_k}^k |^2 \tag{3.4}$$

D'autre part, d'après la formule (iv) de la proposition 3.2.1

$$\varepsilon_{k+1} = \varepsilon_k - 2 | a_{p_k q_k}^k |^2$$

car $a_{p_k q_k}^{k+1} = b_{p_k q_k} = 0$, et de (3.4) on tire que :

$$- | a_{p_k q_k}^k |^2 \leq \frac{-1}{n(n-1)} \varepsilon_k$$

ce qui donne :

$$0 \leq \varepsilon_{k+1} \leq r\varepsilon_k \leq r^2 \varepsilon_{k-1} \leq \cdots \leq r^k \varepsilon_1$$

avec

$$0 < r = (1 - \frac{2}{n(n-1)}) < 1$$

Par conséquent,

$$\lim_{k \to \infty} \varepsilon_k = 0$$

Ce qui prouve que

$$\lim_{k \to \infty} E_k = 0$$

Soit $i \in \{1, 2, .., n\}$ fixé. D'après les formules (3.1), (3.3), pour tout $k \in \mathbb{N}$, si $i \neq p_k$ et $i \neq q_k$ on a :

$$a_{ii}^{k+1} - a_{ii}^k = 0$$

si $i = p_k$ on a :

$$\begin{aligned} a_{p_k p_k}^{k+1} - a_{p_k p_k}^k &= a_{p_k p_k}^k cos^2(\theta_k) + a_{q_k q_k}^k sin^2(\theta_k) - a_{p_k q_k}^k sin(2\theta_k) - a_{p_k p_k}^k \\ &= (a_{q_k q_k}^k - a_{p_k p_k}^k) sin^2(\theta_k) - a_{p_k q_k}^k sin(2\theta_k) \\ &= 2a_{p_k q_k}^k cotg(2\theta_k) sin^2(\theta_k) - a_{p_k q_k}^k sin(2\theta_k) \\ &= [2cotg(2\theta_k) sin^2(\theta_k) - sin(2\theta_k)] a_{p_k q_k}^k \\ &= -tg(\theta_k) a_{p_k q_k}^k \end{aligned}$$

et enfin si $i = q_k$ on a :

$$\begin{aligned} a_{q_k q_k}^{k+1} - a_{q_k q_k}^k &= a_{p_k p_k}^k sin^2(\theta_k) + a_{q_k q_k}^k cos^2(\theta_k) + a_{p_k q_k}^k sin(2\theta_k) - a_{p_k p_k}^k \\ &= (a_{q_k q_k}^k - a_{p_k p_k}^k) cos^2(\theta_k) + a_{p_k q_k}^k sin(2\theta_k) \\ &= 2a_{p_k q_k}^k cotg(2\theta_k) cos^2(\theta_k) + a_{p_k q_k}^k sin(2\theta_k) \\ &= [2cotg(2\theta_k) cos^2(\theta_k) + sin(2\theta_k)] a_{p_k q_k}^k \\ &= tg(\theta_k) a_{p_k q_k}^k \end{aligned}$$

or $| \theta_k | \leq \frac{\pi}{4}$ ce qui donne $|tg(\theta_k)| \leq 1$, d'où :

$$|a_{p_k p_k}^{k+1} - a_{p_k p_k}^k| \leq |a_{p_k q_k}^k| \text{ et } |a_{q_k q_k}^{k+1} - a_{q_k q_k}^k| \leq |a_{p_k q_k}^k|$$

D'autre part

$$\forall k \in \mathbb{N},\ \mid a^k_{p_k q_k} \mid \leq \sqrt{\varepsilon_k} \leq (\sqrt{r})^{k-1}\sqrt{\varepsilon_1}$$

par conséquent

$$\forall k \in \mathbb{N},\ \ \mid a^{k+1}_{ii} - a^k_{ii} \mid \leq (\sqrt{r})^{k-1}\sqrt{\varepsilon_1}$$

d'où, la série numérique $\sum_k \mid a^{k+1}_{ii} - a^k_{ii} \mid$ est convergente, ce qui prouve que la série $\sum_k (a^{k+1}_{ii} - a^k_{ii})$ est convergente ou encore, la suite des sommes partielles

$$S_k = \sum_{l=1}^{k} (a^{l+1}_{ii} - a^l_{ii}) = a^{k+1}_{ii} - a^1_{ii}$$

est convergente. Par conséquent, pour tout i fixé, la suite (a^k_{ii}) converge, ce qui prouve que la suite de matrices (D_k) est convergente vers une matrice diagonale D. D'autre part, on a :

$$det(A_k - \lambda I) = det(A - \lambda I)$$

et par continuité, en utilisant le fait que $E_k \to 0$ et $D_k \to D$, on a :

$$\lim_{k\to\infty} det(A_k - \lambda I) = det(D - \lambda I)$$

Par conséquent, A et D ont le même polynôme caractéristique et la suite de Jacobi (A_k) converge vers D.

THÉORÈME 3.2.2 *On suppose que toutes les valeurs propres de A sont distinctes. Alors la suite de matrices orthogonales*

$$U_k = \Omega_k \Omega_{k-1} \cdots \Omega_1$$

converge vers une matrice orthogonale U dont les colonnes de la matrice U^T sont des vecteurs propres de la matrice A.

DÉMONSTRATION : Soit

$$\delta = \frac{1}{3} \min_{i\neq j} \mid \lambda_i - \lambda_j \mid > 0$$

avec

$$D = diag\{\lambda_1, \lambda_2, \cdots, \lambda_n\} = \lim_{k\to\infty} D_k$$

et $D_k = diag(a^k_{ii})$. Soit $k_0 \geq 1$ tel que pour tout $1 \leq i \leq n$, on a :

$$\forall k \geq k_0,\ \ \mid a^k_{ii} - \lambda_i \mid \leq \delta$$

d'où

$$\forall k \geq k_0,\ \forall i \neq j,\ \mid a^k_{ii} - a^k_{jj} \mid \geq \delta$$

en particulier

$$\forall k \geq k_0,\ \mid a^k_{q_k q_k} - a^k_{p_k p_k} \mid \geq \delta$$

ou encore

$$\forall k \geq k_0,\ 2 \mid a^k_{p_k q_k} \mid\mid cotg(2\theta_k) \mid \geq \delta$$

ce qui prouve que

$$\forall k \geq k_0, \quad \mid tg(2\theta_k) \mid \leq 2\frac{\mid a^k_{p_k q_k} \mid}{\delta}$$

D'autre part,

$$\mid \alpha \mid \leq \mid tg(\alpha) \mid, \ \forall \alpha \in]-\pi/2, \pi/2[$$

d'où

$$\forall k \geq k_0, \quad \mid \theta_k \mid \leq \frac{\mid a^k_{p_k q_k} \mid}{\delta} \leq \frac{\sqrt{\varepsilon_k}}{\delta}$$

D'autre part, en utilisant le fait que la norme de *Frobenius* est invariante par transformation ortogonale et que $U_{k+1} = \Omega_{k+1} U_k$:

$$\begin{array}{rcl} \parallel U_{k+1} - U_k \parallel_F^2 & = & \parallel (\Omega_{k+1} - I) U_k \parallel_F^2 \\ & = & \parallel (\Omega_{k+1} - I) \parallel_F^2 \\ & = & 2(\cos\theta_{k+1} - 1)^2 + 2(\sin(\theta_{k+1}))^2 \\ & = & 8\sin^2(\theta_{k+1}/2) \\ & \leq & 2 \mid \theta_{k+1} \mid^2 \end{array}$$

Par conséquent

$$\forall k \geq k_0, \quad \parallel U_{k+1} - U_k \parallel_F \leq \sqrt{2}\frac{\sqrt{\varepsilon_{k+1}}}{\delta} \leq \frac{\sqrt{2\varepsilon_1}}{\delta}(\sqrt{r})^k$$

d'où, la série $\sum \parallel U_{k+1} - U_k \parallel_F$ est convergente, donc, la série $\sum(U_{k+1} - U_k)$ est convergente. Ce qui prouve que la suite des sommes partielles

$$S_k = \sum_{l=1}^{k} U_{l+1} - U_l = U_{k+1} - U_1$$

est convergente. Donc, la suite de matrices orthogonales (U_k) est convergente vers une matrice orthogonale U et par passage à la limite on a :

$$D = UAU^T$$

d'où, la $i$ème colonne de U^T est un vecteur propre de A associé à λ_i.

3.2.4 Mise en oeuvre de la méthode de Jacobi

Dans la pratique on ne cherche pas à calculer le nombre θ_k pour déterminer les coefficients des lignes et des colonnes p_k, q_k de A_{k+1} ; on peut les obtenir par les relations algébriques suivantes : soit

$$R = \frac{a^k_{q_k q_k} - a^k_{p_k p_k}}{2a^k_{p_k q_k}}$$

et

$$t = tg\theta_k$$

on a les relations trigonométriques suivantes :

$$c = cos\theta_k = \frac{1}{\sqrt{1+t^2}}$$

$$s = sin\theta_k = \frac{t}{\sqrt{1+t^2}}$$

et pour que $a_{p_k q_k}^{k+1}$ soit nul, il faut que

$$cotg2\theta_k = \frac{1}{tg2\theta_k} = \frac{1-t^2}{2t} = R$$

avec la condition $t \in]-1,0[\cup]0,1]$. D'où, t est égal à l'unique solution du trinôme

$$t^2 + 2Rt - 1 = 0, \quad t \in]-1,0[\cup]0,1].$$

On obtient alors les formules suivantes :

$$\begin{cases} a_{p_k i}^{k+1} = ca_{p_k i}^k - sa_{q_k i}^k \quad si \quad i \neq p_k, q_k, \\ a_{q_k i}^{k+1} = sa_{p_k i}^k + ca_{q_k i}^k \quad si \quad i \neq p_k, q_k, \\ a_{p_k p_k}^{k+1} = a_{p_k p_k}^k - ta_{p_k q_k}^k \\ a_{q_k q_k}^{k+1} = a_{q_k q_k}^k + ta_{p_k q_k}^k \\ a_{p_k q_k}^{k+1} = a_{q_k p_k}^{k+1} = 0 \end{cases}$$

REMARQUE 3.2.1 *Un élément annulé à l'itération k peut devenir non nul à une itération $l > k$.*

On distingue trois stratégies pour le choix du couple (p, q) :

1. *Méthode de Jacobi classique : on choisit l'un des couples pour lesquels*

$$\mid a_{pq}^k \mid = \max_{i \neq j} \mid a_{ij}^k \mid .$$

2. *Méthode de Jacobi cyclique : on annule successivement tous les éléments hors-diagonaux par un balayage cyclique, toujours le même : par exemple, on choisit les couples (p, q) dans l'ordre suivant*

$$(1,2),(1,3),\cdots,(1,n);(2,3),\cdots,(2,n);\cdots;(n-1,n)$$

3. *Méthode de Jacobi avec seuil : on procède comme dans le cas de la méthode de Jacobi cyclique mais on omet d'annuler les éléments hors-diagonaux dont le module est inférieur à un certain seuil.*

EXEMPLE
Soit

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

On pose $A_1 = (a_{ij}^1) = A$. La matrice $A_2 = (a_{ij}^2)$ de la première itération de la méthode de Jacobi classique est obtenue comme suit : on cherche d'abord la position (p, q) du plus grand coefficient en valeur absolue en dehors de la diagonale de A_1 : c'est la position (2,3), on calcule le nombre :

$$R = \frac{a_{33}^1 - a_{22}^1}{2a_{23}^1} = \frac{3-3}{-8} = 0$$

on résout l'équation : $t^2+2Rt-1=t^2-1=0$, on prend l'unique solution qui est dans $]-1,0[\cup]0,1]$ (dans notre cas c'est $t=1$), on pose $c=1/\sqrt{1+t^2}=1/\sqrt{2}$, $s=t/\sqrt{1+t^2}=1/\sqrt{2}$ et enfin on a :

$$\begin{array}{l} a_{11}^2=a_{11}^1=2 \\ a_{21}^2=a_{12}^2=c\times a_{21}^1-s\times a_{31}^1=-1/\sqrt{2}-1/\sqrt{2}=-2/\sqrt{2} \\ a_{31}^2=a_{13}^2=s\times a_{21}^1+c\times a_{31}^1=-1/\sqrt{2}+1/\sqrt{2}=0 \\ a_{32}^2=a_{23}^2=0 \\ a_{22}^2=a_{22}^1-t\times a_{23}^1=3+4=7 \\ a_{33}^2=a_{33}^1+t\times a_{23}^1=3-4=-1 \end{array}$$

ce qui donne :

$$A_2=\begin{pmatrix} 2 & \frac{-2}{\sqrt{2}} & 0 \\ \frac{-2}{\sqrt{2}} & 7 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

3.3 Transformation de Householder et la factorisation QR

3.3.1 Transformation de Householder

A tout vecteur $u\in\mathbb{R}^n$, on associe la matrice

$$H=\begin{cases} I-\dfrac{2uu^T}{\parallel u\parallel_2^2}=I-\dfrac{uu^T}{\beta},\ \text{avec } \beta=\dfrac{1}{2}\parallel u\parallel_2^2\ \text{ si } u\neq 0, \\ I,\ \text{si } u=0 \end{cases},$$

la matrice H est appelée la matrice de Householder et le vecteur u est appelé le vecteur de Householder. (On note parfois $H(u)$ pour montrer la dépendance du vecteur u). Il est facile de voir que les matrices de Householder sont symétriques ($H=H^T$), orthogonales ($H^TH=HH^T=I$) et dépendent seulement de la direction du vecteur de Householder.

Les matrices de Householder ont deux propriétés importantes :

– pour tout vecteur $a\neq 0$ et tout vecteur $0\neq b\neq a$, avec $\parallel a\parallel_2=\parallel b\parallel_2$, il existe un vecteur de Householder u tel que :

$$Ha=(I-\frac{uu^T}{\beta})a=b$$

soit encore

$$(-\frac{u^Ta}{\beta})u=b-a \tag{3.5}$$

d'où u est un multiple de $b-a\neq 0$. D'autre part, si $u=b-a$, en utilisant le fait que $b^Tb=a^Ta$, on obtient :

$$\begin{array}{rl} Ha & =a-2\dfrac{(b-a)^Ta}{(b-a)^T(b-a)}(b-a) \\ & =b+\{-1-2\dfrac{(b-a)^Ta}{(b-a)^T(b-a)}\}(b-a)=b \end{array} \tag{3.6}$$

– tout vecteur transformé par H a une forme spéciale. Si on applique H à un vecteur c, on obtient :

$$Hc = (I - \frac{uu^T}{\beta})c = c - (\frac{u^T c}{\beta})u \tag{3.7}$$

d'où, le vecteur Hc est la différence entre le vecteur c et un multiple du vecteur de Householder u. On déduit aussi de (3.7) que l'application de H à c ne change pas les composantes qui correspondent à des composantes nulles de u et que c est invariant par H si $u^T c = 0$. Enfin, le calcul de Hc nécessite seulement le vecteur u et le scalaire β.

Par exemple, soit $u = (-1, -2)^T$, on a

$$\| u \|_2^2 = 5$$

$$H = I - \frac{uu^T}{\beta} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{2}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{8}{5} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{-4}{5} \\ \frac{-4}{5} & \frac{-3}{5} \end{pmatrix}.$$

3.3.2 Factorisation QR

Soit $A = (a_{ij})$ une matrice carrée réelle quelconque. La factorisation QR consiste à trouver une matrice Q orthogonale et une matrice R triangulaire supérieure telles que : $A = QR$. En utilisant les propriétés des matrices de Householder, on va construire une suite de $(n-1)$ matrices de Householder telles que :

$$H_{n-1} H_{n-2} \cdots H_2 H_1 A = R$$

où R est une matrice triangulaire supérieure.

La première étape consiste à construire une matrice de Householder H_1 qui transforme a_1 (la première colonne de A) en un multiple de e_1, ce qui revient à annuler les composantes $a_{21}, a_{31}, \cdots, a_{n1}$. La norme Euclidienne est invariante par transformation orthogonale, par conséquent, on cherche u_1 tel que

$$H_1 a_1 = (I - \frac{u_1 u_1^T}{\beta_1})a_1 = \pm \| a_1 \|_2 \, e_1 = \begin{pmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

avec $| r_{11} | = \| a_1 \|_2$. Nous savons que u_1 doit être un multiple de $\pm \| a_1 \|_2 \, e_1 - a_1$ (voir (3.5)). Comme H_1 dépend uniquement de la direction de u_1, on prendra u_1 comme suit :

$$u_1 = \begin{pmatrix} a_{11} - r_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}.$$

Par définition de u_1, le signe de r_{11} peut-être positif ou négatif. Pour éviter les erreurs numériques, le signe de r_{11} est en général choisi de la manière suivante :

$$sign(r_{11}) = -sign(a_{11}).$$

pour que $\beta = \frac{1}{2} \parallel u \parallel_2^2$ soit le plus grand possible, car on divise par β. Après application de la matrice de Householder H_1, la première colonne de $A_2 = H_1 A$ est un multiple de e_1 :

$$A_2 = H_1 A = \begin{pmatrix} r_{11} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}.$$

Tous les éléments de A sont modifiés. Pour i allant de 2 à n, on a : *la colonne a_i est remplacée par* $a_i - \frac{u_1^T a_i}{\beta_1} u_1$.

A la deuxième étape, il faut que la matrice de Householder H_2 ne change pas la première colonne et la première ligne de A_2. Pour cela, il suffit de choisir le vecteur de Householder u_2 orthogonal à e_1, c'est à dire la première composante de u_2 nulle. Avec un tel choix, l'application de H_2 à un vecteur quelconque ne change pas sa première composante et l'application de H_2 à un multiple de e_1 (comme le cas de la première colonne de A_2) le laisse inchangé, d'où :

$$a = \begin{pmatrix} a_{12}^{(2)} \\ a_{22}^{(2)} \\ a_{32}^{(2)} \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix}, \; b = \begin{pmatrix} a_{12}^{(2)} \\ r_{22} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; u_2 = b - a = \begin{pmatrix} 0 \\ a_{22}^{(2)} - r_{22} \\ a_{32}^{(2)} \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix}$$

avec $r_{22} = \pm\sqrt{(a_{22}^{(2)})^2 + \cdots + (a_{n2}^{(2)})^2}$ pour que $\|a\|_2 = \|b\|_2$.

Au bout de $(n-1)$ étapes de réduction de Householder, nous obtenons :

$$H_{n-1} \cdots H_1 A = R$$

où R est une matrice triangulaire supérieure.

On pose

$$Q^T = H_{n-1} \cdots H_2 H_1$$

soit

$$Q = H_1 H_2 \cdots H_{n-1}$$

ce qui donne la factorisation QR de A :

$$Q^T A = R \text{ ou } A = QR,$$

avec Q est une matrice orthogonale et R est une matrice triangulaire supérieure.

REMARQUE 3.3.1 *Dans le cas d'une matrice A carrée à coefficients complexes, on peut refaire les mêmes calculs, en remplaçant u^T par u^* dans la définition de la matrice de Householder, pour obtenir une factorisation $A = QR$ avec Q une matrice unitaire et R une matrice triangulaire et dans ce cas les matrices de Householder sont hermitiennes et unitaires.*

EXEMPLE
Soit

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

le calcul suivant donne la factorisation QR de la matrice A. La norme 2 de la première colonne de A est égale à $\sqrt{1+2^2+2^2} = 3$, d'où, le vecteur u_1 de la matrice H_1 est égal à :

$$u_1 = \begin{pmatrix} 1+3 \\ 2 \\ 2 \end{pmatrix}$$

ce qui donne : $\beta_1 = \frac{1}{2}\|u_1\|_2^2 = 12$ et

$$H_1 = I_3 - \frac{u_1 u_1^T}{\beta_1} = \begin{pmatrix} -1/3 & -2/3 & -2/3 \\ -2/3 & 2/3 & -1/3 \\ -2/3 & -1/3 & 2/3 \end{pmatrix}, \; H_1 A = \begin{pmatrix} -3 & -1/3 & -8/3 \\ 0 & -8/3 & -7/3 \\ 0 & -5/3 & -7/3 \end{pmatrix}$$

Pour calculer le vecteur u_2 de la matrice H_2, on doit calculer : $r_{22} = \pm\sqrt{64/9 + 25/9} = \pm\sqrt{89}/3$, d'où :

$$u_2 = \begin{pmatrix} 0 \\ -8/3 - \sqrt{89}/3 \\ -5/3 \end{pmatrix} = \begin{pmatrix} 0 \\ -5.81 \\ -1.66 \end{pmatrix}$$

ce qui donne : $\beta_2 = \frac{1}{2}\|u_2\|_2^2 = 18.27$ et

$$H_2 H_1 = (I_3 - \frac{u_2 u_2^T}{\beta_2}) H_1 = \begin{pmatrix} -0.33 & -0.66 & -0.66 \\ 0.91 & -0.38 & -0.07 \\ -0.21 & -0.63 & 0.74 \end{pmatrix}, \; R = H_2 H_1 A = \begin{pmatrix} -3 & -0.33 & -2.66 \\ 0 & 3.14 & 3.21 \\ 0 & 0 & -0.74 \end{pmatrix}$$

et la matrice

$$Q = H_1 H_2 = \begin{pmatrix} -0.33 & 0.91 & -0.21 \\ -0.66 & -0.38 & -0.63 \\ -0.66 & -0.07 & 0.74 \end{pmatrix}$$

3.4 Méthode QR

Soit $A_1 = A$ une matrice quelconque ; on écrit sa factorisation QR, soit $A_1 = Q_1 R_1$, puis on forme $A_2 = R_1 Q_1$; on écrit sa factorisation QR, soit $A_2 = Q_2 R_2$, puis on forme la matrice $A_3 = R_2 Q_2$. A l'étape k, on écrit la factorisation QR de A_k, soit $A_k = Q_k R_k$ puis on forme $A_{k+1} = R_k Q_k$ et ainsi de suite. On obtient ainsi une suite de matrices A_k qui sont toutes semblables à la matrice A, puisque :

$$\begin{array}{lcll} A_2 & = & R_1 Q_1 & = & Q_1^* A Q_1 \\ \vdots & & & & \\ A_{k+1} & = & R_k Q_k & = & Q_k^* A_k Q_k \\ & = & \cdots & = & (Q_1 Q_2 \cdots Q_k)^* A (Q_1 Q_2 \cdots Q_k). \end{array}$$

Dans la pratique, avant d'appliquer la méthode QR, on commence par mettre la matrice A sous la forme d'une matrice de Hessenberg supérieure :

$$\begin{pmatrix} \times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times & \times & \times \\ & & & & & \times & \times & \times & \times & \times \\ & & & & & & \times & \times & \times & \times \\ & & & & & & & \times & \times & \times \\ & & & & & & & & \times & \times \end{pmatrix}$$

semblable à A, en utilisant les matrices de Householder (voir exercice 3).

L'intérêt de la mise sous la forme de Hessenberg de la matrice A est que la suite des matrices A_k de la méthode QR d'une matrice de Hessenberg restent sous la forme de Hessenberg (voir exercice 4), ce qui réduit considérablement le temps de calcul. Vu la forme des matrices de Hessenberg, il suffit d'annuler les coefficients de la sous-diagonale. Si à l'étape k un élément de la sous-diagonale de la matrice A_k est numériquement nul, on partage la matrice en deux sous-matrices

$$A_k = \begin{pmatrix} \times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \times & & A_k^1 & & \times & \times & & & & \times \\ & \times & & & \times & \times & & & & \times \\ & & \times & & \times & \times & & & & \times \\ & & & \times & \times & \times & \times & \times & \times & \times \\ & & & & 0 & \times & \times & \times & \times & \times \\ & & & & & \times & & A_k^2 & & \times \\ & & & & & & \times & & & \times \\ & & & & & & & \times & & \times \\ & & & & & & & & \times & \times \end{pmatrix}$$

et il est clair que

$$spectre(A_k) = spectre(A_k^1) \cup spectre(A_k^2)$$

ce qui permet de réduire l'ordre de la matrice. On garde le bloc A_k^1 en mémoire et recommence la méthode QR avec le bloc A_k^2, ainsi de suite, jusqu'à ce que le bloc inférieur soit d'ordre deux (resp. un), dans ce cas on a déterminé deux valeurs propres (resp. une valeur propre). Puis on redémarre avec le bloc qui est juste avant.

3.5 Exercices : chapitre 3

Exercice 1 Soit

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

Déterminer la matrice semblable à A donnée par la première itération de la méthode de Jacobi.

Exercice 2 Soit

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

1. Déterminer la matrice A_1 semblable à A et qui a la forme d'une matrice de Hessenberg.
2. Déterminer la matrice A_2 semblable à A_1 donnée par la première itération de la méthode QR.

Exercice 3 (Forme de Hessenberg)

Soient $A = (a_{ij}) = A_1 = (a^1_{ij})$ une matrice carrée d'ordre n et $(e_i)_{i=1,n}$ la base canonique de $\mathbb{R}^n$.

1. Montrer qu'il existe un vecteur $u_1 \neq 0$ de $\mathbb{R}^n$ orthogonal à e_1 tel que :

$$H_1 A_1 = \begin{pmatrix} \times & \times & \cdots & \times \\ \times & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \times & \cdots & \times \end{pmatrix}$$

   avec $H_1 = H(u_1) =$ la matrice de Householder associée au vecteur u_1.
2. On pose $A_2 = H_1 A_1 H_1^T = H_1 A_1 H_1$. Montrer que A_2 a la forme :

$$A_2 = \begin{pmatrix} a^2_{11} & a^2_{12} & \cdots & \cdots & a^2_{1n} \\ a^2_{21} & a^2_{22} & \cdots & \cdots & a^2_{2n} \\ 0 & a^2_{32} & \cdots & \cdots & a^2_{3n} \\ \vdots & \vdots & & & \vdots \\ 0 & a^2_{n2} & \cdots & \cdots & a^2_{nn} \end{pmatrix}$$

3. On suppose qu'à la suite de l'étape $k-1$, on a une matrice A_k de la forme

$$A_k = \begin{pmatrix} a^k_{11} & a^k_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & a^k_{1n} \\ a^k_{21} & a^k_{22} & \cdots & \cdots & \cdots & \cdots & \cdots & a^k_{2n} \\ 0 & a^k_{32} & \ddots & \cdots & \cdots & \cdots & \cdots & a^k_{3n} \\ 0 & 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & a^k_{kk-1} & a^k_{kk} & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & a^k_{k+1k} & \cdots & \cdots & \vdots \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & a^k_{nk} & \cdots & \cdots & a^k_{nn} \end{pmatrix}$$

   Montrer qu'il existe un vecteur $u_k \neq 0$ de $\mathbb{R}^n$ orthogonal à e_i, $i = 1, k$, tel que

H_kA_k a la forme suivante :

$$H_kA_k = \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots & \cdots & \cdots & b_{2n} \\ 0 & b_{32} & \cdots & \cdots & \cdots & \cdots & \cdots & b_{3n} \\ 0 & 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & b_{k+1k} & b_{k+1k+1} & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & b_{k+2k+1} & \cdots & \cdots & \vdots \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & b_{nk+1} & \cdots & \cdots & b_{nn} \end{pmatrix}$$

avec $H_k = H(u_k)$ = la matrice de Householder associée au vecteur u_k.

4. On pose $A_{k+1} = H_kA_kH_k^T = H_kA_kH_k$. Montrer que A_{k+1} a la forme suivante :

$$A_{k+1} = \begin{pmatrix} a_{11}^{k+1} & a_{12}^{k+1} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1n}^{k+1} \\ a_{21}^{k+1} & a_{22}^{k+1} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{2n}^{k+1} \\ 0 & a_{32}^{k+1} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{3n}^{k+1} \\ 0 & 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & a_{k+1k}^{k+1} & a_{k+1k+1}^{k+1} & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{k+2k+1}^{k+1} & \cdots & \cdots & \vdots \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & a_{nk+1}^{k+1} & \cdots & \cdots & a_{nn}^{k+1} \end{pmatrix}$$

5. Déduire que toute matrice carrée A est semblable à une matrice qui a la forme d'une matrice de Hessenberg :

$$HS = \begin{pmatrix} \times & \times & \cdots & \cdots & \times \\ \times & \times & \cdots & \cdots & \times \\ 0 & \times & \ddots & & \times \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \times & \times \end{pmatrix}$$

c'est à dire pour tout $j = 1, n-2$, le vecteur $HSe_j \in < e_1, e_2, \cdots, e_{j+1} >$ l'espace engendré par $e_1, e_2, \cdots, e_{j+1}$.

6. Montrer que si la matrice A est symétrique alors la matrice HS semblable à A obtenue par la méthode ci-dessus est symétrique. Déduire que toute matrice symétrique est semblable à une matrice tridiagonale.

Exercice 4

Soit A une matrice carrée d'ordre n qui a la forme d'une matrice de Hessenberg (voir exercice précédent). Soient $H_1, H_2, \cdots, H_{n-1}$ les matrices de Householder de la factorisation QR de la matrice A, c'est à dire : $Q = H_1H_2 \cdots H_{n-1}$=matrice unitaire et $R = Q^*A$=matrice triangulaire.

1. Montrer que pour tout $k = 1, 2, \cdots, n-1$, on a :

$$\begin{cases} H_ke_k \in < e_k, e_{k+1} > \\ H_ke_{k+1} \in < e_k, e_{k+1} > \\ \forall i \neq k, i \neq k+1, \ H_ke_i = e_i \end{cases}$$

2. Montrer que pour tout $k = 1, 2, \cdots, n-1$:
   (a) $Qe_k = H_1 H_2 \cdots H_k e_k$
   (b) $Qe_k \in < e_1, e_2, \cdots, e_{k+1} >$
3. Déduire que la matrice $B = RQ$ a la forme d'une matrice de Hessenberg.
4. Conclusion : Montrer que si une matrice A a la forme d'une matrice de Hessenberg, alors, les matrices de la suite (A_k) de la méthode QR ont la forme d'une matrice de Hessenberg.

3.6 Corrigé des exercices : chapitre 3

Réponse 1 On trouve que $(p, q) = (1, 3)$, d'où : $R = -1/6$, ce qui donne le trinôme :

$$t^2 + 2Rt - 1 = t^2 - t/3 - 1 = 0$$

l'unique solution qui appartient à $]-1, 0[\cup]0, 1]$ est égale à $t = (1 - \sqrt{37})/6 \approx -0.847$, ce qui donne :

$$s = t/\sqrt{1+t^2} \approx -0.646, \quad c = 1/\sqrt{1+t^2} \approx 0.76$$

par conséquent, si on note par $B = (b_{ij})$ la matrice donnée par la première itération de la méthode de Jacobi classique, on obtient :

$$\begin{array}{ll}
b_{12} = b_{21} & = c \times a_{12} - s \times a_{32} \approx 0.76 \times (-1) - (-0.646) \times 2 \\
b_{14} = b_{41} & = c \times a_{14} - s \times a_{34} \approx 0.76 \times (1) - (-0.646) \times (-1) \\
b_{32} = b_{23} & = s \times a_{12} + c \times a_{32} \approx (-0.646) \times (-1) + (0.76) \times 2 \\
b_{34} = b_{43} & = s \times a_{14} + c \times a_{34} \approx (-0.646) \times (1) + (0.76) \times (-1) \\
b_{22} = a_{22} & = 3 \\
b_{44} = a_{44} & = 2 \\
b_{31} = b_{13} & = 0 \\
b_{11} & = a_{11} - t \times a_{13} \approx 1 - (-0.847) \times 3 \\
b_{33} & = a_{33} + t \times a_{13} \approx 2 - (-0.847) \times 3
\end{array}$$

Réponse 2

1. On trouve :
$$A_1 \approx \begin{pmatrix} 2 & 2.45 & 1.22 & 1.22 \\ 2.45 & -0.33 & -3.96 & -0.7 \\ 0 & 2.09 & -2.65 & 0.005 \\ 0 & 0 & 1.39 & 2.5 \end{pmatrix}$$
2. On trouve :
$$R_1 \approx \begin{pmatrix} -3.16 & -1.29 & -2.3 & 0.2 \\ 0 & 2.97 & -0.38 & 1.08 \\ 0 & 0 & -4.16 & 2.31 \\ 0 & 0 & 0 & 2.27 \end{pmatrix}$$
$$Q_1 \approx \begin{pmatrix} -0.63 & 0.55 & 0.53 & -0.11 \\ -0.77 & -0.45 & -0.43 & 0.09 \\ 0 & 0.7 & -0.69 & 0.15 \\ 0 & 0 & 0.21 & 0.97 \end{pmatrix}$$
$$A_2 = R_1 Q_1 = \begin{pmatrix} 3 & -2.77 & 0.51 & 0.09 \\ -2.3 & -1.6 & -0.79 & 1.28 \\ 0 & -2.93 & 3.38 & 1.62 \\ 0 & 0 & 0.49 & 2.22 \end{pmatrix}$$

Réponse 3 (Forme de Hessenberg)

1. On veut que $H_1 \begin{pmatrix} a^1_{11} \\ a^1_{21} \\ a^1_{31} \\ \vdots \\ a^1_{n1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ (voir cours) il suffit que $\| \begin{pmatrix} a^1_{11} \\ a^1_{21} \\ a^1_{31} \\ \vdots \\ a^1_{n1} \end{pmatrix} \|_2 = \| \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \|_2$ et $u_1 = \begin{pmatrix} a^1_{11} \\ a^1_{21} \\ a^1_{31} \\ \vdots \\ a^1_{n1} \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

   Pour que u_1 soit orthogonal à e_1, il faut que $x_1 = a^1_{11}$ et pour avoir la même norme 2, il faut que

   $$\sum_{i=1}^{n} (a^1_{i1})^2 = (a^1_{11})^2 + (x_2)^2$$

   d'où

   $$x_2 = \pm \sqrt{\sum_{i=2}^{n} (a^1_{i1})^2}$$

   ce qui donne $u_1 = \begin{pmatrix} 0 \\ a^1_{21} \pm \sqrt{\sum_{i=2}^{n} (a^1_{i1})^2} \\ a^1_{31} \\ \vdots \\ a^1_{n1} \end{pmatrix}$.

