Lecture V : Numerical Analysis

A) Aims of this class

After this class,

- I can characterize a positive definite symmetric matrix.
- I can determine a region of the plane containing all the eigenvalues of a matrix.
- I can code a method that gives me the spectral radius of a matrix.
- I know the difference between a direct method and an iterative method that solve linear systems, and how to use them.
- I know how to use the notion of condition number of a matrix.
- I know how to evaluate the complexity of a numerical method.

B) To become familiar with this class' concepts (to prepare before the examples class)

Questions V.1 and V.2 must be done before the 8th lab. The solutions are available online.

Question V.1 (Applications of the Schur theorem)
Let $A \in \mathcal{M}_q(\mathbb{C})$, where $q \geq 1$.

Q. V.1.1 Recall the decomposition theorem.

Q. V.1.2 Show that A is normal, that is, $AA^* = A^*A$, if and only if there exists a unitary matrix U and a diagonal matrix D containing the eigenvalues of A such that $A = UDU^*$.
HINT: Compute the diagonal elements of TT^* and T^*T where T is the upper triangular matrix such that $A = UTU^*$ and U is unitary.

Q. V.1.3 Show that if A is Hermitian, it is diagonalizable in an orthogonal basis of eigenvectors and its eigenvalues are real.

Q. V.1.4 Show that, if A is unitary, it is diagonalizable in an orthogonal basis of eigenvectors and that its eigenvalues are of modulus 1.

Question V.2 (Solving triangular systems)
Let $b \in \mathbb{K}^q$, where $q \geq 1$.

Q. V.2.1 Let $L \in T_{q,inf}(\mathbb{K})$. Write the resolution algorithm of the linear system $Lx = b$.

Q. V.2.2 Let $U \in T_{q,sup(\mathbb{K})}$. Write the resolution algorithm of the linear system $Ux = b$.

Q. V.2.3 Compute the number of operations needed to perform these resolutions.

C) Exercises

Exercise V.1 (An intuitive approach to the resolution of linear systems)
Let $A \in GL_q(\mathbb{R})$. Let $b \in \mathbb{R}^q$. What is the complexity of the resolution of the linear system $A^2x = b$ with a direct method? Propose a less costly method.

Exercise V.2 (LU decomposition of the matrix appearing in the Theorem ??)

Let $A \in \mathcal{M}_q(\mathbb{R})$ be a tridiagonal matrix, defined as:

$$A = \begin{pmatrix} 2+c_1 & -1 & & & & & \\ -1 & 2+c_2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2+c_{q-1} & -1 & \\ & & & & -1 & 2+c_q & \end{pmatrix},$$

with $c_i \geq 0, \quad \forall 1 \leq i \leq q$.

E. V.2.1 Let $v \in \mathbb{R}^q$. Show that:

$$v^T A v = \sum_{i=1}^{q} c_i v_i^2 + \left\{ v_1^2 + v_q^2 + \sum_{i=2}^{q} (v_i - v_{i-1})^2 \right\}.$$

E. V.2.2 Deduce that A admits a Cholesky decomposition.

E. V.2.3 In the case $c_i = 0$, $(1 \leq i \leq q)$, show that the matrix B such that $A = BB^T$ is bidiagonal and given by:

$$\begin{cases} B_{i,i} &= \sqrt{\dfrac{i+1}{i}} \quad (1 \leq i \leq q); \\ \\ B_{i+1,i} &= -\sqrt{\dfrac{i}{i+1}} \quad (1 \leq i \leq q-1). \end{cases}$$

Exercise V.3 (QR decomposition)
Let $A \in GL_q(\mathbb{R})$.

E. V.3.1 Show that there exists an upper triangular matrix R with positive diagonal such that:

$$A^T A = R^T R.$$

E. V.3.2 Deduce that there exists an orthogonal matrix $Q \in O_q(\mathbb{R})$ such that: $A = QR$.

E. V.3.3 Show that this decomposition $A = QR$, R being upper triangular with positive diagonal and Q being orthogonal, is unique.

