Analyse numérique matricielle

1 Rappels et compléments d'analyse linéaire

1.1 Notations

$\mathbb{K}$ désigne le corps $\mathbb{R}$ ou $\mathbb{C}$.

Adjoint. V désigne un espace vectoriel de dimension finie n sur $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Lorsqu'une base $e_1, \ldots, e_n$ est fixée sans ambiguité, on identifiera V avec $\mathbb{K}^n$ et l'on notera le vecteur $v = \sum_{i=1}^n v_i e_i,\ v_i \in \mathbb{K}$ par le vecteur colonne

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

vecteur transposé de v : $v^T = (v_1\ \ldots\ v_n)$
vecteur adjoint de v : $v^* = (\overline{v_1}\ \ldots\ \overline{v_n})$

Structures euclidienne et hermitienne Si $\mathbb{K} = \mathbb{R}$, on munit V d'une *structure euclidienne* en définissant sur V le *produit scalaire euclidien* $(\ ,\ ) : V \times V \to \mathbb{R}$ en posant :

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

Si $\mathbb{K} = \mathbb{C}$, on munit V d'une *structure hilbertienne* en définissant sur V le *produit scalaire hermitien* $(\ ,\ ) : V \times V \to \mathbb{C}$ en posant :

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

On parle de produit scalaire canonique si $\mathbb{K}$ n'est pas précisé. La norme $\|u\| = (u, u)^{1/2}$, $u \in \mathbb{K}^n$, qui dérive du produit scalaire canonique est appelée *norme hermitienne* ou *norme euclidienne* suivant que $\mathbb{K} = \mathbb{C}$ ou $\mathbb{K} = \mathbb{R}$.

Orthogonalité. Deux vecteurs u et v de $\mathbb{K}^n$ sont *orthogonaux* si $(u, v) = 0$. Une famille $\{v_1, v_2, \ldots, v_k\}$ de $k \leq n$ vecteurs de $\mathbb{K}^n$ est dite *orthonormale* si $(v_i, v_j) = \delta_{ij},\ i, j = 1, \ldots, k$. Un système orthonormal de $k = n$ vecteurs est une *base orthonormale* de $\mathbb{K}^n$.

Matrices associées On note $M_{m,n}(\mathbb{K})$ le $\mathbb{K}$-espace vectoriel des matrices m lignes et n colonnes et à coefficients dans $\mathbb{K}$.

$$A = \left(a_{ij} \right), 1 \leq i \leq m,\ 1 \leq j \leq n$$

à coefficients dans $\mathbb{K}$ et on note $M_n(\mathbb{K})$ pour $M_{n,n}(\mathbb{K})$.
Soit $A \in M_{m,n}(\mathbb{R})$. On note A^T et on appelle *matrice transposée* de A l'unique matrice de $M_{n,m}(\mathbb{R})$ telle que

$$\forall u \in \mathbb{R}^n, \forall v \in \mathbb{R}^m, (Au, v) = (u, A^T v)$$

Soit $A \in M_{m,n}(\mathbb{C})$. On note A^* et on appelle *matrice adjointe* de A l'unique matrice de $M_{n,m}(\mathbb{C})$ telle que

$$\forall u \in \mathbb{C}^n, \forall v \in \mathbb{C}^m, (Au, v) = (u, A^* v)$$

Exercice 1. Montrer que si $A = \left(a_{ij} \right)_{1 \leq i \leq m\ 1 \leq j \leq n}$, alors $A^* = \left(\overline{a_{ji}} \right)_{1 \leq j \leq n\ \ 1 \leq i \leq m}$

Exercice 2. Pour tous $A \in M_{m,n}(\mathbb{C})$ et $B \in M_{n,p}(\mathbb{C})$, $m, n, p \geq 1$, montrer que

1. $(AB)^* = B^* A^*$
2. Si $A \in M_n(\mathbb{C})$ est inversible, alors A^* est inversible et $(A^*)^{-1} = (A^{-1})^*$.

Matrices particulières Une matrice carrée $A \in M_n(\mathbb{R})$ est :
- *symétrique* si $A = A^T$
- *orthogonale* si $AA^T = A^T A = I$

Une matrice carrée $A \in M_n(\mathbb{C})$ est :
- *hermitienne* ou *auto-adjointe* si $A = A^*$
- *unitaire* si $AA^* = A^* A = I$
- *normale* si $AA^* = A^* A$

Exercice 3. Soit $U \in M_n(\mathbb{C})$. Démontrer les équivalences suivantes :

1. U est unitaire
2. $\forall x, y \in \mathbb{C}^n, (Ux, Uy) = (x, y)$
3. Si $e_1, \ldots, e_n$ est une base orthonormale de $\mathbb{C}^n$, alors $Ue_1, \ldots, Ue_n$ est aussi une base orthonormale de $\mathbb{C}^n$.

Déterminant Soit σ_n le groupe des permutations de $\{1, \ldots, n\}$. On définit le déterminant de $A \in M_n(\mathbb{C})$ par :

$$\det A = \sum_{\sigma \in \sigma_n} \epsilon_\sigma a_{\sigma(1)1} a_{\sigma(2)2} \ldots a_{\sigma(n)n}$$

où ϵ_σ désigne la signature de σ.

Trace Soit $A = \left(a_{ij} \right) \in M_n(\mathbb{C})$. On définit la *trace* de A par :

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

Spectre. Le *spectre* de $A \in M_n(\mathbb{K})$, $n \geq 1$, est l'ensemble des valeurs propres de A :

$$\sigma(A) = \{\lambda \in \mathbb{K},\ A - \lambda I \text{ singulière}\}.$$

Le *rayon spectral* de A est le plus grand des modules des valeurs propres de A :

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

Exercice 4.
1. Démontrer que les valeurs propres d'une matrice hermitienne sont des nombres réels.
2. Démontrer que les valeurs propres d'une matrice unitaire sont des complexes de module 1.

1.2 Réduction matricielle

Le problème de réduction. Soit V un $\mathbb{K}$-e.v. de dimension n et $f : V \to V$ un endomorphisme. Trouver une base dans laquelle la matrice de f soit la plus simple possible.

Matrices semblables. Les matrices A et B de $M_n(\mathbb{K})$ sont dites *semblables* s'il existe $P \in M_n(\mathbb{K})$ inversible telle que $B = P^{-1}AP$.
Interprétation : soit V un $\mathbb{K}$-e.v. de dimension n ; A et B sont matrices d'une même application linéaire $f : V \to V$ dans deux bases resp. $e_1, \ldots, e_n$ et $\epsilon_1, \ldots, \epsilon_n$ de V, et P est la matrice de passage de la base $e_1, \ldots, e_n$ à la base $\epsilon_1, \ldots, \epsilon_n$.

Théorème 1. (Lemme de Schur) Soit $A \in M_n(\mathbb{C})$.
(a) Il existe $U \in M_n(\mathbb{C})$ unitaire telle que $U^{-1}AU$ soit triangulaire.
(b) Si A est normale, il existe $U \in M_n(\mathbb{C})$ unitaire telle que $U^{-1}AU$ soit diagonale.
(c) Si A est hermitienne, il existe $U \in M_n(\mathbb{C})$ unitaire telle que $U^{-1}AU$ soit diagonale réelle.

Décomposition en valeurs singulières. Les matrices A et B de $M_{m,n}(\mathbb{K})$ sont dites *équivalentes* s'il existe $Q \in M_m(\mathbb{K})$ et $P \in M_n(\mathbb{K})$ inversibles telles que $B = QAP$.
Soit $A \in M_n(\mathbb{C})$. Les valeurs propres de la matrice hermitienne A^*A sont des réels ≥ 0. On appelle *valeurs singulières* de A les racines carrées des valeurs propres de A^*A .

Exercice 5. Montrer qu'une matrice carrée est inversible si et seulement si ses valeurs singulières sont > 0.

Théorème 2. Pour tout $A \in M_{m,n}(\mathbb{C})$ où $m \geq n$, il existe deux matrices unitaires $U \in M_m(\mathbb{C})$ et $V \in M_n(\mathbb{C})$ telles que

$$UAV^* = \begin{pmatrix} \mu_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & & \mu_n \\ 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix},$$

les μ_i étant les valeurs singulières de A.

Exercice 6. A désignant une matrice de $M_{m,n}(\mathbb{C})$ où $m \geq n$, on rappelle qu'il existe deux matrices unitaires $U \in M_m(\mathbb{C})$ et $V \in M_n(\mathbb{C})$ telles que

$$A = U\Sigma V^*,\ \Sigma = \begin{pmatrix} \mu_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & & \mu_n \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}.$$

Dans la suite, $u_i \in \mathbb{C}^m$, $1 \leq i \leq m$, désigne la $i^{\text{ième}}$ colonne de la matrice U et $v_k \in \mathbb{C}^n$, $1 \leq k \leq n$, la $k^{\text{ième}}$ colonne de V. On note $(e_i)_{1\leq i\leq n}$ la base canonique de $\mathbb{C}^n$ et $(\epsilon_k)_{1\leq k\leq m}$ celle de $\mathbb{C}^m$.

1. Montrer que $V^* = \sum_{i=1}^n e_i v_i^*$, $\Sigma V^* = \sum_{i=1}^n \mu_i \epsilon_i v_i^*$ et $A = \sum_{i=1}^n \mu_i u_i v_i^*$.
2. Montrer que $A^*A = \sum_{i=1}^n \mu_i^2 v_i v_i^*$.
3. Supposant que $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_r > \mu_{r+1} = \ldots = \mu_n = 0$, $r \geq 1$, en déduire que
   (a) $\ker A = \ker A^*A = \text{vect}\{v_{r+1}, \ldots, v_n\}$
   (b) $\text{Im}\, A = \text{vect}\{u_1, \ldots, u_r\}$
   (c) $\text{Im}\, A^* = \text{vect}\{v_1, \ldots, v_r\} = (\ker A)^\perp$.

1.3 Matrices hermitiennes

Décomposition spectrale. D'après Théorème 1 (c), si $A \in M_n(\mathbb{C})$ est hermitienne, alors A est diagonalisable, ses valeurs propres sont réelles et il existe une b.o.n. de $\mathbb{C}^n$ formée de vecteurs propres de A.

Quotient de Rayleigh. Le *quotient de Rayleigh* de $A \in M_n(\mathbb{C})$ est l'application :

$$R_A : \begin{array}{ccc} \mathbb{C}^n - \{0\} & \to & \mathbb{C} \\ v & \mapsto & \frac{(Av,v)}{(v,v)}. \end{array}$$

Dans la suite, pour alléger les notations, on omettra de préciser que dans l'écriture $R_A(v)$, v ne saurait être nul.

Remarques. 1) Si A est hermitienne, alors $R_A(v) \in \mathbb{R}$ pour tout $v \in \mathbb{C}^n - \{0\}$.
2) Pour tout $v \in \mathbb{C}^n$ et pour tout $\alpha \in \mathbb{C} \setminus \{0\}$, $R_A(\alpha v) = R_A(v)$. Donc si V est un s.e.v de $\mathbb{C}^n$, alors

$$R_A(V \setminus \{0\}) = R_A(\{v \in V \ / \ v^*v = 1\}).$$

En particulier $R_A(V \setminus \{0\})$ est un compact de $\mathbb{C}$ (de $\mathbb{R}$ si A est hermitienne).

Le résultat suivant permet d'exprimer le spectre d'une matrice hermitienne en fonction des valeurs du quotient de Raleight. C'est le *théorème de Courant-Fisher*, appelé aussi *principe du min-max*. Une application de ce résultat sera donnée par la démonstration du théorème 11 :

Théorème 3. Soit $A \in M_n(\mathbb{C})$ une matrice hermitienne dont les valeurs propres sont $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. On note $\{v_1, v_2, \ldots, v_n\}$ une b.o.n. de vecteurs propres associés ($Av_i = \lambda_i v_i$, $i = 1, 2, \ldots, n$), et l'on pose :

$$V_0 = \{0\} \text{ puis } V_k = \text{vect}\{v_1, \ldots, v_k\} \text{ et } \mathcal{V}_k = \{\text{sev de dim } k \text{ de } \mathbb{C}^n\},\ k = 1, \ldots, n.$$

Alors, pour tout $k = 1, \ldots, n$, on a :

(i) $\lambda_k = R_A(v_k) = \max_{u \in V_k} R_A(u) = \min_{u \in V_{k-1}^{\perp}} R_A(u)$.