2. $A_2 e_1 = H_1 A_1 H_1 e_1 = H_1 A_1 e_1 = H_1 \begin{pmatrix} a^1_{11} \\ a^1_{21} \\ a^1_{31} \\ \vdots \\ a^1_{n1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a^2_{11} \\ a^2_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ car

   $H_1 e_1 = e_1$, $(u_1 \perp e_1)$.

3. Même raisonnement que la première étape, on cherche

   $$u_k \perp e_i, i = 1, 2, \cdots, k \quad / \quad H_k \begin{pmatrix} a^k_{k1} \\ \vdots \\ a^k_{kk} \\ a^k_{k+1k} \\ \vdots \\ a^k_{n,k} \end{pmatrix} = \begin{pmatrix} b_{k1} \\ \vdots \\ b_{kk} \\ b_{k+1k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

   on trouve :

   $$u_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a^k_{k+1k} \pm \sqrt{\sum_{i=k+1}^{n} (a^k_{ik})^2} \\ a^k_{k+2k} \\ \vdots \\ a^k_{nk} \end{pmatrix}$$

On a $H_k e_i = e_i$, $i = 1, 2, \cdots, k$, d'où :

$$\forall x \in < e_1, e_2, \cdots, e_k >,\ H_k x = x$$

Par conséquent :

$$\forall i = 1, 2, \cdots, k-1, H_k A_k e_i = A_k e_i$$

et

$$H_k A_k e_k = H_k \begin{pmatrix} a^k_{1k} \\ \vdots \\ a^k_{kk} \\ a^k_{k+1k} \\ \vdots \\ a^k_{n,k} \end{pmatrix} = \begin{pmatrix} b_{1k} \\ \vdots \\ b_{kk} \\ b_{k+1k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

4. $A_{k+1} e_i = H_k A_k H_k e_i = H_k A_k e_i = A_k e_i$, $i = 1, 2, \cdots, k-1$ et
$$A_{k+1} e_k = H_k A_k H_k e_k = H_k A_k e_k = \begin{pmatrix} b_{1k} \\ \vdots \\ b_{kk} \\ b_{k+1k} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a^{k+1}_{1k} \\ \vdots \\ a^{k+1}_{kk} \\ a^{k+1}_{k+1k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$
5. A la suite de l'étape $n-2$, on obtient la forme HS de Hessenberg.
6. Si A est symétrique, alors A_k est symétrique pour tout $k = 1, 2, \cdots, n-2$. Donc, $HS = A_{n-2}$ est symétrique et il est clair qu'une matrice symétrique qui a la forme de Hessenberg est nécessairement tridiagonale.

Réponse 4

1. Par récurrence sur k. Pour $k = 1$, on a $u_1 = \begin{pmatrix} a^1_{11} \pm \sqrt{\sum_{i=1}^{2} (a^1_{i1})^2} \\ a^1_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ car $A_1 = A$
a la forme d'une matrice de Hessenberg, d'où, $H_1 e_1 \in < e_1, e_2 >$ car $u_1 \in < e_1, e_2 >$ et on a bien : $H_1 e_i = e_i$ pour tout $i \geq 3$. Supposons que le résultat est vrai pour $1 \leq k \leq l-1$ et montrons qu'il est vrai pour $k = l$. On a :

$$\begin{aligned} A_l e_l &= H_{l-1} H_{l-2} \cdots H_1 A H_1 \cdots H_{l-2} H_{l-1} e_l \\ &\in H_{l-1} H_{l-2} \cdots H_1 A H_1 \cdots H_{l-2} < e_{l-1}, e_l > \\ &\vdots \\ &\in H_{l-1} H_{l-2} \cdots H_1 A < e_1, e_2, \cdots, e_l > \\ &\in H_{l-1} H_{l-2} \cdots H_1 < e_1, e_2, \cdots, e_{l+1} > \\ &\in < e_1, e_2, \cdots, e_{l+1} > \end{aligned}$$

Donc, $u_l \in < e_l, e_{l+1} >$ et par conséquent H_l vérifie :

$$\begin{aligned} H_l e_l &= e_l - \tfrac{(u_l^T e_l)}{\beta} u_l \in < e_l, e_{l+1} > \\ H_l e_{l+1} &= e_{l+1} - \tfrac{(u_l^T e_{l+1})}{\beta} u_l \in < e_l, e_{l+1} > \\ H_l e_i &= e_i - \tfrac{(u_l^T e_i)}{\beta} u_l = e_i \quad \forall i \neq l, \forall i \neq l+1 \end{aligned}$$

2. a) D'après (1) $H_l e_k = e_k$ pour tout $l \geq k+1$, d'où : $Qe_k = H_1 H_2 \cdots H_{n-1} e_k = H_1 H_2 \cdots H_k e_k$
   b)
   $$\begin{aligned} Qe_k &= H_1 H_2 \cdots H_k e_k \\ &\in H_1 H_2 \cdots H_{k-1} < e_k, e_{k+1} > \\ &\in < e_1, e_2, \cdots, e_{k+1} > \end{aligned}$$

3. La matrice R est triangulaire supérieure, d'où :
   $$\forall l = 1, \cdots, n,\ R < e_1, e_2, \cdots, e_l > \subset < e_1, e_2, \cdots, e_l >$$
   ce qui donne :
   $$\forall k = 1, \cdots, n-2,\ RQ < e_1, \cdots, e_k > \subset R < e_1, \cdots, e_{k+1} > \subset < e_1, \cdots, e_{k+1} >$$
   ce qui prouve que $B = RQ$ est une matrice de Hessenberg.

4. A chaque étape k de la méthode QR, on applique la factorisation QR à $A_k = Q_k R_k$ et on pose $A_{k+1} = R_k Q_k$. D'après les questions précédentes, si A_k a la forme d'une matrice de Hessenberg alors $R_k Q_k = A_{k+1}$ a aussi la forme d'une matrice de Hessenberg. Comme $A_1 = A$ a la forme d'une matrice de Hessenberg, un raisonnement par récurrence nous donne ce qu'il faut.

Chapitre 4

Méthodes d'interpolation et d'approximation

4.1 Formule d'interpolation de Lagrange

Soient (x_i, f_i),$i = 0, \cdots, n$, avec $x_i \neq x_j$ pour $i \neq j$, $(n+1)$-couples réels ou complexes. Le problème d'interpolation polynomiale consiste à trouver un polynôme $P(x)$ de degré inférieur ou égal à n tel que :

$$P(x_i) = f_i,\ i = 0, \cdots, n \tag{4.1}$$

THÉORÈME 4.1.1 *Formule d'interpolation de Lagrange*

Soient (x_i, f_i),$i = 0, \cdots, n$, avec $x_i \neq x_j$ pour $i \neq j$, $(n+1)$-couples réels ou complexes alors, il existe un unique polynôme $P(x)$ de degré inférieur ou égal à n vérifiant (4.1).

DÉMONSTRATION: *Unicité :* Supposons l'existence de deux polynômes $P_1(x)$ et $P_2(x)$ de degré inférieur ou égal à n et vérifiant (4.1). Par conséquent $P_1(x) - P_2(x)$ est un polynôme de degré inférieur ou égal à n et possède $(n+1)$-racines ce qui donne $P_1 - P_2 \equiv 0$.

Existence : On pose

$$\begin{aligned} L_i(x) &= \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} \\ &= \frac{v(x)}{(x-x_i)v'(x_i)} \end{aligned}$$

avec

$$v(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$$

Le polynôme

$$P(x) = \sum_{i=0}^{n} f_i L_i(x)$$

est un polynôme de degré inférieur ou égal à n et vérifie les conditions (4.1).

EXEMPLE: On cherche un polynôme de degré inférieur ou égal à 2 tel que :

$$P(0) = 1,\ P(1) = -1,\ P(4) = 5$$

D'après le théorème précédent, il y a un seul polynôme donné par :

$$P(x) = 1 \times L_0(x) + (-1) \times L_1(x) + 5 \times L_2(x)$$

avec

$$\begin{aligned} L_0(x) &= \frac{(x-1)(x-4)}{(0-1)(0-4)} \\ L_1(x) &= \frac{(x-0)(x-4)}{(1-0)(1-4)} \\ L_2(x) &= \frac{(x-0)(x-1)}{(4-0)(4-1)} \end{aligned}$$

d'où

$$P(x) = x^2 - 3x + 1$$

En général les couples (x_i, f_i), $i = 0, \cdots, n$ sont les résultats d'une expérience. Par conséquent les nombres f_i, $i = 0, \cdots, n$ sont les valeurs d'une fonction f aux points x_i, $i = 0, \cdots, n$ et le polynôme donné par la formule d'interpolation de Lagrange coïncide avec cette fonction aux points x_i, $i = 0, \cdots, n$. Donc la formule d'interpolation de Lagrange est un moyen pour approcher une fonction f lorsque on connait les valeurs prises par f aux points x_i, $i = 0, \cdots, n$. La question qui se pose maintenant est de savoir estimer l'erreur commise en approchant $f(x)$ par $P(x)$.

THÉORÈME 4.1.2 *Si f est $(n+1)-$fois dérivable, alors, pour tout $\bar{x} \in \mathbb{R}$, il existe $c \in I$ (I est le plus petit intervalle contenant $x_0, x_1, \cdots, x_n, \bar{x}$) tel que :*

$$f(\bar{x}) - P(\bar{x}) = v(\bar{x}) \frac{f^{(n+1)}(c)}{(n+1)!} \tag{4.2}$$

avec $v(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$.

DÉMONSTRATION: Si $\bar{x}$ est égal à l'un des x_i, il n'y a rien à démontrer. Supposons que $\bar{x} \neq x_i$, $i = 0, \cdots, n$. Soit $K \in \mathbb{R}$ tel que :

$$F(x) = f(x) - P(x) - Kv(x)$$

vérifie $F(\bar{x}) = 0$. Donc, la fonction $F(x)$ possède $(n+2)-$zéros dans l'intervalle I. En appliquant le théorème de Rolle à $F(x)$ puis à $F'(x)$ puis à $F"(x)$ $\cdots$, on obtient que $F^{(n+1)}(x)$ possède au moins un zéro dans l'intervalle I, d'où :

$$\exists c \in I / F^{(n+1)}(c) = 0$$

or $P^{(n+1)}(x) = 0$ et $v^{(n+1)}(x) = (n+1)!$, ce qui donne :

$$f^{(n+1)}(c) - K(n+1)! = 0$$

d'où

$$K = \frac{f^{(n+1)}(c)}{(n+1)!}$$

En remplacant la valeur de K dans l'expression de $F(x)$ et en utilisant le fait que $F(\bar{x}) = 0$, on obtient le résultat donné par le théorème.

4.2 Algorithme de Neville

L'algorithme de Neville est une méthode très pratique pour le calcul sur ordinateur de $P(\bar{x})$, $\bar{x} \in \mathbb{R}$ fixé, où $P(x)$ est le polynôme d'interpolation de Lagrange qui vérifie (4.1).

Pour tout $0 \leq i \leq j \leq n$, on note par $P_{i,j}(x)$ le polynôme d'interpolation de Lagrange de degré inférieur ou égal à $j - i$ tel que :

$$P_{i,j}(x_k) = f_k, \quad i \leq k \leq j \tag{4.3}$$

En particulier le polynôme $P_{i,i}$ n'est autre que la constante f_i et le polynôme $P_{0,n}(x) = P(x)$.

PROPOSITION 4.2.1 *Pour tout $0 \leq i < j \leq n$, on a :*

$$P_{i,j}(x) = \frac{(x - x_i)P_{i+1,j}(x) - (x - x_j)P_{i,j-1}(x)}{x_j - x_i}$$

DÉMONSTRATION: Soit $0 \leq i < j \leq n$, d'après l'unicité du polynôme d'interpolation de Lagrange, il suffit de prouver que le polynôme :

$$R(x) = \frac{(x - x_i)P_{i+1,j}(x) - (x - x_j)P_{i,j-1}(x)}{x_j - x_i}$$

est un polynôme de degré inférieur ou égal à $j - i$ et vérifie (4.3). Le degré de $P_{i+1,j}(x)$ est inférieur ou égal à $j - i - 1$ et le degré de $P_{i,j-1}(x)$ est inférieur ou égal à $j - i - 1$, ce qui donne que le degré de $R(x)$ est inférieur ou égal à $j - i$. D'autre part, par définition on a :

$$P_{i+1,j}(x_k) = P_{i,j-1}(x_k) = f_k \quad \forall i + 1 \leq k \leq j - 1$$

et

$$P_{i+1,j}(x_j) = f_j, \quad P_{i,j-1}(x_i) = f_i$$

d'où

$$\begin{aligned} R(x_k) &= f_k \quad \forall i + 1 \leq k \leq j - 1 \\ R(x_i) &= \frac{-(x_i - x_j)P_{i,j-1}(x_i)}{x_j - x_i} = f_i \\ R(x_j) &= \frac{(x_j - x_i)P_{i+1,j}(x_j)}{x_j - x_i} = f_j \end{aligned}$$

Soit $\bar{x} \in \mathbb{R}$ fixé. Pour calculer $P(\bar{x}) = P_{o,n}(\bar{x})$, l'algorithme de Neville est constitué de n étapes :

Etape 1 : On calcule tous les $P_{i,i+1}(\bar{x})$, $i = 0, \cdots, n - 1$ par la formule

$$P_{i,i+1}(\bar{x}) = \frac{(\bar{x} - x_i)f_{i+1} - (\bar{x} - x_{i+1})f_i}{x_{i+1} - x_i}$$

Etape 2 : On calcule tous les $P_{i,i+2}(\bar{x})$, $i = 0, \cdots, n - 2$ par la formule

$$P_{i,i+2}(\bar{x}) = \frac{(\bar{x} - x_i)P_{i+1,i+2}(\bar{x}) - (\bar{x} - x_{i+2})P_{i,i+1}(\bar{x})}{x_{i+2} - x_i}$$

$\vdots$

Etape $n-1$: On calcule tous les $P_{i,i+n-1}(\bar{x})$, $i = 0, 1$ (seulement deux valeurs à calculer) par la formule

$$P_{i,i+n-1}(\bar{x}) = \frac{(\bar{x} - x_i)P_{i+1,i+n-1(\bar{x})} - (\bar{x} - x_{i+n-1})P_{i,i+n-2}(\bar{x})}{x_{i+n-1} - x_i}$$

Etape n : Une seule valeur à calculer $P_{i,i+n}(\bar{x})$, $i = 0$ par la formule

$$P_{0,n}(\bar{x}) = \frac{(\bar{x} - x_0)P_{1,n}(\bar{x}) - (\bar{x} - x_n)P_{0,n-1}(\bar{x})}{x_n - x_0}$$

On peut résumer ces étapes dans un tableau, par exemple dans le cas $n = 3$

$$\begin{array}{cccccccc}
 & 0 & & 1 & & 2 & & 3 \\
x_0 & f_0 & & & & & & \\
 & & > & P_{0,1}(\bar{x}) & & & & \\
x_1 & f_1 & & & > & P_{0,2}(\bar{x}) & & \\
 & & > & P_{1,2}(\bar{x}) & & & > & P_{0,3}(\bar{x}) = P(\bar{x}) \\
x_2 & f_2 & & & > & P_{1,3}(\bar{x}) & & \\
 & & > & P_{2,3}(\bar{x}) & & & & \\
x_3 & f_3 & & & & & &
\end{array}$$

EXEMPLE: Soit le polynôme $P(x) = x^3 - 2x^2 + x + 1$. On veut calculer $P(-1)$ par l'algorithme de Neville en utilisant $f_0 = P(0) = 1$, $f_1 = P(1) = 1$, $f_2 = P(2) = 3$, $f_3 = P(3) = 13$:

$$\begin{array}{cccccccc}
 & 0 & & 1 & & 2 & & 3 \\
0 & 1 & & & & & & \\
 & & > & 1 & & & & \\
1 & 1 & & & > & 3 & & \\
 & & > & -3 & & & > & -3 = P(-1) \\
2 & 3 & & & > & 21 & & \\
 & & > & -27 & & & & \\
3 & 13 & & & & & &
\end{array}$$

4.3 Méthode des différences divisées

L'algorithme de Neville permet de calculer $P(x)$ pour une valeur donnée $x = \bar{x}$. Si on veut calculer $P(x)$ pour plusieurs valeurs de x, on doit appliquer l'algorithme de Neville plusieurs fois. Dans ce cas, il vaut mieux calculer toute la fonction polynôme $P(x)$.

Soient $(n+1)$-couples réels ou complexes (x_i, f_i),$i = 0, \cdots, n$, avec $x_i \neq x_j$ pour $i \neq j$ et soit $P(x)$ le polynôme de degré inférieur ou égal à n d'interpolation de Lagrange tels que :

$$P(x_i) = f_i, \quad i = 0, \cdots, n$$

REMARQUE 4.3.1 *Il est facile de voir que la famille :* $1, (x - x_0)$,$(x - x_0)(x - x_1)$,$\cdots$,$(x - x_0)(x - x_1) \cdots (x - x_{n-1})$ *est une base de l'espace vectoriel*

des polynômes de degré inférieur ou égal à n, même sans l'hypothèse : $x_i \neq x_j$ pour $i \neq j$. En effet, le nombre de vecteurs est égal à la dimension de l'espace vectoriel, il suffit donc de prouver que cette famille de vecteurs est libre. Soient $\alpha_0,\ \alpha_1, \cdots, \alpha_n$, $(n+1)$-réels tels que :

$$\alpha_0 + \alpha_1(x - x_0) + \cdots + \alpha_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \equiv 0$$

C'est une égalité entre polynômes, donc, elle est vraie pour tout $x \in \mathbb{R}$. En particulier, si on remplace $x = x_0$ on obtient que $\alpha_0 = 0$ puis on divise par $(x - x_0)$ la nouvelle égalité et on remplace $x = x_1$ on obtient que $\alpha_1 = 0$, etc..., jusqu'à $\alpha_n = 0$, ce qui prouve que la famille en question est libre.

Par conséquent tout polynôme de degré inférieur ou égal à n s'écrit :

$$\begin{aligned} P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots \\ + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned}$$

La méthode des différences divisées permet de calculer les coefficients a_k, $k = 0, \cdots, n$ du polynôme $P(x)$ avec le minimum d'opérations et pour calculer $P(x)$ pour une valeur $x = \bar{x}$ donnée, on utilise le schéma de Horner :

$$\begin{aligned} P(\bar{x}) = (((\cdots(a_n(\bar{x} - x_{n-1}) + a_{n-1})(\bar{x} - x_{n-2}) \\ + a_{n-2})\cdots) + a_1)(\bar{x} - x_0) + a_0 \end{aligned}$$

On rappelle que, pour $0 \leq i \leq j \leq n$, $P_{i,j}(x)$ est le polynôme de degré inférieur ou égal à $j - i$ d'interpolation de Lagrange qui vérifie :

$$P_{i,j}(x_k) = f_k, \quad k = i, \cdots, j$$

THÉORÈME 4.3.1 *Pour tout $0 \leq i \leq j \leq n$, le polynôme $P_{i,j}(x)$ est égal à :*

$$\begin{aligned} P_{i,j}(x) = c_{i,i} + c_{i,i+1}(x - x_i) + c_{i,i+2}(x - x_i)(x - x_{i+1}) + \cdots \\ + c_{i,j}(x - x_i)(x - x_{i+1})\cdots(x - x_{j-1}) \end{aligned}$$

où les coefficients $c_{i,j}$ sont définis par

$$\begin{aligned} c_{i,i} &= f_i, \quad i = 0, \cdots, n \\ c_{i,j} &= \frac{c_{i+1,j} - c_{i,j-1}}{x_j - x_i}, \quad 0 \leq i < j \leq n \end{aligned}$$

En particulier

$$\begin{aligned} P(x) = P_{0,n}(x) = c_{0,0} + c_{0,1}(x - x_0) + c_{0,2}(x - x_0)(x - x_1) + \\ \cdots + c_{0,n}(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned}$$

DÉMONSTRATION: D'après la Proposition 4.2.1, on a :

$$P_{i,i}(x) = f_i, \quad i = 0, \cdots, n$$

et pour tout $0 \leq i < j \leq n$

$$P_{i,j}(x) = \frac{(x - x_i)P_{i+1,j}(x) - (x - x_j)P_{i,j-1}(x)}{x_j - x_i} \tag{4.4}$$

d'où pour tout $0 \leq i < j \leq n$

$$\begin{array}{rcl} P_{i,j}(x) & = & \dfrac{x-x_i}{x_j-x_i}P_{i+1,j}(x) - \dfrac{x-x_i}{x_j-x_i}P_{i,j-1}(x) + P_{i,j-1}(x) \\ & = & P_{i,j-1}(x) + R_{i,j}(x) \end{array} \tag{4.5}$$

avec

$$R_{i,j}(x) = \frac{x-x_i}{x_j-x_i}(P_{i+1,j}(x) - P_{i,j-1}(x)) \tag{4.6}$$

On vérifie facilement que le degré de $R_{i,j}(x)$ est inférieur ou égal à $j-i$ et que

$$R_{i,j}(x_i) = R_{i,j}(x_{i+1}) = \cdots = R_{i,j}(x_{j-1}) = 0$$

d'où

$$R_{i,j}(x) = c_{i,j}(x-x_i)(x-x_{i+1})\cdots(x-x_{j-1}) \tag{4.7}$$

ce qui donne

$$P_{i,j}(x) = P_{i,j-1}(x) + c_{i,j}(x-x_i)(x-x_{i+1})\cdots(x-x_{j-1}) \tag{4.8}$$

On applique le même raisonnement à $P_{i,j-1}(x)$ puis à $P_{i,j-2}(x)$, etc..., on obtient :

$$\begin{array}{r} P_{i,j}(x) = c_{i,i} + c_{i,i+1}(x-x_i) + c_{i,i+2}(x-x_i)(x-x_{i+1}) + \cdots \\ +c_{i,j}(x-x_i)(x-x_{i+1})\cdots(x-x_{j-1}) \end{array}$$

Il est clair que $c_{i,i} = f_i$ (il suffit de remplacer x par x_i). D'après les formules (4.5) et (4.7), le coefficient du monôme x^{j-i} du polynôme $P_{i,j}(x)$ est égal à $c_{i,j}$ car le polynôme $P_{i,j-1}(x)$ est de degré inférieur ou égal à $j-i-1$. D'après la formule (4.4), le coefficient du monôme x^{j-i} du polynôme $P_{i,j}(x)$ est égal à

$$\frac{c_{i+1,j} - c_{i,j-1}}{x_j - x_i}$$

d'où la relation, dite des différences divisées, entre les coefficients $c_{i,j}$.

Dans la pratique, on utilise une méthode analogue à la méthode de Neville pour calculer les différences divisées :

$$\begin{array}{cccccccc} & 0 & & 1 & & 2 & \cdots & n \\ x_0 & \underline{f_0 = c_{0,0}} & & & & & & \\ & & > & \underline{c_{0,1}} & & & & \\ x_1 & f_1 & & & > & \underline{c_{0,2}} & & \\ & & > & c_{1,2} & & & \searrow & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \underline{c_{0,n}} \\ & & > & c_{n-2,n-1} & & & & \\ x_{n-1} & f_{n-1} & & & > & c_{n-2,n} & \nearrow & \\ & & > & c_{n-1,n} & & & & \\ x_n & f_n & & & & & & \end{array}$$

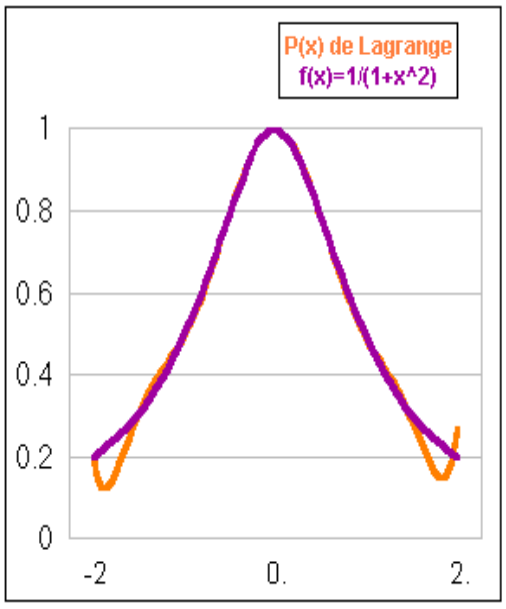
Donc, on lit les coefficients de $P(x)$ sur la diagonale supérieure.

EXEMPLE: On cherche le polynôme d'interpolation de Lagrange par la méthode des différences divisées pour approcher la fonction

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

sur l'intervalle $[-2, 2]$ en utilisant les couples : $(-2, f(-2))$, $(-1.5, f(-1.5))$, $(-1, f(-1))$, $(-0.5, f(-0.5))$, $(0, f(0))$, $(0.5, f(0.5))$, $(1, f(1))$, $(1.5, f(1.5))$, $(2, f(2))$.

	0	1	2	3	4	5	6	7	8
-2.0	0.200								
		0.215							
-1.5	0.308		0.169						
		0.385		0.031					
-1.0	0.500		0.215		-0.154				
		0.600		-0.277		0.037			
-0.5	0.800		-0.200		-0.062		0.049		
		0.400		-0.400		0.185		-0.049	
0.0	1.000		-0.800		0.400		-0.123		0.025
		0.400		0.400		-0.185		0.049	
0.5	0.800		-0.200		-0.062		0.049		
		-0.600		0.277		-0.037			
1.0	0.500		0.215		-0.154				
		-0.385		-0.031					
1.5	0.308		0.169						
		-0.215							
2.0	0.200								



Le théorème suivant sera utilisé dans le prochain paragraphe pour généraliser la méthode des différences divisées.

THÉORÈME 4.3.2 *Soit $f : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction suffisamment dérivable telle que : $f_k = f(x_k)$, $0 \leq k \leq n$. Pour tout $0 \leq i < j \leq n$, on a :*

$$\exists c \in I / c_{i,j} = \frac{f^{(j-i)}(c)}{(j-i)!}$$

où I est le plus petit intervalle de $\mathbb{R}$ contenant $x_i,\ x_{i+1}, \cdots, x_j$ et les $c_{i,j}$ sont données par le théorème 4.3.1.

DÉMONSTRATION: On note par $v_{i,j}(x) = (x - x_i) \cdots (x - x_{j-1})$. D'après le théorème 4.1.2 :

$$\forall \bar{x} \in \mathbb{R}, \exists c \in I / f(\bar{x}) = P_{i,j-1}(\bar{x}) + v_{i,j}(\bar{x}) \frac{f^{(j-i)}(c)}{(j-i)!}$$

avec I est le plus petit intervalle de $\mathbb{R}$ contenant $\bar{x}$, x_i,$\cdots, x_{j-1}$. En particulier pour $\bar{x} = x_j$, on obtient :

$$\exists c \in I / f(x_j) = P_{i,j-1}(x_j) + v_{i,j}(x_j) \frac{f^{(j-i)}(c)}{(j-i)!}$$

or, par définition du polynôme P_{ij}, on a :

$$P_{ij}(x_j) = f_j = f(x_j)$$

et d'après la formule (4.8) :

$$P_{ij}(x_j) = P_{ij-1}(x_j) + c_{ij} v_{ij}(x_j)$$

d'où :

$$c_{ij} v_{ij}(x_j) = v_{ij}(x_j) \frac{f^{(j-i)}(c)}{(j-i)!}$$

or $v_{ij}(x_j) \neq 0$, ce qui prouve le théorème.

4.4 Interpolation de Hermite

Soient f_i^k, α_i, $i = 0, 1, \cdots, m$ et $k = 0, 1, \cdots, n_i - 1$, des nombres réels tels que :

$$\alpha_0 < \alpha_1 < \cdots < \alpha_m$$

Le problème d'interpolation de Hermite consiste à trouver un polynôme $P(x)$ de degré inférieur ou égal à $n = -1 + \sum_{i=0}^{m} n_i$ qui vérifie :

$$P^{(k)}(\alpha_i) = f_i^k \quad \forall 0 \leq i \leq m, \forall 0 \leq k \leq n_i - 1 \tag{4.9}$$

THÉORÈME 4.4.1 *Il existe un unique polynôme de degré inférieur ou égal à $n = -1 + \sum_{i=0}^{m} n_i$ qui vérifie (4.9)*

DÉMONSTRATION: On note par $\mathbb{R}_n[X]$ l'espace vectoriel des polynômes de degré inférieur ou égal à n. Soit $\mathcal{L}$ l'application linéaire de $\mathbb{R}_n[X]$ dans $\mathbb{R}^{n+1}$ définie par :

$$\mathcal{L}(P) = (P(\alpha_0), P'(\alpha_0), \cdots, P^{(n_0-1)}(\alpha_0), \cdots, P(\alpha_m), \cdots, P^{(n_m-1)}(\alpha_m))$$

Montrons que $\mathcal{L}$ est injective. Soit $P(x)$ un polynôme de degré inférieur ou égal à n tel que $\mathcal{L}(P) = 0$, alors le polynôme $P(x)$ vérifie :

$$\forall 0 \leq k \leq n_i - 1,\ P^{(k)}(\alpha_i) = 0 \quad \forall 0 \leq i \leq m$$

Par conséquent, pour tout $i = 0, 1, \cdots, m$, α_i est une racine de $P(x)$ de multiplicité n_i. Si on compte les racines avec leurs multiplicités, on obtient $n + 1$ racines , or $P(x)$ est de degré inférieur ou égal à n, donc, nécessairement $P \equiv 0$. D'autre part la dimension de l'espace vectoriel $\mathbb{R}_n[X]$ est égale à $n + 1$, par conséquent, $\mathcal{L}$ est bijective, d'où l'existence et l'unicité de $P \in \mathbb{R}_n[X]$ qui vérifie (4.9).

