Université de Rennes 1
U.F.R. Mathématiques
Licence Math

2006 - 2007
Support de cours
F04

ALGEBRE LINEAIRE NUMERIQUE

Eric DARRIGRAND, Grégory VIAL

Table des matières

Chapitre 1

ALGÈBRE LINÉAIRE : RAPPELS ET COMPLEMENTS

1.1 Notations et Rappels

Nous nous plaçons dans le corps $\mathbb{K}$, en général $\mathbb{R}$ ou $\mathbb{C}$.

1.1.1 Quelques définitions et notations

Définition 1.

On notera

$\mathbb{K}^d$ l'ensemble des vecteurs de taille d.

$\mathbb{K}^{m\times p}$ ou $\mathcal{M}_{m,p}(\mathbb{K})$ l'ensemble des matrices carrées de taille $m \times p$ (à m lignes et p colonnes).

$\mathcal{M}_m(\mathbb{K}) = \mathcal{M}_{m,m}(\mathbb{K})$ l'ensemble des matrices carrées de taille $m \times m$.

Propriété 1.

$\mathcal{M}_{m,p}(\mathbb{K})$ muni de l'addition et du produit par un scalaire définit un espace vectoriel (cad, $\forall A, B \in \mathcal{M}_{m,p}(\mathbb{K})$ et $\alpha \in \mathbb{K}$, $\alpha A + B \in \mathcal{M}_{m,p}(\mathbb{K})$).

Propriété 2.

$\mathcal{M}_m(\mathbb{K})$ muni de plus du produit entre matrices définit une algèbre. Rappel sur le produit : la matrice $C = AB$ est définie par $C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$.

Définition 2.

Soit A une matrice de $\mathcal{M}_m(\mathbb{K})$. A est dite inversible ou régulière si il existe une matrice B telle que $AB = BA = I$. La matrice B est alors notée A^{-1} et appelée inverse de A.
I désigne la matrice identité de $\mathcal{M}_m(\mathbb{K})$: $I_{ij} = \delta_{ij}$ où δ_{ij} est le symbole de Kronecker (il vaut 1 si $i = j$ et 0 sinon).

Propriété 3.

Soit A une matrice de $\mathcal{M}_m(\mathbb{K})$. Les propositions suivantes sont équivalentes :

1. *A est inversible,*
2. $\ker A = \{0\}$,
3. $\mathrm{Im}A = \mathbb{K}^m$,
4. *il existe $B \in \mathcal{M}_m(\mathbb{K})$ telle que $AB = I$,*
5. *il existe $B \in \mathcal{M}_m(\mathbb{K})$ telle que $BA = I$,*

Définition 3. (et Propriété)

L'ensemble des matrices de $\mathcal{M}_m(\mathbb{K})$ inversibles est noté $GL_m(\mathbb{K})$ et constitue un groupe pour la multiplication dans $\mathcal{M}_m(\mathbb{K})$ qu'on appelle groupe linéaire.
L'ensemble des matrices de $\mathcal{M}_m(\mathbb{K})$ inversibles et de déterminant égal à 1 est noté $SL_m(\mathbb{K})$ et constitue un groupe pour la multiplication dans $\mathcal{M}_m(\mathbb{K})$ qu'on appelle groupe spécial linéaire.

Définition 4.

Soit A une matrice de $\mathcal{M}_{m,p}(\mathbb{K})$. On appelle respectivement matrice transposée de A, notée A^T, et matrice adjointe de A, notée A^*, les matrices de $\mathcal{M}_{p,m}(\mathbb{K})$ définies par

$$A^T_{ij} = A_{ji} \ \ \forall (i,j) \in \{1,...,m\} \times \{1,...,p\} \qquad \text{et} \qquad \begin{cases} A^* = A^T & \text{si } \mathbb{K} = \mathbb{R}\,, \\ A^* = \overline{A}^T & \text{si } \mathbb{K} = \mathbb{C}\,. \end{cases}$$

Proposition 1.

Soit A une matrice de $\mathcal{M}_{m,p}(\mathbb{K})$. Alors :
$\dim \mathrm{Im}A = \dim \mathrm{Im}A^*$,
$\ker A^* = (\mathrm{Im}A)^\perp$,
$\mathrm{Im}A^* = (\ker A)^\perp$.

Définition 5.

Soit A une matrice de $\mathcal{M}_m(\mathbb{K})$. La matrice A est dite
- diagonale si $A_{ij} = 0$ pour tout (i,j) tel que $i \neq j$
  (on désigne alors A par diag($\lambda_1, \cdots, \lambda_m$), où $\lambda_i = A_{ii}$ pour tout $i \in \{1,...,m\}$),
- symétrique si $A = A^T$,
- orthogonale $A^{-1} = A^T$,
- unitaire si $A^{-1} = A^*$,
- normale si $AA^* = A^*A$,
- hermitienne si $A = A^*$ et $\mathbb{K} = \mathbb{C}$,
- auto-adjointe si $A = A^*$,
- symétrique positive, lorsque $\mathbb{K} = \mathbb{R}$, si A est symétrique et si pour tout vecteur v de $\mathbb{K}^m$, $v^T Av \geq 0$,
- symétrique définie positive, lorsque $\mathbb{K} = \mathbb{R}$, si A est symétrique et si pour tout vecteur v de $\mathbb{K}^m \setminus \{0\}$, $v^T Av > 0$.
- hermitienne positive, lorsque $\mathbb{K} = \mathbb{C}$, si A est hermitienne et si pour tout vecteur v de $\mathbb{K}^m$, $\overline{v}^T Av \geq 0$,
- hermitienne définie positive, lorsque $\mathbb{K} = \mathbb{C}$, si A est hermitienne et si pour tout vecteur v de $\mathbb{K}^m \setminus \{0\}$, $\overline{v}^T Av > 0$.
- triangulaire inférieure si $a_{ij} = 0$ pour tout (i,j) tel que $i < j$.
- triangulaire supérieure si $a_{ij} = 0$ pour tout (i,j) tel que $i > j$.

Quelques exemples de matrices

- Une matrice symétrique non hermitienne :

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

- Une matrice hermitienne non symétrique :

$$\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$$

- Une matrice symétrique et hermitienne :

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

· Matrice de Householder :
Soit v un vecteur non nul. On définit la matrice de Householder associée $H(v)$ par $H(v) = I - 2\frac{vv^T}{\|v\|^2}$. Cette matrice est symétrique et orthogonale. Elle est associée à la symétrie orthogonale par rapport à l'hyperplan orthogonal à v. Si x est un vecteur quelconque et u le vecteur de cet hyperplan tel que $x = \alpha v + u$, alors $H(v)x = -\alpha v + u$.
Elle a les bonnes propriétés suivantes :
Si e est un vecteur unitaire vérifiant $v \neq \pm \|v\| e$, alors

$$H(v + \|v\| e)v = -\|v\| e \quad \text{et} \quad H(v - \|v\| e)v = +\|v\| e \,.$$

On peut déjà constater que $H(v + \|v\| e)$ transforme un vecteur v qui a priori peut ne pas avoir de composantes nulles, en un vecteur qui n'aura qu'une composante non nulle pour e dans la base canonique. Les matrices de Householder nous permettront ainsi de mettre en place des méthodes de réduction des matrices (voir les chapitres suivants).

· Matrice de Givens :
La matrice de Givens est une matrice de rotation. Soit θ un angle. La matrice associée à la rotation d'angle θ est la matrice de Givens :

$$\begin{pmatrix}
1 & & & & & & & & & \\
& \ddots & & & & & & & & \\
& & 1 & & & & & & & \\
& & & c & & & & s & & \\
& & & & 1 & & & & & \\
& & & & & \ddots & & & & \\
& & & & & & 1 & & & \\
& & & -s & & & & c & & \\
& & & & & & & & 1 & \\
& & & & & & & & & \ddots \\
& & & & & & & & & & 1
\end{pmatrix}$$

avec $c = \cos(\theta)$ et $s = \sin(\theta)$. Les éléments non diagonaux non représentés sont nuls. Cette matrice est orthogonale et vérifie des propriétés semblables à celles de la matrice de Householder.

1.1.2 Théorie spectrale - Premières réductions

Soient $A, B \in \mathcal{M}_m(\mathbb{C})$; $\mathrm{tr}\,(A)$: trace de A ; $\det(A)$: déterminant de A.

Définition 6.

On appelle permutation d'ordre m, toute bijection de l'ensemble $\{1, \cdots, m\}$ dans lui-même. L'ensemble des permutations d'ordre m est noté $\mathcal{S}_m$. On appelle signature d'une permutation $\sigma \in \mathcal{S}_m$, le nombre

$$\varepsilon(\sigma) = (-1)^{p(\sigma)} \qquad \text{avec} \qquad p(\sigma) = \sum_{1 \leq i \leq j \leq m} \mathrm{Inv}_\sigma(i,j),$$

où $\mathrm{Inv}_\sigma(i,j) = \begin{cases} 0 & \text{si } \sigma(i) \leq \sigma(j)\,, \\ 1 & \text{si } \sigma(i) \geq \sigma(j)\,. \end{cases}$

Définition 7.

On appelle trace de A et on note $\mathrm{tr}\,A$ la somme de ses éléments diagonaux. On appelle déterminant de A le nombre

$$\det A = \sum_{\sigma \in \mathcal{S}_m} \varepsilon(\sigma) \prod_{i=1}^{m} a_{i\sigma(j)}\,.$$

Propriété 4.

La trace et le déterminant vérifient les propriétés suivantes :

$\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$,

$\det(AB) = \det A\ \det B = \det(BA)$,

La trace et le déterminant sont invariants par changement de base.

Propriété 5.

Désignons par $\delta_{m-1}(i,j)$ le déterminant de la matrice carrée de taille $(m-1)\times(m-1)$ extraite de A par suppression de la i-ème ligne et de la j-ème colonne. Alors, pour tout $i \in \{1,\cdots,m\}$

$$\det A = \sum_{j=1}^{m} (-1)^{i+j} a_{ij} \delta_{m-1}(i,j)\,.$$

Cette relation offre une première technique de calcul d'un déterminant. Mais le coût en temps de calcul d'une telle méthode est rédhibitoire : de l'ordre de $(m!)$. La programmation de cette technique est alors fortement déconseillée dès lors que m n'est plus de l'ordre de l'unité.

Définition 8.

On appelle polynôme caractéristique et on note $P_A(\lambda)$ (ou $\chi_A(\lambda)$) le polynôme

$$\chi_A(\lambda) = P_A(\lambda) = \det(A - \lambda I)\,.$$

Ses n racines complexes sont appelées valeurs propres de A. Soit λ_i une valeur propre de A. On dit que λ_i est une valeur propre de multiplicité n_i si λ_i est une racine de $P_A(\lambda)$ de multiplicité n_i. L'ensemble des valeurs propres de A est appelé spectre de A et est noté $\sigma(A)$.

Définition 9.

Soit λ une valeur propre de A. On dit que x est un vecteur propre de A associé à λ si $x \neq 0$ et $Ax = \lambda x$.

Définition 10.

Soit λ une valeur propre de A. On appelle sous-espace propre associé à λ, le sous-espace $E_\lambda = \ker(A - \lambda I)$. On appelle sous-espace spectral ou caractéristique associé à λ le sous-espace $F_\lambda = \cup_{k\geq 1} \ker(A-\lambda I)^k$.

Remarque 1.

Il existe k_0 tel que $F_\lambda = \cup_{1\leq k\leq k_0} \ker(A-\lambda I)^k = \ker(A-\lambda I)^{k_0}$.

Définition 11.

Soit $P(X) = \sum_{i=1}^{d} \alpha_i X^i$ un polynôme sur $\mathbb{C}$. On note $P(A)$, polynôme de la matrice A, la matrice $\sum_{i=1}^{d} \alpha_i A^i$.

Remarque 2.

Soient $P(X)$ et $Q(X)$ deux polynômes sur $\mathbb{C}$. Alors $P(A)Q(A) = Q(A)P(A)$. Si λ est valeur propre de A alors $P(\lambda)$ est valeur propre de $P(A)$.

Démonstration : Soit x vecteur propre associé à λ. Alors, $A^2x = A(\lambda x) = \lambda^2 x$. Par récurrence, on montre que $A^p x = \lambda^p x$ et ainsi $P(A)x = P(\lambda)x$.

Théorème 1. (lemme des noyaux)

Soient $\lambda_1,\cdots,\lambda_p$ les p valeurs propres distinctes de $A \in \mathcal{M}_m(\mathbb{K})$. On note n_i leurs multiplicités ($1 \leq n_i \leq m$ et $\sum_{i=1}^{p} n_i = m$), alors

$$\mathbb{C}^m = \oplus_{i=1}^{p} F_{\lambda_i} \ \text{ où } \ F_{\lambda_i} = \ker(A-\lambda_i)^{n_i},\ \ n_i = \dim F_{\lambda_i}, \quad \text{et} \quad P_A(\lambda) = \prod_{i=1}^{p} (\lambda_i - \lambda)^{n_i}\ .$$

Définition 12.

$A \in \mathcal{M}_m(\mathbb{C})$ est toujours triangularisable, cad : Il existe T triangulaire et $P \in \mathcal{M}_m(\mathbb{C})$ inversible telles que $PTP^{-1} = A$.
$A \in \mathcal{M}_m(\mathbb{C})$ est diagonalisable (cad, il existe D diagonale et $P \in \mathcal{M}_m(\mathbb{C})$ inversible telles que $PDP^{-1} = A$) ssi $F_{\lambda_i} = E_{\lambda_i}$ (autrement dit, $\dim(E_{\lambda_i}) = n_i$) pour toute valeur propre λ_i de A.

Démonstration de la première partie (par récurrence sur la taille m de la matrice) :
P_A admet une racine dans $\mathbb{C}$ notée λ. A admet un vecteur propre v associé à la valeur propre λ. Il existe alors une matrice de changement de base P_1 contenant v telle que $A = P_1 \widetilde{A} P_1^{-1}$, avec $\widetilde{A}$ de la forme

$$\widetilde{A} = \begin{pmatrix} \lambda & \alpha_2 & \cdots & \alpha_m \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}$$

où B est une matrice de dimension $(m-1) \times (m-1)$.
Hypothèse de récurrence : $B = P_2 T_B P_2^{-1}$ avec T_B triangulaire et P_2 inversible. En posant finalement $P = P_1 P_3$ avec P_3 définie par

$$P_3 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & P_2 & \\ 0 & & & \end{pmatrix}$$

et $(\beta_2, \cdots, \beta_m) = (\alpha_2, \cdots, \alpha_m) P_2$, on obtient

$$P^{-1} A P = \begin{pmatrix} \lambda & \beta_2 & \cdots & \beta_m \\ 0 & & & \\ \vdots & & T_B & \\ 0 & & & \end{pmatrix}$$

Théorème 2. (théorème de Cayley-Hamilton)

$$P_A(A) = 0.$$

Démonstration : Si A est diagonalisable : Soit x un vecteur propre et λ la valeur propre associée. Alors on a $P_A(A)x = P_A(\lambda)x$ = 0. On en déduit que tout vecteur propre de A est dans le noyau de $P_A(A)$, ce qui implique que $P_A(A) = 0$, puisqu'on peut trouver une base de vecteurs propres.

Si A est non diagonalisable, on utilise deux résultats non triviaux : Densité de l'ensemble des matrices diagonalisables dans l'ensemble des matrices et continuité de l'application $A \mapsto P_A(A)$.

Définition 13.

On appelle polynôme minimal de A le polynôme de plus petit degré et de coefficient de plus haut degré égal à 1, qui s'annule en A.

Remarque 3.

Si A admet m valeurs propres distinctes deux à deux, alors le polynôme minimal est égal au polynôme caractéristique. La réciproque est fausse.

Exercice : Qu'en est-il de la matrice $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$?

Remarque 4.

Soit B_i une base de F_{λ_i}, alors $B = \cup_{i=1}^{p} B_i$ est une base de $\mathbb{C}^m$. Notons P la matrice de changement de base associée, alors

$$P^{-1}AP = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{pmatrix}$$

où A_i est une matrice carrée de taille n_i ayant pour unique valeur propre λ_i avec la multiplicité n_i, et pouvant être réduite selon la forme de Jordan. Chaque A_i peut s'écrire sous la forme

$$\begin{pmatrix} \lambda_i & \varepsilon_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon_{n_i-1} \\ & & & \lambda_i \end{pmatrix}$$

avec $\varepsilon_k \in \{0, 1\}$ pour $k = 1, \cdots, n_i - 1$.

Si la matrice est diagonalisable, la forme de Jordan donne la matrice diagonale. Sinon, cette forme peut être qualifiée "forme diagonalisée des matrices non diagonalisables".

Théorème 3. (théorème de Schur)

Pour tout $A \in \mathcal{M}_m(\mathbb{C})$, il existe U unitaire ($U^{-1} = U^*$) telle que U^*AU soit triangulaire.

Démonstration : $\exists P$ matrice de changement de base et T triangulaire telles que $A = PTP^{-1}$. Soient v_i les vecteurs colonnes de P. Soient u_i les vecteurs obtenus par orthonormalisation des v_i respectant la relation $\text{Vect}\{v_1, \cdots, v_k\} = \text{Vect}\{u_1, \cdots, u_k\}$ pour tout k. Une telle opération est toujours possible, par le procédé d'orthonormalisation de Gram-Schmidt par exemple. La matrice U engendrée par les u_i est donc unitaire et U^*AU est triangulaire.
En effet, $\exists T$ triangulaire telle que $AP = PT \iff \text{Vect}\{Av_1, \cdots, Av_k\} \subset \text{Vect}\{v_1, \cdots, v_k\}$
$\implies \text{Vect}\{Au_1, \cdots, Au_k\} \subset \text{Vect}\{u_1, \cdots, u_k\} \iff \exists R$ triangulaire telle que $AU = UR$.

Autre démonstration : On aurait pu reprendre la démonstration relative à la définition des matrices triangularisables en complétant v par une famille de vecteurs orthogonaux à v.

Théorème 4.

La matrice A, de valeurs propres $\lambda_1, \cdots, \lambda_m$, est normale ($AA^* = A^*A$) si et seulement s'il existe une matrice unitaire U telle que $A = U\text{diag}(\lambda_1, \cdots, \lambda_m)U^*$. De plus, A peut s'écrire

$$A = \sum_{i=1}^{m} \lambda_i u_i u_i^* ,$$

où les u_i désignent les colonnes de U, autrement dit, les vecteurs propres de A.

Remarque 5.

Cette écriture permet la mise en place d'une technique de réduction de matrice lorsque certaines valeurs propres sont petites par rapport à d'autres.

Théorème 5.

La matrice A, de valeurs propres $\lambda_1, \cdots, \lambda_m$, est auto-adjointe ($A = A^*$) si et seulement s'il existe une matrice unitaire U telle que $A = U\text{diag}(\lambda_1, \cdots, \lambda_m)U^*$, avec $\lambda_i \in \mathbb{R}$.

Démonstration : corollaire du précédent.

Théorème 6.

Soit A une matrice auto-adjointe ($A = A^*$). A est positive si et seulement si toutes ses valeurs propres sont positives ou nulles. A est définie positive si et seulement si toutes ses valeurs propres sont strictement positives.

Théorème 7.

Soit A une matrice auto-adjointe positive. Alors $\exists B$ notée $A^{1/2}$ telle que $A = B^2$.

Démonstration : Existence : $A = UDU^*$ avec D positive. On prend alors Δ telle que $D = \Delta^2$. Immédiat puisque D est diagonale à éléments positifs. On a alors $\Delta = \Delta^*$ et on peut écrire $A = B^2$ avec $B = U\Delta U^*$. L'unicité sera vue en TD.

1.1.3 Normes et suites de matrices

Définition 14.

Une norme sur $\mathbb{C}^m$ est une application, notée $\|\cdot\|$, de $\mathbb{C}^m$ dans $\mathbb{R}^+$ qui vérifie les propriétés suivantes

1. $\forall x \in \mathbb{C}^m,\ \|x\| = 0 \Longrightarrow x = 0$,
2. $\forall x \in \mathbb{C}^m,\ \forall \lambda \in \mathbb{C},\ \|\lambda x\| = |\lambda|\,\|x\|$,
3. $\forall x \in \mathbb{C}^m,\ \forall y \in \mathbb{C}^m,\ \|x + y\| \leq \|x\| + \|y\|$.