(ii) $\lambda_k = \min_{W \in \mathcal{V}_k} \max_{w \in W} R_A(w)$.

(iii) $\lambda_k = \max_{W \in \mathcal{V}_{k-1}} \min_{w \in W^{\perp}} R_A(w)$.

(iv) $R_A(\mathbb{C}^n) = [\lambda_1, \lambda_n]$

Positivité. Une matrice $A \in M_n(\mathbb{K})$ supposée hermitienne si $\mathbb{K} = \mathbb{C}$ et symétrique si $\mathbb{K} = \mathbb{R}$, est dite *positive* si elle vérifie

$$(Av, v) \geq 0,\ \forall v \in \mathbb{K}^n,$$

ce qui se note $A \geq 0$. Elle est dite *définie positive* si elle satisfait de plus l'implication :

$$\forall v \in \mathbb{K}^n,\ (Av, v) = 0 \Longrightarrow v = 0, ;$$

et l'on écrit alors $A > 0$.

Proposition 4. Si A est hermitienne alors :

(i) $A \geq 0 \Longleftrightarrow \sigma(A) \subset \mathbb{R}_+$.

(ii) $A > 0 \Longleftrightarrow \sigma(A) \subset \mathbb{R}_+^*$.

Exercice 7. Démontrer la proposition.

1.4 Norme matricielle

Norme vectorielle. Une norme $\|.\|$ sur $\mathbb{K}^n$, $n \geq 1$, est une application de $\mathbb{K}^n$ dans $\mathbb{R}_+$ satisfaisant les trois axiomes suivants :

(i) $\|v\| = 0 \Longleftrightarrow v = 0$.

(ii) $\|\alpha v\| = |\alpha| \|v\|,\ \forall \alpha \in \mathbb{K},\ v \in \mathbb{K}^n$.

(iii) $\|u + v\| \leq \|u\| + \|v\|,\ \forall u, v \in \mathbb{K}^n$.

Pour tout $p \geq 1$, $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$ définit une norme appelée *norme p* sur $\mathbb{K}^n$. Si $p = 2$ il s'agit évidemment de la norme euclidienne ou hermitienne suivant que $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$. De même $\|v\|_\infty = \max_{i=1,\ldots,n} |v_i|$, est une norme appelée *norme infinie*. Elle peut être vue comme la limite des normes p en ce sens que $\|v\|_\infty = \lim_{p \to +\infty} \|v\|_p,\ \forall v \in \mathbb{K}^n$. A toutes fins utiles rappelons l'*inégalité de Hölder*

$$\sum_{i=1}^n |u_i v_i| \leq \|u\|_p \|v\|_q,\ p > 1,\ \frac{1}{p} + \frac{1}{q} = 1,$$

qui, dans le cas particulier où $p = 2$, est appelée *inégalité de Cauchy-Schwarz* :

$$\sum_{i=1}^n |u_i v_i| \leq \|u\|_2 \|v\|_2.$$

Norme matricielle. Une *norme matricielle* sur $M_n(\mathbb{K})$ est une norme vectorielle sur $\mathbb{K}^{n^2}$ qui est continue pour le produit matriciel, en ce sens qu'elle vérifie la condition supplémentaire suivante :

$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in M_n(\mathbb{K}).$$

Norme matricielle induite ou subordonnée. Soit $\|.\|$ une norme vectorielle sur $\mathbb{K}^n$. Pour tout $A \in M_n(\mathbb{K})$ on pose :

$$\|A\| = \sup_{v \in \mathbb{K}^n - \{0\}} \frac{\|Av\|}{\|v\|} = \sup_{v \in \mathbb{K}^n - \{0\}, \ \|v\| \leq 1} \|Av\| = \sup_{v \in \mathbb{K}^n, \ \|v\|=1} \|Av\|.$$

Cette égalité définit une norme matricielle sur $M_n(\mathbb{K})$ que l'on appelle la *norme induite par* $\|.\|$ ou la *norme subordonnée* à $\|.\|$.

Exercice 8. Montrer qu'il existe $u \in \mathbb{K}^n - \{0\}$ tel que $\|Au\| = \|A\|\|u\|$.

Dans le cas particulier où $\|.\| = \|.\|_p$ et $p = 1, 2, \infty$, il est possible de donner une expression explicite de la norme induite associée :

Proposition 5. Si $A = (a_{ij})_{1 \leq i,j \leq n} \in M_n(\mathbb{K})$, $n \geq 1$, alors :

(i) $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|$.

(ii) $\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|$.

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

De plus $\|.\|_2$ est unitairement invariante :

$$\|AU\|_2 = \|UA\|_2 = \|U^*AU\|_2 = \|A\|_2, \ \forall U \in M_n(\mathbb{K}) \text{ unitaire.}$$

Exercice 9. 1) Soit $A \in M_n(\mathbb{C})$. Démontrer que $\rho(A^*A) = \rho(AA^*)$. En déduire que $\|A\|_2 = \|A^*\|_2$.
2) On suppose A normale. Démontrer que $\|A\|_2 = \rho(A)$.

Exercice 10. (*Norme de Frobenius*)

1. Montrer que
$$\|A\|_F = \left(\sum_{1 \leq i,j \leq n} |a_{ij}|^2 \right)^{1/2}$$
définit une norme matricielle sur $M_n(\mathbb{K})$ qui si $n \geq 2$ n'est induite par aucune norme sur $\mathbb{K}^n$.
2. Vérifier que
$$\|A\|_F = \sqrt{tr(A^*A)}$$

La norme $\|.\|_F$ s'appelle aussi norme de Schur, ou norme de Hilbert-Schmidt ou norme 2 de Schatten.

Norme matricielle et rayon spectral.

Théorème 6. Soit $A \in M_n(\mathbb{C})$. Alors

1. Pour toute norme matricielle $\|.\|$ (subordonnée ou non) sur $M_n(\mathbb{C})$,

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

2. Pour tout $\epsilon > 0$, il existe au moins une norme matricielle subordonnée telle que

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

Exercice 11. Démonstration du théorème 6

1. Soit p un vecteur propre de A associé à une valeur propre λ telle que $|\lambda| = \rho(A)$. Soit v le n-vecteur dont toutes les composantes sont égales à 1. Majorer $\|Apv^*\|$ et en déduire 1.
2. Soient $\delta > 0$ et D_δ la matrice diagonale de taille n telle que $(D_\delta)_{i,i} = \delta^{i-1}$. On rappelle que A est unitairement semblable à une matrice triangulaire T (supérieure pour fixer les idées) : il existe U unitaire telle que $U^*AU = T = (t_{i,j})_{1\leq i,j\leq n}$. On pose alors $U_\delta = UD_\delta$, puis l'on définit la norme suivante :

$$\|A\|_\delta = \|U_\delta^{-1} A U_\delta\|_\infty.$$

   (a) Montrer que $\|.\|_\delta$ est la norme matricielle induite par la norme vectorielle $\|U_\delta^{-1} v\|_\infty$.
   (b) Calculer $U_\delta^{-1} A U_\delta$ puis $\|A\|_\delta$ en fonction de δ, des valeurs propres de A et des éléments de la matrice T.
   (c) Etant donné $\epsilon > 0$, montrer que l'on peut choisir $\delta > 0$ pour que $\|A\|_\delta \leq \rho(A) + \epsilon$.

Convergence d'une suite de matrices.

Théorème 7. Soit $A \in M_n(\mathbb{C})$. Les propositions suivantes sont équivalentes :

(i) $A^k \overset{k\to\infty}{\longrightarrow} 0$

(ii) Quel que soit $v \in \mathbb{C}^n$, $A^k v \overset{k\to\infty}{\longrightarrow} 0$

(iii) $\rho(A) < 1$

(iv) Il existe une norme matricielle induite $\|.\|$ telle que $\|A\| < 1$.

Exercice 12. (Inversibilité d'une matrice). Montrer que :

1. Si $I_n - A$ est singulière, alors $\|A\| \geq 1$ pour toute norme matricielle $\|.\|$.
2. Si $\|A\| < 1$ pour une norme matricielle induite $\|.\|$, alors $I_n - A$ est inversible et $\|(I_n - A)^{-1}\| \leq (1 - \|A\|)^{-1}$.

2 Conditionnement

2.1 Conditionnement d'un système linéaire

Un exemple Considérons le système linéaire

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

On modifie "très légèrement" le second membre, et l'on obtient :

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 + \delta u_1 \\ u_2 + \delta u_2 \\ u_3 + \delta u_3 \\ u_4 + \delta u_4 \end{pmatrix} = \begin{pmatrix} 32,1 \\ 22,9 \\ 33,1 \\ 30,9 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ -1,1 \end{pmatrix}$$

Maintenant, modifions "très légèrement" la matrice du système :

$$\begin{pmatrix} 10 & 7 & 8,1 & 7,2 \\ 7,08 & 5,04 & 6 & 5 \\ 8 & 5,98 & 9,89 & 9 \\ 6,99 & 4,99 & 9 & 9,98 \end{pmatrix} \begin{pmatrix} u_1 + \Delta u_1 \\ u_2 + \Delta u_2 \\ u_3 + \Delta u_3 \\ u_4 + \Delta u_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}, \text{ de solution } \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

Conditionnement. Soit $\|.\|$ une norme sur $\mathbb{C}^n$, $n \geq 1$. On note encore $\|.\|$ la norme matricielle subordonnée correspondante. Le *conditionnement* d'une matrice inversible $A \in M_n(\mathbb{K})$ associé à la norme $\|.\|$ est défini par le produit :

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

Les deux résultats suivants donnent des majorations de l'erreur relative sur la solution d'un système linéaire $Au = b$ d'inconnue $u \in \mathbb{C}^n$, avec $A \in M_n(\mathbb{C})$ inversible et $b \in \mathbb{C}^n$ en fonction des variations relatives sur la matrice A et sur le vecteur b. Ils sont énoncés pour une norme vectorielle quelconque sur $\mathbb{C}^n$ et pour sa norme matricielle subordonnée.

Théorème 8. Soit $A \in M_n(\mathbb{C})$ une matrice inversible, $b \neq 0$ et δb des vecteurs de $\mathbb{C}^n$. On note u et $u + \delta u$ les solutions des systèmes linéaires

$$Au = b \quad \text{et} \quad A(u + \delta u) = b + \delta b$$

Alors l'inégalité

$$\frac{\|\delta u\|}{\|u\|} \leq \text{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

est satisfaite, et c'est la meilleure possible : A étant fixée, on peut trouver des vecteurs non nuls b et δb tels que l'égalité soit satisfaite.

Théorème 9. Soient A et $\Delta A \in M_n(\mathbb{C})$ avec A inversible, et soit $b \neq 0$ un vecteur de $\mathbb{C}^n$. On note u et $u + \Delta u$ les solutions des systèmes linéaires

$$Au = b$$

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

Alors l'inégalité

$$\frac{\|\Delta u\|}{\|u + \Delta u\|} \leq \text{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

est satisfaite, et c'est la meilleure possible : A étant fixée, on peut trouver un vecteur $b \neq 0$ et une matrice $\Delta A \neq 0$ tels que l'égalité soit satisfaite.

Les conditionnements de matrice utilisés dans la pratique correspondent aux trois normes usuelles $\|.\|_p,\ p = 1, 2, \infty$. On note : $\text{cond}_p(A) = \|A\|_p\|A^{-1}\|_p$.