Si on suppose qu'il existe une fonction $(n+1)$-dérivable f telle que pour tout $i = 0, \cdots, m$ et tout $k = 0, \cdots, n_i$ le nombre $f_i^k = f^{(k)}(\alpha_i)$, alors, le théorème suivant permet d'estimer l'erreur commise en approchant $f(x)$ par le polynôme de Hemite $P(x)$:

THÉORÈME 4.4.2 *Pour tout $x \in [a, b]$ il existe un unique $\xi \in I$ tel que :*

$$f(x) = P(x) + \frac{v(x) f^{(n+1)}(\xi)}{(n+1)!}$$

avec $v(x) = (x - \alpha_0)^{n_0}(x - \alpha_1)^{n_1} \cdots (x - \alpha_m)^{n_m}$ et I est le plus petit intervalle contenant $x,\ \alpha_0,\ \alpha_1,\ \cdots,\ \alpha_m$.

La démonstration du théorème précédent est exactement la même que la démonstration du théorème 4.1.2 du même chapitre.

Dans ce qui suit, on se propose de généraliser la méthode des différences divisées pour déterminer le polynôme d'interpolation de Hermite. Pour cela, on écrit les couples (α_i, f_i^k), $i = 0, \cdots, m$, $k = 0, \cdots, n_i - 1$ sous la forme :

$$(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$$

avec

$$\begin{array}{rcll}
x_i & = & \alpha_0, & f_i = f_0^i \quad i = 0, 1, \cdots, n_0 - 1 \\
x_{i+n_0} & = & \alpha_1, & f_{i+n_0} = f_1^i \quad i = 0, 1, \cdots, n_1 - 1 \\
x_{i+n_0+n_1} & = & \alpha_2, & f_{i+n_0+n_1} = f_2^i \quad i = 0, 1, \cdots, n_2 - 1 \\
 & \vdots & & \\
x_{i+n-n_m+1} & = & \alpha_m, & f_{i+n-n_m+1} = f_m^i \quad i = 0, 1, \cdots, n_m - 1
\end{array}$$

REMARQUE 4.4.1 *Pour tout $0 \leq i \leq n$, il existe un entier $0 \leq r(i) \leq m$ tel que : $x_i = \alpha_{r(i)}$.*

En utilisant la remarque 4.3.1, on écrit le polynôme d'interpolation de Hermite $P(x)$, associé aux couples (α_i, f_i^k) sous la forme :

$$\begin{aligned}
P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots \\
+ a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1})
\end{aligned}$$

Le but est de calculer les coefficients $a_0, a_1, \cdots, a_n$. On ne peut pas appliquer la méthode des différences divisées car on n'a pas que $x_i \neq x_j$ pour $i \neq j$. Soient ξ_i, $i = 0, 1, 2, \cdots, n$, des nombres réels tels que :

$$\xi_0 < \xi_1 < \cdots < \xi_n$$

Le polynôme d'interpolation de Lagrange de degré $\leq n$ associé aux couples $(\xi_i, P(\xi_i))$, $i = 0, 1, \cdots, n$ coïncide avec le polynôme $P(x)$ grâce à l'unicité et la méthode des différences divisées nous donne alors :

$$P(x) = d_{0,0} + d_{0,1}(x - \xi_0) + d_{0,2}(x - \xi_0)(x - \xi_1) + \cdots \\ + d_{0,n}(x - \xi_0)(x - \xi_1)\cdots(x - \xi_{n-1})$$

avec

$$d_{i,i} = P(\xi_i), i = 0, 1, \cdots, n$$

et

$$d_{i,j} = \frac{d_{i+1,j} - d_{i,j-1}}{\xi_j - \xi_i}, \quad 0 \leq i < j \leq n \tag{4.10}$$

THÉORÈME 4.4.3 *Pour tout $0 \leq i \leq j \leq n$, la limite de $d_{i,j}$ lorsque ξ_i tend vers x_i ,ξ_{i+1} tend vers x_{i+1},...et ξ_j tend vers x_j existe et si on note cette limite par $c_{i,j}$ on obtient*

$$P(x) = c_{0,0} + c_{0,1}(x - x_0) + c_{0,2}(x - x_0)(x - x_1) + \\ \cdots + c_{0,n}(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

et

$$\begin{aligned} c_{i,i} &= f^0_{r(i)} \quad i = 0, 1, \cdots, n \\ c_{i,j} &= \begin{cases} \dfrac{f^{j-i}_{r(i)}}{(j-i)!}, & si\ x_i = x_j \\ \dfrac{c_{i+1,j} - c_{i,j-1}}{x_j - x_i}, & sinon \end{cases}, \quad 0 \leq i < j \leq n \end{aligned}$$

avec $0 \leq r(i) \leq m$ tel que $\alpha_{r(i)} = x_i$.

DÉMONSTRATION: Montrons par récurrence sur $k = j - i$ avec $k = 0, 1, \cdots, n$, que les limites des $d_{i,j}$ existent. Pour $k = 0$, $d_{i,i} = P(\xi_i)$, $i = 0, 1, \cdots, n$, donc, par continuité de la fonction polynôme $P(x)$ on obtient que $d_{i,i}$ admet une limite lorsque ξ_i tend vers x_i et cette limite est égale à $P(x_i) = f^0_{r(i)}$ avec $0 \leq r(i) \leq m$ tel que $\alpha_{r(i)} = x_i$. Supposons que les limites des $d_{l,s}$ existent pour tous les l, s tels que $0 \leq s - l \leq k$ et montrons que les limites des $d_{i,j}$ pour les i, j tels que $j - i = k + 1$ existent. En effet, d'après le théorème 4.3.2 et la formule (4.10), on a :

$$\begin{aligned} d_{i,j} &= \frac{d_{i+1,j} - d_{i,j-1}}{\xi_j - \xi_i} \\ &= \frac{P^{(j-i)}(c)}{(j-i)!} \end{aligned}, \quad 0 \leq i < j \leq n$$

avec $c \in I =$ le plus petit intervalle de $\mathbb{R}$ contenant $\xi_i,\ \xi_{i+1}, \cdots,\ \xi_j$. D'après l'hypothèse de récurrence $d_{i+1,j}$ (resp. $d_{i,j-1}$) admet une limite. Par conséquent, si $x_i \neq x_j$ la limite de $d_{i,j}$ existe et on a :

$$c_{i,j} = \frac{c_{i+1,j} - c_{i,j-1}}{x_j - x_i}.$$

D'autre part, si $x_i = x_j$ tous les $\xi_i,\ \xi_{i+1}, \cdots,\ \xi_j$ tendent vers la même limite x_i et par suite c tend vers x_i car l'intervalle I à la limite est réduit à $\{x_i\}$. Dans ce

cas on utilise la continuité de $P^{(j-i)}(x)$, on obtient que la limite de $d_{i,j}$ existe et on a :

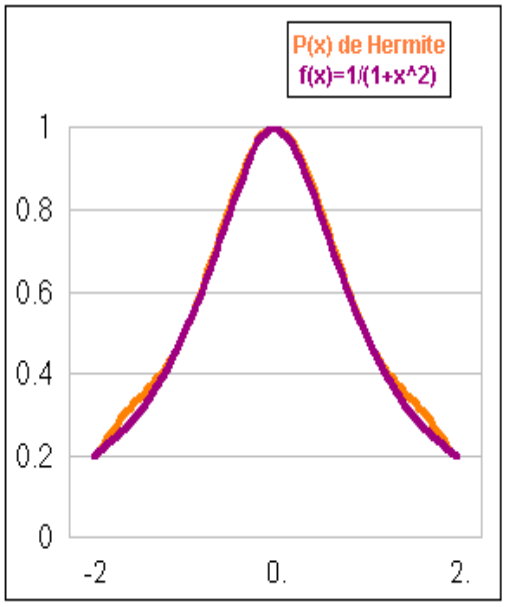
$$c_{i,j} = \frac{P^{(j-i)}(x_i)}{(j-i)!} = \frac{f^{j-i}_{r(i)}}{(j-i)!}$$

avec $0 \leq r(i) \leq m$ tel que $\alpha_{r(i)} = x_i$.
EXEMPLE: On cherche le polynôme d'interpolation de Hermite par la méthode des différences divisées généralisée pour approcher la fonction

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

sur l'intervalle $[-2, 2]$ en utilisant les couples : $(-2, f(-2))$, $(-2, f'(-2))$, $(-1, f(-1))$, $(-1, f'(-1))$, $(0, f(0))$, $(0, f'(0))$, $(1, f(1))$, $(1, f'(1))$, $(2, f(2))$, $(2, f'(2))$.

	0	1	2	3	4	5	6	7	8	9
-2.0	0.200									
		0.160								
-2.0	0.200		0.140							
		0.300		0.060						
-1.0	0.500		0.200		-0.080					
		0.500		-0.100		-0.060				
-1.0	0.500		0.000		-0.200		0.070			
		0.500		-0.500		0.150		-0.040		
0.0	1.000		-0.500		0.250		-0.050		0.010	
		0.000		0.000		0.000		0.000		0.000
0.0	1.000		-0.500		0.250		-0.050		0.010	
		-0.500		0.500		-0.150		0.040		
1.0	0.500		0.000		-0.200		0.007			
		-0.500		0.100		0.060				
1.0	0.500		0.200		-0.080					
		-0.300		-0.060						
2.0	0.200		0.140							
		-0.160								
2.0	0.200									



4.5 Polynômes de Tchebychev

Soient $x \in [-1,1]$ et $\alpha = Arccos(x) \in [0,\pi]$. Pour tout $n \in \mathbb{N}$, on a :

$$\begin{aligned}\cos(n\alpha) + i\sin(n\alpha) &= (\cos(\alpha) + i\sin(\alpha))^n \\ &= (\cos(Arccos(\alpha)) + i\sin(Arccos(\alpha)))^n \\ &= (x + i\sqrt{1-x^2})^n\end{aligned}$$

La partie réelle de cette égalité nous donne :

$$\cos(nArccos(x)) = x^n + C_n^2 x^{n-2}(x^2-1) + C_n^4 x^{n-4}(x^2-1)^2 + \cdots$$

DÉFINITION 4.5.1 *Pour tout $n \in \mathbb{N}$, on appelle polynôme de Tchebychev de degré n le polynôme $T_n(x)$ tel que sa restriction à l'intervalle $[-1,1]$ est égale à la fonction $\cos(nArccos(x))$.*

REMARQUE 4.5.1 *En utilisant la propriété : $\cos[(n+2)\alpha] + \cos(n\alpha) = 2\cos[(n+1)\alpha]\cos(\alpha)$, on obtient la relation de récurrence :*

$$T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x) \tag{4.11}$$

et de cette relation, on vérifie facilement par un raisonnement par récurrence que le coefficient du monôme du plus haut degré du polynôme T_n est égal à 2^{n-1}. On a également :

$$T_n(\cos(\theta)) = \cos(n\theta).$$

Le tableau suivant nous donne les polynômes de Tchebychev T_n pour $n = 0$ à 7 :

n	$T_n(x)$
0	1
1	x
2	$2x^2 - 1$
3	$4x^3 - 3x$
4	$8x^4 - 8x^2 + 1$
5	$16x^5 - 20x^3 + 5x$
6	$32x^6 - 48x^4 + 18x^2 - 1$
7	$64x^7 - 112x^5 + 56x^3 - 7x$

THÉORÈME 4.5.1 *Pour tout $n \in \mathbb{N}$, le polynôme de Tchebychev T_n de degré n possède n racines :*

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

La fonction T_n sur $[-1, 1]$ possède un extremum local en $(n+1)$-points :

$$x'_k = \cos(\frac{k}{n}\pi), k = 0, 1, \cdots, n.$$

de plus, pour tout $k = 0, 1, \cdots, n$, $T_n(x'_k) = (-1)^k$.

DÉMONSTRATION: Il est clair que $1 > x_0 > x_1 > \cdots > x_n > -1$ et en utilisant la propriété $T_n(\cos(\theta)) = \cos(n\theta)$, on obtient :

$$T_n(x_k) = T_n(\cos(\frac{2k+1}{2n}\pi)) = \cos[(2k+1)\frac{\pi}{2}] = 0$$

D'autre part, :

$$T'_n(x) = (\cos(nArccos(x)))' = \frac{n}{\sqrt{1-x^2}}\sin(nArccos(x))$$

d'où

$$\begin{aligned} T'_n(x'_k) &= \frac{n}{\sqrt{1-(x'_k)^2}}\sin(nArccos(x'_k)) \\ &= \frac{n}{\sqrt{1-(x'_k)^2}}\sin(nArccos(\cos(\frac{k}{n}\pi))) \\ &= \frac{n}{\sqrt{1-(x'_k)^2}}\sin(k\pi) = 0 \end{aligned}$$

et on a : $T_n(x'_k) = T_n(\cos(\frac{k}{n}\pi)) = \cos(k\pi) = (-1)^k$.

PROPOSITION 4.5.1 *Pour tout $n \in \mathbb{N}$:*

$$\max_{x\in[-1,1]} |T_n(x)| = 1$$

DÉMONSTRATION:

On a : $|T_n(x)| = |\cos(nArccos(x))| \leq 1$ pour tout $x \in [-1, 1]$ et $T_n(x_k) = (-1)^k$.

4.6 Meilleure approximation au sens de Tchebychev

Soit f une fonction continue sur l'intervalle $[a,b]$. On se propose de déterminer les points x_i, $i = 0, 1, \cdots, n$, pour lesquels le polynôme d'interpolation de Lagrange $P_n(x)$ associé aux $(n+1)$-couples $(x_i, y_i)_{i=0,n}$, avec $y_i = f(x_i)$, réalise la meilleure approximation de la fonction f. D'après le théorème de l'estimation d'erreur de l'interpolation :

$$\forall x \in [a,b], \exists c \in]a,b[/ f(x) - P_n(x) = (x-x_0)(x-x_1)\cdots(x-x_n)\frac{f^{(n+1)}(c)}{(n+1)!}$$

DÉFINITION 4.6.1 *On appelle meilleure approximation au sens de Tchebychev de la fonction f par un polynôme de degré $\leq n$, le polynôme de l'interpolation de Lagrange associé aux points d'interpolation x_i, $i = 0, 1, \cdots, n$ pour lesquels la quantité :*

$$E(x_0, x_1, \cdots, x_n) = \max_{x\in[a,b]} |(x-x_0)(x-x_1)\cdots(x-x_n)|$$

est minimale.

LEMME 4.6.1 *Soit P un polynôme de degré n tel que le coefficient de x^n est égal à 1. Alors :*

$$\max_{x\in[-1,1]} |P(x)| \geq \frac{1}{2^{n-1}}$$

DÉMONSTRATION: On pose :

$$Q(x) = P(x) - \frac{T_n(x)}{2^{n-1}}$$

Les points extremums $(x'_k)_{k=0,n}$ de T_n vérifient :

$$1 = x'_0 > x'_1 > \cdots > x'_n = -1$$

Si on suppose que :

$$\exists k \in \{0, 1, \cdots, n-1\} / Q(x'_k)Q(x'_{k+1}) \geq 0 \tag{4.12}$$

on obtient l'existence de $k \in \{0, 1, \cdots, n-1\}$ tel que :

$$Q(x'_k) = P(x'_k) - \frac{(-1)^k}{2^{n-1}} \geq 0 \text{ et } Q(x'_{k+1}) = P(x'_{k+1}) - \frac{(-1)^{k+1}}{2^{n-1}} \geq 0 \tag{4.13}$$

ou

$$Q(x'_k) = P(x'_k) - \frac{(-1)^k}{2^{n-1}} \leq 0 \text{ et } Q(x'_{k+1}) = P(x'_{k+1}) - \frac{(-1)^{k+1}}{2^{n-1}} \leq 0 \tag{4.14}$$

Si k est un entier pair, alors :

$$P(x'_k) \geq \frac{1}{2^{n-1}} \text{ ou } P(x'_{k+1}) \leq \frac{-1}{2^{n-1}}$$

et si k est un entier impair, alors :

$$P(x'_{k+1}) \geq \frac{1}{2^{n-1}} \text{ ou } P(x'_k) \leq \frac{-1}{2^{n-1}}$$

Dans tous les cas, il existe $i \in \{k, k+1\}$ tel que :

$$|P(x'_i)| \geq \frac{1}{2^{n-1}}$$

Si on suppose que la proposition (4.12) est fausse, alors :

$$\forall k \in \{0, 1, \cdots, n-1\} / Q(x'_k)Q(x'_{k+1}) < 0$$

Dans ce cas, la fonction continue $Q(x)$ change de signe n-fois dans l'intervalle [-1,1], donc, elle possède n-racines dans l'intervalle $[-1, 1]$. D'autre part, $Q(x)$ est un polynôme de degré inférieur ou égal à $n-1$, car le coefficient de x^n du polynôme de Tchebychev T_n est égal à 2^{n-1}. D'où :

$$\forall x \in \mathbb{R},\ Q(x) = P(x) - \frac{T_n(x)}{2^{n-1}} = 0$$

ce qui donne :

$$\max_{x\in[-1,1]} |P(x)| = \max_{x\in[-1,1]} |\frac{T_n(x)}{2^{n-1}}| = \frac{1}{2^{n-1}}.$$

Dans les deux cas, on a ce qu'il faut pour démontrer le lemme.

THÉORÈME 4.6.1 *La meilleure approximation de la fonction f au sens de Tchebychev est réalisée par le choix suivant :*

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} s_i,\ i = 0, 1, \cdots, n$$

où s_i, $i = 0, 1, \cdots, n$ sont les racines du polynôme de Tchebychev T_{n+1} de degré $n+1$, à savoir :

$$s_i = \cos(\frac{2i+1}{2n+2}\pi),\ i = 0, 1, \cdots, n.$$

DÉMONSTRATION: Soient $x_0,\ x_1, \cdots,\ x_n$, $(n+1)$-points de l'intervalle $[a, b]$ et $P_n(x)$ le polynôme d'interpolation de Lagrange associé aux $(n+1)$-couples $(x_i, y_i)_{i=0,n}$, avec $y_i = f(x_i)$. Il est clair que l'application :

$$\varphi : \begin{array}{c} [-1,1] \longrightarrow [a,b] \\ t \longmapsto \frac{b-a}{2}t + \frac{b+a}{2} \end{array}$$

est une bijective et que :

$$\max_{x\in[a,b]} |f(x) - P_n(x)| = \max_{t\in[-1,1]} |f(\varphi(t)) - P_n(\varphi(t))|$$

Si on note par $t_i = \varphi^{-1}(x_i)$, $i = 0, 1, \cdots, n$, $Q_n(t) = P_n(\varphi(t))$ et $g(t) = f(\varphi(t))$, on vérifie facilement que le polynôme d'interpolation de Lagrange associé aux $(n+1)$-couples $(t_i, y_i)_{i=0,n}$, avec $y_i = g(t_i) = f(x_i)$, est égal à $Q_n(t)$. Par conséquent, la recherche des points $(x_i)_{i=0,n}$ qui réalisent la meilleure approximation de f sur l'intervalle $[a, b]$ est équivalente à la recherche des points $(t_i)_{i=0,n}$ qui réalisent la meilleure approximation de g sur l'intervalle $[-1, 1]$.

Etant donné des points $(t_i)_{i=0,n}$ de l'intervalle $[-1,1]$, le polynôme

$$P(t) = (t-t_0)(t-t_1)\cdots(t-t_n)$$

est de degré $n+1$ et le coefficient de t^{n+1} est égal à 1, d'après le lemme précédent :

$$\max_{t\in[-1,1]} |(t-t_0)(t-t_1)\cdots(t-t_n)| \geq \frac{1}{2^n} = \max_{t\in[-1,1]} |\frac{T_{n+1}(t)}{2^n}|$$

D'autre part, on a :

$$T_{n+1}(t) = 2^n(t-s_0)(t-s_1)\cdots(t-s_n)$$

avec $(s_i = \cos(\frac{2i+1}{2n+2}\pi))_{i=0,n}$ les $(n+1)$-racines du ploynôme de Tchebychev T_{n+1}. D'où :

$$\max_{t\in[-1,1]} |(t-t_0)(t-t_1)\cdots(t-t_n)| \geq \max_{t\in[-1,1]} |(t-s_0)(t-s_1)\cdots(t-s_n)|$$

Ce qui prouve que les points $(s_i)_{i=0,n}$ réalisent la meilleure approximation de g et par conséquent les points :

$$x_i = \varphi(s_i) = \frac{a+b}{2} + \frac{b-a}{2}s_i, i = 0, 1, \cdots, n$$

réalisent la meilleure approximation de f.

4.7 Approximation au sens des moindres carrés

4.7.1 Cas général

Étant donné $\omega_i \in]0, +\infty[$, $i = 1, 2, \cdots, m$, on définit le produit scalaire sur $\mathbb{R}^m$ comme suit :

$$y \in \mathbb{R}^m, z \in \mathbb{R}^m, < y, z >= \sum_{i=1}^{m} \omega_i y_i z_i = y^T D z$$

avec $D = diag(\omega_1, \omega_2, \cdots, \omega_m)$ et on note par $\|.\|_D$ la norme associée à ce produit scalaire :

$$y \in \mathbb{R}^m, \|y\|_D = \sqrt{< y, y >} = \sqrt{y^T D y}$$

Soient A une matrice réelle $m \times k$ et $y \in \mathbb{R}^m$. On considère le problème de minimisation :

$$I = \min_{z\in Im(A)} \|y - z\|_D^2 \tag{4.15}$$

REMARQUE 4.7.1 *Il est connu que le problème de minimisation (4.15) admet un point minimal unique (un point qui réalise le minimum de I) $z_0 = Ax_0 \in Im(A)$ qui est égal à la projection orthogonale de y sur le sous-espace vectoriel $Im(A)$. Le vecteur x_0 est unique si et seulemnet si la matrice A est injective (ce qui est équivalent à : les colonnes de A sont linéairement indépendantes).*

THÉORÈME 4.7.1 *Étant donné A une matrice $m \times k$ réelle et un vecteur $y \in \mathbb{R}^m$, alors :*

1. *$z_0 = Ax_0$, $x_0 \in \mathbb{R}^k$, est un point minimal du problème de minimisation (4.15) si et seulement si x_0 est une solution du système linéaire :*

$$A^T DAx = A^T Dy \tag{4.16}$$

   en particulier, le système linéaire (4.16) admet au moins une solution et cette solution est unique si et seulement si A est injective.

2. *la matrice symétrique $A^T DA$ est toujours positive et elle est définie positive (en particulier inversible) si et seulement si la matrice A est injective.*

DÉMONSTRATION:

1. Le problème de minimisation (4.15) est équivalent à la minimisation de la fonction :

$$x \in \mathbb{R}^k,\ F(x) = \|y - Ax\|_D^2 = (y - Ax)^T D(y - Ax)$$

   qui est une fonction dérivable et on a :

$$x \in \mathbb{R}^k,\ F'(x) = -2A^T D(y - Ax)$$

   D'autre part, il est facile de vérifier que la fonction F est convexe :

$$\begin{array}{ll} \lambda \in [0,1],\ (x_1, x_2) \in \mathbb{R}^k \times \mathbb{R}^k,\ F(\lambda x_1 + (1-\lambda)x_2) & = \|y - A(\lambda x_1 + (1-\lambda)x_2)\|_D^2 \\ & = \|\lambda(y - Ax_1) + (1-\lambda)(y - Ax_2)\|_D^2 \\ & \leq (\lambda\|y - Ax_1\|_D + (1-\lambda)\|y - Ax_2\|_D)^2 \\ & \leq \lambda\|y - Ax_1\|_D^2 + (1-\lambda)\|y - Ax_2\|_D^2 \\ & = \lambda F(x_1) + (1-\lambda)F(x_2) \end{array}$$

   (la dernière inégalité est donnée par la convexité de la fonction réelle à variable réelle : $x \longrightarrow x^2$). La fonction F est minorée sur $\mathbb{R}^k$, donc son minimum global est atteint si et seulement si l'équation :

$$F'(x) = 0$$

   possède une solution dans $\mathbb{R}^k$, ce qui est équivalent au système linéaire : :

$$A^T Dy - A^T DAx = 0$$

   et d'après la remarque précédente, le minimum global de F est atteint en un point $x_0 \in \mathbb{R}^k$ et ce point est unique si et seulement si A est injective.

2. Il est clair que la matrice $A^T DA$ est une matrice symétrique et on a :

$$\forall x \in \mathbb{R}^k,\ x^T(A^T DA)x = (Ax)^T D(Ax) = \|Ax\|_D^2 \geq 0$$

   et $Ax \neq 0$ pour tout $x \neq 0$ si et seulement si A est injective.

4.7.2 Approximation au sens des moindres carrés d'une fonction par un polynôme

A la suite d'une expérience, on obtient des résultats qui sont en général sous la forme :

i	1	2	3	$\cdots\cdots\cdots$	100
x_i	0	0.3	0.1		-0.5
y_i	-2.3	1.6	5.2		3.1

Dans cette expérience, on refait une même manipulation n-fois (n =100 dans le cas du tableau ci-dessus) et à chaque fois on introduit une valeur x_i, $i = 1, \cdots, n$, $x_i \neq x_j$ pour $i \neq j$, et on mesure le résultat y_i. Le problème d'approximation consiste à trouver une fonction g qui approche le mieux possible les points expérimentaux, c'est à dire, une fonction g telle que sa courbe passe le plus proche possible des points expérimentaux $(x_i, y_i)_{i=1,\cdots,n}$. Soit V un sous-espace vectoriel de l'espace des fonctions continues (Vest en général de dimension finie) et $\omega_i \in]0, +\infty[$, $i = 1, 2, \cdots, m$, l'approximation au sens des moindres carrés consiste à déterminer $g \in V$ qui réalise le minimum de :

$$\min_{g \in V} \sum_{i=1}^{n} \omega_i (g(x_i) - y_i)^2 \tag{4.17}$$

Si on prend $V = \mathbb{R}_p[X]$, $p \in \mathbb{N}$, alors $g(t) = a_0 + a_1 t + \cdots + a_p t^p$ est un polynôme de degré $\leq p$ et le problème de minimisation de l'approximation au sens des moindres carrés (4.17) devient :

$$\min_{x \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \omega_i (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_p x_i^p - y_i)^2 \tag{4.18}$$

avec $x^T = (a_0 \ a_1 \ \cdots \ a_p) \in \mathbb{R}^{p+1}$.

Si on note par $D = diag(\omega_1, \cdots, \omega_n)$ et par :

$$A_p = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^p \\ \vdots & & & & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^p \end{pmatrix}, \ x = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \text{ et } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

le problème de minimisation (4.18) s'écrit sous la forme générale :

$$\min_{x \in \mathbb{R}^{p+1}} \|y - A_p x\|_D^2 \tag{4.19}$$

On applique, alors, les résultats du paragraphe précédent pour calculer les coefficients du polynôme g.

Exercice - Montrer que si $p = n-1$, alors la matrice A_{n-1} du problème de minimisation de l'approximation au sens des moindres carrés (4.19) est inversible. Déduire que pour tout $p \leq n-1$, les colonnes de la matrice A_p sont linéairement indépendantes. (Ind. On utilisera le théorème d'interpolation de Lagrange.)

Réponse : Pour $p = n-1$, le polynôme d'interpolation de Lagrange de degré $\leq n-1$ associé aux couples (x_i, y_i), $i = 1, 2, \cdots, n$, réalise le minimum de

(4.18) car il donne une somme égale à zéro. Donc, tout polynôme de degré $\leq n-1$ qui réalise le minimum de (4.18) donne une somme égale à zéro et par conséquent, il vérifie les hypothèses du théorème de Lagrange. D'après l'unicité du polynôme d'interpolation de Lagrange, le problème de minimisation (4.18) possède un point minimal unique. D'après le théorème 4.7.1, la matrice A_{n-1} est injective, donc, ses colonnes sont linéairement indépendantes, ce qui prouve qu'elle est inversible car c'est une matrice carrée. On remarque que les colonnes de la matrice A_p associée au problème de minimisation pour $p \leq n-1$ sont les p premières colonnes de la matrice A_{n-1} associée au problème de minimisation pour $p = n-1$ qui est une matrice inversible.

REMARQUE 4.7.2 *L'exercice précédent montre que dans le cas $p \leq n-1$, le polynôme de degré p qui réalise l'approximation au sens des moindres carrés est unique et que pour $p \geq n-1$, le polynôme d'interpolation de Lagrange associé aux couples $(x_i, y_i)_{i=1,n}$ est l'un des polynômes qui réalise le minimum de (4.18).*

Exemple 1. On considère l'expérience suivante :

[(Au module i de coefficient α_i)$\longrightarrow$(Salah obtient $\alpha_i y_i$, y_i=note du module i)]

i	1	2	3	4
x_i	$module1$	$module2$	$module3$	$module4$
y_i	y_1	y_2	y_3	y_4

Le problème est de trouver une note globale de Salah. On cherche une note x la plus proche possible de ses quatre notes : y_1, y_2, y_3 et y_4. L'approximation au sens des moindres carrés permet de calculer cette note x, il suffit de minimiser les carrés :

$$\min_{x\in\mathbb{R}} \sum_{i=1}^{4} \alpha_i (x - y_i)^2$$

(le coefficient α_i représente le poids du module i). Donc $p = 0$, $n = 4$ et

$$A = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \ x = x \in \mathbb{R}, \ y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \text{ et } D = diag(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$$

ce qui donne $A^T DA = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$ et $A^T Dy = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \alpha_4 y_4$. D'après ce qui précède, le point $x \in \mathbb{R}$ qui réalise le minimum est la solution du système linéaire :

$$A^T DAx = A^T Dy$$

d'où :

$$x = \frac{\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \alpha_4 y_4}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}$$

ce qui donne la moyenne.

Exemple 2. Lors d'une expérimentation, on a obtenu les résultats suivants :

i	1	2	3	4	5
x_i	0	0.1	0.2	0.3	0.4
y_i	-0.5	0.2	0.6	1.2	1.3

et on sait que la courbe de la fonction associée à cette expérience est une droite ($y = ax + b$). Le problème de minimisation de l'approximation au sens des moindres carrés est comme suit :

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{5} (y_i - ax_i - b)^2 \tag{4.20}$$

(les $\omega_i = 1$). Dans cet exemple $p = 1$ et $n = 5$ et

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0.1 \\ 1 & 0.2 \\ 1 & 0.3 \\ 1 & 0.4 \end{pmatrix}, \; x = \begin{pmatrix} b \\ a \end{pmatrix}, \; y = \begin{pmatrix} -0.5 \\ 0.2 \\ 0.6 \\ 1.2 \\ 1.3 \end{pmatrix} \text{ et } D = I$$

D'après le paragraphe précédent, le vecteur $x \in \mathbb{R}^2$ qui réalise le minimum de (4.20) est une solution du système linéaire :

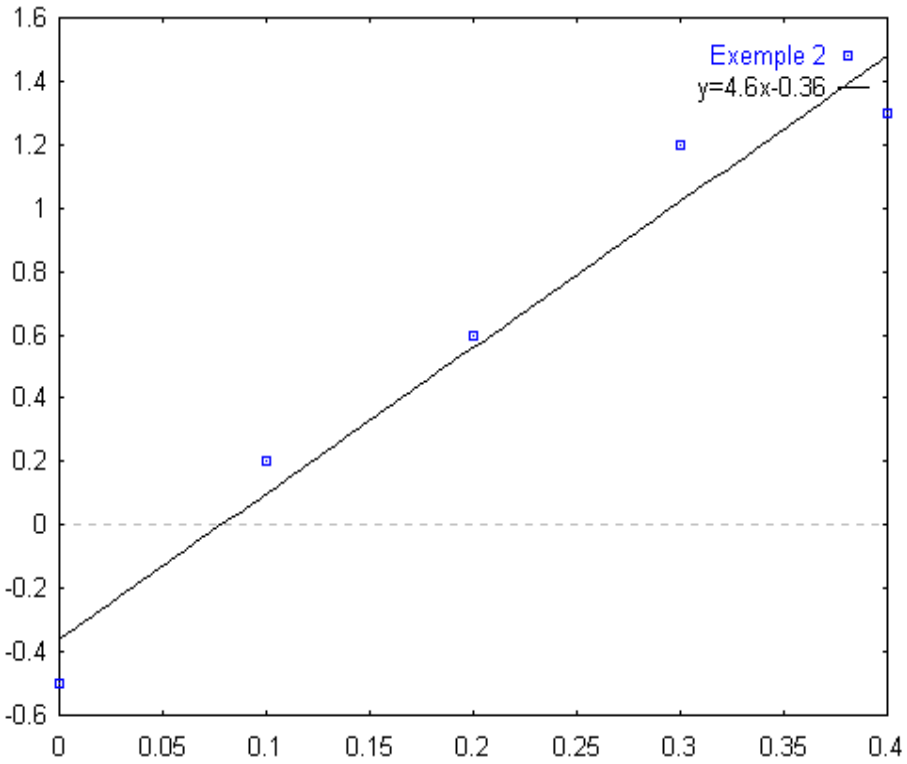
$$A^T A x = A^T y \tag{4.21}$$

Un calcul simple nous donne :

$$A^T A = \begin{pmatrix} 5 & 1 \\ 1 & 0.3 \end{pmatrix} \text{ et } A^T y = \begin{pmatrix} 2.8 \\ 1.02 \end{pmatrix}$$

d'où, la solution du système linéaire (4.21) : $a = 4.6$ et $b = -0.36$. Donc, la droite la plus proche des résultats expérimentaux de cet exemple au sens des moindres carrés est :

$$y = 4.6x - 0.36$$



Exemple 3.

