Analyse Numérique

Notes de Cours

27 février 2026

Table des matières

Chapitre 1

Intégration numérique et points de Gauss

1.1 Préliminaires

On souhaite évaluer numériquement une intégrale $I(f) = \int_a^b f(x)\,dx$ où f est une fonction continue sur un intervalle $[a, b]$ avec $a < b$.

On cherche une approximation de $I(f)$, notée $J(f)$ sous la forme :

$$J(f) = \sum_{i=0}^{N} \lambda_i f(x_i), \qquad \text{où les points } x_i \in [a, b].$$

Les x_i sont appelés : les points d'intégration, et les coefficients λ_i : les poids d'intégration.

1.2 Intégration des polynômes

Théorème 1.1 :

Supposons que les points d'intégration x_i sont imposés avec :

$$a \leq x_0 < \cdots < x_n \leq b.$$

Dans ce cas, il existe un jeu unique de λ_i tel que $I(f)$ soit exactement égale à $J(f)$ pour tout polynôme de degré n.

Démonstration. Par linéarité de I et J, il suffit que $I(f) = J(f)$ pour tout $f : x \mapsto x^i$ avec $0 \leq i \leq n$.

Les poids d'intégration sont alors solution du système linéaire ci-dessous, qui admet une unique solution, car
$\det(A) = \prod_{i=1, j\neq i}^{n} (x_i - x_j) \neq 0.$

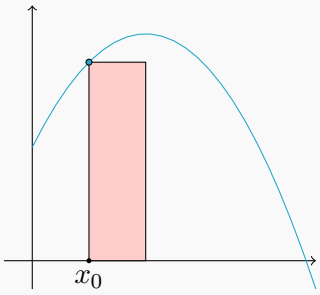
$$\begin{pmatrix} 1 & \dots & 1 \\ x_0 & \dots & x_n \\ \vdots & \vdots & \vdots \\ x_0^n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \vdots \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} b-a \\ \dfrac{b^2-a^2}{2} \\ \vdots \\ \dfrac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix} \quad \Leftrightarrow \quad A\lambda = S.$$

Remarque : A est la matrice de Vandermonde. □

Pour $n = 0$, $x_0 = a$.

On trouve $\lambda_0 = b - a$.

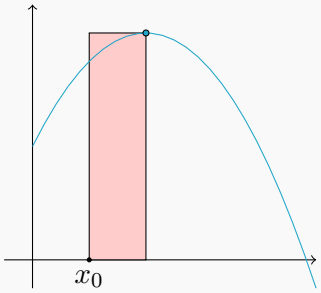
Soit $J(f) = (b-a) \cdot f(a)$,
c'est la méthode des rectangle à gauche.



Pour $n = 0$, $x_0 = b$.

On trouve $\lambda_0 = b - a$.

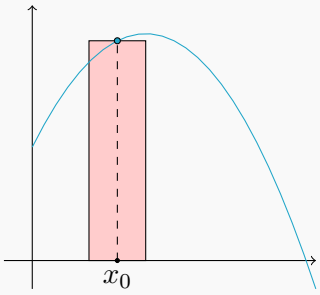
Soit $J(f) = (b-a) \cdot f(b)$,
c'est la méthode des rectangles à droite.



Pour $n = 0$, $x_0 = \dfrac{a+b}{2}$.

On trouve $\lambda_0 = b - a$.

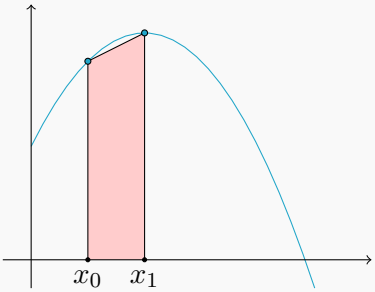
Soit $J(f) = (b-a) \cdot f\left(\dfrac{a+b}{2}\right)$,
c'est la méthode du point milieu.



Pour $n = 1$, $x_0 = a$, $x_1 = b$;

On trouve $\lambda_0 = \lambda_1 = \dfrac{b-a}{2}$.

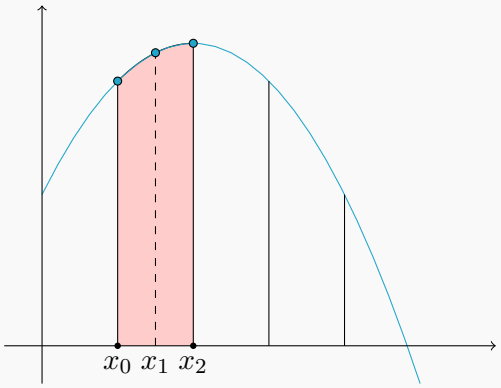
Soit $J(f) = (b-a) \cdot \dfrac{f(a)+f(b)}{2}$,
c'est la méthode des trapèzes.



Pour $n = 2$, $x_0 = a$, $x_1 = \dfrac{a+b}{2}$ et $x_2 = b$.

On trouve $\lambda_0 = \lambda_2 = \dfrac{b-a}{6}$, et $\lambda_1 = \dfrac{2(b-a)}{3}$.

Soit $J(f) = \dfrac{b-a}{6} \cdot (f(a) + 4f(\dfrac{a+b}{2}) + f(b))$, c'est la méthode de Simpson.



Remarque : On peut constater qu'une méthode avec $n+1$ points distincts peut être exacte pour des polynômes de degré $\leq n$.

Méthode	**Nombre de points**	**Degré d'exactitude max (d)**
Rectangle (gauche/droite)	1	0
Point Milieu	1	1
Trapèzes	2	1
Simpson	3	3

TABLE 1.1 – Degré maximal des polynômes pour lesquels l'intégration est exacte.

