Analyse numérique

Benjamin Boutin

Semestre 6 2014-2015

Table des matières

Chapitre 1

Compléments d'algèbre linéaire et bilinéaire

1.1 Réduction de matrices carrés

Théorème 1.1 (*CNS de diagonalisabilité*)

Soit $A \in \mathcal{M}_n(\mathbb{K})$. Les assertions suivantes sont équivalentes :

(i) A est diagonalisable sur $\mathbb{K}$.

(ii) χ_A est scindé sur $\mathbb{K}$ et la multiplicité des algébrique des racines de χ_A est égale à la dimension des sous-espace propres associés (multiplicité géométrique).

(iii) μ_A est scindé à racines simples.

Théorème 1.2 (*CNS de trigonalisabilité*)

Soit $A \in \mathcal{M}_n(\mathbb{K})$. A est trigonalisable si et seulement si χ_A est scindé sur $\mathbb{K}$.

Corollaire 1.3

Toute matrice de $\mathcal{M}_n(\mathbb{C})$ est trigonalisable sur $\mathbb{C}$.

Théorème 1.4 (*Forme réduite de Jordan*)

Soit $A \in \mathcal{M}_n(\mathbb{K})$ supposée trigonalisable sur $\mathbb{K}$. On peut trouver $P \in GL_n(\mathbb{K})$ telle que

$$P^{-1}AP = \begin{pmatrix} J_{k_1}(\lambda_1) & & & 0 \\ & J_{k_2}(\lambda_2) & & \\ & & \ddots & \\ 0 & & & J_{k_m}(\lambda_m) \end{pmatrix} \quad \text{avec } J_k(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}$$

où les $(\lambda_i)_{1 \leqslant i \leqslant m}$ sont les valeurs propres de A avec éventuellement des répéti-

tions.

1.2 Réduction en base orthonormée

1.2.1 Notations

On travaille sur $\mathbb{R}^n$ muni du produit scalaire euclidien canonique et de la norme $\|.\|_2$ associée, ou sur $\mathbb{C}^n$ muni du produit scalaire hermitien canonique

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

et de la norme associée.

Définition 1.5

Une base $(e_1, \dots, e_n)$ de $\mathbb{K}^n$ est dite orthonormée si $(e_i, e_j) = \delta_{i,j},\ i, j \in [\![1, n]\!]$.

Exemples : Soit $O \in \mathcal{M}_n(\mathbb{R})$. $O \in \mathcal{O}_n(\mathbb{R})$ si et seulement les colonnes de O forment une base orthonormée.

Soit $U \in \mathcal{M}_n(\mathbb{C})$. $U \in \mathcal{U}_n(\mathbb{C})$ si et seulement si les colonnes de U forment une base orthonormée.

Proposition 1.6

Soit $(v_i)_{1 \leqslant i \leqslant n}$ une base de $\mathbb{K}^n$. On peut construire par le procédé de Gramm-Schmidt une base orthonormée $(e_i)_{1 \leqslant i \leqslant n}$ de la manière suivante : on pose

$$e_1 = \frac{v_1}{\|v_1\|_2},$$

pour $i[\![2, n]\!]$,

$$u_i = v_i - p_{Vect(e_1, \dots, e_{i-1})}(v_i)$$

avec p_F la projection orthogonale sur F, puis

$$e_i = \frac{u_i}{\|u_i\|_2}.$$

En fait :

$$u_i = v_i - \sum_{j=1}^{i-1} (e_j, v_i) e_j = v_i - \sum_{j=1}^{i-1} \frac{(e_j, u_j)}{(u_j, u_j)} u_j$$

Par construction, on a $Vect(e_1, \dots, e_i) = Vect(v_1, \dots, v_i)$ pour tout i donc la matrice de passage est triangulaire supérieure.

Proposition 1.7 (*Factorisation QR*)

Soit $A \in GL_n(\mathbb{C})$. Alors il existe une matrice $Q \in \mathcal{U}_n(\mathbb{C})$ et une matrice $R \in \mathcal{T}_n^s(\mathbb{C})$ (triangulaire supérieure) telles que $A = QR$. On a de plus unicité si on suppose les coefficients diagonaux de R réels positifs.

▷ − Unicité : Soient deux décompositions $A = QR = \tilde{Q}\tilde{R}$. On a :

$$\tilde{Q}^{-1}Q = \tilde{R}R^{-1} = B \in \mathcal{U}_n(\mathbb{C}) \cap \mathcal{T}_n^s(\mathbb{C})$$

En particulier, cette matrice est normale ($BB^* = B^*B$) et triangulaire donc diagonale (admis pour le moment). Les coefficients diagonaux de B sont précisément

$$B_{i,i} = \frac{\tilde{R}_{i,i}}{R_{i,i}} \in \mathbb{R}_+$$

Comme B est unitaire, $|B_{i,i}| = 1$ donc $B_{i,i} = 1$ et $B = I_n$. Ainsi $Q = \tilde{Q}$ et $R = \tilde{R}$.

− Existence : On applique Gramm-Schmidt aux colonnes $(v_i)_{1 \leqslant i \leqslant n}$ de A.

$$v_1 = \|v_1\| \, e_1$$

$$v_i = \|u_i\| \, e_i + \sum_{j=1}^{i-1} (v_i, e_j) e_j \qquad 2 \leqslant i \leqslant n$$

donc la matrice de changement de base R est triangulaire supérieure. Ainsi

$$A = QR$$

avec $A = (v_1|v_2|\dots|v_n)$, $Q = (e_1|e_2|\dots|e_n)$ et $R = \begin{pmatrix} \|v_1\| & & & (v_i, e_j) \\ & \|u_2\| & & \\ & & \ddots & \\ 0 & & & \|u_n\| \end{pmatrix}$ □

Remarque : L'algorithme de Gramm-Schmidt n'est pas celui utilisé en pratique pour des raisons de stabilité numérique (*cf.* chapitre 2).

Théorème 1.8 (*Schur, trigonalisation en base orthonormée*)

Soit $A \in \mathcal{M}_n(\mathbb{C})$. Alors il existe $U \in \mathcal{U}_n(\mathbb{C})$ et $T \in \mathcal{T}_n^s(\mathbb{C})$ telle que

$$A = UTU^*.$$

▷ On sait qu'il existe $P \in GL_n(\mathbb{C})$ et $S \in \mathcal{T}_n^s(\mathbb{C})$ telle que $A = PSP^{-1}$. On utilise la factorisation QR de P : $P = QR$ avec $Q \in \mathcal{U}_n(\mathbb{C})$ et $R \in \mathcal{T}_n^s(\mathbb{C})$. On obtient

$$A = QRSR^{-1}Q^* = UTU^*$$

avec $U = Q$ et $T = RSR^{-1} \in \mathcal{T}_n^s(\mathbb{C})$. □

1.2.2 Le cas des matrices normales

Définition 1.9

Une matrice réelle ou complexe A est normale si $AA^* = A^*A$.

Exemples : Sont normales les matrices diagonales, symétriques ou antisymétriques, hermitienne ou antihermitiennes, unitaires, orthogonales...

Lemme 1.10

Toute matrice triangulaire et normale est diagonale.

▷ On peut prendre $A \in \mathcal{T}_n^s(\mathbb{C})$ normale. On examine les coefficients diagonaux de AA^* et A^*A.

$$A^*A = \begin{pmatrix} |A_{1,1}|^2 & \\ & * \end{pmatrix}$$

$$AA^* = \begin{pmatrix} \sum_{j=1}^{n} |A_{1,j}|^2 & \\ & * \end{pmatrix}$$

Comme A est normale, $A_{1,j} = 0$ pour $2 \leqslant j \leqslant n$. On prouve de même que A est diagonale. □

Corollaire 1.11 (*du théorème de Schur*)

Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ de valeurs propres $\lambda_1, \ldots, \lambda_n$ est normale si et seulement si elle est diagonalisable en base orthonormée ie. il existe $U \in \mathcal{U}_n(\mathbb{C})$ telle que

$$U^*AU = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

▷ Partant du théorème de Schur : $A = UTU^*$. A étant normale et U unitaire, on a $AA^* = A^*A \iff TT^* = T^*T$ et par le lemme 1.10, T est diagonale :

$$T = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

à l'ordre près (il faut éventuellement utiliser une matrice de permutation, qui est orthogonale et unitaire si l'ordre est important). □

Remarque : On peut en fait écrire

$$A = \sum_{j=1}^{n} \lambda_i u_i u_i^* \qquad \text{où } u_i \text{ est la } i^e \text{ colonne de } U.$$

Les matrices $u_i u_i^*$ sont de rang 1. En effet,

$$(u_i u_i^*) \left(\sum_{j=1}^{n} \alpha_j u_j \right) = u_i \sum_{j=1}^{n} \alpha_j \underbrace{(u_i^* u_j)}_{\delta_{i,j}} = \alpha_i u_i$$

donc $Im(u_i u_i^*) = \text{Vect}(u_i)$.

Exemples : Les matrices déjà rencontrées dans l'exemple précédent sont diagonalisables en base orthonormée. De plus, les valeurs propres sont :

– réelles si A est symétrique réelle ou, plus généralement, hermitienne ;

– de module 1 si A est orthogonale réelle ou, plus généralement, unitaire ;

– imaginaires pures si A est antisymétrique réelle ou antihermitienne ;

– réelles (strictement) positives si A est symétrique (définie) positive ou hermitienne (définie) positive.

En effet, la matrice diagonale des valeurs propres de A conserve les propriétés de symétrique, unitaire, etc. de A. Par exemple,

$$\text{diag}(\lambda_1, \ldots, \lambda_n) = \overline{\text{diag}(\lambda_1, \ldots, \lambda_n)}^T \Rightarrow \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

1.3 Propriétés spectrales des matrices hermitiennes

Soit $A \in \mathcal{H}_n(\mathbb{C})$ $(A = A^*)$.

$$\forall x \in \mathbb{C}^n \qquad (Ax, x) \in \mathbb{R}.$$

En effet, $(Ax, x) = (x, A^*x) = (x, Ax) = \overline{(Ax, x)}$.

Définition 1.12 (*Quotient de Rayleigh*)

On définit l'application

$$r_A : \begin{array}{ccc} \mathbb{C}^n & \to & \mathbb{R} \\ x & \mapsto & \dfrac{(Ax, x)}{(x, x)} \end{array}$$

Propriété 1.13

Soit $\lambda \in \mathbb{C}\backslash\{0\}$. Alors $\forall x \in \mathbb{C}^n\backslash\{0\},\ r_A(\lambda x) = r_A(x)$.

Pour étudier cette fonction, il suffit de l'étudier sur la sphère unité (hermitienne) de $\mathbb{C}^n$ qui est compacte.

Théorème 1.14

Supposons ordonnées les valeurs propres de A : $\lambda_1 \leqslant \cdots \leqslant \lambda_n$. Alors

$$\min_{\mathbb{C}^n\backslash\{0\}} r_A = \lambda_1 \qquad \max_{\mathbb{C}^n\backslash\{0\}} r_A = \lambda_n$$

▷ On réduit A...(exercice) □

Théorème 1.15 (*Caractérisation min-max de Courant-Fisher*)

Soit $k \in [\![1, n]\!]$. On note $\mathcal{E}_k$ l'ensemble des sous-espaces vectoriels de $\mathbb{C}^n$ de dimension k. On a :

$$\lambda_k = \min_{W \in \mathcal{E}_k} \max_{\substack{x \in W \\ x \neq 0}} r_A(x).$$

▷ Notons $u_1, \ldots, u_n$ une base orthonormée de vecteurs propres de A. Soient $k \in [\![1, n]\!]$ et $W \in \mathcal{E}_k$. On considère le sous-espace vectoriel $V = W \cap \mathrm{Vect}(u_k, \ldots, u_n)$ qui est de dimension au moins égale à 1. On considère alors $x \in V\backslash\{0\}$ qu'on écrit sous la forme $x = \sum_{j=k}^{n} \alpha_j u_j$. Alors,

$$(Ax, x) = \sum_{j=k}^{n} \lambda_j |\alpha_j|^2, \qquad (x, x) = \sum_{j=k}^{n} |\alpha_j|^2,$$

de sorte que $r_A(x) \geqslant \lambda_k$. Par suite, il vient que $\max_{x \in W} r_A(x) \geqslant \max_{x \in V} r_A(x) \geqslant \lambda_k$. Alors,

$$\inf_{W \in \mathcal{E}_k} \max_{x \in W} r_A(x) \geqslant \lambda_k.$$

Montrons que la borne inférieure est atteinte. Soit $W = \mathrm{Vect}(u_1, \ldots, u_k)$. Alors pour $x = u_k$, on obtient $r_A(x) = \lambda_k$ de sorte que $\max_{x \in W} r_A(x) = \lambda_k$. Ainsi, la borne inférieure est un minimum. □

Théorème 1.16 (*Weyl (perturbation de valeurs propres)*)

Soient $A, B \in H_n(\mathbb{C})$ et $\lambda_i(A),\ \lambda_i(B),\ \lambda_i(A+B)$ les n valeurs propres (réelles) de A, B et $A + B$ respectivement, rangées dans l'ordre croissant. Alors pour

tout $k \in [\![1, n]\!]$ *:*

$$\lambda_k(A) + \lambda_1(B) \leqslant \lambda_k(A+B) \leqslant \lambda_k(A) + \lambda_n(B).$$

▷ Soit $x \in \mathbb{C}^n$. Alors

$$\lambda_1(B) \leqslant r_B(x) \leqslant \lambda_n(B).$$

Or $r_{A+B}(x) = r_A(x) + r_B(x)$ et en utilisant la caractérisation min-max de Courant-Fisher des valeurs propres de A, il vient les inégalités annoncées. □

1.4 Norme matricielle et rayon spectral

1.4.1 Rappels et compléments

Définition 1.17

Soit E un $\mathbb{K}$-espace vectoriel. On appelle norme sur E, notée $\|.\|$ toute application définie sur E à valeurs dans $\mathbb{R}_+$ telle que

(i) $\forall x \in E, \quad \|x\| = 0 \quad \Rightarrow \quad x = 0,$

(ii) $\forall \alpha \in \mathbb{K},\ \forall x \in E, \quad \|\alpha x\| = |\alpha|\,\|x\|$

(iii) $\forall x, y \in E, \quad \|x + y\| \leqslant \|x\| + \|y\|.$

Définition 1.18

Une norme $\|.\|$ sur $\mathcal{M}_n(\mathbb{K})$ est dite matricielle (ou d'algèbre) si elle vérifie la propriété de sous-multiplicativité :

$$\forall A, B \in \mathcal{M}_n(\mathbb{K}), \qquad \|AB\| \leqslant \|A\|\,\|B\|\,.$$

On peut construire facilement des normes matricielles en se basant sur une norme sur $\mathbb{K}^n$ et en considérant un norme induite sur $\mathcal{M}_n(\mathbb{K})$, qui sera dite également subordonnée.

Définition 1.19

L'application $A \mapsto |||A||| = \sup_{\|x\|=1} \|Ax\|$ définit une norme matricielle sur $\mathcal{M}_n(\mathbb{K})$ dite norme subordonnée à $\|.\|$ (ou induite par $\|.\|$).

Avec un léger abus de notation, on note parfois $\|A\|$ (au lieu de $|||A|||$) la norme de A subordonnée à la norme vectorielle $\|x\|$.

Proposition 1.20

Une norme subordonnée $|||.|||$ vérifie les propriétés suivantes :

(i) $\forall A \in \mathcal{M}_n(\mathbb{K}),\ \forall x \in \mathbb{K}^n, \qquad \|Ax\| \leqslant |||A||| \, \|x\|.$

(ii) $|||A||| = \sup_{x \neq 0} \dfrac{\|Ax\|}{\|x\|}.$

(iii) $|||A||| = \max_{x \neq 0} \dfrac{\|Ax\|}{\|x\|}.$

En d'autres termes, $|||A|||$ est la plus petite constante C telle que pour tout $x \in \mathbb{K}^n$ on ait $\|Ax\| \leqslant C \, \|x\|$. La recherche d'une telle constante passe généralement par les deux étapes suivantes : (i) obtenir une majoration de cette forme et donc un majorant de $|||A|||$, puis (ii) prouver que le calcul est optimal (s'il l'est bien) en exhibant un vecteur x non nul qui réalise l'égalité (ou une suite de vecteurs x_n qui donne à la limite l'égalité).

Remarque : Attention, toutes les normes matricielles ne sont pas subordonnées (il suffit de choisir une norme matricielle telle que $\|I_n\| \neq 1$). Même sous la contrainte supplémentaire que $\|I_n\| = 1$, on peut encore trouver des contre-exemples.

Proposition 1.21

On considère $x \in \mathbb{C}^n$ et $A \in \mathcal{M}_n(\mathbb{C})$.

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{i,j}|$$

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| \qquad \|A\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{i,j}|.$$

▷ Considérons par exemple le cas de la norme $\|.\|_1$. Soit $x \in \mathbb{C}^n$ et

$$Ax = \left(\sum_{j=1}^{n} a_{i,j} x_j \right)_{1 \leqslant i \leqslant n}.$$

Alors

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{i,j} x_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}||x_j|,$$

par l'inégalité triangulaire. Dès lors, on peut intervertir les deux sommes et mettre en facteur le terme $|x_j|$ dans la somme sur i. On a donc

$$\|Ax\|_1 \leqslant \sum_{j=1}^{n} \left(|x_j| \sum_{i=1}^{n} |a_{i,j}| \right).$$

Il suffit alors de majorer chacune des sommes sur i par leur maximum (atteint pour une certaine valeur de j). On a :

$$\|Ax\|_1 \leqslant \left(\sum_{j=1}^{n} |x_j|\right) \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}| = \|x\|_1 \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|.$$

Ainsi, on a l'inégalité

$$\|A\|_1 \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|.$$

Pour obtenir l'égalité, il suffit de montrer que le maximum est atteint pour un certain choix du vecteur x. Pour ce faire, il faut qu'il y ait égalité au niveau de l'inégalité triangulaire (coefficients positifs ou nuls) et au niveau du remplacement de chaque somme par leur maximum (égalité avec le maximum de chaque composante ou nullité de $|x_j|$). Ces considérations mènent à poser k l'indice tel que $\max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}| = \sum_{i=1}^{n} |a_{i,k}|$, et x de composantes $x_j = 0$ si $j \neq k$ et $x_k = 1$. Alors, $\|Ax\|_1 = \sum_{i=1}^{n} |a_{i,k}$ tandis que $\|x\|_1 = 1$. □

1.4.2 Rayon spectral d'une matrice

Définition 1.22 (*Rayon spectral*)

Soit $A \in \mathcal{M}_n(\mathbb{K})$. On appelle rayon spectral de A la quantité suivante :

$$\rho(A) = \max\{|\lambda|,\ \lambda \in \sigma(A)\}.$$

Lemme 1.23

Si A est diagonalisable en base orthonormée, alors $\|A\|_2 = \rho(A)$. Plus généralement, pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$, on a

$$\|A\|_2 = \sqrt{\rho(A^*A)}.$$

▷ Dans un premier temps, si A est normale, alors on écrit la diagonalisation $A = UDU^*$ de sorte que pour tout $x \in \mathbb{K}^n$,

$$\|Ax\|_2 = \|UDU^*x\|_2 = \|DU^*x\|_2$$

puisque U préserve la norme 2. Alors

$$\|A\|_2 = \max\{\|Ax\|_2,\ \|x\|_2 = 1\} = \max\{\|DU^*x\|_2,\ \|x\|_2 = 1\}$$

donc

$$\|A\|_2 = \max\{\|DU^*x\|_2\,,\ \|U^*x\|_2 = 1\} = \|D\|_2\,.$$

Il est alors facile de voir, D étant diagonale, que $\|D\|_2 = \rho(D) = \rho(A)$.