A la suite d'une expérience, on a obtenu les résultats suivants :

i	1	2	3	4	5	6	7	8	9	10
x_i	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
y_i	3.1	2.5	2.4	2.3	2	1.7	1.6	1.5	1.2	1.3

et on sait que la courbe de la fonction associée à cette expérience est de la forme : $y = \beta e^{-\alpha x}$ avec α et β deux réels positifs. On remarque que $Log(y) = -\alpha x + Log(\beta) = ax + b$ qui est une fonction affine. Au lieu de chercher une fonction qui approche y, on cherche une fonction qui approche $Log(y)$, ce qui revient à appliquer la même méthode de l'exemple précédent avec les résultats suivants :

i	1	2	3	4	5	6	7	8	9	10
x_i	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
$Log(y_i)$	1.13	0.92	0.88	0.83	0.69	0.53	0.47	0.41	0.18	0.26

ce qui permet de calculer les nombres a et b et par suite $\alpha = -a$ et $\beta = e^b$.

4.8 Exercices : chapitre 4

Exercice 1

1. Calculer le polynôme d'interpolation de Lagrange p de la fonction cosinus aux points suivants :
$$x_1 = 0,\ x_2 = \frac{\pi}{3},\ x_3 = \frac{\pi}{2}$$
Tracer soigneusement les courbes représentatives de cosinus et de P dans le même repère . Quelle conclusion en tirer ?
2. Calculer le polynôme d'interpolation passant par les points
$$(0,0),\quad (1,3),\quad (3,1),\quad (5,2),\quad (8,2)$$
par la méthode des différences divisées.

Exercice 2

Nous savons que l'erreur pour l'interpolation de Lagrange de f aux points x_0, x_1 est
$$f(x) - p(x) = (x - x_0)(x - x_1)\frac{f''(\xi(x))}{2}, \quad x_0 < x < x_1,$$
si f est de classe C^2 sur $[x_0, x_1]$. Déterminer la fonction $\xi(.)$ explicitement dans le cas où $f(x) = 1/x,\ x_0 = 1,\ x_1 = 2$ et trouver le maximum (respectivement le minimum) de $\xi(.)$ sur $[1, 2]$.

Exercice 3

On interpole la fonction sinus aux points équidistants $x_j = jh,\ j \geq 0$.

1. Majorer l'erreur d'interpolation dans l'intervalle $[x_i, x_{i+1}]$ lorsqu'on fait passer un polynôme de degré 3 par $x_{i-1}, x_i, x_{i+1}, x_{i+2}$.
2. Quel h peut-on prendre pour que cette erreur soit $\leq 10^{-8}$, pour tout $i > 1$?

Exercice 4

1. Déterminer explicitement un polynôme p de degré 3 tel que :

$$p(1) = 4,\ p(2) = 5,\ p'(1) = 3,\ p'(2) = 2$$

2. Trouver un prolongement de classe C^2 par un polynôme de la fonction

$$f(x) = \begin{cases} x^2 e^x & si \quad x \leq 0 \\ (x^2-1)sin(\pi x) & si \quad x \geq 1 \end{cases}$$

Exercice 5

Soient f de classe C^1 sur l'intervalle $[0,1]$ et $x_0 = 0 < x_1 < x_2 < \cdots < x_n = 1$ une subdivision de l'intervalle $[0,1]$, $n \in \mathbb{N}^*$. On note par $\mathbb{R}_{2n+1}[X]$ l'espace vectoriel des polynômes de degré $\leq 2n+1$ et par P_f le polynôme d'interpolation de Hermite tel que :

$$P_f(x_i) = f(x_i), P_f'(x_i) = f'(x_i) \quad \forall i = 0, 1, \cdots, n.$$

1. Montrer que l'application $\mathcal{L} : \mathbb{R}_{2n+1}[X] \longrightarrow \mathbb{R}^{2n+2}$ définie par

$$\mathcal{L}(P) = (P(x_0), P(x_1), \cdots, P(x_n), P'(x_0), P'(x_1), \cdots, P'(x_n))$$

est un isomorphisme.

2. Déduire l'existence de (2n+2) polynômes de degré $\leq 2n+1$, $P_0(x), P_1(x), \cdots, P_n(x)$, $Q_0(x), Q_1(x), \cdots, Q_n(x)$, indépendants de f tels que :

$$P_f(x) = \sum_{i=0}^{n} f(x_i)P_i(x) + \sum_{i=0}^{n} f'(x_i)Q_i(x)$$

3. On suppose que $x_0 = 0$ et $x_1 = 1$. Montrer que pour tout $\overline{x} \in [0,1]$, il existe $c \in [0,1]$ tel que

$$f(\overline{x}) = P_f(\overline{x}) + \frac{\overline{x}^2(1-\overline{x})^2}{24} f^{(4)}(c)$$

4.9 Corrigé des exercices : chapitre 4

Réponse 1

1) $x_1 = 0$, $f_1 = \cos(x_1) = 1$, $x_2 = \frac{\pi}{3}$, $f_2 = \cos(x_2) = 0.5$, $x_3 = \frac{\pi}{2}$, $f_3 = \cos(x_3) = 0$

	0	1	2
0	1		
		$\frac{-3}{2\pi}$	
$\frac{\pi}{3}$	$\frac{1}{2}$		$\frac{-3}{\pi^2}$
		$\frac{-3}{\pi}$	
$\frac{\pi}{2}$	0		

$P(x) = 1 - \frac{3}{2\pi}x - \frac{3}{\pi^2}x(x - \frac{\pi}{3}) = 1 - \frac{1}{2\pi}x - \frac{3}{\pi^2}x^2$

2)

	0	1	2	3	4
0	0				
		3			
1	3		$\frac{-4}{3}$		
		-1		$\frac{-23}{15}$	
3	1		$\frac{+3}{8}$		$\frac{-131}{1680}$
		$\frac{1}{2}$		$\frac{-19}{280}$	
5	2		$\frac{-1}{10}$		
		0			
8	2				

$P(x) = 3x - \frac{4}{3}x(x-1) - \frac{23}{15}x(x-1)(x-3) - \frac{131}{1680}x(x-1)(x-3)(x-5)$

Réponse 2 $p(x) = \frac{3}{2} - \frac{x}{2}$, $f''(x) = \frac{2}{x^3} \Longrightarrow \xi(x) = \sqrt[3]{2x}$,
1) $\forall x \in [x_i, x_{i+1}], \exists c \in]x_{i-1}, x_{i+2}[/$

$$\sin(x) - P_i(x) = (x - x_{i-1})(x - x_i)(x - x_{i+1})(x - x_{i+2})\frac{\sin^{(4)}(c)}{4!}$$

$P_i(x)$=le polynôme de degré ≤ 3 qui passe par $(x_{i-1}, \sin(x_{i-1}))$, $(x_i, \sin(x_i))$, $(x_{i+1}, \sin(x_{i+1}))$, $(x_{i+2}, \sin(x_{i+2}))$, avec $x_i = ih$. Comme pour tout $x \in [x_i, x_{i+1}]$ on a $|x - x_{i-1}| \leq 2h$, $|x - x_i| \leq h$, $|x - x_{i+1}| \leq h$, $|x - x_{i+2}| \leq 2h$, on obtient :

$$|\sin(x) - P_i(x)| \leq \frac{h^4}{6}$$

2) Il suffit que $h \leq \frac{\sqrt[4]{6}}{100}$.

Réponse 3
1)

	0	1	2	3
1	4			
		3		
1	4		-2	
		1		3
2	5		1	
		2		
2	5			

$P(x) = 4 + 3(x-1) - 2(x-1)^2 + 3(x-1)^2(x-2)$
2) $f(0-) = 0$, $f'(0-) = 0$, $f''(0-) = 2$, $f(1+) = 0$, $f'(1+) = 0$, $f''(1+) = -4\pi$

	0	1	2	3	4	5
0	0					
		0				
0	0		1			
		0		-1		
0	0		0		1	
		0		0		-2π $-$ 1
1	0		0		-2π	
		0		-2π		
1	0		-2π			
		0				
1	0					

$P(x) = x^2 - x^3 + x^3(x-1) - (2\pi+1)x^3(x-1)^2$

$$f(x) = \begin{cases} x^2 e^x six \leq 0 \\ x^2 - x^3 + x^3(x-1) - (2\pi+1)x^3(x-1)^2 si 0 < x < 1 \\ (x^2-1)\sin(\pi x) six \geq 1 \end{cases}$$

Réponse 4
1) $\dim(\mathbb{R}^{2n+2}) = \dim(\mathbb{R}_{2n+1}[X]) = 2n+2$ et il est facile de vérifier que $\mathcal{L}$ est linéaire. Il suffit donc de prouver que $\mathcal{L}$ est injective :

$$\mathcal{L}(P) = 0 \Longrightarrow P(x_i) = P'(x_i) = 0 \forall i = 0, 1, \cdots, n$$

d'où, pour tout $i = 0, 1, \cdots, n$, x_i est une racine double de P, ce qui prouve que P possède $2n+2$ racines et comme $d^0 P \leq 2n+1$, nécessairement $P \equiv 0$.

2) On note par $(e_j)_{0 \leq j \leq 2n+1}$la base canonique de $\mathbb{R}^{2n+2}$. Il suffit de prendre :

$$\forall i = 0, 1, \cdots, n, P_i(x) = \mathcal{L}^{-1}(e_i)$$
$$\forall i = 0, 1, \cdots, n, Q_i(x) = \mathcal{L}^{-1}(e_{i+n+1})$$

Il est clair que les P_i et les Q_i sont indépendants de f et comme :

$$\begin{aligned} \mathcal{L}(P_f) &= (P_f(x_0), \cdots, P_f(x_n), P'_f(x_0), \cdots, P'_f(x_n)) \\ &= (f(x_0), \cdots, f(x_n), f'(x_0), \cdots, f'(x_n)) \end{aligned}$$

alors

$$\begin{aligned} P_f &= \mathcal{L}^{-1}(f(x_0), \cdots, f(x_n), f'(x_0), \cdots, f'(x_n)) \\ &= \mathcal{L}^{-1}(\textstyle\sum_{i=0}^{n} f(x_i)e_i + \sum_{i=0}^{n} f'(x_i)e_{i+n+1}) \end{aligned}$$

en utilisant la linéarité de $\mathcal{L}^{-1}$, on obtient ce qu'il faut.

3) On applique le théorème 4.4.2.

Chapitre 5

Méthodes numériques d'intégration d'une fonction et d'intégration d'un système différentiel

5.1 Formule d'intégration de Newton-côtes

Soit f une fonction continue d'un intervalle $[a, b]$ dans $\mathbb{R}$. Pour calculer, d'une manière approchée, la valeur de

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

la méthode de Newton-côtes consiste à utiliser un polynôme d'interpolation $P(x)$ pour approcher la fonction $f(x)$ sur l'intervalle $[a, b]$ et à prendre

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

comme valeur approchée de l'intégrale de $f(x)$ entre a et b.

On pose $h = \frac{b-a}{n}$, $n \in \mathbb{N}^*$ et on considère la subdivision uniforme :

$$x_k = a + kh, \quad k = 0, 1, \cdots, n$$

Soit $P(x)$ un polynôme d'interpolation de degré inférieur ou égal à n tel que :

$$f(x_k) = P(x_k) = f_k, \quad k = 0, 1, \cdots, n$$

on obtient alors :

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

Si $P(x)$ est le polynôme d'interpolation de Lagrange, on obtient :

$$\begin{aligned}\int_a^b P(x)dx &= \int_a^b (\sum_{k=0}^{n} f_k \frac{v(x)}{(x-x_k)v'(x_k)})dx \\ &= \sum_{k=0}^{n} h f_k (\frac{1}{h}\int_a^b \frac{v(x)}{(x-x_k)v'(x_k)}dx)\end{aligned}$$

avec $v(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$. En utilisant le changement de variable $x = a + ht$, on obtient

$$\alpha_k = \frac{1}{h}\int_a^b \frac{v(x)}{(x-x_k)v'(x_k)}dx = \int_0^n \prod_{i=0, i\neq k}^{n} \frac{t-i}{k-i}dt$$

par conséquent les nombres α_k, $k = 0, 1, \cdots, n$ sont des nombres rationnels indépendants de la fonction f et de l'intervalle $[a, b]$, ils dépendent seulement de n. Si f est la fonction constante 1 et $a = 0$, $b = 1$, on obtient que le polynôme d'interpolation de Lagrange P est aussi égal la constante 1, d'où :

$$1 = \int_0^1 dx = \sum_{k=0}^{n} h f_k \alpha_k = h \sum_{k=0}^{n} \alpha_k$$

ce qui donne

$$\sum_{k=0}^{n} \alpha_k = n$$

car dans ce cas particulier $h = 1/n$.

Pour $n = 1$, on obtient

$$\alpha_0 = \alpha_1 = 1/2$$

d'où

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(a) + f(b)) = \frac{b-a}{2}(f(a) + f(b))$$

Pour $n = 2$, on obtient $\alpha_0 = 1/3, \quad \alpha_1 = 4/3, \quad \alpha_2 = 1/3$. D'où

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{3}(f(a) + 4f(a+h) + f(b)) \\ &= \frac{b-a}{6}(f(a) + 4f(a+h) + f(b))\end{aligned}$$

5.2 Méthode du trapèze

Soit $N \in \mathbb{N}^*$. On considère la subdivision uniforme : $x_i = a + ih$, $i = 0, 1, \cdots, N$, avec $h = (b-a)/N$. La méthode du trapèze consiste à appliquer la méthode de Newton-côtes avec $n = 1$ sur chaque intervalle $[x_i, x_{i+1}]$, $i = 0, 1, \cdots, N-1$. On obtient

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx \approx T(h)$$

avec

$$\begin{aligned} T(h) &= \sum_{i=0}^{N-1} \frac{h}{2}(f(a+ih)+f(a+(i+1)h)) \\ &= h(\frac{f(a)}{2}+f(a+h)+\cdots+f(a+(n-1)h)+\frac{f(b)}{2}) \end{aligned}$$

THÉORÈME 5.2.1 *La méthode du trapèze est d'ordre deux, c'est à dire :*

$$\int_a^b f(x)dx - T(h) = O(h^2)$$

Plus précisément, si f est suffisamment dérivable et $h=(b-a)/N$, on obtient :

$$\exists c \in]a,b[\ / \ \int_a^b f(x)dx - T(h) = -\frac{b-a}{12}h^2 f"(c)$$

DÉMONSTRATION :

LEMME 5.2.1 *On suppose que la fonction $f : [a,b] \to \mathbb{R}$ est de classe C^3. Alors*

$$\exists c \in]a,b[\ / \ \int_a^b f(x)dx = \frac{b-a}{2}(f(b)+f(a)) - \frac{(b-a)^3}{12}f"(c)$$

Démonstration (du lemme)
On définit la fonction $g : [a,b] \to \mathbb{R}$ par :

$$g(x) = \int_a^x f(t)dt - \frac{x-a}{2}(f(x)+f(a))$$

Le calcul des dérivées de g nous donne :

$$g(a) = g'(a) = g"(a) = 0 \text{ et } g^{(3)}(x) = -\frac{1}{2}(f"(x)+(x-a)f^{(3)}(x))$$

D'après la formule de Taylor avec reste intégrale, on a :

$$\begin{aligned} g(b) &= \int_a^b \frac{(b-t)^2}{2!}g^{(3)}(t)dt \\ &= -\frac{1}{4}[\int_a^b (b-t)^2 f"(t)dt + \int_a^b (b-t)^2(t-a)f^{(3)}(t)dt] \end{aligned}$$

en intégrant par parties le deuxième terme, on obtient :

$$g(b) = -\frac{1}{2}\int_a^b (b-t)(t-a)f"(t)dt$$

D'après la formule de la moyenne et le fait que $(b-t)(t-a)$ ne change pas de signe sur l'intervalle $[a,b]$, on obtient :

$$\exists c \in]a,b[/ g(b) = \frac{-f"(c)}{2}\int_a^b (b-t)(t-a)dt = \frac{-f"(c)}{12}(b-a)^3$$

Revenons à la démonstration du théorème.

$$\begin{aligned}\int_a^b f(x)dx - T(h) &= \sum_{i=0}^{N-1}(\int_{x_i}^{x_{i+1}} f(x)dx \\ &\quad -\frac{x_{i+1}-x_i}{2}(f(x_{i+1})+f(x_i))) \\ &= -\frac{1}{12}\sum_{i=0}^{N-1}(x_{i+1}-x_i)^3 f"(c_i)\end{aligned}$$

avec $c_i \in]x_i, x_{i+1}[\subset]a, b[$, $i = 0, 1, \cdots, N-1$.
D'après la continuité de $f"$ sur l'intervalle $[a, b]$ et le fait que :

$$\min_{x\in[a,b]} f"(x) \leq \frac{f"(c_0)+f"(c_1)+\cdots+f"(c_{N-1})}{N} \leq \max_{x\in[a,b]} f"(x)$$

on obtient

$$\exists c \in]a, b[/ \frac{f"(c_0)+f"(c_1)+\cdots+f"(c_{N-1})}{N} = f"(c)$$

d'où

$$\exists c \in]a, b[/ \int_a^b f(x)dx - T(h) = -\frac{b-a}{12}h^2 f"(c)$$

5.3 Méthode de Simpson

Soit $N \in \mathbb{N}^*$. On considère la subdivision uniforme : $x_i = a+ih$, $i = 0, 1, \cdots, 2N$, avec $h = (b-a)/2N$. La méthode de Simpson consiste à appliquer la méthode de Newton-côtes avec $n = 2$ sur chaque intervalle $[x_{2i}, x_{2i+2}]$, $i = 0, 1, \cdots, N-1$. On obtient

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1}\int_{x_{2i}}^{x_{2i+2}} f(x)dx \approx S(h)$$

avec

$$\begin{aligned}S(h) &= \sum_{i=0}^{N-1}\frac{h}{3}(f(a+2ih)+4f(a+(2i+1)h)+f(a+(2i+2)h)) \\ &= \frac{h}{3}(f(a)+4f(a+h)+2f(a+2h)+4f(a+3h)+2f(a+4h) \\ &\qquad +\cdots+2f(a+2(N-2)h)+4f(a+(2N-1)h)+f(b))\end{aligned}$$

THÉORÈME 5.3.1 *La méthode de Simpson est d'ordre quatre, c'est à dire :*

$$\int_a^b f(x)dx - S(h) = O(h^4)$$

Plus précisément, si f est suffisamment dérivable et $h = (b-a)/2N$, on obtient :

$$\exists c \in]a, b[\ / \ \int_a^b f(x)dx - S(h) = -\frac{b-a}{180}h^4 f^{(4)}(c)$$

DÉMONSTRATION :

LEMME 5.3.1 *On suppose que la fonction $f : [a,b] \to \mathbb{R}$ est de classe C^5. Alors, il existe $c \in]a,b[$ tel que :*

$$\int_a^b f(x)dx = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b)) - \frac{(\frac{b-a}{2})^5}{90} f^{(4)}(c)$$

Démonstration (du lemme)
On pose $A = a + \frac{a+b}{2}$, $B = \frac{b-a}{2}$ et $D = \frac{a+b}{2}$. On définit la fonction $h : [a,D] \to \mathbb{R}$ par

$$h(x) = \int_{A-x}^{B+x} f(t)dt - \frac{x-a}{3}(f(A-x) + 4f(D) + f(B+x))$$

le calcul des dérivées de h nous donne :

$$h(a) = h'(a) = h"(a) = h^{(3)}(a) = h^{(4)}(a) = 0$$

et

$$h^{(5)}(x) = -\frac{2}{3}\psi(x) - \frac{x-a}{3}\psi'(x)$$

avec

$$\psi(x) = f^{(4)}(A-x) + f^{(4)}(B+x)$$

D'après la formule de Taylor avec reste intégrale, on a :

$$\begin{aligned} h(D) &= \int_a^D \frac{(D-t)^4}{4!} h^{(5)}(t)dt \\ &= -\frac{1}{36}\int_a^D (D-t)^4\psi(t)dt \\ &\quad -\frac{1}{72}\int_a^D (D-t)^4(t-a)\psi'(t)dt \end{aligned}$$

en intégrant par parties le deuxième terme, on obtient :

$$h(D) = -\frac{1}{72}\int_a^D [3(D-t)^4 - 4(D-a)(D-t)^3]\psi(t)dt$$

Comme $t \in [a,D]$, on a :

$$3(D-t)^4 - 4(D-a)(D-t)^3 \leq -(D-t)^4 \leq 0$$

D'après la formule de la moyenne :

$$\exists c_1 \in]a,b[/ h(D) = \frac{\psi(c_1)}{72}\int_a^D [3(D-t)^4 - 4(D-a)(D-t)^3]dt$$

Après avoir calculé l'intégrale, on obtient :

$$\exists c_1 \in]a,b[\ / \ h(D) = -\frac{\psi(c_1)}{180}(\frac{b-a}{2})^5$$

D'autre part, en utilisant la continuité de $f^{(4)}$ et le fait que

$$\min_{x\in[a,b]} f^{(4)}(x) \leq \frac{1}{2}(f^{(4)}(A-c_1) + f^{(4)}(B+c_1)) \leq \max_{x\in[a,b]} f^{(4)}(x)$$

on obtient

$$\exists c \in]a,b[\ / \ \frac{\psi(c_1)}{2} = f^{(4)}(c)$$

d'où

$$\exists c \in]a,b[\ / \ h(D) = -\frac{1}{90}(\frac{b-a}{2})^5 f^{(4)}(c)$$

Revenons à la démonstration du théorème.

$$\begin{aligned}\int_a^b f(x)dx - S(h) &= \sum_{i=0}^{N-1}(\int_{x_{2i}}^{x_{2i+2}} f(x)dx \\ &\quad -\frac{x_{2i+2}-x_{2i}}{6}(f(x_{2i}) + 4f(x_{2i+1}) \\ &\quad +f(x_{2i+2}))) \\ &= \frac{-1}{90}\sum_{i=0}^{N-1}(\frac{x_{2i+2}-x_{2i}}{2})^5 f^{(4)}(c_i)\end{aligned}$$

avec $c_i \in]a,b[$, $i = 0, 1, \cdots, N-1$. D'après la continuité de $f^{(4)}$ sur l'intervalle $[a,b]$ et le fait que :

$$\min_{x\in[a,b]} f^{(4)}(x) \leq \frac{f^{(4)}(c_0) + \cdots + f^{(4)}(c_{N-1})}{N} \leq \max_{x\in[a,b]} f^{(4)}(x)$$

on obtient

$$\exists c \in]a,b[\ / \ \frac{f^{(4)}(c_0) + f^{(4)}(c_1) + \cdots + f^{(4)}(c_{N-1})}{N} = f^{(4)}(c)$$

d'où

$$\exists c \in]a,b[\ / \ \int_a^b f(x)dx - S(h) = -\frac{b-a}{180}h^4 f^{(4)}(c)$$

5.4 Problème de Cauchy

Soit $[a,b]$ un intervalle fermé de $\mathbb{R}$. Soit f une application de $\mathbb{R} \times \mathbb{R}^n$ dans $\mathbb{R}^n$. On appelle système différentiel du premier ordre la relation :

$$\frac{dy}{dx}(x) = f(x, y(x)) \tag{5.1}$$

avec $y : [a,b] \longrightarrow \mathbb{R}^n$ une application dérivable sur $]a,b[$. On appelle problème de Cauchy le système défini par les équations (5.1) et la condition initiale :

$$y(x_0) = y_0 \tag{5.2}$$

avec $x_0 \in [a,b]$
et $y_0 \in \mathbb{R}^n$.
Une application y de $[a,b]$ dans $\mathbb{R}^n$ est dite solution du problème de Cauchy (5.1)+(5.2) si elle est dérivable sur $]a,b[$ et vérifie les équations (5.1) et (5.2). Nous admettons le théorème suivant :

THÉORÈME 5.4.1 *Soit $f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ une application continue. Si f vérifie la condition de Lipschitz par rapport à la deuxième variable : il existe $L > 0$ tel que*

$$\| f(x,y) - f(x,z) \| \leq L \| y - z \|, \quad \forall y, z \in \mathbb{R}^n, \quad \forall x \in]a,b[$$

avec $\| \ . \ \|$ est une norme quelconque de $\mathbb{R}^n$. Alors, le problème de Cauchy (5.1)+(5.2) admet une solution et une seule sur $[a,b]$ pour tout $x_0 \in [a,b]$ et $y_0 \in \mathbb{R}^n$.

COROLLAIRE 5.4.1 *Soit $f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ une application continue, dérivable par rapport à y et vérifie :*

$$\sum_{i=1}^{n} | \frac{\partial f_j}{\partial y_i}(x,y) | \leq M, \quad \forall x \in [a,b], \ \forall y \in \mathbb{R}^n, \forall j = 1,n.$$

Alors, le problème de Cauchy (5.1)+(5.2) admet une solution et une seule sur $[a,b]$ pour tout $x_0 \in [a,b]$ et $y_0 \in \mathbb{R}^n$.

DÉMONSTRATION : Soient $z \in \mathbb{R}^n$ et $y \in \mathbb{R}^n$. Pour tout $x \in [a,b]$ et tout $j = 1,n$, on applique le lemme de Rolle à l'application $H_j(t) = f_j(x, y+t(z-y))$ sur l'intervalle $[0,1]$:

$$\exists c \in]0,1[\ / \ f_j(x,z) - f_j(x,y) = H_j(1) - H_j(0) = H_j'(c)$$

or

$$H_j'(t) = \sum_{i=1}^{n} \frac{\partial f_j}{\partial x_i}(x, y + t(z-y)).(z_i - y_i)$$

ce qui donne :

$$\begin{array}{ll} |H_j'(t)| & \leq \sum_{i=1}^{n} |\frac{\partial f_j}{\partial x_i}(x, y + t(z-y))||z_i - y_i| \\ & \leq \|z-y\|_\infty \sum_{i=1}^{n} |\frac{\partial f_j}{\partial x_i}(x, y + t(z-y))| \\ & \leq M \|z-y\|_\infty \end{array}$$

d'où

$$\|f(x,z) - f(x,y)\|_\infty = \max_j |f_j(x,z) - f_j(x,y)| \leq M \|z-y\|_\infty$$

il suffit donc d'appliquer le théorème précédent.

Dans toute la suite de ce chapitre nous nous intéressons à la résolution numérique du problème de Cauchy. Nous supposons donc que le problème de Cauchy (5.1)+(5.2) admet une solution unique définie sur $[a,b]$. Etant donné une subdivision de l'intervalle $[a,b]$, $\{x_0, x_1, \cdots, x_k\}$, l'idée principale des méthodes numériques est d'approcher numériquement les valeurs $\{y(x_0), y(x_1), \cdots, y(x_k)\}$ par des valeurs approchées $\{y_0, y_1, \cdots, y_k\}$, ($y(x_0) = y_0$ et $y(.)$ est la solution exacte du problème de Cauchy (5.1)+(5.2)).

5.5 Intégration approchée

5.5.1 Principe d'une méthode numérique

Notons par $x_0 = a$ et par y_0 la condition initiale du problème de Cauchy. A l'étape m, $0 \leq m \leq k-1$, on pose :

$$y_{m+1} = y_m + h_m \Phi(x_m, y_m, h_m) \tag{5.3}$$

où Φ est l'application qui caractérise la méthode d'intégration et $h_m = x_{m+1} - x_m$. Le choix de la fonction Φ dépend de la précission avec laquelle on veut approcher les vraies valeurs $(y(x_m))_{m=1,k}$. Pour simplifier les notations, on supposera dans toute la suite de ce chapitre que la subdivision de l'intervalle $[a, b]$ est uniforme, c'est à dire :

$$\forall m = 1,\ k,\ h_m = h = \frac{b-a}{k}$$

5.5.2 Consistance d'une méthode numérique

La consistance est une notion théorique qui exprime la relation entre le système différentiel et la fonction Φ.

DÉFINITION 5.5.1 *On dit que la méthode d'intégration (5.3) est consistante avec le système différentiel (5.1) si :*

$$\lim_{h \to 0} \max_{m} \| \frac{y(x_{m+1}) - y(x_m)}{h} - \Phi(x_m, y(x_m), h) \| = 0 \tag{5.4}$$

pour toute solution y du problème de Cauchy (5.1)+(5.2).

THÉORÈME 5.5.1 *Une condition nécessaire et suffisante pour que la méthode d'intégration approchée (5.3) soit consistante est que :*

$$\Phi(x, z, 0) = f(x, z), \qquad \forall x \in [a, b], \quad \forall z \in \mathbb{R}^n \tag{5.5}$$

DÉMONSTRATION : Montrons que la condition (5.5) est nécessaire ; en effet, soit $x \in [a, b]$ et $z \in \mathbb{R}^n$, d'après le théorème 5.4.1 il existe une solution unique y du système (5.1) avec la condition $y(x) = z$. On considère la subdivision de l'intervalle $[a, b]$ définie par

$$\begin{aligned} &h = \frac{x-a}{p}, \quad x_0 = a \\ &x_1 = x_0 + h, x_2 = x_1 + h, \cdots, x_p = x, x_{p+1} = x_p + h, \\ &\qquad \cdots, x_{k-1} = a + (k-1)h \end{aligned}$$

où k est l'entier vérifiant l'inégalité suivante :

$$a + (k-1)h < b \leq a + kh.$$

et on pose $x_k = b$. Utilisons le fait que y est une solution du système (5.1), d'où

$$\int_{x_m}^{x_{m+1}} f(t, y(t))dt = y(x_{m+1}) - y(x_m).$$

Si la méthode est consistante, on obtient que

$$\lim_{h\to 0}\max_{m} \parallel \frac{1}{h}\int_{x_m}^{x_{m+1}} f(t,y(t))dt - \Phi(x_m,y(x_m),h) \parallel = 0.$$

En particulier pour $m = p$

$$\lim_{h\to 0} \parallel \frac{1}{h}\int_{x}^{x+h} [f(t,y(t)) - \Phi(x,y(x),h)]dt \parallel = 0.$$

Utilisons le fait que les fonctions $f,\ \Phi$ sont continues et que $z = y(x)$, alors

$$f(x,z) = \Phi(x,z,0).$$

Montrons que la condition (5.5) est suffisante ; d'après (5.5), la condition (5.3) est la même que la condition suivante :

$$\lim_{h\to 0}\max_{m} \parallel \frac{1}{h}\int_{x_m}^{x_{m+1}} [\Phi(t,y(t),0) - \Phi(x_m,y(x_m),h)]dt \parallel = 0$$

or, cette condition est une conséquence immédiate de la continuité uniforme de Φ sur le compact $\{(t,y(t))/t \in [a,b]\} \times [0, b-a]$.