Préambule : La formule de Taylor-Lagrange

La formule de Taylor-Lagrange est un outil central en analyse numérique. Elle permet d'approcher une fonction régulière par un polynôme, tout en quantifiant l'erreur commise via un terme appelé le « reste ». Contrairement à la formule de Taylor-Young, qui décrit un comportement local (limite), la version de Lagrange fournit une égalité valable sur tout un intervalle, généralisant ainsi le Théorème des Accroissements Finis.

Théorème 1.2 : - Formule de Taylor-Lagrange

Soit f une fonction de classe $\mathcal{C}^{n+1}$ sur un intervalle I et soient $a, x \in I$. Il existe un réel c compris strictement entre a et x tel que :

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(c).$$

Théorème 1.3 : - Formule de Taylor avec reste intégral

Soit f une fonction de classe $\mathcal{C}^{n+1}$ sur un intervalle I et soient $a, x \in I$. Alors :

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\,\mathrm{d}t.$$

1.3 Intégration des fonctions régulières

On aura pas nécessairement $I(f) = J(f)$, on se demandera alors dans quelle mesure $J(f)$ est une bonne approximation de $I(f)$.

Notation : On notera $E(f) = J(f) - I(f)$ l'erreur de la méthode d'intégration.

Théorème 1.4 : - Erreur pour la méthode des rectangles à gauche

Si $n = 0$, $x_0 = a$ et f est de classe $\mathcal{C}^1$ sur $[a, b]$, alors

$$|E(f)| \leq \frac{(b-a)^2}{2} \sup_{x\in[a,b]} |f'(x)|.$$

Démonstration. Rappelons que pour la méthode des rectangles à gauche (avec un seul rectangle), l'approximation vaut $J(f) = (b-a)f(a)$. On peut réécrire ce terme sous forme intégrale : $(b-a)f(a) = \displaystyle\int_a^b f(a)\,dx$. L'erreur s'écrit alors :

$$E(f) = \int_a^b f(a)\,dx - \int_a^b f(x)\,dx = \int_a^b (f(a) - f(x))\,dx.$$

D'après l'inégalité des accroissements finis appliquée à f entre a et x (pour $x \in [a, b]$) :

$$|f(x) - f(a)| \leq |x-a| \cdot \sup_{t\in[a,b]} |f'(t)|.$$

En passant à la valeur absolue dans l'intégrale (inégalité triangulaire) :

$$\begin{aligned} |E(f)| &= \left| \int_a^b (f(a) - f(x))\,dx \right| \\ &\leq \int_a^b |f(a) - f(x)|\,dx \\ &\leq \left(\sup_{t\in[a,b]} |f'(t)| \right) \int_a^b (x-a)\,dx \end{aligned}$$

Or, le calcul de la primitive donne :

$$\int_a^b (x-a)\,dx = \left[\frac{(x-a)^2}{2} \right]_a^b = \frac{(b-a)^2}{2}.$$

On obtient bien le résultat annoncé :

$$|E(f)| \leq \frac{(b-a)^2}{2} \sup_{x\in[a,b]} |f'(x)|.$$

□

Théorème 1.5 : - Erreur pour la méthode du point milieu

Si $n = 0$, $x_0 = \dfrac{a+b}{2}$ et f est de classe $\mathcal{C}^2$ sur $[a,b]$, alors

$$|E(f)| \leq \frac{(b-a)^3}{24} \sup_{x\in[a,b]} |f''(x)|.$$

Démonstration. Posons $m = \dfrac{a+b}{2}$. D'après l'inégalité de Taylor-Lagrange à l'ordre 2 appliquée en m, pour tout $x \in [a,b]$, on a :

$$|f(x) - (f(m) + (x-m)f'(m))| \leq \frac{(x-m)^2}{2} \sup_{t\in[a,b]} |f''(t)|.$$

Rappelons que $J(f) = (b-a)f(m) = \displaystyle\int_a^b f(m)\, dx$. L'erreur s'écrit :

$$E(f) = J(f) - I(f) = \int_a^b f(m)\, dx - \int_a^b f(x)\, dx.$$

Remarquons que l'intégrale du terme impair $(x-m)$ est nulle par symétrie autour de m :

$$\int_a^b (x-m)f'(m)\, dx = f'(m)\left[\frac{(x-m)^2}{2}\right]_a^b = 0.$$

On peut donc introduire artificiellement ce terme nul dans l'expression de l'erreur pour faire apparaître la formule de Taylor :

$$\begin{aligned}
|E(f)| &= \left|\int_a^b (f(m) + (x-m)f'(m) - f(x))\ dx\right| \\
&\leq \int_a^b |f(x) - (f(m) + (x-m)f'(m))|\ dx \\
&\leq \int_a^b \frac{(x-m)^2}{2}\left(\sup_{t\in[a,b]} |f''(t)|\right) dx
\end{aligned}$$

Il ne reste plus qu'à calculer l'intégrale du polynôme. En posant le changement de variable $u = x - m$:

$$\int_a^b (x-m)^2\, dx = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} u^2\, du = \left[\frac{u^3}{3}\right]_{-\frac{b-a}{2}}^{\frac{b-a}{2}} = \frac{2}{3}\left(\frac{b-a}{2}\right)^3 = \frac{(b-a)^3}{12}.$$

En n'oubliant pas le facteur $\dfrac{1}{2}$ présent dans l'inégalité, on obtient finalement :

$$|E(f)| \leq \frac{1}{2}\cdot\frac{(b-a)^3}{12} \sup_{x\in[a,b]} |f''(x)| = \frac{(b-a)^3}{24} \sup_{x\in[a,b]} |f''(x)|.$$

□

Théorème 1.6 : Démo. admise

Soit une méthode d'intégration J à n points exactes pour des polynômes de degré $p \geq n$. Alors si f est de classe $\mathcal{C}^{p+1}$ sur $[a,b]$,telle que :

$$|E(f)| = K(b-a)^{p+2} \sup_{x\in[a,b]} |f^{(p+1)}(x)|,$$

où $K = K(p)$ est une constante indépendante de f, a et b.