Dans le cas général, remarquons dans un premier temps que A^*A étant hermitienne, donc normale, on a d'après ce qui précède $\|A^*A\|_2 = \rho(A^*A)$. Par ailleurs, on montre que $\|A\|_2 = \|A^*\|_2 = \sqrt{\|A^*A\|_2}$ ce qui suffit à conclure. Pour cela, soit $x \in \mathbb{K}^n$ non nul :

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax) \leqslant \|x\|_2 \|A^*Ax\|_2$$

grâce à l'inégalité de Cauchy-Schwarz. Il vient alors

$$\|Ax\|_2^2 \leqslant \|A^*A\|_2 \|x\|_2^2\,.$$

On en déduit donc que $\|A\|_2^2 \leqslant \|A^*A\|_2$. Par ailleurs, par l'inégalité de norme d'algèbre, on a $\|A^*A\|_2 \leqslant \|A^*\|_2 \|A\|_2$ d'où (en supposant que $A \neq 0$ sans quoi le résultat est trivial) $\|A\|_2 \leqslant \|A^*\|_2$. Par le rôle symétrique de A et A^* on a de même $\|A^*\|_2 \leqslant \|A\|_2$, d'où l'égalité $\|A^*\|_2 = \|A\|_2$ et, au passage, l'égalité

$$\|A\|_2^2 = \|A^*A\|_2\,.$$

□

Théorème 1.24

Soit $A \in \mathcal{M}_n(\mathbb{K})$ fixée.
(i) Pour toute norme subordonnée $\|.\|$, $\rho(A) \leqslant \|A\|$.
(ii) $\forall \varepsilon > 0,\ \exists\, \|.\|_\varepsilon$ subordonnée, $\quad \|A\|_\varepsilon \leqslant \rho(A) + \varepsilon$.

▷ (*i*) Par définition $\rho(A) = |\lambda|$ où $\lambda \in \sigma(A)$ associée à un vecteur propre $x \in \mathbb{K}^n$. Ainsi, $\|Ax\| = |\lambda| \|x\|$ et donc $\|A\| \geqslant |\lambda| = \rho(A)$.

(*ii*) Par le théorème de Schur : $A = UTU^*$, $U \in U_n \in (\mathbb{C})$, T triangulaire supérieure.

$$\|T\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |T_{i,j}| = \max_{1\leqslant j\leqslant n} \left(|\lambda_j| + \sum_{i=1}^{j-1} |T_{i,j}| \right).$$

On va « effacer » les termes non diagonaux de T. Soit $\eta > 0$, on pose

$$D(\eta) = \begin{pmatrix} \eta & & & 0 \\ & \eta^2 & & \\ & & \ddots & \\ 0 & & & \eta^n \end{pmatrix}$$

et

$$T(\eta) = D(\eta)^{-1} T D(\eta) = \begin{pmatrix} \lambda_1 & & (\eta^{j-i} T_{i,j}) \\ & \ddots & \\ & & \ddots \\ 0 & & \lambda_n \end{pmatrix}.$$

Alors

$$\|T(\eta)\|_1 = \max_{1 \leqslant j \leqslant n} \left(|\lambda_j| + \sum_{i=1}^{j-1} \eta^{j-i} |T_{i,j}| \right) \xrightarrow[\eta \to 0]{} \rho(A).$$

Pour $\varepsilon > 0$ fixé, il existe $\eta > 0$ tel que $\|T(\eta)\|_1 \leqslant \rho(A) + \varepsilon$. On a :

$$A = UTU^* = (UD(\eta))T(\eta)(D(\eta)^{-1}U^*).$$

On considère sur $\mathbb{C}^n$ la norme

$$\|x\|_\varepsilon = \left\| D(\eta)^{-1} U * x \right\|_1.$$

Alors, pour $x \in \mathbb{C}^n$,

$$\|Ax\|_\varepsilon = \left\| (D(\eta)^{-1}U*)(UD(\eta)T(\eta)D(\eta)^{-1}U*)x \right\|_1 = \left\| T(\eta)(D(\eta)^{-1}U * x) \right\|_1 \leqslant \|T(\eta)\|_1 \|x\|_\varepsilon.$$

Ainsi,

$$\|A\|_\varepsilon \leqslant \|T(\eta)\|_1 \leqslant \rho(A) + \varepsilon.$$

□

1.5 Caractérisation de la convergence de suites et séries de matrices

Théorème 1.25

Soit $A \in \mathcal{M}_n(\mathbb{K})$. Les assertions suivantes sont équivalentes :
(i) $\lim_{k \to +\infty} A^k = 0$
(ii) $\rho(A) < 1$
(iii) Il existe une norme subordonnée $\|.\|$ telle que $\|A\| < 1$.

▷ $(ii) \Leftrightarrow (iii)$ d'après le théorème précédent.
$(i) \Rightarrow (ii)$ Supposons que $A^k \xrightarrow[k \to +\infty]{} 0$. Soit $x \in \mathbb{K}^n$ un vecteur propre non nul associé à la valeur propre λ tel que $\rho(A) = |\lambda|$. Alors $A^k x = \lambda^k x$ et donc

$$\left\| A^k x \right\| = |\lambda^k| \|x\| = \rho(A)^k \|x\| \xrightarrow[k \to +\infty]{} 0$$

implique que $\rho(A) < 1$.

$(iii) \Rightarrow (i)$ S'il existe une norme subordonnée telle que $\|A\| < 1$ alors

$$\left\|A^k\right\| \leqslant \|A\|^k \xrightarrow[k\to+\infty]{} 0$$

donc $A^k \xrightarrow[k\to+\infty]{} 0$. □

Remarque : Sous ces conditions, $I - A$ est inversible, d'inverse donné par la série absolument convergente $\sum_{k\geqslant 0} A^k$.

Corollaire 1.26

Pour tout norme subordonnée,

$$\lim_{k\to+\infty} \left\|A^k\right\|^{\frac{1}{k}} = \rho(A).$$

▷ Si λ est valeur propre de A alors λ^k est valeur propre de A^k donc

$$\rho(A^k) = \rho(A)^k.$$

Ainsi,

$$\rho(A) = \rho(A^k)^{\frac{1}{k}} \leqslant \left\|A^k\right\|^{\frac{1}{k}}$$

car $\rho(A^k) \leqslant \|A\|_k$.

De plus, pour $\varepsilon > 0$, en posant $A_\varepsilon = (\rho(A)+\varepsilon)^{-1}A$, on a $\rho(A_\varepsilon) = \dfrac{\rho(A)}{\rho(A)+\varepsilon} < 1$ de sorte que $A_\varepsilon^k \xrightarrow[k\to 0]{} 0$. En particulier, il existe $k_\varepsilon \in \mathbb{N}$ tel que $k \geqslant k_\varepsilon \Rightarrow \left\|A_\varepsilon^k\right\| < 1$. Or $\left\|A_\varepsilon^k\right\| = (\rho(A)+\varepsilon)^{-1}\left\|A^k\right\|$ d'où

$$\left\|A^k\right\| < (\rho(A)+\varepsilon)^k.$$

□

Remarque : − De manière plus générale, la suite $(A^k)_{k\in\mathbb{N}}$ est bornée indépendamment de k si et seulement si $\rho(A) \leqslant 1$ et les valeurs propres de A de module 1 sont non défectives (semi-simple, la multiplicité algébrique est égale à la multiplicité géométrique).

Exemple : $A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$, $|\lambda| = 1$. Alors

$$A^k = \begin{pmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{pmatrix}$$

non bornée à cause du terme $|k\lambda^{k-1}| = k \xrightarrow[k\to+\infty]{} +\infty$

– Si $\|.\|$ est une norme quelconque sur $\mathcal{M}_n(\mathbb{K})$ on a toujours

$$\left\|A^k\right\|^{\frac{1}{k}} \xrightarrow[k\to+\infty]{} \rho(A).$$

En effet, toutes les normes sont équivalentes donc il existe $\|.\|_s$ et $C_1, C_2 > 0$ telles que

$$\forall B \in \mathcal{M}_n(\mathbb{K}), \qquad C_1 \|B\|_s \leqslant \|B\| \leqslant C_2 \|B\|_s .$$

Alors, pour $B = A^k$:

$$C_1 \left\|A^k\right\|_s \leqslant \left\|A^k\right\| \leqslant C_2 \left\|A^k\right\|_s$$

et

$$\underbrace{C_1^{\frac{1}{k}} \left\|A^k\right\|_s^{\frac{1}{k}}}_{\longrightarrow \rho(A)} \leqslant \left\|A^k\right\|^{\frac{1}{k}} \leqslant \underbrace{C_2^{\frac{1}{k}} \left\|A^k\right\|_s^{\frac{1}{k}}}_{\longrightarrow \rho(A)} .$$

Application : Soit $U_0 \in \mathbb{K}^n$. Pour $k \geqslant 0$ on pose $U_{k+1} = AU_k + V_k$ où $(V_k)_{k\geqslant 0}$ est une suite de $\mathbb{K}^n$ donnée et $A \in \mathcal{M}_n(\mathbb{K})$. En réalité, ce qui est calculé est

$$\tilde{U}_{k+1} = A\tilde{U}_k + V_k + \varepsilon_k$$

où $\varepsilon_k \in \mathbb{K}^n$ représente les « erreurs » à l'étape k. On a aussi

$$\tilde{U}_0 = U_0 + \eta$$

où $\eta \in \mathbb{K}^n$. On cherche à estimer $\left\|U_k - \tilde{U}_k\right\|$ à partir de ces données. La suite $W_k = \tilde{U}_k - U_k$ vérifie la relation de récurrence :

$$\begin{cases} W_0 = \eta \\ \forall k \geqslant 0,\ W_{k+1} = AW_k + \varepsilon_k. \end{cases}$$

Par une formule de Duhamel (obtenue par récurrence sur k), on a :

$$\forall k \geqslant 0, \qquad W_k = A^k\eta + \sum_{j=0}^{k-1} A^{k-1-j}\varepsilon_j.$$

Ainsi, pour une norme subordonnée,

$$\|W_k\| \leqslant \left\|A^k\right\| \|\eta\| + \sum_{j=0}^{k-1} \left\|A^{k-1-j}\right\| \|\varepsilon_j\| .$$

Si $\rho(A) \leqslant 1$ et les valeurs propres de A de module 1 sont semi-simples, il existe $C > 0$ tel que $\forall k \geqslant 0,\ \left\|A^k\right\| \leqslant C$. Alors

$$\forall k \geqslant 0, \qquad \|W_k\| \leqslant C(\|\eta\| + k \max_{0\leqslant j\leqslant k-1} \|\varepsilon_j\|).$$

L'erreur augmente donc au plus comme k.

Si $\rho(A) > 1$, on peut s'attendre à une explosion de l'erreur en $k\rho(A)^k$.

1.6 Conditionnement

On se place sur $\mathbb{K}^n$ muni d'une norme $\|.\|$ et $\mathcal{M}_n(\mathbb{K})$ muni de la norme subordonnée $\|.\|$.

Définition 1.27

On appelle conditionnement d'une matrice $A \in GL_n(\mathbb{K})$ dans la norme $\|.\|$ la quantité

$$\mathrm{Cond}(A) = \|A\| \left\|A^{-1}\right\|.$$

Proposition 1.28

Soient $A, B \in GL_n(\mathbb{K})$ et $\alpha^{\mathbb{K}}$. On a :
(i) $\mathrm{Cond}(A) > 0$.
(ii) $\mathrm{Cond}(A) \geqslant 1$.
(iii) si $\alpha \neq 0$, $\mathrm{Cond}(\alpha A) = \mathrm{Cond}(A)$.
(iv) $\mathrm{Cond}(AB) \leqslant \mathrm{Cond}(A)\,\mathrm{Cond}(B)$.

▷ *(ii)* $I_n = AA^{-1}$ et $\|I_n\| = 1$ donc $1 = \|AA^{-1}\| \leqslant \|A\|\,\|A^{-1}\|$. □

Théorème 1.29

Soient $A \in GL_n(\mathbb{K})$ et $b \in \mathbb{K}^n$. On considère $x \in \mathbb{K}^n$ solution de $Ax = b$.
(i) Soient $\delta x \in \mathbb{K}^n$ et $\delta b \in \mathbb{K}^n$ tels que $A(x + \delta x) = b + \delta b$. Alors

$$\frac{\|\delta x\|}{\|x\|} \leqslant \mathrm{Cond}(A)\frac{\|\delta b\|}{\|b\|}.$$

(ii) Soient $\delta x \in \mathbb{K}^n$ et $\delta A \in \mathcal{M}_n(\mathbb{K})$ tels que $A + \delta A \in GL_n(\mathbb{K})$ et

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

Alors,

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leqslant \mathrm{Cond}(A)\frac{\|\delta A\|}{\|A\|}.$$

Le conditionnement quantifie l'influence relative d'une perturbation des données sur le résultat.
▷ *(i)* On a $A\delta x = \delta b$ donc $\delta x = A^{-1}\delta b$ et $\|\delta x\| \leqslant \|A^{-1}\|\,\|\delta b\|$. Par ailleurs, $\|b\| = \|Ax\| \leqslant \|A\|\,\|x\|$. Par produit

$$\|\delta x\|\,\|b\| \leqslant \mathrm{Cond}(A)\,\|x\|\,\|\delta b\|.$$

(*ii*) On a $A\delta x + \delta A(x+\delta x) = 0$ donc $\delta x = -A^{-1}\delta A(x+\delta x)$ d'où

$$\|\delta x\| \leqslant \left\|A^{-1}\right\| \|\delta A\| \|x+\delta x\| = \mathrm{Cond}(A)\frac{\|\delta A\|}{\|A\|}\|x+\delta x\|\,.$$

□

On parle de système bien conditionné si $\mathrm{Cond}(A) \simeq 1$ et de système mal conditionné si $\mathrm{Cond}(A) >> 1$.

Théorème 1.30 (*conditionnement en norme 2*)

(*i*) *Soit* $A \in GL_n(\mathbb{K})$.

$$\mathrm{Cond}_2(A) = \sqrt{\frac{\max(|\lambda|,\ \lambda \in \sigma(A^*A))}{\min(|\lambda|,\ \lambda \in \sigma(A^*A))}}$$

(*ii*) *Si* A *est normale,*

$$\mathrm{Cond}_2(A) = \frac{\max(|\lambda|,\ \lambda \in \sigma(A))}{\min(|\lambda|,\ \lambda \in \sigma(A))}$$

(*iii*) *En particulier, si* A *est unitaire, alors* $\mathrm{Cond}_2(A) = 1$ *car* $\sigma(A) \subset \{z \in \mathbb{C},\ |z| = 1\}$. *Il y a en fait équivalence.*

L'intérêt des matrices orthogonales ou unitaires est donc de ne pas amplifier les erreurs. Par exemple, plutôt que de résoudre $Ax = b$, on cherche $Q \in O_n(\mathbb{R})$ et $R \in \mathcal{T}_n^s(\mathbb{R})$ telles que $A = QR$ et on résout $Rx = Q^*b$. Les erreurs sur x seront « réduites ».

Chapitre 2

Résolution de systèmes linéaires $Ax = b$

Exemple pratique : On cherche à résoudre le problème différentiel

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

On recherche u solution (supposée régulière). Formellement, on pose pour tout $0 \leqslant i \leqslant n$ $x_i = ih = \frac{i}{n}$. La famille $(x_i)_{0 \leqslant i \leqslant n}$ est une subdivision uniforme de $[0,1]$. On définit $U = (u(x_i))_{1 \leqslant i \leqslant n-1}$ qui est la nouvelle inconnue du problème. On construit un système linéaire dont U est solution (approchée). Si h est suffisamment petit,

$$u'(x_i) \simeq \frac{u(x_{i+1} - x_i)}{h} \qquad u''(x_i) \simeq \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}$$

car $u(x_{i+1}) = u(x_i + h) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \dots$ et $u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) + \dots$ u vérifie le système linéaire :

$$\begin{cases} \forall 1 \leqslant i \leqslant n-1,\ \dfrac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} + u(x_i) \simeq f(x_i) \\ u(x_0) = u(x_n) = 0. \end{cases}$$

ou bien

$$AU = F$$

avec $F = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_{n-1}) \end{pmatrix}$, $A = I_{n-1} - \frac{1}{h^2}B$, $B = \begin{pmatrix} -2 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix} \in \mathcal{M}_{n-1}(\mathbb{R})$.

Après avoir résolu $AU = F$, on s'attend à avoir U proche de la solution exacte.

Remarque : Par les formules de Cramer, pour résoudre $Ax = b$, $A \in GL_n(\mathbb{R})$ il faut calculer $(n+1)$ déterminants de taille n. Cela nécessite le l'ordre de $(n+1)!$ opérations. Pour $n = 50$, sur une machine de 1 exaflop (10^{18} opérations par seconde), il faut 10^{39} années.

Les méthodes utilisables nécessitent de l'ordre de n^3 opérations.

Remarque : Pour résoudre $Ax = b$, on ne calcule jamais A^{-1}. Dans Scilab, on écrit $x = A\backslash b$ et non $x = \text{inv}(A) * b$. Pour calculer l'inverser, il suffit de résoudre les n systèmes linéaires $Ax^{(i)} = e_i$ où les (e_i) sont les vecteurs de la base canonique. $x^{(i)}$ est alors la i^e colonne de A^{-1}.

On distingue deux types de méthodes de résolution : les méthodes directes permettent d'obtenir la solution exacte (théoriquement) après un nombre fini d'étapes de calcul (en général de l'ordre de n^3) ; les méthodes itératives consistent à construire des suites (x_k) de $\mathbb{K}^n$ qui converge vers la solution x.