Exercice 13. Démontrer les propriétés du conditionnement énoncées ci-dessous :

1. $\text{cond}(A) \geq 1$,
   $\text{cond}(A) = \text{cond}(A^{-1})$,
   Pour tout scalaire $\alpha \neq 0$, $\text{cond}(\alpha A) = \text{cond}(A)$
2. $$\text{cond}_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$$
   où $\mu_1(A) > 0$ et $\mu_n(A) > 0$ désignent respectivement la plus petite et la plus grande des valeurs singulières de A.
3. Si A est une matrice normale, alors
   $$\text{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}$$
   où les $\lambda_i(A)$ sont les valeurs propres de A.
4. Le conditionnement $\text{cond}_2(A)$ d'une matrice unitaire ou orthogonale vaut 1.
5. Le conditionnement $\text{cond}_2(A)$ est invariant par transformation unitaire : soit $A, U \in M_n(\mathbb{C})$, U unitaire. Alors
   $$\text{cond}_2(A) = \text{cond}_2(AU) = \text{cond}_2(UA) = \text{cond}_2(U^*AU)$$

2.2 Conditionnement d'un problème de valeurs propres

Un exemple. Pour $\epsilon \in \mathbb{R}$, considérons la matrice d'ordre n :

$$A(\epsilon)\begin{pmatrix} 0 & & & & & \epsilon \\ 1 & 0 & & & & \\ 0 & 1 & 0 & & & \\ & 0 & 1 & 0 & & \\ & & & \cdots & & \\ & & & 0 & 1 & 0 \end{pmatrix}$$

Pour $\epsilon = 0$, toutes les valeurs propres de la matrice sont nulles. Pour $n = 40$ et $\epsilon = 10^{-40}$, les n valeurs propres de $A(10^{-40})$ sont de module $0, 1$, c'est-à-dire 10^{39} fois la variation de ϵ !
Il est possible sous certaines conditions de contrôler la variation du spectre d'une matrice de référence A, induite par la modification ΔA de ses éléments.

Cas général.

Théorème 10. (Théorème de Bauer-Fike) Soient $A \in M_n(\mathbb{C})$, $n \geq 1$, une matrice diagonalisable : $P^{-1}AP = D = \text{diag}(\lambda_i)_{1\leq i\leq n}$, et $\|.\|$ une norme matricielle telle que pour toute matrice diagonale $diag(d_1, \ldots, d_n)$,

$$\|diag(d_1, \ldots, d_n)\| = \max_i |d_i|$$

Alors pour toute matrice $\Delta A \in M_n(\mathbb{C})$ on a

$$\sigma(A + \Delta A) \subset \cup_{i=1}^n \Delta_i,\ \Delta_i = \{z \in \mathbb{C},\ |z - \lambda_i| \leq (\text{cond}\, P)\|\Delta A\|\}.$$

Exercice 14. Vérifier que les trois normes matricielles $\|.\|_1, \|.\|_2$ et $\|.\|_\infty$ vérifient l'hypothèse de l'énoncé du théorème 10.

Cas hermitien.

Théorème 11. Soient A et $B = A + \Delta A \in M_n(\mathbb{C})$ deux matrices hermitiennes de valeurs propres

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n \text{ et } \beta_1 \leq \beta_2 \leq \ldots \leq \beta_n$$

respectivement. Alors

$$|\beta_k - \alpha_k| \leq \|\Delta A\|_2,\ k = 1, \ldots, n.$$

Exercice 15. (Inégalité de Kantorovitch). On considère une matrice $A \in M_n(\mathbb{R})$, $n \geq 1$, symétrique définie positive. On note $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ ses valeurs propres et $\{v_1, v_2, \ldots, v_n\}$ une base orthonormale de $\mathbb{R}^n$ de vecteurs propres associés ($Av_k = \lambda_k v_k$ pour $k = 1, 2, \ldots, n$).
Le but de cet exercice est d'établir l'inégalité de Kantorovitch :

$$1 \leq \frac{(Ax, x)(A^{-1}x, x)}{\|x\|^4} \leq \frac{(1 + \text{cond}_2 A)^2}{4\text{cond}_2 A},\ \forall x \in \mathbb{R}^n - \{0\}. \tag{1}$$

(a) Rappeler l'expression de $\text{cond}_2 A$ en fonction de λ_1 et λ_n.

(b) On pose $p(\lambda) = \lambda^2 - (\lambda_1 + \lambda_n)\lambda + \lambda_1\lambda_n$ pour tout $\lambda \in \mathbb{R}$.

  (i) Montrer que $p(\lambda_k) \leq 0$ pour tout $k = 1, 2, \ldots, n$.

  (ii) Calculer les valeurs propres de la matrice $B = A^{-1}p(A)$ et en déduire que $(Bx, x) \leq 0$ pour tout $x \in \mathbb{R}^n$.

(c) Pour $x \in \mathbb{R}^n$ fixé, on pose $f(\lambda) = \lambda^2(Ax, x) - (\lambda_1 + \lambda_n)\|x\|^2\lambda + \lambda_1\lambda_n(A^{-1}x, x)$ pour tout $\lambda \in \mathbb{R}$.

  (i) En comparant $f(1)$ et (Bx, x), montrer que $f(0)f(1) \leq 0$, puis que

$$(\lambda_1 + \lambda_n)^2\|x\|^4 - 4(Ax, x)(A^{-1}x, x)\lambda_1\lambda_n \geq 0.$$

  (ii) En déduire alors que :

$$\frac{(Ax, x)(A^{-1}x, x)}{\|x\|^4} \leq \frac{(1 + \text{cond}_2 A)^2}{4\text{cond}_2 A}.$$

(iii) Montrer en examinant le cas particulier du vecteur $x = v_1 + v_n$ que l'inégalité précédente est optimale.

(d) Etablir enfin l'inégalité de gauche dans (1) et montrer qu'elle est optimale.

Exercice 16. Considérons le système linéaire $Au = b$:

$$\begin{pmatrix} 10 & 1 & 4 & 0 \\ 1 & 10 & 5 & -1 \\ 4 & 5 & 10 & 7 \\ 0 & -1 & 7 & 9 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 15 \\ 15 \\ 26 \\ 15 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

et le système linéaire perturbé $A(u + \delta u = b + \delta b$:

$$\begin{pmatrix} 10 & 1 & 4 & 0 \\ 1 & 10 & 5 & -1 \\ 4 & 5 & 10 & 7 \\ 0 & -1 & 7 & 9 \end{pmatrix} \begin{pmatrix} u_1 + \delta u_1 \\ u_2 + \delta u_2 \\ u_3 + \delta u_3 \\ u_4 + \delta u_4 \end{pmatrix} = \begin{pmatrix} 16 \\ 16 \\ 25 \\ 16 \end{pmatrix}, \text{ de solution } \begin{pmatrix} 832 \\ 1324 \\ -2407 \\ 2021 \end{pmatrix}$$

On donne la plus petite et la plus grande valeur propre de la matrice : $\lambda_1 = 0,0005343$ et $\lambda_2 = 19,1225$ Calculer le conditionnement de la matrice et le comparer avec le rapport des erreurs relatives sur u et b.

Exercice 17. (Conditionnement du problème de l'inversion d'une matrice). Soit A une matrice inversible donnée. Si $A + \delta A$ est une matrice inversible, alors

$$\frac{\|(A + \delta A)^{-1} - A^{-1}\|}{\|(A + \delta A)^{-1}\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}$$

et

$$\frac{\|(A + \delta A)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|} (1 + o(\|\delta A\|))$$

Exercice 18. Soit $A = \left(a_{ij} \right)$ une matrice carrée complexe d'ordre n.
1) (Théorème de Gerschgorin-Hadamard) Montrer que

$$\sigma(A) \subset \bigcup_{i=1}^{n} \{z \in \mathbb{C}; |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}$$

2) Montrer que si la matrice A est *strictement diagonalement dominante*, c'est-à-dire si pour tout $i = 1, \ldots, n$, $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, alors elle est inversible.
3) Montrer que si la matrice A est strictement diagonalement dominante, on a l'inégalité

$$|\det A| \geq \prod_{i=1}^{n} (|a_{ii}| - \sum_{j \neq i} |a_{ij}|)$$

Indication : Appliquer 1) à la matrice $A' = DA$ où $D = diag(\delta_i)$ avec

$$\delta_i = \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}|\right)^{-1}, \quad i = 1, \ldots, n$$

3 Méthodes itératives de résolution de systèmes linéaires

3.1 Généralités sur les méthodes itératives

Soit $A \in M_n(\mathbb{C})$ un matrice inversible et soit $b \in \mathbb{C}^n$. On cherche à résoudre le système linéaire

$$Au = b$$

Méthode itérative. Supposons que l'on ait trouvé une matrice B et un vecteur c tels que la matrice $I - B$ soit inversible et tels que la solution du système linéaire $Au = b$ soit aussi la solution du système linéaire $u = Bu + c$.
On obtient alors une *méthode itérative* de résolution du système $Au = b$: on se donne un *vecteur initial* u_0 et on définit une suite de vecteurs $(u_k)_{k\geq 0}$ par :

$$\forall k \geq 0,\ u_{k+1} = Bu_k + c \tag{2}$$

On dit que cette méthode itérative est *convergente* si pour tout vecteur initial u_0,

$$\lim_{k\to\infty} u_k = u$$

Critère de convergence.

Théorème 12. Les assertions suivantes sont équivalentes :

1. La méthode itérative (2) est convergente
2. $\rho(B) < 1$
3. $\|B\| < 1$ pour au moins une norme matricielle $\|.\|$

Calcul de l'erreur.

Proposition 13. Soit $\|.\|$ une norme vectorielle quelconque. Alors

$$\lim_{k\to\infty} \sup_{\|u_0-u\|=1} \|u_k - u\|^{1/k} = \rho(B)$$

Conclusion : le vecteur erreur $e_k = u_k - u$ se comporte "au pire" comme $\rho(B)^k$

Comparaison des méthodes itératives.

Proposition 14. Soit $\|.\|$ une norme vectorielle quelconque. Soit $B, B' \in M_n(\mathbb{C})$ inversibles et $c, c' \in \mathbb{C}$ tels que les systèmes $u = Bu + c$ et $u = B'u + c'$ aient la même solution u. On suppose $\rho(B) < \rho(B')$.
On considère les méthodes itératives

$$\forall k \geq 0,\ u_{k+1} = Bu_k + c,\ \text{et}\ \ u'_{k+1} = B'u'_k + c',\ \text{avec}\ u_0 = u'_0$$

Alors pour tout $\epsilon > 0$, il existe un rang $K \in \mathbb{N}$ tel que

$$\forall k \geq K,\ \sup_{\|u_0-u\|=1} \left(\frac{\|u'_k - u\|}{\|u_k - u\|}\right)^{1/k} \geq \frac{\rho(B')}{\rho(B) + \epsilon}$$

Conclusion : la méthode la plus rapide est celle dont la matrice a le plus petit rayon spectral.

3.2 Description des méthodes de jacobi, de Gauss-Seidel, de relaxation

Les trois méthodes sont basées sur le même principe ; chacune est une amélioration de la précédente.

Le principe Pour résoudre le système $Au = b$, on décompose A sous la forme

$$A = M - N$$

où M est une matrice inversible facile à inverser (diagonale, triangulaire, diagonale par blocs, triangulaire par blocs). Alors $Au = B$ est équivalent à $u = Bu + c$ avec

$$B = M^{-1}N \text{ et } c = M^{-1}b,$$

et l'on applique la méthode itérative associée à ce système.

La méthode de Jacobi Cette méthode s'applique à une matrice d'ordre n $A = \left(\ a_{ij} \ \right)$ inversible et telle que $\forall i, a_{ii} \neq 0$.
On décompose A en $A = D - E - F$ où

$D = \text{diag}(a_{11}, \ldots, a_{nn})$

$(-E)_{ij} = a_{ij}$ si $i > j$, 0 sinon

$(-F)_{ij} = a_{ij}$ si $i < j$, 0 sinon

Autrement dit, avec une notation légèrement abusive :

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & -F & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \mathbf{D} & \ldots & \vdots \\ \vdots & -E & \vdots & & \ldots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & \ldots & a_{nn} \end{pmatrix}$$

La matrice $J = D^{-1}(E + F)$ s'appelle *la matrice de Jacobi* (par points).

La méthode de Gauss-Seidel Elle consiste à poser $M = D - E$ et $N = F$, c'est-à dire à considérer la méthode itérative :

$$u_{k+1} = (D - E)^{-1}Fu_k + (D - E)^{-1}b$$

La matrice $\mathcal{L}_1 = (D - E)^{-1}F$ s'appelle *la matrice de Gauss-Seidel* (par points).

Avantages de la méthode de Gauss-Seidel sur celle de Jacobi :

1. Utilise deux fois moins de mémoire de l'ordinateur
2. Heuristiquement plus efficace car la matrice M prend en compte davantage de coefficients de A.