5.5.3 Stabilité d'une méthode numérique

Soient $(y_m)_{m=1,k},\ (z_m)_{m=1,k}$ les solutions respectivement de :

$$\begin{cases} y_{m+1} = y_m + h\Phi(x_m,y_m,h) \\ y_0 \in \mathbb{R}^n \text{ quelconque donné} \end{cases} \tag{5.6}$$

et de

$$\begin{cases} z_{m+1} = z_m + h[\Phi(x_m,z_m,h) + \varepsilon_m] \\ z_0 \in \mathbb{R}^n \text{ quelconque donné} \end{cases} \tag{5.7}$$

avec

$$\begin{aligned} z_0 &= y_0 \\ x_0 &= a \\ x_{m+1} &= x_m + h \end{aligned}$$

DÉFINITION 5.5.2 *On dit que la méthode d'intégration (5.3) est stable s'il existe M_1 indépendant de h tel que :*

$$\max_{m} \parallel y_m - z_m \parallel \leq M_1 \max_{m} \parallel \varepsilon_m \parallel \tag{5.8}$$

La notion de stabilité veut dire en "gros" qu'une petite perturbation de la fonction Φ, c'est à dire de la méthode entraîne au plus une petite perturbation sur le résultat. Comme les valeurs (y_m) sont calculées d'une manière approchée, il faut que la méthode utilisée soit stable, sinon on risque d'obtenir des résultats complètement faux.

THÉORÈME 5.5.2 *S'il existe $M > 0$ indépendant de h tel que pour h assez petit Φ vérifie*

$$\parallel \Phi(x,y,h) - \Phi(x,z,h) \parallel \leq M \parallel y - z \parallel, \quad \forall y,z \in \mathbb{R}^n, \quad \forall x \in [a,b].$$

Alors, la méthode d'intégration (5.3) est stable.

DÉMONSTRATION : Pour simplifier la démonstration, on suppose que le pas de la subdivision $(x_m)_{m=1,k}$ est uniforme, c'est-à-dire que $h = \dfrac{b-a}{k}$. On a

$$\begin{aligned} \| y_{m+1} - z_{m+1} \| \quad &\leq \| y_m - z_m \| + \\ &h \| \Phi(x_m, y_m, h) - \Phi(x_m, z_m, h) \| + h \| \varepsilon_m \| \end{aligned}$$

soit encore

$$\| y_{m+1} - z_{m+1} \| \leq (1 + Mh) \| y_m - z_m \| + h \| \varepsilon_m \| .$$

Montrons par récurrence que

$$\| y_{m+1} - z_{m+1} \| \leq \frac{(1+hM)^{m+1} - 1}{M} \max_{p=1,k} \| \varepsilon_p \| .$$

En effet

$$\| y_{m+2} - z_{m+2} \| \leq (1 + Mh) \| y_{m+1} - z_{m+1} \| + h \| \varepsilon_{m+1} \|,$$

d'où

$$\| y_{m+2} - z_{m+2} \| \leq [(1 + hM) \frac{(1+hM)^{m+1} - 1}{M} + h] \max_{p=1,k} \| \varepsilon_p \|,$$

soit encore

$$\| y_{m+2} - z_{m+2} \| \leq \frac{(1+hM)^{m+2} - 1}{M} \max_{p=1,k} \| \varepsilon_p \| .$$

D'autre part, pour $x \geq 0$ on a :

$$1 + x \leq e^x$$

d'où

$$(1 + hM)^m \leq e^{mhM} \leq e^{(b-a)M}, \quad \forall m = 1, \cdots, k.$$

Par conséquent

$$\max_{m=1,k} \| y_m - z_m \| \leq M_1 \max_{m=1,k} \| \varepsilon_m \|,$$

avec $M_1 = \dfrac{e^{(b-a)M} - 1}{M}$.

5.5.4 Convergence d'une méthode numérique

DÉFINITION 5.5.3 *On dit que la méthode d'intégration (5.3) est convergente si :*

$$\lim_{h \to 0} \max_{m} \| y_m - y(x_m) \| = 0.$$

THÉORÈME 5.5.3 *Supposons que la méthode d'intégration (5.3) est consistante et stable. Alors, elle est convergente.*

DÉMONSTRATION : Soit y la solution du système différentiel (5.1) avec la condition initiale $y(a) = y_0$, on pose

$$\varepsilon_m = \frac{y(x_{m+1}) - y(x_m)}{h} - \Phi(x_m, y(x_m), h).$$

Si la méthode est consistante, par définition on a

$$\lim_{h\to 0} \max_m \| \varepsilon_m \| = 0.$$

D'autre part, si on applique la définition de la stabilité à

$$(z_m)_{m=1,k} = (y(x_m))_{m=1,k}$$

et

$$(y_m)_{m=1,k},$$

on obtient :

$$\max_m \| y_m - y(x_m) \| \leq M_1 \max_m \| \varepsilon_m \|,$$

d'où la convergence en faisant tendre h vers zéro.

5.5.5 Ordre d'une méthode numérique

L'ordre d'une méthode est un moyen théorique pour mesurer la précision avec laquelle les valeurs approchées sont calculées.

DÉFINITION 5.5.4 *On dit que la méthode d'intégration est d'ordre p si :*

$$\max_m \| y_m - y(x_m) \| = O(h^p).$$

où $y(.)$ est la solution du problème de Cauchy (5.1)+(5.2) pour y_0 quelconque.

THÉORÈME 5.5.4 *Supposons que la méthode d'intégration (5.3) est stable et que*

$$\max_m \| \frac{1}{h}[y(x_{m+1}) - y(x_m)] - \Phi(x_m, y(x_m), h) \| = O(h^p).$$

pour toute solution $y(.)$ du problème de Cauchy (5.1)+(5.2). Alors, la méthode d'intégration (5.3) est d'ordre p.

DÉMONSTRATION : Soit $y(.)$ la solution exacte du problème (5.1) avec la condition initiale $y(x_0) = y_0$. Soit $(y_m)_{m=1,k}$ la solution de la méthode d'intégration (5.3). On a

$$\begin{array}{rl} y(x_{m+1}) - y(x_m) & = h[\Phi(x_m, y(x_m), h) + (\frac{y(x_{m+1})-y(x_m)}{h} - \Phi(x_m, y(x_m), h))] \\ y_{m+1} - y_m & = h\Phi(x_m, y_m, h) \end{array}$$

Si on pose $\varepsilon_m = \dfrac{y(x_{m+1}) - y(x_m)}{h} - \Phi(x_m, y(x_m), h)$ et $z_m = y(x_m)$, par définition de la stabilité on a :

$$\max_m \| z_m - y_m \| = \max_m \| y(x_m) - y_m \| \leq M_1 \max_m \| \varepsilon_m \|,$$

il suffit donc de prouver que $\varepsilon_m = O(h^p)$, ce qui est vrai par hypothèse.

La proposition suivante est très utile pour le calcul de l'ordre d'une méthode numérique.

PROPOSITION 5.5.1 *Soient $x \in [a,b]$ et $z \in K$ avec K un compact de $\mathbb{R}^n$. Supposons que f et Φ sont suffisamment régulières, alors :*

1. *la solution y du problème de Cauchy (5.1)+(5.2) vérifie :*

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!}y"(x) + \cdots + \frac{h^p}{p!}y^{(p)}(x) + O(h^{p+1})$$

2. *la fonction Φ vérifie :*

$$\Phi(x,z,h) = \Phi(x,z,0) + h\frac{\partial \Phi}{\partial h}(x,z,0) + \cdots + h^p\frac{\partial^p \Phi}{\partial h^p}(x,z,0) + O(h^{p+1})$$

et dans les deux cas le reste est une quantité uniformément par rapport à (x,z).

La démonstration est une application simple de la formule de Taylor avec reste intégrale à chaque fonction $h \to y_i(x+h)$ et à chaque fonction $h \to \Phi_i(x,z,h)$, $i = 1, \cdots, n$.

5.6 Méthode d'Euler

Soit $y(.)$ la solution exacte du problème (5.1). La méthode d'Euler est une conséquence de la remarque suivante : pour h assez petit, on a :

$$y(x+h) - y(x) \simeq h\ f(x, y(x))\ ;$$

par conséquent, pour $h = \frac{b-a}{k}$, en tout point $x_m = a + mh$, $m = 0, 1, \cdots, k$, on obtient une approximation y_m de la valeur $y(x_m)$ de la solution exacte $y(x)$ de la manière suivante :

$$y_0 = y(x_0)$$

$$y_{m+1} = y_m + hf(x_m, y_m), \quad x_{m+1} = x_m + h.$$

Ce qui revient à prendre

$$\Phi(x,y,h) = f(x,y), \quad \forall x, y, h.$$

Il est clair que la méthode d'Euler est consistante et si on suppose que f est lipschitzienne par rapport à y uniformément par rapport à $x \in [a,b]$, on obtient une méthode stable et convergente.

Exercice d'application : Vérifier que la méthode d'Euler est d'ordre 1 lorsqu'on suppose que f est lipschitzienne par rapport à y.
Réponse : Soit $h > 0$. La fonction $t \to f(t, y(t))$ est continue sur [a,b], donc, elle est uniformément continue sur [a,b], d'où :

$$\max_{|t-t'| \leq h} \|f(t, y(t)) - f(t', y(t'))\| = O(h)$$

D'autre part, d'après (5.1) :

$$y(x_{m+1}) - y(x_m) = \int_{x_m}^{x_{m+1}} f(x, y(x))dx$$

d'où

$$\begin{aligned}\|\frac{1}{h}[y(x_{m+1}) - y(x_m)] \quad & -\Phi(x_m, y(x_m), h)\| \\ & = \frac{1}{h}\| \int_{x_m}^{x_{m+1}} (f(x, y(x)) - f(x_m, y(x_m)))dx\| \\ & \leq \frac{1}{h} \int_{x_m}^{x_{m+1}} \|f(x, y(x)) - f(x_m, y(x_m)))\|dx \\ & \leq \frac{1}{h} \int_{x_m}^{x_{m+1}} O(h)dx = O(h)\end{aligned}$$

5.7 Méthodes de Runge-Kutta

Pour obtenir des méthodes d'ordre supérieur à un, on introduit les méthodes dites de Runge-Kutta :

$$\begin{aligned} k_1 &= f(x, y), \\ k_2 &= f(x + \theta_2 h, y + a_{21} h k_1), \\ k_3 &= f(x + \theta_3 h, y + a_{31} h k_1 + a_{32} h k_2), \\ &\vdots \\ k_r &= f(x + \theta_r h, y + h \sum_{i=1}^{r-1} a_{ri} k_i), \end{aligned}$$

et on pose $\Phi(x, y, h) = \sum_{j=1}^{r} c_j k_j$.

Il est clair que

$$\Phi(x, y, 0) = f(x, y) \sum_{j=1}^{r} c_j;$$

par conséquent, les méthodes de Runge-Kutta sont consistantes si et seulement si

$$\sum_{j=1}^{r} c_j = 1.$$

Si on suppose que f est lipschitzienne par rapport à y uniformément par rapport à $x \in [a, b]$, on vérifie facilement que pour h petit $\Phi(x, ., h)$ est aussi lipschitzienne par rapport à y uniformément par rapport à $x \in [a, b]$, avec une constante de Lipschitz indépendante de h et de x, ce qui prouve que les méthodes de Runge-Kutta sont stables. Les méthodes de Runge-Kutta sont donc convergentes.

Etudions l'ordre de quelques unes de ces méthodes. Un calcul simple nous donne :

$$\begin{aligned} y'(x) &= f(x, y(x)), \\ y"(x) &= \frac{df(x, y(x))}{dx} = \frac{\partial f(x, y)}{\partial x} + \frac{\partial f(x, y)}{\partial y} f(x, y), \end{aligned}$$

et dans le cas $\Phi(x, z, h) = \sum_{i=1}^{2} c_i k_i$:

$$\begin{aligned} \Phi(x, z, 0) &= (c_1 + c_2) f(x, z) \\ \frac{\partial \Phi}{\partial h}(x, z, 0) &= c_2[\theta_2 \frac{\partial f(x, z)}{\partial x} + a_{21} \frac{\partial f(x, z)}{\partial y} f(x, z)] \end{aligned}$$

ce qui donne, en utilisant la proposition 5.5.1 :

$$\frac{y(x+h)-y(x)}{h} = f(x,y(x)) + \frac{1}{2}h[\frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}f(x,y)] + O(h^2)$$

et

$$\begin{aligned}\Phi(x,y,h) &= c_1 f(x,y) + c_2 f(x+\theta_2 h, y + a_{21} h f(x,y)),\\ &= (c_1+c_2)f(x,y) + c_2 h[\theta_2 \frac{\partial f(x,y)}{\partial x}\\ &\qquad + a_{21}\frac{\partial f(x,y)}{\partial y} f(x,y)] + O(h^2).\end{aligned}$$

D'où, pour que la méthode de Runge-Kutta soit d'ordre deux, il faut que

$$c_1 + c_2 = 1, \quad c_2\theta_2 = 1/2, \quad c_2 a_{21} = 1/2.$$

Une solution de cette équation est

$$c_1 = c_2 = 1/2, \quad \theta_2 = a_{21} = 1,$$

ce qui donne la méthode de Heun (1900) :

$$\Phi(x,y,h) = \frac{1}{2}[f(x,y) + f(x+h, y+hf(x,y))],$$

elle demande le calcul de f seulement deux fois par itération. Une autre solution est

$$c_1 = 0, \quad c_2 = 1, \quad \theta_2 = a_{21} = 1/2,$$

elle est connue sous le nom "Méthode d'Euler modifiée" [Collatz 1960]

$$\Phi(x,y,h) = f(x + \frac{h}{2}, y + \frac{h}{2}f(x,y)),$$

La méthode de Runge-Kutta d'ordre quatre est la suivante

$$\begin{aligned}k_1 &= f(x,y),\\ k_2 &= f(x+h/2, y+hk_1/2),\\ k_3 &= f(x+h/2, y+hk_2/2)\\ k_4 &= f(x+h, y+hk_3),\end{aligned}$$

$$\Phi(x,y,h) = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

On montre que la méthode de Runge-Kutta 4 est d'ordre 4.

5.8 Exercices corrigés

Exercice 5.1

Soit $a \leq x_0 < x_1 < \cdots < x_n \leq b$ une subdivision de l'intervalle $[a,b]$. Soit $P(x)$ un polynôme de degré $\leq n$.

1) Montrer que le polynôme d'interpolation de Lagrange Q de degré $\leq n$ tel que : $Q(x_i) = P(x_i)$ pour $i = 0, 1, \cdots, n$, coïncide avec $P(x)$.

2) Montrer que les coefficients du polynôme $P(x)$ vérifient un système de la forme $Ax = b$ avec A une matrice à déterminer. Déduire du théorème d'interpolation de Lagrange que A est inversible.

3) En utilisant l'expression du polynôme d'interpolation de Lagrange, montrer qu'il existe $\gamma_0, \gamma_1, \cdots, \gamma_n$ des constantes indépendantes de $P(x)$ tels que :

$$\int_a^b P(x)dx = \sum_{i=0}^{n} \gamma_i P(x_i)$$

4) En remplacant le polynôme $P(x)$, dans l'équation des γ_i, par les monômes $1, x, x^2, \cdots, x^n$, montrer que le vecteur formé par les γ_i est une solution d'un système de la forme $A^T x = d$ (la matrice A de la question 2)). En déduire que les γ_i sont uniques.

Réponse 5.1

Soit $Q(x)$ le polynôme d'interpolation de Lagrange de degré $\leq n$ tel que :

$$Q(x_i) = P(x_i), \;\; i = 0, 1, \cdots, n$$

On sait que l'expression de $Q(x)$ est la suivante :

$$Q(x) = \sum_{i=0}^{n} P(x_i) L_i(x)$$

avec

$$L_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}$$

D'autre part, d'après l'unicité du polynôme d'interpolation de Lagrange et le fait que degré de $P(x) \leq n$, on obtient que $P(x) = Q(x)$. D'où :

$$\int_a^b P(x)dx = \sum_{i=0}^{n} P(x_i) \int_a^b L_i(x)dx$$

On pose $\gamma_i = \int_a^b L_i(x)dx$ pour $i = 0, 1, \cdots, n$, ce qui donne l'existence des γ_i. Montrons l'unicité des γ_i. Le polynôme d'interpolation de Lagrange $P(x) = a_0 + a_1 x + \cdots + a_n x^n$ est unique, donc les coefficients $a_0, a_1, \cdots, a_n$ sont uniques. Par conséquent, le système suivant admet une solution unique :

$$\begin{cases} a_0 + a_1 x_0 + \cdots + a_n x_0^n & = P(x_0) \\ a_0 + a_1 x_1 + \cdots + a_n x_1^n & = P(x_1) \\ \vdots & \\ a_0 + a_1 x_n + \cdots + a_n x_n^n & = P(x_n) \end{cases}$$

quelque soit les nombres $P(x_0), P(x_1), \cdots, P(x_n)$. D'où, la matrice :

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}$$

est inversible. D'autre part, si on remplace, dans l'équation des γ_i, le polynôme $P(x)$ par les monômes $1, x, x^2, \cdots, x^n$, on obtient le système :

$$\begin{cases} \gamma_0 + \gamma_1 + \cdots + \gamma_n & = \int_a^b 1dx \\ \gamma_0 x_0 + \gamma_1 x_1 + \cdots + \gamma_n x_n & = \int_a^b x dx \\ \vdots & \\ \gamma_0 x_0^n + \gamma_1 x_1^n + \cdots + \gamma_n x_n^n & = \int_a^b x^n dx \end{cases}$$

qui s'écrit sous la forme matricielle :

$$A^T\gamma = b$$

avec $\gamma = (\gamma_0 \quad \gamma_1 \quad \cdots \gamma_n)^T$. Comme la matrice A est inversible, alors, A^T est aussi inversible. Ce qui prouve l'unicité des γ_i.

Exercice 5.2

Si $f \in C^2[a,b]$ alors il existe $\bar{x} \in [a,b]$ tel que l'erreur de la méthode du trapèze est égale à :

$$\int_a^b f(x)dx - \frac{1}{2}(b-a)(f(a)+f(b)) = \frac{1}{12}(b-a)^3 f"(\bar{x})$$

Déduire ce résultat en utilisant un résultat du cours du chapitre précédent.

Réponse 5.2

D'après un théorème du chapitre précédent et pour $n=1$, $x_0 = a$ et $x_1 = b$:

$$\forall x \in [a,b], \exists c(x) \in [a,b] / f(x) = P(x) + \frac{(x-a)(x-b)f"(c(x))}{2!}$$

avec $P(x) = f(a) + (f(b)-f(a))(x-a)/(b-a)$ le polynôme d'interpolation de Lagrange de degré $\leq n = 1$ et qui vérifie $P(a) = f(a)$ et $P(b) = f(b)$. Montrons que $f"(c(x))$ est continue sur $[a,b]$. On a :

$$f"(c(x)) = \frac{2(f(x)-P(x))}{(x-a)(x-b)}$$

Il est clair que la fonction $f"(c(x))$ est continue sur $[a,b]$ sauf peut être en a et b, mais il est facile de voir qu'on peut la prolonger en ces points car :

$$\lim_{x\to a^+} f"(c(x)) = \lim_{x\to a^+} \frac{2(f(x)-P(x))}{(x-a)(x-b)} = \frac{2}{a-b}(f'(a)-P'(a))$$

et

$$\lim_{x\to b^-} f"(c(x)) = \lim_{x\to b^-} \frac{2(f(x)-P(x))}{(x-a)(x-b)} = \frac{2}{b-a}(f'(b)-P'(b))$$

On peut donc supposer que $f"(c(x))$ est continue sur $[a,b]$. D'autre part, $(x-a)(x-b)$ ne change pas de signe sur $[a,b]$, d'après la formule de la moyenne :

$$\exists \alpha \in [a,b] / \int_a^b (x-a)(x-b)f"(c(x))dx = f"(c(\alpha)) \int_a^b (x-a)(x-b)dx$$

En calculant $\int_a^b P(x)dx$ et $\int_a^b (x-a)(x-b)dx$, on obtient ce qu'il faut.

Exercice 5.3

Si $f \in C^4[a,b]$ alors il existe $\bar{x} \in [a,b]$ tel que :

$$\int_a^b f(x)dx - \frac{1}{2}(b-a)(f(a)+f(b)) - \frac{(b-a)^2}{12}(f'(a)-f'(b)) = \frac{(b-a)^5}{720} f^{(4)}(\bar{x})$$

Déduire ce résultat du cours du chapitre précédent.

Réponse 5.3

D'après un théorème du chapitre précédent et pour $m = 1$, $\xi_0 = a$, $\xi_1 = b$ et les conditions$P(a) = f(a),\ P'(a) = f'(a),\ P(b) = f(b),\ P'(b) = f'(b)$:

$$\forall x \in [a,b], \exists c(x) \in [a,b] / f(x) = P(x) + \frac{(x-a)^2(x-b)^2 f^{(4)}(c(x))}{4!}$$

avec $P(x) = c_{00} + c_{01}(x-a) + c_{02}(x-a)^2 + c_{03}(x-a)^2(x-b)$ le polynôme d'interpolation de Hermite de degré ≤ 3 et les coefficients donnés par la méthode des différences divisées généralisée :

$$c_{00} = f(a)$$

$$c_{01} = f'(a)$$

$$c_{02} = \frac{f(b) - f(a) - (b-a)f'(a)}{(b-a)^2}$$

$$c_{03} = \frac{f'(b) + f'(a) - 2(f(b) - f(a))/(b-a)}{(b-a)^2}$$

Montrons que $f^{(4)}(c(x))$ est continue sur $[a,b]$. On a :

$$f^{(4)}(c(x)) = \frac{4!(f(x) - P(x))}{(x-a)^2(x-b)^2}$$

Il est clair que la fonction $f^{(4)}(c(x))$ est continue sur $[a,b]$ sauf peut être en a et b, mais il est facile de voir qu'on peut la prolonger en ces points car :

$$\lim_{x \to a^+} f^{(4)}(c(x)) = \lim_{x \to a^+} \frac{4!(f(x) - P(x))}{(x-a)^2(x-b)^2} = \frac{2}{(a-b)^2}(f"(a) - P"(a))$$

et

$$\lim_{x \to b^-} f^{(4)}(c(x)) = \lim_{x \to b^-} \frac{4!(f(x) - P(x))}{(x-a)^2(x-b)^2} = \frac{2}{(b-a)^2}(f"(b) - P"(b))$$

On peut donc supposer que $f^{(4)}(c(x))$ est continue sur $[a,b]$. D'autre part, $(x-a)^2(x-b)^2$ ne change pas de signe sur $[a,b]$, d'après la formule de la moyenne :

$$\exists \alpha \in [a,b] / \int_a^b (x-a)^2(x-b)^2 f^{(4)}(c(x))dx = f^{(4)}(c(\alpha)) \int_a^b (x-a)^2(x-b)^2 dx$$

En calculant $\int_a^b P(x)dx$ et $\int_a^b (x-a)^2(x-b)^2 dx$, on obtient ce qu'il faut.

Exercice 5.4

Soient $f \in C^6[-1,1]$ et $P \in \mathbb{R}^5[X]$ le polynôme d'interpolation de Hermite avec

$$P(x_i) = f(x_i),\ P'(x_i) = f'(x_i),\ \ x_i = -1, 0, +1$$

.

1) Montrer que pour tout $Q \in \mathbb{R}^5[X]$ on a :

$$\int_{-1}^{1} Q(x)dx = \frac{7}{15}Q(-1) + \frac{16}{15}Q(0) + \frac{7}{15}Q(+1) + \frac{1}{15}Q'(-1) - \frac{1}{15}Q'(+1)$$

2) La formule de la question précédente est une méthode pour approcher l'intégrale de la fonction f sur $[-1,1]$ par l'intégrale du polynôme de Hermite $P(x)$:

$$\int_{-1}^{1} f(x)dx \approx \frac{7}{15}f(-1) + \frac{16}{15}f(0) + \frac{7}{15}f(+1) + \frac{1}{15}f'(-1) - \frac{1}{15}f'(+1)$$

Lorsque f est un polynôme de degré ≤ 5 la formule est exacte. Montrer que cette formule n'est pas, en général, exacte pour les polynômes de degré ≤ 6.
3) Utiliser un résultat du chapitre précédent pour déterminer une estimation de l'erreur de la méthode de la question 1).

Réponse 5.4

1) Il suffit de vérifier la formule pour les monômes : $1,\ x,\ x^2,\ x^3,\ x^4,\ x^5$, car l'intégrale d'une combinaison linéaire est une combinaison linéaire des intégrales et le polynôme est une combinaison linéaire des monômes.
2) La formule n'est pas exacte pour x^6.
3) D'après un théorème du chapitre précédent :

$$\forall x \in [-1,1], \exists c(x) \in [-1,1] / f(x) = P(x) + \frac{(x+1)^2x^2(x-1)^2f^{(6)}(c(x))}{6!}$$

De la même manière que l'exercice précédent, on vérifie que la fonction $f^{(6)}(c(x))$ est continue sur $[-1,1]$ et on a que $(x+1)^2x^2(x-1)^2$ ne change pas de signe sur $[-1,1]$. D'après la formule de la moyenne, il existe $\alpha \in [-1,1]$ tel que :

$$\int_{-1}^{1} x^2(x+1)^2(x-1)^2f^{(6)}(c(x))dx = f^{(6)}(c(\alpha)) \int_{-1}^{1} x^2(x+1)^2(x-1)^2dx$$

Tout calcul fait, on obtient :

$$\begin{array}{ll} \int_{-1}^{1} f(x)dx & = \frac{7}{15}f(-1) + \frac{16}{15}f(0) + \frac{7}{15}f(+1) \\ & + \frac{1}{15}f'(-1) - \frac{1}{15}f'(+1) + \frac{f^{(6)}(c(\alpha))}{4725} \end{array}$$

Exercice 5.5

Pour approcher la solution d'un système différentiel avec condition initiale, on utilise la méthode numérique associée à la fonction Φ suivante :

$$\Phi(x,y,h) = \frac{1}{6}[k_1 + 4k_2 + k_3]$$

$$k_1 = f(x,y)$$

$$k_2 = f(x + \frac{h}{2}, y + \frac{h}{2}k_1)$$

$$k_3 = f(x+h, y + h(-k_1 + 2k_2))$$

Montrer que cette méthode est au moins d'ordre 3.

Réponse 5.5

Pour (x,y) fixé, on pose :

$$\begin{array}{ll} k_1(h) & = f(x,y) \\ k_2(h) & = f(x + \frac{h}{2}, y + \frac{h}{2}k_1(h)) \\ k_3(h) & = f(x+h, y + h(-k_1(h) + 2k_2(h))) \end{array}$$

Pour simplifier les notations, on notera par f à la place de $f(.,.)$, par f_x à la place de $\frac{\partial}{\partial x}f(.,.)$, par f_y à la place de $\frac{\partial}{\partial y}f(.,.)$, par f_{xy} à la place de $\frac{\partial^2}{\partial x \partial y}f(.,.)$, par f_{xx} à la place de $\frac{\partial^2}{\partial x^2}f(.,.)$ et par f_{yy} à la place de $\frac{\partial}{\partial y^2}f(.,.)$.
On a alors :

$$\begin{array}{ll} y'(x) & = f \\ y"(x) & = f_x + f.f_y \\ y"'(x) & = f_{xx} + 2f.f_{xy} + f_x.f_y + f.(f_y)^2 + f^2.f_{yy} \end{array}$$

et

$$\begin{array}{ll} k_1'(h) & = 0 \\ k_2'(h) & = \frac{1}{2}f_x + \frac{1}{2}k_1(h)f_y \\ k_2"(h) & = \frac{1}{4}[f_{xx} + 2k_1(h)f_{xy} + k_1^2(h)f_{yy} \\ k_3'(h) & = f_x + (-k_1(h) + 2k_2(h) + 2hk_2'(h))f_y \\ k_3"(h) & = f_{xx} + 2(-k_1(h) + 2k_2(h) + 2hk_2'(h))f_{xy} + (-k_1(h) \\ & \quad +2k_2(h) + 2hk_2'(h))f_{yy} + (4k_2'(h) + 2hk_2"(h))f_y \end{array}$$

D'où :

$$\begin{array}{ll} k_2'(0) & = \frac{1}{2}(f_x + f.f_y) \\ k_2"(0) & = \frac{1}{4}(f_{xx} + 2f.f_{xy} + f^2.f_{yy}) \\ k_3'(0) & = f_x + f.f_y \\ k_3"(0) & = f_{xx} + 2f.f_{xy} + f.f_{yy} + 2f_y(f_x + f.f_y) \end{array}$$

Un développement limité à l'ordre 2 par rapport à h au voisinage de zéro, (x, y) fixé, de $\Phi(x, y, h)$ nous donne :

$$\begin{array}{ll} \Phi(x,y,h) & = f + \frac{h}{2}(f_x + f.f_y) + \\ & \frac{h^2}{6}[f_{xx} + 2f.f_{xy} + f^2 f_{xy} + (f_x + f.f_y)f_y] + O(h^3) \end{array}$$

et si $x \in [a, b]$ et y dans un compact, alors, le reste est une quantité $O(h^3)$ uniformément par rapport à (x, y).
De même, un développement limité à l'ordre 3 par rapport à h au voisinage de zéro, x fixé de $y(x + h)$ nous donne :

$$y(x+h) - y(x) = y'(x)h + \frac{h^2}{2}y"(x) + \frac{h^3}{6}y"'(x) + O(h^4)$$

et si $x \in [a, b]$ alors, le reste est une quantité $O(h^4)$ uniformément par rapport à x.
Si on remplace les expressions de $y'(x),\ y"(x),\ y"'(x)$, on obtient que :

$$\frac{y(x+h) - y(x)}{h} - \Phi(x, y(x), h) = O(h^3)$$

avec le reste est une quantité $O(h^3)$ uniformément par rapport à $x \in [a, b]$ car $y(x) \in y([a, b]) =$ un compact. D'où, pour toute subdivision $(x_m)_{0 \leq m \leq N}$ de pas uniforme h de l'intervalle $[a, b]$, on a :

$$\max_m || \frac{y(x_{m+1}) - y(x_m)}{h} - \Phi(x_m, y(x_m), h)|| = O(h^3)$$

ce qui prouve, d'après un théorème du cours, que la méthode est au moins d'ordre 3.

Exercice 5.6
Pour approcher la solution d'un système différentiel avec condition initiale, on utilise la méthode numérique associée à la fonction Φ suivante :

$$\Phi(x, y, h) = f(x, y) + \frac{h}{2} g(x + \frac{h}{3}, y + \frac{h}{3} f(x, y))$$

avec

$$g(x, y) = \frac{\partial}{\partial x} f(x, y) + f(x, y) \frac{\partial}{\partial y} f(x, y)$$

Montrer que cette méthode est au moins d'ordre 3.
Réponse 5.6
Même raisonnement que l'exercice précédent.

Chapitre 6

Programmation Linéaire

6.1 Introduction

Un problème de programmation linéaire (on dit aussi un programme linéaire) consiste à minimiser ou à maximiser une fonction linéaire sous contraintes linéaires :

$$(P_0) \qquad \begin{cases} Min\ c^T x = \sum_{i=1}^{n} c_i x_i \\ a_j^T x = d_j,\ j = 1, \cdots, p\ \text{(contraintes-égalités)} \\ a_j^T x \leq d_j,\ j = p+1, \cdots, m\ \text{(contraintes-inégalités)} \end{cases}$$

où $x, c, a_1, a_2, \cdots, a_m$ sont des vecteurs de $\mathbb{R}^n$ et $d_1, d_2, \cdots, d_m$ sont des nombres réels.

On peut toujours exprimer un programme linéaire sous la forme suivante :

$$(P) \qquad \begin{cases} Min\ z(x) = c^T x = \sum_{i=1}^{n} c_i x_i \\ Ax = d \\ x = (x_1\ x_2\ \cdots\ x_n)^T \geq 0 \end{cases}$$

où A est une matrice m lignes (m contraintes-égalités), n colonnes (n variables) et $d \in \mathbb{R}^m$. En effet, il est toujours possible de se ramener à ce cas :

– une contrainte-inégalité se transforme en une contrainte-égalité en ajoutant une variable dite d'écart :

$$a_j^T x \leq d_j \Longleftrightarrow a_j^T x + v_j = d_j, v_j \geq 0$$

– une variable x_i de signe quelconque est remplacée par $x_i^+ - x_i^-$, avec $x_i^+, x_i^- \geq 0$

On dit que (P) est la forme standard de (P_0).