2.1 Méthodes directes

2.1.1 Résolution d'un système triangulaire

Si $A \in GL_n(\mathbb{K})$ est triangulaire supérieure, on obtient x par méthode de remontée : on pose

$$\begin{cases} x_n = \dfrac{b_n}{a_{n,n}} \\ \\ \forall i = n-1 \dots 1,\ x_i = \dfrac{b_i - \displaystyle\sum_{k=i+1}^{n} a_{i,k} x_k}{a_{i,i}} \end{cases}$$

Si $A \in GL_n(\mathbb{K})$ est triangulaire inférieure, on obtient x par méthode de descente : on pose

$$\begin{cases} x_1 = \dfrac{b_1}{a_{1,1}} \\ \\ \forall i = 2 \dots n,\ x_i = \dfrac{b_i - \displaystyle\sum_{k=i+1}^{n} a_{i,k} b_k}{a_{i,i}} \end{cases}$$

On effectue de l'ordre de n^2 opérations élémentaires.

2.1.2 Méthode de Gauss

Opérations élémentaires : – permutation : matrice de transposition,

$$I_n - E_{i,i} - E_{j,j} + E_{i,j} + E_{j,i}.$$

– dilatation :

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \lambda & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \qquad \lambda \in \mathbb{K}^*$$

– transvection :

$$I_n + \lambda E_{i,j} = T(\lambda, i, j)$$

Multiplier à gauche consiste à opérer sur les lignes. Multiplier à droite consiste à opérer sur les colonnes.

$$T(\lambda, i, j)A \quad : \quad \ell_i \leftarrow \ell_i + \lambda \ell_j.$$

Pour résoudre $Ax = b$ on opère à gauche ! C'est le procédé d'élimination de Gauss.

2.1.3 Méthode LU

On recherche une factorisation de A sous la forme $A = LU$ avec $L \in \mathcal{T}_n^-(\mathbb{K})$, $U \in \mathcal{T}_n^+(\mathbb{K})$ et $L_{i,i} = 1,\ \forall i \in [\![1, n]\!]$.

Définition 2.1

Soit $A \in \mathcal{M}_n(\mathbb{K})$. On appelle mineurs fondamentaux de A les déterminants suivants :

$$k \in [\![1, n]\!], \qquad \det((a_{i,j})_{1 \leqslant i,j \leqslant k})$$

Théorème 2.2

Soit $A \in \mathcal{M}_n(\mathbb{K})$ dont tous les mineurs fondamentaux sont non nuls. Alors il existe un unique couple $(L, U) \in \mathcal{T}_n^-(\mathbb{K}) \times \mathcal{T}_n^+(\mathbb{K})$ tel que

$$\forall i \in [\![1, n]\!] \quad L_{i,i} = 1 \qquad A = LU.$$

Conséquence : Une fois L et U connues

$$Ax = b \iff \begin{cases} Ly = b & \text{(descente)} \\ Ux = y & \text{(remontée)} \end{cases}$$

⚠Les coefficients diagonaux ne sont pas les valeurs propres de A ! Mais

$$\det A = \prod_{i=1}^{n} U_{i,i}$$

L'hypothèse du théorème n'est pas équivalente à $A \in GL_n(\mathbb{K})$. Contre-exemple :

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

est inversible mais ne vérifie pas l'hypothèse.

Dans le cas où $A \in GL_n(\mathbb{K})$ seulement, on peut utiliser la variante : il existe P une matrice de permutation telle que $P^T A$ vérifie les hypothèses du théorème. Alors, $A = PLU$.

Remarque : Variante plus symétrique : sous les hypothèses du théorème, on peut écrie $A = LDU$ où $\forall i \in [\![1, n]\!],\ L_{i,i} = U_{i,i} = 1$ et $D \in D_n(\mathbb{K})$ (diagonale).

▷ Unicité : On suppose que $A = L_1U_1 = L_2U_2$ sont deux décompositions LU de A. Alors,

$$L_2^{-1}L_1 = U_2U_1^{-1} \in \mathcal{T}_n^+(\mathbb{K}) \cap \mathcal{T}_n^-(\mathbb{K}) = D_n(\mathbb{K}).$$

Par ailleurs, $(L_2^{-1}L_1)_{i,i} = 1\ \forall i$ donc $L_2^{-1}L_1 = U_2U_1^{-1} = I_n$ d'où l'unicité.

Existence (Algorithme LU) :

Lemme 2.3

Soit $A \in \mathcal{M}_n(\mathbb{K})$ ayant ses mineurs fondamentaux tous non nuls et soit $T(\lambda, i, j)$ une matrice de transvection avec $i > j$. Alors $T(\lambda, i, j)$ a ses mineurs fondamentaux également non-nuls.

▷ Soit $k \in [\![1, n]\!]$. En écrivant A par blocs avec un bloc $A_k \in \mathcal{M}_n(\mathbb{K})$ on obtient :

$$T(\lambda, i, j)A = \begin{pmatrix} T_k & 0 \\ \tilde{T} & T_{n-k} \end{pmatrix} \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix} = \begin{pmatrix} T_kA_k & * \\ * & * \end{pmatrix}$$

Le mineur fondamental de $T(\lambda, i, j)A$ d'ordre k est $\det(T_kA_k) = \det(T_k) = \det(A_k) \neq 0$. □

Soit $k \in [\![1, n-1]\!]$. Supposons construites des matrices $T_1, \dots T_{k_1}$ chacune produit de transvections telles que

$$T_{k-1} \dots T_1 A = \begin{pmatrix} a_{1,1}^{(k)} & & & a_{i,j}^{(k)} \\ & \ddots & & \\ & & a_{k,k}^{(k)} & \\ & & \vdots & \\ 0 & & a_{n,k}^{(k)} & \end{pmatrix}$$

On remarque que $a_{k,k}^{(k)} \neq 0$ par le lemme précédent. On peut donc éliminer, sur la colonne k, les lignes $k+1$ à n à partir de la ligne k. On pose

$$\forall j \in [\![k+1, n]\!], \quad x_{j,k}^{(k+1)} = \frac{a_{j,k}^{(k)}}{a_{k,k}^{(k)}}$$

et

$$T_k = \prod_{j=k+1}^{n} T(-x_{j,k}^{(k+1)}, j, k)$$

$(\ell_j \leftarrow \ell_j - x_{j,k}^{(k+1)} \ell_k, \ \forall j)$. Alors, $T_k \dots T_1 A$ est de la forme :

$$\begin{pmatrix} a_{1,1}^{(k+1)} & & & a_{i,j}^{(k+1)} \\ & \ddots & & \\ & & a_{k+1,k+1}^{(k+1)} & \\ & & \vdots & \\ 0 & & a_{n,k+1}^{(k+1)} & \end{pmatrix}$$

et $\forall i \in [\![1, k]\!]$ $a_{i,j}^{(k+1)} = a_{i,j}^{(k)}$.

À la dernière étape, $T_{n-1} \dots T_1 A$ est triangulaire supérieure. On l'appelle U et on pose

$$L = (T_{n-1} \dots T_1)^{-1}.$$

L est bien triangulaire inférieure et ses termes diagonaux sont égaux à 1. □

Remarque : L est connue directement :

$$L = T_1^{-1} \dots T_n^{-1} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ x_{i,j}^{(j+1)} & & 1 \end{pmatrix}$$

Synthèse de l'algorithme :
Pour k de 1 à $n-1$
  Pour $i = k+1$ à n
    $a_{i,k} \leftarrow \dfrac{a_{i,k}}{a_{k,k}}$ calcul du coefficient d'élimination
    Pour $j = k+1$ à n
      $a_{i,j} \leftarrow a_{i,j} - a_{i,k}a_{k,j}$ élmination
    Fin
  Fin
Fin

Le résultat de cet algorithme est une matrice B avec $(B_{i,j})_{1 \leqslant i \leqslant j \leqslant n}$ sont les coefficients non nuls de U et $(B_{i,j})_{1 \leqslant j < i \leqslant n}$ sont les coefficients non nuls et non diagonaux de L.

Remarque : Conservation du profil : si A est une matrice bande de demi-largeur de bande $p \in \mathbb{N}$ (*ie.* $a_{i,j} = 0$ pour $|i-j| > p$) alors $A = LU$ avec L et U de même profil. Si A a une bande de largeur $2p+1$ alors L a une bande de largeur p (sous la diagonale) et U une bande de largeur $p+1$ (au-dessus de la diagonale).

Nombre d'opérations dans le cas général :

$$\sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \left(1 + \sum_{j=k+1}^{n} 2\right) = \sum_{k=1}^{n-1} k(1+2k) = \frac{2n^3}{3} + o(n^3).$$

2.1.4 Méthode de Cholesky

Théorème 2.4

Si $A \in \mathcal{M}_n(\mathbb{R})$ est symétrique définie positive. Alors il existe une unique matrice $B \in \mathcal{T}_n^-(\mathbb{K})$ telle que $\forall i\ B_{i,i} > 0$ et $A = BB^T$.

Rappel : Les mineurs fondamentaux de A sont strictement positifs (considérer la restriction de la forme bilinéaire symétrique définie positive à $\mathbb{R}^k \times \mathbb{R}^k$).

$\triangleright$ Existence : On peut obtenir $A = LDU$ et comme $A^T = A$ on a $U = L^T$ (par unicité). Donc $A = LDL^T$. De plus, les coefficients diagonaux de D sont réels positifs strictement : si $L^T x = e_i$ (vecteur de la base canonique) alors

$$0 < (Ax, x) = (LDL^T x, x) = (DL^T x, L^T x) = D_{i,i}.$$

On pose finalement $B = L\text{diag}(\sqrt{D_{1,1}}, \ldots, \sqrt{D_{n,n}})$ et $A = BB^T$.

Unicité : si $A = B_1 B_1^T = B_2 B_2^T$ alors $X = B_2^{-1} B_1 = B_2^T (B_1^T)^{-1}$ est diagonale de coefficients strictement positifs et

$$X^2 = XX^T = B_2^{-1} B_1 B_1^{-1} B_2 = I_n$$

donc $X = I_n$. □

Le calcul de B peut être obtenu à partir de $A = LU$, mais en pratique on procède par identification (pour des raisons de stabilité) : on résout les équations données par les coefficients de $BB^T = A$

$$\forall i, j \qquad a_{i,j} = \sum_{k=1}^{\min(i,j)} b_{i,k} b_{j,k}.$$

On obtient

$$\begin{cases} b_{1,1} = \sqrt{a_{1,1}} \\ \forall i = 2 \ldots n, \quad b_{i,1} = \dfrac{a_{i,1}}{b_{1,1}} \\ \forall j = 2 \ldots n, \qquad b_{j,j} = \sqrt{a_{j,j} - \displaystyle\sum_{k=1}^{j-1} b_{j,k}^2} \\ \qquad\qquad\qquad \forall i = j+1 \ldots n, \qquad b_{i,j} = \dfrac{a_{i,j} - \displaystyle\sum_{k=1}^{k-1} b_{i,k} b_{j,k}}{b_{j,j}}. \end{cases}$$

Nombre d'opérations : $\dfrac{n^3}{3}$ (le caractère symétrique réapparait : gain d'un facteur 2)

2.1.5 Factorisation QR par les matrices de Householder

A est multipliée à gauche par des matrices orthogonales particulières, les matrices dites de Householder, dans une stratégie d'élimination.

Définition 2.5

Soit $v \in \mathbb{R}^n \backslash \{0\}$. On pose

$$H(v) = I_n - 2\frac{vv^T}{v^T v}$$

et $H(0) = I_n$.

Proposition 2.6

(i) $H(v) \in \mathcal{S}_n(\mathbb{R}) \cap \mathcal{O}_n(\mathbb{R})$.
(ii) $H(v)$ *est la matrice de la symétrie orthogonale par rapport à* $v^\perp$.
(iii) Soit $e \in \mathbb{R}^n$ *un vecteur unitaire. Alors,*

$$\forall v \in \mathbb{R}^n,\ v \neq \pm \|v\|_2 e, \qquad \begin{cases} H(v + \|v\| e)v = -\|v\| e \\ H(v - \|v\| e)v = \|v\| e. \end{cases}$$

La propriété *(iii)* sera essentielle pour éliminer des coefficients dans A :

$$H(c_1 - \|c_1\| e_1)A = \begin{pmatrix} \|c_1\| & & \\ 0 & & * \\ \vdots & & \\ 0 & & \end{pmatrix}$$

▷ *(i)* $H(v)^T = H(v)$ car $(vv^T)^T = vv^T$.

$$H(v)H(v)^T = H(v)^2 = I_n - \frac{4}{\|v\|_2^2} vv^T + \frac{4}{\|v\|_2^4}(vv^T)(vv^T) = I_n$$

car $(vv^T)(vv^T) = v(v^Tv)v^T = v\|v\|^2 v^T = \|v\|^2 vv^T$.

(ii) Soit $x = \alpha v + w$ avec $w \in v^\perp$ ($\mathbb{R}^n = \mathbb{R}v \oplus v^\perp$). On doit montrer

$$H(v)x = -\alpha v + w.$$

Or

$$H(v)w = w - \frac{2}{\|v\|^2} v \underbrace{v^T w}_{=(v,w)=0} = w$$

$$H(v)v = v - 2v = -v.$$

(iii) On écrit

$$v = \frac{1}{2}(v + \|v\| e) + \frac{1}{2}(v - \|v\| e)$$

avec $(v + \|v\| e, v - \|v\| e) = 0$, donc

$$H(v + \|v\| e)v = -\frac{1}{2}(v + \|v\| e) + \frac{1}{2}(v - \|v\| e) = -\|v\| e.$$

□

Algorithme d'élimination : À l'étape k : supposons la matrice $A^{(k)}$ de la forme

$$\begin{pmatrix} a_{1,1}^{(k+1)} & & & (a_{i,j}^{(k+1)}) \\ & \ddots & & \\ & & a_{k+1,k+1}^{(k+1)} & \\ & & \vdots & \\ (0) & & a_{n,k+1}^{(k+1)} & \end{pmatrix}$$

On note $v^{(k)} = (a_{k+1,k+1}^{(k+1)}, \dots, a_{n,k+1}^{(k+1)})^T \in \mathbb{R}^{n-k}$. Si $v^(k) = (a_{k+1,k+1}, 0 \dots, 0)^T$, on pose $H^{(k)} = I_n$ et $A^{(k+1)} = H^{(k)}A^{(k)}$. Sinon, on pose

$$H^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 & H(v^{(k)} - \left\|v^{(k)}\right\| e^{(k)}) \end{pmatrix}$$

où $e^{(k)} = (1, 0 \dots 0) \in \mathbb{R}^{n-k}$. On pose $A^{(k+1)} = H^{(k)}A^{(k)}$. Dans tous les cas, $A^{(k+1)}$ est de la forme

$$\begin{pmatrix} * & & & (*) \\ & \ddots & & \\ & & \left\|v^{(k)}\right\| & \\ & & 0 & \\ & & \vdots & \\ (0) & & 0 & \end{pmatrix}$$

car $H(v^{(k)} - \left\|v^{(k)}\right\| e^{(k)})v^{(k)} = \left\|v^{(k)}\right\| e^{(k)}$. À la dernière étape, la matrice prend la forme triangulaire supérieure $R = H^{(n-1)} \dots H^{(1)}A = Q^T A$ où $Q \in \mathcal{O}_n(\mathbb{R})$. Alors $A = QR$.

2.2 Méthodes itératives

On se limite ici aux méthodes définies par une récurrence simple : on définit x_{k+1} à partir de $x_k \in \mathbb{R}^n$ avec $\lim\limits_{k \to +\infty} x_k = x$ où x est l'unique solution de $Ax = b$.

2.2.1 Généralités

Définition 2.7

Soit $A \in \mathcal{M}_n(\mathbb{K})$ inversible. On appelle décomposition régulière de A un couple (M, N) de $\mathcal{M}_n(\mathbb{K})^2$ telles que

(i) $M - N = A$,

(ii) M inversible (facile à inverser).
Une méthode itérative est alors associée à cette décomposition : pour $x_0 \in \mathbb{R}^n$, pour $k \geqslant 0$, x_{k+1} est solution de $M_{x+1} = Nx_k + b$.

Remarque : L'unique limite possible d'une telle suite $(x_k)_{k\geqslant 0}$ est x tel que $Ax = b$.

On dira que la méthode itérative converge si pour tout $x_0 \in \mathbb{R}^n$, la suite $(x_k)_{k\geqslant 0}$ tend vers x.

Définition 2.8

On appelle résidu à l'étape k la quantité $r_k = b - Ax_k$ et erreur à l'étape k la quantité $e_k = x_k - x$.

Remarque : $r_k = Ax - Ax_k = -Ae_k$. Donc $\|r_k\| \leqslant \|A\| \, \|e_k\|$ et $\|e_k\| \leqslant \|A^{-1}\| \, \|r_k\|$. Sur le critère d'arrêt d'un algorithme itératif, on ne peut qu'employer $\|r_k\|$: arrêt si $\|r_k\| < 10^{-15}$ mais si $\|A^{-1}\| >> 1$ on ne peut pas conclure sur la petitesse de $\|e_k\|$.

En pratique, on résout à chaque étape $Mx_{k+1} = Nx_k + b$, l'inverse de M n'est jamais explicitement calculé, mais dans la théorie, on analyse la convergence en écrivant $x_{k+1} = M^{-1}Nx_k + M^{-1}b$. La matrice $M^{-1}N$ est appelée matrice des itérations.

Théorème 2.9

La méthode converge si et seulement si $\rho(M^{-1}N) < 1$.

▷ On a

$$e_{k+1} = x_{k+1} - x = (M^{-1}Nx_k + M^{-1}b) - (M^{-1}Nx + M^{-1}b) = M^{-1}Ne_k$$

donc $e_k = (M^{-1}N)^k e_0$. On a vu que $(M^{-1}N)^k \xrightarrow[k\to+\infty]{} 0$ si et seulement si $\rho(M^{-1}N) < 1$. □

Un premier exemple : Méthode de Richardson (ou gradient à pas fixe). On prend $M = \frac{1}{\alpha} I_n$ et $N = \frac{1}{\alpha} I_n - A$ avec $\alpha \neq 0$. On a $M^{-1}N = I_n - \alpha A$ donc les valeurs propres de $M^{-1}N$ sont les $1 - \alpha\lambda_i$ et

$$\rho(M^{-1}N) = \max_{1\leqslant i\leqslant n} |1 - \alpha\lambda_i|$$

Hypothèse ; A est symétrique définie positive. Alors, $\lambda_i \in \mathbb{R}_+^*$.