La méthode de relaxation C'est une amélioration de la méthode de Gauss-Seidel qui consiste à faire intervenir un paramètre réel $\omega \neq 0$ dont la valeur sera choisie de façon à obtenir une plus grande vitesse de convergence. Précisément, on pose :

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

La matrice

$$\mathcal{L}_\omega = \big(\frac{1}{\omega}D - E\big)^{-1}\big(\frac{1-\omega}{\omega}D + F\big)$$

s'appelle *la matrice de relaxation* (par points).

3.3 Convergence de la méthode de relaxation lorsque $\mathbf{A > 0}$

(sous forme de problème)

Dans les parties 1 et 2, on considère une matrice A hermitienne définie positive.

Partie 1

Où l'on démontre une condition suffisante de convergence d'une méthode itérative $M-N$.
1) Vérifier que la matrice $M^* + N$ est hermitienne
2) Montrer que

$$\|v\| = (v^*Av)^{1/2}$$

définit une norme sur $\mathbb{C}^n$. On note encore $\|.\|$ la norme matricielle subordonnée.
3) Soit $v \in \mathbb{C}^n$ tel que $\|v\| = 1$. On pose $w = M^{-1}Av$. Démontrer que

$$\|v - w\|^2 = 1 - w^*(M^* + N)w$$

4) En déduire que si la matrice $M^* + N$ est définie positive, alors $\|M^{-1}N\| < 1$.
5) En déduire que si la matrice $M^* + N$ est définie positive, alors $\rho(M^{-1}N) < 1$.

Partie 2

Où l'on démontre une condition suffisante de convergence de la méthode de relaxation.
On considère les matrices M et N de la méthode de relaxation.
1) Démontrer que

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

2) Démontrer que la matrice D est définie positive.
3) En déduire que $M^* + N$ est définie positive si et seulement si $0 < w < 2$.
4) Montrer que si $0 < w < 2$, alors la méthode de relaxation converge.

Partie 3

Où l'on démontre une condition suffisante de divergence de la méthode de relaxation
Dans cette partie, on ne fait aucune hypothèse sur A
1) Démontrer que $\det \mathcal{L}_\omega = (1-\omega)^n$.
2) Démontrer que $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$.
3) Conclure

3.4 Convergence comparée des trois méthodes ; cas A tridiagonale

Comparaison des méthodes de Jacobi et de Gauss-Seidel

Théorème 15. Soit A une matrice tridiagonale. Alors les rayons spectraux des matrices de Jacobi et de Gauss-Seidel sont liés par la relation :

$$\rho(\mathcal{L}_1) = \rho(J)^2$$

En conséquence, les deux méthodes convergent ou divergent simultanément. Lorsqu'elles convergent, la méthode de Gauss-Seidel converge plus rapidement que la méthode de Jacobi.

Comparaison des méthodes de Jacobi et de relaxation

Théorème 16. Soit A une matrice tridiagonale dont tous les termes diagonaux sont réels. Alors la méthode de Jacobi et la méthode de relaxation convergent ou divergent simultanément. Lorsqu'elles convergent, le graphe de la fonction

$$\omega \in]0, 2[\longmapsto \rho(\mathcal{L}_\omega)$$

a l'allure indiquée ci-dessous. Le minimum est atteint en

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

et

$$\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1$$

Problème

Étude détaillée d'un problème aux limites en dimension 1

On s'intéresse à la situation physique suivante : une poutre horizontale de longueur 1 appuyée simplement à ses extrémités est étiré selon son axe par une force P et est soumise à une charge transversale $f(x)dx$ par unité de longueur dx. Alors l'ordonnée $u(x)$ du point d'abcisse x de la poutre, appelée *moment fléchissant* en x, est solution du *problème aux limites* :

$$A = \begin{cases} -u"(x) + c(x)u(x) = f(x),\ 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases}$$

où $x \mapsto c(x)$ est une fonction positive dépendant du matériau constituant la poutre. On peut démontrer (on l'admettra ici) que ce problème aux limites possède une unique solution $\phi : [0,1] \to \mathbb{R}^4$ de classe C^4.
On ne connait pas de méthode qui permettrait de trouver une formule exacte pour décrire ϕ. L'objet de ce problème est la mise en œuvre de la méthode des différences finies afin d'approcher la solution ϕ d'aussi près que l'on veut.

Partie 1

Méthode des différences finies

Étant donné un entier $N \geq 1$, on pose :

$$h = \frac{1}{N+1}$$

et on définit un *maillage uniforme* de l'intervalle $[0,1]$ comme étant l'ensemble des points $x_i = ih,\ i \in \{0,1,2,\ldots,N,N+1\}$.
La méthode des différences finies est un moyen d'obtenir une approximation de ϕ aux nœuds du maillages, d'une qualité d'autant meilleure que le pas h est petit. Autrement dit, on cherche un vecteur

$$u^u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \in \mathbb{R}^N$$

tel que u_i soit "voisin" de $\phi(x_i)$ pour tout $i = 1,\ldots,N$.
1) Soit $i = 1,\ldots,N$. Écrire les formules de Taylor-Lagrange à l'ordre 3 pour la fonction ϕ sur les intervalles $[x_{i-1}, x_i]$ et $[x_i, x_{i+1}]$ pour exprimer $\phi(x_{i-1})$ et $\phi(x_{i+1})$ en fonction de $\phi(x_i), \phi'(x_i), \phi"(x_i)$ et $\phi^{(3)}(x_i)$.
2) Montrer qu'il existe $\xi_i \in]x_{i-1}, x_{i+1}[$ tel que

$$-\phi(x_{i+1}) + 2\phi(x_i) - \phi(x_{i-1}) = -h^2\phi"(x_i) - \frac{h^4}{12}\phi^{(4)}(\xi_i)$$

3) En déduire

$$-\phi"(x_i) = \frac{-\phi(x_{i+1}) + 2\phi(x_i) - \phi(x_{i-1})}{h^2} + \frac{h^2}{12}\phi^{(4)}(\xi_i)$$

4) Pour alléger les notations, posons pour $1 < i < N, \phi_i = \phi(x_i), c_i = c(x_i)$ et $f_i = f(x_i)$. En exprimant que ϕ est solution du problème aux limites, montrer que pour tout $i = 1, \ldots, N$,

$$\frac{-\phi_{i-1} + 2\phi_i - \phi_{i+1}}{h^2} + c_i\phi_i = f_i - \frac{h^2}{12}\phi^{(4)}(\xi_i)$$

5) Vérifier que le système d'équations obtenu en négligeant les termes $\frac{h^2}{12}\phi^{(4)}(\xi_i)$, s'écrit sous la forme matricielle :

$$A_h\phi_h = b_h$$

avec

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2+c_1h^2 & -1 & & & \\ -1 & 2+c_2h^2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+c_{N-1}h^2 & -1 \\ & & & -1 & 2+c_Nh^2 \end{pmatrix}$$

$$\phi_h = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N-1} \\ \phi_N \end{pmatrix}, \qquad b_h = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}$$

Partie 2

Étude spectrale détaillée d'une matrice de différences finies

Dans cette partie, nous allons comparer les méthodes de Jacobi, de Gauss-Seidel et de relaxation dans le cas où $c = 0$, c'est-à-dire lorsque la force P est nulle. Le système $A_h\phi_h = b_h$ est alors équivalent au système $A\phi_h = h^2b_h$, où

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

1) Notons $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A. Montrer que les valeurs propres de la matrice de Jacobi J associée à A sont les réels

$$1 - \frac{\lambda_i}{2},\ i = 1, \ldots, n$$

2) On considère la suite récurrente linéaire double définie par :

$$(*) \qquad x_0, x_1,\ \forall k \geq 1, x_{k+1} = -x_{k-1} + (2-\lambda)x_k$$

a) Démontrer que λ est valeur propre de A si et seulement si il existe $x_1 \in \mathbb{R}$ tel que la suite récurrente $(*)$ de données initiales $x_0 = 0, x_1$ vérifie $x_{n+1} = 0$.

b) La suite récurente $(*)$ peut s'écrire sous forme matricielle

$$y_0 = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} \text{ et } \forall k \geq 0, y_{k+1} = Cy_k$$

avec

$$\forall k, y_k = \begin{pmatrix} x_k \\ x_{k+1} \end{pmatrix} \text{ et } C = \begin{pmatrix} 0 & 1 \\ -1 & 2-\lambda \end{pmatrix}$$

On suppose $\lambda \in]0,4[$. Démontrer que les valeurs propres de C sont deux complexes conjugués μ et $\overline{\mu}$ distincts de module 1.
c) Soit $(x_k)_k$ une suite définie par $(*)$. Démontrer qu'il existe deux complexes a et b tels que

$$\forall k \geq 0,\ x_k = a\mu^k + b\overline{\mu}^k$$

4) Montrer que les valeurs propres de la matrice de Jacobi J sont les n réels

$$\cos \frac{m\pi}{(n+1)}, m = 1, \ldots, n$$

et que les valeurs propres de la matrice de Jacobi
5) Montrer que

$$\rho(J) = 1 - \frac{\pi^2}{2n^2} + o\big(\frac{1}{n^4}\big) \text{ et } \rho(\mathcal{L}_1) = 1 - \frac{\pi^2}{n^2} + o\big(\frac{1}{n^4}\big)$$

6) En déduire un équivalent du paramètre optimal ω_0 et de $\rho(\mathcal{L}_{\omega_0})$ pour les grandes valeurs de n.

Partie 3

Estimation des erreurs

1) Soit

$$u_0,\ u_{k+1} = Bu_k + c$$

une méthode itérative convergente de matrice B.
a) Soit u la solution du système $u = Bu + c$. On considère le vecteur erreur $e_k = u_k - u$ pour $k \geq 1$. Montrer que

$$\frac{\| e_k\|_2}{\|e_0\|_2} \leq \|B\|_2^k$$

b) En déduire que si B est normale, alors le nombre de pas d'itérations nécessaire pour diviser l'erreur par 2 est :

$$k \geq -\frac{\ln 2}{\ln \rho(B)}$$

2) Pour n grand, donner un équivalent du nombre de pas d'itérations nécessaire pour diviser l'erreur par 2 dans les méthodes de Jacobi, de Gauss-Seidel et de relaxation.

4 Calculs de valeurs propres

Nous étudions dans ce chapitre quatre méthodes itératives pour calculer des approximation des valeurs propres d'une matrice et éventuellement de ses vecteurs propres :

Pour une matrice symétrique :

- Méthode de Jacobi
- Méthode de Givens-Householder

Pour une matrice quelconque :

- Méthode QR
- Méthode de la puissance

4.1 La méthode de Jacobi

C'est une méthode que l'on met en œuvre lorsque que l'on cherche à calculer toutes les valeurs propres, et éventuellement tous les vecteurs propres d'une matrice symétrique.
Soit $A \in M_n(\mathbb{R})$, $n \geq 1$, une matrice symétrique dont les valeurs propres sont notées $\lambda_1, \lambda_2, \ \ldots, \lambda_n$.

Le principe La méthode de Jacobi consiste à construire une suite $(O_k)_{k\geq 1}$ de matrices orthogonales telle que

$$\lim_{k\to+\infty} O_k^T A O_k = \mathrm{diag}(\lambda_i)_{1\leq i\leq n}.$$

De plus, si les valeurs propres de A sont 2 à 2 distinctes, la suite de matrices $(O_k)_{k\in\mathbb{N}}$ converge vers une matrice O dont la i-ème colonne est un vecteur propre de A associé à λ_i.
Les matrices O_k sont des produits $O_k = \Omega_1\Omega_2\ldots\Omega_k$ de matrices de rotation élémentaires. Posons $A_k = O_k^T A O_k$; le principe de chaque transformation $A_k \to A_{k+1} = \Omega_k^T A_k \Omega_k$ est d'annuler deux éléments hors diagonaux de A_k en position symétrique.

Élimination d'éléments hors diagonaux On considère la matrice de rotation élémentaire dans $\mathbb{R}^n$ des p-ième et q-ième vecteurs de base, $1 \leq p < q \leq n$:

$$\Omega = \Omega(p,q,\theta) \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \cdots & s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & 1 & & & \\ & & & -s & \cdots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}, \ c = \cos\theta, \ s = \sin\theta.$$

Etant donnée une matrice symétrique $A = (a_{ij})_{1\leq i,j\leq n}$, on définit :

$$B = \Omega^T A \Omega = (b_{ij})_{1\leq i,j\leq n}.$$

Proposition 17. Si $a_{pq} \neq 0$, alors il existe un unique $\theta \in]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}]$ tel que $b_{pq} = 0$.