EXEMPLE

$$(P_0) \qquad \begin{cases} Min\ z = 5x_1 - 3x_2 \\ x_1 - x_2 \geq 2 \\ 2x_1 + 3x_2 \leq 4 \\ -x_1 + 6x_2 = 10 \\ x_1 \geq 0, x_2 \in \mathbb{R} \end{cases}$$

En introduisant les variables d'écart, on obtient la forme standard de (P_0) :

$$(P) \quad \begin{cases} Min\ z = 5x_1 - 3(x_2 - x_3) + 0x_4 + 0x_5 \\ x_1 - (x_2 - x_3) - x_4 = 2 \\ 2x_1 + 3(x_2 - x_3) + x_5 = 4 \\ -x_1 + 6(x_2 - x_3) = 10 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0 \end{cases}$$

$$c^T = (5\ \ -3\ 3\ 0\ 0), d^T = (2\ 4\ 10)$$

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

REMARQUE 6.1.1 *On peut toujours supposer que $rang(A) = m$. En effet, si $rang(A) < m$, une ou plusieurs lignes de A peuvent s'exprimer comme combinaison linéaire des autres lignes, alors ou bien l'ensemble des solutions de $Ax = d$ est vide, ou bien les équations qui correspondent à ces lignes sont redondantes et peuvent être éliminées. On obtient alors une nouvelle matrice telle que son rang est égal au nombre de contraintes.*

6.2 Formulation d'un programme linéaire

6.2.1 Exemple : le problème du portefeuille

Supposons que c'est le début de l'an 2000, et que vous êtes un investisseur avec un portefeuille composé de trois types d'actions : 75 actions avec la STB, 1000 actions avec la BNA et 25 actions avec la BIAT. Les prix actuels sont les suivants : 20D par action de la STB, 2D par action de la BNA et 100D par action de la BIAT.

On suppose que l'étude du marché et de certains facteurs économiques vous permettent d'estimer les bénéfices de chaque action durant l'an 2000 et le prix de chaque action à la fin de l'an 2000. Par exemple : le bénéfice de la STB est 5D par action, pas de bénéfice pour la BNA, le bénéfice de la BIAT est de 2D par action, et à la fin de l'an 2000, la valeur de la STB est 18D par action, de la BNA est de 3D par action et de la BIAT est 102D par action.

Vous n'avez pas des frais à payer pour acheter ou vendre des actions et vous pouvez vendre ou acheter une fraction d'action.

Vous devez ajuster votre portefeuille pour maximiser vos bénéfices de l'an 2000. Pour ce faire, vous devez satisfaire les restrictions suivantes :

1. Après ajustement, la valeur de votre portefeuille reste la même ; c'est-à-dire, investir la valeur actuelle de votre portefeuille.

2. Il faut tenir compte de l'$inflation$, la valeur de votre portefeuille à la fin de l'an 2000 doit être augmenter au moins de 5% de sa valeur actuelle.

3. Pour éviter les risques du marché, il faut que votre portefeuille soit équilibré : après ajustement, la valeur de chaque type d'action doit être supérieure ou égale à 25% de la valeur totale de votre portefeuille.

4. L'ajustement des actions se fait une seule fois au début de l'année et le nombre d'actions par type d'action doit être positive.

6.2.2 Programme linéaire du problème du portefeuille

On note par x_1 la variation des actions de type STB, x_2 la variation des actions de type BNA et par x_3 la variation des actions de type BIAT ; c'est-à-dire, si $x_1 = 50$, votre investissement avec la STB devient $75+50 = 125$ actions, si $x_2 = -30$, votre investissement avec la BNA devient $1000-30 = 970$ actions.

Vous devez maximiser le bénéfice de votre portefeuille, c'est-à-dire, maximiser la fonction affine :

$$5(75 + x_1) + 0(1000 + x_2) + 2(25 + x_3)$$

ou encore

$$5x_1 + 2x_3 + 425$$

ce qui revient à minimiser la fonction linéaire

$$\min_{x_1,x_2,x_3} -5x_1 - 2x_3$$

La première contrainte est donnée par la restriction 1 : la valeur actuelle est de

$$20 \times 75 + 2 \times 1000 + 100 \times 25 = 6000$$

le portefeuille ajusté doit avoir la même valeur

$$20(75 + x_1) + 2(1000 + x_2) + 100(25 + x_3) = 6000$$

ou encore

$$20x_1 + 2x_2 + 100x_3 = 0$$

La seconde contrainte est donnée par la restriction 2 : la valeur du portefeuille à la fin de l'an 2000 doit être supérieure ou égale à $6000 + 5 \times 6000/100 = 6300$

$$18(75 + x_1) + 3(1000 + x_2) + 102(25 + x_3) \geq 6300$$

ou encore

$$18x_1 + 3x_2 + 102x_3 \geq -600$$

La troisième contrainte est donnée par la restriction 3 : la valeur de chaque type d'action est supérieure ou égale à 1500

$$\begin{array}{rcl} 20(75 + x_1) & \geq & 1500 \\ 2(1000 + x_2) & \geq & 1500 \\ 100(25 + x_3) & \geq & 1500 \end{array}$$

ou encore

$$\begin{array}{c} x_1 \geq 0 \\ x_2 \geq -250 \\ x_3 \geq -10 \end{array}$$

La quatrième restriction est contenue dans la troisième.

La formulation du problème du portefeuille est donnée par le programme linéaire suivant :

$$\begin{cases} Min\ z = -5x_1 - 2x_3 \\ 20x_1 + 2x_2 + 100x_3 = 0 \\ 18x_1 + 3x_2 + 102x_3 \geq -600 \\ x_1 \geq 0 \\ x_2 \geq -250 \\ x_3 \geq -10 \end{cases}$$

6.3 Définitions et propriétés

Un ensemble convexe de la forme

$$U = \{x \in \mathbb{R}^n / Ax = d,\ x \geq 0\}$$

avec A une matrice $m \times n$, $d \in \mathbb{R}^m$ et $rang(A) = m$, est appelé un polyèdre. Les éléments de U sont appelés les solutions réalisables du programme linéaire (P). On appelle sommet du polyèdre U tout point de U qui ne s'écrit pas comme une combinaison convexe d'autres points de U. On appelle base toute sous matrice B régulière $(m \times m)$ extraite de A (il y a au moins une puisque $rang(A) = m$). On note $1 \leq b_1, b_2, \cdots, b_m \leq n$ les indices des colonnes, appelées colonnes de base, qui forment la sous matrice B, $1 \leq h_1, h_2, \cdots, h_{n-m} \leq n$ les indices des autres colonnes de A, appelées colonnes hors-base et H la sous matrice formée par les colonnes hors-base, c'est à dire, si on note par A_j, $j = 1, \cdots, n$, les colonnes de A, on a :

$$\text{la } k\text{ième colonne de } B = A_{b_k}, 1 \leq k \leq m$$

$$\text{la } k\text{ième colonne de } H = A_{h_k}, 1 \leq k \leq n - m$$

$$\{1, 2, \cdots, n\} = \{b_1, b_2, \cdots, b_m, h_1, h_2, \cdots, h_{n-m}\}$$

Soit $x = (x_1\ x_2\ \cdots\ x_n)^T \in \mathbb{R}^n$ un vecteur quelconque et B une base de A, la notation $x = [x_B, x_H]$ veut dire :

$$x_B = (x_{b_1}\ x_{b_2}\ \cdots\ x_{b_m})^T, x_H = (x_{h_1}\ x_{h_2}\ \cdots\ x_{h_{n-m}})^T$$

d'où :

$$Ax = \sum_{j=1}^{n} x_j A_j = \sum_{k=1}^{m} x_{b_k} A_{b_k} + \sum_{k=1}^{n-m} x_{h_k} A_{h_k}$$

Le système $Ax = d$ est donc équivalent à

$$Bx_B + Hx_H = d \tag{6.1}$$

On appelle solution de base (associée à la base B), la solution particulière de (6.1) obtenue en faisant $x_H = 0$. Le vecteur x_B est alors détérminé d'une façon unique par :

$$x_B = B^{-1}d.$$

Une base B est dite réalisable si la solution de base est réalisable c'est-à-dire $x_B \geq 0$. Une base réalisable est dite dégénérée si l'une des composantes de $x_B = B^{-1}d$ est nulle. Une solution réalisable (resp. une base réalisable) est dite optimale si elle (resp. la solution de base associée) réalise le minimum de (P).

Pour $x = (x_1\ x_2\ \cdots\ x_n)^T \in \mathbb{R}^n$, on note

$$I^*(x) = \{j\ /\ x_j > 0\}.$$

THÉORÈME 6.3.1 *L'ensemble des sommets du polyèdre*

$$U = \{x \in \mathbb{R}^n / Ax = d,\ \geq 0\}$$

correspond à l'ensemble des solutions de base des bases réalisables. En particulier, l'ensemble des sommets est fini.

LEMME 6.3.1 *Soit u un sommet de U. Alors, la famille $\{A_j / j \in I^*(u)\}$ est libre.*

Démonstration (Lemme)

Supposons le contraire, c'est à dire :

$$\exists(\alpha_j)_{j\in I^*(u)} / \sum_{j\in I^*(u)} \alpha_j A_j = 0$$

avec au moins un $\alpha_j \neq 0$. On pose

$$\alpha_j = 0, \forall j \notin I^*(u)$$

$$\alpha^T = (\alpha_1 \ \alpha_2 \ \cdots \ \alpha_n)$$

$$I_1 = \{j/\alpha_j > 0\}, I_2 = \{j/\alpha_j < 0\}$$

Il est facile de voir que

$$A(u - \theta\alpha) = d, \forall\theta \in \mathbb{R}.$$

D'autre part, si $I_1 \neq \emptyset$, alors :

$$\begin{array}{rcl} \forall j \in I_1, u_j - \theta\alpha_j \geq 0 & \Longleftrightarrow & \forall j \in I_1, \theta\alpha_j \leq u_j \\ & \Longleftrightarrow & \forall j \in I_1, \theta \leq u_j/\alpha_j \\ & \Longleftrightarrow & \theta \leq \theta_1 = \min_{j\in I_1}(u_j/\alpha_j) \end{array}$$

sinon, on obtient que $u - \theta\alpha \geq 0$ pour tout $\theta \geq 0$, on pose dans ce cas $\theta_1 = +\infty$. De même, si $I_2 \neq \emptyset$, alors :

$$\begin{array}{rcl} \forall j \in I_2, u_j - \theta\alpha_j \geq 0 & \Longleftrightarrow & \forall j \in I_2, \theta\alpha_j \leq u_j \\ & \Longleftrightarrow & \forall j \in I_2, \theta \geq u_j/\alpha_j \\ & \Longleftrightarrow & \theta \geq \theta_2 = \max_{j\in I_2}(u_j/\alpha_j) \end{array}$$

sinon, on obtient que $u - \theta\alpha \geq 0$ pour tout $\theta \leq 0$, on pose dans ce cas $\theta_2 = -\infty$. Il est clair que $\theta_2 < 0$ et $\theta_1 > 0$ et que

$$\forall\theta \in]\theta_2, \theta_1[, u - \theta\alpha \in U$$

D'autre part, soit $\theta \in]\theta_2, \theta_1[- \{0\}$ tel que $-\theta \in]\theta_2, \theta_1[$ (un tel θ existe car $\theta_2 < 0 < \theta_1$), on a :

$$u \pm \theta\alpha \in U, \text{ et } u = \frac{1}{2}(u - \theta\alpha) + \frac{1}{2}(u + \theta\alpha)$$

par conséquent u s'écrit comme combinaison convexe d'éléments de U, ce qui contredit le fait que u est un sommet.

Démonstration (Théorème)

Soit u un sommet de U, d'après le lemme précédent la famille $\{A_j \ / \ j \in I^*(u)\}$ est libre, donc, on peut compléter cette famille par d'autres colonnes de A pour obtenir une base B. Il est clair que

$$Bu_B = \sum_{i\in I^*(u)} u_i A_i = d$$

d'où u est la solution de base de la base B.

Inversement, soit B une base réalisable et soit $u = [u_B, u_H] = [B^{-1}d, 0]$ la solution de base de la base B. Supposons que u n'est pas un sommet de U :

$$\exists \lambda \in]0,1[, \exists v \neq w \in U / u = \lambda v + (1-\lambda)w$$

d'où

$$\exists \lambda \in]0,1[, \exists v \neq w \in U / \left\{ \begin{array}{l} u_B = \lambda v_B + (1-\lambda)w_B \\ u_H = \lambda v_H + (1-\lambda)w_H \end{array} \right.$$

Or, $u_H = 0$, ce qui prouve que $v_H = w_H = 0$, car $v_H \geq 0$ et $w_H \geq 0$. D'autre part :

$$\begin{array}{rcl} u \in U & \Longrightarrow & Bu_B + Hu_H = d \\ v \in U & \Longrightarrow & Bv_B + Hv_H = d \\ w \in U & \Longrightarrow & Bw_B + Hw_H = d \end{array}$$

Par conséquent

$$u_B = v_B = w_B = B^{-1}d$$

d'où $u = v = w$. Donc u est un sommet de U. Enfin, la matrice A est constituée de n colonnes, et une base est constituée de m colonnes, donc, le nombre de bases réalisables est au plus égal à C_n^m.

COROLLAIRE 6.3.1 *Soit $u \in U$. Alors, u est un sommet de U si et seulement si la famille $\{A_j ; j \in I^*(u)\}$ est une famille libre.*

DÉMONSTRATION :

Soit $u \in U$ un sommet de U, d'après le théorème 6.3.1, il existe m- colonnes de A : $A_{b_1}, A_{b_2}, \cdots, A_{b_m}$, qui forment une base réalisable B telle que :

$$u = [u_B \quad 0] \quad avec \quad u_B = (u_{b_1} \quad u_{b_2} \cdots u_{b_m})^T = B^{-1}d$$

d'où, $I^*(u) \subset \{b_1, b_2, \cdots, b_m\}$. Comme B est inversible, donc, ses colonnes $A_{b_1}, A_{b_2}, \cdots, A_{b_m}$ sont linéairement indépendantes et par suite la famille $\{A_j ; j \in I^*(u)\}$ est libre (car la sous-famille d'une famille libre est une famille libre).

Soit $u \in U$ tel que $\{A_j ; j \in I^*(u)\}$ est une famille libre, donc, on peut compléter cette famille par d'autres colonnes de A pour obtenir une base $B = \{A_{b_1}, A_{b_2}, \cdots, A_{b_m}\}$ avec $I^*(u) \subset \{b_1, b_2, \cdots, b_m\}$. D'autre part :

$$d = Au = \sum_{j \in I^*(u)} u_j A_j = \sum_{i=1}^{m} u_{b_i} A_{b_i}$$

ce qui prouve que $u = [u_B \; 0]$ est la solution de base associée à la base B formée par les colonnes $\{A_{b_1}, A_{b_2}, \cdots, A_{b_m}\}$.

THÉORÈME 6.3.2 *Un polyèdre non vide*

$$U = \{x \in \mathbb{R}^n / Ax = d, x \geq 0\}$$

possède au moins un sommet.

DÉMONSTRATION : Soit $u \in U$. Supposons que la famille $\{A_j; j \in I^*(u)\}$ est liée, donc

$$\exists(\alpha_j)_{j\in I^*(u)} / \sum_{j\in I^*(u)} \alpha_j A_j = 0 \tag{6.2}$$

avec au moins un $\alpha_j \neq 0$, $j \in I^*(u)$. Quitte à multiplier (6.2) par -1, on peut supposer que $I = \{j \in I^*(u)/\alpha_j > 0\} \neq \emptyset$. Soit

$$\theta = \min_{j\in I} \frac{u_j}{\alpha_j} = \frac{u_{j_0}}{\alpha_{j_0}}$$

on considère le vecteur $w \in \mathbb{R}^n$:

$$w_j = \begin{cases} u_j - \theta\alpha_j, j \in I^*(u) \\ 0, sinon \end{cases}$$

Montrons que $w \in U$; en effet, on a :

$$\begin{cases} Aw = Au - \theta \sum_{j\in I^*(u)} \alpha_j A_j = Au = d \\ w_j = u_j - \theta\alpha_j \geq 0, \forall j \in I \\ w_j = u_j - \theta\alpha_j \geq u_j > 0, \forall j \in I^*(u)\backslash I \\ w_j = 0, \forall j \notin I^*(u) \end{cases}$$

D'autre part,

$$\begin{array}{ccc} j_0 & \in & I^*(u) \\ w_{j_0} & = & u_{j_0} - \theta\alpha_{j_0} = 0 \\ I^*(w) & \subset & I^*(u) \end{array}$$

Par conséquent

$$I^*(w) \subset I^*(u)\backslash\{j_0\}$$

Si la famille $\{A_j; j \in I^*(w)\}$ est liée, on recommence le même raisonnement avec w au lieu de u. A chaque étape le nombre d'éléments de la famille liée décroît au moins de 1 ; alors, on obtient nécessairement après un nombre fini d'étapes un vecteur $w \in U$ tel que la famille $\{A_j; j \in I^*(w)\}$ soit libre ou $I^*(w) = \emptyset$, dans ce dernier cas $w = 0$. Il suffit donc de prouver que si $0 \in U$, 0 est un sommet de U. Supposons qu'il existe $0 < \lambda < 1$ et $v_1 \neq v_2 \in U$ tel que :

$$0 = \lambda v_1 + (1 - \lambda)v_2$$

comme $v_1 \geq 0$ et $v_2 \geq 0$, nécessairement $v_1 = v_2 = 0$.

PROPOSITION 6.3.1 *On suppose que U est non vide et que la fonction linéaire $z(x) = c^T x$ est minorée sur U. Soit $v \in U$. Si v n'est pas un sommet de U, alors, il existe $\bar{v} \in U$ tel que :*

$$z(\bar{v}) \leq z(v)$$

et

$$card(I^*(\bar{v})) < card(I^*(v))$$

DÉMONSTRATION : Si v n'est pas un sommet, il existe $u_1 \neq u_2 \in U$ et $0 < \lambda < 1$ tels que :

$$v = \lambda u_1 + (1 - \lambda)u_2$$

L'ensemble des indices des composantes non nulles de $y = u_1 - u_2$ est non vide car $u_1 \neq u_2$ et il est contenu dans $I^*(v)$ car si $y_i \neq 0$ nécessairement $v_i \neq 0$. D'autre part, $Ay = 0$ d'où :

$$z(v + \theta y) = z(v) + \theta z(y), A(v + \theta y) = Av = d, \forall \theta \in \mathbb{R},$$

On peut supposer que

$$z(y) \geq 0 \; et \; I^*(y) \neq \emptyset$$

en effet, dans le cas $z(y) \leq 0$ et y possède des composantes négatives, on remplace y par $-y$, c'est à dire, on prend $y = u_2 - u_1$ au lieu de $u_1 - u_2$. Le cas $z(y) < 0$ et les composantes de y sont toutes positives est impossible car :

$$v + \theta y \in U, \forall \theta \geq 0$$

et comme $z(y) < 0$, on obtient

$$\lim_{\theta \longrightarrow +\infty} z(v + \theta y) = \lim_{\theta \longrightarrow +\infty} z(v) + \theta z(y) = -\infty$$

ce qui contredit le fait que z est minorée. Il reste seulement le cas $I^*(y) = \emptyset$ et $z(y) > 0$ qui est aussi impossible car :

$$v - \theta y \in U, \forall \theta \geq 0$$

et comme $z(y) > 0$, on obtient

$$\lim_{\theta \longrightarrow +\infty} z(v - \theta y) = \lim_{\theta \longrightarrow +\infty} z(v) - \theta z(y) = -\infty$$

ce qui contredit le fait que z est minorée.

On pose alors :

$$\bar{\theta} = \min\{\frac{v_i}{y_i} / i \in I^*(y)\} = \frac{v_{i_0}}{y_{i_0}} > 0$$

et

$$\bar{v} = v - \bar{\theta} y$$

Montrons que $\bar{v} \in U$; en effet, on a $A\bar{v} = d$, il suffit donc de prouver que $\bar{v} \geq 0$: pour $i \in I^*(y)$:

$$\begin{array}{rcl} \bar{v}_i & = & v_i - \bar{\theta} y_i \\ & = & y_i(v_i/y_i - \bar{\theta}) \\ & \geq & y_i(\bar{\theta} - \bar{\theta}) \\ & = & 0 \end{array}$$

et pour $i \notin I^*(y)$

$$\bar{v}_i = v_i - \bar{\theta} y_i \geq v_i \geq 0, (\bar{\theta} \geq 0)$$

D'autre part, $I^*(y) \subset I^*(v)$, d'où :

$$I^*(\bar{v}) \subset I^*(v)$$

et on a :

$$i_0 \in I^*(v) \text{ et } \bar{v}_{i_0} = 0.$$
$$z(\bar{v}) = z(v - \bar{\theta}y) = z(v) - \bar{\theta}z(y) \leq z(v)$$

THÉORÈME 6.3.3 *On suppose que $U \neq \emptyset$. Alors on a l'une des assertions suivantes :*

a) inf$\{z(x); x \in U\} = -\infty$

ou bien

b) il existe un sommet w de U tel que :

$$z(w) = \min_{x \in U} z(x).$$

DÉMONSTRATION : L'ensemble des sommets de U est fini et non vide, il existe donc un sommet w de U tel que

$$z(w) = \min\{z(v) \ / \ v \text{ est un sommet de } U\}$$

Si z est non minorée, c'est alors l'assertion a) qui est vérifiée. Supposons que z est minorée. Soit $\varepsilon > 0$. Par définition de l'inf., il existe $v \in U$ tel que

$$z(v) \leq \inf\{z(x); x \in U\} + \varepsilon$$

Si v n'est pas un sommet, on applique plusieurs fois la proposition précédente jusqu'à ce que l'on obtienne un $\bar{v} \in U$ tel que :

$$\begin{cases} z(\bar{v}) \leq z(v) \leq \inf\{z(x); x \in U\} + \varepsilon \\ I^*(\bar{v}) = \emptyset \text{ ou } \bar{v} \text{ est un sommet} \end{cases}$$

or, si $I^*(\bar{v}) = \emptyset$, nécessairement $\bar{v} = 0$ et donc $\bar{v}$ est un sommet de U. Par conséquent :

$$z(w) \leq z(\bar{v}) \leq \inf\{z(x); x \in U\} + \varepsilon$$

Comme le sommet w est indépendant de ε, alors $z(w) = \inf\{z(x); x \in U\}$.

6.4 Résolution des programmes linéaires : Algorithme du simplexe

LEMME 6.4.1 *Soit C un convexe fermé de $\mathbb{R}^k$, $k \geq 1$, et soit $y \in \mathbb{R}^k \backslash C$. Alors, il existe $a \in \mathbb{R}^k$ tel que :*

$$a^T y < \inf\{a^T x; x \in C\}$$

DÉMONSTRATION : On pose

$$\delta = \inf_{x \in C} \| x - y \|_2 \tag{6.3}$$

L'inf. de (6.3) est atteint, en effet, C est un fermé de $\mathbb{R}^k$ et $y \notin C$, alors, $\delta > 0$; d'où, l'inf. sur C est aussi égal à l'inf. sur $\{x \in C / \ \| x - y \|_2 \leq 2\delta\}$ qui est un compact, ce qui prouve que l'inf. est atteint, soit

$$\delta = \| \bar{x} - y \|_2 = \inf_{x \in C} \| x - y \|_2$$

On pose $a = \bar{x} - y$. Comme C est convexe, alors, pour tout $0 < \alpha < 1$ et $x \in C$, le vecteur $\bar{x} + \alpha(x - \bar{x})$ est dans C ; d'où

$$\delta^2 \leq \| \bar{x} + \alpha(x - \bar{x}) - y \|_2^2 = \delta^2 + 2\alpha(\bar{x} - y)^T(x - \bar{x}) + \alpha^2 \| x - \bar{x} \|_2^2$$

soit

$$2(\bar{x} - y)^T(x - \bar{x}) + \alpha \| x - \bar{x} \|_2^2 \geq 0$$

par conséquent, lorsque α tend vers 0+, on obtient que :

$$a^T(x - \bar{x}) = (\bar{x} - y)^T(x - \bar{x}) \geq 0$$

d'où

$$a^T x \geq a^T \bar{x} = a^T(\bar{x} - y) + a^T y = \delta^2 + a^T y$$

ce qui prouve que

$$\inf_{x \in C} a^T x \geq a^T y + \delta^2 > a^T y.$$

PROPOSITION 6.4.1 *On suppose que le programme linéaire* (P) *possède une solution optimale* $x^* \in U$*, alors :*

$$\exists \lambda^* \in \mathbb{R}^m / \begin{cases} c - A^T \lambda^* \geq 0 \\ d^T \lambda^* = c^T x^* \end{cases}$$

DÉMONSTRATION : Soit $z^* = c^T x^*$. On pose

$$C = \{(r, w) \in \mathbb{R} \times \mathbb{R}^n \ / \ r = tz^* - c^T x, \ w = td - Ax \ ; \ t \in \mathbb{R}_+ \ et \ x \in \mathbb{R}_+^n\}$$

Il est facile de vérifier que C est un convexe fermé, de plus, C est un cône :

$$\forall t \geq 0, \forall (r, w) \in C \Longrightarrow t(r, w) = (tr, tw) \in C$$

Montrons que $(1, 0) \notin C$; en effet, supposons le contraire, alors :

$$\exists (t_0, x_0) \in \mathbb{R}_+ \times \mathbb{R}_+^n / w = t_0 d - Ax_0 = 0, 1 = t_0 z^* - c^T x_0$$

Si $t_0 > 0$, alors, x_0/t_0 est une solution réalisable du programme linéaire (P) et on a

$$c^T(x_0/t_0) = z^* - 1/t_0 < z^* = c^T x^* = \min_{x \in U} c^T x$$

ce qui est absurde. Si $t_0 = 0$, alors,

$$Ax_0 = 0, \ x_0 \geq 0 \ et \ c^T x_0 = -1$$

d'où, pour tout x solution réalisable du programme (P), $x + \alpha x_0$ est aussi une solution réalisable du programme (P) pour tout α positif, ce qui est absurde, car :

$$c^T(x + \alpha x_0) = c^T x - \alpha \longrightarrow -\infty$$

quand α tend vers $+\infty$. Par conséquent, C est un fermé et $y = (1, 0) \notin C$, d'après le lemme précédent, il existe $a = (s, \lambda) \in \mathbb{R} \times \mathbb{R}^m$ tel que

$$a^T y = s < \delta = \inf\{a^T x = s.r + \lambda^T w, (r, w) \in C\}$$

Le nombre δ est nécessairement positif ; sinon,

$$\exists x = (r, w) \in C / a^T x = s.r + \lambda^T w < 0$$

et comme C est un cône, on obtient que l'inf. est égal à $-\infty$ ce qui contredit le fait que $\delta > s$. D'autre part, $(0, 0) \in C$, d'où

$$\delta \leq 0$$

ce qui prouve que $\delta = 0$. Donc, $s < 0$ et on a :

$$s.r + \lambda^T w \geq 0, \forall (r, w) \in C$$

Si on note $\lambda^* = \lambda/(-s)$, on obtient :

$$-r + (\lambda^*)^T w \geq 0, \forall (r, w) \in C$$

soit encore, en utilisant que $(r, w) \in C$:

$$-(tz^* - c^T x) + (\lambda^*)^T (td - Ax) \geq 0$$

d'où

$$t(d^T \lambda^* - z^*) + (c - A^T \lambda^*)^T x \geq 0, \forall t \geq 0, \forall x \geq 0$$

Pour $t = 0$, on obtient que

$$A^T \lambda^* \leq c \tag{6.4}$$

Pour $x = 0$ et $t = 1$, on obtient :

$$d^T \lambda^* \geq z^*$$

Utilisons le fait que $x^* \geq 0$ et que $Ax^* = d$, et multiplions l'équation (6.4) par x^*, on obtient :

$$d^T \lambda^* \leq c^T x^* = z^*$$

d'où

$$d^T \lambda^* = c^T x^*$$

THÉORÈME 6.4.1 *Soit B une base réalisable. Une condition suffisante pour que B soit une base réalisable optimale est que :*

$$\bar{c}_H = (\bar{c}_{h_k})_{k=1,n-m} = c_H - (c_B^T B^{-1} H)^T \geq 0, \tag{6.5}$$

avec $c = [c_B, c_H]$. Si de plus, la solution de base est non dégénérée, alors, la condition (6.5) est nécessaire.

DÉMONSTRATION :

Soit B une base réalisable. Supposons que la condition (6.5) est vérifiée. Par définition, la solution de base $x^0 = [x_B^0, x_H^0] = [B^{-1}d, 0]$ est réalisable. Soit $x = [x_B, x_H] \in U$, on a :

$$x_B \geq 0, x_H \geq 0, \text{ et } Bx_B + Hx_H = d$$

d'où :

$$\begin{array}{rcl} z(x) & = & c_B^T x_B + c_H^T x_H \\ & = & c_B^T (B^{-1}d - B^{-1}Hx_H) + c_H^T x_H \\ & = & c_B^T B^{-1}d + (c_H^T - c_B^T B^{-1}H)x_H \\ & \geq & c_B^T B^{-1}d \\ & = & z(x^0) \end{array}$$

Supposons que la base réalisable B est optimale, c'est à dire que la solution de base, qu'on note x^*, est optimale et que x^* est non dégénérée. D'après la proposition ci-dessus, il existe $\lambda^* \in \mathbb{R}^m$ tel que :

$$\left\{ \begin{array}{l} c - A^T\lambda^* \geq 0 \\ d^T\lambda^* = c^T x^* \end{array} \right.$$

de la première équation, on déduit que :

$$\left\{ \begin{array}{l} c_{b_k} - A_{b_k}^T\lambda^* \geq 0 \ \forall k = 1, \cdots, m \\ c_{h_k} - A_{h_k}^T\lambda^* \geq 0 \ \forall k = 1, \cdots, n-m \end{array} \right.$$

et de la deuxième équation, en remplaçant d par Ax^*, en utilisant les inégalités données par la première équation et le fait que la solution x^* est non dégénérée, on déduit que :

$$c_{b_k} - A_{b_k}^T\lambda^* = 0 \ \forall k = 1, \cdots, m$$

d'où

$$c_B = B^T\lambda^* \text{ et } c_H - H^T\lambda^* \geq 0$$

par conséquent

$$(B^T)^{-1}c_B = \lambda^*$$

d'où

$$c_H - H^T(B^T)^{-1}c_B \geq 0.$$

COROLLAIRE 6.4.1 *Soit B une base réalisable quelconque et x^0 la solution de base correspondante. On suppose qu'il existe une composante strictement négative de $\bar{c}_H$ d'indice hors-base h_s, $\bar{c}_H$ est le vecteur défini par (6.5) et on pose*

$$\bar{d} = B^{-1}d = (\bar{d}_k)_{k=1,m}, \bar{A}_{h_s} = B^{-1}A_{h_s} = (\bar{a}_{i,h_s})_{i=1,m} \tag{6.6}$$

alors :

(a) *ou bien les composantes de $\bar{A}_{h_s}$ sont toutes négatives, et dans ce cas l'inf. sur U est égal à $-\infty$.*

(b) *ou bien*

$$\emptyset \neq I^*(\bar{A}_{h_s}) \subset I^*(\bar{d})$$

et dans ce cas on met en évidence une autre base $\bar{B}$ réalisable et une solution de base $\bar{x}$ (associée à $\bar{B}$) telle que : $z(\bar{x}) < z(x^0)$

(c) *ou bien*

$$\exists i \in I^*(\bar{A}_{h_s}) / i \notin I^*(\bar{d})$$

et dans ce cas on met en évidence une autre base $\bar{B}$ réalisable avec la même solution de base x_0.