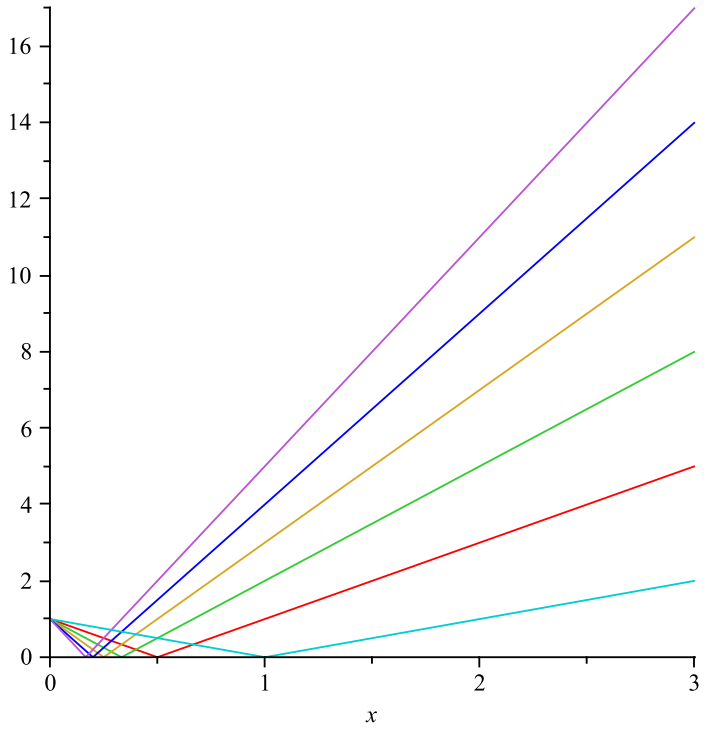

FIGURE 2.1 – $x \mapsto |1 - x\lambda_i|$

donc $\rho(M^{-1}N) \begin{cases} \alpha\lambda_n - 1 & \text{si } \alpha \geqslant \frac{2}{\lambda_1+\lambda_n} \\ 1 - \alpha\lambda_1 & \text{si } 0 \leqslant \alpha \leqslant \frac{2}{\lambda_1+\lambda_n} \\ 1 - \alpha\lambda_n & \text{si } \alpha \leqslant 0 \end{cases}$. En particulier, $\rho(M^{-1}N) < 1$ si et seulement si $\alpha \in]0, \frac{2}{\rho A}[$.

Remarque : $\rho(M^{-1}N)$ est minimal lorsque $\alpha = \frac{2}{\lambda_1+\lambda_n}$. Il vaut

$$1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} = \frac{\mathrm{Cond}_2(A) - 1}{\mathrm{Cond}_2(A) + 1}.$$

Sous certaines hypothèses, il n'est pas nécessaire d'évaluer $\rho(M^{-1}N)$ pour assurer la convergence.

Théorème 2.10

Soient A hermitienne définie positive et $A = M - N$ une décomposition régulière avec M inversible. Alors $M^* + N$ est hermitienne. Si de plus, $M^* + N$ est définie

positive, alors $\rho(M^{-1}N) < 1$.

▷ $-\ (M^* + N)^* = M + N^* = A + N + N^* = A^* + N^* + N = M^* + N$.

– Si $M^* + N$ est définie positive, on construit une norme subordonnée telle que $\|M^{-1}N\| < 1$, car alors, on aura $\rho(M^{-1}N) < 1$. On considère le produit scalaire induit par A :

$$\forall x \in \mathbb{C}^n, \qquad \|x\|_A = \sqrt{(Ax, x)}.$$

Soit $x \in \mathbb{C}^n$ tel que $\|x\|_A = 1$ et $\|M^{-1}Nx\|_A = \|M^{-1}N\|_A$. On calcule $\|M^{-1}Nx\|_A^2$:

$$\begin{aligned}
\|M^{-1}Nx\|_A^2 &= (AM^{-1}Nx, M^{-1}Nx) = (AM^{-1}(M-A)x, M^{-1}(M-A)x) \\
&= (Ax - AM^{-1}Ax, (I - M^{-1}A)x) \\
&= (Ax, x) - (AM^{-1}Ax, x) + (AM^{-1}Ax, M^{-1}Ax) - (Ax, \underbrace{M^{-1}Ax}_{w}) \\
&= 1 - (Aw, x) + (Aw, w) - (\underbrace{Ax}_{Mw}, w) \\
&= 1 - (w, Ax) + (Aw, w) - (Mw, w) \\
&= 1 - ((M^* + N)w, w) \leqslant 1 - \lambda_{\min} \|w\|_2^2
\end{aligned}$$

avec $\lambda_{\min} = \sigma(M^* + N)$. Ainsi $\|M^{-1}N\|_A < 1$. □

Exemple : $A = \begin{pmatrix} -2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$ est définie positive. On choisit

$$M = \begin{pmatrix} 2 & & 0 \\ & \ddots & \\ 0 & & 2 \end{pmatrix} \qquad N = \begin{pmatrix} 0 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 0 \end{pmatrix}$$

$$A = M - N.$$

Alors,

$$M^{-1}N = \begin{pmatrix} 0 & \frac{1}{2} & & 0 \\ \frac{1}{2} & \ddots & \ddots & \\ & \ddots & \ddots & \frac{1}{2} \\ 0 & & \frac{1}{2} & 0 \end{pmatrix}.$$

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

$$M^* + N = \begin{pmatrix} 2 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 2 \end{pmatrix}$$

est symétrique, définie positive car

$$\langle (M^* + N)x, x \rangle = x_1^2 + \sum_{i=1}^{n-1} (x_i + x_{i+1})^2 + x_n^2 \geqslant 0$$

et nul si et seulement si $x = 0$.

Remarque : Lorsque $n \to +\infty$, $\rho(M^{-1}N) \to 1$. La méthode itérative converge de moins en moins vite pour de grands systèmes.

Théorème 2.11 (*stabilité numérique*)

On suppose qu'à chaque étape k de la méthode itérative, le calcul est en fait

$$x_{n+1} = M^{-1}Nx_k + M^{-1}b + \varepsilon_k \qquad \text{avec } \varepsilon_k \in \mathbb{R}^n.$$

On suppose que $\varphi(M^{-1}N) < 1$ et l'existence d'une norme $\|.\|$ sur $\mathbb{R}^n$ et de $\varepsilon > 0$ tel que $\forall k \in \mathbb{N}$, $\|\varepsilon_k\| \leqslant \varepsilon$. Alors il existe une norme $\|.\|_s$ sur $\mathcal{M}_n(\mathbb{R})$ telle que $\|M^{-1}N\|_s < 1$ et

$$\exists c > 0, \qquad \limsup_{k \to +\infty} \|x - x_k\| \leqslant \Lambda\varepsilon$$

avec $\Lambda = \frac{c^2}{1 - \|M^{-1}N\|_s}$.

Remarque : Si $\rho(M^{-1}N) \approx 1$, alors $\Lambda >> 1$ et les itérations $(x_k)_k$ sont éventuellement à grande distance de la solution x.

▷ $e_k = x_k - x$, $e_{k+1} = M^{-1}Ne_k + \varepsilon_k$ $\forall k \geqslant 0$. Par récurrence, on obtient donc

$$\forall k \geqslant 0, \qquad e_k = (M^{-1}N)^k e_0 + \sum_{i=0}^{k-1} (M^{-1}N)\varepsilon_{k-1-i}.$$

On travaille avec la norme $\|.\|_s$ sur $\mathbb{R}^n$ associée à une norme $\|.\|_s$ sur $\mathcal{M}'_n\mathbb{R})$ telle que $\|M^{-1}N\|_s < 1$.

$$\|e_k\|_s \leqslant \left\|M^{-1}N\right\|_s^k \|e_0\|_s + \sum_{i=0}^{k-1} \left\|M^{-1}N\right\|_s^i \|e_{k-1-1}\|_s .$$

Par ailleurs, il existe $c > 0$ tel que $\forall y \in \mathbb{R}^n$, $\|y\|_s \leqslant c\|y\|$e et $c^{-1}\|y\| \leqslant \|y\|_s$ (dimension finie).

$$\|e_k\| \leqslant c\|e_k\|_s \leqslant c\left\|M^{-1}N\right\|_s^k \|e_0\|_s + c\sum_{i=0}^{k-1}\left\|M^{-1}N\right\|_s^i c\varepsilon \leqslant c\left\|M^{-1}N\right\|_s^k \|e_0\|_s + \frac{c^2}{1-\left\|M^{-1}N\right\|_s}\varepsilon.$$

Par passage à la limite sup, on obtient le résultat. □

2.2.2 Méthode de Jacobi

Soit $A = (a_i)_{1\leqslant i,j\leqslant n}$. On note D la diagonale de A. $D \in \mathcal{M}_n(\mathbb{R})$, $D = (a_{i,j}\delta_{i,j})_{1\leqslant i,j\leqslant n}$. $A = M - N$ avec $M = D$ et $N = D - A$. La matrice des itérations de Jacobi est $J = I - D^{-1}A$. Bien sûr, il faut s'assurer que D est inversible. En pratique, cela revient à remplacer le système d'équations $Ax = b$:

$$\forall i \in \{1,\dots,n\}, \qquad x_i = \frac{1}{a_{i,i}}\left(b_i - \sum_{j\neq i} a_{i,j}x_j\right)$$

par les itérations sur $k \geqslant 0$:

$$\forall i \in \{1,\dots,n\}, \qquad x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)}\right)$$

$(x^{(k)})_{k\geqslant 0}$ est une suite récurrente associée au point fixe.

Conditions de convergence : − Si A est hermitienne, la méthode converge. Si de plus A et $2D - A$ sont définies positives ($M^* + N = D = D + D - A = 2D - N$)

$$J_{i,j} = \begin{cases} 0 & \text{si } i = j \\ -\frac{a_{i,j}}{a_{i,i}} & \text{si } i \neq j \end{cases}$$

On remarque que $\|J\|_\infty = \max_i \sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|}$. En particulier, $\|J\|_\infty < 1$ lorsque A est à diagonale dominante strictement : $\forall i,\ \sum_{j\neq i}|a_{i,j}| < |a_{i,i}|$.

2.2.3 Méthode de Gauss-Seidel

$M = D - E$ $N = F$ avec

$$A = \begin{pmatrix} & & -F \\ & D & \\ -E & & \end{pmatrix}$$

$E_{i,j} = -a_{i,j}$ si $j < i$. D est la diagonale de A et $F_{i,j} = -a_{i,j}$ si $i < j$.

Remarque : $-$ M est inversible si et seulement si D est inversible.
$-$ M est facile à inverser.

Conditions de convergence : Si A est hermitienne, définie, positive, alors $M^* + N = \overline{D} - E^* + F = \overline{D} = D$ car $F = E^*$. Or D est définie positive, $\langle Ae_i, e_i \rangle = D_{i,i} > 0$. Donc la méthode converge.
Exprimée sur les composantes de $x^{(k)}$, la méthode s'écrit

$$\forall i, \qquad x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i + \sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{j>i} a_{i,j} x_j^{(k)} \right)$$

(M est triangulaire inférieure donc méthode de descente).

2.2.4 Méthode de relaxation

(SOR : Successive over relaxation)
Soit $\omega \in \mathbb{R}_+^*$. On pose

$$M = \frac{1}{\omega} D - E \qquad \text{et} \qquad N = \frac{1-\omega}{\omega} D + F.$$

On a

$$A = D - E - F = M - N.$$

Pour appliquer la méthode, il faut et il suffit que $D \in GL_n(\mathbb{R})$.
$-$ $\omega = 1$, Gauss-Seidel,
$-$ $\omega < 1$, sous-relaxation,
$-$ $\omega > 1$, sur-relaxation.

Objectif : Identifier les valeurs de ω telles que la méthode converge, et les valeurs pour lesquelles elle converge le plus rapidement. Minimiser

$$\rho(g_\omega), \qquad g_\omega = \left(\frac{1}{\omega} D - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right)$$

Supposons par exemple A hermitienne, définie positive.

$$M^* + N = \frac{D}{\omega} - E^* + \frac{1-\omega}{\omega} D + F = \frac{2-\omega}{\omega} D$$

car $E^* = F$.

En particulier, $M^* + N$ est définie positive si $\omega \in]0, 2[$. Alors la méthode converge.

Théorème 2.12

Si $\omega \in]0, 2[$ et A est hermitienne définie positive, alors SOR converge.

Proposition 2.13

Pour toute matrice A avec D inversible. On a $\forall \omega > 0,\ \rho(g_\omega) \geqslant |1 - \omega|$.

$\triangleright \quad \det(g_\omega) = \dfrac{\det\left(\frac{1-\omega}{\omega} D + F\right)}{\det\left(\frac{1}{\omega} D - E\right)} = (1-\omega)^n \left(= \dfrac{\prod \frac{1-\omega}{\omega} D_{i,i}}{\prod \frac{1}{\omega} D_{i,i}} \right).$

$$\rho(g_\omega)^n \geqslant \left| \prod_{i=1}^{n} \lambda_i(g_\omega) \right| = |\det(g_\omega)| = |1 - \omega|^n$$

donc $\rho(g_\omega) \geqslant |1 - \omega|$. □

Si $\omega \geqslant 2$, $\rho(g_\omega) \geqslant 1$ donc la méthode itérative ne converge pas. $\omega_{\text{opt}} > 1$ en général : sur-relaxation.

2.3 Méthodes variationnelles (optimisation, calcul des variations, calcul différentiel)

Ici, A est symétrique.

2.3.1 Problème de minimisation

$f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$.
On note $\nabla f(x)$ le vecteur gradient (dans $\mathbb{R}^n$).

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

$$f(x + h) = f(x) + \langle \nabla f(x), h\rangle + o(\nabla h).$$

Plus précisément, f étant quadratique,

$$\begin{array}{rcl} f(x+h) & = & \frac{1}{2}\langle A(x+h), x+h\rangle - \langle b, x+h\rangle \\ & = & \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + \frac{1}{2}\langle Ax, h\rangle + \langle Ah, x\rangle - \langle b, h\rangle + \frac{1}{2}\langle Ah, h\rangle \\ & = & f(x) + \langle Ax - b, h\rangle + \frac{1}{2}\langle Ah, h\rangle \end{array}$$

donc $\nabla f(x) = Ax - b$.

f a un minimum global en un point $x \in \mathbb{R}^n$ si et seulement si A est semi-définie positive et $Ax = b$.

Si A est semi-définie positive, $\langle Ah, h\rangle \geqslant 0$ mais $\ker A \neq \{0\}$. Supposons que $b \in \text{Im } A$, f atteint son minimum global sur $x + \ker A$ (il n'y a pas d'autre minimum local).

Si A n'est pas positive ou si $b \notin$ Im , alors $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.

– Si A n'est pas positive, alors il existe $(\lambda, e) \in \mathbb{R}_-^* \times \mathbb{R}^n$ tel que $Ae = \lambda e$ et en posant $x = 0$, $h = \alpha e$ $\alpha > 0$, on trouve

$$f(x + h) = -\alpha\langle b, e\rangle + \frac{1}{2}\alpha^2\lambda \|e\|_2^2 \sim \frac{1}{2}\lambda \|e\|^2 \alpha^2 \xrightarrow[\alpha\to+\infty]{} \infty$$

– Si $b \notin \text{Im } A$ mais $A \geqslant 0$, dans la direction $\ker A$, $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.

▷ Dans le cas A symétrique définie positive,

– si $Ax = b$ alors $f(x + h) = f(x) + \frac{1}{2}\langle Ah, h\rangle > f(x)$ si $h \neq 0$ donc x est un minimum global de f.

– si x minimise f alors nécessairement $Ax = b$. Si $Ax \neq b$, on pose

$$h = -\eta(Ax - b), \qquad \eta > 0.$$

Alors,

$$f(x + h) = f(x) - \eta \|Ax - b\|_2^2 + o(\eta) < f(x)$$

pour $\eta > 0$ suffisamment petit.

– Si $b \in \text{Im } A$ et $\ker A \neq \{0\}$, il existe $y \in \mathbb{R}^n$ tel que $Ay = b$.

$$\forall x \in \mathbb{R}^n,\ Ax = b \iff Ax = Ay \iff x \in y + \ker A.$$

Notons H le supplémentaire orthogonal de $\ker A$ dans $\mathbb{R}^n$: $\mathbb{R}^n = \ker A \oplus H$. Soit $x \in \mathbb{R}^n$, $x = y + h_1 + h_2$ avec $h_1 \in \ker A$ et $h_2 \in H$. Alors,

$$f(x) = f(y) + \langle Ay - b, h_1 + h_2\rangle + \frac{1}{2}\langle A(h_1 + h_2), h_1 + h_2\rangle = f(y) + \frac{1}{2}\langle Ah_2, h_2\rangle.$$

A est symétrique définie positive sur H donc $f(x) > f(y)$ si $h_2 \neq 0$. □

2.3.2 Gradient à pas fixe

Soit $x_0 \in \mathbb{R}^n$. $\forall k \geqslant 0$ $x_{k+1} = x_k - \alpha\nabla f(x_k)$, α est le pas, indépendant de k.

Observations : On a

$$f(x_{k+1}) = f(x_k) - \alpha(\nabla f(x_k), \nabla f(x_k)) + \frac{\alpha^2}{2}(A\nabla f(x_k), \nabla f(x_k)).$$

Si $\alpha > 0$ suffisamment petit, $f(x_{k+1}) = f(x_k) - \alpha \left\|\nabla f(x_k)\right\|^2 + \mathcal{O}(\alpha^2) < f(x_k)$. En fait

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

méthode de Richardson (itérative). On a vu que si $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$ sont les valeurs propres de A, il y a convergence (indépendamment de x_0) si et seulement si $\alpha \in]0, \frac{2}{\lambda_n}[$. De plus, il y a un paramètre optimal $\alpha_{\text{opt}} = \frac{2}{\lambda_1+\lambda_n}$ avec convergence au plus géométrique de raison $\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}$:

$$\exists C > 0,\ \forall k \in \mathbb{N}, \qquad \left\|x_k - x\right\| \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^k C \left\|x_0 - x\right\|.$$

Si $\text{Cond}_2(A) = \frac{\lambda_n}{\lambda_1} >> 1$ alors $\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1} \approx 1$ et la convergence est plus lente.

Remarque : Si $\alpha \in]0, \frac{2}{\lambda_n}[$,

$$f(x_{k+1}) - f(x_k) \leqslant -\alpha \left\|\nabla f(x_k)\right\|_2^2 + \frac{\alpha^2}{2}\lambda_n \left\|\nabla f(x_k)\right\|_2^2 \leqslant \alpha(-1 + \frac{\alpha\lambda_n}{2}) \left\|\nabla f(x_k)\right\|_2^2 < 0.$$

2.3.3 Méthode du gradient à pas optimal

$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, on choisit $\alpha_k \in \mathbb{R}$ tel que $\alpha_k = \arg\min_{\alpha \in \mathbb{R}} g(\alpha)$ où

$$g : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ \alpha & \mapsto & f(x_k) - \alpha(\nabla f(x_k), \nabla f(x_k)) + \frac{\alpha^2}{2}(A\nabla f(x_k), \nabla f(x_k)) \end{array}$$

g est strictement convexe car A est symétrique définie positive et pourvu que $\nabla f(x_k) \neq 0$.

$$g'(\alpha) = -\left\|\nabla f(x_k)\right\|_2^2 + \alpha(A\nabla f(x_k), \nabla f(x_k)) = 0 \iff \alpha = \frac{\left\|\nabla f(x_k)\right\|}{(A\nabla f(x_k), \nabla f(x_k))} = \alpha_k$$

à chaque étape, on minimise f sur la droite affine $x_k + \text{Vect}(\nabla f(x_k))$.