Exercise V.4
Let $\omega \in \mathbb{R} \setminus \{-1\}$. Let $A \in \mathcal{M}_q(\mathbb{R})$ be a non-singular matrix such that

$$A = (1+\omega)P - (N + \omega P),$$

with P invertible and $P^{-1} N$ having real eigenvalues $1 > \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q$.

E. V.4.1 Find the values of ω such that the iterative method

$$\forall n \geq 0, \quad (1+\omega)Px^{(n+1)} = (N+\omega P)x^{(n)} + b$$

converges for all initial vector x_0 to the solution of the system $Ax = b$.

E. V.4.2 Find the value of ω that guarantees a minimal convergence rate, so that the convergence speed is optimized.

D) Going further

These exercises can be found on edunao as Jupyter notebooks.

Chapter VIII: Solutions

Solution de Q. V.1.1 Let $A \in \mathcal{M}_q$. Then, there exist

- an upper triangular matrix T
- a unitary matrix U

such that $A = UTU^{-1}$.

Solution de Q. V.1.2 Apply the Schur decomposition: let $U \in U_q(\mathbb{C})$ and $T \in T_{q,sup}(\mathbb{C})$ such that $A = UTU^*$. So $AA^* = UTU^*UT^*U^* = UTT^*U^*$ and $A^*A = UT^*TU^*$.

- Let us show the converse: Assume that $T = D$ with D diagonal. One has $D^* = \overline{D}$. Then $AA^* = UD\overline{D}U^*$ and $A^*A = U\overline{D}DU^*$. We conclude that A is normal.
- Let us show the direct sense: assume that A is normal. Let us examine the diagonal elements $TT^* = T^*T$. One has, for all $i \in \{1, \dots, q\}$,

$$(TT^*)_{ii} = \sum_{k=1}^{q} (T)_{ik}(T^*)_{ki} = \sum_{k=1}^{q} |T_{ik}|^2 = \sum_{k=i}^{q} |T_{ik}|^2$$
$$(T^*T)_{ii} = \sum_{k=1}^{q} (T^*)_{ik}(T)_{ki} = \sum_{k=1}^{q} |T_{ki}|^2 = \sum_{k=1}^{i} |T_{ki}|^2$$

Let us use an induction to show that the extra-diagonal elements of T vanish until the line i:

- for $i = 1$, one has:
$$\sum_{k=1}^{q} |T_{1k}|^2 = |T_{11}|^2.$$
So $\sum_{k=2}^{q} |T_{1k}|^2 = 0$ and $\forall k \geq 2$, $T_{1k} = 0$.
- Inductive hypothesis: assume that, for $i \geq 1$, for all $p \leq i$, $k \neq p$, $T_{pk} = 0$.
Let us show that this property is true for $i+1$, that is, for all $k \leq i$, $T_{i+1,k} = 0$:
Knowing that for all $k \neq i+1$, $T_{k,i+1} = 0$,
$$\sum_{k=i+1}^{q} |T_{i+1,k}|^2 = \sum_{k=1}^{i+1} |T_{k,i+1}|^2 = |T_{i+1,i+1}|^2$$
so $\sum_{k=i+2}^{q} |T_{i+1,k}|^2 = 0$. One concludes that, for all $k \neq i+1$, $T_{i+1,k} = 0$.
- Conclusion : We showed that T is a diagonal matrix. Let us denote $T = D$.

Let us show that D contains the eigenvalues of A. For all $i \in \{1, \dots, q\}$, let v_i the i-th column of the matrix U, that is, $v_i = Ue_i$, with e_i the i-th canonical vector of $\mathbb{C}^q$. Recall that U unitary implies $v_i^* v_j = \delta_{ij}$. Then

$$Av_i = UDU^* v_i = UD \begin{pmatrix} v_1^* v_i \\ \vdots \\ v_i^* v_i \\ \vdots \\ v_q^* v_i \end{pmatrix} = UDe_i = D_{ii} Ue_i = D_{ii} v_i.$$

So the vectors v_i are eigenvectors of A associated with D_{ii} an eigenvalue and the vectors v_i form an orthonormal basis of $\mathbb{C}^q$, since they are columns of U. We found q eigenvectors forming a basis of $\mathbb{C}^q$ and q associted eigenvalues: we conclude that $\mathrm{Sp}(A) = \{D_{ii},\, i \in \{1, \dots, q\}\}$.