Exercice 19. Démonstration de la proposition
1) Montrer que :

$$\left\{\begin{array}{lcll} b_{ij} & = & a_{ij} & \text{si } i \neq p,q \text{ et } j \neq p,q \\ b_{pj} & = & ca_{pj} - sa_{qj} & \text{si } j \neq p,q \\ b_{qj} & = & sa_{pj} + ca_{qj} & \text{si } j \neq p,q \\ b_{pp} & = & c^2 a_{pp} + s^2 a_{qq} - 2csa_{pq} & \\ b_{qq} & = & s^2 a_{pp} + c^2 a_{qq} + 2csa_{pq} & \\ b_{pq} & = & (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}). & \end{array}\right.$$

2) On suppose $a_{pq} \neq 0$. Montrer que $b_{pq} = 0$ équivaut à :

$$cotan(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}}.$$

3) Conclure

Exercice 20.

1. Démontrer que quel que soit θ,

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

indication : utiliser la norme de Frobenius et l'exercice 10.

2. On choisit θ tel que $b_{pq} = 0$. Démontrer que

$$\sum_{i=1}^{n} b_{ii}^2 = \sum_{i=1}^{n} a_{ii}^2 + 2a_{pq}^2$$

et que

$$\begin{array}{rcl} b_{pp} & = & a_{pp} - \tan\theta\ a_{pq} \\ b_{qq} & = & a_{qq} + \tan\theta\ a_{pq} \end{array}$$

Description de la méthode de Jacobi classique. L'idée est donc de diagonaliser A par une suite (infinie) de transformations semblables orthogonales :

$$\left\{\begin{array}{rcl} A_1 & = & A \\ A_{k+1} & = & \Omega_k^T A_k \Omega_k,\ k \geq 1, \end{array}\right.$$

où Ω_k est une matrice de rotation élémentaire choisie pour annuler un élément non diagonal de A_k. Dans la méthode décrite ici, appelée méthode de Jacobi classique, on choisit un élément de plus grand module. Plus précisément, pour chaque matrice $A_k = (a_{ij}^{(k)})_{1\leq i,j\leq n}$, on détermine les entiers $1 \leq p_k < q_k \leq n$ pour lesquels

$$|a_{p_k q_k}^{(k)}| = \max_{1\leq i\neq j\leq n} |a_{ij}^{(k)}|,$$

puis on ajuste $\theta_k \in]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}]$ de sorte que $a_{p_k q_k}^{(k+1)} = 0$ avec $\Omega_k = \Omega(p_k, q_k, \theta_k)$.

Convergence de la méthode. On rappelle les notations :

$$A_{k+1} = \Omega_k^T A_k \Omega_k = O_k^T A O_k \quad \text{où } O_k = \Omega_1 \Omega_2 \dots \Omega_k$$

Théorème 18. 1. La suite $(A_k)_{k\geq 1}$ est convergente, et à une permutation près σ des valeurs propres λ_i, on a :

$$\lim_{k\to+\infty} A_k = \operatorname{diag}(\lambda_{\sigma(i)}).$$

2. Si toutes les valeurs propres de A sont distinctes, alors la suite $(O_k)_{k\geq 1}$ converge vers une matrice orthogonale O dont les vecteurs colonnes constituent un ensemble orthonormal de vecteurs propres de la matrice A.

Exercice 21. (Démonstration du théorème)

1. Démontrer le lemme suivant :
   Lemme. *Soit X un espace vectoriel normée de dimension finie et soit $(x_k)_k$ une suite bornée dans X, admettant un nombre fini de valeurs d'adhérence, et telle que $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$. Alors la suite (x_k) est convergente.*
2. Dans cette question, on munit l'espace vectoriel $M_n(\mathbb{R})$ de la norme de Frobenius

$$\|A\| = \left(\sum_{1\leq i,j\leq n} |a_{ij}|^2 \right)^{1/2}$$

   Pour $k \geq 1$, on pose $D_k = \operatorname{diag}(a_{ii}^{(k)})$.
   (a) Montrer que la suite (D_k) est bornée.
   (b) Posons pour tout entier k : $A_k = D_k + B_k$. Démontrer que

$$\|B_{k+1}\|^2 \leq \left(1 - \frac{2}{n(n-1)}\right) \|B_k\|^2$$

   et en déduire que

$$\lim_{k\to\infty} \|B_k\| = 0$$

   (c) Soit $(D_{\varphi(k)})_k$ une suite extraite de $(D_k)_k$ convergeant vers D (forcément diagonale !).
      (i) Montrer que A et D ont même polynôme caractéristique puis qu'il existe une permutation σ de $\{1, 2, \dots, n\}$ telle que

$$D = \operatorname{diag}(\lambda_{\sigma(i)})_{1\leq i\leq n}.$$

      (ii) En déduire que la suite $(D_k)_k$ n'a qu'un nombre fini de valeurs d'adhérence.
   (d) Montrer que $D_{k+1} - D_k \overset{k\to+\infty}{\longrightarrow} 0$.
   (e) En déduire que la suite (D_k) converge vers l'une de ses valeurs d'adhérence D.
   (f) En déduire la partie 1 du théorème.

3. Dans cette question, on munit l'espace vectoriel $M_n(\mathbb{R})$ de la norme subordonnée à la norme euclidienne. Soit O une matrice orthogonale telle que $O^TAO = \text{diag}(\lambda_i)$.
   (a) Montrer que la suite O_k est bornée
   (b) Notons $p_1, \ldots, p_n$ ses vecteurs colonnes. Démontrer que la suite (O_k) n'a qu'un nombre fini de valeurs d'adhérence, dont les vecteurs colonnes sont nécessairement de la forme $\pm p_{\sigma(1)}, \ldots, \pm p_{\sigma(n)}$.
   (c) (i) Montrer qu'il existe un entier l tel que pour tous $\alpha, \beta \in \{1, \ldots, n\}$,
   $$k \geq l \Rightarrow |a^k_{\beta\beta} - a^k_{\alpha\alpha}| \geq \frac{1}{2} \min_{i \neq j} |\lambda_i - \lambda_j| > 0$$
   (ii) En déduire que
   $$\lim_{k\to\infty} \theta_k = 0 \text{ et que } \lim_{k\to\infty} \Omega_k = I$$
   (iii) En déduire que $\lim_{k\to\infty}(O_{k+1} - O_k) = 0$.
   (d) En déduire la partie 2 du théorème.

Exercice 22. Soit $a \in \mathbb{R}$ et soient λ_1 et λ_2 les valeurs propres de la matrice

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$$

On pose

$$J = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

1) Trouver c et s, tels que $c^2 + s^2 = 1$ et tels $(J^TAJ)_{11} = a$. Montrer qu'il faut $a \in [\lambda_1, \lambda_2]$ pour avoir des solutions réelles.
2) En déduire que si $A \in M_n(\mathbb{R})$ est symétrique, il existe U orthogonale telle que $U^TAU = \frac{1}{n}tr(A)I + B$ avec $B_{ii} = 0, i = 1, \ldots, n$.

4.2 La méthode de Givens-Householder.

Il s'agit d'une méthode bien adaptée à la recherche de valeurs propres sélectionnées d'une matrice symétrique A, par exemple toutes les valeurs propres situées dans un intervalle fixé à l'avance. En revanche, cette méthode ne fournit pas les vecteurs propres.

Le principe La méthode de Givens-Householder comprend deux étapes :

(a) *Réduction.* On détermine, par la méthode de réduction de Householder, une matrice O orthogonale (obtenue comme produit de $n-2$ matrices de Householder) telle que la matrice O^TAO soit tridiagonale, i.e.

$$O^TAO = \begin{pmatrix} b_1 & c_1 & 0 & \ldots & 0 \\ c_1 & b_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c_{i-2} & b_{n-1} & c_{n-1} \\ 0 & \ldots & 0 & c_{n-1} & b_n \end{pmatrix}, \; b_i \in \mathbb{R}, \; c_j \in \mathbb{R}^*.$$

(b) *Bissection.* On est ainsi ramené au calcul des valeurs propres d'une matrice symétrique tridiagonale, qui s'effectue par la méthode de bissection de Givens.

Réduction de Householder Soit p un entier ≥ 1. On appelle *matrice de Householder* toute matrice de $M_p(\mathbb{R})$ du type

$$H_w^{(p)} = I - 2\frac{ww^T}{w^T w},$$

où $w \in \mathbb{R}^p$.

Exercice 23. Vérifier que toute matrice de Householder est symétrique et orthogonale.

Théorème 19. Étant donnée une matrice symétrique A, il existe une matrice O, produit de $(n-2)$ matrices de Householder, telle que la matrice $O^T AO$ soit tridiagonale.

Exercice 24. (Démonstration du théorème)

1. Pour tout $x \in \mathbb{R}^p$, $p \geq 1$, on pose

$$u_\mp = x \mp \|x\| e_1, \text{ où } x = \sum_{i=1}^{p} x_i e_i,$$

la famille $\{e_1, e_2, \ldots, e_p\}$ désignant la base canonique de $\mathbb{R}^p$. En remarquant que $\|u_\mp\|^2 = 2(\|x\|^2 \mp \|x\| x_1)$, montrer que $H_{u_\mp}^{(p)} x = \pm \|x\| e_1$.

2. Soit $A = (a_{ij})_{1 \leq i,j \leq n} \in M_n(\mathbb{R})$, $n \geq 1$, une matrice symétrique.

   (a) Soient $c = (a_{21}, a_{31}, \ldots, a_{n1})^T \in \mathbb{R}^{n-1}$ et

$$O_1 = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & & & & \\ 0 & & & & \\ \vdots & & & H_{c \mp \|c\| e_1}^{(n-1)} & \\ 0 & & & & \\ 0 & & & & \end{pmatrix}.$$

   Montrer que O_1 est une matrice de Householder et que $A_1 = O_1^T A O_1$ est de la forme :

$$A_1 = \begin{pmatrix} a_{11} & \pm\|c\| & 0 & \ldots & 0 \\ \pm\|c\| & & & & \\ 0 & & & & \\ \vdots & & & A' & \\ 0 & & & & \\ 0 & & & & \end{pmatrix} \text{ avec } A' = (a'_{ij})_{1 \leq i,j \leq n_1} = (A')^T \in M_{n-1}(\mathbb{R}).$$

   (b) Si $c' = (a'_{21}, a'_{31}, \ldots, a'_{(n-1)1})^T \in \mathbb{R}^{n-2}$ et

$$O_2 = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & \\ 0 & 0 & & & \\ \vdots & \vdots & & H_{c' \mp \|c'\| e_1}^{(n-2)} & \\ 0 & 0 & & & \\ 0 & 0 & & & \end{pmatrix},$$

montrer que $(O_1O_2)^T A O_1 O_2$ est de la forme :

$$A_2 = \begin{pmatrix} * & * & 0 & 0 & \dots & 0 \\ * & * & * & 0 & \dots & 0 \\ 0 & * & & & & \\ 0 & 0 & & & & \\ \vdots & \vdots & & A'' & & \\ 0 & 0 & & & & \\ 0 & 0 & & & & \end{pmatrix} \text{ avec } A'' = (A'')^T \in M_{n-2}(\mathbb{R}).$$

(c) En déduire qu'il existe une matrice orthogonale O telle que $O^T AO$ est tridiagonale.

Maintenant, nous allons décrire la méthode de Givens pour trouver les valeurs propres d'une matrice tridiagonale. On remarque que si l'un des c_i est nul, alors la matrice $B = O^T AO$ se décompose en deux sous matrices tridiagonales dont on peut chercher indépendamment les valeurs propres pour obtenir les valeurs propres de B. On peut donc supposer srlg que les c_i sont tous non nuls.