Quelques normes usuelles :

$$\text{La norme euclidienne : } \|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2\right)^{\frac{1}{2}}$$

$$\text{La norme } l^p : \|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}},\ p \geq 1,$$

$$\text{La norme } l^\infty : \|x\|_\infty = \max_{1\leq i\leq m} |x_i|.$$

$\mathbb{C}^m$ est de dimension finie donc toutes les normes sur $\mathbb{C}^m$ sont équivalentes. En particulier, rappelons les relations d'équivalence :

$$\|x\|_\infty \leq \|x\|_p \leq m^{1/p}\,\|x\|_\infty \quad \text{et} \quad \|x\|_2 \leq \|x\|_1 \leq \sqrt{m}\,\|x\|_2 .$$

Définition 15. (Norme matricielle)

Une norme $\|\cdot\|$ sur $\mathcal{M}_m(\mathbb{C})$ est dite matricielle si elle vérifie pour toutes matrices $A, B \in \mathcal{M}_m(\mathbb{C})$

$$\|AB\| \leq \|A\|\,\|B\| .$$

Définition 16. (Norme subordonnée)

Soit $\|\cdot\|$ une norme vectorielle sur $\mathbb{C}^m$. On lui associe sur $\mathcal{M}_m(\mathbb{C})$ une norme matricielle dite subordonnée à cette norme vectorielle et définie par : pour toute matrice $A \in \mathcal{M}_m(\mathbb{C})$

$$\|A\| = \sup_{x\in\mathbb{C}^m, x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Quelques relations d'équivalence pour les normes subordonnées usuelles :

$$m^{-1/p}\,\|A\|_\infty \leq \|A\|_p \leq m^{1/p}\,\|A\|_\infty , \quad m^{-1/2}\,\|A\|_2 \leq \|A\|_1 \leq m^{1/2}\,\|A\|_2 .$$

Proposition 2.

Soit $\|\cdot\|$ une norme matricielle subordonnée sur $\mathcal{M}_m(\mathbb{C})$. Alors

1. *Pour toute matrice A, la norme $\|A\|$ est aussi définie par*

$$\|A\| = \sup_{x\in\mathbb{C}^m, \|x\|=1} \|Ax\| = \sup_{x\in\mathbb{C}^m, \|x\|\leq 1} \|Ax\|,$$

2. *Il existe $x_A \in \mathbb{C}^m, x_A \neq 0$ tel que*

$$\|A\| = \frac{\|Ax_A\|}{\|x_A\|},$$

3. *La matrice identité vérifie*

$$\|I\| = 1,$$

4. *Une norme subordonnée est bien une norme matricielle.*

Définition 17. (Norme de Frobenius)

$$\|A\|_F = \sum_{1\leq i,j\leq m} |a_{ij}|^2 = (\mathrm{tr}\,(A^T A))^{1/2}$$

La norme de Frobenius est une norme matricielle non subordonnée.

Définition 18. (Conditionnement)

Soit A inversible. On appelle conditionnement de A relativement à la norme $\|\cdot\|$, le nombre $\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\|$.

Propriété 6.

Si A est carrée symétrique définie positive, $\mathrm{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$, où $\lambda_{\max}(A)$ et $\lambda_{\min}(A)$ désignent respectivement la plus grande et la plus petite valeurs propres de A.

Remarque 6.

Le conditionnement est une quantité qui joue un rôle important dans la résolution numérique des systèmes linéaires. En effet, on verra qu'il est un indicateur de la stabilité numérique des méthodes directes et de la vitesse de convergence des méthodes itératives.

Définition 19. (Rayon spectral)

On appelle rayon spectral de A : $\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda|$.

Théorème 8.

Pour toute norme subordonnée $\|\cdot\|$, $\rho(A) \leq \|A\|$.
Réciproquement : $\forall \varepsilon > 0,\ \exists \|\cdot\|_\varepsilon$ une norme subordonnée telle que $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$.

Proposition 3. (Autres résultats)

- $\|(A^n)\|^{1/n} \longrightarrow \rho(A)$ *quand* $n \to \infty$.
- $(A^n)_n$ *converge ssi* $\rho(A) < 1$.
- $\exp(tA) \longrightarrow 0$ *quand* $t \to \infty$ *ssi* $\forall \lambda \in \sigma(A), \Re\mathrm{e}(\lambda) < 0$
- $\exp(tA)$ *bornée ssi (*$\forall \lambda \in \sigma(A), \Re\mathrm{e}(\lambda) \leq 0$ *et* $[\Re\mathrm{e}(\lambda) = 0 \Rightarrow \mathrm{E}_\lambda = \mathrm{F}_\lambda]$*).*

Ce type de résultat intervient dans la résolution des systèmes d'équations différentielles à coefficients constants.

1.2 Décompositions usuelles

Nous avons jusqu'à présent introduit quelques objets relatifs à la théorie de l'algèbre linéaire. Nous proposons ici quelques premiers outils qui joueront un rôle fondamental dans la résolution de systèmes linéaires.

Une matrice peut souvent être difficile à inverser. Il est donc appréciable de pouvoir la décomposer en plusieurs matrices dont les propriétés sont plus avantageuses.

1.2.1 Décomposition polaire

Théorème 9.

Soit $A = (A_{ij})_{1\leq i,j\leq m} \in Gl_m(\mathbb{C})$, alors il existe un unique couple de matrices (H, U), avec H hermitienne définie positive, et U unitaire ($U^{-1} = U^*$), tel que $A = HU$.

Démonstration : Supposons que l'on peut écrire $A = HU$ avec U unitaire et H hermitienne alors $AA^* = HUU^*H = H^2$. On en déduit que H est unique et hermitienne positive nécessairement donnée par $H = (AA^*)^{1/2}$.
On peut toujours choisir $H = (AA^*)^{1/2}$ et poser $A = HU$. Il suffit de prendre $U = H^{-1}A$. Montrons alors que $H^{-1}A$ est unitaire : $UU^* = H^{-1}AA^*(H^{-1})^* = H^{-1}H^2(H^{-1})^*$, or H est hermitienne donc $UU^* = I$. L'unicité de H implique celle de U.

Remarque 7.

Si A est non inversible, on a existence de H et U (il suffit de regarder $A_\varepsilon = A + \varepsilon I$). Mais on n'a pas l'unicité.

Remarque 8.

En dim 1, pour $z \in \mathbb{C}^*$, $\exists!\rho > 0$ et u tels que $|u| = 1$ et $z = \rho u$. Il existe un unique $\theta \in [0, 2\pi[$ tel que $u = e^{i\theta}$. D'où le nom de la décomposition.

1.2.2 Décompositions LU, LDU, de Cholesky

Théorème 10. (Factorisation LU)

Soit une matrice $A \in \mathcal{M}_m(\mathbb{K})$ dont toutes les sous-matrices d'ordre $k \in \{1, \cdots, m\}$, de la forme $(A_{ij})_{1\leq i,j\leq k}$, sont inversibles. Il existe un unique couple de matrices (L, U), avec U triangulaire supérieure, et L triangulaire inférieure à diagonale unité (i.e. $l_{ii} = 1$), tel que

$$A = LU.$$

Démonstration : Par récurrence sur m.
Si $m = 1$, évident.
Supposons $A \in \mathcal{M}_m(\mathbb{K})$, $m > 1$ et la propriété vraie au rang $m - 1$. On écrit alors

$$\begin{pmatrix} \widetilde{A} & X \\ Y^T & a \end{pmatrix}$$

avec $a \in \mathbb{K}$, $X \in \mathbb{K}^{m-1}$, $Y \in \mathbb{K}^{m-1}$ et $\widetilde{A} \in \mathcal{M}_{m-1}(\mathbb{K})$.
Par hypothèse de récurrence, $\exists(\widetilde{L}, \widetilde{U})$ tel que $\widetilde{A} = \widetilde{L}\widetilde{U}$ avec $\widetilde{U}$ triangulaire supérieure et $\widetilde{L}$ triangulaire inférieure à diaginale unité.
Posons alors

$$L = \begin{pmatrix} \widetilde{L} & 0 \\ l^T & 1 \end{pmatrix} \qquad U = \begin{pmatrix} \widetilde{U} & z \\ 0 & u \end{pmatrix}$$

où $l, z \in \mathbb{K}^{m-1}$ et $u \in \mathbb{K}$ sont à déterminer. En identifiant, on trouve qu'il faut et qu'il suffit de prendre l, z et u tels que :

- $\widetilde{L}\widetilde{U} = \widetilde{A}$ déjà vérifié.
- $l^T\widetilde{U} = Y^T$ cad $l^T = Y^T\widetilde{U}^{-1}$ possible car $\widetilde{U}$ inversible par hypothèse sur $\widetilde{A}$ sous-matrice diagonale de A.
- $\widetilde{L}z = X$ cad $z = \widetilde{L}^{-1}X$ possible car $\widetilde{L}$ inversible par hypothèse sur $\widetilde{A}$ sous-matrice diagonale de A.
- $l^Tz + u = a$ cad $u = a - l^Tz$.

Remarque 9.

On voit que l'unicité provient du fait que l'on impose une diagonale unité à L. On verra un algorithme efficace de détermination de L et U. La technique qui apparaît dans cette démonstration n'est pas optimale.

Remarque 10.

$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ n'admet pas de décomposition LU.

Corollaire 1. (Factorisation LDU)

Soit une matrice $A \in \mathcal{M}_m(\mathbb{K})$ dont toutes les sous-matrices d'ordre $k \in \{1, \cdots, m\}$, de la forme $(A_{ij})_{1 \leq i,j \leq k}$, sont inversibles. Il existe un unique triplet de matrices (L, D, U), avec U triangulaire supérieure à diagonale unité (i.e. $u_{ii} = 1$), L triangulaire inférieure à diagonale unité (i.e. $l_{ii} = 1$), et D diagonale, tel que

$$A = LDU.$$

Démonstration : Soit (L_0, U_0) les matrices associées à la décomposition LU de A. On pose alors $L = L_0$, $D = diag(U_0)$ et $U = D^{-1}U_0$. Alors L, D et U satisfont les propriétés souhaitées. Leur unicité provient de l'unicité de la décomposition LU.

Remarque 11.

D ne contient pas en général les valeurs propres de A. (i.e. $L \neq U^{-1}$). Il ne s'agit pas d'une diagonalisation.

Remarque 12.

Ces théorèmes s'appliquent aux matrices définies positives car leurs mineurs fondamentaux sont non nuls (i.e. les déterminants des sous-matrices diagonales sont non nuls).

Corollaire 2. (Factorisation de Cholesky)

Soit A une matrice symétrique réelle, définie positive. Aors il existe une unique matrice réelle B triangulaire inférieure, telle que tous ces éléments diagonaux soient positifs, et qui vérifie

$$A = BB^T.$$

Démonstration : Soit A symétrique définie positive $\in \mathcal{M}_m(\mathbb{R})$.
On sait alors qu'il existe un unique (L, D, U) tel que L triangulaire inférieure à diagonale unité, U triangulaire supérieure à diagonale unité et D diagonale.
• Puisque A est réelle symétrique, on a $A = A^* = U^*D^*L^*$ avec U^* triangulaire inférieure à diagonale unité, L^* triangulaire supérieure à diagonale unité et D^* diagonale.
Puisque la décomposition LDU de A est unique, on en déduit : $L = U^*$ et $D = D^*$. Montrons alors que D est définie positive : $(Ax, x) = (LDUx, x) = (DUx, L^*x) = (DL^*x, L^*x)$. Or A est définie positive et L inversible, donc pour tout y, il existe x tel que $y = L^*x$ et donc $(Dy, y) > 0$ si $y \neq 0$.
On peut alors prendre $D_0 = D^{1/2}$, puis poser $B = LD_0$ et $C = D_0U$. Ainsi B est triangulaire inférieure de diagonale $diag(D_0)$ et C est triangulaire supérieure de diagonale $diag(D_0)$ aussi et telles que $A = BC$.
• Il reste à montrer que $C = B^T$. Pour cela, on écrit d'abord $A = A^T$. Donc ${C^T}^{-1}B = B^T C^{-1}$. On vérifie aisément que ${C^T}^{-1}B$ est triangulaire inférieure à diagonale unité et que $B^T C^{-1}$ est triangulaire supérieure à diagonale unité. Puisqu'elles sont égales, on en déduit que ${C^T}^{-1}B = B^T C^{-1} = I$. D'où $C = B^T$.
Cette démonstration utilise un raisonnement qu'il est bon de savoir utiliser mais on peut faire plus simple ici : La démonstration de $C = B^T$ découle rapidement de $L = U^*$.
• Démontrons maintenant l'unicité : Soient B_1 et B_2 réelles triangulaires inférieures telles que $B_1B_1^T = B_2B_2^T$. Alors $B_2^{-1}B_1 = B_2^T{B_1^T}^{-1}$. L'une est réelle triangulaire inférieure et l'autre réelle triangulaire supérieure. Elles sont donc diagonales. Et il existe D une matrice réelle diagonale telle que $B_2^{-1}B_1 = D$.
Puisque $A = B_1B_1^T$, on a alors $A = B_2D(B_2D)^T = B_2DD^TB_2^T = B_2D^2B_2^T$. Or $A = B_2B_2^T$, donc $D = I$. On en déduit $B_1 = B_2$.

Remarque 13.

Pour des matrices symétriques définies positives, on a $|B_{ij}|^2 \leq A_{ii}$ pour tout (i,j).

Remarque 14.

Si une matrice A admet une décomposition de Cholesky, alors A est symétrique définie positive. Cela est un moyen algorithmique numérique de vérifier qu'une matrice est symétrique définie positive. En effet : $(BB^T)^T = BB^T$ et $(BB^Tx, x) = (B^Tx, B^Tx) = \left\|B^Tx\right\| \geq 0 (= 0 \text{ ssi } x = 0)$.

Remarque 15. Intérêt des matrices triangulaires ou diagonales

Les matrices diagonales sont inversibles de manière immédiate.
Les systèmes associés à des matrices triangulaires se résolvent rapidement par des techniques dites de remontée ou de descente suivant que la matrice est triangulaire supérieure ou inférieure.

1.2.3 Décomposition QR

Théorème 11. (Factorisation QR)

Soit A une matrice réelle inversible. Il existe un unique couple de matrices (Q,R), où Q est une matrice unitaire ($Q^{-1} = Q^*$), et R une matrice triangulaire supérieure dont tous les éléments diagonaux sont positifs, tel que

$$A = QR.$$

Démonstration : A est inversible donc ses vecteurs colonnes $\{c_1, \cdots, c_m\}$ constituent une base de $\mathbb{R}^m$. Par orthonormalisation de Gram-Schmidt, on construit $\{u_1, \cdots, u_m\}$ tel que pour tout $k \leq m$, $\text{Vect}\{u_1, \cdots, u_k\} = \text{Vect}\{c_1, \cdots, c_k\}$. Notons R la matrice de passage de $\{u_1, \cdots, u_m\}$ à $\{c_1, \cdots, c_m\}$, Q celle de passage de $\{e_1, \cdots, e_m\}$ à $\{u_1, \cdots, u_m\}$.
Alors, $A = QR$; par le procédé de Gram-Schmidt, R est triangulaire supérieure ; Q est unitaire puisque $\{u_1, \cdots, u_m\}$ est orthonormée. Il reste à assurer que R est à diagonale positive. Ceci vient du procédé d'orthonormalisation où l'on peut choisir un vecteur ou son opposé de façon à imposer une composante positive dans la direction voulue. Cette condition joue un rôle essentiel dans l'unicité de la décomposition :
Supposons qu'il existe deux décompositions QR : (Q_1, R_1) et (Q_2, R_2). Alors $R_2R_1^{-1}$ est orthogonale (car égale à $Q_1Q_2^{-1}$). Elle est aussi triangulaire supérieure. Comme toute matrice triangulaire orthogonale est égale à l'unité, on en déduit que $R_1 = R_2$ et ainsi $Q_1 = Q_2$.

Remarque 16.

Le procédé de Gram-Schmidt donne un premier exemple d'algorithme de factorisation LU mais on en verra un plus intéressant car plus stable numériquement (par rapport à la propagation des erreurs d'arrondi).

Corollaire 3.

Soit $A \in \mathcal{M}_{m,n}(\mathbb{R})$ avec $m \geq n$ et $rang(A) = n$; alors il existe (Q,R) tel que $Q \in \mathcal{M}_m(\mathbb{R})$ orthogonale et $R \in \mathcal{M}_{m,n}(\mathbb{R})$ triangulaire supérieure avec $A = QR$.

Démonstration : Appliquer la décomposition QR à la matrice $\widetilde{A} = \begin{pmatrix} A & B \end{pmatrix}$ où B complète A tel que $\widetilde{A} \in \mathcal{M}_m(\mathbb{R})$ inversible. On alors $(\widetilde{Q}, \widetilde{R})$ avec $\widetilde{R} = \begin{pmatrix} R_1 & R_2 \end{pmatrix}$ tel que $R_1 \in \mathcal{M}_{m,n}(\mathbb{R})$. Il suffit alors de prendre $Q = \widetilde{Q}$ et $R = R_1$. On remarque que, le choix de B n'étant pas unique, la décomposition QR n'est pas unique dans ce cas.

1.2.4 Décomposition en valeurs singulières

Définition 20.

Les valeurs singulières de $A \in \mathcal{M}_{m,n}(\mathbb{C})$ sont les racines carrées positives des valeurs propres de A^*A (qui est hermitienne positive).

Théorème 12. (Décomposition SVD)

Soit $A \in \mathcal{M}_{m,n}(\mathbb{C})$ de rang k. Elle admet alors k valeurs singulières strictement positives : $\sigma_1 \geq \cdots \geq \sigma_k > 0$. Et il existe $U \in \mathcal{M}_m(\mathbb{C})$ unitaire et $V \in \mathcal{M}_n(\mathbb{C})$ unitaire telles que $A = USV^*$ avec $S \in \mathcal{M}_{m,n}(\mathbb{C})$ de la forme :

$$\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{où } D = diag(\sigma_1, \cdots, \sigma_k)$$

Démonstration :

$x \in \ker(A^*A) \Longrightarrow x^*A^*Ax = 0 \Longrightarrow \|Ax\| = 0 \Longrightarrow x \in \ker(A)$.
$x \in \ker(A) \Longrightarrow Ax = 0 \Longrightarrow A^*Ax = 0 \Longrightarrow x \in \ker(A^*A)$.
Ainsi, $\ker(A^*A) = \ker(A)$ et donc A^*A a exactement k valeurs propres strictement positives non nécessairement distinctes et $(n-k)$ valeurs propres nulles. Notons σ_i^2, $i = 1, \cdots, k$ ces valeurs propres. Soit $\{v_1, \cdots, v_k\}$ une famille orthonormée de vecteurs propres de A^*A associés aux σ_i^2.
Soit $u_i = \sigma_i^{-1} A v_i$.
Alors $AA^*u_i = \sigma_i^{-1} AA^*Av_i = \sigma_i^{-1}\sigma_i^2 Av_i = \sigma_i^2 u_i$.
et $(u_i, u_j) = (\sigma_i^{-1}Av_i, \sigma_j^{-1}Av_j) = \sigma_i^{-1}\sigma_j^{-1}\sigma_i^2(v_i, v_j) = \delta_{ij}$.
Ainsi, les u_i sont des valeurs propres de AA^* associés aux σ_i^2 qui constituent une famille orthonormale. On complète les 2 familles par des vecteurs pour constituer des familles orthonormales :
$u_1, \cdots, u_k$ complétée par $u_{k+1}, \cdots, u_m \longrightarrow U$ unitaire.
$v_1, \cdots, v_k$ complétée par $v_{k+1}, \cdots, v_n \longrightarrow V$ unitaire.
Regardons enfin U^*AV :

$$U^*AV = U^* \begin{pmatrix} \sigma_1 u_1 & \cdots & \sigma_k u_k & 0 & \cdots & 0 \end{pmatrix}$$

car $Av_i = \sigma_i u_i$ par définition des u_i. Ainsi,

$$U^*AV = \begin{pmatrix} \sigma_1 & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & \sigma_k & \ddots & & \vdots \\ \vdots & & \ddots & 0 & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & 0 \end{pmatrix}$$

puisque U est unitaire.

1.3 Théorie Spectrale

Cette section a pour but de donner quelques caractérisations des valeurs propres d'une matrice. On s'intéresse dans un premier temps à des matrices hermitiennes.

Définition 21.

Soit $A \in \mathcal{M}_m(\mathbb{C})$ une matrice hermitienne et $x \in \mathbb{C}^m \setminus \{0\}$. On note $r_A(x)$ le nombre $\frac{(Ax,x)}{(x,x)}$ appelé quotient de Rayleigh-Ritz.