Solution de Q. V.1.3 If A is Hermitian, then $A = A^*$ so A is normal. Thanks to Question V.1.2, there exists $D \in diag(\mathbb{C})$ and $U \in U_q(\mathbb{C})$ such that $A = UDU^*$. Moreover, as $A = A^*$, $D = U^* AU = U^* UD^* U^* U = D^*$. We conclude that the diagonal coefficients of D are conjugated with themselves, so they are real.

Solution de Q. V.1.4 If A is unitary, then $AA^* = I_q = A^* A$, then A is normal. According to Question V.1.2, A is diagonalizable in an orthogonal basis of eigenvectors because there exists $U \in U_q(\mathbb{C})$ and D diagonal such that $A = UDU^*$. as $AA^* = I_q$, one gets $DD^* = I_q$ and for all $i \in \{1, \dots, q\}$, $|D_{ii}|^2 = 1$, so $|D_{ii}| = 1$: the spectrum of A lies in in the unit circle.

Solution de Q. V.2.1 Let us fix the convention that a vector with an empty set of indices is zero: $x(1:0) = 0$.

Assume that L is invertible. Let x the solution to the system $Lx = b$. Then

$$\begin{aligned} L_{11} x_1 &= b_1 \\ &\vdots \\ \sum_{k=1}^{i} L_{ik} x_k &= b_i \\ &\vdots \\ \sum_{k=1}^{q} L_{qk} x_k &= b_q \end{aligned}$$

x can be computed simply by induction, since

$$\forall i \in \{1, \dots, q\}, \quad L_{ii} x_i = b_i - \sum_{k=1}^{i-1} L_{ik} x_k.$$

The algorithm writes

Algorithme 1 : Solving a lower triangular system

Donnée : L, b

Paramètre : $q = size(A, 1)$

1 **Pour** i=1 **to** q **faire**
2 **si** $L(i,i) \neq 0$ **alors**
3 $x(i) \leftarrow \dfrac{b(i) - L(i, 1:i-1) * x(1:i-1)}{L(i,i)}$
4 **sinon**
5 break: the matrix L is not invertible!
6 **finsi**
7 **finpour**

Résultat : x

Solution de Q. V.2.2 The only difference with the previous question lies in the fact that the algorithm must begin with the last component of x !

Algorithme 2 : Solving an upper triangular system

Donnée : U, b

Paramètre : $q = size(A, 1)$

1 **Pour** i=q **to** 1 **faire**
2 **si** $U(i,i) \neq 0$ **alors**
3 $x(i) \leftarrow \dfrac{b(i) - U(i, i+1:q) * x(i+1:q)}{U(i,i)}$
4 **sinon**
5 break: the matrix U is not invertible!
6 **finsi**
7 **finpour**

Résultat : x

Solution de Q. V.2.3 We only give here the computation for the lower triangular matrix, since the computation for an upper one is the same.
La ligne 3 nécessite

- $i-1$ products and $i-1$ additions (in fact, $i-2$ additions and one subtraction)
- one division.

Summing up these operations i, one gets: $q(q-1)/2$ additions, $q(q+1)/2$ products (in fact, $q(q-1)/2$ products and q divisions). So the complexity is of order $O(q^2)$ (quadratic).

Solution de Q. V.1 Let $A = PLU$ where P is a permulation matrix and L and U the Lower- and Upper-triangular matrices.

Multiplying by P is linear ($O(q)$) if we carry out this operation before solving the system. Solving $PLUx = b$ is done by solving $Pz = b$ then $Ly = z$ and $Ux = y$. Solving for y and x has a quadratic complexity $O(q^2)$. In we showed that sovling two triangular systems has a complexity in $O(q^2)$. Let us note $N_{triang} = q^2$. See also Exercise V.2.

Computing A^2 requires q^2 times $q-1$ sums and q multiplications. Therefore we have $2q^3$ operations. Let us denote by $N_{mult} = 2q^3$.

The complexity to decompose A in PLU is $N_{PLU} = q^3$.