Remarque : On a $|E(f)| \leq |K|(b-a)^{p+2} \sup_{c\in[a,b]} |f^{(p+1)}(c)|$, mais on ne peut pas en déduire que $E(f) \to 0$ quand le nombre de points d'intégration augmente.

Nous somme donc conduit à introduire des méthodes dites d'intégrations composites.

1.4 Méthodes composites

Principe : L'idée est de découper l'intervalle $[a, b]$ en plusieurs petits sous-intervalles, et d'appliquer une méthode d'intégration simple (dite « locale ») sur chacun d'eux. Soient des points d'intégration fixés sur l'intervalle de référence $[0, 1]$, tels que $0 \leq \xi_0 < \cdots < \xi_n \leq 1$. Soit J une méthode d'intégration élémentaire, exacte pour les polynômes de degré n sur $[0, 1]$:

$$J(f) = \sum_{i=0}^{n} \mu_i f(\xi_i),$$

où les μ_i sont les poids d'intégration associés. **Passage à un intervalle quelconque** :

Considérons un polynôme Q de degré $\leq n$ sur un intervalle général $[a, b]$. Par le changement de variable affine $t = (b-a)\xi + a$ (où $dt = (b-a)d\xi$), on obtient :

$$\int_a^b Q(t)\, dt = (b-a) \int_0^1 Q((b-a)\xi + a)\, d\xi.$$

L'approximation devient alors :

$$\int_a^b Q(t)\, dt \approx (b-a) \sum_{i=0}^{n} \mu_i Q((b-a)\xi_i + a).$$

Définition 1.1

Soit $\sigma = (x_0, x_1, \ldots, x_m)$ une subdivision de $[a, b]$ telle que $a = x_0 < x_1 < \cdots < x_m = b$. Une méthode d'intégration composite consiste à approcher l'intégrale $I(f)$ par la somme des approximations sur chaque sous-intervalle $[x_{k-1}, x_k]$:

$$\begin{aligned} I(f) &= \int_a^b f(x)dx \\ &= \sum_{k=1}^{m} \int_{x_{k-1}}^{x_k} f(x)dx \\ &= \sum_{k=1}^{m} (x_k - x_{k-1}) \int_0^1 f((x_k - x_{k-1})\xi + x_{k-1})d\xi \\ &\approx \sum_{k=1}^{m} (x_k - x_{k-1}) \sum_{i=0}^{n} \mu_i f\left((x_k - x_{k-1})\xi_i + x_{k-1}\right) \\ &= H(f). \end{aligned}$$

Cas particulier : Subdivision régulière

Si la subdivision est régulière, on a $x_k = a + kh$ avec le pas constant $h = \dfrac{b-a}{m}$. La formule se simplifie en :

$$H(f) = h \sum_{k=1}^{m} \sum_{i=0}^{n} \mu_i f\left(a + (k - 1 + \xi_i)h\right).$$

Théorème 1.7 : - Erreur globale pour les méthodes composites

On suppose que la méthode locale est d'ordre p. On a alors la majoration globale :

$$|H(f) - I(f)| \leq |K|(b-a) \sup_{x\in[a,b]} |f^{(p+1)}(x)|h^{p+1},$$

où K ne dépend que de p et h le pas de la subdivision.

Démonstration. Notons m le nombre de sous-intervalles de la subdivision régulière, tel que $h = \dfrac{b-a}{m}$. L'erreur globale $E(f) = H(f) - I(f)$ est la somme des erreurs commises sur chaque sous-intervalle $[x_k, x_{k+1}]$ (pour k allant de 0 à $m-1$). D'après l'inégalité triangulaire :

$$|E(f)| = \left|\sum_{k=0}^{m-1} e_k(f)\right| \leq \sum_{k=0}^{m-1} |e_k(f)|,$$

où $e_k(f)$ représente l'erreur locale sur l'intervalle $[x_k, x_{k+1}]$. D'après le théorème sur l'erreur locale, nous savons que sur un intervalle de largeur h, l'erreur est majorée par :

$$|e_k(f)| \leq |K|h^{p+2} \sup_{t\in[x_k,x_{k+1}]} |f^{(p+1)}(t)|.$$

On peut majorer le sup local par le sup global sur tout l'intervalle $[a, b]$:

$$\sup_{t\in[x_k,x_{k+1}]} |f^{(p+1)}(t)| \leq \sup_{x\in[a,b]} |f^{(p+1)}(x)| = M_{p+1}.$$

En injectant cette majoration dans la somme :

$$\begin{aligned} |E(f)| &\leq \sum_{k=0}^{m-1} \left(|K|h^{p+2}M_{p+1}\right) \\ &= m \cdot |K|h^{p+2}M_{p+1} \\ &= (m \cdot h) \cdot |K|h^{p+1}M_{p+1}. \end{aligned}$$

Or, la longueur totale de l'intervalle est donnée par $b - a = m \cdot h$. On obtient finalement :

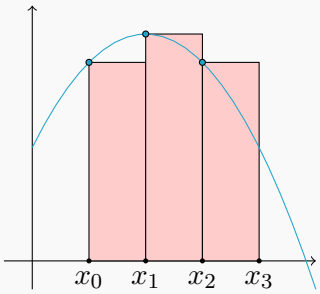
$$|H(f) - I(f)| \leq (b-a)|K|h^{p+1} \sup_{x\in[a,b]} |f^{(p+1)}(x)|.$$

□

Remarque : Cette fois-ci, l'erreur tend bien vers 0 quand $h \to 0$, i.e. quand $m \to +\infty$ à p **fixé**.