Suite de Sturm Pour tout $i = 1, 2, \dots, n$, on définit la famille de polynômes p_i à l'aide de la relation de récurrence suivante :

$$\begin{cases} p_0(x) = 1 \\ p_1(x) = b_1 - x \\ p_i(x) = (b_i - x)p_{i-1}(x) - c_{i-1}^2 p_{i-2}(x) \text{ si } 2 \leq i \leq n. \end{cases}$$

Exercice 25. Démontrer que p_i est le polynôme caractéristique de la matrice

$$B_i = \begin{pmatrix} b_1 & c_1 & 0 & \dots & 0 \\ c_1 & b_2 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c_{i-2} & b_{i-1} & c_{i-1} \\ 0 & \dots & 0 & c_{i-1} & b_i \end{pmatrix}.$$

Le théorème suivant exprime que les racines des polynômes p_i ont des propriétés d'emboitement remarquables, illustrées sur la figure ci-dessous.

Proposition 20.
1. $\lim_{x \to -\infty} p_i(x) = +\infty$, $1 \leq i \leq n$.
2. $p_i(t) = 0 \Longrightarrow p_{i-1}(t)p_{i+1}(t) < 0$, $1 \leq i \leq n-1$.
3. p_i a i racines réelles distinctes qui séparent les $(i+1)$ racines du polynôme $p_{i+1}, 1 \leq i \leq n-1$ comme l'indique la figure ci-dessous.

Exercice 26. démontrer la proposition

Une famille de polynômes vérifiant les propriétés 1, 2 et 3 de la proposition s'appelle une *suite de Sturm*. La méthode de Givens repose sur une proporiété remarquable d'une telle suite, énoncée dans le théorème suivant :

Théorème 21.

Le nombre de racines de $p_i < \mu \in \mathbb{R}$ est le nombre $N(i, \mu)$ de changements de signes dans la suite $(1, p_1(\mu), \ldots, p_i(\mu))$.

Bissection de Givens Soit A une matrice tridiagonale symétrique réelle et soient $\lambda_1 \leq \ldots \leq \lambda_n$ ses valeurs propres. Fixons $i \in \{1, \ldots, n\}$. Supposons que l'on veuille approcher la valeur propre λ_i. On procède par dichotomie en appliquant de façon répétée le théorème précédent :
On commence par choisir un intervalle $[a_0, b_0]$ qui contient λ_i, par exemple $[a_0, b_0] = [-\|A\|, \|A\|]$ (car $\rho(A) \leq \|A\|$). notons $c_0 = \frac{a_0+b_0}{2}$ le milieu de l'intervalle $[a_0, b_0]$.

- si $N(n, c_0) \geq i$, alors $\lambda_i \in [a_0, c_0[$ et l'on itère le procédé sur l'intervalle $[a_1, b_1] = [a_0, c_0]$
- si $N(n, c_0) < i$, alors $\lambda_i \in [c_0, b_0]$ et l'on itère le procédé sur l'intervalle $[a_1, b_1] = [c_0, b_0]$

On détermine ainsi une suite d'intervalles emboités $[a_k, b_k], k \geq 0$ tels que

$$\forall k \geq 0,\ \lambda_i \in [a_k, b_k] \text{ et } b_k - a_k = 2^{-k}(b_0 - a_0)$$

On peut donc encadrer λ_i avec une précision arbitrairement grande.

4.3 La méthode QR

4.3.1 La factorisation QR d'une matrice

Préliminaire : matrices de Householder. A tout vecteur $v \in \mathbb{C}^n - \{0\}$ on associe la *matrice de Householder* :

$$H(v) = I - 2\frac{vv^*}{v^*v} = I - \frac{2}{\|v\|^2}vv^*,$$

où $\|v\| = \|v\|_2$ est la norme hermitienne de v.

(a) Vérifier que $H(v)$ est hermitienne et unitaire.

(b) Etant donné $x = \sum_{i=1}^n x_i e_i$ un vecteur de $\mathbb{C}^n$ (la famille $\{e_1, e_2, \ldots, e_n\}$ désignant la base canonique de $\mathbb{C}^n$) tel que $x_1 = e^{i\theta}|x_1| \neq 0$ et $\sum_{i=2}^n |x_i| > 0$, vérifier que

$$H(x \mp \|x\|e^{i\theta}e_1)x = \pm\|x\|e^{i\theta}e_1.$$

Théorème 22. (Factorisation QR d'une matrice) Soit $A \in M_n(\mathbb{C})$ une matrice quelconque. Il existe une matrice unitaire Q, produit de $n-1$ matrices de Householder, et une matrice triangulaire supérieure R telles que

$$A = QR$$

De plus, il est possible de choisir les coefficients diagonaux de R tous réels ≥ 0, auquel cas, si A est inversible, alors la factorisation QR est unique.

Démonstration du théorème Soit $A = (a_{ij})_{i,j} \in M_n(\mathbb{C})$. Nous allons trouver $n-1$ matrices de Householder $H_1, H_2, \ldots, H_{n-1}$ telle que $H_{n-1}\ldots H_2H_1A$ soit triangulaire supérieure.

Construction de H_1. La matrice H_1 doit permettre de "faire apparaître des zéros" à la place des $n-1$ dernières composantes de la première colonne $a_1 = (a_{11}, a_{21}, a_{31}, \cdots, a_{n1}) \in \mathbb{C}^n$ de la matrice A :

$$A_2 = H_1A_1 = \begin{pmatrix} * & * & \ldots & * \\ 0 & * & \ldots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix}.$$

(a) Si $\sum_{i=2}^n |a_{i1}| > 0$, alors il existe $w_1 \in \mathbb{C}^n$ tel que $H_{w_1}^{(n)}a_1$ a toutes ses composantes nulles à l'exception de la première. Ainsi $H_1 = H_{w_1}^{(n)}$ convient.

(b) Si $\sum_{i=2}^n |a_{i1}| = 0$, alors il suffit de choisir $H_1 = I_n$.

Construction de H_2. Soit $a_2 = (a_{22}^{(2)}, a_{32}^{(2)}, \cdots, a_{n2}^{(2)})$ le vecteur de $\mathbb{C}^{n-1}$ formé par les $n-1$ dernières composantes de la deuxième colonne de A_2.

(a) Si $\sum_{i=3}^{n} |a_{i2}^{(2)}| > 0$, il existe $\tilde{w}_2 \in \mathbb{C}^{n-1}$ tel que $H_{\tilde{w}_2}^{(n-1)} a_2$ a toutes ses composantes nulles à l'exception de la première. On pose alors :

$$H_2 = \begin{pmatrix} 1 & 0 \\ 0 & H_{\tilde{w}_2}^{(n-1)} \end{pmatrix}.$$

En notant w_2 le vecteur de $\mathbb{C}^n$ dont la première composante est nulle et les $n-1$ dernières sont celles de $\tilde{w}_2$, on a :

$$H_2 = H_{w_2}^{(n)}.$$

(b) Si $\sum_{i=3}^{n} |a_{i2}^{(2)}| = 0$, on pose $H_2 = I_n$.

Dans les deux cas, on obtient :

$$A_3 = H_2 A_2 = \begin{pmatrix} * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & 0 & * & \dots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{pmatrix}.$$

Exercice 27. 1) Donner un vecteur $w_1 \in \mathbb{C}^n$ pour l'étape 1 de telle sorte que le coefficient $a_{11}^{(2)}$ de A_2 soit positif.
2) idem pour w_2 à l'étape 2 et pour le coefficient $a_{22}^{(3)}$ de A_3.
3) Décrire ensuite l'étape k, $k \geq 3$ de cette méthode, et en déduire l'existence d'une factorisation QR de la matrice A.
4)

(a) Montrer que toute matrice unitaire et triangulaire (supposée supérieure pour fixer les idées) est forcément de la forme $\mathrm{diag}(d_i)_{1 \leq i \leq n}$ où $|d_i| = 1$ pour tout $i = 1, 2, \dots, n$.

(b) En déduire que si $Q_1 R_1$ et $Q_2 R_2$ désignent deux décompositions QR d'une matrice $A \in GL_n(\mathbb{C})$ $(n \geq 1)$, alors il existe $D = \mathrm{diag}(d_i)_{1 \leq i \leq n}$ avec $|d_i| = 1$ pour tout $i = 1, 2, \dots, n$, telle que

$$\begin{cases} Q_2 &= Q_1 D \\ R_2 &= D^{-1} R_1. \end{cases}$$

(c) Montrer que $D = I$ et en déduire l'unicité de la décomposition QR lorsque A est inversible.

4.3.2 La méthode QR de recherche des valeurs propres.

Description de la méthode QR. A partir de $A \in M_n(\mathbb{C})$, on définit une suite de matrices $(A_k)_{k \geq 1}$ comme suit. Posant $A_1 = A$, le successeur A_{k+1} de A_k, $k \geq 1$, s'obtient en :

(i) Ecrivant la factorisation QR de A_k, $A_k = Q_k R_k$.

(ii) Formant ensuite la matrice $A_{k+1} = R_k Q_k$.

Pour tout $k \geq 1$, on a donc :

$$A_{k+1} = R_k Q_k = Q_k^* R_k Q_k = \dots = (Q_1 Q_2 \dots Q_k)^* A (Q_1 Q_2 \dots Q_k)$$

Convergence de la méthode QR. Nous donnons ici un résultat de convergence qui n'est pas le plus général, mais dont la démonstration a l'avantage de la simplicité.

Théorème 23. On suppose que A est inversible et que toutes ses valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$, sont de modules différents :

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0.$$

La matrice A est donc diagonalisable : il existe P inversible telle que $A = P\Lambda P^{-1}$ avec $\Lambda = \mathrm{diag}(\lambda_i)_{1\leq i\leq n}$. On suppose de plus que P^{-1} possède une factorisation LU : il existe L triangulaire inférieure et U triangulaire supérieure telles que :

$$P^{-1} = LU, \text{ avec } (L)_{ii} = 1,\ i = 1, 2, \ldots, n.$$

Alors la suite de matrices $(A_k)_{k\geq 1}$ est telle que :

$$\begin{cases} (A_k)_{ii} \overset{k\to+\infty}{\longrightarrow} \lambda_i, & 1 \leq i \leq n \\ (A_k)_{ij} \overset{k\to+\infty}{\longrightarrow} 0, & 1 \leq j < i \leq n. \end{cases}$$

Exercice 28. (démonstration du théorème)
Posons pour tout $k \geq 1$,

$$q_k = Q_1 Q_2 \ldots Q_k \text{ et } r_k = R_k \ldots R_2 R_1$$

Nous avons déjà observé que $A_{k+1} = q_k^* A q_k$. L'objectif est alors détudier le comportement de la suite $(q_k)_{k\geq 1}$.

1. Démontrer que pour tout $k \geq 1$, $A^k = q_k r_k$
   Pour étudier la suite (q_k), nous allons étudier la suite $(A^k)_k$ en déterminant une autre décomposition QR de A^k et en la comparant avec celle obtenue en 1.
2. *Deuxième décomposition QR de A^k.*

   (a) Considérons $P = QR$, l'unique factorisation QR de P telle que $(R)_{ii} > 0$, $i = 1, 2, \ldots, n$ et la décomposition $P^{-1} = LU$ de P^{-1}. Montrer que

   $$A^k = QR(\Lambda^k L \Lambda^{-k})\Lambda^k U$$

   où $\Lambda^{-k} := \mathrm{diag}(\lambda_i^{-k})$.

   (b) Montrer que $\Lambda^k L \Lambda^{-k} \overset{k\to+\infty}{\longrightarrow} I$.

   (c) On pose $E_k = \Lambda^k L \Lambda^{-k} - I$. Montrer qu'il existe $k_0 \geq 1$ tel que $I + RE_kR^{-1}$ est inversible pour tout $k \geq k_0$.
   Pour $k \geq k_0$, on note $\tilde{Q}_k \tilde{R}_k$ l'unique décomposition QR de $I + RE_kR^{-1}$ telle que $(\tilde{R}_k)_{ii} > 0$ pour tout $i = 1, 2, \ldots, n$.

   (d) Montrer que $A^k = (Q\tilde{Q}_k)(\tilde{R}_k R \Lambda^k U)$ pour tout $k \geq k_0$.

3. *Comparaison des deux décompositions QR de A^k.*
   En déduire qu'il existe une matrice D_k diagonale vérifiant $|(D_k)_{ii}| = 1$, telle que $q_k = Q\tilde{Q}_k D_k$ et $r_k = D_k^{-1}\tilde{R}_k R \Lambda^k U$.