Eventually, we need to carry out $N_{mult} + N_{LU} + N_{triang}$ operations. This yields a complexity in $O(q^3)$.

Doing the decomposition $A = P_2L_2U_2$ first then solving $P_2L_2U_2L_2U_2x = b$ leads to $N_{LU} + 2N_{triang}$ operations. The complexity is the same but the computing time is divided by 2. You can experiment using Python or Matlab.

Solution de Q. V.2.1 Let us compute $v^T Av$. In what follows, we will note $v_0 = 0$ and $v_{q+1} = 0$.

$$\begin{aligned}
v^T Av &= \sum_{i=1}^{q} v_i (Av)_i = \sum_{i=1}^{q} v_i \left(\sum_{j=1}^{q} A_{ij} v_j \right) \\
&= v_1(A_{1,1}v_1 + A_{1,2}v_2) + \sum_{i=2}^{q-1} v_i(A_{i,i-1}v_{i-1} + A_{i,i}v_i + A_{i,i+1}v_{i+1}) + v_q(A_{q,q-1}v_{q-1} + A_{q,q}v_q) \\
&= \sum_{i=1}^{q} v_i(-v_{i-1} + (2+c_i)v_i - v_{i+1}) = \sum_{i=1}^{q} c_i v_i^2 + \sum_{i=1}^{q} v_i(v_i - v_{i-1}) + \sum_{i=1}^{q} v_i(v_i - v_{i+1}) \\
&= \sum_{i=1}^{q} c_i v_i^2 + \sum_{i=1}^{q} v_i(v_i - v_{i-1}) + \sum_{i=2}^{q+1} v_{i-1}(v_{i-1} - v_i) \\
&= \sum_{i=1}^{q} c_i v_i^2 + \sum_{i=2}^{q} (v_i - v_{i-1})^2 + v_1^2 + v_q^2.
\end{aligned}$$

Solution de Q. V.2.2 We have

- The matrix A is symmetric (obvious).

- The matrix A is positive definite. Indeed, for all $v \in \mathbb{R}^q$, the previous question yields
  - $v^T Av \geq 0$
  - $v^T\, Av = 0$ implies $v_1 = v_q = 0$ and $\forall i \in \{2, \ldots, q\}$, $v_i = v_{i-1} = v_1$ therefore $v = 0$.

The Cholesky Theorem **??** yields there exists a lower triangular matrix B with with positive diagonal terms such that $A = B\, B^*$.

Solution de Q. V.2.3

$$(BB^T)_{i,j} = \sum_{k=1}^{q} B_{i,k}B_{j,k} = B_{i,j-1}B_{j,j-1} + B_{i,j}B_{j,j}.$$

- if $j - i \geq 2$, $(BB^T)_{i,j} = 0$,
- if $i \leq q-1$ and $j = i+1$, $(BB^T)_{i,i+1} = B_{i,i}B_{i+1,i} = -\sqrt{\dfrac{i+1}{i}}\sqrt{\dfrac{i}{i+1}} = -1$,
- si $i = j$, $(BB^T)_{i,i} = B^2_{i,i-1} + B^2_{i,i} = \dfrac{i-1}{i} + \dfrac{i+1}{i} = 2$.

B is the one which is provided thanks to the uniqueness of the Cholesky decomposition.

Solution de Q. V.3.1
What follows is important!

Let us prove the matrix $A^T A$ is symmetric positive definite (SPD).

- Symmetry is obvious.
- Let $v \in \mathbb{R}^q$. We have $v^T A^T Av = (Av)^T Av = \|Av\|_2^2$. Thus $v^T A^T Av > 0$ if $Av \neq 0$. Since A is not singular, $Av = 0$ implies $v = 0$.

Since $A^T A$ is a SPD matrix, there exists $R \in T_{q,sup}$ with positive diagonal terms such that $A^T A = R^T R$ (Cholewsky Theorem **??**).

Solution de Q. V.3.2 Let $Q = AR^{-1}$

$$Q^T Q = (R^{-1})^T A^T AR^{-1} = (R^{-1})^T R^T RR^{-1} = I_q.$$

Therefore O is an orthogonal matrix.