Propriété 2.14

$\forall k \geqslant 0$ on pose $r_k = b - Ax_k = -\nabla f(x_k)$. On a :

$$(r_k, r_{k+1}) = 0$$

▷ $r_{k+1} = b - Ax_{k+1} = b - Ax_k + \alpha_k A(-r_k) = r_k - \alpha_k A r_k$ donc

$$(r_{k+1}, r_k) = \left\|r_k\right\|^2 - \alpha_k(Ar_k, r_k) = 0$$

par définition de α_k. □

2.3.4 Méthodes de Krylov et gradient conjugué

Définition 2.15

Soient $r_0 \in \mathbb{R}^n$ et $k \in \mathbb{N}$. On appelle espace de Krylov d'ordre k associé à r_0 :

$$K_k = Vect(r_0, Ar_0, \dots, A^{k-1}r_0).$$

Proposition 2.16

Soit $r_0 \in \mathbb{R}^n$ fixé. $\forall k \in \mathbb{N}$, $K_k \subset K_{k+1}$. Il existe $k_{max} \in [\![1, n]\!]$ tel que

$$\dim K_k = \begin{cases} k & si\ k \leqslant k_{max} \\ k_{max} & si\ k \geqslant k_{\max} \end{cases}$$

▷ $P = \{k \in \mathbb{N},\ (r_0, Ar_0, \dots, A^{k-1}r_0) \text{ est libre}\}$. $k_{\max} = \max P$. On obtient alors le résultat. □

Proposition 2.17

Considérons une méthode de gradient de la forme

$$\begin{cases} x_0 \in \mathbb{R}^n \\ \forall k \in \mathbb{N},\ x_{k+1} = x_k + \alpha_k r_k,\ r_k = b - A_k \end{cases}$$

Alors, $\forall k \geqslant 0,\ r_k \in K_{k+1}$, associé à $r_0 = b - Ax_0$ et $x_k \in x_0 + K_k$.

▷ $-$ $r_0 \in K_1 = \text{Vect}(r_0)$ et $x_1 = x_0 + \alpha_0 r_0 \in x_0 + K_1$.

$-$ Supposons pour $k \geqslant 0$ fixé que $r_k \in K_{k+1}$ et $x_{k+1} \in x_0 + K_{k+1}$. Alors,

$r_{k+1} = \underbrace{r_k}_{\in K_{k+1} \subset K_{k+2}} - \underbrace{\alpha_k}_{\in AK_{k+1} \subset K_{k+2}} Ar_k \in K_{k+2}$ et $x_{k+2} = x_{k+1} + \alpha_{k+1} r_{k+1} \in x_0 + K_{k+2}$

par hypothèse de récurrence. □

Principe de la méthode du gradient conjugué : À chaque étape $k \geqslant 0$, on minimise la fonctionnelle f sur le sous-espace $x_0 + K_k$.

Admettons que l'on sache traiter ce problème. Alors, on a le résultat suivant.

Théorème 2.18

Si A est symétrique définie positive, la méthode converge en au plus n étapes.

▷ Pour tout $k \in \mathbb{N}$, la restriction de f à $x_0 + K_k$ est une fonction strictement convexe et admet un unique minimum x_k. De plus, la suite $(x_0 + K_k)_{k \geqslant 0}$ étant stationnaire à partir de $k = k_{\max}$ qui dépend de r_0, on a :

$$\forall k \geqslant k_{\max}, \qquad x_k = x_{k_{\max}}.$$

□

⚠L'algorithme du gradient conjugué peut converger en strictement moins de n itérations. Elle converge toujours vers x la solution de $Ax = b$.

Difficulté : Comment construire x_{k+1} effectivement à partir de $x_k, x_{k-1}, \dots, x_0$? On utilise pour cela le produit scalaire associé à la forme bilinéaire symétrique A. On montre que nécessairement, en notant $r_k = b - Ax_k$ (résidu) et $p_k = x_{k+1} - x_k$ (direction de descente),

(i) $K_k = \text{Vect}(r_0, Ar_0, \dots, A^{k-1}r_0) = \text{Vect}(r_0, r_1, \dots, r_{k-1}) = \text{Vect}(p_0, p_1, \dots, p_{k-1})$,

(ii) $(r_k)_{k\geqslant 0}$ forme une famille orthogonale,

(iii) $(p_k)_{k\geqslant 0}$ forme une famille A-conjuguée (ou A-orthogonale) *ie.* $\forall k \neq l,\ (Ap_k, p_l) = 0$.

Pour construire x_{k+1}, on détermine la nouvelle direction de descente p_k de la manière suivante : p_k complète $(p_0, p_1, \dots, p_{k-1})$ en une base A-orthogonale de K_{k+1}. On recherche p_k sous la forme $p_k = r_k - \beta_{k-1}p_{k-1}$ avec

$$0 = (Ap_k, p_{k-1}) = (Ar_k, p_{k-1}) - \beta_{k-1}(Ap_{k-1}, p_{k-1})$$

donc avec

$$\beta_{k-1} = \frac{(Ar_k, p_{k-1})}{(Ap_{k-1}, p_{k-1})}.$$

Algorithme : On se donne $x_0 \in \mathbb{R}^n$ et on pose $r_0 = b - Ax_0$ et $p_0 = r_0$.
À l'étape $k \geqslant 0$: on pose

$$\alpha_k = \frac{(p_k, r_k)}{(Ap_k, p_k)},$$
$$x_{k+1} = x_k + \alpha_k p_k,$$
$$r_{k+1} = r_k - \alpha_k Ar_k,$$
$$\beta_k = \frac{(Ap_k, r_{k+1})}{(Ap_k, p_k)},$$
$$p_{k+1} = r_{k+1} - \beta_k p_k.$$

La minimisation se fait à chaque étape sur $x_k + \text{Vect}(p_k)$ mais elle revient à minimiser sur $x_0 + K_k$ grâce au choix des directions $(p_k)_{k\geqslant 0}$ qui sont A-conjuguées.

Proposition 2.19 (*vitesse de convergence (admise)*)

$$\forall k \geqslant 0, \qquad \|x_k - x\|_2 \leqslant 2\sqrt{\text{Cond}_2(A)}r^k \|x_0 - x\|_2$$

où $r = \frac{\sqrt{\text{Cond}_2(A)}-1}{\sqrt{\text{Cond}_2(A)}+1}$.

2.4 Systèmes surdéterminés et moindre carrés

$A \in \mathcal{M}_{n,p}(\mathbb{R})$, $b \in \mathbb{R}^n$, $n > p$ (plus d'équations que d'inconnues) $Ax = b$. Si $b \notin \text{Im } A$, on définit x comme solution de $Ax = b$ au sens des moindres carrés si x minimise $\|Ax - b\|_2^2$.

2.4.1 Existence et unicité de la solution x

On pose $F = \text{Im } A = \{Ay,\ y \in \mathbb{R}^p\}$ qui est un sous-espace vectoriel de $\mathbb{R}^n$. On note p_F la projection orthogonale sur F pour $(.,.)_{\mathbb{R}^n}$. Alors $z = p_F(b)$ est l'unique élément de F tel que $\|z - b\|_2 = \inf_{y \in \mathbb{R}^p} \|Ay - b\|$. En effet, par le théorème de Pythagore, si $v \in F$

$$\|v - b\|^2 = \|v - z\|^2 + \|z - b\|^2 + 2(v - z, z - b)$$

et, comme $z = p_F(b)$, $\forall w \in F,\ (z - b, w) = 0$ donc $\|v - b\|^2 \geqslant \|z - b\|^2$ avec égalité si et seulement si $v = z$. Puisque $z \in F$, il existe $x \in \mathbb{R}^p$ tel que $Ax = z$. Il y a unicité de x si et seulement si $\ker(A) = \{0\}$.

2.4.2 Équation normale

Lemme 2.20

x est solution du problème aux moindres carré si et seulement si $A^*Ax = A^*b$ ($A * A \in \mathcal{M}_{p,p}(\mathbb{R})$).

$\triangleright$ Soit $y \in \mathbb{R}^p$.

$$\begin{aligned} \|Ay - b\|_{\mathbb{R}^n}^2 &= \|Ay - Ax + Ax - b\|^2 = \|Ay - Ax\|^2 + \|Ax - b\|^2 + 2(Ay - Ax, Ax - b)_{\mathbb{R}^n} \\ &= \|Ay - Ax\|^2 + \|Ax - b\|^2 + 2(y - x, A^*Ax - A^*b)_{\mathbb{R}^p} \end{aligned}$$

– Si $A^*Ax = A^*b$ alors $\|Ay - b\|^2 \geqslant \|Ax - b\|^2$ avec égalité si et seulement si $A(x - y) = 0$ *ie.* $x = y$ (si $\ker A = 0$).

– Réciproquement, soient $z \in \mathbb{R}^p$, $t \in \mathbb{R}$, $y = x + tz$.

$$\|Ay - b\|^2 = \|Ax - b\|^2 + 2t(z, A^*Ax - A^*b) + t^2 \|Az\|^2 .$$

Pour $t > 0$ proche de 0, $2(z, A^*Ax - A^*b) + t\|Az\|^2 \geqslant 0$ et pour $t < 0$ proche de 0, $2(z, A^*Ax - A^*b) - t\|Az\|^2 \leqslant 0$ et à la limite, on obtient $(z, A^*Ax - A^*b) = 0$ ceci pour tout $z \in \mathbb{R}^p$ donc x est solution de $A^*Ax = A^*b$. $\square$

A^*A est hermitienne, positive et définie si $\ker A = \{0\}$. Notamment,

$$\ker(A^*A) = \ker(A).$$

2.4.3 Résoudre l'équation normale par factorisation QR

Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$. Par multiplication à gauche par des matrices de Householder, on construit $Q \in \mathcal{O}_n(\mathbb{R}), R \in \mathcal{M}_{n,p}(\mathbb{R})$ telles que $A = QR$ avec $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ avec $R_1 \in \mathcal{M}_p(\mathbb{R})$ triangulaire supérieure.

Soit $x \in \mathbb{R}^p$,

$$\|Ax - b\|_{\mathbb{R}^n} = \|QRx - b\|_{\mathbb{R}^n} = \left\|Rx - Q^T b\right\|_{\mathbb{R}^n}.$$

Supposons $\ker A = \{0\} \Leftrightarrow R_1 \in GL_d(\mathbb{R})$. On écrit

$$Q = (Q_1|Q_2)$$

avec $Q_1 \in \mathcal{M}_{n,p}(\mathbb{R})$ de sorte que

$$\|Ax - b\|^2_{\mathbb{R}^n} = \left\|R_1 - x - Q_1^T b\right\|_{\mathbb{R}^p} + \left\|0 - Q_2^T b\right\|^2_{\mathbb{R}^{n-p}} = \left\|R_1 x - Q_1^T b\right\|^2_{\mathbb{R}^p} + \left\|Q_2^T b\right\|^2_{\mathbb{R}^{n-p}}.$$

La quantité est minimisée lorsque $x = R_1^{-1} Q_1^T b$. La distance minimale est $\left\|Q_2^T b\right\|^2_{\mathbb{R}^{n-p}}$.

Si $\ker A \neq \{0\}$, on écrit

$$A = (Q_1|Q_2) \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix}$$

avec $Q_1 \in \mathcal{M}_{n,r}(\mathbb{R})$, $Q_2 \in \mathcal{M}_{n,n-r}(\mathbb{R})$, $R_1 \in GL_r(\mathbb{R})$, $R_2 \in \mathcal{M}_{r,p-r}(\mathbb{R})$ où $r = \mathrm{rg}(A)$. Alors, pour tout $x \in \mathbb{R}^p$,

$$\|Ax - b\|^2_{\mathbb{R}^n} = \left\|(R_1|R_2)x - Q_1^T b\right\|^2_{\mathbb{R}^r} + \left\|Q_2^T b\right\|^2_{\mathbb{R}^{n-r}}.$$

La quantité est minimale lorsque

$$(R_1|R_2)x = Q_1^T b,$$

obtenu lorsque $x = R_1^{-1}(Q_1^T b - R_2 \tilde{x})$, $\forall \tilde{x} \in \mathbb{R}^{p-r}$. Le minimum est toujours $\left\|Q_2^T b\right\|^2_{\mathbb{R}^{n-r}}$.

Chapitre 3

Approximation spectrale

Problème : Trouver les éléments propres d'une matrice. Ce n'est pas un problème linéaire, c'est plutôt la recherche des racines d'un polynôme.

3.1 Localisation des valeurs propres

Remarque : $\forall \lambda \in \mathrm{Sp}(A),\ |\lambda| \leqslant \rho(A) \leqslant |||A|||$ et ce pour tout norme subordonnée.

Théorème 3.1 (*Premier théorème de Gershgörin*)

Soient $A \in \mathcal{M}_n(\mathbb{C})$ et $\lambda \in Sp(A)$. Alors

$$\lambda \in \bigcup_{i=1}^{n} D_i \qquad \text{où } D_i = \left\{ z \in \mathbb{C},\ |z - a_{i,i}| \leqslant \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \right\}.$$

D_i est appelé disque (ligne) de Gershgörin.

▷ Soit $x \in \mathbb{C}^n \backslash \{0\}$ un vecteur propre associé à λ. $Ax = \lambda x$ si et seulement si

$$\forall i \in [\![1, n]\!], \qquad \lambda x_i - a_{i,i} x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} x_j$$

donc

$$|\lambda - a_{i,i}||x_i| \leqslant \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}||x_j| \leqslant \|x\|_\infty \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|.$$

Pour $i_0 \in [\![1, n]\!]$ tel que $|x_{i_0}| = \|x\|_\infty \neq 0$, on obtient

$$|\lambda - a_{i_0,i_0}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}|$$

donc $\lambda \in D_{i_0} \subset \bigcup_{i=1}^{n} D_i$. □

Par ailleurs, $\mathrm{Sp}(A) = \mathrm{Sp}(A^T)$ donc

$$\mathrm{Sp}(A) \subset \bigcup_{j=1}^{n} C_j \qquad \text{où } C_j = \left\{ z \in \mathbb{C},\ |z - a_{j,j}| \leqslant \sum_{i\neq j} |a_{i,j}| \right\}.$$

C_j est appelé disque-colonne de Gershgörin.

Finalement,

$$\mathrm{Sp}(A) \subset \left(\bigcup_{i=1}^{n} D_i \right) \cap \left(\bigcup_{j=1}^{n} C_j \right).$$

Remarque : − Certains disque peuvent finalement ne contenir aucune valeur propre.

Exemple : $A = \begin{pmatrix} 1 & -1 \\ 2 & -1 \end{pmatrix}$, $\mathrm{Sp}(A) = \{i, -i\}$ et D_1 ne contient pas de valeurs propres.

$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $\mathrm{Sp}(A) = \{i, -i\}$, $D_1 = D_2$ est les valeurs propres sont au bord des disques.

On a la conséquence importante suivante :

Lemme 3.2 (*Hadamard*)

Soit $A \in \mathcal{M}_n(\mathbb{C})$ à diagonale strictement dominante :

$$\forall i \in [\![1, n]\!], \qquad |a_{i,j}| > \sum_{j\neq i} |a_{i,j}|$$

alors A est inversible.

▷ $\forall i,\ 0 \notin D_i$. □

Théorème 3.3 (*Deuxième théorème de Gershgörin*)

Soit $m \in [\![1, n]\!]$ tel qu'on puisse trouver m disques de Gershgörin dont la réunion est disjointe des $n - m$ disques restants. Alors cette réunion de m

disques contient exactement m valeurs propres (avec multiplicité) et les autres en contiennent $n - m$.

Corollaire 3.4

– Si un disque est isolé, il contient exactement une valeur propre, qui est simple.
– Si les disques sont deux à deux disjoints alors chacun contient exactement une valeur propre simple.

▷ Posons $S_1 = \bigcup_{i=1}^{m} D_i$, $S_2 = \bigcup_{i=m+1}^{n} D_i$, $S_1 \cap S_2 = \varnothing$. Pour $t \in [0,1]$, on définit $B(t) = (b_{i,j}(t))_{1 \leqslant i,j \leqslant n}$ avec

$$b_{i,j}(t) = \begin{cases} a_{i,i} & \text{si } i = j \\ t a_{i,j} & \text{sinon} \end{cases}$$

$B(0) = \operatorname{diag}(a_{i,i})$, $B(1) = A$. $\sigma(B(0)) = \{a_{i,i},\ i \in \{1, \dots, n\}\}$, $\sigma(B(1)) = \sigma(A)$. Pour chaque $t \in [0,1]$, $\sigma(B(t))$ est l'ensemble des racines d'un polynôme dont les coefficients dépendent continûment de t. On admet le résultat suivant : par un principe de sélection continue des racines, il existe n fonctions continues sur $[0,1]$ à valeurs dans $\mathbb{C}$, $\lambda_1, \dots, \lambda_n$ tel que pour tout t,

$$\sigma(B(t)) = \{\lambda_1(t), \dots, \lambda_n(t)\}.$$

En particulier, $\lambda_i(0) = a_{i,i}$ pour tout i (quitte à renuméroter) et $\lambda_i(1) = \alpha_{\sigma(i)}$ où $\sigma(A) = \{\alpha_1, \dots, \alpha_n\}$ et σ est une permutation de $\{1, \dots, n\}$. En appliquant le premier théorème,

$$\forall t \in [0,1], \qquad \sigma(B(t)) \subset \bigcup_{i=1}^{n} D_i(t)$$

avec

$$D_i(t) = \{z \in \mathbb{C},\ |z - a_{i,i}| \leqslant t \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|.$$

$t \mapsto D_i(t)$ est croissante au sens de l'inclusion et

$$\bigcup_{t=0}^{1} D_i(t) = D_i.$$

De plus,

$$\forall t \in [0,1], \qquad \left(\bigcup_{i=1}^{m} D_i(t) \right) \cap \left(\bigcup_{i=m+1}^{t} D_i(t) \right) = \varnothing.$$

Par continuité des fonctions λ_i, on trouve à chaque instante $t \in [0,1]$, un ensemble $\{\lambda_1(t), \dots, \lambda_m(t)\}$ dans S_1 et $\{\lambda_{m+1}, \dots, \lambda_n\}$ dans S_2. □

3.2 Méthode de la puissance

Principe : Étant donné $x_0 \in \mathbb{C}^n$, on calcule itérativement

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|}, \qquad k \geqslant 0.$$

$(x_k)_{k\geqslant 0}$ se comporte comme sa composante le long du vecteur propre pour la valeur propre de module maximal.