Théorème 13. (Rayleigh - Ritz)

Soit $A \in \mathcal{M}_m(\mathbb{C})$ hermitienne dont les valeurs propres sont ordonnées :

$$\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m = \lambda_{\max}$$

Alors :

$$\lambda_{\max} = \max_{x \neq 0} r_A(x) \qquad \text{et} \qquad \lambda_{\min} = \min_{x \neq 0} r_A(x).$$

Remarque 17.

$x = \|x\|\,\hat{x}$ avec $\hat{x}$ de norme 1. Et ainsi, $\frac{(Ax,x)}{(x,x)} = (A\hat{x},\hat{x})$.

On en déduit que $\sup_{x\in\mathbb{C}^m, x\neq 0} \frac{(Ax,x)}{(x,x)} = \sup_{x\in\mathbb{C}^m,\|x\|=1}(Ax,x)$. Or, $\{x \in \mathbb{C}^m, \|x\| = 1\}$ est compact. Le sup et l'inf sont alors atteints.

Démonstration du théorème : $\exists U$ unitaire telle que $A = UDU^*$ et $D = diag(\lambda_1, \cdots, \lambda_m)$. Alors on voit facilement que $r_A(x) = r_D(y)$ avec $y = U^*x$.
D'autre part,

$$\lambda_{\min}\sum_{i=1}^{m}|y_i|^2 \leq r_D(y) = \sum_{i=1}^{m}\lambda_i|y_i|^2 \leq \lambda_{\max}\sum_{i=1}^{m}|y_i|^2.$$

Si y vérifie $\|y\| = 1$ alors $\lambda_{min} \leq r_D(y) \leq \lambda_{max}$ et λ_{min} est atteint pour $y = (1, 0, \cdots, 0)$ et λ_{max} est atteint pour $y = (0, \cdots, 0, 1)$.
Enfin, puisque U est unitaire, $\max_{\|x\|=1} r_A(x) = \max_{\|y\|=1} r_D(y)$.

Corollaire 4.

Soit $A \in \mathcal{M}_m(\mathbb{C})$ une matrice hermitienne et $x \in \mathbb{C}^m$. On pose $\alpha = \frac{(Ax,x)}{(x,x)} \in \mathbb{R}$. Alors

$$\sigma(A)\cap]-\infty,\alpha] \neq \emptyset \qquad \textit{et} \qquad \sigma(A)\cap[\alpha,+\infty[\neq\emptyset\,.$$

Théorème 14. (Caractérisation min-max de Courant-Fisher)

Caractérisation de toutes les valeurs propres :
Soit $A \in \mathcal{M}_m(\mathbb{C})$ hermitienne de valeurs propres $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$.
Soit $k \in \{1, \cdots, m\}$. Alors :

$$\min_{w_1,\cdots,w_{m-k}\in\mathbb{C}^m} \quad \max_{x\neq 0,\, x\perp w_1,\cdots,w_{m-k}} r_A(x) = \lambda_k\,.$$

$$\max_{w_1,\cdots,w_{k-1}\in\mathbb{C}^m} \quad \min_{x\neq 0,\, x\perp w_1,\cdots,w_{k-1}} r_A(x) = \lambda_k\,.$$

Démonstration : On diagonalise $A = UDU^*$ avec $D = diag(\lambda_1, \cdots, \lambda_m)$.
Alors

$$\begin{aligned}
\max_{\substack{x\neq 0\\ x\perp w_1,\cdots,w_{m-k}}} r_D(U^*x) &= \max_{\substack{\|y\|=1\\ y\perp(U^*w_j)_{j=1,\cdots,m-k}}} \sum_{i=1}^{m}\lambda_i|y_i|^2 \\
&\geq \max_{\substack{\|y\|=1\\ y\perp(U^*w_j)_{j=1,\cdots,m-k}\\ y_1=\cdots=y_{k-1}=0}} \sum_{i=1}^{m}\lambda_i|y_i|^2 \\
&= \max_{\substack{\|y\|=1\\ y\perp(U^*w_j)_{j=1,\cdots,m-k}\\ y_1=\cdots=y_{k-1}=0}} \sum_{i=k}^{m}\lambda_i|y_i|^2 \\
&\geq \lambda_k \max_{\substack{\|y\|=1\\ y\perp(U^*w_j)_{j=1,\cdots,m-k}\\ y_1=\cdots=y_{k-1}=0}} \sum_{i=k}^{m}|y_i|^2 = \lambda_k.
\end{aligned}$$

Ainsi

$$\min_{w_1,\cdots,w_{m-k}\in\mathbb{C}^m} \quad \max_{\substack{\|y\|=1\\ y\perp(U^*w_j)_{j=1,\cdots,m-k}}} r_D(y) \geq \lambda_k\,.$$

Regardons le cas particulier $w_j = U_{m-j+1}$, pour $j = 1, \cdots, m-k$:

$$\min_{w_1,\cdots,w_{m-k}\in\mathbb{C}^m} \max_{\substack{\|y\|=1 \\ y\perp(U^*w_j)_{j=1,\cdots,m-k}}} r_D(y) \;\leq\; \max_{\substack{\|y\|=1 \\ y\perp(U^*U_{m-j+1})_{j=1,\cdots,m-k}}} r_D(y) \;=\; \max_{\substack{\|y\|=1 \\ y_{k+1}=\cdots=y_m=0}} \sum_{i=k}^{m} \lambda_i |y_i|^2 = \lambda_k.$$

En effet, $\{U^*U_{m-j+1}\}_{j=1,\cdots,m-k} = \{e_{m-j+1}\}_{j=1,\cdots,m-k} = \{e_l\}_{l=k+1,\cdots,m}$.

Remarque 18.

On écrit aussi :

$$\min_{S\subset\mathbb{C}^m\,;\,\dim S=k} \max_{x\neq 0,\, x\in S} r_A(x) = \lambda_k\,.$$

On peut désormais établir un premier résultat de perturbation :

Théorème 15. (Weyl)

Soient $A, B \in \mathcal{M}_m(\mathbb{C})$ hermitiennes et $\lambda_i(A)$, $\lambda_i(B)$, $\lambda_i(A+B)$ rangées dans l'ordre croissant. Alors, pour tout $k = 1, \cdots, m$:

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_m(B).$$

Conséquence : Si B a des valeurs propres petites devant celles de A, la matrice $A+B$ a des valeurs propres proches de celles de A.

Démonstration : Pour tout $x \in \mathbb{C}^d$, $\lambda_1(B) \leq r_B(x) \leq \lambda_m(B)$.
Ainsi

$$\begin{aligned}
\lambda_k(A+B) &= \min_{\dim S=k} \max_{x\neq 0\,;\,x\in S} r_{A+B}(x) \\
&= \min_{\dim S=k} \max_{x\neq 0\,;\,x\in S} (r_A(x) + r_B(x)) \\
&\geq \min_{\dim S=k} \max_{x\neq 0\,;\,x\in S} (r_A(x) + \lambda_1(B)) = \lambda_k(A) + \lambda_1(B).
\end{aligned}$$

La démonstration de l'autre inégalité se fait à l'aide de l'autre caractérisation.

Corollaire 5.

Soient $A, E \in \mathcal{M}_m(\mathbb{C})$ hermitiennes avec $\|E\| \leq \varepsilon$ où $\|\cdot\|$ est une norme subordonnée. Alors $|\lambda_k(A+E) - \lambda_k(A)| \leq \varepsilon$.

Intéressons nous maintenant à des perturbations non diagonales, cad au cas non hermitien.

Théorème 16. (Gershgorin)

Soit $A \in \mathcal{M}_m(\mathbb{C})$ et λ une valeur propre de A. Alors

$$\lambda \in \cup_{i=1}^m D_i \qquad \textit{où} \qquad D_i = \{z \in \mathbb{C}\,;\; |z - A_{ii}| \leq \sum_{j\neq i} |A_{ij}|\}.$$

Remarque 19.

Les D_i sont appelés disques de Gershgorin.

Démonstration du théorème : Soit $x \in \mathbb{C}^m$ tel que $Ax = \lambda x$ et $|x_i| = \max_j(|x_j|)$.
Alors $\lambda x_i = \sum_{j=1}^m A_{ij}x_j$ d'où $(\lambda - A_{ii})x_i = \sum_{j\neq i} A_{ij}x_j$.
Ainsi $|\lambda - A_{ii}||x_i| \leq \sum_{j\neq i} |A_{ij}||x_j|$.

Corollaire 6.

Soit $A \in \mathcal{M}_m(\mathbb{C})$ à diagonale strictement dominante, alors A est inversible.

Démonstration : A non inversible $\iff$ $\lambda = 0$ valeur propre de $A \implies$ il existe i tel que $|A_{ii}| \leq \sum_{j\neq i} |A_{ij}|$, ce qui contredit l'hypothèse sur A.

Remarque 20.

En utilisant A^T, on a le même type de résultat sur les colonnes :

$$\lambda \in \cup_{i=1}^{m} \{z \in \mathbb{C}\,;\ |z - A_{ii}| \leq \sum_{j \neq i} |A_{ji}|\}.$$

Théorème 17. (Gershgorin précisé)

Si l'union de k des m disques de Gershgorin est un ensemble disjoint des autres $m-k$ disques, alors cette union contient exactement k valeurs propres.

Démonstration : Posons $A = D + B$ avec $D = diag(A)$ et $A_\varepsilon = D + \varepsilon B$. Supposons que les k disques sont les k premiers pour simplifier l'écriture. On pose alors

$$G_k(\varepsilon) = \cup_{i=1}^{k} \{z \in \mathbb{C}\,;\ |z - A_{ii}| \leq \sum_{j \neq i} \varepsilon |A_{ij}|\}\ .$$

Pour $0 \leq \varepsilon \leq 1$, on a $G_k(\varepsilon) \subset G_k(1)$. D'autre part, pour tout $i \leq k$, on a $\lambda_i(A_0) \in G_k(0) \subset G_k(1)$. Etant donné que $\lambda_i(A_0)$ et $\lambda_i(A_1)$ sont jointes continûment à l'intérieur de $G_k(1)$, celui-ci contient au moins k valeurs propres. De même pour les autres sous ensembles. D'où la conclusion. En fait, $G_k(0) = \cup_{i=1}^{k} \{\lambda_i\}$.

Corollaire 7.

Soit D diagonale et E telle que $\|E\|_\infty \leq \varepsilon$, alors

$$\forall \lambda_\varepsilon \in \sigma(D+E),\ \exists \lambda \in \sigma(D)\ \textit{tel que}\ |\lambda_\varepsilon - \lambda| \leq \varepsilon$$

Démonstration :

$$\lambda_\varepsilon \in \cup_{i=1}^{m} \{z \in \mathbb{C}\,;\ |z - \lambda_i - E_{ii}| \leq \sum_{j \neq i} |E_{ji}|\} \subset \cup_{i=1}^{m} \{z \in \mathbb{C}\,;\ |z - \lambda_i| \leq \sum_{j} |E_{ji}|\}\,.$$

Attention : Ce résultat ne s'étend pas au cas de matrices quelconques mais il s'étend au cas des matrices diagonalisables. On le connaissait déjà dans le cas où D et E sont toutes les deux hermitiennes.

Proposition 4.

Soit $A \in \mathcal{M}_m(\mathbb{C})$ diagonalisable avec $A = PDP^{-1}$. Soit $E \in \mathcal{M}_m(\mathbb{C})$ telle que $\|E\|_\infty \leq \varepsilon$. Alors,

$$\forall \lambda_\varepsilon \in \sigma(A+E),\ \exists \lambda \in \sigma(A)\ \textit{tel que}\ |\lambda_\varepsilon - \lambda| \leq K(P)\varepsilon$$

où $K(P) = \|P\|_\infty \left\|P^{-1}\right\|_\infty$ est le conditionnement en norme infinie de P.

Démonstration : $A+E = P(D+P^{-1}EP)P^{-1}$. Il suffit alors de considérer les valeurs propres de $D + P^{-1}EP$. Puis, $\|\cdot\|_\infty$ étant matricielle, $\left\|P^{-1}EP\right\|_\infty \leq K(P)\varepsilon$.

Remarque 21.

On a écrit la proposition précédente pour la norme infinie mais elle est vraie pour toute norme $\|\cdot\|$ vérifiant que $\|\Lambda\| = \max_i(\Lambda_{ii})$ pour toute matrice Λ diagonale. Elle est donc en particulier vraie pour toutes les matrices usuelles $\|\cdot\|_p$.

Corollaire 8.

Si A est normale et E telle que $\|E\|_2 \leq \varepsilon$, alors

$$\exists \lambda_\varepsilon \in \sigma(A+E),\ \exists \lambda \in \sigma(A)\ \textit{tels que}\ |\lambda_\varepsilon - \lambda| \leq \varepsilon\,.$$

Théorème 18.

Soit $A \in \mathcal{M}_m(\mathbb{C})$ diagonalisable avec $A = PDP^{-1}$; $D = diag(\lambda_i)$.
Soit $\hat{x} \in \mathbb{C}^m$ et $\hat{\lambda} \in \mathbb{C}$. On pose $r = A\hat{x} - \hat{\lambda}\hat{x}$. Alors

$$\exists \lambda \in \sigma(A) \text{ tel que } |\hat{\lambda} - \lambda| \leq K(P)\frac{\|r\|_p}{\|\hat{x}\|_p}.$$

Démonstration : Supposons $\hat{\lambda} \notin \sigma(A)$ (sinon, c'est évident).
Alors $r = P(D - \hat{\lambda}I)P^{-1}\hat{x}$. Alors,

$$\begin{aligned} \|\hat{x}\| \;=\; \left\|P(D-\hat{\lambda}I)^{-1}P^{-1}r\right\| \;&\leq\; K(P)\left\|(D-\hat{\lambda}I)^{-1}\right\|\|r\| \\ &\leq\; K(P)\|r\|\left(\min_{\lambda\in\sigma(A)}|\lambda-\hat{\lambda}|\right)^{-1} \end{aligned}$$

Remarque 22.

On a vu un certain nombre de résultats qui donnent des estimations des valeurs propres d'une matrice perturbée. Le cas des vecteurs propres est plus délicat.
Soit $A_\varepsilon = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$. Ses valeurs propres sont $1 \pm \varepsilon$ qui sont proches de 1, la valeur propre double de A_0. Mais les vecteurs propres de A_ε sont $(1\ 1)^T$ et $(-1\ 1)^T$. Les vecteurs propres de A_0 ne permettraient pas d'avoir une estimation de ceux de A_ε.
Si on prend $\hat{\lambda} = 1$ et $\hat{x} = (1\ 0)^T$, on a $r = (0\ \varepsilon)^T$. On peut en déduire que $\hat{\lambda}$ est proche de $\lambda(A_\varepsilon)$ mais on voit bien qu'il serait faux de penser qu'un vecteur propre de A est proche de $\hat{x}$.

Remarque 23.

Soit $A_\varepsilon = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$. On alors $A_\varepsilon \longrightarrow A_0$ quand $\varepsilon \to 0$, $\sigma(A_\varepsilon) = \{\pm\sqrt{\varepsilon}\}$ et $\sigma(A_0) = \{0\}$.
Dans ce cas, $A_0 = A_\varepsilon + E$ avec $\|E\|_\infty = \varepsilon$. Mais $\exists C$ tel que $|\lambda_\varepsilon - \lambda| \leq C\|E\|_\infty$ équivaudrait à dire qu'il existe une constante C telle que $\sqrt{\varepsilon} \leq \varepsilon\,,\ \forall \varepsilon > 0$. Ceci est FAUX. Le résultat n'est pas vrai dans ce cas. L'hypothèse non vérifiée est le fait que A_0 ne soit pas diagonalisable.
On peut aussi regarder de plus près A_ε. Elle est diagonalisable, de matrice de passage $P_\varepsilon = \begin{pmatrix} 1 & 1 \\ \sqrt{\varepsilon} & -\sqrt{\varepsilon} \end{pmatrix}$, avec $\|P_\varepsilon\|_\infty = 2$. Et $P_\varepsilon^{-1} = \frac{1}{-2\sqrt{\varepsilon}}\begin{pmatrix} -\sqrt{\varepsilon} & -1 \\ -\sqrt{\varepsilon} & 1 \end{pmatrix}$, avec $\left\|P_\varepsilon^{-1}\right\|_\infty = \frac{1+\sqrt{\varepsilon}}{2\sqrt{\varepsilon}}$. Ainsi $K(P)$ tend vers l'infini quand ε tend vers 0.

1.4 Applications

1.4.1 Imagerie numérique

L'imagerie numérique a longtemps été un exemple d'application de la SVD. Maintenant, il existe des moyens plus élaborés et plus efficaces. Regardons ici l'utilisation de la SVD pour le transfert d'image. Une photo en noir et blanc peut être assimilée à une matrice A dont les éléments correspondent à des niveaux de gris pour les différents pixels qui constituent l'image. Chaque A_{ij} donne le niveau de gris du pixel (i, j) de l'image. $A_{ij} \in [0, 1]$, avec par exemple $A_{ij} = 0$ pour un pixel (i, j) blanc et $A_{ij} = 1$ pour un pixel (i, j) noir.

Dans le livre de Allaire et Kaber, les auteurs présentent le cas d'un portrait noir et blanc avec une image originale de taille 558×768 pixels dont la matrice A représentative est générée par un logiciel de lecture de l'image (via un scanner ou un appareil photo numérique). Le résultat renvoyé correspond à une matrice A de rang 555. La matrice admet donc 555 valeurs singulières non nulles. Soit $\sigma_{\max}$ la valeur maximale de ces valeurs singulières. Soient $\sigma_1, \cdots, \sigma_{555}$ ces valeurs singulières. On sait alors qu'il existe $U \in \mathcal{M}_{558}(\mathbb{C})$, $V \in \mathcal{M}_{768}(\mathbb{C})$ tels que $A = U[diag(\sigma_1, \cdots, \sigma_{555}, 0, 0, 0)\ \ 0_{\mathbb{C}^{558\times 210}}]V$.

En désignant par u_i les vecteurs colonnes de U et par v_i^* les vecteurs lignes de V, on obtient que $A = \sum_{i=1}^{555} \sigma_i u_i v_i^*$.

La connaissance de A elle-même équivaut à la connaissance de 558x768 nombres (i.e. 428544 nombres). Si on regarde l'expression de droite, la connaissance de A est donnée par la connaissance de 555 valeurs singulières, plus 555 vecteurs de taille 558, plus 555 vecteurs de taille 768. On n'y est pas gagnant.

Par contre, si on regarde de plus près les valeurs singulières non nulles $\sigma_1, \cdots, \sigma_{555}$, on constate que nombreuses d'entre elles sont très petites devant les autres. On peut alors décider que si $\frac{\sigma_i}{\sigma_{\max}} < \varepsilon$ (où ε est un filtre choisi par l'utilisateur, exple $\varepsilon = 10^{-3}$), alors σ_i est considérée comme nulle. On réduit ainsi le nombre de σ_i à conserver dans l'expression de A selon la décomposition SVD.

Notons $\sigma_1, \cdots, \sigma_k$ les valeurs singulières ainsi conservées. Dans l'exemple proposé dans le livre de Allaire et Kaber, avec $k = 10$, l'image est floue mais on constate aisément qu'il s'agit d'un portrait. Pour $k = 40$, l'image est déjà ressemblante et l'individu reconnaissable. Pour $k = 80$, il est impossible de dire, de l'image obtenue ou de l'image originale, laquelle est la plus vraie. Avec $k = 80$, on a alors $A \approx \sum_{i=1}^{80} \sigma_i u_i v_i^*$. La connaissance de A est alors donnée par la connaissance de 80 valeurs singulières, plus 80 vecteurs de taille 558, plus 80 vecteurs de taille 768, soit un total de 106160 nombres.

1.4.2 La corde vibrante

Déflection d'une corde tenue en ses deux extrémités ($x = 0$ et $x = 1$). On note $u(t, x)$ sa déflection en son point $x \in [0, 1]$, au temps t.

On se place dans le cas où la corde n'est pas soumise à des forces extérieures. Son évolution dépend donc de sa position au temps initial $u(0, x)$. Elle vérifie le problème aux limites :

$$\begin{cases} m\frac{\partial^2 u}{\partial t^2}(t,x) - k\frac{\partial^2 u}{\partial x^2}(t,x) = 0 \\ u(t,0) = 0 \quad , \qquad u(t,1) = 0 \end{cases}$$

où k et m désignent respectivement la raideur et la masse linéique de la corde que l'on suppose constantes le long de la corde.