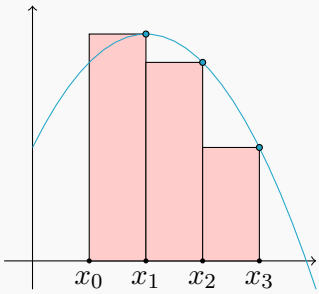
Méthode des rectangle à gauche

$$H(f) = h \sum_{k=1}^{m} f(a + (k-1)h)$$



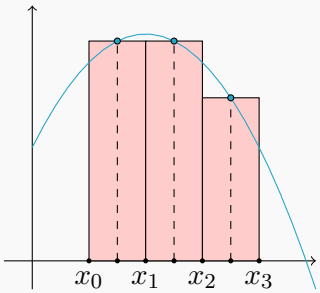
Méthode des rectangle à droite

$$H(f) = h \sum_{k=1}^{m} f(a + kh)$$



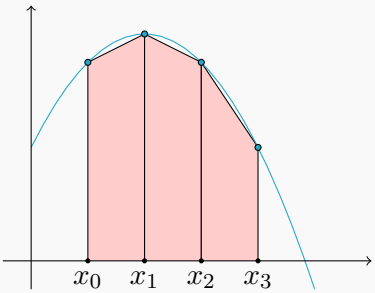
Méthode des points milieu

$$H(f) = h \sum_{k=1}^{m} f(a + (k - \frac{1}{2})h)$$



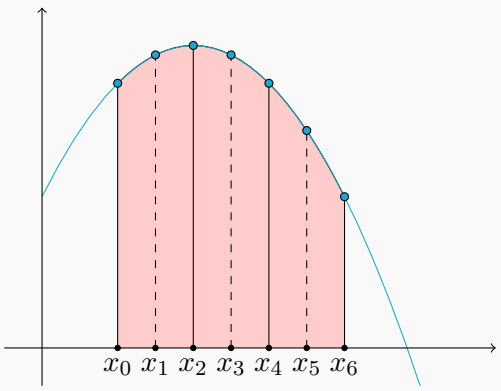
Méthode des trapèzes

$$H(f) = h \sum_{k=1}^{m} \frac{1}{2}(f(a+(k-1)h)+f(a+kh)).$$



Méthode de Simpson

$$H(f) = h \sum_{k=1}^{m} \frac{1}{6}(f(a + (k-1)h) + f(a + (k - \frac{1}{2})h) + f(a + kh)).$$



1.5 Ordre d'une méthode

Définition 1.2

On dit que la méthode est d'ordre ℓ pour la fonction f s'il existe c non nul telle que :

$$|H(f) - I(f)| \sim ch^{\ell}, \qquad \text{quand } h \to 0.$$

Remarque : Dans le théorème précédent, si l'estimation est optimale, dans ce cas $\ell = p+1$.

Pour $n = 0$, si $x_0 = \dfrac{a+b}{2}$ (méthode du point milieu), la méthode est d'ordre 2.

Par contre si $x_0 = a$ (méthode des rectangles à gauche), elle est d'ordre 1 (avec f une fonction adéquate).

La position des points d'intégration joue un rôle crucial.

1.6 Méthode de Gauss

La méthode de Gauss détermine les points d'intégration ξ_i sur l'intervalle $[0, 1]$ de sorte que la méthode $J(f) = \sum_{i=0}^{n} \mu_i f(\xi_i)$, déjà exacte pour les polynômes de degré $\leq n$, soit exacte pour des polynômes de degré le plus élevé possible. Considérons sur l'ensemble des polynômes le produit scalaire $\langle P, Q\rangle = \int_0^1 P(t)Q(t)\, dt$.

On construit une suite de polynômes orthogonaux. Pour l'exemple, on prendra la famille issue de l'orthogonalisation de la suite $x \mapsto x^n$ par le procédé de Gram-Schmidt (ce sont les polynômes de Legendre translatés sur $[0, 1]$) :

- $L_0(x) = 1$
- $L_1(x) = \sqrt{3}(2x - 1)$
- $L_2(x) = \sqrt{5}(6x^2 - 6x + 1)$
- $L_3(x) = \sqrt{7}(20x^3 - 30x^2 + 12x - 1)$

Note : Les facteurs racines carrées servent à normaliser les polynômes (norme 1), mais n'affectent pas leurs zéros.

On admet que L_k possède k zéros distincts dans $]0, 1[$.

Théorème 1.8 : - Quadrature de Gauss

Si les ξ_i (pour $i = 0 \dots n$) sont les zéros du polynôme orthogonal L_{n+1}, alors la méthode $J(f) = \sum_{i=0}^{n} \mu_i f(\xi_i)$ est exacte pour les polynômes de degré $\leq 2n + 1$.

Démonstration. Soit P un polynôme de degré $\leq 2n + 1$.

Effectuons la division euclidienne de P par L_{n+1} :

$$P = q \cdot L_{n+1} + r$$

Comme $\deg(P) \leq 2n + 1$ et $\deg(L_{n+1}) = n + 1$, le quotient q est de degré $\leq n$ et le reste r est de degré $\leq n$.