Théorème 3.5

Supposons que $A \in \mathcal{M}_n(\mathbb{C})$ admet une valeur propre dominante ie.

$$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_p|$$

avec éventuellement des multiplicités ($\sigma(A) = \{\lambda_1, \ldots, \lambda_p\}$).

En plus de la suite (x_k) on définit $\nu_k = \langle x_k, Ax_k\rangle$, $k \geqslant 0$. Alors, pour presque tout $x_0 \in \mathbb{C}^n$,

$$\lim_{k\to+\infty} \nu_k = \lambda_1$$

et

$$\lim_{k\to+\infty} \left(\frac{|\lambda_1|}{\lambda_1}\right)^k x_k = e \in \ker(A - \lambda_1 I_n) \qquad (\textit{vecteur propre}).$$

▷ Supposons ici A diagonalisable et λ_1 simple. Soient $e_1, \ldots, e_n$ une base de vecteurs propres tels que

$$\forall i \in \{1, \ldots, n\}, \qquad Ae_i = \lambda_i, e_i$$

Alors, en décomposant $x_0 = \sum_{i=1}^{n} \alpha_i e_i$ on a, pour tout k,

$$A^k x_0 = \sum_{i=1}^{n} \alpha_i A^k e_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k e_i = \lambda_1^k \left(\alpha_1 e_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k e_i\right).$$

Par ailleurs,

$$x_k = \frac{A^k x_0}{\|A_k x_0\|} = \frac{\lambda_1 \left(x_1 e_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)}{\left\|\lambda_1^k \left(\alpha_1 e_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)\right\|}.$$

Si $\alpha_1 \neq 0$, alors

$$x_k \underset{k\to+\infty}{\sim} \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \underbrace{\frac{\alpha_1}{|\alpha_1|}\frac{e_1}{\|e_1\|}}_{\in \ker(A-\lambda_1 I_n)}.$$

Alors,

$$\nu_k = \langle x_k, Ax_k \rangle \sim \left\langle \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\alpha_1}{|\alpha_1|}\frac{e_1}{\|e_1\|}, \lambda_1 \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\alpha_1}{|\alpha_1|}\frac{e_1}{\|e_1\|} \right\rangle \sim \lambda_1.$$

Si A n'est pas diagonalisable, il faut considérer des blocs de Jordan et leurs puissances successives :

$$J_p = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \qquad J_p^k = \begin{pmatrix} \lambda & k\lambda_{k-1} & & \binom{k}{p-1} \\ & \ddots & \ddots & \\ & & \ddots & k\lambda^{k-1} \\ 0 & & & \lambda^k \end{pmatrix}$$

□

Concernant la vitesse de convergence, on pose $q_k = \left(\frac{|\lambda_1|}{\lambda_1}\right)^k x_k$, de limite e.

$$\|q_k - e\| = \begin{cases} \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) & \text{si } \lambda_1 \text{ est non défective ainsi que toutes les valeurs propres de module } |\lambda_2| \\ \mathcal{O}\left(k^{r-1}\left(\frac{\lambda_2}{\lambda_1}\right)^k\right) & \text{si } \lambda_1 \text{ est non défective et } r \text{ désigne la taille du plus grand bloc de Jordan associé à une valeur propre de module } |\lambda_2| \\ \mathcal{O}\left(\frac{1}{k}\right) & \text{si } \lambda_1 \text{ est défective.} \end{cases}$$

Remarque : – Si λ_1 n'est pas la seule valeur propre de module $\rho(A)$, il n'y a pas convergence en général. Plus sournois, (ν_k) peut converger vers un complexe qui n'est pas dans $\sigma(A)$.

– Si $\alpha_1 = 0$, la théorie ne prévoit pas la convergence de (ν_k) ou de (q_k) mais en pratique, les erreurs d'arrondis vont la récupérer !

$$x_0 = \alpha_1 e_1 + \sum_{i \geqslant 2} \alpha_i e_i \qquad \text{avec } |\alpha_i| = \varepsilon > 0.$$

Méthode de la puissance inverse.

Supposons A inversible, on obtient la valeur propre de module minimal en appliquant la méthode de la puissance à A^{-1}.

Algorithme : $x_0 \in \mathbb{C}^n$,

$$\forall k \geqslant 0, \qquad Ay_k = x_k \qquad x_{k+1} = \frac{y_k}{\|y_k\|}.$$

En pratique on peut déterminer une factorisation $A = PLU$ Chaque itération ne coûte pas plus cher que le produit Ax.

Méthode de la puissance inverse avec translation.

Soit $\mu \in \mathbb{C}$, avec $\mu \notin \sigma(A)$. On applique l'algorithme précédent à $A - \mu I_n$ de valeur propre $\{\lambda_i - \mu,\ i \in \{1, \ldots, n\}\}$. On peut faire ensuite circuler μ dans $D(0, \|A\|_\infty)$. Elle converge vers $\lambda_i - \mu$ tel que $|\lambda_i - \mu|$ est minimal.

3.3 Méthode QR (utilisée dans Scilab/Matlab) (1961)

Algorithme : Soit $A \in \mathcal{M}_n(\mathbb{R})$ quelconque. $A_0 = A$. Pour $k \geqslant 0$, on calcule la décomposition QR de $A_k = Q_k R_k$. On pose $A_{k+1} = R_k Q_k$. On remarque que $A_{k+1} = Q_k^T A_k Q_k$ donc A_k a le même spectre que A.

Si (A_k) converge, alors à la limite $Q_k R_k - R_k Q_k \to 0$ et la limite est en fait triangulaire supérieure.

Théorème 3.6

Si A est diagonalisable inversible de valeurs propres $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$, alors $\lim_{k \to +\infty} A_k = T$ existe avec T triangulaire supérieure.

$$\forall i \in \{1, \ldots, n\}, \qquad T_{i,i} = \lambda_i \qquad \textit{(les valeurs propres sont ordonnées).}$$

La convergence est au plus géométrique, de raison

$$\max_{i \geqslant 2} \left| \frac{\lambda_i}{\lambda_{i+1}} \right|.$$

Variante avec translation : Soit $\mu \in \mathbb{C}$ tel que (en renumérotant les λ_i)

$$|\lambda_1 - \mu| > \cdots > |\lambda_n - \mu|$$

et réduisant la quantité

$$\max_{i \geqslant 2} \left| \frac{\lambda_i - \mu}{\lambda_{i-1} - \mu} \right|.$$

Cela accélère la convergence de l'algorithme QR appliqué à $A - \mu I_n$. $Q_k R_k = A_k - \mu I_n$. On pose $A_{k+1} = R_k Q_k + \mu I_n$.

Référence : Quateroni-Sacco-Saleri.

Remarque : Si A est symétrique, à chaque étape de la méthode QR, la matrice A_k est symétrique car $A_{k+1} = Q_k^T A_k Q_k$ et en cas de convergence, la limite est diagonale.

3.4 Méthode de Jacobi

Soit A une matrice symétrique réelle.

Principe : Effectuer des changements de bases orthodromies de sorte à converger vers la diagonalisation en base orthonormée de A (par les rotations de Givens).

3.4.1 Exemple en dimension 2

Soient

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \qquad P = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

Alors,

$$PAP^T = \begin{pmatrix} \alpha' & \beta' \\ \beta' & \gamma' \end{pmatrix}$$

et

$$\beta' = \beta(\cos^2\theta - \sin^2\theta) - (\alpha - \gamma)\sin\theta\cos\theta = \beta\cos 2\theta - \frac{\alpha - \gamma}{2}\sin 2\theta.$$

Si on choisit $\operatorname{cotan}(2\theta) = \frac{\alpha-\gamma}{2\beta}$ alors $\beta' = 0$ et $\sigma(A) = \{\alpha', \gamma'\}$.

Remarque : On

$$\|A\|_F^2 = \operatorname{Tr}(A^T A) = \alpha^2 + \gamma^2 + 2\beta^2 = \operatorname{Tr}((P^T AP)^T (P^T AP)) = \left\|P^T AP\right\|_F^2 = \alpha'^2 + \gamma'^2.$$

En particulier, $\alpha'^2 + \gamma'^2 \geqslant \alpha^2 + \gamma^2$.

En dimension supérieure, de la même manière, on annule successivement des termes non-diagonaux, on perd certains des termes précédemment annulés mais la somme des carrés des termes diagonaux augmente avec $\sum_{i \neq j} |a_{i,j}^{(k)}| \xrightarrow[k \to +\infty]{} 0$.

3.4.2 Cas général

On définit une suite $(A^{(k)})_{k\geqslant 0}$ de la manière suivante. $A^{(0)} = A$. Pour $k \geqslant 0$, on choisit $(p,q) \in \{1,\dots,n\}^2$, $p \neq q$, tel que

$$|a_{p,q}^{(k)}| = \max_{i\neq j} |a_{i,j}^{(k)}|.$$

On pose θ solution de

$$\operatorname{cotan}(2\theta) = \frac{a_{p,p}^{(k)} - a_{q,q}^{(k)}}{2a_{p,q}^{(k)}}$$

et $P^{(k)} = (p_{i,j}^{(k)})$ avec

$$p_{i,j} = \begin{cases} 1 & \text{si } i = j \notin \{p,q\} \\ \cos\theta & \text{si } i = j \in \{p,q\} \\ \sin\theta & \text{si } (i,j) = (p,q) \\ -\sin\theta & \text{si } (i,j) = (q,p) \\ 0 & \text{sinon.} \end{cases}$$

Alors,

$$A^{(k+1)} = P^{(k)} A^{(k)} {P^{(k)}}^T$$

est telle que $a_{p,q}^{(k+1)} = 0$. De plus

$$\sum_i \left(a_{i,i}^{(k+1)}\right)^2 \geqslant \sum_i \left(a_{i,i}^{(k)}\right)^2.$$

On peut prouver (admis ici) que

$$\sum_{i\neq j} \left(a_{i,j}^{(k)}\right)^2 \xrightarrow[k\to+\infty]{} 0$$

de sorte que $A^{(k)} \xrightarrow[k\to+\infty]{} \operatorname{diag}(\lambda_{\sigma(i)})$ où $\sigma \in \mathfrak{S}_n$ et $\lambda_i \in \sigma(A)$.

3.5 Méthode de Givens-Housholder

Principe : A est symétrique réelle. On en détermine une réduction de la forme

$$\tilde{A} = O^T A O = \begin{pmatrix} b_1 & c_1 & & 0 \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & c_{n-1} & b_n \end{pmatrix}$$

par des transformations de Householder. On détermine les valeurs propres de $\tilde{A}$ par la bissection de Givens : on construit une suite de Sturm de polynômes qui sont

$$p_0(x) = 1 \qquad p_1(x) = b_1 - x \qquad p_i(x) = (b_i - x)p_{i-1}(x) - c_{i-1}^2 p_{i-2}(x).$$

On trouve que p_i a i racines qui séparent les $(i+1)$ racines de p_{i+1}.

Chapitre 4

Résolution de systèmes d'équations non-linéaires

4.1 Introduction

4.1.1 Position du problème

On considère une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ continue. On cherche $x^* \in \mathbb{R}^n$ tel que $f(x^*) = 0$.

En pratique, x^* sera approché par une suite $(x^{(k)})$ d'éléments de $\mathbb{R}^n$ définie par une récurrence d'ordre 1 :

$$x_{k+1} = \phi(x^{(k)}), \qquad k \geqslant 0$$

où ϕ reste à définir, mais admet x^* pour point fixe.

Définition 4.1

On dira que $(x^{(k)})_{k\in\mathbb{N}}$ converge vers x^* à l'ordre p si $\lim\limits_{k\to+\infty} x^{(k)} = x^*$ et s'il existe $C > 0$ et $k_0 \in \mathbb{N}$ tels que

$$\forall k \geqslant k_0, \qquad \left\|x^{(k+1)} - x^*\right\| \leqslant C \left\|x^{(k)} - x^*\right\|^p .$$

On appelle taux de convergence la plus petite constante C accessible.

Remarque : Dans le cas non-linéaire, les résultats de convergence sont en général seulement locaux :

$$\exists \delta > 0,\ \forall x^{(0)} \in B(x^*, \delta), \qquad x^{(k)} \xrightarrow[k\to+\infty]{} x^*.$$

4.1.2 Théorème du point fixe

Théorème 4.2

Soit (E, d) un espace métrique complet. Soit $\phi : E \to E$ contractante (ie. L-lipschitzienne, avec $L < 1$). Alors ϕ admet un unique point fixe $x^* \in E$ et pour tout $x^{(0)} \in E$, la suite $(x^{(k)})_{k\geqslant 0}$ définie par $x^{(k+1)} = \phi(x^{(k)})$ converge vers x^*.

▷ − Unicité : Soient $x^* = \phi(x^*)$ et $y^* = \phi(y^*)$. Alors,

$$d(x^*, y^*) = d(\phi(x^*), \phi(y^*)) \leqslant L d(x^*, y^*)$$

donc $d(x^*, y^*) = 0$ et $x^* = y^*$.

− Existence : Soit $x^{(0)} \in E$. On montre que $(x^{(k)})_{k\geqslant 0}$ est de Cauchy dans (E, d), donc elle converge vers $x^* \in E$ avec, par continuité de ϕ, $x^* = \phi(x^*)$. En effet,

$$d(x^{(k+1)}, x^{(k)}) \leqslant L d(x^{(k)}, x^{(k-1)}) \leqslant \cdots \leqslant L^k d(x^{(1)}, x^{(0)})$$

donc

$$d(x^{(k+p)}, x^{(k)}) \leqslant \sum_{i=0}^{p-2} d(x^{(k+i+i)}, x^{(k+i)}) \leqslant L^k \frac{1 - L^{p-1}}{1 - L} d(x^{(1)}, x^{(0)}) \xrightarrow[k\to+\infty]{} 0.$$

□

Remarque : Pour résoudre $f(x^*) = 0$, on cherche $\phi : \mathbb{R}^n \to \mathbb{R}^n$ continue avec $\phi(x^*) = x^*$, on cherche $E \subset \mathbb{R}^n$ fermé avec :
− $x^* \in E$,
− $\phi(E) \subset E$,
− ϕ contractante sur E.
Alors, les itérées $(x^{(k)})_{k\geqslant 0}$ convergent vers x^* pour tout $x^{(0)}$ dans E.

4.1.3 Conditions suffisantes de convergence en dimension 1

Théorème 4.3

Soit $\phi \in \mathcal{C}^1([a, b])$ telle que
− $\phi([a, b]) \subset [a, b]$,
− $\exists L < 1,\ \forall x \in [a, b],\ |\phi'(x)| \leqslant L$.
Alors, il existe un unique $x^* \in]a, b[$ avec $\phi(x^*) = x^*$ et

$$\forall x^{(0)} \in [a, b], \qquad \lim_{k\to+\infty} x^{(k)} = x^*.$$

De plus,

$$\lim_{k\to+\infty} \frac{x^{(k+1)}-x^*}{x^{(k)}-x^*} = \phi'(x^*).$$

▷ On montre que ϕ est L-lipschitzienne (accroissements finis) et ensuite

$$\frac{x^{(k+1)}-x^*}{x^{(k)}-x^*} = \frac{\phi(x^{(k)})-\phi(x^*)}{x^{(k)}-x^*} \xrightarrow[k\to+\infty]{} \phi'(x^*).$$

□

Corollaire 4.4 (*Convergence locale*)

Soit ϕ une application continue sur $[a,b]$ de classe $\mathcal{C}^1$ sur un voisinage de $x^* \in]a,b[$ avec $\phi(x^*) = x^*$ et $|\phi'(x^*)| < 1$. Alors,

$$\exists \delta > 0,\ \forall x^{(0)} \in [x^*-\delta, x^*+\delta] \subset]a,b[, \qquad \lim_{k\to+\infty} x^{(k)} = x^*.$$

▷ On construit $\delta > 0$ de sorte que :
- $[x^*-\delta, x^*+\delta] \subset]a,b[$,
- $\exists L < 1,\ \forall x \in [x^*-\delta, x^*+\delta],\ |\phi'(x)| \leqslant L$,

Alors on a bien $\phi([x^*-\delta, x^*+\delta]) \subset [x^*-\delta, x^*+\delta]$ car

$$|\phi(y)-x^*| = |\phi(y)-\phi(x^*)| \leqslant L(y-x^*)$$

donc

$$|y-x^*| \leqslant \delta \qquad \Rightarrow \qquad |\phi(y)-x^*| \leqslant \delta.$$

On applique donc le théorème précédent. □

Remarque : Si $|\phi'(x^*)| < 1$, x^* est un point fixe attractif. Si $|\phi'(x^*)| > 1$, x^* est un point fixe répulsif (localement, les itérées "s'éloignent" de x^*). Si $|\phi'(x^*)| = 1$, c'est un cas pour lequel l'analyse précédente ne peut pas permettre de conclure : tout peut arriver.

Proposition 4.5

Soient $\phi \in \mathcal{C}^p([a,b])$, $x^* \in]a,b[$ un point fixe de ϕ. On suppose

$$\begin{cases} \phi'(x^*) = \phi^{(2)}(x^*) = \cdots = \phi^{(p-1)}(x^*) = 0 \\ \phi^{(p)}(x^*) \neq 0. \end{cases}$$

Alors il y a convergence locale au voisinage de x^*. Cette convergence est d'ordre p et de taux $\frac{\phi^{(p)}(x^*)}{p!}$.

▷ On a

$$x^{(k+1)} - x^* = \phi(x^{(k)}) - \phi(x^*) = \phi(x^* + (x^{(k)} - x^*)) - \phi(x^*) = \frac{1}{p!}(x^{(k)} - x^*)^p \phi^{(p)}(\eta^{(k)})$$

avec $\eta_k \in [x^*, x^{(k)}]$. Lorsque k tend vers $+\infty$, $x^{(k)}$ tend vers x^* et $\eta^{(k)}$ tend vers x^* donc

$$x^{(k+1)} - x^* \sim \frac{\phi^{(p)}(x^*)}{p!}(x^{(k)} - x^*)^p.$$

□

4.2 Méthodes usuelles en dimension 1

La méthode la plus élémentaire rencontrée est la dichotomie. On considère $f : [a, b] \to \mathbb{R}$ continue avec $f(a)f(b) < 0$. Par le théorème des valeurs intermédiaires, on construit deux suites $(a_k)_{k\geqslant 0}$, $(b_k)_{k\geqslant 0}$ adjacentes qui convergent vers x^* tel que $f(x^*) = 0$.