On cherche des solutions périodiques en temps : $u(t, x) = v(x)e^{i\omega t}$ où ω est appelée fréquence de vibration de la corde et est une inconnue du problème.

On obtient le système en v :

$$\begin{cases} -v''(x) = \frac{m\omega^2}{k} v(x) \\ v(0) = 0 \quad , \qquad v(1) = 0 \end{cases}$$

Si k et m sont non constantes et variables selon x, il est en général impossible de trouver une solution exacte du problème. On "discrétise" alors l'équation différentielle. On va en fait chercher une approximation de la solution en un nombre fini de points $x_1, \cdots, x_n \in [0, 1]$. Prenons par exemple $x_i = \frac{i}{n}$. Soit v_i la valeur approchée de $v(x_i)$.

Ensuite, on approche la dérivée seconde de v via la formule de Taylor :

$$-v''(x_i) \approx \frac{2v(x_i) - v(x_{i-1}) - v(x_{i+1})}{n^{-2}}$$

On obtient de la sorte un système d'équations :

$$\begin{cases} \frac{2v_i - v_{i-1} - v_{i+1}}{n-2} = \lambda v_i \qquad \forall i \in \{1, n-1\} \\ v_0 = 0 \quad , \qquad v_n = 0 \end{cases}$$

ce qui s'écrit $A_n v = \lambda v$ avec $\lambda = \frac{m\omega^2}{k}$, $v = (v_1, \cdots, v_n)^T$ et

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

La détermination de la fréquence de vibration ω se ramène ainsi à un problème aux valeurs propres.

Chapitre 2

RÉSOLUTIONS NUMÉRIQUES

2.1 Introduction

Nous nous intéressons dans ce chapitre à des méthodes de résolution d'un système linéaire $Ax = b$ où A est une matrice donnée, b un vecteur donné et x l'inconnue du problème.

Sauf mention contraire, A est carrée inversible de taille $m \times m$.

Une première méthode d'inversion du système est la considération des formules de Cramer vues en L1/L2 :

$$x_i = \frac{\det\begin{pmatrix} a_1 & \cdots & a_{i-1} & b & a_{i+1} & \cdots & a_m \end{pmatrix}}{\det(A)}$$

En procédant par développement selon les lignes ou les colonnes, le nombre d'opérations pour calculer un déterminant d'ordre m est $\geq m!$. On doit ici calculer $(m+1)$ déterminants. Soyons généreux et considérons un seul déterminant dans le sens où les déterminants au numérateur diffèrent de celui au dénominateur d'une seule colonne et qu'il est donc possible de faire certaines choses une seule fois. Le coût d'une telle méthode est donc supérieur à $m!$.

Considérons que notre ordinateur calcule à 1 Gflops (cad 1 milliard d'opérations à la seconde – ce qui est déjà très bien). Considérons le cas très petit d'une matrice 50×50. Le temps de calcul serait, en années :

$$\frac{50!}{10^9 \times 60 \times 60 \times 24 \times 365}$$

$$= \frac{50}{60} \times \frac{49}{60} \times \frac{48}{24} \times \frac{47}{365} \times \frac{46 \times 45}{1000} \times \frac{44 \times 43}{1000} \times \frac{42 \times 41}{1000} \times 40! \approx 40! \approx 10^{48}.$$

Il s'agit bien d'un petit cas. Aujourd'hui, certaines techniques numériques nous permettent de considérer la résolution de systèmes pleins de taille $10^6 \times 10^6$. On appelle matrice pleine une matrice dont le nombre d'éléments nécessairement nuls n'est pas important. A contrario, on appelle matrice creuse, une matrice dont le nombre d'éléments nécessairement nuls est grand (exemple : les matrices tridiagonales comme celle intervenant dans la discrétisation du problème de la corde vibrante du chapitre précédent). Les matrices creuses de taille $10^6 \times 10^6$ sont assimilables à des objets de taille 10^6 alors que les matrices pleines de taille $10^6 \times 10^6$ sont assimilables à des objets de taille 10^{12}.

Les méthodes que nous allons étudier dans un premier temps sont des méthodes de résolution qui ont un coût de calcul en m^3. Avec la matrice précédente, de taille 50×50, le système se résout alors par ces techniques en quelques dix-millièmes de seconde.

Analyse numérique de la courbe de temps en fonction de m : Si on fait des tests avec plusieurs valeurs de m (10, 100, 1000, 10000, ...) et que l'on regarde la courbe du temps de calcul $t(m)$ en fonction de m, la largeur de la matrice, il est utile et recommandé de considérer des courbes en

échelle logarithmique lorsque le comportement est du type $t(m) = \alpha m^p$. Lorsque le comportement est de ce type, le paramêtre essentiel est l'exposant p. En échelle logarithmique, on a

$$\ln(t(m)) = p\ln(m) + \ln(\alpha)$$

On voit que $\ln(t(m))$ est une fonction affine de $\ln(m)$ et sa représentation est une droite dont la pente est exactement l'exposant p. Une approximation de celui-ci est donc tout simplement donnée par :

$$\frac{\ln(t(m_{max})) - \ln(t(m_{min}))}{\ln(m_{max}) - \ln(m_{min})}.$$

On peut utiliser cette technique pour étudier la convergence des méthodes itératives de recherche d'un zéro ou d'un point fixe d'une fonction comme on vous le propose en TP, avec la méthode de dichotomie.

2.2 Méthodes directes pour la résolution de systèmes linéaires carrés

2.2.1 Méthode de Gauss

On veut résoudre $Ax = b$. A est carrée de taille $m \times m$.
Idée : Se ramener à un système triangulaire par combinaisons successives des équations.

Théorème 19. (Théorème d'élimination de Gauss)
Soit A carrée inversible ou non. Il existe M inversible telle que $T = MA$ soit triangulaire supérieure.

Remarque 24.
Il s'agira ensuite de résoudre $Tx = Mb$ où M ne sera pas calculée explicitement mais Mb obtenu au fur et à mesure de la construction de T.

Démonstration du théorème : On va en fait construire $T = MA$ et Mb de manière itérative sur les lignes en construisant une suite de matrices : A^1, A^2, $\cdots$, jusqu'à $A^m = MA = T$. On donne par la même occasion l'algorithme de construction de T et Mb sans avoir en fait à connaître M à aucun moment.

Etape 1 : $A^1 = A$, on construit $\widetilde{A}^1 = P^1 A^1$ où P^1 est une matrice de permutation telle que $\widetilde{A}^1_{1,1} \neq 0$: Si $A^1_{1,1} \neq 0$, on prend $P^1 = I$. Sinon, si il existe i tel que $A^1_{i,1} \neq 0$, on permute les lignes i et 1. On fait subir le même traitement au second membre $b \to \widetilde{b}^1 = P^1 b$. $A^1_{i,1}$ devient $\widetilde{A}^1_{1,1}$ que l'on appelle pivot de $\widetilde{A}^1$.
Ensuite, on multiplie $\widetilde{A}^1$ par une matrice E^1 telle que $A^2 = E^1\widetilde{A}^1$ ne contienne que des zéros sur la première colonne à partir de la deuxième ligne. La i^{eme} ligne de A^2 est en fait obtenue par la combinaison des i^{eme} et 1^{ere} lignes de $\widetilde{A}^1$ qui fait apparaître 0 dans la première colonne ($A^2_{i,j} = \widetilde{A}^1_{i,j} - \frac{\widetilde{A}^1_{i,1}}{\widetilde{A}^1_{1,1}}\widetilde{A}^1_{1,j}$). On fait subir la même combinaison des lignes au second membre $\widetilde{b}^1 \to b^2 = E^1\widetilde{b}^1$.
La matrice E^1 est de la forme

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\frac{\widetilde{A}^1_{i,1}}{\widetilde{A}^1_{1,1}} & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & 0 \\ -\frac{\widetilde{A}^1_{m,1}}{\widetilde{A}^1_{1,1}} & 0 & 0 & 1 \end{pmatrix}$$

Etape k : A^k connue avec que des zéros en dessous de la diagonale sur les $(k-1)$ premières colonnes. On construit $\widetilde{A}^k = P^k A^k$ où P^k est une matrice de permutation telle que $\widetilde{A}^k_{k,k} \neq 0$: Si $A^k_{k,k} \neq 0$, on prend $P^k = I$. Sinon, si il existe $i > k$ tel que $A^k_{i,k} \neq 0$, on permute les lignes i et k. On fait subir le même traitement au second membre $b \to \widetilde{b}^k = P^k b$. $A^k_{i,k}$ devient $\widetilde{A}^k_{k,k}$ que l'on appelle pivot de $\widetilde{A}^k$.
La matrice P^k est définie par

$$\left|\begin{array}{l} P^k_{j,j} = 1 \text{ si } j \neq i \text{ et } j \neq k. \\ P^k_{k,i} = P^k_{i,k} = 1 \\ \text{Tous les autres éléments de la matrice sont nuls.} \end{array}\right.$$

Ensuite, on multiplie $\widetilde{A}^k$ par une matrice E^k telle que $A^{k+1} = E^k \widetilde{A}^k$ ne contienne que des zéros sur la k^{eme} colonne à partir de la $(k+1)^{eme}$ ligne. Pour $i > k$, la i^{eme} ligne de A^{k+1} est en fait obtenue par la combinaison des i^{eme} et k^{eme} lignes de $\widetilde{A}^k$ qui fait apparaître 0 dans la k^{eme} colonne ($A^k_{i,j} = \widetilde{A}^k_{i,j} - \frac{\widetilde{A}^k_{i,k}}{\widetilde{A}^k_{k,k}} \widetilde{A}^k_{k,j}$). On fait subir la même combinaison des lignes au second membre $\widetilde{b}^k \to b^{k+1} = E^k \widetilde{b}^k$.
La matrice E^k est définie par :

$$\left|\begin{array}{l} E^k_{j,j} = 1 \quad \forall j. \\ \text{La } k^{eme} \text{ colonne de } E \text{ est } \begin{pmatrix} 0 & \cdots & 0 & 1 & -\frac{\widetilde{A}^k_{k+1,k}}{\widetilde{A}^k_{k,k}} & \cdots & -\frac{\widetilde{A}^k_{n,k}}{\widetilde{A}^k_{k,k}} \end{pmatrix}^T \\ \text{Tous les autres éléments de la matrice sont nuls.} \end{array}\right.$$

Après l'étape $(n-1)$: On obtient A^n et b^n tels que $A^n x = b^n$ et A^n triangulaire surpérieure.

$$A^n = (E^{n-1}P^{n-1}\cdots E^1 P^1)A \qquad b^n = (E^{n-1}P^{n-1}\cdots E^1 P^1)b$$
$$M = E^{n-1}P^{n-1}\cdots E^1 P^1.$$

$\det(M) = \pm 1$ par définition des E^i et P^i.

Remarque 25.
Algorithmiquement : On ne calculera jamais les E^i ni les P^i. On se contente de faire subir des échanges et combinaisons de lignes à A et b de manière successive.

Remarque 26.
La résolution de $Ax = b$ se fait par la résolution du système triangulaire supérieur $A^n x = b^n$:

$$x_n = \frac{b^n_n}{a^n_{n,n}}$$
$$x_k = \frac{b^n_k - \sum_{l=k+1}^{n} a^n_{k,l} x_l}{a^n_{k,k}} \qquad k \text{ allant de } n-1 \text{ à } 1.$$

Notion de stabilité numérique

Les erreurs d'arrondis se propagent au fil des étapes. D'une étape à l'autre, les erreurs sont amplifiées par un coefficient multiplicatif, $\frac{\widetilde{A}^k_{i,k}}{\widetilde{A}^k_{k,k}}$. Par conséquent, pour limiter les erreurs, il convient de prendre pour pivot de $\widetilde{A}^k$, l'élément $A^k_{i,k}$, $i > k$ maximal en valeur absolue. On parle alors de pivot partiel : $\left|\widetilde{A}^k_{k,k}\right| = \max_{i \geq k} \left|A^k_{i,k}\right|$.

On peut aussi s'autoriser des permutations de colonnes pour le choix du pivot (Attention : Les échanges de colonnes de la matrice du système doivent s'accompagner d'échanges de lignes de l'inconnue et non du second membre.). On peut alors faire le choix du pivot total : $\left|\widetilde{A}^k_{k,k}\right| = \max_{i,j\geq k}\left|A^k_{i,j}\right|$.

2.2.2 Décomposition LU

Rappel : Si A est carrée et si tous ses mineurs fondamentaux sont non nuls, alors il existe un unique (L,U) tel que L soit triangulaire inférieure à diagonale unité, U triangulaire supérieure et $A = LU$.

Proposition 5.

La condition sur les mineurs fondamentaux implique que la méthode de Gauss est applicable sans échange de lignes ni de colonnes, cad sans pivot (i.e., on peut se permettre le choix : $P^{n-1} = \cdots = P^1 = I$).

Démonstration : Par récurrence sur l'étape, supposons la relation vraie à l'ordre $k-1$, alors $A = (E^1)^{-1}\cdots(E^{k-1})^{-1}A^k$. Il est facile de vérifier les identités remarquables :

$$E^k = \begin{pmatrix} 1 & 0 & \cdots & 0 & & \cdots & & 0 \\ 0 & \ddots & \ddots & \vdots & & & & \\ & \ddots & \ddots & 0 & & & & \vdots \\ \vdots & & 0 & 1 & & & & \\ & & & -l_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & & \\ & & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -l_{n,k} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

$$(E^k)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & & \cdots & & 0 \\ 0 & \ddots & \ddots & \vdots & & & & \\ & \ddots & \ddots & 0 & & & & \vdots \\ \vdots & & 0 & 1 & & & & \\ & & & +l_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & & \\ & & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & +l_{n,k} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

En notant $\mathcal{E}^k = (E^1)^{-1}\cdots(E^k)^{-1}$, alors :

$$\mathcal{E}^k = \begin{pmatrix} 1 & 0 & \cdots & 0 & & \cdots & & 0 \\ l_{2,1} & \ddots & \ddots & \vdots & & & & \\ & \ddots & \ddots & 0 & & & & \vdots \\ \vdots & & l_{k,k-1} & 1 & & & & \\ & & & l_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & & \\ & & & \vdots & \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \cdots & l_{n,k-1} & l_{n,k} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

On peut donc écrire la relation $\mathcal{E}^k A^k = A$ par bloc :

$$\left(\begin{array}{c|c} \mathcal{E}^k_{1,1} & 0 \\ \hline \star & Id_k \end{array}\right) \left(\begin{array}{c|c} A^k_{1,1} & A^k_{1,2} \\ \hline A^k_{2,1} & A^k_{2,2} \end{array}\right) = \left(\begin{array}{c|c} A_{1,1} & A_{1,2} \\ \hline A_{2,1} & A_{2,2} \end{array}\right)$$

où

$$\mathcal{E}^k_{1,1} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ \star & & 1 \end{pmatrix} \text{ et } Id_k = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

$A_{1,1}$ est un bloc de taille $k \times k$, et $A_{2,2}$ est un bloc de taille $(n-k) \times (n-k)$.
L'égalité par blocs nous donne donc la relation : $A_{1,1} = \mathcal{E}^k_{1,1} A^k_{1,1}$. Or $\mathcal{E}^k_{1,1}$ est une matrice triangulaire de déterminant égal à 1, et $A_{1,1}$ est de déterminant non nul par hypothèse sur les mineurs fondamentaux de A. On en déduit que le déterminant de $A^k_{1,1}$ est aussi non nul. Au regard de la matrice A^k, le bloc $A^k_{1,1}$ est triangulaire inférieur. Ses éléments diagonaux sont alors tous non nuls, ce qui assure que $a^k_{k,k}$ est non nul et que la méthode de Gauss est applicable sans pivot.

Théorème 20.
La décomposition LU correspond à une méthode de Gauss sans pivot.

Démonstration : La méthode de Gauss sans pivot renvoie $A = LU$ avec $U = A^n$ et $L = (E^1)^{-1} \cdots (E^{n-1})^{-1}$. U est bien triangulaire supérieure et L est bien triangulaire inférieure avec uniquement des 1 sur la diagonale.

Remarque 27.
$\det A = \det(U) = \prod_{i=1}^n u_{ii}$. Calcul efficace en $(n-1)$ produits, une fois U construite.

L'algorithme numérique

On peut stocker la partie supérieure de U dans celle de A et la partie strictement inférieure de L dans celle de A :

Pour $k = 1$ à $n-1$ (étape k)
  Pour $i = k+1$ à n (ligne i)
    $a_{i,k} = \frac{a_{i,k}}{a_{k,k}}$
    Pour $j = k+1$ à n (nouvelle colonne de L)
      $a_{i,j} = a_{i,j} - a_{i,k} a_{k,j}$
    Fin pour j
  Fin pour i
Fin pour k

Estimation du nombre d'opérations

Détermination de LU : $\sum_{k=1}^{n-1} \sum_{i=k+1}^{n} (1 + \sum_{j=k+1}^{n} 1)$, qui vaut asymptotiquement $\frac{n^3}{3}$.

Une fois L et U calculées, on a les coûts suivants :
– Résolution de $LUx = b$, par les résolutions de $Ly = b$ et $Ux = y$: $2 \times \frac{n^2}{2}$.
– Déterminant de LU ($\det(U) = \prod_{i=1}^n u_{ii}$) : n.
Ainsi, les coûts totaux sont :
– Déterminant de A ($\det(A) = \det(U) = \prod_{i=1}^n u_{ii}$) : $\frac{n^3}{3} + n$.
– Résolution de $Ax = b$: $\frac{n^3}{3} + n^2$.
– Inversion de A : $\frac{4n^3}{3} = \frac{n^3}{3} + n \times n^2$.

Calcul pratique de LU

L et U vérifient $l_{i,k} = 0$ pour $k > i$ et $u_{k,j} = 0$ pour $k > j$ et $A = LU$. Ainsi :

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k}u_{k,j} = \sum_{k=1}^{\min(i,j)} l_{i,k}u_{k,j}.$$

Etape 1 : On regarde la colonne 1 de A. On détermine ainsi les colonnes 1 de U et L :

$$a_{i,1} = l_{i,1}u_{1,1} \quad \forall i.$$

De $l_{1,1} = 1$, on déduit $u_{1,1} = a_{1,1}$ puis l'obtention de $(l_{i,1})_{i=2,\cdots,n}$ est immédiate : $l_{i,1} = \frac{a_{i,1}}{u_{1,1}}$.

Etape j : On connaît les colonnes 1 à $j-1$ de U et de L. On regarde la colonne j de A pour déterminer les colonnes j de U et L :

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k}u_{k,j} = \sum_{k=1}^{\min(i,j)} l_{i,k}u_{k,j}.$$

Ceci se récrit pour $i \leq j$,

$$\begin{aligned} a_{1,j} &= l_{1,1}u_{1,j} \\ a_{2,j} &= l_{2,1}u_{1,j} + l_{2,2}u_{2,j} \\ &\vdots \\ a_{j,j} &= l_{j,1}u_{1,j} + l_{j,2}u_{2,j} + \cdots + l_{j,j}u_{j,j} \end{aligned}$$

Ainsi, le vecteur $U_j = \begin{pmatrix} u_{1,j} & \cdots & u_{j,j} \end{pmatrix}^T$ est solution du système linéaire triangulaire inférieur $L^{(j)}U_j = A_j^{(1)}$ où $A_j^{(1)} = \begin{pmatrix} a_{1,j} & \cdots & a_{j,j} \end{pmatrix}^T$ et $L^{(j)}$ est la partie déjà connue de L :

$$L^{(j)} = \begin{pmatrix} l_{1,1} & & 0 \\ \vdots & \ddots & \\ l_{j,1} & \cdots & l_{j,j} \end{pmatrix}$$

En écrivant la relation pour $i > j$, on obtient les $l_{i,j}$:

$$l_{i,j} = \frac{1}{u_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} l_{i,k}u_{k,j} \right).$$

Ainsi, si on a déjà une fonction de résolution d'un système triangulaire inférieur, l'étape j du calcul de L et U se ramène à la résolution d'un système linéaire triangulaire inférieur de taille $j \times j$ et à un simple calcul vectoriel de taille $(n-j)$. Cela permet une programmation élégante et sûre de la méthode.