Calculons l'intégrale. Par orthogonalité, L_{n+1} est orthogonal à tout polynôme de degré inférieur (donc orthogonal à q) :

$$\int_0^1 q(t)L_{n+1}(t)\,dt = \langle q, L_{n+1}\rangle = 0$$

Ainsi :

$$\begin{aligned}\int_0^1 P(t)\,dt &= \int_0^1 q(t)L_{n+1}(t)\,dt + \int_0^1 r(t)\,dt \\ &= 0 + \int_0^1 r(t)\,dt \\ &= \sum_{i=0}^{n} \mu_i r(\xi_i) \quad \text{(car la méthode est exacte pour } r \in \mathbb{R}_n[X])\end{aligned}$$

Or, pour les points de quadrature ξ_i (qui sont les racines de L_{n+1}), on a $L_{n+1}(\xi_i) = 0$. Donc :

$$P(\xi_i) = q(\xi_i)\underbrace{L_{n+1}(\xi_i)}_{0} + r(\xi_i) = r(\xi_i).$$

On conclut :

$$\int_0^1 P(t)\,dt = \sum_{i=0}^{n} \mu_i P(\xi_i).$$

□

Rappelons que l'ordre de la méthode est le degré maximal exact +1.

- **Pour n=1 (2 points)** : On cherche les racines de L_2.

  $$6x^2 - 6x + 1 = 0 \implies \xi = \frac{6 \pm \sqrt{12}}{12} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$$

  On obtient $\xi_0 \approx 0.211$ et $\xi_1 \approx 0.789$. La méthode est d'ordre $2(1) + 2 = 4$.
- **Pour n=2 (3 points)** : On cherche les racines de L_3.

  $$20x^3 - 30x^2 + 12x - 1 = 0$$

  $1/2$ est racine évidente. En factorisant, on trouve les deux autres racines :

  $$\xi_0 = \frac{1}{2} - \frac{\sqrt{15}}{10}, \quad \xi_1 = \frac{1}{2}, \quad \xi_2 = \frac{1}{2} + \frac{\sqrt{15}}{10}.$$

  La méthode est d'ordre $2(2) + 2 = 6$.

1.7 Polynômes Orthogonaux : définition par récurrence

On rappelle l'orthogonalisation de Gram-Schmidt :

Soit E un espace vectoriel muni d'un produit scalaire $\langle\cdot,\cdot\rangle$. Si $(f_1, \ldots, f_n)$ est une famille libre de E, on construit par récurrence une famille orthonormale $(e_1, \ldots, e_n)$ en posant :

$$e_k = \frac{f_k - \sum_{j=1}^{k-1}\langle e_j, f_k\rangle e_j}{\|f_k - \sum_{j=1}^{k-1}\langle e_j, f_k\rangle e_j\|}$$

Alors pour tout $k \in [\![1, n]\!]$, on a :

— $\text{Vect}(f_1, \dots, f_k) = \text{Vect}(e_1, \dots, e_k)$;
— $\langle e_k, f_k \rangle > 0$.

Méthode non normalisée :

$$e_k = f_k - \sum_{j=1}^{k-1} \langle e_j, f_k \rangle \frac{e_j}{|e_j|^2}$$

Théorème 1.9 : A6

Existence et Récurrence à 3 termes

Il existe une unique famille orthogonale de polynômes $(p_n)_{n\in\mathbb{N}}$ telle que pour tout n, p_n soit de degré n et de coefficient dominant 1 (polynôme unitaire).

Cette famille est donnée par la relation de récurrence :

$$\begin{cases} p_0(x) & = 1 \\ p_1(x) & = x - \alpha_0 \\ p_{n+1}(x) & = (x - \alpha_n)p_n(x) - \lambda_n p_{n-1}(x), \quad \forall n \geq 1 \end{cases}$$

avec les coefficients :

$$\alpha_n = \frac{\langle xp_n, p_n \rangle}{\|p_n\|^2} \quad (n \geq 0) \qquad \text{et} \qquad \lambda_n = \frac{\|p_n\|^2}{\|p_{n-1}\|^2} \quad (n \geq 1)$$

Démonstration. **1. Existence et Unicité** L'existence découle du procédé de Gram-Schmidt appliqué à $(1, x, x^2, \dots)$. Comme on impose le coefficient dominant égal à 1, chaque p_n est défini de manière unique par projection : $p_n = x^n - \text{proj}_{\mathbb{R}_{n-1}[X]}(x^n)$.

2. Formule de récurrence (Le "coeur" de la preuve) Soit $n \geq 1$. Le polynôme xp_n est de degré $n+1$ et unitaire. Décomposons-le dans la base $(p_0, \dots, p_{n+1})$:

$$xp_n = \sum_{k=0}^{n+1} c_k p_k$$

Comme xp_n et p_{n+1} sont unitaires, le coefficient devant p_{n+1} est $c_{n+1} = 1$. Pour les autres coefficients, par orthogonalité :

$$c_k = \frac{\langle xp_n, p_k \rangle}{\|p_k\|^2}$$

L'argument de symétrie : On sait que $\langle xp_n, p_k \rangle = \int_a^b xp_n(x)p_k(x)dx = \langle p_n, xp_k \rangle$.

— Si $k < n-1$, alors $\deg(xp_k) = k+1 < n$.
— Or, p_n est orthogonal à tout polynôme de degré strictement inférieur à n.
— Donc $\langle p_n, xp_k \rangle = 0$ pour tout $k \in \{0, \dots, n-2\}$.