4.2.1 Méthode de la corde

Soit $f \in \mathcal{C}^1([a, b])$ telle que $f(a) \neq f(b)$. $f(x^*) = 0$. On définit

$$x^{(k+1)} = x^{(k)} - \frac{b-a}{f(b) - f(a)} f(x^{(k)}).$$

petit dessin des premières itérations. On construit la droite parallèle à la corde passant par les points $(a, f(a))$ et $(b, f(b))$ qui passe par $(x^{(k)}, f(x^{(k)}))$ et $x^{(k+1)}$ est l'abscisse du point d'intersection entre cette droite et l'axe des abscisses.

$$\phi(x) = x - \frac{b-a}{f(b) - f(a)} f(x)$$

donc

$$\phi'(x^*) = 1 - \frac{b-a}{f(b) - f(a)} f'(x^*).$$

Condition suffisante de convergence locale : $|\phi'(x^*)| < 1$ *ie.*

$$\frac{b-a}{f(b) - f(a)} f'(x^*) \in]0, 2[.$$

On peut ajuster a et b de sorte que cette condition soit réalisée. Alors, il y a convergence locale d'ordre 1 en général (convergence (au plus) géométrique).

4.2.2 Méthode de Newton

Soit $f \in \mathcal{C}^2([a,b])$. $f(x^*) = 0$. On suppose que $f'(x^*) \neq 0$. On définit

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}.$$

petit dessin À l'itération k, on cherche l'abscisse du point d'intersection de la tangente à f en $x^{(k)}$ et l'axe des abscisses.

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

est définie au voisinage de x^* (car $f'(x^*) \neq 0$). On a $\phi(x^*) = x^*$ et

$$\phi'(x^*) = 1 - \frac{f'(x^*)^2 - f(x^*)f''(x^*)}{f'(x^*)^2} = 0.$$

La méthode de Newton converge localement vers x^*, à l'ordre 2 en général (convergence quadratique). En pratique

$$|x^{(k)} - x^*| \leqslant Cr^{2^k} \qquad \text{avec } 0 < r < 1$$

(tend vers 0 beaucoup plus vite que r^k)

4.3 Critère d'attractivité en dimension n

Soient Ω un ouvert de $\mathbb{R}^n$ et $\phi \in \mathcal{C}^1(\Omega, \mathbb{R}^n)$, $\mathrm{d}\phi(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ la différentielle de ϕ en $x \in \Omega$.

Lemme 4.6

(i) Si ϕ est L-lipschitzienne sur Ω relativement à une norme $\|.\|$ alors

$$\forall x \in \Omega, \qquad \|\!|d\phi(x)|\!\| \leqslant L.$$

(ii) Si Ω est convexe et si $\forall x \in \Omega$, $\|\!|d\phi(x)|\!\| \leqslant L$ alors ϕ est L-lipschitzienne sur Ω pour $\|.\|$.

▷ (i) Par définition de la différentielle,

$$\forall \delta > 0, \exists r > 0, \qquad \|h\| \leqslant r \quad \Rightarrow \quad \|\phi(x+h) - \phi(x) - \mathrm{d}\phi(x)h\| \leqslant \delta \|h\|$$

donc

$$\|\mathrm{d}\phi(x)h\| \leqslant \|\phi(x+h) - \phi(x)\| + \delta \|h\| \leqslant (L + \delta) \|h\|$$

donc $|||\mathrm{d}\phi(x)||| \leqslant L + \delta$, ceci pour tout $\delta > 0$, d'où le résultat.

(ii) Si Ω est convexe

$$\forall x, y \in \Omega, \qquad \phi(y) - \phi(x) = \int_0^1 \mathrm{d}\phi(x + t(y-x))(y-x)\mathrm{d}t.$$

Donc, pour $x, y \in \Omega$,

$$\begin{aligned} \|\phi(y) - \phi(x)\| &\leqslant \int_0^1 \|\mathrm{d}\phi(x + t(y-x))(y-x)\| \,\mathrm{d}t \\ &\leqslant \int_0^1 |||\mathrm{d}\varphi(x + t(y-x))||| \, \|y - x\| \,\mathrm{d}t \\ &\leqslant L \|y - x\| . \end{aligned}$$

□

Théorème 4.7

Soit $\phi \in \mathcal{C}^1(\Omega, \mathbb{R}^n)$ et soit $x^* \in \Omega$ un point fixe de ϕ. Les assertions suivantes sont équivalentes :

(i) Il existe un voisinage fermé V de x^* dans Ω tel que $\phi(V) \subset V$ et une norme $\|.\|$ sur $\mathbb{R}^n$ tels que $\phi_{|V}$ soit contractante pour $\|.\|$.

(ii) $\rho(d\phi(x^*)) < 1$.

Sous ces conditions, toute suite $(x^{(k)})_{k\geqslant 0}$ initialisée en $x^{(0)} \in V$ converge vers x^* : il y a convergence locale.

▷ $(i) \Rightarrow (ii)$: $\phi_{|V}$ est contractante pour $\|.\|$ de rapport $L < 1$. Alors,

$$\rho(\mathrm{d}\phi(x^*)) \leqslant |||\mathrm{d}\phi(x^*)||| \leqslant L$$

par le lemme.

$(ii) \Rightarrow (i)$: $\rho(\mathrm{d}\phi(x^*)) < 1$ donc il existe une norme $\|.\|$ sur $\mathbb{R}^n$ telle que

$$|||\mathrm{d}\phi(x^*)||| \leqslant L < 1.$$

Par continuité de $\mathrm{d}\phi$ en x^*, il existe un voisinage $V = \overline{B(x^*, r)}$ tel que $V \subset \Omega$ et

$$\forall x \in V, \qquad |||\mathrm{d}\phi(x)||| \leqslant \tilde{L} < 1.$$

Par le lemme (il faudrait peut-être prendre un ouvert un peu plus gros et appliquer le lemme ?), V étant convexe, ϕ est contractante sur V de rapport $\tilde{L}$. Par ailleurs $\phi(V) \subset V$ car

$$\forall x \in V, \qquad \|\phi(x) - \phi(x^*)\| \leqslant \tilde{L} \|x - x^*\| \leqslant \tilde{L} r < r$$

donc $\phi(x) \in V$ (car $\phi(x^*) = x^*$). □

Remarque : L'analyse de convergence des méthodes itératives pour "$Ax = b$" est complètement traitée par cette approche.

4.4 Méthode de Newton-Raphson

4.4.1 Construction de la méthode

Soit $f \in \mathcal{C}^2(\Omega, \mathbb{R}^n)$, où Ω est un ouvert de $\mathbb{R}^n$, et $x^* \in \Omega$ tel que $f(x^*) = 0$.
Pour résoudre cette équation, on résout l'équation approchante :

$$f(x^{(k)}) + \mathrm{d}f(x^{(k)})(x - x^{(k)}) = 0$$

ce qui revient à résoudre un système linéaire. Il faut que $\mathrm{d}f(x^{(k)}) \in GL_n(\mathbb{R})$. Alors, on obtient :

$$x^{(k+1)} = x^{(k)} - (\mathrm{d}f(x^{(k)}))^{-1} f(x^{(k)}).$$

En pratique, on résout

$$\mathrm{d}f(x^{(k)})\delta^{(k)} = -f(x^{(k)}), \qquad \delta^{(k)} \in \mathbb{R}^n$$

puis on pose

$$x^{(k+1)} = x^{(k)} + \delta^{(k)}.$$

Le système linéaire est résolu par une méthode directe ou itérative...

4.4.2 Théorème de convergence locale quadratique

Théorème 4.8

On suppose $f \in \mathcal{C}^1(\Omega, \mathbb{R}^n)$ et $x^* \in \Omega$ tels que
- *$f(x^*) = 0$;*
- *$df(x^*) \in GL_n(\mathbb{R})$, on note $C > 0$ tel que $\|\|df(x^*)^{-1}\|\| \leqslant C$;*
- *il existe $R, L > 0$ tels que $B(x^*, R) \subset \Omega$ et*

$$\forall x, y \in B(x^*, R), \qquad \|\|df(x) - df(y)\|\| \leqslant L \|x - y\| .$$

Alors, il existe $r > 0$ tel que, pour tout $x^{(0)} \in B(x^*, r) \subset \Omega$, la méthode est bien définie et converge vers x^* avec

$$\forall k \geqslant 0, \qquad \left\|x^{(k+1)} - x^*\right\| \leqslant CL \left\|x^{(k)} - x^*\right\|^2 .$$

De plus, r est explicite en fonction des données.

$\triangleright \quad \mathrm{d}f(x) = \mathrm{d}f(x^*) - \mathrm{d}f(x^*) + \mathrm{d}f(x) = \mathrm{d}f(x^*)(I - \underbrace{\mathrm{d}f(x^*)^{-1}(\mathrm{d}f(x^*) - \mathrm{d}f(x))}_{A(x)})$.

$$|||A(x)||| \leqslant |||\mathrm{d}f(x^*)^{-1}||| \, |||\mathrm{d}f(x^*) - \mathrm{d}f(x)||| \leqslant CL \|x^* - x\|.$$

On pose $r = \min(R, \frac{1}{2CL})$. Alors $\forall x \in B(x^*, r)$, $x \in \Omega$ et $|||A(x)||| \leqslant \frac{1}{2}$ donc

$$\mathrm{d}f(x) \in GL_n(\mathbb{R}).$$

De plus,

$$|||\mathrm{d}f(x)^{-1}||| \leqslant |||\mathrm{d}f(x^*)^{-1}||| \, |||(I - A(x))^{-1}||| \leqslant C\frac{1}{1 - \frac{1}{2}} = 2C.$$

Construction de la suite $(x^{(k)})_{k \geqslant 0}$. Soit $x^{(k)} \in B(x^*, r)$. $\mathrm{d}f(x^{(k)}) \in GL_n(\mathbb{R})$ donc on peut définir

$$x^{(k+1)} = x^{(k)} - \mathrm{d}f(x^{(k)})^{-1}.f(x^{(k)}).$$

$$\begin{aligned} \left\|x^{(k+1)} - x^*\right\| &= \left\|x^{(k)} - x^* - \mathrm{d}f(x^{(k)})^{-1}.(f(x^{(k)}) - f(x^*))\right\| \\ &= \left\|\mathrm{d}f(x^{(k)})^{-1}(f(x^{(k)}) - f(x^*) - \mathrm{d}f(x^{(k)})(x^{(k)} - x^*))\right\| \end{aligned}$$

$$\leqslant 2C \left\| \int_0^1 \mathrm{d}f(x^{(k)} + t(x^* - x^{(k)}))(x^* - x^{(k)})\mathrm{d}t - \int_0^1 \mathrm{d}f(x^{(k)})(x^* - x^{(k)})\mathrm{d}t \right\|$$

$$\leqslant 2C \int_0^1 |||\mathrm{d}f(x^{(k)} + t(x^* - x^{(k)})) - \mathrm{d}f(x^{(k)})||| \left\|x^* - x^{(k)}\right\| \mathrm{d}t.$$

$$\leqslant 2C \int_0^1 Lt \left\|x^* - x^{(k)}\right\|^2 \mathrm{d}t \leqslant CL \left\|x^* - x^{(k)}\right\|^2 \leqslant CLr^2 \leqslant \frac{1}{2}r.$$

Donc $x^{(k+1)} \in B(x^*, r)$.

Conclusion : Si $x^{(0)} \in B(x^*, r)$, alors on peut définir $(x^{(k)})_{k \geqslant 0}$ avec $\forall k \in \mathbb{N}$, $x^{(k)} \in B(x^*, r)$ et

$$\left\|x^{(k+1)} - x^*\right\| \leqslant CL \left\|x^{(k)} - x^*\right\|^2.$$

De plus, si $u_k = CL \left\|u^{(k)} - x^*\right\|$, $k \geqslant 0$ alors $0 \leqslant u_{k+1} \leqslant u_k^2$ avec de plus

$$u_0 = CL \left\|x^{(0)} - x^*\right\| \leqslant CLr \leqslant \frac{1}{2}.$$

On montre que

$$\forall k \geqslant 0, u_k \leqslant (u_0)^{2k} \leqslant \left(\frac{1}{2}\right)^{2^k} \xrightarrow[k \to +\infty]{} 0.$$

□

4.4.3 Variantes pratiques

Mise à jour cyclique de la jacobiennne :

$$a^{(k+1)} = x^{(k)} - \mathrm{d}f(x^{(k)})^{-1} f(x^{(k)})$$

remplacé par $x^{(k+1)} = x^{(k)} - \mathrm{d}f(x^{(n_k)})^{-1} f(x^{(k)})$ avec par exemple

$$n_k = (1, \dots, 1, 10, \dots, 10, 20, \dots, 20, \dots).$$

Résolution approchée des systèmes linéaires : $\mathrm{d}f(x^{(k)})\delta^{(k)} = -f(x^{(k)})$ résolu par Jacobi, Gauss-Seidel, gradient, etc.

Approximation de la jacobine par des taux d'accroissement :

$$\forall j \in \{1, \dots, n\}, \qquad \mathrm{d}f(x^{(k)})e_j = \frac{f(x^{(k)} + h_j e_j^{(k)}) - f(x^{(k)})}{h_j^{(k)}}$$

paramètres de calcul $(h_j^{(k)})_{1 \leqslant j \leqslant n, k \geqslant 0}$.

Chapitre 5

Intégration numérique

5.1 Introduction et premiers exemples

5.1.1 Principe

Soit f une fonction réelle, intégrable (éventuellement continue) sur $[a,b]$. On approche $I(f)=\displaystyle\int_a^b f(t)\mathrm{d}t$ par $\displaystyle\int_a^b f_n(t)\mathrm{d}t$ (facile à calculer). f_n approche f. Typiquement, f_n est un polynôme avec $\|f-f_n\|_{L^\infty([a,b])} \xrightarrow[n\to+\infty]{} 0$. Alors $|I(f)-I_n(f)| \leqslant (b-a)\|f-f_n\|_\infty \xrightarrow[n\to+\infty]{} 0$.

Un principe : Utiliser le polynôme d'interpolation de f

$$f_n=\pi_n f : t\mapsto f_n(t)=\sum_{i=0}^{n}\ell_i(t)f(x_i),$$

polynôme d'interpolation de Lagrange aux points $a\leqslant x_0<\cdots<x_n\leqslant b$, $(\ell_i)_{0\leqslant i\leqslant n}$ base de Lagrange de $\mathbb{R}_n[X]$ associée. Alors $I(\pi_n f)=\displaystyle\sum_{i=0}^{n}\alpha_i f(x_i)$ avec $\alpha_i=\displaystyle\int_a^b \ell_i(t)\mathrm{d}t$. Par exemple, avec $n=0$, $x_0=a$, $\pi_n f(t)=f(a)$, $I_n(f)=(b-a)f(a)$: méthode des rectangles à gauche.

Attention : Pour l'interpolation de Lagrange, il n'est pas acquis que

$$\|f-f_n\|_\infty \xrightarrow[n\to+\infty]{} 0 \quad \text{(phénomène de Runge).}$$

Définition 5.1

On appelle formule de quadrature élémentaire sur $[a, b]$ une formule

$$Q(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$$

avec $n \in \mathbb{N}$ fixé, $x_i \in [a, b]$ distincts deux à deux et $\alpha_i \in \mathbb{R}^*$.

5.1.2 Ordre d'une quadrature

Définition 5.2

On appelle ordre d'une quadrature Q le plus grand entier $N \geqslant 0$ tel que $\forall p \in \mathbb{R}_N[X]$, $Q(p) = I(p)$.

Remarque : Pour déterminer N, il suffit d'évaluer $E(f) = I(f) - Q(f)$, l'erreur de quadrature pour f parcourant une base de polynômes telle que la base canonique.

Exemples : Rectangles à gauche : on a

$$Q(f) = (b-a)f(a),$$

$$E(t \mapsto 1) = \int_a^b t\mathrm{d}t - (b-a) = 0,$$

$$E(t \mapsto t) = \int_a^b t\mathrm{d}t - (b-a)a = (b-a)\frac{b-a}{2} \neq 0.$$

Donc ordre 0.

Point milieu : On a

$$Q(f) = (b-a)f\left(\frac{a+b}{2}\right),$$

$$E(t \mapsto 1) = 0,$$

$$E(t \mapsto t) = 0,$$

$$E(t \mapsto t^2) \neq 0$$

Ordre 1.

Proposition 5.3

Une quadrature à $(n+1)$ points est d'ordre au moins n si et seulement si elle est interpolatoire : $\forall f \in \mathcal{C}([a,b])$, $Q(f) = 2(\pi_n f)$ où l'interpolation π_n est effectuée aux points de quadrature $(x_i)_{0\leqslant i\leqslant n}$. À ce moment là, on a nécessairement :

$$\alpha_i = \int_a^b \ell_i(x)dx.$$

▷ Si la formule est interpolatoire, alors soit $p \in \mathbb{R}_n[X]$, on montre que $E(p) = 0$. En effet, $E(p) = I(p) - Q(p) = I(p) - I(\pi_n p) = I(p) - I(p) = 0$. Ainsi, Q est d'ordre n au moins.

Réciproquement, supposons que la quadrature Q est d'ordre au moins n. Alors soit π_n l'interpolation aux points (x_i) de la quadrature. $\pi_n f$ est dans $\mathbb{R}_n[X]$ pour tout f

$$\forall f \in \mathcal{C}([a,b]), \qquad Q(f) = \sum_{i=0}^{n} \alpha_i f(x_i) = \sum_{i=0}^{n} \alpha_i \pi_n f(x_i) = Q(\pi_n f).$$

$Q(f) = Q(\pi_n f) = I(\pi_n f)$ car ordre n au moins et la valeur des (α_i) est obtenue par exemple par : soit $j \in \{0, \ldots, n\}$ $Q(\ell_j) = \sum \alpha_i \ell_j(x_i) = \alpha_j$. Par ailleurs

$$Q(\ell_j) = I(\ell_j) = \int_a^b \ell_j(x)\mathrm{d}x.$$

□

Remarque : On verra que l'ordre d'une quadrature élémentaire à $(n+1)$ points est maximal pour les méthodes de Gauss et vaut $N = 2n + 1$.

5.1.3 Quadratures composées

Soient $k \in \mathbb{N}^*$, $a = a_0 < a_1 < \cdots < a_k = b$ une subdivision de $[a,b]$.

$$\begin{aligned}
\int_a^b f(t)\mathrm{d}t &= \sum_{i=0}^{k-1} \int_{a_i}^{a_{i+1}} f(t)\mathrm{d}t \\
&= \sum_{i=0}^{k-1} (a_{i+1} - a_i) \int_0^1 f(a_i + t(a_{i+1} - a_i))\mathrm{d}t \\
&\simeq \sum_{i=0}^{k-1} (a_{i+1} - a_i) \sum_{j=0}^{n} \alpha_j f(a_i + x_j(a_{i+1} - a_i))
\end{aligned}$$

quadrature associée à la quadrature élémentaire $\int_0^1 g(t)\mathrm{d}t \simeq \sum_{j=0}^{n} \alpha_j g(x_j)$.