Solution de Q. V.3.3 Suppose there exist two decompositions QR of the matrix A: (Q_1, R_1) and (Q_2, R_2).

$$A = Q_1 R_1 = Q_2 R_2$$

therefore $Q_1^T Q_2 = R_1 R_2^{-1}$.

- The matrix $Q_1^T Q_2$ on the left hand side is the product of two orthogonal matrices, therefore it is orthogonal.
- The inverse of an upper triangular matrices with positive coefficients on the diagonal is an upper triangular matrices with positive coefficients on the diagonal. The matrix $R_1 R_2^{-1}$ on the right hand side is the product of two upper triangular matrices with positive coefficients on the diagonal, therefore it is an upper triangular matrix with positive coefficients on the diagonal.

The only orthogonal matrix, which is triangular is the identity matrix. Therefore $Q_1 = Q_2$ and $R_1 = R_2$. This proves the unicity of the decomposition.

Solution de Q. V.4.1 The integration matrix of the method is:

$$\mathcal{M}(\omega) = \frac{1}{1+\omega} P^{-1}(N + \omega P) = \frac{1}{1+\omega}(P^{-1}N + \omega I_q).$$

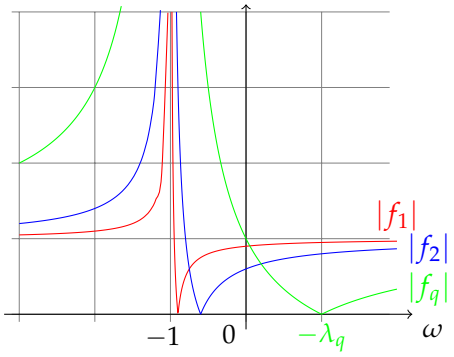
According to the Theorem , the method converges iff the associated iteration matrix M satisfies: $\rho(\mathcal{M}(\omega)) < 1$.
Let us characterize the eigenvalues of M:

$$\begin{aligned} \mu \in \mathrm{Sp}(\mathcal{M}(\omega)) &\iff \mu \in \mathrm{Sp}\left(\frac{1}{1+\omega}(P^{-1}N + \omega I_q)\right) \\ &\iff \mu \in \left\{\frac{\lambda_i + \omega}{1+\omega},\ i \in \{1,\dots,q\}\right\} \end{aligned}$$

For a given $\omega \in \mathbb{R} \setminus \{-1\}$, the method converges iff

$$\max_{i\in\{1,\dots,q\}} \left|\frac{\lambda_i + \omega}{1+\omega}\right| < 1.$$

For $i \in \{1,\dots,q\}$, let us note $f_i : \omega \mapsto (\lambda_i + \omega)/(1+\omega)$. Then $i \in \{1,\dots,q\}$, $|f_i|_{]-\infty,-1[}| > 1$.



Moreover

$$\{\omega \in]-1,+\infty[/ |f_i(\omega)| < 1\} = \left]-\frac{1+\lambda_i}{2}, +\infty\right[.$$

Therefore

$$\{\omega \in]-1,+\infty[: \rho(\mathcal{M}(\omega)) < 1\} = \bigcap_{1\leq i\leq q} \left]-\frac{1+\lambda_i}{2},+\infty\right[= \left]-\frac{1+\lambda_q}{2},+\infty\right[.$$

Solution de Q. V.4.2 Let us remark that

$$\rho(\mathcal{M}) : \omega \mapsto \begin{cases} |f_q(\omega)| & \text{si } \omega \in]-\infty,\omega_0[\setminus\{-1\} \\ |f_1(\omega)| & \text{si } \omega \in]\omega_0,+\infty[\end{cases}$$

where ω_0 it the value for which f_1 and f_q are equal, satisfaying

$$|\lambda_1+\omega_0| = \lambda_1+\omega_0 = |\lambda_q+\omega_0| = -\lambda_q-\omega_0$$

which is

$$\omega_0 = -\frac{\lambda_1+\lambda_q}{2}.$$

The minimal converence rate is

$$\rho(\mathcal{M}(\omega_0)) = \frac{\lambda_1-\lambda_q}{2-(\lambda_1+\lambda_q)}.$$