D'où $c_k = 0$ pour $k < n-1$. La somme se réduit à trois termes :

$$xp_n = p_{n+1} + c_n p_n + c_{n-1} p_{n-1}$$

3. Calcul des coefficients On pose $\alpha_n = c_n$. Par projection : $\alpha_n = \frac{\langle xp_n, p_n \rangle}{\|p_n\|^2}$.

Pour $\lambda_n = c_{n-1}$, on utilise $\langle xp_n, p_{n-1}\rangle = \langle p_n, xp_{n-1}\rangle$. Comme p_{n-1} est unitaire de degré $n-1$, on peut écrire $xp_{n-1} = p_n + R$ où $R \in \mathbb{R}_{n-1}[X]$.

$$\langle p_n, xp_{n-1}\rangle = \langle p_n, p_n + R\rangle = \langle p_n, p_n\rangle + \underbrace{\langle p_n, R\rangle}_{=0} = \|p_n\|^2$$

On obtient donc : $\lambda_n = \dfrac{\|p_n\|^2}{\|p_{n-1}\|^2}$.

En réordonnant $p_{n+1} = xp_n - c_n p_n - c_{n-1}p_{n-1}$, on trouve bien la relation annoncée. □

Théorème 1.10 : A7

Zéros des polynômes orthogonaux

Soit $n \geq 1$. Le polynôme p_n admet exactement n racines réelles distinctes, et elles sont toutes situées dans l'intervalle ouvert $]a, b[$.

Démonstration. Soient $x_1, \ldots, x_k$ les racines distinctes de p_n situées dans $]a, b[$ où p_n **change effectivement de signe** (racines de multiplicité impaire).

Par définition, ces racines sont dans l'intervalle, donc $k \leq n$. On veut montrer que $k = n$.

Construisons le polynôme $Q(x) = \prod_{i=1}^{k}(x - x_i)$ (si $k = 0$, on pose $Q = 1$). Considérons alors le produit $p_n(x)Q(x)$ sur $]a, b[$:

— En chaque x_i, p_n change de signe et $(x - x_i)$ change de signe simultanément. Leur produit ne change donc pas de signe en ces points.

— Aux éventuelles autres racines de p_n (multiplicité paire), p_n ne change pas de signe.

Ainsi, la fonction $f(x) = p_n(x)Q(x)$ garde un **signe constant** sur $[a, b]$ et n'est pas identiquement nulle.

Par conséquent, son intégrale est strictement non nulle :

$$I = \langle p_n, Q\rangle = \int_a^b p_n(x)Q(x)dx \neq 0$$

Supposons par l'absurde que $k < n$. Alors $\deg(Q) = k < n$. Par la propriété d'orthogonalité, p_n est orthogonal à tout polynôme de degré strictement inférieur à n. On devrait donc avoir :

$$\langle p_n, Q\rangle = 0$$

Ceci contredit $I \neq 0$. On en conclut que l'hypothèse $k < n$ est impossible, donc $k = n$.

Comme p_n est de degré n et possède n racines où il change de signe dans $]a, b[$, ces racines sont nécessairement **simples** et toutes contenues dans l'intervalle ouvert. □

Chapitre 2

Résolution numérique d'EDO

Soit $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue.

Pour $u_0 \in \mathbb{R}^d$, le problème de Cauchy en x_0 pour f est de trouver $u \in \mathcal{C}^1([0,T], \mathbb{R}^d)$ solution de l'équation différentielle :

$$\begin{cases} u' & = f(t,u) \\ u(0) & = u_0. \end{cases}$$

Théorème 2.1 :

- Cauchy-Lipschitz On suppose de plus que pour tout compact $K \subset \mathbb{R}^d$, il existe $L > 0$ tel que :

$$\forall t \in [0,T],\ \forall x, y \in K, \qquad |f(t,x) - f(t,y)| \leq L|x-y|.$$

Alors, pour tout $u_0 \in \mathbb{R}^d$, il existe $T' \leq T$, et une unique solution maximale $u \in \mathcal{C}([0,T'], \mathbb{R}^d)$ au problème de Cauchy.

Remarque : Une manière simple de satisfaire cette condition est de supposer que f est de classe $\mathcal{C}^1$.

Question : Comment calculer (ou approcher) u de manière approchée ?

Plus précisément, on s'intéresse à l'intervalle $[0,T]$, que l'on échantillonne en $N+1$ points : $0 = t_0 < t_1 < \cdots < t_N = T$, où $t_N = Nh$ est l'incrément avec $h = \dfrac{T}{N}$ et on cherche à évaluer $u(t_j)$ pour tout $j \in [\![0, N]\!]$. On notera u_n la valeur approchée de $u(t_n)$.

2.1 Idée principale de la méthode

Pour résoudre le problème de Cauchy, on constate qu'il est équivalent à trouver u tel que :

$$u(t) = u_0 + \int_0^t f(s, u(s))\, ds.$$

En particulier, entre deux points de la grille de temps t_n and t_{n+1}, on a la relation exacte :

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} f(t, u(t))\, dt.$$

L'idée fondamentale des méthodes à une évaluation de f est d'approcher cette intégrale par une formule de quadrature numérique simple du type :

$$\int_{t_n}^{t_{n+1}} g(t)\, dt \approx h \cdot g(\xi), \quad \xi \in [t_n, t_{n+1}].$$

2.1.1 Méthode d'Euler explicite (ou Euler progressif)

On utilise la méthode des rectangles à gauche pour l'intégrale :

$$\int_{t_n}^{t_{n+1}} f(t, u(t))\, dt \approx h \cdot f(t_n, u(t_n)).$$

On obtient la récurrence suivante :

$$u_{n+1} = u_n + hf(t_n, u_n)$$

Caractéristique : La méthode est dite *explicite* car u_{n+1} se calcule directement à partir des données connues au temps t_n.