DÉMONSTRATION : L'équation (6.5) s'écrit sous la forme suivante :

$$\bar{c}_{h_j} = c_{h_j} - \sum_{k=1}^{m} \bar{a}_{k,h_j} c_{b_k}, 1 \leq j \leq n-m$$

et par hypothèse, on a :

$$\bar{c}_{h_s} < 0$$

Soit

$$\bar{d} = B^{-1}d = (\bar{d}_k)_{k=1,m} \geq 0$$

On note par $(e_i)_{i=1,n}$ la base canonique de $\mathbb{R}^n$ et pour $\theta \in \mathbb{R}^+$ on pose

$$x(\theta) = (\sum_{k=1}^{m} (\bar{d}_k - \theta \bar{a}_{k,h_s}) e_{b_k}) + \theta e_{h_s} \tag{6.7}$$

on obtient que

$$\begin{array}{rcl} Ax(\theta) & = & A(\sum_{k=1}^{m} \bar{d}_k e_{b_k}) + \theta(A_{h_s} - \sum_{k=1}^{m} \bar{a}_{k,h_s} A_{b_k}) \\ & = & Ax^0 + \theta(A_{h_s} - B\bar{A}_{h_s}) \\ & = & d \end{array} \tag{6.8}$$

D'autre part

$$\begin{array}{rcl} z(x(\theta)) & = & z(x^0) + \theta(z(e_{h_s}) - \sum_{k=1}^{m} \bar{a}_{k,h_s} z(e_{b_k})) \\ & = & z(x^0) + \theta(c_{h_s} - \sum_{k=1}^{m} \bar{a}_{k,h_s} c_{b_k})) \\ & = & z(x^0) + \theta \bar{c}_{h_s} \end{array} \tag{6.9}$$

(a) Si toutes les composantes de $\bar{A}_{h_s}$ sont négatives, d'après (6.7) et (6.8) $x(\theta)$ est une solution réalisable quelque soit le nombre $\theta > 0$ et d'après (6.9) $z(x(\theta))$ tend vers $-\infty$ quand θ tend vers $+\infty$.

(b) Dans ce cas, la valeur de θ ne peut pas être augmentée indéfiniment, la plus grande valeur de θ est donnée par :

$$\hat{\theta} = \min_{i \in I^*(\bar{A}_{h_s})} \frac{\bar{d}_i}{\bar{a}_{i,h_s}} = \frac{\bar{d}_r}{\bar{a}_{r,h_s}}$$

La nouvelle solution $\hat{x} = x(\hat{\theta}) \in U$ et vérifie :

$$\begin{array}{rcl} z(\hat{x}) & < & z(x^0) \\ \hat{x}_{b_r} & = & 0 \\ \hat{x}_{h_s} & = & \hat{\theta} > 0 \end{array}$$

il suffit donc de prouver que $\hat{B} = B + \{A_{h_s}\} - \{A_{b_r}\}$ est une base. Supposons que $\hat{B}$ n'est pas une base, soit $I = \{1, 2, \cdots, m\} \backslash \{r\}$, alors

$$\exists (\alpha_k)_{k \in I} / A_{h_s} = \sum_{k \in I} \alpha_k A_{b_k} \tag{6.10}$$

D'autre part, $\bar{a}_{r,h_s} > 0$ et d'après (6.6) on a :

$$A_{h_s} = \sum_{k=1}^{m} \bar{a}_{k,h_s} A_{b_k} \tag{6.11}$$

d'où, en combinant (6.11) et (6.10), on obtient :

$$\sum_{k\in I}(\bar{a}_{k,h_s}-\alpha_k)A_{b_k}+\bar{a}_{r,h_s}A_{b_r}=0$$

ce qui est impossible car $(A_{b_k})_{k=1,m}$ est une base et $\bar{a}_{r,h_s}\neq 0$.

(c) Même raisonnement que (b), sauf que $\hat{\theta}=0$, on change de base sans que la solution de base change.

L'intérêt du corollaire ci-dessus vient du fait qu'une des méthodes de résolution des programmes linéaires en découle directement : *l'algorithme du simplexe*

Algorithme du simplexe

On suppose qu'on dispose d'une base réalisable de départ B^0

(a) $B=B^0$ *base réalisable de départ. Itération* $k=0$.

(b) $k\leftarrow k+1$

(c) *à l'itération* k *soit* B *la base courante calculer :*

$$\begin{aligned}\bar{d} &= B^{-1}d\\ \bar{H} &= B^{-1}H\\ \bar{c}_H &= c_H-\bar{H}^Tc_B=(\bar{c}_{h_k})_{k=1,n-m},\end{aligned}$$

(d) *Si* $\bar{c}_H\geq 0$, *STOP : l'optimum est atteint.*
Sinon, il existe h_s *tel que :* $\bar{c}_{h_s}<0$ *alors :*

(e) *Poser* $\bar{A}_{h_s}=B^{-1}A_{h_s}=(\bar{a}_{i,h_s})_{i=1,m}$=*la sième colonne de* $\bar{H}$.
Si $\bar{a}_{i,h_s}\leq 0,\forall i=1,\cdots,m$, *STOP : optimum non borné* $(-\infty)$.
Sinon, calculer :

$$\bar{\theta}=\frac{\bar{d}_r}{\bar{a}_{r,h_s}}=\min_{i\in I^*(\bar{A}_{h_s})}\frac{\bar{d}_i}{\bar{a}_{i,h_s}}$$

(f) *Soit* $\bar{x}$ *tel que :*

$$\begin{aligned}\bar{x}_{b_k} &= \bar{d}_k-\bar{\theta}\bar{a}_{k,h_s},\forall k=1,m\\ \bar{x}_{h_s} &= \bar{\theta}\\ \bar{x}_{h_k} &= 0,\forall k\neq s\end{aligned}$$

(x_{b_r} quitte la base et x_{h_s} entre en base). $\bar{x}$ est la solution de base de la nouvelle base $\hat{B}=B+\{A_{h_s}\}-\{A_{b_r}\}$. *Retourner en (b).*

La méthode du simplexe, due à G.B.Dantzig, est une procédure qui permet de passer, le long de la frontière de U, d'un sommet à un autre sommet adjacent, ce qui permet d'atteindre le sommet optimal. Algébriquement, elle permet de construire une suite de bases adjacentes

$$B^0,B^1,\cdots,B^k,\cdots$$

et des sommets

$$x^0,x^1,\cdots,x^k,\cdots$$

avec x^k, $k=0,1,2,\cdots$, une solution de base de B^k tel que :

$$z(x^0)\geq z(x^1)\geq z(x^2)\geq\cdots\geq z(x^k)\geq\cdots$$

REMARQUE 6.4.1 *Dans l'algorithme du simplexe, le choix en (d) de la variable x_{h_s} qui va entrer en base, dans le cas où plusieurs composantes de $\bar{c}_H$ sont strictement négatives, est quelconque. Dans la pratique, l'expérience montre qu'il est plus intéressant de choisir la variable qui correspond à la composante la plus négative de $\bar{c}_H$.*

REMARQUE 6.4.2 *Si la solution de l'étape k est non dégénérée le nombre $\bar{\theta}$ est strictement positif, dans ce cas $z(x^{k+1}) < z(x^k)$. Sinon, le nombre $\bar{\theta}$ peut-être nul (c'est l'assertion (c) du corollaire 6.4.1), dans ce cas la base change mais la valeur de z ne change pas et il est possible donc de retrouver, après un certain nombre de changement de base à une base déjà rencontrée et de cycler indéfiniment. On appelle ce phénomène "le phénomène de cyclage". L'expérience montre que le phénomène de cyclage est très rare et la plupart des programmes utilisés dans l'industrie ne sont pas munis d'une stratégie contre le cyclage. Une des stratégies contre le cyclage est donnée par Bland (1977) :*

— parmi toutes les variables x_{h_s} qui peuvent entrer en base (qui correspondent à des composantes strictement négatives de $\bar{c}_H$) choisir celle de plus petit indice.

— parmi toutes les variables x_{b_r} qui peuvent quitter la base, choisir celle de plus petit indice.

Il est clair que si le phénomène de cyclage ne se produit pas, l'algorithme du simplexe converge en un nombre fini d'itérations, car le nombre de sommets de U est fini. On montre qu'avec la stratégie de Bland, la méthode du simplexe converge au bout d'un nombre fini d'itérations, même dans le cas de dégénérescence.

REMARQUE 6.4.3 *La base B de l'étape k et $\hat{B}$ de l'étape $k+1$ ne diffèrent que d'une colonne : la colonne r de B $(= A_{b_r})$ est remplacée par la colonne h_s de A $(= A_{h_s})$. Soit C la matrice $m \times m$ définie par :*

$$\begin{array}{lcl} c_{i,j} & = & \delta_{i,j}, \forall 1 \leq i \leq m, \forall 1 \leq j \leq m, j \neq r \\ c_{i,r} & = & -\bar{a}_{i,h_s}/\bar{a}_{r,h_s}, \forall 1 \leq i \leq m, i \neq r \\ c_{r,r} & = & 1/\bar{a}_{r,h_s} \end{array}$$

Il est facile de voir que la matrice C est inversible car ses colonnes sont linéairement indépendantes. Si on note par $(f_i)_{i=1,m}$ la base canonique de $\mathbb{R}^m$, il est clair que :

$$Cf_i = f_i, \forall 1 \leq i \leq m, i \neq r$$
$$C\bar{A}_{h_s} = f_r$$

d'où

$$\begin{cases} \hat{B}f_i = Bf_i = BC^{-1}f_i, & i \neq r \\ \hat{B}f_r = A_{h_s} = B\bar{A}_{h_s} = BC^{-1}f_r, \end{cases}$$

ce qui prouve que

$$\hat{B} = BC^{-1}$$

donc

$$\hat{B}^{-1}d = CB^{-1}d = C\bar{d}$$
$$\hat{B}^{-1}A = CB^{-1}A$$

et de la dernière équation, on déduit facilement $\bar{H}$ puisque les nouveaux indices de la matrice hors-base $h_1,\ h_2, \cdots,\ h_{n-m}$ sont connus.

6.5 Base de départ

Deux questions se posent avant de lancer la méthode du simplexe pour résoudre un programme linéaire : la première est de savoir si l'ensemble des solutions réalisables

$$U = \{x \in \mathbb{R}^n_+ / Ax = d\}$$

avec A une matrice $m \times n$ de rang m et $d \in \mathbb{R}^m$, est vide ou non. La deuxième question, dans le cas où $U \neq \emptyset$, est alors de déterminer un sommet de U, ce qui est équivalent à une base réalisable de A, pour initialiser la méthode du simplexe.

Pour répondre aux questions précédentes, une des méthodes consiste à construire un autre programme linéaire tel que l'un de ses sommets se calcule facilement, il possède une solution optimale et en fonction de sa solution on donnera une réponse. Le cas $m = 1$ est évident. Pour $m \geq 2$, il y a au moins une colonne de A, A_{i_0}, qui n'est pas un multiple de d, sinon $rang(A) = 1 < m$. On pose

$$r_0 = d - A_{i_0}$$

$$\bar{A} = [A \;\; r_0], \text{ matrice } m \times (n+1)$$

$$\bar{c} = (0 \;\cdots\; 0 \; 1)^T \in \mathbb{R}^{n+1}$$

$$\tilde{x} \in \mathbb{R}^{n+1} / \tilde{x}_{i_0} = \tilde{x}_{n+1} = 1, \tilde{x}_i = 0, \forall i_0 \neq i \neq n+1$$

On définit le programme linéaire $(\bar{P})$ par :

$$(\bar{P}) \qquad \begin{cases} \min \bar{z}(\bar{x}) = \bar{c}^T \bar{x} = \bar{x}_{n+1} \\ \bar{x} \in \bar{U} = \{\bar{x} \in \mathbb{R}^{n+1} / \bar{A}\bar{x} = d, \bar{x} \geq 0\} \end{cases}$$

On vérifie facilement que $\tilde{x}$ est un sommet de $\bar{U}$ et que $\bar{z}$ est minorée par zéro sur $\bar{U}$. La méthode du simplexe permet donc de déterminer un sommet $\xi = (\xi_1 \; \xi_2 \; \cdots \; \xi_n \; \xi_{n+1})^T$ de $\bar{U}$ solution optimale du programme linéaire $(\bar{P})$. Il est clair que $\xi_{n+1} = 0$ si et seulement si $U \neq \emptyset$. De plus, si $\xi_{n+1} = 0$ la famille $\{A_j, j \in I^*(\xi)\}$ est libre ; d'où, $(\xi_1 \; \xi_2 \; \cdots \; \xi_n)^T$ est un sommet de U (pour plus de détails, consulter les exercices de ce chapitre).

6.6 Etude d'un exemple

Pour résoudre, par la méthode du simplexe, un programme linéaire

$$(P) \qquad \begin{cases} \min z(x) = c^T x = \sum_{i=1}^n c_i x_i \\ x \in U = \{x \in \mathbb{R}^n / Ax = d \; et \; x = (x_1 \; x_2 \; \cdots \; x_n)^T \geq 0\} \end{cases}$$

on met toutes les informations de chaque itération dans un tableau de la forme suivante :

x_1	x_2	x_3	x_4	$\cdots\cdots$	x_n	
$B^{-1}A$						$\bar{d}$
c_1	c_2	c_3	c_4	$\cdots\cdots$	c_n	
h_2	b_3	h_1	h_3	$\cdots\cdots$	b_7	
$\bar{c}_{h_2}$	$*$	$\bar{c}_{h_1}$	$\bar{c}_{h_3}$	$\cdots\cdots$	$*$	

avec B la base réalisable de l'itération en cours, $b_1, b_2, \cdots, b_m$ les indices des colonnes de A qui forment la base réalisable B et $h_1, h_2, \cdots, h_{n-m}$ les indices des colonnes hors base de A qui forment la matrice H. On remarque que les indices b_k et les indices h_k ne sont pas nécessairement ordonnés, il suffit que les colonnes $A_{b_1},\ A_{b_2},\ \cdots,\ A_{b_m}$ forment une base réalisable. Il est facile de reconnaître les colonnes de base dans ce tableau, car, à une permutation près, les colonnes de $B^{-1}A$ qui forment l'identité de $\mathbb{R}^m$ sont les colonnes de base et les autres colonnes forment, à une permutation près, la matrice $\bar{H}$. Ce qui permet de remplir la dernière ligne du tableau :

$$\bar{c}_H = (\bar{c}_{h_k}) = c_H - \bar{H}^T c_B.$$

EXEMPLE

$$\begin{cases} \min z = -x_1 - 2x_2 \\ \textit{sous contraintes} \ : \\ -3x_1 + 2x_2 + x_3 = 2 \\ -x_1 + 2x_2 + x_4 = 4 \\ x_1 + x_2 + x_5 = 5 \\ x_1, x_2, x_3, x_4, x_5 \geq 0 \end{cases}$$

On a

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

On remarque que les trois dernières colonnes de A forment une base. On peut donc prendre :

$$b_1 = 4, b_2 = 3, b_3 = 5, h_1 = 2, h_2 = 1$$

soit

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B^{-1}A = \begin{pmatrix} -1 & 2 & 0 & 1 & 0 \\ -3 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}, \bar{d} = \begin{pmatrix} 4 \\ 2 \\ 5 \end{pmatrix},$$

ce qui donne

x_1	x_2	x_3	x_4	x_5	
-1	2	0	1	0	4
-3	2	1	0	0	2
$+1$	$+1$	0	0	1	5
-1	-2	0	0	0	
h_2	h_1	b_2	b_1	b_3	
-1	-2	$*$	$*$	$*$	

,

$$\bar{c}_H = \begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} 2 & 2 & 1 \\ -1 & -3 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$x_{h_1} = x_2$ entre en base car $\bar{c}_{h_1} = -2 < \bar{c}_{h_2} = -1 < 0$

$$\bar{\theta} = \min_{i \in I^*(\bar{A}_2)} \frac{\bar{d}_i}{\bar{a}_{i,2}} = \frac{\bar{d}_2}{\bar{a}_{2,2}}, (r = 2)$$

donc $x_{b_2} = x_3$ quitte la base. La nouvelle base est donc

$$b_1 = 4, b_2 = 2, b_3 = 5, h_1 = 3, h_2 = 1$$

Pour calculer le nouveau vecteur $\bar{d}$ et la nouvelle matrice $\bar{H}$, il suffit de multiplier à gauche la matrice formée par les trois premières lignes du tableau par la matrice C de changement de base :

$$C = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1/2 & 0 \\ 0 & -1/2 & 1 \end{pmatrix}$$

ce qui revient à :

$$\begin{array}{lcl} L_1 & \longleftarrow & L_1 - L_2 \\ L_2 & \longleftarrow & 1/2 L_2 \\ L_3 & \longleftarrow & L_3 - 1/2 L_2 \end{array}$$

ce qui donne le tableau suivant

x_1	x_2	x_3	x_4	x_5	
2	0	-1	1	0	2
$-3/2$	1	$1/2$	0	0	1
$5/2$	0	$-1/2$	0	1	4
-1	-2	0	0	0	
h_2	b_2	h_1	b_1	b_3	
-4	$*$	1	$*$	$*$	

$$\begin{aligned} \bar{c}_H &= \begin{pmatrix} 0 \\ -1 \end{pmatrix} - \begin{pmatrix} -1 & 1/2 & -1/2 \\ 2 & -3/2 & 5/2 \end{pmatrix} \begin{pmatrix} 0 \\ -2 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ -4 \end{pmatrix} \end{aligned}$$

$x_{h_2} = x_1$ entre en base car $\bar{c}_{h_2} = -4 < 0 < \bar{c}_{h_1} = 1$

$$\bar{\theta} = \min_{i \in I^*(\bar{A}_1)} \frac{\bar{d}_i}{\bar{a}_{i,1}} = \frac{\bar{d}_1}{\bar{a}_{1,1}}, (r = 1)$$

donc $x_{b_1} = x_4$ quitte la base. La nouvelle base est donc

$$b_1 = 1, b_2 = 2, b_3 = 5, h_1 = 3, h_2 = 4$$

La matrice C de changement de base est égale à :

$$C = \begin{pmatrix} 1/2 & 0 & 0 \\ 3/4 & 1 & 0 \\ -5/4 & 0 & 1 \end{pmatrix}$$

ce qui revient à :

$$\begin{array}{lcl} L_1 & \longleftarrow & 1/2 L_1 \\ L_2 & \longleftarrow & L_2 + 3/4 L_1 \\ L_3 & \longleftarrow & L_3 - 5/4 L_1 \end{array}$$

ce qui donne le tableau suivant

x_1	x_2	x_3	x_4	x_5	
1	0	$-1/2$	$1/2$	0	1
0	1	$-1/4$	$3/4$	0	$5/4$
0	0	$3/4$	$-5/4$	1	$3/2$
-1	-2	0	0	0	
b_1	b_2	h_1	h_2	b_3	
$*$	$*$	-1	2	$*$	

$$\begin{aligned}\bar{c}_H &= -\begin{pmatrix} -1/2 & -1/4 & 3/4 \\ 1/2 & 3/4 & -5/4 \end{pmatrix}\begin{pmatrix} -1 \\ -2 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -1 \\ 2 \end{pmatrix}\end{aligned}$$

$x_{h_1} = x_3$ entre en base car $\bar{c}_{h_1} = -1 < 0$

$$\bar{\theta} = \min_{i \in I^*(\bar{A}_3)} \frac{\bar{d}_i}{\bar{a}_{i,3}} = \frac{\bar{d}_3}{\bar{a}_{3,3}}, (r = 3)$$

donc $x_{b_3} = x_5$ quitte la base. La nouvelle base est donc

$$b_1 = 1, b_2 = 2, b_3 = 3, h_1 = 5, h_2 = 4$$

La matrice C de changement de base est égale à :

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

ce qui revient à :

$$\begin{aligned} L_1 &\longleftarrow L_1 + 2/3L_3 \\ L_2 &\longleftarrow L_2 + 1/3L_3 \\ L_3 &\longleftarrow 4/3L_3 \end{aligned}$$

ce qui donne le tableau suivant

x_1	x_2	x_3	x_4	x_5	
1	0	0	$-1/3$	$2/3$	2
0	1	0	$1/3$	$1/3$	3
0	0	1	$-5/3$	$4/3$	2
-1	-2	0	0	0	
b_1	b_2	b_3	h_2	h_1	
$*$	$*$	$*$	$1/3$	$4/3$	

$$\begin{aligned}\bar{c}_H &= -\begin{pmatrix} -1/3 & 1/3 & -5/3 \\ 2/3 & 1/3 & 4/3 \end{pmatrix}\begin{pmatrix} -1 \\ -2 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 \\ 4/3 \end{pmatrix}\end{aligned}$$

L'optimum est atteint car $\bar{c}_H \geq 0$. La solution de (P) est donc la solution de base de la dernière base :

$$x_{b_1} = \bar{d}_1 = 2, x_{b_2} = \bar{d}_2 = 3, x_{b_3} = \bar{d}_3 = 2, x_{h_1} = x_{h_2} = 0$$

Or, $b_1 = 1, b_2 = 2, b_3 = 3, h_1 = 5, h_2 = 4$, ce qui donne :

$$x = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 0 \\ 0 \end{pmatrix} \text{ solution optimale de } (P).$$

6.7 Dualité

Considérons le programme linéaire sous la forme standard

$$(P) \qquad \begin{cases} \min z(x) = c^T x \\ \textit{sous contraintes} \ : \\ \quad Ax = d \\ \quad x \geq 0 \end{cases}$$

On associe à ce programme linéaire un autre programme linéaire :

$$(D) \qquad \begin{cases} \max w(u) = d^T u \\ \textit{sous contraintes} \ : \\ \quad A^T u \leq c \end{cases}$$

Le programme linéaire (P) est appelé le primal et le programme linéaire (D) est appelé le dual de (P).

Exercice d'application : Déterminer la forme standard de (D) et montrer que le dual de (D) est (P).

Réponse : On note par $A_j \in \mathbb{R}^m$, $j = 1, n$, les colonnes de la matrice A, par $u^T = (u_1 \ u_2 \ \cdots u_m) \in \mathbb{R}^m$ et par $d^T = (d_1 \ d_2 \ \cdots d_m) \in \mathbb{R}^m$. On peut toujours écrire :

$$u = \begin{pmatrix} v_1 - v_{m+1} \\ v_2 - v_{m+2} \\ \vdots \\ v_m - v_{2m} \end{pmatrix}$$

avec $v^T = (v_1 \ v_2 \ \cdots v_{2m}) \in \mathbb{R}^{2m}_+$, ce qui donne :

$$d^T u = -f^T v \quad et \quad A^T u = Bv$$

avec $f^T = (-d_1 \ -d_2 \ \cdots -d_m \ d_1 \ d_2 \ \cdots d_m)$ et B est une matrice n lignes $2m$ colonnes telle que sa ième ligne $= (A_i^T \ -A_i^T)$. On obtient alors un programme équivalent au programme (D) :

$$(\tilde{D}) \begin{cases} Min \ f^T v \\ Bv \leq c \\ v \in \mathbb{R}^{2m}_+ \end{cases}$$

Enfin, on obtient la forme standard de (D) en ajoutant n-variables positives à v, ce qui donne :

$$(\hat{D}) \begin{cases} Min \ \hat{f}^T \hat{v} \\ \hat{B}\hat{v} = c \\ \hat{v} \in \mathbb{R}^{2m+n}_+ \end{cases}$$

avec $\hat{f}^T = (f^T \ \ 0_{\mathbb{R}^n})$, $\hat{v}^T = (v^T \ v_{2m+1} \ \cdots v_{2m+n})$ et $\hat{B} = [B \ I_n]$. Par conséquent le dual de $(\hat{D})$ est :

$$(\hat{P}) \begin{cases} Max \ \ c^T y \\ \hat{B}^T y \leq \hat{f} \\ y \in \mathbb{R}^n \end{cases}$$

Si on remplace B et f par leurs valeurs en fonction de A et d et y par $-x$, on obtient :

$$(\hat{P}) \begin{cases} Max \ \ c^T(-x) \\ A(-x) \leq -d \\ -A(-x) \leq d \\ (-x) \leq 0 \\ x \in \mathbb{R}^n \end{cases}$$

ce qui donne le programme (P).

Les relations entre le problème primal (P) et son dual (D) sont données par le théorème suivant :

THÉORÈME 6.7.1 **(a)** *Si $\bar{x}$ et $\bar{u}$ sont respectivement des solutions réalisables du primal et du dual, alors :*

$$z(\bar{x}) = c^T \bar{x} \geq w(\bar{u}) = d^T \bar{u}$$

(b) *Si l'ensemble des solutions réalisables du primal et l'ensem-ble des solutions réalisables du dual sont non vides, alors, le programme primal admet une solution optimale.*

(c) *Si x^* et u^* sont respectivement des solutions réalisables du primal et du dual telles que :*

$$c^T x^* = d^T u^*$$

alors, x^* est une solution optimale du primal et u^* est une solution optimale du dual.

(d) *Si (P) admet une solution optimale x^*, nécessairement, (D) admet une solution optimale u^* et on a :*

$$c^T x^* = d^T u^*$$

DÉMONSTRATION :

(a)

$$A\bar{x} = d \Longrightarrow \bar{u}^T A\bar{x} = \bar{u}^T d = d^T \bar{u} = (A\bar{x})^T \bar{u} = \bar{x}^T (A^T \bar{u})$$

et comme $\bar{x} \geq 0$, $A^T \bar{u} \leq c$, alors, $d^T \bar{u} \leq \bar{x}^T c = c^T \bar{x}$.

(b) Si l'ensemble des solutions réalisables du primal et l'ensemble des solutions réalisables du dual sont non vides, d'après (a), la fonction z est minorée et le théorème 6.3.3 montre que (P) admet une solution.

(c) On utilise (a) et (b).

(d) Voir proposition 6.4.1

COROLLAIRE 6.7.1 *Etant donné un programme linéaire (P) et son dual (D) :*

(a) *Si (P) et (D) ont des solutions réalisables, alors, chacun d'eux a une solution optimale et :*

$$z^* = \min(P) = \max(D) = w^*$$

(b) *Si l'un des programmes* (P) *ou* (D) *admet une solution optimale, alors, l'autre admet aussi une solution optimale et :*

$$z^* = \min(P) = \max(D) = w^*$$

(c) *Si l'un d'eux a un optimum non borné l'autre n'a pas de solution réalisable.*

DÉMONSTRATION :

(a) D'après (b) du théorème précédent, le programme (P) admet une solution optimale x^* et d'après (d) du même théorème le programme linéaire (D) admet une solution optimale u^* telle que : $c^T x^* = d^T u^*$.

(b) Si (P) admet une solution optimale, on utilise (d) du théorème précédent, on obtient ce qu'il faut. Si (D) admet une solution optimale alors sa forme standard admet une solution optimale et d'après l'exercice d'application, le dual de la forme standard de (D) n'est autre que (P), donc encore une fois d'après (d) du théorème pécédent, on obtient que (P) admet une solution optimale.

(c) Si l'optimum de (P) est non borné, c'est à dire : $inf(P) = -\infty$, en utilisant (a) du théorème précédent, on vérifie facilement que l'ensemble des solutions réalisables de (D) est vide. Même raisonnement si l'optimum de (D) est non borné, c'est à dire : $sup(D) = +\infty$.

6.8 Exercices du chapire 6 :

Exercice 1

Soit

$$(\mathcal{P}_0) \begin{cases} \min -x_1 - 2x_2 \\ -x_1 + x_2 \leq 2 \\ x_1 + x_2 \leq 4 \\ x_1 \geq 0, x_2 \geq 0 \end{cases}$$

1. Résoudre graphiquement le programme linéaire $(\mathcal{P}_0)$.
2. Donner un équivalent du programme linéaire $(\mathcal{P}_0)$ sous la forme standard. On notera par $(\mathcal{P})$ le nouveau programme.
3. Résoudre par l'algorithme du simplexe le programme linéaire $(\mathcal{P})$. Conclure.

Exercice 2

Soit

$$(\mathcal{P}_\gamma) \begin{cases} \min\ 2x_1 - 3x_2 + x_3 + 3x_4 \\ 2x_1 - x_2 - x_3 - x_4 = 2\gamma \\ -x_1 + x_2 + x_3 + 2x_4 = -\gamma \\ 3x_1 - x_2 - x_3 = 2\gamma + 1 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \end{cases}$$

1. Déterminer le nombre γ pour lequel l'ensemble des $x \in \mathbb{R}^4$ qui vérifient les contraintes soit non vide.

2. On notera par $(\mathcal{P})$ le programme linéaire pour la valeur de γ obtenue en 1.). Montrer que le programme $(\mathcal{P})$ s'écrit sous la forme :

$$(\mathcal{P}) \begin{cases} \min 2x_1 - 3x_2 + x_3 + 3x_4 \\ x \in \mathcal{U} = \{x \in \mathbb{R}^4 / Ax = d; x \geq 0\} \end{cases}$$

où A et d sont des données à déterminer avec $rang(A) = 2$.

3. Résoudre par l'algorithme du simplexe le programme linéaire $(\mathcal{P})$.

Exercice 3

On considère le programme linéaire :

$$(\mathcal{P}_0) \begin{cases} \min 2x_1 - x_2 + x_3 \\ x_1 + x_2 - x_3 \geq -2 \\ -x_1 + x_2 + 2x_3 \leq 1 \\ x_1 + x_3 = 1 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \end{cases}$$

1. Ecrire le programme $(\mathcal{P}_0)$ sous la forme standard $(\mathcal{P})$:

$$(\mathcal{P}) \begin{cases} \min z(x) = c^T x \\ x \in \mathcal{U} = \{x \in \mathbb{R}^n / Ax = d \text{ et } x_i \geq 0, i = 1, \cdots, n\} \end{cases}$$

où n, c, d, A sont des données à déterminer.

2. Vérifier que $\bar{x} = (\frac{1}{3}\ 0\ \frac{2}{3}\ \frac{5}{3}\ 0)^T$ est un sommet de $\mathcal{U}$.
3. Résoudre par la méthode du simplexe le programme linéaire $(\mathcal{P})$ et déduire un vecteur de $\mathbb{R}^3$ qui réalise le minimum de $(\mathcal{P}_0)$.
4. Transformer le programme $(\mathcal{P}_0)$ en un programme linéaire de dimension 2.
5. Résoudre graphiquement le programme linéaire de dimension 2. Comparer la solution obtenue par la méthode graphique et la solution obtenue par la méthode du simplexe.

Exercice 4

Une usine fabrique deux types de jouets en bois : des soldats et des trains. Les données de ce problème sont représentées dans le tableau suivant :

	P. vente	Mat. prem.	Frais gén.	Menuiserie	Finition
1 soldat	27DT	10DT	14DT	1h de travail	2h de travail
1 train	21DT	9DT	10DT	1h de travail	1h de travail

Par semaine l'usine dispose de toutes les matières premières nécessaires à la fabrication et ne dispose que de 100h de finition et 80h de menuiserie. La demande des trains et des soldats est illimitée. Déterminer le plan de production qui maximise le profit de l'usine.

Exercice 5

Le self-service d'un hotel offre chaque jour à ses clients quatre plats : plat1, plat2, plat3, plat4. Le prix d'une unité du plat1 vaut 0.5DT, du plat2 vaut 0.2DT, du plat3 vaut 0.3DT et du plat4 vaut 0.8DT. Le tableau suivant nous donne la quantité de vitamines V1, V2, V3 et V4 dans une unité de chaque plat :

par unité	V1	V2	V3	V4
plat1	400	3	2	2
plat2	200	2	2	4
plat3	150	0	4	1
plat4	500	0	4	5

Un client suit un régime alimentaire doit manger au moins : 500 unités de V1, 6 unités de V2, 10 unités de V3 et 8 unités de V4. Déterminer le régime qui coûte le moins cher.

Exercice 6

Soit $(\mathcal{P})$ le programme linéaire :

$$(\mathcal{P})\begin{cases} \min z(x) = c^T x \\ x \in \mathcal{U} = \{x \in \mathbb{R}^n / Ax = d; x \geq 0\} \end{cases}$$

avec A une matrice $m \times n$ de $rang(A) = m$, $c^T = (c_1 c_2 \cdots c_n) \in \mathbb{R}^n$ et $d \in \mathbb{R}^m$. On notera par $A_j \in \mathbb{R}^m$, $j = 1, \cdots, n$, les colonnes de la matrice A et on suppose que $m \geq 2$.