Exemple : Si $a_i = a + ih$ avec $h = \frac{b-a}{k}$, la formule des rectangles à gauche :

$$\int_a^b f(t)\mathrm{d}t \simeq \sum_{i=0}^{k-1} \frac{b-a}{k} f\left(a + \frac{i(b-a)}{k}\right) \qquad \text{somme de Riemann}$$

Proposition 5.4

Si Q est une quadrature élémentaire d'ordre $N \geqslant 0$, ie. si $\sum_{i=0}^{n} \alpha_i = (b-a)$ et si f est continue sur $[a,b]$ alors la quadrature composée suivante converge :

$$Q_k(f) = \sum_{i=0}^{k-1} (a_{i+1} - a_i) Q(t \mapsto f(a_i + t(a_{i+1} - a_i)))$$

$$Q_k(f) \to \int_a^b f(t)dt \qquad \textit{quand } h = \max_i (a_{i+1} - a_i) \to 0.$$

Remarque : On verra que plus l'ordre N est élevé, plus la convergence est rapide.

5.1.4 Exemples usuels

Méthode des trapèzes : $\pi_n f$ interpole f en a et en b

$$\pi_n f(t) = f(a) + \frac{t-a}{b-a}(f(b) - f(a)).$$

$$Q(t) = (b-a)\frac{f(a) + f(b)}{2}.$$

Ordre exactement 1.

Méthode de Simpson : Interpolation en a, b et $\frac{a+b}{2}$.

$$Q(f) = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b)).$$

Ordre 3.

Méthodes de Newton-Cotes : On intègre le polynômes d'interpolation de f aux points équidistants $x_j = a + \frac{j}{n}(b-a)$. La quadrature composée de Newton-Cotes à $(n+1)$ points est :

$$Q_k(t) = \sum_{i=0}^{k-1}(a_{i+1} - a_i) \sum_{j=0}^{n} \alpha_i f\left(a_i + \frac{a_{i+1} - a_i}{n} j\right)$$

$$\alpha_j = \int_0^1 \prod_{k=0, k\neq j}^{n} \frac{t - \frac{k}{n}}{\frac{j}{n} - \frac{k}{n}} \mathrm{d}t.$$

Ordre n si n impair, $n+1$ si n pair.

Exemple de calcul d'erreur : point milieu

$$Q(f) = (b-a) f\left(\frac{a+b}{2}\right)$$

Développement de Taylor en $\frac{a+b}{2}$ à l'ordre de 1 pour $f \in \mathcal{C}^2([a,b])$. Pour tout $x \in [a,b]$, il existe $\xi_x \in]a,b[$ tel que

$$f(x) = f\left(\frac{a+b}{2}\right) + \left(x - \frac{a+b}{2}\right) f'\left(\frac{a+b}{2}\right) + \frac{1}{2}\left(x - \frac{a+b}{2}\right)^2 f''(\xi_x).$$

$$I(f) = (b-a) f\left(\frac{a+b}{2}\right) + 0 + \int_a^b \frac{1}{2}(x - \frac{a+b}{2})^2 f''(\xi_x) \mathrm{d}x$$

donc

$$|E(f)| \leqslant \frac{1}{2} \|f''\|_\infty \int_a^b |x - \frac{a+b}{2}|^2 \mathrm{d}x = \frac{(b-a)^3}{24} \|f''\|_\infty .$$

De plus,

$$|E_k(f)| = |Q_k(f) - \int_a^b f(t)\mathrm{d}t| \leqslant \sum_{i=0}^{n} \frac{(a_{i+1} - a_i)^3}{24} \max_{[a_i, a_{i+1}]} |f''| \leqslant \frac{h^2(b-a)}{24} \|f''\|_{\infty,[a,b]} .$$

Analyse d'erreur pour la méthode des trapèzes formulée sur $[a,b]$:

$$Q(f) = (b-a)\frac{f(a) + f(b)}{2}.$$

$$\pi_1 f(x) = f(a) + (x-a)\frac{f(b) - f(a)}{b-a}.$$

$$Q(f) = \int_a^b \pi_1 f(x) \mathrm{d}x.$$

Première étape pour obtenir l'erreur :

$$E(f) = \int_a^b f(t)\mathrm{d}t - \int_a^b \pi_1 f(t)\mathrm{d}t = \int_a^b \underbrace{(f(t) - \pi_1 f(t))}_{\text{à identifier}}\,\mathrm{d}t.$$

Deuxième étape : calcul de l'erreur d'interpolation avec hypothèse de régularité sur f. Si $f \in \mathcal{C}^2([a,b])$.

$$\forall x \in [a,b], \exists \xi_x \in]a,b[, \qquad f(x) - \pi_1 f(x) = \frac{f''(\xi_x)}{2}(x-a)(x-b).$$

Prouvons ce résultat : Soit $x \in]a,b[$ fixé, on considère $q \in \mathbb{R}_2[X]$ qui interpole f en a, b et x.

$$q(t) = \pi_1 f(t) + \alpha(t-a)(t-b).$$

On recherche $\alpha \in \mathbb{R}$ tel que $q(a) = f(a)$, $q(b) = f(b)$ et $q(x) = f(x)$. On trouve

$$\alpha = \frac{f(x) - \pi_1 f(x)}{(x-a)(x-b)}.$$

Par ailleurs, $q - f \in \mathcal{C}^2([a,b])$ et s'annule en a, b et x, donc par le théorème de Rolle, $(q-f)'$ s'annule en deux points distincts de $]a,b[$ et donc $(q-f)''$ s'annule en un point noté $\xi_x \in]a,b[$.

$$0 = (q-f)''(\xi_x) = 2\alpha - f''(\xi_x)$$

d'où le résultat. Finalement

$$E(f) = \frac{1}{2}\int_a^b f''(\xi_x)(x-a)(x-b)\mathrm{d}x$$

donc

$$|E(f)| \leqslant \frac{1}{2}\|f''\|_\infty \int_a^b (x-a)(b-x)\mathrm{d}x.$$

5.2 Étude générale de l'erreur

5.2.1 Noyau de Peano

Définition 5.5

Soit Q une quadrature élémentaire sur $[a,b]$, d'ordre N. On définit sur $[a,b]$:

$$K_N(t) = E(x \mapsto (x-t)_+^N)$$

avec

$$(x-t)_+ = \begin{cases} x-t & si\ x > t \\ 0 & si\ x \leqslant t. \end{cases}$$

Théorème 5.6

Soit $f \in \mathcal{C}^{N+1}([a,b])$ alors

$$E(f) = \frac{1}{N!}\int_a^b K_n(t) f^{(N+1)}(t)dt.$$

▷ La formule de Taylor avec reste intégral en a à l'ordre N s'écrit :

$$f(x) = p_N(x) + \frac{1}{N!}\int_a^x (x-t)^N f^{(N+1)}(t)\mathrm{d}t = p_N(x) + \frac{1}{N!}\int_a^b (x-t)_+^N f^{(N+1)}(t)\mathrm{d}t.$$

Ainsi,

$$\begin{aligned} E(f) &= \underbrace{E(p_N)}_{=0} + E\left(x \mapsto \frac{1}{N!}\int_a^b (x-t)_+^N f^{(N+1)}(t)\mathrm{d}t\right) \\ &= \frac{1}{N!}\int_a^b E(x \mapsto (x-t)_+^N) f^{(N+1)}(t)\mathrm{d}t. \end{aligned}$$

En effet, par exemple, on a

$$\begin{aligned} Q\left(x \mapsto \frac{1}{N!}\int_a^b (x-t)_+^N f^{(N+1)}(t)\mathrm{d}t\right) &= \sum_{i=0}^n \alpha_i \frac{1}{N!}\int_a^b (x_i-t)_+^N f^{(N+1)}(t)\mathrm{d}t \\ &= \frac{1}{N!}\int_a^b \underbrace{\left(\sum_{i=0}^n \alpha_i (x_i-t)_+^N\right)}_{Q(x\mapsto(x-t)_+^N)} f^{(N+1)}(t)\mathrm{d}t. \end{aligned}$$

□

Corollaire 5.7

Si $f \in \mathcal{C}^{N+1}([a,b])$ alors

$$|E(f)| \leqslant \frac{1}{N!}\left\|f^{(N+1)}\right\|_\infty \int_a^b |K_N(t)|dt.$$

De plus, si K_N est de signe constant, il existe $\xi \in]a,b[$ tel que

$$E(f) = \frac{f^{(N+1)}(\xi)}{N!} \int_a^b K_N(t)dt.$$

Si on choisit alors $f(x) = x^{N+1}$,

$$E(x \mapsto x^{N+1}) = \frac{(N+1)!}{N!} \int_a^b K_N(t)dt$$

et donc pour toute fonction $f \in \mathcal{C}^{N+1}([a,b])$, il existe $\xi \in]a,b[$, tel que

$$E(f) = \frac{f^{(N+1)}(\xi)}{N+1} E(x \mapsto x^{N+1}).$$

5.2.2 Erreur pour les quadratures composées

Soient $a = a_0 < a_1 < \cdots < a_k = b$ une subdivision et $Q_{\{a_0,\ldots,a_k\}}(f)$ la quadrature composée correspondante :

$$Q_{\{a_0,\ldots,a_k\}}(f) = \sum_{j=0}^{k-1} (a_{j+1} - a_j) Q(x \mapsto f(a_j + (a_{j+1} - a_j)x))$$

avec Q la quadrature élémentaire sur $[0,1]$.

L'erreur pour la quadrature composée est :

$$\begin{aligned}
E_{\{a_0,\ldots,a_k\}}(f) &= \sum_{j=0}^{k-1} \left(\int_{a_j}^{a_{j+1}} f(t)\mathrm{d}t - (a_{j+1} - a_j) Q(x \mapsto f(a_j + (a_{j+1} - a_j)x)) \right) \\
&= \sum_{j=0}^{k-1} (a_{j+1} - a_j) E(x \mapsto f(a_j + (a_{j+1} - a_j)x)) \\
&= \sum_{j=0}^{k-1} (a_{j+1} - a_j) \frac{1}{N!} \int_0^1 K_N(t) \frac{\mathrm{d}^{N+1}}{\mathrm{d}t^{N+1}} (f(a_j + (a_{j+1} - a_j)t))\mathrm{d}t \\
&= \sum_{j=0}^{k-1} (a_{j+1} - a_j)^{N+2} \frac{1}{N!} \int_0^1 K_N(t) f^{(N+1)}(a_j + (a_{j+1} - a_j)t)\mathrm{d}t
\end{aligned}$$

On pose $h = \max_{0\leqslant j\leqslant k-1}(a_{j+1} - a_j)$. Alors,

$$\begin{aligned} |E_{\{a_0,\dots,a_k\}}(f)| &\leqslant \sum_{j=0}^{k-1} h^{N+2}\frac{1}{N!}\int_0^1 |K_N(t)|\left\|f^{(N+1)}\right\|_{L^\infty([a,b])}\mathrm{d}t \\ &\leqslant \frac{kh^{N+2}\left\|f^{(N+1)}\right\|_\infty}{N!}\int_0^1 |K_N(t)|\mathrm{d}t. \end{aligned}$$

Si le pas est uniforme : $\forall j,\ a_{j+1} - a_j = \frac{b-a}{k}$, alors

$$|E(f)| \leqslant \frac{(b-a)h^{N+1}\left\|f^{(N+1)}\right\|_\infty}{N!}\int_0^1 |K_N(t)|\mathrm{d}t.$$

L'erreur tend vers 0 lorsque $h \to 0$ comme h^{N+1}.

5.3 Méthodes de Gauss

Objectif : – Rechercher les méthodes de quadrature d'ordre maximal (avec un nombre de points fixés).

– Intégrer sur $\mathbb{R}$, $\mathbb{R}_+$, ou en présence de singularités au bord d'un domaine borné.

On considère $]\alpha,\beta[$ un intervalle fini ou non-fini, une fonction w définie sur $]\alpha,\beta[$:

– positive,

– telle que $\displaystyle\int_\alpha^\beta w(x)|x|^n\mathrm{d}x < \infty$ pour tout $n \in \mathbb{N}$,

– et $\displaystyle\int_\alpha^\beta w(x)f(x)\mathrm{d}x = 0$ avec f continue et positive $\Longrightarrow f \equiv 0$.

On appelle w une fonction poids. On veut approcher, pour f telle que

$$\int_\alpha^\beta |w(x)f(x)|\mathrm{d}x < \infty,$$

la quantité

$$\int_\alpha^\beta f(x)w(x)\mathrm{d}x \simeq \sum_{i=0}^n \alpha_i f(x_i).$$

Exemple : sur $]-1,1[$, $w(x) = \frac{1}{\sqrt{1-x^2}}$, $\displaystyle\int_{-1}^1 \frac{e^{-t}+\cos(t)}{\sqrt{1-t^2}}\mathrm{d}t$.

Théorème 5.8

Il existe un unique $(x_0, \dots, x_n) \in]\alpha, \beta[^{n+1}$, *et un unique* $(\alpha_0, \dots, \alpha_n) \in \mathbb{R}^{n+1}$ *tels que* Q *soit d'ordre maximal* N. *Alors* $N = 2n+1$ *et les* $(x_i)_{0 \leqslant i \leqslant n}$ *sont les* $n+1$ *racines du polynôme* P_{n+1}, *de degré* $n+1$, *issu d'une famille de polynômes orthogonaux pour le produit scalaire*

$$(f, g) = \int_\alpha^\beta w(x) f(x) g(x) dx$$

et les α_i *sont obtenus par la formule :*

$$0 \leqslant i \leqslant n, \qquad \alpha_i = \int_\alpha^\beta w(x) \ell_i(x) dx$$

où ℓ_i *est le* i*-ième polynôme de Lagrange associé aux* (x_i).

Rappel : Il existe une unique famille de polynômes $(P_k)_{k \geqslant 0}$ tel que
- $\deg P_k = k$;
- le monôme de plus haut degré de P_k est exactement x^k ;
- $(P_i, P_j) = 0$ si $i \neq j$.

De plus, chaque polynôme P_k admet exactement k racines simples dans $]\alpha, \beta[$.

Exemples : $-$ $[\alpha, \beta] = [-1, 1]$, $w(x) = 1$: polynômes de Legendre. Dans cet exemple, on retrouve la formule du point milieu avec l'unique racine de $P_1 = X$: 0. $Q(f) = 2f(0)$ approchant $\displaystyle\int_{-1}^{1} f(t)\mathrm{d}t$: ordre 1.
Avec $P_2 = X^2 - \frac{1}{3}$ et ses racines $\pm\frac{1}{\sqrt{3}}$ on obtient la quadrature suivante :

$$Q(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

d'ordre 3 : "formule de Gauss-Legendre à deux points".
Pour $n = 2$, les racines de P_3 sont 0 et $\pm\sqrt{\frac{3}{5}}$

$$Q(f) = \frac{1}{9}\left(5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right)$$

d'ordre 5.

– Sur $]-1,1[$, $w(x)=\frac{1}{\sqrt{1-x^2}}$ on obtient pour $(P_k)_{k\geqslant 0}$ les polynômes de Tchebyshev. Les racines de P_{n+1} sont $x_i=\cos\left(\frac{2i+1}{2n+2}\pi\right)$, $0\leqslant i\leqslant n$ et $\alpha_i=\frac{\pi}{n+1}$ $\forall i$.

– $\mathbb{R}_+$, $w(x)=e^{-x}$: polynômes de Laguerre. Pour $n=1$, quadrature de Gauss-Laguerre à 2 points :

$$\int_0^{+\infty} f(x)e^{-x}\mathrm{d}x \simeq \frac{1}{4}(2+\sqrt{2})f(2-\sqrt{2})+\frac{1}{4}(2-\sqrt{2})f(2+\sqrt{2})$$

(formule exacte sur $\mathbb{R}_3[X]$)

– $\mathbb{R}$, $w(x)=e^{-x^2}$, polynômes de Hermite. $n=1$:

$$Q(f)=\frac{\sqrt{\pi}2}{f}\left(\frac{\sqrt{2}}{2}\right)+\frac{\sqrt{\pi}}{2}f\left(-\frac{\sqrt{2}}{2}\right).$$

▷ (du théorème) Considérons P_{n+1} le $n+1$-ième polynôme orthogonal sur $]-1,1[$ vérifiant les hypothèses énoncées dans le rappel. Soit $P\in\mathbb{R}_{2n+1}[X]$, on montre que $E(P)=0$. On effectue la division euclidienne de P par P_{n+1} :

$$P=P_{n+1}Q+R$$

avec $Q\in\mathbb{R}_n[X]$ et $R\in\mathbb{R}_n[X]$. Alors,

$$\int_\alpha^\beta P(x)w(x)\mathrm{d}x=\underbrace{\int_\alpha^\beta P_{n+1}(x)Q(x)w(x)\mathrm{d}x}_{(P_{n+1},Q)=0}+\int_\alpha^\beta R(x)w(x)\mathrm{d}x$$

Or $P(x_i)=R(x_i)$ car $P_{n+1}(x_i)=0$ donc $\mathrm{Quad}(P)=\mathrm{Quad}(R)=\displaystyle\int_\alpha^\beta R(x)w(x)\mathrm{d}x$ car les poids (α_i) sont choisis de sorte que la quadrature soit interpolatoire (exacte sur au moins $\mathbb{R}_n[X]$). Ainsi, $E(P)=0$ donc la quadrature est d'ordre au moins $2n+1$. Elle n'est pas d'ordre $2n+2$ car pour $P=(P_{n+1})^2\in\mathbb{R}_{2n+2}[X]$, on a

$$\int_\alpha^\beta (P_{n+1})^2 w\mathrm{d}x>0$$

mais

$$\mathrm{Quad}(P_{n+1}^2)p=\sum_{i=0}^{n}\alpha_i P_{n+1}(x_i)^2=0.$$

Preuve de l'unicité : Soit $\widetilde{\text{Quad}} = \sum_{i=0}^{n} \beta_i f(y_i)$ une formule d'ordre supérieur ou égal à $2n+1$. On montre que cette formule coincide avec celle du théorème. Soit

$$\pi_{n+1}(x) = \prod_{i=0}^{n}(x - y_i).$$

Pour tout $p \in \mathbb{R}_n[X]$, $\deg(P\pi_{n+1}) \leqslant 2n+1$ donc

$$\int_{\alpha}^{\beta} P(x)\pi_{n+1}(x)w(x)\mathrm{d}x = \widetilde{\text{Quad}}(P\pi_{n+1}) = 0$$

donc $\pi_{n+1} \perp \mathbb{R}_n[X]$. Or π_{n+1} unitaire de degré $n+1$. Ainsi, par unicité de P_{n+1}, $\pi_{n+1} = P_{n+1}\ \{x_0, \ldots, x_n\} = \{y_0, \ldots, y_n\}$ et par suite, les coefficients coïncident également. □