4. *Comportement asymptotique de* $(\tilde{Q}_k)_k$ *et* $(\tilde{R}_k)_k$.
   (a) On va montrer que la suite $(\tilde{Q}_k)_k$ a I pour unique valeur d'adhérence.
      (i) Justifier l'existence d'une sous-suite $(\tilde{Q}_{\varphi(k)})_k$ de $(\tilde{Q}_k)_k$ qui converge (et dont la limite sera notée $\tilde{Q}$).
      (ii) Montrer que $(\tilde{R}_{\varphi(k)})_k$ converge vers une matrice triangulaire $\tilde{R}$.
      (iii) Montrer alors que $\tilde{Q} = \tilde{R} = I$.
      (iv) Conclure.
   (b) En déduire que $\tilde{R}_k \overset{k\to+\infty}{\longrightarrow} I$ et $\tilde{Q}_k \overset{k\to+\infty}{\longrightarrow} I$.
5. Montrer pour tout $k \geq 1$ que $A_{k+1} = D_k^* \Omega_k D_k$ où $\Omega_k = \tilde{Q}_k^* R \Lambda R^{-1} \tilde{Q}_k$.
6. Conclure.

4.3.3 Mise en œuvre pratique et variantes

La méthode QR décrite précédemment a deux défauts :

1. La décomposition QR d'une matrice d'ordre n est un procédé lent : elle demande n extractions de racines carrées et un nombre d'opérations élémentaires équivalent à $2n^3$.
2. On peut démontrer que pour $i > j$, la suite $((A_k)_{ij})_{k\geq 1}$ se comporte assymptotiquement comme une suite géométrique de raison
$$\frac{|\lambda_i|}{|\lambda_j|}$$
Si deux valeurs $|\lambda_i|$ et $|\lambda_j|$ sont très proches, alors la convergence est très lente.

En pratique, on améliore la méthode de la façon suivante :

1. Pour rendre les décompositions QR plus rapides, on commence par déterminer une matrice de Hessenberg supérieure A' semblable à A (i.e. une matrice A' telle que $A'_{ij} = 0$ pour $i > j+1$), en utilisant des matrices de Householder (même méthode que Householder décrit à la section 4.2 pour une matrice symétrique). Les matrices A_k sucessives sont alors aussi de Hessenberg. Or la décomposition QR d'une matrice hermitienne nécessite $O(n^3/3)$ opérations élémentaires et $O(n)$ extractions de racines carrées. (Dans le cas où A est hermitienne, toutes les matrices manipulées sont alors tridiagonales hermitiennes.)
2. Une fois que A est sous forme Hessenberg, on met alors en œuvre une méthode QR avec translation afin d'améliorer la vitesse de convergence :
Méthode QR translatée On considère une suite de paramètres $(\sigma_k)_k$ de translation, et l'on considère la suite récurrente (A_k) définie par $A_1 = A$ et pour tout $k \geq 1$,
$$A_k - \sigma_k I = Q_k R_k \text{ et } A_{k+1} = \sigma_k I + R_k Q_k$$

(a) **(translation)** On choisit pour les premières itérations un paramètre commun σ aussi proche que possible de λ_n. Alors les coefficients d'indice $(n-1, n)$ convergent comme
$$\frac{|\lambda_n - \sigma|}{|\lambda_{n-1} - \sigma|}$$
c'est-à-dire très vite. (choix possibles : consulter la littérature).

(b) **Déflation** Lorsque l'on considère que la dernière ligne a suffisamment convergé, on la supprime ainsi que la dernière colonne, et l'on travaille alors sur une matrice d'ordre $n-1$ en itèrant le procédé.

4.4 La méthode de la puissance inverse

Utilisée pour le calcul de vecteurs propres, une fois que l'on a obtenu des approximations des valeurs propres correspondantes par une méthode appropriée (Givens-Householder, QR, etc.)

La méthode de la puissance Supposons que $A \in M_n(\mathbb{C})$ soit diagonalisable et qu'elle possède une unique valeur propre λ_1 de module maximal :

$$\forall \lambda \in \sigma(A) - \{\lambda_1\},\ |\lambda| < \rho(A) = |\lambda_1|.$$

Soit $\{p_1, p_2, \ldots, p_n\}$ une base de $\mathbb{C}^n$ de vecteurs propres de A telle que $Ap_i = \lambda_i p_i$, $i = 1, 2, \ldots, \mu$.
On choisit un vecteur initial $u_0 \in \mathbb{C}^n$ et on considère la suite récurrente suivante

$$u_{k+1} = Au_k,\ k \geq 0,$$

Si $u_0 = \sum_{i=1}^n \alpha_i p_i$ est tel que $\alpha_1 \neq 0$, alors on peut calculer, à partir de cette suite, une approximation de λ_1 ainsi qu'un vecteur propre associé. En effet, on a pour tout $k \geq 1$,

$$u_k = \sum_{i=1}^n \alpha_i \lambda_i^k p_i = \lambda_1^k \left(\alpha_1 p_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k p_i \right)$$

D'où

$$\lim_{k \to \infty} \frac{1}{\lambda_1^k} u_k = \alpha_1 p_1$$

Autrement dit la suite $(u_k)_k$ se comporte comme la suite $(\lambda_1^k \alpha_1 p_1)_k$, et si $u_k^{(j)}$ désigne la j-ème composante de u_k dans la base canonique, alors

$$\lambda_1 = \lim_{k \to \infty} \frac{u_{k+1}^{(j)}}{u_k^{(j)}}$$

Exemple. Si l'on cherche par cette méthode la plus grande valeur propre de

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

à partir du vecteur $u_0 = (1, 0)^T$, la suite des vecteurs itérés et les valeurs approchées de la valeur propre calculées à partir de la deuxième composante sont :

k	1	2	3	4	5	6
u_k	2	5	14	41	122	365
	1	4	13	40	121	364
$\lambda^{(k)}$		2,5	2,8	2,93	2,98	2,99

Cette méthode très simple a deux défauts majeurs, qui vont être circonscrits par la méthode de la puissance inverse avec décalage :

1. Les composantes des vecteurs u_k peuvent devenir très grandes ou très petites. Il faut normaliser pour éviter tout problème de dépassement de capacité de la machine.
2. Elle ne permet de traîter que la plus grande des valeurs propres, ou la plus petite, si l'on met en œuvre la *méthode de la puissance inverse* :

$$Au_{k+1} = u_k,\ k \geq 0; \quad u_0 \text{ vecteur arbitraire non nul}$$

Pour avoir accès aux autres sous-espaces propres, il faut faire intervenir un décalage.

La méthode de la puissance inverse avec décalage Soit A une matrice diagonalisable et soit λ une valeur propre de A. On suppose que l'on connait une approximation $\tilde{\lambda} \in \mathbb{C}$ de λ telle que $\tilde{\lambda} \neq \lambda$. La méthode de la puissance inverse avec décalage $\tilde{\lambda}$ et vecteur initial u_0 est définie par :

$$(A - \tilde{\lambda}I)u_{k+1} = u_k,\ \forall k \geq 0$$

Théorème 24. Si pour tout $\mu \in \sigma(A) \setminus \{\lambda\}$,

$$|\tilde{\lambda} - \lambda| < |\tilde{\lambda} - \mu|,$$

et si u_0 n'est pas contenu dans le sous-espace vectoriel engendré par les vecteurs propres correspondant aux valeurs propres $\neq \lambda$, alors pour toute norme vectorielle,

$$\lim_{k\to\infty} \left(\frac{(\lambda - \tilde{\lambda})^k}{|\lambda - \tilde{\lambda}|^k} \frac{u_k}{\|u_k\|} \right) = q$$

où q est un vecteur propre associé à λ.

Exercice 29. Soit $A \in M_n(\mathbb{R})$ symétrique admettant une valeur propre positive λ_1 telle que $\lambda_1 = \rho(A)$. Posons $B = A + \alpha I$. Trouver α donnant la convergence la plus rapide vers la valeur propre maximale dans la méthode de la puissance.

5 Extrema des fonctions réelles

5.1 Extrema et dérivée première

Soient E et G des e.v.n. et U un ouvert de E. Si $F : U \to G$ est une application différentiable en $a \in U$, on notera $F'(a)$ ou $DF(a)$ la différentielle de F au point a. On rappelle que $F'(a)$ est une application linéaire continue de E dans $\mathbb{F}$.
On emploiera indifféremment les termes "différentiable", "dérivable", "différentielle", "dérivée".
Soit U un ouvert d'un espace vectoriel normé E. On dit qu'une fonction $F : U \to \mathbb{R}$ admet au point $a \in U$ un *minimum local* (resp. *maximum local*) s'il existe un ouvert $U' \subset U$ contenant a tel que $\forall x \in U', F(x) \geq F(a)$ (resp. $F(x) \leq F(a)$). S'il n'y a pas lieu de distinguer entre minimum et maximum, on parle d'*extremum local*.

Théorème 25. (Condition nécessaire d'extremum local) Soit $F : U \to \mathbb{R}$ différentiable en a. Une condition nécessaire pour que F admette en a un extremum local est que $F'(a) = 0$.

La relation $F'(a)$ est parfois appelée *équation d'Euler*. Un point $a \in U$ tel que $F'(a) = 0$ s'appelle un *point critique* de f.

Théorème 26. (Condition nécessaire d'extremum local lié)
Soit U un ouvert de $\mathbb{R}^n$ et soit $g = (g_1, \ldots, g_k) : U \to \mathbb{R}^k$ de classe C^1. On note $M = \{x \in U \ / \ g(x) = 0\}$. Soit $F : U \to \mathbb{R}$ et soit a un point de M tel que :

1. F est dérivable en a
2. Les dérivées $g_1'(a), \ldots, g_k'(a)$ sont linéairement indépendantes

Si la restriction $F_{|M}$ admet un extremum local au point $a \in M$, alors il existe k réels $\lambda_1, \ldots, \lambda_k$ tels que :

$$F'(a) = \lambda_1 g_1'(a) + \ldots \lambda_k g_k'(a)$$

Les réels λ_i s'appellent des *multiplicateurs de Lagrange*.

Exercice 30. La densité d'une surface métallique Σ définie par l'équation $x^2+y^2+z^2 = 4$ est donnée par $\rho(x, y, z) = 2 + xz + y^2$. Déterminer les points de Σ où la densité est la plus faible et ceux où elle est la plus forte.

Exercice 31. Soit A une matrice symétrique d'ordre n, B une matrice symétrique définie positive d'ordre n et b un vecteur de $\mathbb{R}^n$. Énoncer une condition d'extremum local lié de la fonction

$$J : v \in \mathbb{R}^n \mapsto J(v) = \frac{1}{2}(Av, v) - (b, v)$$

par rapport à l'ensemble

$$M = \{v \in \mathbb{R}^n \ / \ (Bv, v) = 1\}$$

Exercice 32. Soit $n \in \mathbb{N}$. On "oublie" que $\|(x_1, \ldots, x_n\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$ définit une norme sur $\mathbb{R}^n$ (ou $\mathbb{C}^n$) et l'on souhaite démontrer l'inégalité triangulaire (dite *inégalité de Minkowski*) : pour $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$,

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \qquad (1)$$

1) Montrer que (1) est vraie lorsque x et y sont colinéaires.
2) On fixe $\alpha, \beta > 0$ et on envisage le problème de maximisation sous contrainte :

$$S = \sup\{\|x+y\|_p^p; \|x\|_p^p = \alpha^p \text{ et } \|y\|_p^p = \beta^p\}$$

a) Montrer que S est atteint en au moins un point (X, Y).

b) Calculer les dérivés partielles de $x \mapsto \|x\|_p^p$ en X, et justifier que les dérivées $D\|x\|_p^p(X,Y)$ et $D\|y\|_p^p(X,Y)$ sont linéairement indépendantes.

c) Appliquer alors le théorème des extrema liés pour établir que X et Y sont colinéaires et conclure.

5.2 Extrema et dérivée seconde

Si $F : U \to \mathbb{F}$ est une application deux fois dérivable en $a \in U$, on notera $F"(a)$ ou $D^2F(a)$ la différentielle seconde de F au point a. On rappelle que $F"(a)$ est une application bilinéaire continue de $E \times E$ dans $\mathbb{F}$.