Matrices bandes

$A \in \mathcal{M}_n(\mathbb{R})$ est dite matrice bande, de demi-largeur de bande $p \in \mathbb{N}$ si $a_{i,j} = 0$ pour $|i-j| > p$, la largeur de la bande est alors $2p+1$.

Ainsi, $\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$ et $\begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & -1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$ ont une largeur de bande égale à 3.

Et la matrice $\begin{pmatrix} -5 & 2 & 0 & 0 \\ 1 & \ddots & \ddots & 0 \\ 1 & \ddots & \ddots & 2 \\ 0 & 1 & 1 & -5 \end{pmatrix}$ a une largeur de bande égale à 5.

Proposition 6.

La factorisation LU conserve la structure bande.

Si $A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$ alors la factorisation aboutira à des matrices de la forme :

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \star & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \star & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} \star & \star & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \star \\ 0 & 0 & 0 & \star \end{pmatrix}$$

Cela permet d'adapter les algorithmes de façon à être très efficace lorsque les matrices ne sont pas très remplies.

Démonstration : Par récurrence sur les éléments. Il suffit d'identifier A et LU selon le calcul pratique défini plus haut.

2.2.3 Méthode de Cholesky

Rappel : Soit A réelle symétrique définie positive. Alors, il existe une unique matrice B triangulaire inférieure à diagonale strictement positive, telle que $A = BB^*$. On ne calcule et ne stocke que B. Avec donc un coût en $\frac{n^3}{6}$ au regard des résultats concernant la factorisation LU.

Calcul pratique (comme pour LU)

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

Etape 1 : On détermine la ligne 1 de B :

$$a_{1,1} = b_{1,1} b_{1,1} \quad \text{donc} \quad b_{1,1} = \sqrt{a_{1,1}}.$$

Etape i : On connaît les lignes 1 à $i-1$ de B :

$$B^{(i-1)} = \begin{pmatrix} b_{1,1} & & 0 \\ \vdots & \ddots & \\ b_{i-1,1} & \cdots & b_{i-1,i-1} \end{pmatrix}$$

On écrit

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

pour tout j afin de déterminer $b_{i,1}, \cdots, b_{i,i}$. En l'écrivant pour $1 \leq j \leq i-1$, on a :

$$\begin{aligned} a_{i,1} &= b_{i,1} b_{1,1} \\ a_{i,2} &= b_{i,1} b_{2,1} + b_{i,2} b_{2,2} \\ &\vdots \\ a_{i,i-1} &= b_{i,1} b_{i-1,1} + \cdots + b_{i,i-1} b_{i-1,i-1} \end{aligned}$$

Ainsi, le vecteur $B_i = \begin{pmatrix} b_{i,1} & \cdots & b_{i,i-1} \end{pmatrix}^T$ est solution du système linéaire triangulaire inférieur $B^{(i-1)}B_i = A_i^{(2)}$ où $A_i^{(2)} = \begin{pmatrix} a_{i,1} & \cdots & a_{i,i-1} \end{pmatrix}^T$.
En écrivant la relation pour $j = i$, on obtient $b_{i,i}$:

$$b_{i,i} = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} b_{i,k}^2}.$$

Algorithme très pratique dès que l'on a une fonction de résolution d'un système triangulaire inférieur.

Structures bande et profil

La décomposition de Cholesky conserve la structure bande ainsi que la structure profil : les zéros d'une ligne figurant avant le premier élément non nul de cette ligne resteront zéros dans la matrice B.

2.2.4 Notion de stabilité numérique

On a vu comment améliorer la stabilité numérique de la méthode de Gauss par un choix judicieux du pivot. Les méthodes LU, LDU et de Cholesky dérivent de la méthode de Gauss sans pivot. Il va donc de soi qu'une telle amélioration de la stabilité numérique de ces méthodes n'est pas envisageable. La stabilité numérique des méthodes LU, LDU et de Cholesky est très liée au conditionnement de la matrice du système. Plus le conditionnement de la matrice du système est grand, plus la solution numérique sera différente de la solution exacte du système.

2.2.5 Factorisation QR

La factorisation QR permet la résolution de $Ax = b$ par la résolution du système triangulaire supérieur $Rx = Q^*b$.

On a vu que Q et R peuvent être obtenues par le procédé de Gram-Schmidt. Le coût de calcul de Q et R est voisin de n^3, ce qui reste de l'ordre de n^3 mais un peu plus coûteux que la méthode de Gauss. En fait, le procédé QR s'avère intéressant dans le cas des systèmes non carrés ou très mal conditionnés.

2.3 Systèmes sur-déterminés

On s'intéresse ici au cas $A \in \mathbb{R}^{n\times p}$ avec $n > p$. Il s'agit du cas où le nombre d'équations est supérieur au nombre d'inconnues. On ne peut alors pas inverser A. On peut par contre toujours chercher la solution au problème de minimisation

$$\text{Trouver } x \in \mathbb{R}^p \text{ tel que } \|Ax - b\| = \min_{y\in\mathbb{R}^p} \|Ay - b\| \qquad (MC\star)$$

En général, on choisit la norme euclidienne. On parle de méthode des moindres carrés.

Remarque 28.
($MC\star$) admet une solution même si $b \notin \mathrm{Im}(A)$. Ce procédé est aussi applicable au cas où A est carrée non inversible.

2.3.1 Résultats préliminaires

Soit $A \in \mathbb{R}^{n\times p}$ et $b \in \mathbb{R}^n$. On cherche $x \in \mathbb{R}^p$ tel que $\|Ax-b\| = \min_{y\in\mathbb{R}^p} \|Ay-b\|$.

Lemme 1.
$x \in \mathbb{R}^p$ est solution de $\|Ax-b\| = \min_{y\in\mathbb{R}^p} \|Ay-b\|$ ssi $x \in \mathbb{R}^p$ est solution de $A^*Ax = A^*b$.

Démonstration :

$$
\begin{array}{ll}
 & \|b-Ax\|^2 \leq \|b-Ay\|^2 \ \forall y \in \mathbb{R}^p \\
\iff & \|b-Ax\|^2 \leq \|b-Ax-tAz\|^2 \ \forall z \in \mathbb{R}^p \ \forall t \in \mathbb{R} \text{ ; chgt de var. } y = x+tz. \\
\iff & \|b-Ax\|^2 \leq \|b-Ax\|^2 + t^2\|Az\|^2 - 2t(b-Ax,Az) \ \forall z \in \mathbb{R}^p \ \forall t \in \mathbb{R}. \\
\iff & 0 \leq t^2\|Az\|^2 - 2t(b-Ax,Az) \ \forall z \in \mathbb{R}^p \ \forall t \in \mathbb{R}. \\
\iff & \left| \begin{array}{l} t\|Az\|^2 - 2(b-Ax,Az) \geq 0 \ \forall z \in \mathbb{R}^p \ \forall t > 0. \\ t\|Az\|^2 - 2(b-Ax,Az) \leq 0 \ \forall z \in \mathbb{R}^p \ \forall t < 0. \end{array} \right. \\
\iff & (b-Ax,Az) = 0 \ \forall z \in \mathbb{R}^p \text{(par passage à la limite pour } \Rightarrow \text{. Evident pour } \Leftarrow\text{).} \\
\iff & (A^*b - A^*Ax, z) = 0 \ \forall z \in \mathbb{R}^p. \\
\iff & A^*b - A^*Ax = 0\ .
\end{array}
$$

Définition 22.
L'équation $A^*Ax = A^*b$ s'appelle équation normale.

Théorème 21.
Pour tout $A \in \mathcal{M}_{n,p}(\mathbb{R})$, avec $n > p$, il existe une solution de l'équation normale. De plus, la solution est unique ssi $\ker(A) = \{0\}$.

Démonstration :
Existence : lemme précédent. Unicité : car $\ker(A) = \ker(A^*A)$.

2.3.2 Résolution de l'équation normale

Si $\ker(A) = \{0\}$, alors A^*A est symétrique définie positive. On peut résoudre l'équation normale par la méthode de Cholesky, par exemple. L'avantage de cette équation est de se ramener au cas carré que l'on a déjà traité.

Un inconvénient majeur mériterait cependant d'être évité. Toutes les méthodes du type de celle de Gauss (Méthodes de Gauss, LU, de Cholesky, QR basée sur le procédé de Gram-Schmidt) utilisent à chaque étape les résultats des étapes précédentes. Elles propagent les erreurs d'étape en étape avec un coefficient multiplicateur qui sera d'autant plus grand que le conditionnement de A^*A est grand.

Le conditionnement mesure le rapport entre les grandes et les petites valeurs propres de la matrice. Plus il est grand, plus le rapport entre les grandeurs des nombres intervenant dans les méthodes de résolution sera grand et plus les erreurs d'arrondis seront amplifiées.

Pour fixer les idées sur la stabilité de la méthode, nous nous permettons une comparaison dans le cas des matrices carrées : Ici, la matrice concernée est A^*A dont le conditionnement est à peu près le carré du conditionnement de A. Finalement, pour une matrice carrée A avec un conditionnement plutôt grand, les méthodes de résolution seront deux fois moins stables pour A^*A que pour A. La résolution de l'équation normale n'est intuitivement pas le choix idéal.

2.3.3 Méthode de factorisation QR

Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$ avec $n \geq p$. On cherche d'abord $R \in \mathcal{M}_{n,p}(\mathbb{R})$ triangulaire inférieure avec $r_{ii} \geq 0$, pour tout i, et $Q \in \mathcal{M}_n(\mathbb{R})$ orthogonale telles que $A = QR$. Ensuite on résout le problème

$$x \in \mathbb{R}^p \text{ tel que } \|Q^*b - Rx\| \leq \|Q^* - Ry\| \ \forall y \in \mathbb{R}^p.$$

On suppose A de rang r. On considère que les r premiers vecteurs colonnes de A constituent une famille libre (toujours possible modulo un certain nombre de permutations de colonnes, cad modulo la multiplication par une suite de matrices orthogonales – comme celles intervenant pour le pivot de Gauss).

On applique alors le procédé de Gram-Schmidt à ces r vecteurs. On obtient une famille orthonormée de r vecteurs de taille n que l'on complète par $n-r$ vecteurs afin d'obtenir une matrice dont les colonnes constituent une base orthonormée de $\mathbb{R}^n$, notée $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ avec $Q_1 \in \mathbb{R}^{n\times r}$ et $Q_2 \in \mathbb{R}^{n\times(n-r)}$. En notant $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, avec $A_1 \in \mathbb{R}^{n\times r}$ et $A_2 \in \mathbb{R}^{n\times(p-r)}$, par le procédé de Gram-Schmidt, on peut écrire $A_1 = Q_1 R_1$ avec R_1 triangulaire et donc :

$$A = \begin{pmatrix} A_1 & A_2 \end{pmatrix} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 & R_2 \\ 0 & R_4 \end{pmatrix}.$$

avec $R_1 \in \mathbb{R}^{r\times r}$ triangulaire (procédé de Gram-Schmidt), et $R_4 \in \mathbb{R}^{(n-r)\times(p-r)}$.

D'après le rang de A et l'organisation de ses colonnes, A_2 est engendrée par A_1 au sens : les colonnes de A_2 sont des combinaisons linéaires des colonnes de A_1. Ceci équivaut à l'existence d'une matrice $M \in \mathbb{R}^{r\times(p-r)}$ telle que $A_2 = A_1 M$. La relation $A_1 = Q_1 R_1$ donne alors $A_2 = Q_1 R_2$ avec $R_2 = R_1 M \in \mathbb{R}^{r\times(p-r)}$. Ainsi, on obtient $R_4 = 0$.

Conclusion : Il existe $R_1 \in \mathbb{R}^{r\times r}$ triangulaire supérieure, $R_2 \in \mathbb{R}^{r\times(p-r)}$ et $Q \in \mathbb{R}^{n\times n}$ orthogonale telles que

$$A = Q \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix}.$$

Théorème 22.

La fonction $x \mapsto \|Ax-b\|_2$ atteint son minimum en $x = \begin{pmatrix} R_1^{-1}(Q^*b)_1 \\ 0 \end{pmatrix}$, et la valeur minimale de $\|Ax-b\|_2$ est $\|(Q^*b)_2\|_{2,\mathbb{R}^{n-r}}$, en définissant $(Q^*b)_1 \in \mathbb{R}^r$ et $(Q^*b)_2 \in \mathbb{R}^{n-r}$ tels que $Q^*b = \begin{pmatrix} (Q^*b)_1 \\ (Q^*b)_2 \end{pmatrix}$.

Démonstration : $\|\cdot\|_2$ étant la norme associée au produit scalaire usuel et Q étant orthogonale, Q conserve la norme $\|\cdot\|_2$. Ainsi :

$$\|Ax-b\|_2^2 = \|Rx - Q^*b\|_{2,\mathbb{R}^n}^2 = \|R_1x_1 + R_2x_2 - (Q^*b)_1\|_{2,\mathbb{R}^r}^2 + \|(Q^*b)_2\|_{2,\mathbb{R}^{n-r}}^2 .$$

en posant $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ avec $x_1 \in \mathbb{R}^r$ et $x_2 \in \mathbb{R}^{p-r}$.

On en déduit qu'une solution du problème de minimisation est donnée par

$$x_1 = R_1^{-1}(Q^*b)_1 \quad \text{et} \quad x_2 = 0.$$

Le minimum est donc $\|(Q^*b)_2\|_{2,\mathbb{R}^{n-r}}$ qui est aussi atteint en tout x vérifiant :

$$x_1 = R_1^{-1}((Q^*b)_1 - R_2x_2) \quad \text{et} \quad x_2 \text{ quelconque}.$$

On vérifie bien l'existence de la solution quelle que soit A, et l'unicité de la solution pour $r = p$, cad rang de A égal à p, cad $\ker(A) = \{0\}$.

Si on prend $n = p = r$, on retrouve bien le résultat connu.

Remarque 29.

$\mathrm{cond}(R)$ est de l'ordre de $\mathrm{cond}(A)$ (lorsque A est carrée inversible) en comparaison avec $\mathrm{cond}(A^*A)$. La détermination de $R_1^{-1}y$, pour y donné, en est mieux conditionnée. Cependant le procédé de Gram-Schmidt est instable : Lorsque n devient grand, Q est numériquement de moins en moins orthogonale et son inverse de moins en moins égale à son adjointe.

2.3.4 Algorithme de Householder - Une autre mise en œuvre de la méthode de factorisation QR

On va établir, ici, un algorithme de factorisation QR plus stable que celle basée sur le procédé de Gram-Schmidt. On utilise toujours le même principe : Multiplier A par une suite de matrices orthogonales simples pour mettre A sous une forme triangulaire supérieure. il s'agit des matrices de Householder.

Définition 23. - Rappel

On appelle matrice de Householder associée au vecteur $v \neq 0$, la matrice

$$H(v) = I - 2\frac{vv^T}{\|v\|_2^2} \quad \text{cad} \quad H_{ij} = \delta_{ij} - 2\frac{v_i v_j}{\sum_{k=1}^{n} |v_k|^2}\,.$$

Par convention $H(0) = I$.

Remarque 30.

- $(vv^T)x = v(v^T x) = (v^T x)v$ puisque $v^T x \in \mathbb{R}$.
- $H(v)$ est la symétrie orthogonale par rapport à l'hyperplan orthogonal à v.
- Pour tout e unitaire et pour tout $v \neq \pm \|v\| e$ (cad v non colinéaire à e), on a :

$$H(v + \|v\| e)v = -\|v\| e \quad H(v - \|v\| e)v = \|v\| e$$
$$(vv^T)(vv^T) = \|v\|^2 vv^T$$

Une conséquence remarquable : On peut transformer un vecteur v quelconque en un vecteur n'ayant qu'une coordonnée non nulle en multipliant par la matrice de Householder qui convient, via le choix du bon vecteur unitaire de la base canonique.

Algorithme de Householder :

Cas des systèmes surdéterminés, $n \geq p$, $A \in \mathcal{M}_{n,p}(\mathbb{R})$. On construit une suite $H^k \in \mathcal{M}_n(\mathbb{R})$ et une suite $A^{k+1} \in \mathcal{M}_{n,p}(\mathbb{R})$, $1 \leq k \leq p$ telles que $A^1 = A$, $A^{k+1} = H^k A^k$, $A^{p+1} = R$ triangulaire supérieure.

Etape 1 : $A^1 = A$. Soit a^1 la première colonne de A^1.

Si $a^1 = \begin{pmatrix} a^1_{1,1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, alors on ne fait rien (i.e. $H^1 = I = H(0)$).

Sinon, $A^2 = H^1 A^1$ avec $H^1 = H(a^1 - \left\|a^1\right\| e_1)$, où e_1 est le premier vecteur de la base canonique de $\mathbb{R}^n$.

La première colonne de A^2 est donc :

$$A^2 e_1 = H^1 A^1 e_1 = H^1 a^1 = H(a^1 - \left\|a^1\right\| e_1)a_1 = \begin{pmatrix} \left\|a^1\right\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Etape k : Les $(k-1)$ premières colonnes de A^k ont des zéros sous la diagonale. Soit a^k le vecteur de taille $(n+1-k)$ dont les composantes sont $a^k_{k,k}, a^k_{k+1,k}, \cdots, a^k_{n,k}$, cad les $(n+1-k)$ dernières composantes de la k^{eme} colonne de A^k.

Si $a^k = \begin{pmatrix} a^k_{k,k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, alors on ne fait rien (i.e. $H^k = I = H(0)$).

Sinon, $A^{k+1} = H^k A^k$ avec $H^k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H(a^k - \|a^k\| e_1) \end{pmatrix}$, où e_1 est le premier vecteur de la base canonique de $\mathbb{R}^{n+1-k}$.
La k^{eme} colonne de A^{k+1} est donc :

$$A^{k+1}e_k = \begin{pmatrix} a^k_{1,k} \\ \vdots \\ a^k_{k-1,k} \\ H(a^k - \|a^k\| e_1)a_k \end{pmatrix} = \begin{pmatrix} a^k_{1,k} \\ \vdots \\ a^k_{k-1,k} \\ \|a^k\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

La multiplication par H^k ne modifie pas les colonnes précédentes (1 à $k-1$). Après p étapes :

$$A^{p+1} = \begin{pmatrix} a^2_{1,1} & \cdots & a^{p+1}_{1,p} \\ 0 & \ddots & \vdots \\ & \ddots & a^{p+1}_{p,p} \\ \vdots & & 0 \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

et $A = QR$ avec $R = A^{p+1}$ et $Q = H^{1*} \cdots H^{p*}$.

Remarque 31.
La non modification des colonnes déjà triangularisées assure une meilleure stabilité qu'à l'algorithme basé sur le procédé de Gram-Schmidt. Les matrices H^k sont tout à fait orthogonale. L'orthogonalité de Q est numériquement optimale.

2.4 Méthodes itératives

2.4.1 Principe

Il s'agit de résoudre un système matriciel de taille $n \times n$ avec un coût inférieur à n^3 en se ramenant à une matrice simple à inverser de manière quasi immédiate.

En effet, inverser une matrice A revient à résoudre n systèmes $Ax = b$ où b parcours l'ensemble des vecteurs de la base canonique. Ce qui se fait avec un coût en n^3 avec une méthode directe.

Alors, on peut imaginer résoudre un système linéaire avec un coût de l'ordre de n^2.

Définition 24. (et principe)
On veut résoudre $Ax = b$, avec $b \in \mathbb{K}^n$ et $A \in \mathbb{K}^{n\times n}$ inversible. On écrit $A = M - N$ avec M facile à inverser. Le couple (M, N) est dit décomposition (régulière) de A. ("splitting" en Anglais) puis on construit la suite :

$$\begin{cases} x_0 \in \mathbb{R}^n \qquad \text{donné.} \\ Mx_{k+1} = Nx_{x_k} + b \qquad \forall k \geq 0 \qquad (\star) \end{cases}$$

On appelle cette construction méthode itérative basée sur la décomposition régulière (M, N).

Remarque 32.
Si $(x_k)_k$ converge vers $\overline{x}$ alors $(\star)$ donne par passage à la limite

$$M\overline{x} = N\overline{x} + b \quad \text{et donc} \quad (M-N)\overline{x} = b \quad \text{c'est à dire} \quad A\overline{x} = b.$$

Ainsi, si $(x_k)_k$ converge alors, la suite converge vers la solution du système linéaire $Ax = b$.

Remarque 33.
En pratique, il est important de savoir quand on peut arrêter la construction du $(x_k)_k$, à quel moment le x_k obtenu est assez proche de la solution exacte (sans connaître la solution exacte).