2.1.2 Méthode d'Euler implicite (ou Euler régressif)

On utilise ici la méthode des rectangles à droite :

$$\int_{t_n}^{t_{n+1}} f(t, u(t))\, dt \approx h \cdot f(t_{n+1}, u(t_{n+1})).$$

Le schéma devient :

$$u_{n+1} = u_n + hf(t_{n+1}, u_{n+1})$$

Caractéristique : La méthode est *implicite*. Pour trouver u_{n+1}, il faut résoudre une équation (souvent non-linéaire) à chaque pas de temps, par exemple avec une méthode de Newton. Bien que plus coûteuse, elle est bien plus stable pour les équations dites "raides" (stiff).

2.1.3 Exemple comparatif

Considérons l'EDO scalaire $u' = \lambda u$ avec $\lambda < 0$.

— **Euler explicite :** $u_{n+1} = (1 + h\lambda)u_n$. La solution n'est stable que si $|1 + h\lambda| \leq 1$, soit $h \leq \dfrac{2}{|\lambda|}$.

— **Euler implicite :** $u_{n+1} = \dfrac{1}{1 - h\lambda} u_n$. La solution est stable pour tout $h > 0$, car $|1 - h\lambda| > 1$.

2.2 Méthodes explicites avec prédicteurs

Dans les méthodes d'Euler, on n'évalue la fonction f qu'une seule fois. Pour gagner en précision, on peut multiplier les évaluations de f entre t_n et t_{n+1}.

- **Méthode du point milieu (RK2)** :

  L'idée est d'utiliser la quadrature du point milieu pour gagner en précision :

  $$\int_{t_n}^{t_{n+1}} f(t, u(t))\ dt \simeq hf\left(t_n + \frac{h}{2},\ u\left(t_n + \frac{h}{2}\right)\right).$$

  On effectue une étape intermédiaire (un "sous-pas") en utilisant Euler explicite sur $[t_n, t_n + \dfrac{h}{2}]$ pour estimer la valeur au centre de l'intervalle :

  $$u_{n+1} = u_n + hf\left(t_n + \frac{h}{2},\ u_n + \frac{h}{2} f(t_n, u_n)\right).$$

  On dit que $u_n + \dfrac{h}{2} f(t_n, u_n)$ est un predicteur au temps $t = t_n + h/2$.

- **Méthode de Heun** :

  On utilise un prédicteur au temps $t = t_{n+1}$ noté $\widetilde{u}_{n+1} := u_n + hf(t_n, u_n)$ que l'on injecte dans la méthode des trapèzes :

$$u_{n+1} = u_n + \frac{h}{2}\left(f(t_n, u_n) + f(t_n + h,\ u_n + hf(t_n, u_n))\right).$$

2.3 Formalisme général d'une méthode

Toutes les méthodes précédentes rentrent dans le cadre général suivant :

$$u_{n+1} = u_n + h\phi(t_n, u_n, h),$$

où la fonction ϕ (appelée fonction d'incrément) encapsule tous les calculs intermédiaires effectués durant l'intervalle $[t_n, t_{n+1}]$.

Définition 2.1

On appelle erreur de convergence les quantités :

$$E_n := u(t_n) - u_n$$

- On dit que la méthode est convergente si $\max_{n=0,\dots,N} |E_n| \to 0$ quand $h \to 0$.
- On dit que la méthode est (convergente) d'ordre p s'il existe $\delta, C > 0$ indépendantes de h tels que si $|h| \leq \delta$ alors :

$$\max_{n=0,\dots,N} |E_n| \leq Ch^p.$$

Pour montrer la convergence, en général, on utilise deux notions intermédiaires : la consistance et la stabilité.

Définition 2.2

L'erreur de consistance (relative à une solution exacte u) est la quantité :

$$e_n = e_n(u) := u(t_{n+1}) - u(t_n) - h\phi(t_n, u(t_n), h).$$

Autrement dit, e_n est l'erreur que commet le schéma sur $[t_n, t_{n+1}]$.

- On dit que la méthode est consistante si pour toute solution sur $[0, T]$,

$$\sum_{n=0}^{N-1} |e_n| \to 0, \quad \text{quand } h \to 0.$$

- On dit que la méthode est consistante d'ordre $p \geq 1$ s'il existe $\delta, C > 0$ indépendantes de h tels que si $|h| \leq \delta$, alors pour toute solution exacte v , $\sum_{n=0}^{N-1} |e_n| \leq Ch^p$.

Définition 2.3

On dit que la méthode est stable s'il existe une constante $S > 0$ (appelée constante de stabilité) tel que si : $(\varepsilon_n)_{n=0,\dots,N-1}$, $(u_n)_{n=0,\dots,N}$, $(v_n)_{n=0,\dots,N}$ tels que $\forall n = 0, \dots, N-1$ on a :

$$\begin{cases} u_{n+1} &= u_n + h\phi(t_n, u_n, h) \\ v_{n+1} &= v_n + h\phi(t_n, v_n, h) + \varepsilon_n. \end{cases}$$

Alors $\forall n = 0, \dots, N$, on a :

$$|u_n - v_n| \leq S \left(|u_0 - v_0| + \sum_{j=0}^{N-1} |\varepsilon_j| \right).$$

Théorème 2.2 :

Une méthode stable et consistante est convergente.