Théorème 27. (Condition nécessaire d'extremum) Soit E un e.v.n., U un ouvert de E et $F : U \to \mathbb{R}$ une application dérivable sur U deux fois dérivable en $a \in U$. Si F a un minimum (resp. maximum) local en x, alors pour tout $v \in E$,

$$F"(a)(v,v) \geq 0 \ \big(\text{ resp. } \ F"(a)(v,v) \leq 0\big)$$

Théorème 28. (Conditions suffisantes d'extremum) Soit E un e.v.n., U un ouvert de E et $F : U \to \mathbb{R}$ une application dérivable sur U telle que $F'(a) = 0$.

1. Si F est deux fois dérivable en a et s'il existe $\alpha > 0$ tel que pour tout $w \in E$,

$$F"(a)(w,w) \geq \alpha\|w\|^2,$$

alors F admet en a un minimum local strict.

2. Si F est deux fois dérivable sur U et s'il existe une boule $B \subset U$ telle que

$$\forall x \in B, \forall w \in E, \ F"(x)(w,w) \geq 0,$$

alors F admet en a un minimum local.

5.3 Extrema et convexité

Une partie U d'un e.v.n. est dite convexe si pour tous points $x, y \in U$, le segment $[x,y] = \{tx + (1-t)y \ ; \ 0 \leq t \leq 1\}$ est inclus dans U.

Théorème 29. (Condition nécessaire de minimum local sur un ensemble convexe) Soit $F : U \to \mathbb{R}$ une fonction définie sur un ouvert U d'un e.v.n. E et soit Ω une partie convexe de U. Si F est dérivable en $a \in \Omega$ et si la restriction $F_{|\Omega}$ admet en a un minimum local, alors

$$\forall x \in \Omega, \ F'(a)(x-a) \geq 0$$

Une fonction $F : \Omega \to \mathbb{R}$ définie sur une partie convexe Ω d'un e.v.n. E est dite *convexe* (resp. *strictement convexe*) sur Ω si pour tous $x, y \in \Omega$ et pour tout $t \in [0, 1]$,

$$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y)$$

$$\big(\text{ resp. } F(tx + (1-t)y) < tF(x) + (1-t)F(y)\big)$$

Proposition 30. (Caractérisation des fonctions convexes) Soit $F : U \to \mathbb{R}$ une fonction définie sur un ouvert U d'un e.v.n. E et soit Ω une partie convexe de U.

1. On suppose F dérivable sur U. Alors F est convexe sur Ω si et seulement si pour tous $x, y \in \Omega$,
$$F(y) \geq F(x) + F'(x)(y - x)$$
2. On suppose F deux fois dérivable sur U. Alors F est convexe sur Ω si et seulement si pour tous $x, y \in \Omega$,
$$F"(x)(y - x, y - x) \geq 0$$

Théorème 31. (Minimum des fonctions convexes) Soit Ω une partie convexe d'une e.v.n. et soit $F : \Omega \to \mathbb{R}$ une fonction convexe sur Ω.

1. Si F a un minimum local en $a \in \Omega$, alors c'est un minimum global sur Ω.
2. Supposons F définie sur un ouvert U contenant Ω et dérivable en $a \in \Omega$. Alors F admet un minimum local en a si et seulement si
$$\forall x \in \Omega,\ F'(a)(x - a) \geq 0$$
3. sous les mêmes hypothèses qu'en (2.), si de plus Ω est un ouvert de E, alors F admet un minimum local en a si et seulement si $F'(a) = 0$.

Exercice 33. Soit $F : U \to \mathbb{R}$ une fonction dérivable sur un ouvert U d'un e.v.n. E et soit Ω une partie convexe de U. Montrer que F est strictement convexe sur Ω si et seulement si pour tous $x, y \in \Omega$,

$$F(y) > F(x) + F'(x)(y - x)$$

Exercice 34. Soit A une matrice symétrique d'ordre n et soit b un vecteur de $\mathbb{R}^n$.
1) Démontrer que la fonction

$$J : v \in \mathbb{R}^n \mapsto J(v) = \frac{1}{2}(Av, v) - (b, v)$$

est convexe si et seulement si la matrice symétrique A est positive, et strictement convexe si et seulement si A est définie positive.
2) Démontrer qu'il existe $y \in \mathbb{R}^n$ tel que $\forall x \in \mathbb{R}^n, J(y) \leq J(x)$ si et seulement si A est positive et $\{z \in \mathbb{R}^n\ ;\ Az = b\} \neq \emptyset$

Exercice 35. Soit B une matrice réelle $m \times n$ et soit $c \in \mathbb{R}^m$. Considérons la fonction $J : \mathbb{R}^n \to \mathbb{R}$ définie par

$$J(v) = \frac{1}{2}\|Bv - c\|_2^2 - \frac{1}{2}\|c\|_2^2$$

1) Démontrer que la fonction J est convexe.
2) Calculer la dérivée de J.
3) Montrer que l'ensemble des minima locaux de J coïncide avec l'ensemble des solutions de l'équation

$$B^T Bu = B^T c$$

Exercice 36. 1) Soit $E = (e_{ij})$ la matrice carrée d'ordre n définie par $e_{ij} = 1,\ 1 \leq i, j \leq n$. Décrire les valeurs propres et les sous-espaces propres de E.
2) Considérons l'ensemble

$$\Omega = \{x \in \mathbb{R}^n;\ x_i > 0, 1 \leq i \leq n\}$$

et la fonction $J : \Omega \to \mathbb{R}$ définie par

$$\forall x \in \Omega,\ J(x) = -\left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

Calculer $J'(x)$ et $J"(x)$.
3) Montrer que J est convexe mais non strictement convexe.
4) Soit $U = \{x \in \Omega;\ \sum_{i=1}^n x_i = n\}$. Démontrer que la restriction $K = J_{|U}$ est strictement convexe.
5) On note e le vecteur de Ω dont toutes les composantes valent 1. Montrer que pour tout $x \in U,\ J'(e)(x - e) = 0$. En déduire qu'il existe un unique $u \in U$ tel que

$$J(y) = \inf_{x \in U} J(x)$$

6) Pour tout $x \in \Omega$, démontrer l'inégalité

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} x_i$$

et décrire le sous-ensemble Ω pour lequel cette inégalité devient une égalité.

5.4 Le théorème de projection de Hilbert

On appelle *espace préhilbertien* un espace vectoriel V muni d'un produit scalaire $(.,.)$. En particulier, V est un e.v.n. pour la norme associée $\|v\| = \sqrt{(v,v)},\ v \in V$. Si V est complet pour cette norme, on dit que V est un *espace de Hilbert*.

Exemples

1) Tout e.v.n. de dimension fini muni d'un produit scalaire est un espace de Hilbert

2) $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. L'espace $l^2(K)$ des suites $x = (x_n)_{n\in\mathbb{N}}$ de $\mathbb{K}$ de carrés sommables (i.e. $\sum_{n=0}^{\infty} |x_n|^2 < \infty$, muni du produit scalaire

$$(x,y) = \sum_{n=0}^{\infty} x_n \bar{y_n}$$

3) L'espace $\mathcal{C}([a,b],\mathbb{K})$ des fonctions $f : [a,b] \to \mathbb{K}$ continues muni du produit scalaire

$$(f,g) = \int_a^b f(t)\overline{g(t)}dt$$

Théorème 32. (Théorème de projection). Soit C un sous-ensemble non vide, convexe, fermé, d'un espace de Hilbert V.

1. Étant donné un élément quelconque $w \in V$, il existe un et un seul élément $P(w)$ tel que
$$P(w) \in C \ \text{ et } \|w - P(w)\| = \inf_{v\in C} \|w - v\|$$
2. Cet élément $P(w) \in C$ est l'unique élément de C vérifiant
$$(P(w) - w, v - P(w)) \geq 0 \ \text{ pour tout } v \in C$$
3. L'application $P : V \to C$ ainsi définie est telle que
$$\|P(w_1) - P(w_2)\| \leq \|w_1 - w_2\|$$
4. L'application P est linéaire si et seulement si le sous-ensemble C est un sous-espace vectoriel de V, auquel cas les inégalités de 2. sont remplacées par les égalités :
$$(P(w) - w, v) = 0 \ \text{ pour tout } v \in C$$

6 Approximation au sens des moindres carrés

Position du problème Soient $x_1, \ldots, x_m$ des réels distincts. À chaque point x_i, on associe un réel c_i, qui peut être par exemple une valeur expérimentale. On cherche à construire une fonction U *d'un type donné* qui approche "au mieux" (dans un sens précisé ultérieurement) les valeurs c_i aux points x_i.
On se donne n fonctions réelles $w_1, \ldots, w_n$ (par exemple polynomiales) linéairement indépendantes définies dans un intervalle de $\mathbb{R}$ contenant les points x_i, et l'on cherche une fonction U de la forme

$$U = \sum_{j=1}^{n} u_j w_j$$

de telle façon que pour tout $i = 1, \ldots, m$, le réel $U(x_i)$ approche au "mieux" c_i.
La façon la plus commune de réaliser cette approximation est d'approcher les égalités $U(x_i) = c_i$ *au sens des moindres carrés* : on cherche une fonction U telle que le nombre

$$\sum_{i=1}^{m} |U(x_i) - c_i|^2$$

soit minimum.

Partie 1

Soit $B = (b_{ij})$ la matrice $m \times n$ définie par $b_{ij} = w_j(x_i)$, et soit $c = (c_1, \ldots, c_m)^T \in \mathbb{R}^m$.

1) Démontrer que l'approximation au sens des moindres carrées équivaut à trouver $u \in \mathbb{R}^n$ tel que

$$\|Bu - c\|_2 = \inf_{v \in \mathbb{R}^n} \|Bv - c\|_2 \qquad (*)$$

2) Considérons le sous-espace vectoriel $Im(B) = \{Bv; v \in \mathbb{R}^n\}$ de $\mathbb{R}^n$. Démontrer qu'il existe un unique $\tilde{u} \in Im(B)$ tel que

$$\|\tilde{u} - c\|_2 = \inf_{\tilde{v} \in Im(B)} \|\tilde{v} - c\|_2$$

3) En déduire l'existence d'une solution au problème $(*)$. À quelle condition sur $B^T B$ cette solution est-elle unique ?
4) Montrer que u est solution de $(*)$ si et seulement si u est solution des *équations normales*

$$B^T Bu = B^T c \qquad (**)$$

5) On pose $S = \{u \in \mathbb{R}^n, B^T Bu = B^T c\}$. Démontrer que S contient un unique élément s de norme minimale.
6) Démontrer que s est caractérisé par

$$s = S \cap \big(ker(B^T B)\big)^{\perp}$$

Partie 2

Matrice pseudo-inverse

Le but de cette question est d'étudier l'application $c \in \mathbb{R}^m \mapsto s \in \mathbb{R}^n$
1) Soit r le rang de B. Démontrer qu'il existe une matrice orthogonale U d'ordre m et une matrice orthogonale V d'ordre n telles que

$$U^T B V = \begin{pmatrix} \mu_1 & & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 \\ 0 & & \mu_r & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix},$$

2) En remarquant que $s \in S$ est équivalent à $V^T B^T B V (V^T s) = V^T B^T c$, démontrer que pour tout $i = 1, \dots, r$, la i-ème composante du vecteur $V^T s$ s'exprime par :

$$(V^T s)_i = \frac{1}{\mu_i^2} (V^T B^T c)_i, \quad i = 1, \dots, r$$

3) Démontrer que $w \in ker(B^T B)$ si et seulement si pour tout $i = 1, \dots, r$, $(V^T w)_i = 0$. En déduire que

$$(V^T s)_i = 0, \quad i = r+1, \dots, n.$$

4) En déduire que l'application $c \in \mathbb{R}^m \mapsto s \in \mathbb{R}^n$ est linéaire et que $s = B^+ c$ où B^+ est la matrice

$$B^+ = V \begin{pmatrix} \frac{1}{\mu_1} & & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 \\ 0 & & \frac{1}{\mu_r} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} U^T$$

5) Démontrer les relations suivantes :

$$(B^+)^+ = B \ ; \ (B^T)^+ = (B^+)^T \ ; \ B = BB^+B \ ; \ B^+ = B^+BB^+ \ ; \ (B^+B)^T = B^+B$$

La matrice B^+ s'appelle la matrice pseudo-inverse de B.