Définition 25.
On dit qu'une méthode itérative converge ssi $(x_k)_k$ converge quel que soit x_0.

Définition 26.
On rappelle résidu (resp. erreur) à l'itération k, le vecteur $r_k = b - Ax_k$ (resp. $e_k = x_k - x$) ; où x désigne la solution exacte.

Remarque 34.
On ne connaît pas e_k mais il est facile de calculer r_k à chaque itération. La méthode est convergente ssi $e_k \underset{k\to\infty}{\longrightarrow} 0$ ssi $r_k \underset{k\to\infty}{\longrightarrow} 0$.

Remarque 35.
$Mx_{k+1} = Nx_k + b \quad \Longleftrightarrow \quad x_{k+1} = M^{-1}Nx_k + M^{-1}b$. On appelle $M^{-1}N$ la matrice d'itération de la méthode itérative associée à la décomposition régulière (M, N).

Théorème 23.
La méthode itérative $(\star)$ *converge ssi* $\rho(M^{-1}N) < 1$.

Démonstration :
Rappel : Pour une matrice B, $\rho(B) < 1$ ssi $\exists \|\cdot\|$ norme matricielle telle que $\|B\| < 1$.
Soit x la solution exacte, alors, puisque $(M-N)x = b$, on a aussi $x = M^{-1}Nx + M^{-1}b$. Ainsi :

$$\begin{aligned} e_k = x_k - x &= (M^{-1}Nx_{k-1} + M^{-1}b) - (M^{-1}Nx + M^{-1}b) \\ &= M^{-1}N(x_{k-1} - x) = M^{-1}Ne_{k-1} \end{aligned}$$

Il en découle que $e_k = (M^{-1}N)^k(x_0 - x)$. Alors $e_k \underset{k\to\infty}{\longrightarrow} 0$ ssi $\exists \|\cdot\|$ norme matricielle telle que $\left\|M^{-1}N\right\|^k \underset{k\to\infty}{\longrightarrow} 0$, cad ssi $\rho(M^{-1}N) < 1$.

Un premier exemple : La méthode itérative de Richardson

Le choix le plus facile pour M est l'identité à un coefficient multiplicatif près. La méthode est :

$$\begin{cases} x_0 \in \mathbb{R}^n, \\ x_{k+1} = x_k + \alpha(b - Ax_k) \qquad \forall k \geq 0 \end{cases}$$

on a alors la récurrence : $\frac{1}{\alpha}x_{k+1} = (\frac{1}{\alpha}I - A)x_k + b$. Il s'agit de la méthode itérative associée à la décomposition (M, N) avec $M = \frac{1}{\alpha}I$ et $N = \frac{1}{\alpha}I - A$. Dans ce cas, la matrice d'itération est $M^{-1}N = I - \alpha A$.
On reverra cette méthode comme une méthode dite "variationnelle".

Regardons maintenant sa convergence : Quand a-t-on $\rho(I-\alpha A)<1$?

$$\begin{array}{rcl}\lambda\in\sigma(I-\alpha A) & \iff & \exists x \text{ tel que } x-\alpha Ax=\lambda x\\ & \iff & \exists x \text{ tel que } Ax=\left(\frac{1-\lambda}{\alpha}\right)x\\ & \iff & \left(\frac{1-\lambda}{\alpha}\right)\in\sigma(A).\end{array}$$

D'autre part,

$$-1<\lambda<1 \iff \left|\begin{array}{ll}0<\frac{1-\lambda}{\alpha}<\frac{2}{\alpha} & \text{si } \alpha>0.\\ \frac{2}{\alpha}<\frac{1-\lambda}{\alpha}<0 & \text{si } \alpha<0.\end{array}\right.$$

Ainsi, on ne peut choisir $\alpha>0$ que si toutes les valeurs propres de A sont positives et dans ce cas on doit avoir $\rho(A)<\frac{2}{\alpha}$. De même, on ne peut choisir $\alpha<0$ que si toutes les valeurs propres de A sont négatives et dans ce cas on doit avoir $\rho(A)<-\frac{2}{\alpha}$.

Conclusion :

$$\rho(M^{-1}N)<1 \iff \left|\begin{array}{c}\sigma(A)\subset\mathbb{R}^+ \text{ et } 0<\alpha<\frac{2}{\rho(A)}.\\ \text{ou}\\ \sigma(A)\subset\mathbb{R}^- \text{ et } 0>\alpha>\frac{-2}{\rho(A)}.\end{array}\right.$$

Remarque 36. Notion de conditionnement

Si A est normale inversible, son conditionnement est lié à son rayon spectral par la relation $\mathrm{cond}_2(A)=\rho(A)\rho(A^{-1})$. Ainsi, plus le conditionnement est grand, plus α est petit en valeur absolue, plus la convergence est lente, et plus la méthode de Richardson est coûteuse.

Théorème 24.

Soit A hermitienne définie positive. Soit (M,N) une décomposition régulière de A ($A=M-N$ avec M inversible). Alors, la matrice M^*+N est hermitienne.

De plus, si (M^*+N) est aussi définie positive, alors $\rho(M^{-1}N)<1$.

Démonstration : Cf TD.

Remarque 37.

Il existe plein d'autres résultats de ce type.

Remarque 38.

Après l'étude de la convergence, il est intéressant de regarder les problèmes de stabilité de ces méthodes : comment se propagent les erreurs numériques au cours des itérations ?

Théorème 25.

Soit une décomposition régulière de A définie par $A=M-N$ avec M inversible. Soit $b\in\mathbb{R}^n$ et $x\in\mathbb{R}^n$ solution du problème $Ax=b$. On suppose qu'à chaque étape k, la méthode itérative est affectée d'une erreur $\varepsilon_k\in\mathbb{R}^n$ au sens où x_{k+1} est donné par

$$x_{k+1}=M^{-1}Nx_k+M^{-1}b+\varepsilon_k.$$

On suppose $\rho(M^{-1}N)<1$ et l'existence d'une norme vectorielle $\|\cdot\|$ et d'une constante $\varepsilon>0$ telle que $\forall k\geq 0$, $\|\varepsilon_k\|<\varepsilon$.

Alors, $\exists\,\|\cdot\|_s$ norme subordonnée et $\exists c\in\mathbb{R}$ telles que

$$\limsup_{k\to+\infty}\|x_k-x\|\leq K\varepsilon \quad \textit{avec} \quad K=\frac{c^2}{1-\|M^{-1}N\|_s}\,.$$

Remarque 39.

Ceci nous donne un résultat de stabilité, mais on n'a pas stabilité inconditionnelle par rapport à la taille du système à résoudre.

Démonstration :
Soit $e_k = x_k - x$. Alors $e_{k+1} = M^{-1}Ne_k + \varepsilon_k$. Ainsi :

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

D'autre part, $\rho(M^{-1}N) < 1$. Ainsi, $\exists \left\|\cdot\right\|_s$ norme subordonnée telle que $\left\|M^{-1}N\right\|_s < 1$. Notons de même la norme vectorielle associée. De plus, toutes les normes étant équivalentes, il existe c tel que :

$$c^{-1} \left\|y\right\| \leq \left\|y\right\|_s \leq c \left\|y\right\| \qquad \forall y \in \mathbb{R}^n.$$

Ainsi

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

Par le choix de $\left\|\cdot\right\|_s$, le premier terme du second membre tend vers 0. Le deuxième est évidemment majoré par $\frac{c\varepsilon}{1-\left\|M^{-1}N\right\|_s}$.

Enfin, puisque $\left\|e_k\right\| \leq c \left\|e_k\right\|_s$, on récupère le résultat avec $K\varepsilon = \frac{c^2\varepsilon}{1-\left\|M^{-1}N\right\|_s}$.

2.4.2 Méthode de Jacobi

Notons $A = (A_{i,j})_{1\leq i,j\leq n}$. Posons alors $D = (A_{i,j}\delta_{i,j})_{1\leq i,j\leq n}$ où $\delta_{i,j}$ est le symbole de Kronecker. D est la diagonale de A.

Définition 27.
La méthode de Jacobi est la méthode itérative associée à la décomposition régulière (M, N) avec $M = D$ et $N = D - A$. On note $\mathcal{J}$ sa matrice d'itération : $\mathcal{J} = I - D^{-1}A$.

Remarque 40.
La méthode a un sens ssi D est inversible. Il faut donc veiller à écrire le système de telle sorte que A n'ait pas de zéro sur sa diagonale.

Remarque 41.
Si A est hermitienne, alors la méthode de Jacobi converge si A et $2D - A$ sont définies positives.

2.4.3 Méthode de Gauss-Seidel

Notons $A = (A_{i,j})_{1\leq i,j\leq n}$. Ecrivons $A = D - E - F$ avec :
D est la partie diagonale de A : $D_{i,j} = A_{i,j}\delta_{i,j}$.
$-E$ est la partie triangulaire inférieure de A : $E_{i,j} = -A_{i,j}$ si $i > j$ et 0 sinon.
$-F$ est la partie triangulaire supérieure de A : $F_{i,j} = -A_{i,j}$ si $i < j$ et 0 sinon.

Définition 28.
La méthode de Gauss-Seidel est la méthode itérative associée à la décomposition régulière (M, N) avec $M = D - E$ et $N = F$. On note $\mathcal{G}_1$ sa matrice d'itération : $\mathcal{G}_1 = (D - E)^{-1}F$.

Remarque 42.
La méthode a un sens ssi D est inversible (comme celle de Jacobi).

Remarque 43.
Si A est hermitienne définie positive, alors $M^* + N = D$ est aussi hermitienne positive et donc la méthode de Gauss-Seidel converge.

Comparaison algorithmique des méthodes de Jacobi et de Gauss-Seidel

Notons l'itéré $x_k = (x_i^k)_{1 \leq i \leq n}$.

Pour la méthode de Jacobi, à l'itération $k+1$, le calcul de x_i^{k+1} nécessite la connaissance de b_i ainsi que la quasi-totalité de x_k car celui-ci est multiplié par $D - A$.

Pour la méthode de Gauss-Seidel, à l'itération $k+1$, le calcul de x_i^{k+1} nécessite :
$x_{i+1}^k, \cdots, x_n^k$ (calcul de Fx_k)
$x_1^{k+1}, \cdots, x_{i-1}^{k+1}$ et b_i (méthode de descente à $(D-E)x_{k+1} = Fx_k + b$)

Algorithme de la méthode de Jacobi :

$$x_i^{k+1} = \frac{1}{A_{i,i}} \left[-A_{i,1}x_1^k - \cdots - A_{i,i-1}x_{i-1}^k + b_i - A_{i,i+1}x_{i+1}^k - \cdots - A_{i,n}x_n^k \right]$$

Algorithme de la méthode de Gauss-Seidel :

$$x_i^{k+1} = \frac{1}{A_{i,i}} \left[-A_{i,1}x_1^{k+1} - \cdots - A_{i,i-1}x_{i-1}^{k+1} + b_i - A_{i,i+1}x_{i+1}^k - \cdots - A_{i,n}x_n^k \right]$$

Pour la méthode de Jacobi, il est nécessaire d'avoir en permanence deux itérés consécutifs. Pour la méthode de Gauss-Seidel, on peut écraser un itéré par son successeur au fur et à mesure du calcul.

2.4.4 Méthode de relaxation (SOR - Successive Over Relaxation)

Même écriture de A que pour la méthode de Gauss-Seidel : $A = D - E - F$.

Définition 29.
Soit $\omega \in \mathbb{R}_*^+$. On appelle méthode de relaxation, de paramètre de relaxation ω, la méthode itérative associée à la décomposition (M, N) avec

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

On note $\mathcal{G}_\omega$ sa matrice d'itération

$$\mathcal{G}_\omega = \left(\frac{D}{\omega} - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right)$$

Remarque 44.
La méthode est bien définie ssi D est inversible (comme pour la méthode de Jacobi).

Remarque 45.

Si $\omega = 1$, il s'agit de la méthode de Gauss-Seidel.
Si $\omega < 1$, on parle de sous-relaxation.
Si $\omega > 1$, on parle de sur-relaxation.
L'efficacité de la méthode est mesurée par $\rho(\mathcal{G}_\omega)$ et dépend de ω. On va donc chercher ω minimisant $\rho(\mathcal{G}_\omega)$. En général, cet ω est > 1. D'où le nom"SOR".

Théorème 26.
Si A est hermitienne définie positive, alors la méthode de relaxation converge pour tout $\omega \in]0, 2[$.

Démonstration :

A est hermitienne déf. pos. donc D aussi et $\frac{D}{\omega} - E$ est inversible.
$M^* + N = \frac{D}{\omega} - E^* + \frac{1-\omega}{\omega}D + F = \frac{2-\omega}{\omega}D$ car A hermitienne implique $E^* = F$. Ainsi, $M^* + N$ est aussi déf. pos. ssi $0 < \omega < 2$. D'où le résultat.

Théorème 27.

Soit $A \in \mathbb{C}^{n\times n}$ inversible quelconque. Alors $\rho(\mathcal{G}_\omega) \geq |1-\omega|$ pour tout $\omega \neq 0$. Ainsi, si la méthode converge, on a nécessairement $0 < \omega < 2$.

Démonstration :

$\det(\mathcal{G}_\omega) = \det(\frac{1-\omega}{\omega}D + F)/\det(\frac{D}{\omega} - E) = (1-\omega)^n$.
$\rho(\mathcal{G}_\omega)^n \geq \prod_{i=1}^n |\lambda_i(\mathcal{G}_\omega)| = |\det(\mathcal{G}_\omega)| = |1-\omega|^n$ où les $\lambda_i(\mathcal{G}_\omega)$ désignent les valeurs propres de $\mathcal{G}_\omega$.

2.4.5 Comparaison des méthodes sur des matrices tridiagonales

Voir feuille d'exercices de TD numéro 3.

2.4.6 Programmation dans le cas général

On s'intéresse ici à la résolution du système $Ax = b$ par une méthode itérative quelconque basée sur la décomposition régulière (M, N) avec M "facilement" inversible.

Critère d'arrêt :

Il est nécessaire de définir un critère qui arrête le calcul lorsque la solution approchée x_k est suffisamment proche de la solution exacte x sachant que x est une inconnue.

Un premier critère consiste à définir la précision ε désirée sur le résidu : Le critère d'arrêt $\|b - Ax_k\| \leq \varepsilon$ est fréquemment utilisé. Il faut cependant garder à l'esprit qu'il peut être trompeur dans le cas où $\left\|A^{-1}\right\|$ serait grand. En effet :

$$\|x - x_k\| \leq \left\|A^{-1}\right\| \|b - Ax_k\| \leq \varepsilon \left\|A^{-1}\right\|.$$

Certains utilisent aussi un critère sur le résidu dit relatif : $\frac{\|b-Ax_k\|}{\|b-Ax_0\|}$.

Un deuxième choix consiste à regarder la différence entre deux itérés consécutifs : Arrêt de l'algorithme lorsque $\|x_{k-1} - x_k\| \leq \varepsilon$. Ce critère est très simple mais dangereux ! La vitesse de convergence n'est pas régulière. Il arrive que la convergence ralentisse sévèrement avant de réaccélérer, ce qui a pour conséquence de faire apparaître des itérés successifs proches même loin de la solution exacte.

Algorithme

On veut résoudre $Ax = b$ selon la méthode

$$x_0 \text{ arbitraire} \quad \text{et} \quad x_{k+1} = M^{-1}Nx_k + M^{-1}b.$$

Puisque $N = M - A$, on peut écrire $x_{k+1} = x_k + M^{-1}r_k$ avec $r_k = b - Ax_k$. Il en découle que $r_{k+1} = r_k - AM^{-1}r_k$.

En utilisant ces relations, on met en place l'algorithme suivant où, à chaque étape, les seuls calculs faisant intervenir une matrice sont $y = M^{-1}r_k$ et Ay.

Choisissons le critère d'arrêt portant sur la norme euclidienne du résidu. L'algorithme s'écrit (les lignes commençant par un signe "%" sont des lignes de commentaire) :

%%%%%%%%%% Début de l'algorithme %%%%%%%%%%

Données : A, b
Sortie : x solution approchée du système $Ax = b$.

Initialisation : Choisir $x \in \mathbb{R}^n$, calculer $r = b - Ax$.
Tant que $\|r\|_2 > \varepsilon$
% A ce moment, $x = x_k$ et $r = r_k$

1. Calculer $y \in \mathbb{R}^n$ solution de $My = r$ (cad trouver $y = M^{-1}r$).
   % cette étape doit prendre en compte la forme de M
   % afin de minimiser le coût de la résolution.

2. Mise à jour de la solution : $x = x + y$.
3. Calcul du résidu : $r = r - Ay$.

% A ce moment, $x = x_{k+1}$ et $r = r_{k+1}$.
Fin tant que

%%%%%%%%%% Fin de l'algorithme %%%%%%%%%%

Remarque 46.
En pratique, on ajoute une condition dans le critère d'arrêt qui consiste à limiter le nombre d'itérations à un nombre $itermax$ afin d'éviter de boucler sans fin dans un cas de non convergence.

Remarque 47.
Le coût de calcul est à chaque itération en n^2. Si la matrice M est triangulaire (resp. diagonale), le coût d'une itération est à peu près $\frac{3}{2}n^2$ (resp. n^2)

Remarque 48.
Dans le cas de la méthode de Gauss-Seidel, il est préférable d'utiliser l'algorithme énoncé lors de son étude et qui s'appuie sur la forme triangulaire de N. L'algorithme que l'on vient de donner ne prend pas en compte cette particularité.

2.5 Méthodes variationnelles

Soit $A \in \mathbb{R}^{n\times n}$ symétrique. Nous présentons ici une série de méthodes itératives pouvant certaines aussi être vues comme des méthodes directes. Pour certaines démonstrations, nous renvoyons le lecteur à la bibliographie.

2.5.1 La méthode du gradient à pas fixe

The steepest descent method (méthode de la plus grande pente).

Définition 30.
La méthode itérative du gradient est définie par la décomposition régulière (M, N) avec $M = (1/\alpha)I_n$ et $N = (1/\alpha)I_n - A$ où α est un paramètre de $\mathbb{R}^*$. Il s'agit de calculer :

$$\begin{cases} x_0 \text{ donné dans } \mathbb{R}^n \\ x_{k+1} = x_k + \alpha(b - Ax_k) \quad k \geq 0 \end{cases}$$

Remarque 49.
Déjà vue sous le nom de méthode de Richardson.

Théorème 28.

Soit A diagonalisable de valeurs propres $\lambda_1 \leq \cdots \leq \lambda_n$.

(i) Si $\lambda_1 \leq 0 \leq \lambda_n$, la méthode diverge quelque soit α.

(ii) Si $0 < \lambda_1 \leq \lambda_n$, la méthode converge quelque soit $0 < \alpha < \frac{2}{\lambda_n}$.

(iii) Si $0 < \lambda_1 \leq \lambda_n$, alors α_{opt} qui minimise $\rho(M^{-1}N)$ est : $\alpha_{opt} = \frac{2}{\lambda_1+\lambda_n}$ et $\min_\alpha \rho(M^{-1}N) = \frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}$.

Remarque 50.

Si A est normale inversible, $\mathrm{cond}_2(A) = \frac{\lambda_n}{\lambda_1}$ et donc : $\min_\alpha \rho(M^{-1}N) = \frac{\mathrm{cond}_2(A)-1}{\mathrm{cond}_2(A)+1}$. Plus le conditionnement est proche de 1, plus la méthode converge vite. Plus le conditionnement est grand, plus la méthode est lente.

2.5.2 Interprétation graphique

Remarque 51. (Rappel)

Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$. On appelle gradient de f en x, la quantité

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n} \right)^T .$$

Désignons désormais par f, la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$f(x) = \frac{1}{2} < Ax, x > - < b, x > .$$

La résolution de $Ax = b$ correspond alors à la minimisation de la fonctionnelle quadratique f. Son gradient est $\nabla f(x) = Ax - b$.

Il faut imaginer que l'on visite une fosse sous-marine et que l'on veut atteindre le point le plus bas de cette fosse, en suivant les fonds marins et par visibilité réduite. La direction de descente à choisir localement est alors celle de la plus grande pente vers le bas.

Proposition 7.

(i) Si A est symétrique définie positive, alors f admet un unique minimum en x_0 solution unique du système linéaire $Ax = b$.

(ii) Si A est symétrique positive non définie, et si $b \in \mathrm{Im}(A)$, alors f atteint son minimum pour tout et seulement tout vecteur solution du système linéaire $Ax = b$.