1. Montrer qu'il existe $j_0 \in \{1, \cdots, n\}$ tel que A_{j_0} n'est pas un multiple du vecteur d.
2. On pose $\widetilde{A}_{n+1} = d - A_{j_0}$, $\widetilde{A}_j = A_j$ pour $j = 1, \cdots, n$, $\widetilde{A}$ la matrice $m \times (n+1)$ dont les colonnes sont $\widetilde{A}_j$, $j = 1, \cdots, n+1$ et $\widetilde{c}^T = (0 \cdots 0\ 1) \in \mathbb{R}^{n+1}$. On considère le nouveau programme linéaire :

$$(\tilde{\mathcal{P}})\begin{cases} \min z(\tilde{x}) = \tilde{c}^T \tilde{x} = \tilde{x}_{n+1} \\ \tilde{x} \in \tilde{\mathcal{U}} = \{\tilde{x} \in \mathbb{R}^{n+1} / \widetilde{A}\tilde{x} = d; \tilde{x} \geq 0\} \end{cases}$$

   (a) Montrer que $\bar{x} \in \mathbb{R}^{n+1}$ avec $\bar{x}_{j_0} = \bar{x}_{n+1} = 1$ et $\bar{x}_i = 0$ pour $i \neq j_0$ et $i \neq n+1$, est un sommet de $\tilde{\mathcal{U}}$.

   (b) Montrer que le minimum de $(\tilde{\mathcal{P}})$ est atteint en un sommet de $\tilde{\mathcal{U}}$. On notera par $\xi^T = (\xi_1\ \xi_2 \cdots \xi_n\ \xi_{n+1})$ un sommet de $\tilde{\mathcal{U}}$ qui réalise le minimum de $(\tilde{\mathcal{P}})$.

   (c) Montrer que :

$$\xi_{n+1} = 0 \Longleftrightarrow \mathcal{U} = \{x \in \mathbb{R}^n / Ax = d; x \geq 0\} \neq \emptyset$$

   (d) Montrer que si $\xi_{n+1} = 0$ alors :
   $x^T = (\xi_1 \cdots \xi_n)$ est un sommet de $\mathcal{U} = \{x \in \mathbb{R}^n / Ax = d; x \geq 0\}$.

6.9 Corrigé des exercices

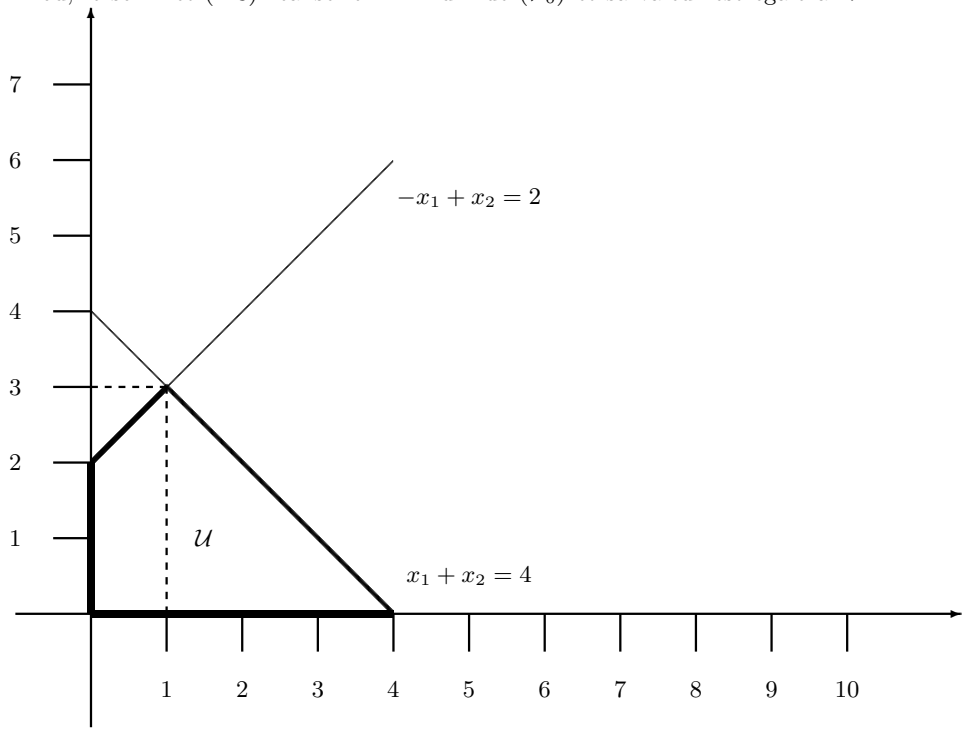
Réponse 1

Soit

$$(\mathcal{P}_0)\begin{cases} \min -x_1 - 2x_2 \\ -x_1 + x_2 \leq 2 \\ x_1 + x_2 \leq 4 \\ x_1 \geq 0, x_2 \geq 0 \end{cases}$$

1. Il est clair que la fonction linéaire $z(x) = -x_1 - 2x_2$ est continue et que l'ensemble $\mathcal{U}$ des points qui vérifient les contraintes est un compact de $\mathbb{R}^2$, donc le minimum est atteint. D'après un théorème du cours, le minimum est atteint en un sommet de $\mathcal{U}$, or les seuls sommets de $\mathcal{U}$ sont (0 2), (1 3), (0 0) et (4 0) et on a respectivement : -4, -7, 0 et -4 comme images par $z(x)$.

D'où, le sommet (1 3) réalise le minimum de $(\mathcal{P}_0)$ et sa valeur est égale à -7.



2. $(\mathcal{P})\begin{cases} \min -x_1 - 2x_2 \\ -x_1 + x_2 + x_3 = 2 \\ x_1 + x_2 + x_4 = 4 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \end{cases}$

3. $A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$, $c^T = (-1\ -2\ 0\ 0)$ et $d = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$. On voit que $B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ est une base ($b_1 = 2$, $b_2 = 4$, $h_1 = 1$, $h_2 = 3$), $B^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ et la solution de base associée est $x = [x_B\ 0]$ avec $x_B = B^{-1}d = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, d'où $x^T = (0\ 2\ 0\ 2)$ et $z(x) = -4$. On calcule : $\bar{H} = B^{-1}H$, on trouve

$$\bar{H} = \begin{pmatrix} -1 & 1 \\ 2 & -1 \end{pmatrix}, \bar{c}_H = c_H - \bar{H}^T c_B = \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} -2 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 2 \end{pmatrix}$$

	x_1	x_2	x_3	x_4	d
A	-1	1	1	0	2
	1	1	0	1	4
	h_1	b_1	h_2	b_2	
c	-1	-2	0	0	
$\bar{c}_H$	-3	$*$	2	$*$	

h_1=1 entre dans la base et $\bar{A}_1 = B^{-1}A_1 = (-1\ 2)^T$, d'où, $I^*(\bar{A}_1) = \{2\}$:

$$\bar{\theta} = \min_{i \in \{2\}} \frac{\bar{d}_i}{\bar{a}_{i1}} = \frac{\bar{d}_2}{\bar{a}_{21}}$$

donc, $b_2 = 4$ quitte la base.

4. La nouvelle base est donc avec $b_1 = 2$, $b_2 = 1$, $h_1 = 4$, $h_2 = 3$ $B = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ et $B^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ d'où $x^T = (1\ 3\ 0\ 0)$ et $z(x) = -7$. On calcule : $\bar{H} = B^{-1}H$, on trouve

$$\bar{H} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \bar{c}_H = c_H - \bar{H}^T c_B = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 3/2 \\ 1/2 \end{pmatrix}$$

	x_1	x_2	x_3	x_4	d
A	-1	1	1	0	2
	1	1	0	1	4
	h_1	b_1	h_2	b_2	
c	-1	-2	0	0	
$\bar{c}_H$	$*$	$*$	$3/2$	$1/2$	

$\bar{c}_H \geq 0$, alors le minimum est atteint. On garde seulement les variables de notre problème initial et on laisse tomber les variables artificielles x_3 et x_4.
La conclusion : on trouve la même solution avec les deux méthodes, donc, il vaut mieux, dans le cas de la dimension 2, utiliser la méthode graphique.

Réponse 2

1. La troisième équation est égale à 2 fois la première + la deuxième. Pour que le système admet une solution, il faut que 2 fois le second membre de la première équation + le second membre de la deuxième est égal au second membre de la troisième équation, ce qui donne : $\gamma = 1$.
2. Pour $\gamma = 1$, la troisième équation est inutile. Le programme linéaire est donc :

$$(\mathcal{P}) \begin{cases} \min c^T x \\ Ax = d \\ x \geq 0 \end{cases}$$

avec $A = \begin{pmatrix} 2 & -1 & -1 & -1 \\ -1 & 1 & 1 & 2 \end{pmatrix}$, $d = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$ et $c^T = (2\ -3\ 1\ 3)$. Il est clair que le rang de A est égal à 2 et on voit que $B = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ est une base réalisable de solution de base $x^T = (1\ 0\ 0\ 0)$ et $z(x) = 2$, donc, $b_1 = 1$, $b_2 = 2$, $h_1 = 3$, $h_2 = 4$. Si on calcule :

$$\bar{c}_H = c_H - \bar{H}^T c_B = \begin{pmatrix} 4 \\ 10 \end{pmatrix} \geq 0$$

donc $x^T = (1\ 0\ 0\ 0)$ réalise le minimum de $(\mathcal{P})$ et la valeur du minimum est 2.

Réponse 3

1. Les variables x_1, x_2, x_3 sont positives d'où :

$$x_1 + x_2 - x_3 \geq -2 \Longleftrightarrow \exists x_4 \geq 0 / x_1 + x_2 - x_3 - x_4 = -2$$

$$-x_1 + x_2 + 2x_3 \leq 1 \Longleftrightarrow \exists x_5 \geq 0 / -x_1 + x_2 + 2x_3 + x_5 = 1$$

D'où :

$$(\mathcal{P}_0) \Longleftrightarrow \begin{cases} \min 2x_1 - x_2 + x_3 \\ x_1 + x_2 - x_3 - x_4 = -2 \\ -x_1 + x_2 + 2x_3 + x_5 = 1 \\ x_1 + x_3 = 1 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0 \end{cases}$$

ce qui est encore équivalent à :

$$\begin{cases} \min c^T x \\ x \in \mathcal{U} = \{x \in \mathbb{R}^5 / Ax = d, x_i \geq 0, i = 1, \cdots, 5\} \end{cases}$$

avec $c^T = (2\ -1\ 1\ 0\ 0)$, $d^T = (-2\ 1\ 1)$ et $A = \begin{pmatrix} 1 & 1 & -1 & -1 & 0 \\ -1 & 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$

2. Il faut d'abord prouver que $\bar{x} \in \mathcal{U}$:

$$\bar{x} \geq 0 \ et$$

$$A\bar{x} = \begin{pmatrix} 1 & 1 & -1 & -1 & 0 \\ -1 & 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 0 \\ 2/3 \\ 5/3 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix} = d$$

Montrons maintenant que $\bar{x}$ est un sommet : d'après un théorème du cours, il faut et il suffit que $\{A_i / \bar{x}_i > 0\}$ est une famille libre (A_i =la ième colonne de A). Dans notre cas la famille en question est $\{A_1, A_3, A_4\}$. Soit $(\alpha, \beta, \gamma) \in \mathbb{R}^3$, on a :

$$\begin{aligned} \alpha A_1 + \beta A_3 + \gamma A_4 = 0 \quad &\Longleftrightarrow \alpha \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + \beta \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} + \gamma \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} = 0 \\ &\Longleftrightarrow \begin{matrix} (1) \\ (2) \\ (3) \end{matrix} \begin{cases} \alpha - \beta - \gamma = 0 \\ -\alpha + 2\beta = 0 \\ \alpha + \beta = 0 \end{cases} \end{aligned}$$

L'équation (3) + l'équation (2) nous donne $3\beta = 0$, d'où $\beta = 0$. De l'équation (3), on déduit que $\alpha = 0$ et de l'équation (1) on déduit que $\gamma = 0$. D'où la famille $\{A_1, A_3, A_4\}$ est une famille libre. Donc, $\bar{x}$ est un sommet de $\mathcal{U}$.

3. La matrice de base réalisable associée au sommet $\bar{x}$ est :

$$B = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 0 \\ 1 & 1 & 0 \end{pmatrix} avec\ b_1 = 1, b_2 = 3, b_3 = 4$$

et la matrice hors-base est :

$$H = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} avec\ h_1 = 2, h_2 = 5$$

Une itération de la méthode du simplexe nécessite le calcul de $\bar{H} = B^{-1}H$ (ce qui est équivalent au calcul de $\bar{A}_2 = B^{-1}A_2 = (\bar{a}_{1,2}\ \ \bar{a}_{2,2}\ \ \bar{a}_{3,2})^T$ et $\bar{A}_5 = B^{-1}A_5 = (\bar{a}_{1,5}\ \ \bar{a}_{2,5}\ \ \bar{a}_{3,5})^T$) et le calcul de $\bar{d} = B^{-1}d$. Pour ce calcul, on peut adapter la méthode de Gauss-Jordan, à gauche la matrice B et à droite la matrice formée par les colonnes A_2, A_5 et d :

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

$$\begin{matrix} \\ l_2 + l_1 \\ l_3 - l_1 \end{matrix} \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -2 \\ 2 & 1 & -1 \\ -1 & 0 & 3 \end{pmatrix}$$

$$\begin{array}{c} l_1 + l_2 \\ \\ l_3 - 2l_2 \end{array} \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 3 & 1 & -3 \\ 2 & 1 & -1 \\ -5 & -2 & 5 \end{pmatrix}$$

$$\begin{array}{c} l_1 + 2/3l_3 \\ l_2 + 1/3l_3 \\ \\ \end{array} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} -1/3 & -1/3 & 1/3 \\ 1/3 & 1/3 & 2/3 \\ -5 & -2 & 5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1/3 & -1/3 & 1/3 \\ 1/3 & 1/3 & 2/3 \\ -5/3 & -2/3 & 5/3 \end{pmatrix}$$

D'où, les deux premières colonnes de la matrice obtenue à droite représente $\bar{H}$ et la troisième colonne est égale à $\bar{d}$. On doit maintenant calculer le vecteur :

$$\bar{c}_H = c_H - \bar{H}^T c_B$$

avec $c_H = \begin{pmatrix} c_2 \\ c_5 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ et $c_B = \begin{pmatrix} c_1 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$. D'où :

$$\bar{c}_H = \begin{pmatrix} -1 \\ 0 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} -1 & 1 & -5 \\ -1 & 1 & -2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2/3 \\ 1/3 \end{pmatrix} = \begin{pmatrix} \bar{c}_{h_1} \\ \bar{c}_{h_2} \end{pmatrix}$$

$c_{h_1} = -2/3 < 0$ alors $\bar{x}$ ne réalise pas le minimum de $(\mathcal{P})$ et $h_1 = 2$ entre en base. D'autre part $I^*(\bar{A}_{h_1}) = I^*(\overline{A}_2) = \{2\}$, d'où :

$$\bar{\theta} = \min_{k \in I(\bar{A}_2)} \frac{\bar{d}_k}{\bar{a}_{k,2}} = \frac{\overline{d_r}}{\bar{a}_{r,2}} = \frac{\overline{d_2}}{\bar{a}_{2,2}}$$

ce qui donne $r = 2$ et $b_r = b_2 = 3$ quitte la base. La nouvelle base réalisable obtenue à la suite de la première itération de la méthode du simplexe est alors : $b_1 = 1, b_2 = 2, b_3 = 4, h_1 = 3, h_2 = 5$.
La matrice de base réalisable associée au nouveau sommet est :

$$B = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \textit{avec } b_1 = 1, b_2 = 2, b_3 = 4$$

et la matrice hors-base est :

$$H = \begin{pmatrix} -1 & 0 \\ 2 & 1 \\ 1 & 0 \end{pmatrix} \textit{avec } h_1 = 3, h_2 = 5$$

Le calcul de $\bar{H} = B^{-1}H$ (est équivalent au calcul de $\bar{A}_3 = B^{-1}A_3 = (\bar{a}_{1,3}\ \bar{a}_{2,3}\ \bar{a}_{3,3})^T$ et $\bar{A}_5 = B^{-1}A_5 = (\bar{a}_{1,5}\ \bar{a}_{2,5}\ \bar{a}_{3,5})^T$) et le calcul de $\bar{d} = B^{-1}d$. Pour ce calcul, on peut adapter la méthode de Gauss-Jordan, à gauche la matrice B et à droite la matrice formée par les colonnes A_2, A_5 et d : :

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

$$\begin{array}{c} \\ l_2 + l_1 \\ l_3 - l_1 \end{array} \begin{pmatrix} 1 & 1 & -1 \\ 0 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & -2 \\ 1 & 1 & -1 \\ 2 & 0 & 3 \end{pmatrix}$$

$$\begin{array}{c} l_1 - 1/2l_2 \\ \\ l_3 + 1/2l_2 \end{array} \begin{pmatrix} 1 & 0 & -1/2 \\ 0 & 2 & -1 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} -3/2 & -1/2 & -3/2 \\ 1 & 1 & -1 \\ 5/2 & 1/2 & 5/2 \end{pmatrix}$$

$$\begin{array}{c} l_1 + l_3 \\ l_2 + 2l_3 \end{array} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 6 & 2 & 4 \\ 5/2 & 1/2 & 5/2 \end{pmatrix}$$

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

D'où, les deux premières colonnes de la matrice obtenue à droite représente $\bar{H}$ et la troisième colonne est égale à $\bar{d}$. On doit maintenant calculer le vecteur :

$$\bar{c}_H = c_H - \bar{H}^T c_B$$

avec $c_H = \begin{pmatrix} c_3 \\ c_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $c_B = \begin{pmatrix} c_1 \\ c_2 \\ c_4 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}$. D'où :

$$\bar{c}_H = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & 3 & 5 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \geq 0$$

Donc, le sommet $\hat{x}$ associé à cette base réalise le minimum de $(\mathcal{P})$. Le vecteur $\hat{x} = [x_B \ x_H]$ avec

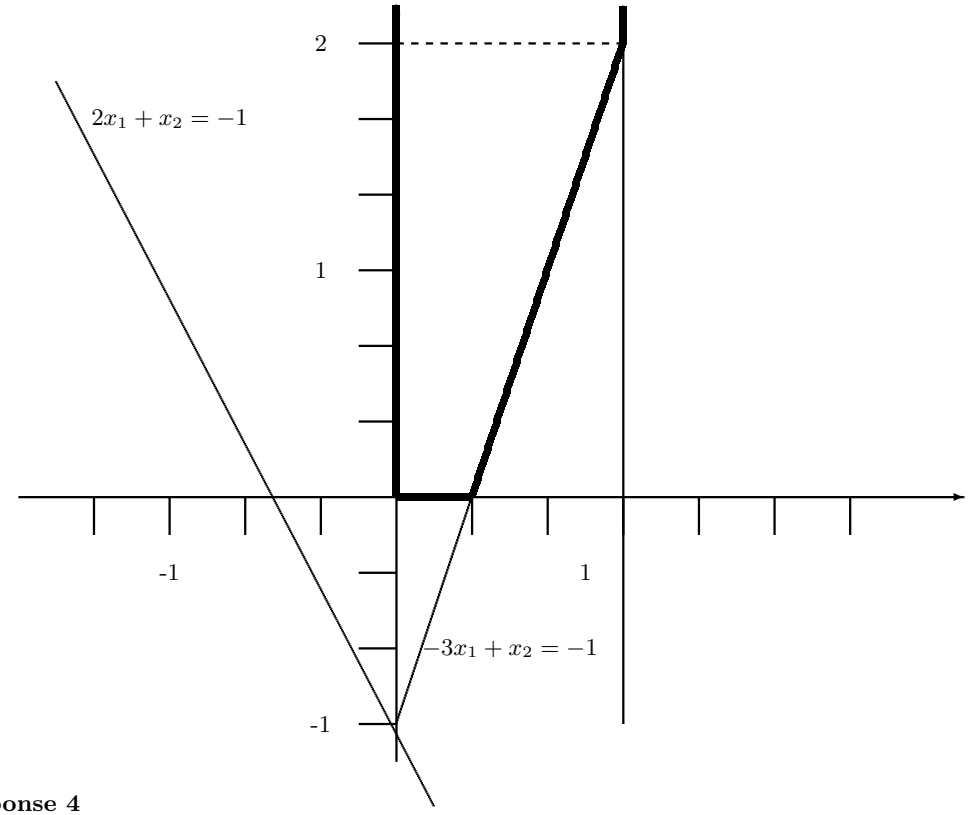
$$x_B = \begin{pmatrix} x_1 \\ x_2 \\ x_4 \end{pmatrix} = B^{-1}d = \bar{d} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}, \ et \ x_H = \begin{pmatrix} x_3 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

D'où $\hat{x} = (1 \quad 2 \quad 0 \quad 5 \quad 0)^T$ et $z(\hat{x}) = c^T\hat{x} = 2 - 2 = 0$. Donc, le vecteur $\tilde{x} = (1 \quad 2 \quad 0)^T$ réalise le minimum de $(\mathcal{P}_0)$.

4. La troisième contrainte du problème $(\mathcal{P}_0)$ est : $x_1 + x_3 = 1$ avec $x_1 \geq 0$ et $x_3 \geq 0$. D'où, $x_3 = 1 - x_1 \geq 0$ ce qui donne $x_1 \leq 1$. Pour obtenir un problème de dimension 2 équivalent à $(\mathcal{P}_0)$, il suffit de remplacer la variable x_3 par $1 - x_1$ et la contrainte $x_3 \geq 0$ par $x_1 \leq 1$, on obtient alors :

$$\left\{ \begin{array}{c} \min 2x_1 - x_2 + (1 - x_1) \\ x_1 + x_2 - (1 - x_1) \geq -2 \\ -x_1 + x_2 + 2(1 - x_1) \leq 1 \\ x_1 \geq 0, x_2 \geq 0, x_1 \leq 1 \end{array} \right. = \left\{ \begin{array}{c} \min x_1 - x_2 + 1 \\ 2x_1 + x_2 \geq -1 \\ -3x_1 + x_2 \leq -1 \\ x_1 \geq 0, x_2 \geq 0, x_1 \leq 1 \end{array} \right.$$

5. On trouve la même solution :



Réponse 4

On note par :

x_1 = le nombre de soldats produits chaque semaine ,

x_2 = le nombre de trains produits chaque semaine.

D'après les données, on a les contraintes suivantes :

$$\begin{cases} x_1 + x_2 & \leq 80 \\ 2x_1 + x_2 & \leq 100 \\ x_1 \geq 0, x_2 & \geq 0 \end{cases}$$

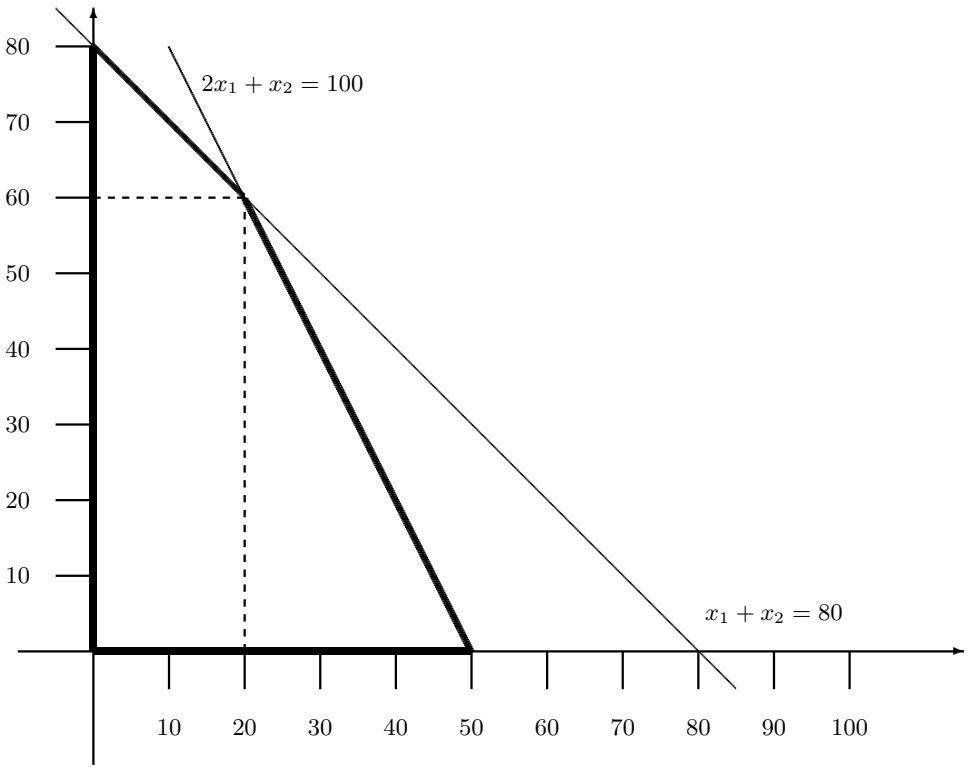
La fonction à maximiser est :

$$(27 - 10 - 14)x_1 + (21 - 9 - 10)x_2$$

Ce qui donne le programme linéaire suivant :

$$\begin{cases} \max\ 3x_1 + 2x_2 \\ x_1 + x_2 \leq 80 \\ 2x_1 + x_2 \leq 100 \\ x_1 \geq 0, x_2 \geq 0 \end{cases}$$

Le programme est de dimension deux, donc, on le resout graphiquement :



le sommet (0 80) donne une valeur égale à : $3 \times 0 + 2 \times 80 = 160$,
le sommet (20 60) donne une valeur égale à : $3 \times 20 + 2 \times 60 = 180$,
le sommet (50 0) donne une valeur égale à : $3 \times 50 + 2 \times 0 = 100$,
et le sommet (0 0) donne une valeur égale à : $3 \times 0 + 2 \times 0 = 0$.
Le meilleur plan de production est donc : 20 soldats et 60 trains.

Réponse 5

On note par :
x_1 = la quantité du plat1 consommée par le client ,
x_2 = la quantité du plat2 consommée par le client,
x_3 = la quantité du plat3 consommée par le client,
x_4 = la quantité du plat4 consommée par le client,
D'après les données, on a les contraintes suivantes :

$$\begin{cases} 400x_1 + 200x_2 + 150x_3 + 500x_4 \geq 500 \\ 3x_1 + 2x_2 \geq 6 \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 \geq 10 \\ 2x_1 + 4x_2 + x_3 + 5x_4 \geq 8 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \end{cases}$$

La fonction à minimiser est :

$$0.5x_1 + 0.2x_2 + 0.3x_3 + 0.4x_4$$

Ce qui donne le programme linéaire suivant :

$$\begin{cases} \min\ 0.5x_1 + 0.2x_2 + 0.3x_3 + 0.4x_4 \\ 400x_1 + 200x_2 + 150x_3 + 500x_4 \geq 500 \\ 3x_1 + 2x_2 \geq 6 \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 \geq 10 \\ 2x_1 + 4x_2 + x_3 + 5x_4 \geq 8 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0,\ x_4 \geq 0 \end{cases}$$

En utilisant l'applet Java de l'exempleinteractif, on trouve :

$$x_1 = 0,\ x_2 = 3,\ x_4 = 1,\ x_5 = 0$$

Réponse 6

1. On a $rang(A) = m \geq 2$, donc il y a m colonnes de A qui sont linéairement indépendantes. Par conséquent deux colonnes au moins sont linéairement indépendantes car $m \geq 2$. Donc, il y a au moins une colonne qui n'est pas colinéaire à d.
2. a) Il est clair que $\bar{x} \geq 0$ et on a : $\tilde{A}\bar{x} = \tilde{A}_{j_0} + \tilde{A}_{n+1} = A_{j_0} + (d - A_{j_0}) = d$. D'où, $\bar{x} \in \tilde{\mathcal{U}}$. D'autre part, d'après un théorème du cours :
   $\bar{x} \in \tilde{\mathcal{U}}$ est un sommet de $\tilde{\mathcal{U}} \Longleftrightarrow \{\widetilde{A}_j/\bar{x}_j > 0\}$ est une famille libre.
   Or $\{\widetilde{A}_j/\bar{x}_j > 0\} = \{\widetilde{A}_{j_0}, \tilde{A}_{n+1}\} = \{A_{j_0}, d - A_{j_0}\}$ avec A_{j_0} est non colinéaire à d. Par conséquent, $d - A_{j_0}$ est non colinéaire à A_{j_0} (sinon $d - A_{j_0} = \alpha A_{j_0}$, ce qui donne $d = (1+\alpha)A_{j_0}$ ce qui contredit le fait que d est non colinéaire à A_{j_0}). D'où, $\{\widetilde{A}_j/\bar{x}_j > 0\} = \{A_{j_0}, d - A_{j_0}\}$ est une famille libre.
   b) La fonction linéaire à minimiser du problème $(\tilde{\mathcal{P}})$ vérifie :

$$\forall \tilde{x} \in \tilde{\mathcal{U}}, z(\tilde{x}) = \tilde{x}_{n+1} \geq 0$$

   donc, elle est minorée sur $\tilde{\mathcal{U}}$ et $\tilde{\mathcal{U}} \neq \emptyset$ (d'après 2)a)). D'après un théorème du cours, le minimum de $(\tilde{\mathcal{P}})$ est atteint par un sommet de $\tilde{\mathcal{U}}$.
   c) On a : $\xi = (\xi_1\ \xi_2\ \cdots \xi_n\ \xi_{n+1})^T$ est un point de $\tilde{\mathcal{U}}$qui réalise le minimum de $(\tilde{\mathcal{P}})$. D'où, pour tout $i = 1, \cdots, n+1$:

$$\xi_i \geq 0$$

   et

$$d = \tilde{A}\xi = \sum_{i=1}^{n+1} \xi_i \tilde{A}_i = \sum_{i=1}^{n} \xi_i A_i + \widetilde{A}_{n+1}\xi_{n+1} = Ax + \widetilde{A}_{n+1}\xi_{n+1}$$

   avec $x = (\xi_1\ \xi_2 \cdots\ \xi_n)^T$ et la valeur du minimum de $(\tilde{\mathcal{P}})$ est égale à $z(\xi) = \xi_{n+1}$. Alors, on a :

$$\xi_{n+1} = 0 \Longrightarrow Ax = d \text{ et } x \geq 0 \Longrightarrow x \in \mathcal{U} \Longrightarrow \mathcal{U} \neq \emptyset$$
$$\mathcal{U} \neq \emptyset \Rightarrow \exists x = (x_1\ x_2\ \cdots\ x_n)^T \in \mathbb{R}^n / Ax = d \text{ et } x \geq 0$$

   On pose $\hat{x} = (x_1\ x_2 \cdots\ x_n\ 0)^T$, on obtient que $\hat{x} \in \tilde{\mathcal{U}}$ et que

$$z(\hat{x}) = 0 \leq z(\tilde{x})\ \forall \tilde{x} \in \tilde{\mathcal{U}}$$

   Ce qui prouve que $\hat{x}$ réalise le minimum de $(\tilde{\mathcal{P}})$ et que la valeur du minimum est égale à zéro. D'où, $\xi_{n+1} = 0$.
   d) $\xi = (\xi_1\ \xi_2\ \cdots\ \xi_n\ 0)^T$ est un sommet de $\tilde{\mathcal{U}}$, alors, la famille $\{\tilde{A}_j/\xi_j > 0\}$ est libre. Or :

$$\{\tilde{A}_j/\xi_j > 0\} = \{A_j/\xi_j > 0\}$$

   car $\xi_{n+1} = 0$ et $\tilde{A}_j = A_j$ pour $j \neq n+1$. Ce qui prouve que $x = (\xi_1\ \xi_2\ \cdots\ \xi_n)^T$ est un sommet de $\mathcal{U}$ car $x \in \mathcal{U}$ et $\{A_j/\xi_j > 0\}$ est une famille libre.

Bibliographie

Ciarlet P.G. : *Introduction à l'analyse numérique matricielle et à l'optimisation.*
Masson, Paris (1985).

Stoer J.; Bulirsch R. : *Introduction to numerical analysis.* Springer-Verlag (1983).

Mathlouthi S.: *Analyse numérique linéaire et non linéaire.* Centre de publication universitaire (2002).