(iii) Si A n'est pas positive, ou si $b \notin \mathrm{Im}(A)$, alors f n'admet pas de minimum, c'est à dire que son infimum est $-\infty$.

Théorème 29.

Soit $A \in \mathbb{R}^{n\times n}$ symétrique définie positive, et $f : x \mapsto \frac{1}{2} < Ax, x > - < b, x >$. Alors

(i) $x \in \mathbb{R}^n$ minimise f ssi $\nabla f(x) = 0$.

(ii) Si $x \in \mathbb{R}^n$ vérifie $\nabla f(x) \neq 0$, alors $\forall \alpha \in]0, \frac{2}{\rho(A)}[$, $f(x - \alpha \nabla f(x)) < f(x)$.

On retrouve, dans ce théorème, des valeurs critiques de la méthode du gradient à pas fixe.

Dans la section précédente, on avait défini x_{k+1} en fonction de x_k par un déplacement dans la direction de $b - Ax_k$ avec un pas fixe égal à α. La direction choisie correspond à la direction du gradient de f en x_k, $\nabla f(x_k)$, c'est à dire la direction de plus grande variation de f en ce point, direction orthogonale en x_k à la courbe de niveau de f passant par x_k. Une question devient alors légitime : Pourquoi ne pas choisir un α différent à chaque étape ? Cela permet d'introduire une variante : la méthode du gradient à pas variable pour laquelle on s'autorise des valeurs différentes pour α, au cours des itérations de la méthode. Mais alors, peut-on faire encore mieux ? Peut-on choisir, à chaque itération, un α qui serait optimal dans la direction de déplacement $\nabla f(x_k)$?

2.5.3 Méthode du gradient à pas optimal

A chaque itération, le paramètre α est choisi de façon à minimiser la valeur de f sur la droite $x_k +$ Vect$\{\nabla f(x_k)\}$. On cherche ainsi à minimiser $g_k : \alpha \mapsto f(x_k + \alpha(b - Ax_k))$ par rapport à α.

La recherche du minimum de f sur $\mathbb{R}^n$ devient donc une suite de problèmes de minimisation sur $\mathbb{R}$ que l'on peut résoudre de manière exacte : la fonction g_k est en fait polynomiale de degré 2, et de coefficient de plus haut degré positif. En effet,

$$g_k(\alpha) = f(x_k + \alpha(b - Ax_k)) = ... = \frac{\alpha^2}{2} < Ar_k, r_k > -\alpha < r_k, r_k > + < \frac{1}{2}Ax_k - b, x_k >$$

en posant $r_k = b - Ax_k$.
Elle atteint alors son minimum au point unique α_k vérifiant $g'_k(\alpha_k) = 0$, ce qui nous amène à l'expression suivante :

$$\alpha_k = \frac{< r_k, r_k >}{< Ar_k, r_k >} = \frac{\|r_k\|^2}{\|r_k\|_A^2}$$

où $\|\cdot\|_A$ désigne la norme associée à la matrice symétrique définie positive A.

On obtient alors le nouvel algorithme :

$$x_0 \text{ donné dans } \mathbb{R}^n$$
$$\text{Pour } k \geq 0 \quad \begin{cases} r_k = b - Ax_k. \\ \alpha_k = \frac{<r_k, r_k>}{<Ar_k, r_k>}. \\ x_{k+1} = x_k + \alpha_k r_k \end{cases}$$

Interprétation graphique

On gravit une montagne, au cœur d'une forêt dense, avec l'objectif d'en atteindre le sommet. A chaque étape, on se rapproche du sommet selon le choix de la direction de plus grande pente au point d'arrêt de la précédente étape. Et on suit cette direction jusqu'à ce que la pente devienne nulle. Au delà de ce point, on reperdrait l'altitude gagnée jusque là. Puis on passe à l'étape suivante ... Le choix du pas est optimal, localement à chaque étape, pour la direction choisie. Cette méthode est tout de même limitée. En effet, le caractère d'optimalité de la méthode n'est que local en chaque point x_k. A chaque instant, on oublie d'ailleurs les choix précédents. On ne peut d'ailleurs a priori pas établir de résultat de convergence vers le sommet en temps fini.

Si on se place maintenant dans le cas d'un montagnard averti doté d'un fort sens de l'orientation, on a envie de choisir un déplacement (certes toujours dans une direction montante) qui prendrait aussi en considération le chemin déjà parcouru globalement. C'est le principe de la méthode du gradient conjugué, basée sur la considération des espaces de Krylov introduits ci-dessous.

2.5.4 Espaces de Krylov

Définition 31.

Soit r_0 un vecteur de $\mathbb{R}^n$. On appelle espace de Krylov associé à r_0 (et à A), et on note K_k, le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les $k + 1$ vecteurs $\{r_0, Ar_0, \cdots, A^k r_0\}$:

$$K_k = \text{Vect}\{r_0, Ar_0, \cdots, A^k r_0\}$$

Proposition 8.

Soit $r_0 \neq 0$. Alors $(K_k)_{k \leq 0}$ est croissante au sens de l'inclusion : $K_k \subset K_{k+1}$.
De plus, il existe k_0 tel que $0 \leq k_0 \leq n - 1$ et :

$$\dim K_k = k + 1 \text{ si } 0 \leq k \leq k_0 \quad \text{et} \quad \dim K_k = k_0 + 1 \text{ si } k_0 \leq k \ .$$

Proposition 9.

Soit une méthode du gradient $\begin{cases} x_0 \in \mathbb{R}^n \\ x_{k+1} = x_k + \alpha_k(b - Ax_k) \end{cases}$
On définit le résidu $r_k = b - Ax_k$*. On a alors le résultat :*

(i) $r_k \in K_k$*, espace de Krylov associé à* $r_0 = b - Ax_0$*.*

(ii) $x_{k+1} \in x_0 + K_k$*.*

Démonstration : Par récurrence sur k.

Remarque 52.

La méthode du gradient à pas optimal implique : $\forall k, \ < r_{k+1}, r_k >= 0$. En effet,

$$\begin{aligned} < r_{k+1}, r_k > &= < b - Ax_{k+1}, b - Ax_k > \\ &= < Ax_k + \alpha_k A(b - Ax_k) - b, Ax_k - b > \\ &= < Ax_k - b, Ax_k - b > -\alpha_k < A(Ax_k - b), Ax_k - b > \\ &= < Ax_k - b, Ax_k - b > - < Ax_k - b, Ax_k - b >= 0 \\ &\quad \text{par définition de } \alpha_k \end{aligned}$$

La méthode du gradient à pas optimal consiste à choisir x_{k+1} minimisant $g_k(\alpha) = f(x_k + \alpha r_k)$ (c'est à dire, minimisant f sur la droite $x_k + \mathrm{Vect}\{r_k\}$), ou encore x_{k+1} tel que r_{k+1} orthogonal à r_k.

2.5.5 Méthode du gradient conjugué

La méthode du gradient à pas optimal optimise le choix du pas en fonction de l'étape précédente uniquement. Ici, il s'agit de mettre en place une méthode qui optimise le choix du pas et de la direction de descente en fonction de l'ensemble des itérations précédentes. Dans la fosse sous-marine, la pente optimale globalement (celle qui se dirige vers le fond), n'est pas forcément celle de plus grande pente localement. En comparaison avec la méthode du gradient à pas optimal (voir remarque 52), la méthode du gradient conjugué consiste à chaque itération, à choisir x_{k+1} minimisant f sur $x_0 + K_k$, avec K_k l'espace de Krylov associé à r_0, ou encore x_{k+1} tel que $r_{k+1} \in x_0 + K_k$ orthogonal à K_k.

Théorème 30.

Soit $A \in \mathbb{R}^{n \times n}$ *symétrique définie positive. La méthode du gradient conjugué converge en au plus* n *itérations.*

Idée de la démonstration : La méthode a convergé lorsque les espaces de Krylov s'arrêtent de croitre. Ainsi, si $k_0 \geq n - 1$, on a nécessairement $K_{n-1} = \mathbb{R}^n$ et $k_0 = n - 1$.

Proposition 10.

Soit $A \in \mathbb{R}^{n \times n}$ *symétrique définie positive. Soit* $(x_k)_{0 \leq k \leq n}$ *la suite construite par la méthode du gradient conjugué. Notons* $r_k = b - Ax_k$*,* $d_k = x_{k+1} - x_k$*. Alors :*

(i) $K_k = \mathrm{Vect}\{r_0, Ar_0, \cdots, A^k r_0\} = \mathrm{Vect}\{r_0, r_1, \cdots, r_k\} = \mathrm{Vect}\{d_0, d_1, \cdots, d_k\}$

(ii) $(r_k)_{0 \leq k \leq n-1}$ *est orthogonale.*

(iii) $(d_k)_{0 \leq k \leq n-1}$ *est "conjuguée par rapport à* A*", c'est à dire qu'elle est orthogonale pour le produit scalaire associé à* $A :< Ad_k, d_l >= 0$ *si* $k \neq l$*.*

Démonstration : admise

Remarque 53.

La méthode du gradient conjugué est a priori compliquée puisqu'elle consiste en une suite de problème de minimisation sur un espace dont la dimension est incrémentée à chaque étape. Cependant, les résultats admis ici permettent de montrer que la méthode se ramène à un algorithme très simple, comparable à la méthode du gradient à pas optimal par ses coûts calcul et mémoire. Ceci fait l'objet de l'exercice 7 de la série de TD3.

Algorithme de la méthode du gradient conjugué :

Initialisation :

$x_0 \in \mathbb{R}^n$, ε précision souhaitée sur le résidu.
$r_0 = b - Ax_0$; $p_0 = r_0$; $\theta_0 =< p_0, r_0 >$;

Itérations successives : $k \geq 0$,

$\alpha_k = \frac{\theta_k}{<Ap_k, p_k>}$ (pas dans la direction p^k)
$x_{k+1} = x_k + \alpha_k p_k$
$r_{k+1} = r_k - \alpha_k A p_k$ (résidu à l'itération $k+1$)
Arrêt des itérations si $\|r_{k+1}\| \leq \varepsilon$.
$\theta_{k+1} =< r_{k+1}, r_{k+1} >$
$\beta_{k+1} = \frac{\theta_{k+1}}{\theta_k}$
$p_{k+1} = r_{k+1} + \beta_{k+1} p_k$ (prochaine direction de descente)

Coût de l'algorithme et convergence

La partie la plus coûteuse de l'algorithme est le produit matrice-vecteur Ap_k de chaque itération. Les autres calculs sont vectoriels ou scalaires et donc peu significatifs devant ce produit. Le coût total est donc de l'ordre de $n_{iter}n^2$.

En théorie, n_{iter} est inférieur à n. On peut ainsi définir une méthode a priori directe, atteignant la solution exacte avec un coût d'un ordre inférieur ou égal à n^3. Cependant, la convergence de la méthode et le résultat théorique sur la convergence en temps fini dépend fortement de la façon dont les résidus numériques ne constituent pas vraiment une famille orthogonale. Ceci dépend du conditionnement de la matrice. On peut énoncer le résultat suivant :

Théorème 31.
La vitesse de convergence est donnée par la relation :

$$\|x_k - x^\star\|_2 \leq 2\sqrt{\mathrm{cond}_2(A)} \left(\frac{\sqrt{\mathrm{cond}_2(A)} - 1}{\sqrt{\mathrm{cond}_2(A)} + 1} \right)^k \|x_0 - x^\star\|_2$$

où $x^\star$ désigne la solution exacte du système $Ax = b$.

Ainsi, dans la pratique, on rencontre parfois des matrices (même denses) très bien conditionnées pour lesquelles la méthode du gradient conjugué atteint une précision suffisante après $n_{iter} \sim \ln(n)$ itérations, ce qui fait de la méthode, une méthode avec un coût de l'ordre de $n^2 \ln(n)$, ce qui est très agréable. A contrario, on trouve aussi des matrices symétriques définies positives mal conditionnées pour lesquelles la méthode ne converge pas en des temps acceptables.

Chapitre 3

APPROXIMATION SPECTRALE

3.1 Introduction

3.1.1 Motivations

On a mentionné au dernier paragraphe du chapitre 1 une application du calcul des valeurs propres d'une matrice de grande taille. On peut en citer de nombreuses, en mécanique des structures comme il a été vu, mais aussi en chimie quantique (calcul de niveaux d'énergie), en économie (détermination d'un taux de croissance dans un modèle de comptabilité nationale), etc. Les problèmes se divisent en deux catégories :

- la détermination de toutes les valeurs propres (et, éventuellement, de tous les vecteurs propres) d'une matrice,
- la détermination des quelques plus grandes valeurs propres d'une matrice (ou des quelques plus petites, ou encore des quelques plus proches d'un nombre complexe donné).

Ces deux types de problèmes conduisent à des méthodes différentes, le premier cas étant *a priori* plus coûteux que le second.

3.1.2 Analyse de sensibilité

On rappelle ici le résultat énoncé dans la proposition 4 du chapitre 1 : si A est diagonalisable telle que $A = PDP^{-1}$ avec D diagonale, et E est une matrice de perturbation telle que $\|E\|_\infty \leq \varepsilon$, alors

$$\forall \lambda_\varepsilon \in \sigma(A+E) \quad \exists \lambda \in \sigma(A) \quad |\lambda - \lambda_\varepsilon| \leq K(P)\varepsilon,$$

où $K(P) = \|P\|_\infty \|P^{-1}\|_\infty$ désigne le conditionnement en norme infinie de P. Ainsi, le facteur d'amplification des erreurs est lié au conditionnement de la matrice de passage, et non à celui de A comme c'était le cas pour la résolution d'un système linéaire. De la sorte, on dira qu'un problème aux valeurs propres est mal conditionné quand la matrice de passage P a un conditionnement important (alors qu'un système linéaire est mal conditionné lorsque la matrice A a un fort conditionnement).

3.2 Méthodes de la puissance

3.2.1 Méthode de la puissance

La méthode de la puissance vise à approcher la valeur propre de plus grand module (supposée unique) d'une matrice complexe donnée $A \in \mathbb{C}^{d\times d}$. Elle s'appuie sur le constat suivant, montré au chapitre 1 : pour toute norme subordonnée, on a

$$\|A^n\|^{1/n} \to \rho(A), \text{ lorsque } n \to \infty.$$

Ainsi, la matrice A^n se comporte asymptotiquement comme $\rho(A)^n$, c'est-à-dire une matrice scalaire. Plus précisément, la méthode de la puissance consiste à construire la suite $(x^n)_n$ à partir d'un vecteur initial $x^0 \in \mathbb{C}^d$ par la relation de récurrence

$$x^{n+1} = Ax^n.$$

Pour des raisons de stabilité numérique (on ne souhaite pas que la suite (x^n) soit non bornée, ou qu'elle tende vers 0), on normalise à chaque étape :

$$x^{n+1} = \frac{Ax^n}{\|Ax^n\|}.$$

Notons que la convergence ne peut avoir lieu lorsque plusieurs valeurs propres sont de module maximal, comme le montre l'exemple suivant :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad x^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

La convergence est fonction des propriétés spectrales de la matrice A, comme le précise le théorème suivant.

Théorème 32.

On suppose que la matrice A admet une unique valeur propre de module maximal ρ_1, notée λ ; on note ρ_2 le second plus grand module. Pour presque tout x^0 de $\mathbb{C}^d$, la méthode de la puissance converge, au sens où il existe un vecteur propre $x \in \mathbb{C}^d$ de A associé à la valeur propre λ tel que

$$q^n \overset{\text{def.}}{=} \left[\frac{|\lambda|}{\lambda}\right]^n \frac{x^n}{\|x^n\|} \longrightarrow x, \text{ lorsque } n \to \infty.$$

De plus, la vitesse de convergence est donnée par

- $\diamond$ $\|q^n - x\| = \mathcal{O}\left(\left|\frac{\rho_2}{\rho_1}\right|^n\right)$ *si λ est non défective, ainsi que toute valeur propre λ' telle que $|\lambda'| = \rho_2$.*
- $\diamond$ $\|q^n - x\| = \mathcal{O}\left(n^{r-1}\left|\frac{\rho_2}{\rho_1}\right|^n\right)$ *si λ est non défective, et r est la dimension du plus grand bloc de Jordan associé à une valeur propre de module égal à ρ_2.*
- $\diamond$ $\|q^n - x\| = \mathcal{O}\left(\frac{1}{n}\right)$ *si λ est défective.*

Rappel *: une valeur propre λ est dite défective s'il existe un bloc de Jordan associé à λ de taille strictement supérieure à 1. De manière équivalente, cela signifie que les sous-espaces propre et caractéristique associés à la valeur propre λ sont distincts.*

PREUVE. En réduisant la matrice A sous sa forme de Jordan, le problème revient à étudier les puissances d'un bloc de Jordan. En effet, si $A = P\mathcal{J}P^{-1}$, avec $\mathcal{J}$ composée de tels blocs, l'étude du vecteur $x^n = A^n x^0$ se ramène à celle du vecteur $y^n = P^{-1}x^n$, qui satisfait $y^n = \mathcal{J}^n y^0$.

Soit donc $J \in \mathbb{C}^{r\times r}$ un bloc de Jordan de taille r, associé à la valeur propre μ. écrivant la matrice J comme somme de μI_r et d'une matrice nilpotente, on obtient via la formule du binôme, l'expression

$$J^n = \begin{pmatrix} \mu^n & n\mu^{n-1} & \cdots & \binom{n}{r-1}\mu^{n-r+1} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & n\mu^{n-1} \\ 0 & \cdots & 0 & \mu^n \end{pmatrix}. \tag{3.1}$$

Notons que le coefficient dominant (quand $n \to \infty$) de la matrice J^n est $\binom{n}{r-1}\mu^{n-r+1}$; les termes d'ordre directement inférieur étant $\binom{n}{r-2}\mu^{n-r}$.

Revenant à l'étude de la suite (y^n), si y^0 est générique, alors le vecteur $y^n = \mathcal{J}^n y^0$ satisfait lorsque n tend vers l'infini,

$$y^n \sim \sum_\ell \binom{n}{r-1}\lambda^{n-r+1} y^0_{r+s_\ell-1}\mathbf{e}_{s_\ell}, \tag{3.2}$$

avec

- λ est la valeur propre de plus grand module de A,
- r la taille du plus grand bloc de Jordan associé à λ,
- à chaque indice s_ℓ correspond un bloc de Jordan associé à λ de taille r, situé entre les indices s_ℓ et $s_\ell + r - 1$ de la base canonique $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$.

Ainsi, lorsque $n \to \infty$, le vecteur q^n défini dans l'énoncé converge vers le vecteur

$$x = P \sum_\ell y^0_{r+s_\ell-1}\mathbf{e}_{s_\ell},$$

qui est bien un vecteur propre de A associé à la valeur propre de module maximal λ.

Pour l'étude de la vitesse de convergence, on reprend le raisonnement qui a conduit à l'équivalent (3.2) : le terme suivant dans le développement asymptotique de y^n provient

- soit du second coefficient dominant de J^n lorsque $r > 1$, à savoir $\binom{n}{r-2}\mu^{n-r}$, cf.(3.1),
- soit du coefficient dominant de J'^n lorsque $r = 1$, où J' est le bloc de Jordan associé à une valeur propre de module ρ_2, de taille maximale r'.

Dans le premier cas, on obtient $x^n = \binom{n}{r-1}\lambda^{n-r+1}x + \binom{n}{r-2}\lambda^{n-r}\tilde{x} + \mathcal{O}\big(\binom{n}{r-2}|\lambda|^{n-r}\big)$ où $\tilde{x}$ est un vecteur de $\mathbb{C}^d$. Ainsi, la différence $q^n - x$ prend la forme

$$q^n - x = \left[\frac{|\lambda|}{\lambda}\right]^n \left[\binom{n}{r-1}|\lambda|^{n-r+1}\|x\|\right]^{-1} \binom{n}{r-2}\lambda^{n-r}\tilde{x} + \mathcal{O}\left(\frac{\binom{n}{r-2}|\lambda|^{n-r}}{\binom{n}{r-1}|\lambda|^{n-r+1}}\right).$$

Il apparaît donc que la vitesse de convergence de q^n vers x satisfait

$$\|q^n - x\| = \mathcal{O}\left(\frac{1}{n}\right).$$

Dans le second cas, on a

$$x^n = \lambda^n x + \binom{n}{r'-1}\sum_{|\mu|=\rho_2} \mu^{n-r'+1}x_\mu + \mathcal{O}\left(\binom{n}{r'-1}|\mu|^{n-r'+1}\right).$$

La vitesse de convergence est donc fournie par

$$\|q^n - x\| = \mathcal{O}\left(n^{r'-1}\left|\frac{\rho_2}{\rho_1}\right|^n\right).$$

■

Quelques remarques s'imposent :

- Il est possible de préciser l'expression "pour presque tout x^0" en indiquant que les composantes de x^0 relatives aux directions principales doivent être non-nulles. En pratique, on choisit souvent le vecteur initial x^0 au hasard, auquel cas la condition est presque sûrement satisfaite.
- Dans le cas où plusieurs valeurs propres partagent le même module maximal, on peut parfois utiliser la méthode de la puissance avec astuce. En effet, supposons que $\lambda = -\lambda$ soient les deux seules valeurs propres de module maximal de A. Alors la méthode de la puissance associée à A^2 converge et permet d'obtenir une approximation de λ^2, et donc de λ.

– La méthode peut être utilisée pour calculer d'autres valeurs propres dans le cas où la matrice A est symétrique. En effet, notons $\lambda_1 < \lambda_2 \leq \cdots < \lambda_d$ les valeurs propres de A, supposées distinctes. La méthode de la puissance conduit à une approximation de λ_d, notée $\tilde{\lambda}_d$, associée au vecteur $\tilde{v}_d$, approximation du vecteur propre v_d. Si l'on applique la méthode de la puissance avec $x^0 \in \{v_d\}^\perp$, alors elle converge théoriquement vers la plus grande valeur propre de A restreinte au sous-espace stable $\{v_d\}^\perp$, c'est-à-dire λ_{d-1}. Malheureusement, on n'a accès numériquement qu'à une approximation de v_d et cette remarque ne s'applique pas en pratique car les erreurs vont s'accroître dans la direction de v_d. Toutefois, il est possible de projeter à chaque étape sur le sous-espace $\{\tilde{v}_d\}^\perp$, ce qui revient à appliquer la méthode de la puissance à la matrice $A(I - \tilde{v}_d\tilde{v}_d^T)$. Cette technique porte le nom de *méthode de déflation*. On peut réitérer cette méthode pour approcher λ_{d-2}, mais il faut garder à l'esprit que les erreurs s'accumulent, si bien que la méthode n'est pas très efficace pour calculer *toutes* les valeurs propre de la matrice A.

3.2.2 Méthode de la puissance inverse

La méthode de la puissance inverse revient à appliquer la méthode de la puissance à l'inverse de la matrice A. Elle permet donc d'obtenir sa valeur propre minimale en module. Bien sûr, on n'inverse pas explicitement la matrice A, mais on résout, à chaque étape de l'algorithme, le système linéaire

$$Ax^{n+1} = x^n.$$

(de même que pour la méthode de la puissance, on normalise x^n à chaque itération). Il est utile de remarquer que seul le second membre change dans les systèmes linéaires successifs, et que l'on a intérêt à calculer la décomposition PLU (par exemple) de la matrice A une fois pour toutes et à résoudre seulement deux systèmes triangulaires à chaque étape.

3.2.3 Méthode de la puissance inverse avec translation

Il s'agit ici d'approcher la valeur propre (et le vecteur propre associé) la plus proche d'un nombre complexe μ donné. On applique donc la méthode de la puissance à la matrice $(A-\mu I_d)^{-1}$, et chaque étape nécessite la résolution du système linéaire

$$(A - \mu I_d)x^{n+1} = x^n,$$

avec les mêmes remarques que précédemment. Un paradoxe semble surgir : plus le nombre μ est proche d'une valeur propre, plus le système linéaire est mal conditionné. Ainsi on pourrait croire qu'il est néfaste d'avoir une très bonne approximation de la valeur propre qu'on recherche. Il n'en est rien car les erreurs dues au mauvais conditionnement de la matrice $A-\mu I_d$ sont principalement dans la direction du (ou des) vecteur(s) propre(s) associé(s) à la valeur de plus petit module de $A - \mu I_d$, direction qu'on cherche justement à approcher. De cette manière, loin d'être pénalisant, le mauvais conditionnement du système linéaire favorise la convergence.

3.3 Méthode de Jacobi

Il s'agit d'une méthode itérative de diagonalisation d'une matrice symétrique basée sur des changements de base associés à des matrices de rotation de Givens. Partant de la matrice symétrique A, on construit une matrice orthogonale Ω_n comme produit de matrice de Givens, telle que $\Omega_n^T A \Omega_n$ converge lorsque n tend vers l'infini vers une matrice diagonale contenant les valeurs propres de A. Les colonnes de Ω_n convergent vers une base de vecteurs propres de A dès que toutes les valeurs propres de A sont distinctes.

3.4 Méthode de Givens-Householder

Cette méthode s'appuie sur les propriétés remarquables des polynômes caractéristiques des matrices tridiagonales symétriques. Tout d'abord, on montre que toute matrice symétrique peut être mise sous forme tridiagonale par un procédé algorithmique simple.

Proposition 11.
Soit $A \in \mathbb{R}^{d\times d}$ une matrice symétrique réelle. Alors on peut construire une matrice orthogonale $\Omega \in O(d)$ telle que $\Omega^T A\Omega$ soit tridiagonale symétrique.

PREUVE. On utilise les matrices de Householder, déjà rencontrées dans le chapitre précédent. La matrice A peut s'écrire par blocs

$$\left(\begin{array}{c|c} \alpha & X^T \\ \hline X & B \end{array}\right) \text{ avec } \alpha \in \mathbb{R}, X \in \mathbb{R}^{d-1} \text{ et } B \in \mathbb{R}^{(d-1)\times(d-1)}.$$

On sait qu'il existe une matrice de Householder $h \in O(d-1)$ telle que hX n'ait que sa première composante non-nulle ($h = H(v)$ avec $v = X/\|X\| \pm \mathbf{e}_1$ où $\mathbf{e}_1$ désigne le premier vecteur de la base canonique de $\mathbb{R}^{d-1}$). Ainsi, on a l'identité par blocs

$$\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & h \end{array}\right)\left(\begin{array}{c|c} \alpha & X^T \\ \hline X & B \end{array}\right)\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & h^T \end{array}\right) = \left(\begin{array}{c|c} \alpha & X^Th^T \\ \hline hX & hBh^T \end{array}\right) = \left(\begin{array}{c|c|c} \alpha & \beta & 0 \\ \hline \beta & \gamma & \tilde{X}^T \\ \hline 0 & \tilde{X} & \tilde{A} \end{array}\right).$$

Il suffit alors d'itérer le procédé avec la matrice de taille $(d-1)\times(d-1)$

$$\left(\begin{array}{c|c} \gamma & \tilde{X}^T \\ \hline \tilde{X} & \tilde{A} \end{array}\right).$$

■

On est désormais ramené au cas où la matrice A est tridiagonale symétrique. Notons A_i la matrice

$$A_i = \begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & \ddots & \ddots & \\ & & \ddots & \ddots & b_{i-1} \\ & & & b_{i-1} & a_i \end{pmatrix}.$$

Pour $i \leq d$, on note $p_i = \det(A_i - \lambda I)$ le polynôme caractéristique de A_i (par convention, on pose $p_0 = 1$).

Proposition 12.
La famille $(p_0, \ldots, p_d)$ est une suite de Sturm, c'est-à-dire

1. $\lim_{\mu\to-\infty} p_i(\mu) = +\infty$, *pour* $i = 1, 2, \ldots, d$;
2. $p_i(\mu) = 0 \Rightarrow p_{i-1}(\mu)p_{i+1}(\mu) < 0$, *pour* $i = 1, 2, \ldots d-1$.

PREUVE. Le coefficient dominant de p_i est $(-1)^i X^i$, ce qui implique directement le point 1.

D'autre part, par développement par rapport aux lignes et colonnes, on détermine la relation de récurrence suivante :

$$p_i = (a_i - \lambda)p_{i-1} - b_i^2 p_{i-2}. \tag{3.3}$$

Soit alors μ tel que $p_{i-1}(\mu) = 0$. Si $p_{i-1}(\mu) = 0$, alors de proche en proche on déduit que $p_0(\mu) = 0$, ce qui est impossible. Ainsi le produit $p_{i-2}(\mu)p_i(\mu)$ est strictement négatif, qui prouve le point 2. ■

Propriété 7.

Soit $(p_0, p_1, \ldots, p_d)$ une suite de Sturm. Pour $\mu \in \mathbb{R}$, on note

$$\mathrm{sgn}_i(\mu) = \begin{cases} \textit{signe de } p_i(\mu) & \textit{si } p_i(\mu) \neq 0, \\ \mathrm{sgn}_{i-1}(\mu) & \textit{sinon.} \end{cases}$$

Le nombre de changements de signes dans le d-uplet

$$\big(\mathrm{sgn}_0(\mu), \ldots, \mathrm{sgn}_d(\mu)\big)$$

est égal au nombre de racines de p_d qui sont strictement inférieures à μ.

PREUVE. On montre tout d'abord que les racines des (p_i) sont emboîtées : le polynôme p_1 est donné par $p_1(\lambda) = a_1 - \lambda$, si bien qu'il admet $\lambda_1^1 = a_1$ comme unique racine. Vu la relation $p_0(\lambda_1^1)p_2(\lambda_1^1) < 0$, il vient $p_2(\lambda_1^1) < 0$. Comme, d'autre part, le polynôme p_2 tend vers $+\infty$ en $\pm\infty$, il admet exactement deux racines $\lambda_1^2 < \lambda_1^1 < \lambda_2^2$. En réitérant le procédé, on obtient que le polynôme p_1 admet exactement i racines réelles $\lambda_1^i, \ldots, \lambda_i^i$ qui satisfont

$$\lambda_1^i < \lambda_1^{i-1} < \lambda_2^i < \lambda_2^{i-1} < \cdots < \lambda_{i-1}^i < \lambda_{i-1}^{i-1} < \lambda_i^i.$$

Par ailleurs, d'après le point 2. de la définition d'une suite de Sturm, la définition de sgn_i est correcte. Montrons, par récurrence sur i, que le nombre de changements de signes dans la suite $\big(\mathrm{sgn}_0(\mu), \ldots, \mathrm{sgn}_i(\mu)\big)$ est égal au nombre de racines de p_i qui sont strictement inférieures à μ. Pour $i = 0$, le résultat est clair ; supposons-le acquis au rang $i-1$. On considère alors $\mu \in \mathbb{R}$. Par hypothèse de récurrence, le nombre de changements de signe dans $\big(\mathrm{sgn}_0(\mu), \ldots, \mathrm{sgn}_{i-1}(\mu)\big)$ est égal au nombre k de racines de p_{i-1} strictement inférieures à μ. D'après ce qui précède, on en déduit que p_i admet au moins k racines strictement inférieures à μ, et au plus $k+1$. Il reste à considérer le changement de signe éventuel entre $\mathrm{sgn}_{i-1}(\mu)$ et $\mathrm{sgn}_i(\mu)$. Tout dépend de la position relative de μ avec la racine λ_{k+1}^i. :

- si $\mu = \lambda_{k+1}^i$, alors $\mathrm{sgn}_i(\mu) = \mathrm{sgn}_{i-1}(\mu)$ par définition, il n'y a donc pas de changement de signe et p_i admet bien exactement k racines strictement inférieures à μ ;
- si $\mu < \lambda_{k+1}^i$, alors le point 2. permet de montrer que p_{i-1} et p_i sont de même signe sur l'intervalle $(\lambda_k^{i-1}, \lambda_{k+1}^{i-1})$ donc on n'ajoute pas de changement de signe ;
- si $\mu > \lambda_{k+1}^i$, c'est la situation inverse : p_i et p_{i-1} sont de signes contraires sur $(\lambda_k^i, \lambda_k^{i-1})$ et donc on ajoute un changement de signe entre $\mathrm{sgn}_{i-1}(\mu)$ et $\mathrm{sgn}_i(\mu)$,

ce qui achève la preuve. ■

Grâce aux propriétés des suites de Sturm, on peut mettre en place une méthode dichotomique d'approximation des valeurs propres $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ de la matrice tridiagonale symétrique $A = A_d$. En effet, fixons un entier i entre 1 et d ; voici comment approcher λ_i.

- ◇ on détermine tout d'abord un intervalle $[a_0, b_0]$ contenant le spectre de A (on peut choisir $b_0 = -a_0 = \|A\|$, où $\|\cdot\|$ désigne une norme subordonnée quelconque, par exemple la norme 1 ou ∞ pour faciliter de calcul, ou bien encore utiliser les disques de Gershgorin).
- ◇ on pose $c_0 = \frac{a_0+b_0}{2}$ et on note N_0 le nombre de changements de signes dans la suite $(\mathrm{sgn}_0(c_0), \ldots, \mathrm{sgn}_d(c_0))$.
  - si $N_0 \geq i$, alors au moins i valeurs propres sont inférieures à c_0, donc $\lambda_i \in [a_0, c_0]$,
  - sinon $\lambda_i \in [c_0, b_0]$.
- ◇ on recommence avec l'intervalle de taille moitié ainsi obtenu.

L'algorithme obtenu converge de façon géométrique, avec une raison égale à $\frac{1}{2}$.

Remarque 54.

- Insistons sur le fait que l'évaluation des polynômes est effectuée à l'aide de la formule de récurrence (3.3) et non par sommation des monômes dans la base canonique, ce dernier calcul se révélant souvent instable numériquement (on lui préfère le schéma de Horner dans le cas général).
- La méthode permet d'obtenir une approximation de toutes les valeurs propres de la matrice A. Si l'on souhaite obtenir les vecteurs propres correspondants, on peut utiliser une méthode de puissance inverse avec translation (qui améliore donc aussi l'approximation de la valeur propre obtenue).

3.5 Méthode QR

Cette méthode est très simple à mettre en œuvre, et très performante (c'est celle qui est utilisée – à quelques optimisations près – par la commande `eig` de matlab). Elle est basée sur la décomposition QR d'une matrice quelconque. Soit $A \in \mathbb{R}^{d\times d}$, on construit la suite (A_n) comme suit :

- on note $A_1 = A$ et Q_1, R_1 telles que $A_1 = Q_1R_1$ avec Q_1 orthogonale et R_1 triangulaire supérieure à diagonale positive ;
- la matrice A_2 est définie comme le produit $A_2 = R_1Q_1$.
- $\cdots$
- si A_n est construite, on note $A_n = Q_nR_n$ et on définit A_{n+1} comme R_nQ_n.

Remarquons qu'à chaque étape, on reste dans la classe de similitude de A. En effet, $A_{n+1} = R_nQ_n = Q_n^TA_nQ_n$. De la sorte, si la matrice A_n prend une forme pour laquelle les valeurs propres sont facilement accessibles, on aura déterminé celles de A. En fait, la suite (A_n) devient triangulaire supérieure, comme le montre le résultat suivant (dont les hypothèses ne sont malheureusement pas minimales au regard de la convergence observée numériquement).

Théorème 33.
Soit $A \in \mathbb{R}^{d\times d}$ une matrice diagonalisable inversible. On note $\lambda_1, \ldots, \lambda_d$ ses valeurs propres sur lesquelles on fait l'hypothèse

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_d|.$$

On suppose, de plus, que A s'écrit $A = PDP^{-1}$, avec une matrice P admettant une décomposition LU et D la matrice diagonale telle que $D_{ii} = \lambda_i$.

Alors la méthode QR définie plus haut converge, au sens où

$$\lim(A_n)_{ii} = \lambda_i \quad \textit{et} \quad \lim(A_n)_{ij} = 0, \ \textit{pour } i > j.$$

PREUVE. D'après la remarque préliminaire, $A_{n+1} = Q_n^TA_nQ_n$. Par une récurrence immédiate, on obtient

$$A_{n+1} = \Omega_n^TA\Omega_n, \text{ avec } \Omega_n = Q_1Q_2\cdots Q_n. \tag{3.4}$$

Aussi le comportement asymptotique de la suite (A_n) est-il lié à celui de la suite orthogonale (Ω_n). Pour étudier ce dernier, on va écrire la décomposition QR de la matrice puissance A^n (unique, car A – donc A^n – est inversible) de deux manières différentes.

- ◇ On écrit $A^n = (Q_1R_1)^n = Q_1(R_1Q_1)^{n-1}R_1 = Q_1A_2^{n-1}R_1$, dont on déduit immédiatement par récurrence, $A^n = (Q_1Q_2\cdots Q_n)(R_n\cdots R_2R_1)$, qui n'est autre que la décomposition QR de A^n. Ainsi, Ω_n apparaît comme le facteur orthogonal de cette décomposition.

◇ On diagonalise A^n en PD^nP^{-1}, et si on note $P = QR$ et $P^{-1} = LU$ les décompositions QR et LU des matrices de passage P et P^{-1}, on obtient, si $\widetilde{Q}_n\widetilde{R}_n$ désigne la décomposition QR de la matrice $RD^nLD^{-n}R^{-1}$,

$$A^n = QRD^nLU = Q(RD^nLD^{-n}R^{-1})RD^nU = Q\widetilde{Q}_n\widetilde{R}_nRD^nU.$$

La matrice $Q\widetilde{Q}_n$ est orthogonale et la matrice $\widetilde{R}_nRD^nU$ est triangulaire supérieure. C'est donc presque la décomposition QR de la matrice A^n, à ceci près que la matrice triangulaire n'a pas nécessairement tous ses coefficients diagonaux strictement positifs. Il existe donc une matrice diagonale Λ_n, dont toutes les entrées sont de valeur absolue 1, telle que

$$A^n = [Q\widetilde{Q}_n\Lambda_n][\Lambda_n^{-1}\widetilde{R}_nRD^nU].$$

◇ Par identification des deux décompositions QR il vient

$$\Omega_n = Q\widetilde{Q}_n\Lambda_n. \tag{3.5}$$

Étudions à présent le comportement asymptotique de la suite $(\widetilde{Q}_n)$. Par définition, $\widetilde{Q}_n$ est le facteur orthogonal de la décomposition QR de la matrice $RD^nLD^{-n}R^{-1}$. Or, L étant triangulaire inférieure à diagonale unité, la matrice D^nLD^{-n} converge vers l'identité : en effet,

$$(D^nLD^{-n})_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i < j, \\ \lambda_i^n\lambda_j^{-n}L_{ij} & \text{si } i > j. \end{cases}$$

(noter que $|\lambda_i| < |\lambda_j|$ si $i > j$). En conséquence, la matrice $RD^nLD^{-n}R^{-1}$ tend vers l'identité. D'autre part, la suite $(\widetilde{Q}_n)$ étant compacte, elle admet une sous-suite convergente $(\widetilde{Q}_{n_k})$, de limite $\widetilde{Q}$. Mais alors la suite $(\widetilde{R}_{n_k})$ converge elle aussi (car le produit $\widetilde{Q}_{n_k}\widetilde{R}_{n_k}$ converge vers l'identité). La limite $\widetilde{R}$ de $(\widetilde{R}_{n_k})$ est triangulaire supérieure à diagonale positive, donc $\widetilde{Q}\widetilde{R}$ est la décomposition QR de la matrice I_d. Ainsi $\widetilde{Q} = I_d$. Ainsi, la suite compacte $(\widetilde{Q}_n)$ n'a qu'une valeur d'adhérence, donc converge vers l'identité.

On déduit de l'égalité (3.5) que la suite $(\Omega_n\Lambda_n)$ (noter que $\lambda_n = \lambda_n^T = \lambda_n^{-1}$) converge vers la matrice Q, facteur orthogonal de P. Ainsi

$$\lambda^n A_{n+1}\lambda^n \longrightarrow Q^TAQ = Q^T(PDP^{-1})Q = Q^T(QRDR^{-1}Q^T)Q = RDR^{-1},$$

cette dernière matrice étant triangulaire supérieure avec la diagonale comportant les valeurs propres de A. ■

Remarque 55.

D'après la démonstration précédente, la vitesse de convergence est géométrique, de raison

$$\max_{i>j}\left|\frac{\lambda_i}{\lambda_j}\right| = \max_i\left|\frac{\lambda_i}{\lambda_{i-1}}\right|.$$

Bibliographie

[1] Allaire, Kaber. Algèbre linéaire numérique. Paris, Ellipses, 2002.

[2] Lax. Linear Algebra. New-York, Willey and Sons, 1997.

[3] Lascaux, Théodor. Analyse numérique matricielle appliquée à l'art de l'ingénieur. Paris, Masson, 1987.

[4] Horn, Johnson. Matrix Analysis. Cambridge, Cambridge University Press, 1985.