Si de plus, elle est consistante d'ordre p alors la méthode est convergente d'ordre p.

Démonstration. On cherche à majorer l'erreur globale $E_n = u(t_n) - u_n$.

1. Écriture des suites : Le schéma numérique est : $u_{n+1} = u_n + h\phi(t_n, u_n, h)$. La solution exacte vérifie (par définition de l'erreur de consistance e_n) : $u(t_{n+1}) = u(t_n) + h\phi(t_n, u(t_n), h) + e_n$.

2. Stabilité : On applique la propriété de stabilité aux suites u_n et $v_n = u(t_n)$ avec perturbation $\varepsilon_n = e_n$:

$$|u(t_n) - u_n| \leq S \left(|u(0) - u_0| + \sum_{k=0}^{n-1} |e_k| \right).$$

3. Conclusion : Comme la méthode est consistante d'ordre p, on a :

$$\sum_{k=0}^{N-1} |e_k| \leq Ch^p.$$

□

Définition 2.4

Soit v la solution de $v' = f(s, v)$ avec comme donnée initiale $v(t) = u$ au temps t.

Alors $\dfrac{d^j f}{dt^j}(t, u)$ est la dérivée d'ordre j de la fonction $s \mapsto f(s, v(s))$ au point t.

Par la formule de dérivation composée, $\dfrac{d^j f}{dt^j}$ est bien défini (et continue) dès que f est de classe $\mathcal{C}^j$. On a par exemple :

$$\begin{aligned} \frac{df}{dt}(t, u) &= v''(t) = \partial_t f(t, v(t)) + v'(t)\partial_u f(t, v(t)) \\ &= \partial_t f(t, v(t)) + f(t, v(t))\partial_u f(t, v(t)). \end{aligned}$$

Théorème 2.3 : Admise

- La méthode est consistante si $\forall t \in [0, T]$, $\forall n \in \mathbb{R}$, $\phi(t, u, 0) = f(t, u)$.
- On suppose f et ϕ de classe $\mathcal{C}^p$. La méthode est consistante d'ordre $p \geq 1$ si et seulement si,

$$\forall j = 0, \dots, p-1, \qquad \frac{\partial^j \phi}{\partial h^j}(t, u, 0) = \frac{1}{j+1} \frac{d^j f}{dt^j}(t, u).$$

Remarque :

- Par exemple, Euler explicite vérifie le premier point, donc elle est consistante.
- Méthode de Heun : $\partial_h \phi(t,u,0) = (\frac{1}{2})(\partial_t f(t,u) + f(t,u)\partial_u f(t,u)) = (\frac{1}{2})\frac{df}{dt}(t,u)$, donc elle est consistante d'ordre 2.
- Le même calcul donne que la méthode du point milieu est consistante d'ordre 2.

Théorème 2.4 : Admise

Si la fonction ϕ est lipschitzienne en la seconde variable, *i.e.* il existe $\delta, L > 0$ telle que :

$$\forall (t,u,h) \in [0,T] \times \mathbb{R}^n \times [0,\delta], \qquad |\phi(t,u,h) - \phi(t,v,h)| \leq L|u-v|,$$

alors la méthode ϕ est stable.

Si f est lipschitzienne en sa seconde variable alors les méthodes d'Euler, du point milieu et de Heun le sont également et donc elles sont stables.

2.4 Méthode de Runge-Kutta (RK)

2.4.1 Construction générale

Le principe des méthodes de Runge-Kutta est de généraliser les méthodes précédentes (Euler, Heun, Point milieu) pour obtenir une meilleure précision *i.e.* un meilleur ordre de consistance. On se donne un entier $q \geq 1$ (le nombre d'étapes). On définit q instants intermédiaires $t_{n,i} = t_n + c_i h$, où $0 = c_1 \leq c_2 \leq \cdots \leq c_q \leq 1$. L'objectif est d'utiliser des prédicteurs de la solution $u_{n,i} \simeq u(t_{n,i})$ en ces points, ainsi que les pentes associées $p_{n,i} = f(t_{n,i}, u_{n,i})$. Pour cela, on repart de la formulation intégrale exacte :

$$u(t_n + c_i h) = u(t_n) + h \int_0^{c_i} f(t_n + sh, u(t_n + sh))\, ds.$$

Approximation : On remplace l'intégrale par une formule de quadrature numérique utilisant les pentes déjà calculées aux étapes précédentes $j < i$ (méthode explicite).

$$\int_0^{c_i} f(\dots)\, ds \simeq \sum_{j=1}^{i-1} a_{ij} f(t_n + c_j h, u(t_n + c_j h)).$$

et donc :

$$u(t_n + c_i h) \simeq u(t_n) + h \sum_{j=1}^{i-1} a_{ij} f(t_n + c_j h, u(t_n + c_j h))\, ds.$$

En injectant les prédicteurs en $t_n + c_j h$ pour $j < i$ calculés précedemment, on obtient le prédicteur en $t_n + c_i$. L'algorithme est défini dans la section suivante.

2.4.2 Algorithme explicite

Pour aller de u_n à u_{n+1}, on calcule successivement pour $i = 1, \dots, q$:

- **L'instant intermédiaire** :

$$t_{n,i} = t_n + c_i h$$

- **Le prédicteur** (par combinaison des pentes précédentes) :

$$u_{n,i} = u_n + h \sum_{j=1}^{i-1} a_{ij} p_{n,j}$$

(Si $i = 1$, la somme est vide, donc $u_{n,1} = u_n$).

- **La pente intermédiaire** :

$$p_{n,i} = f(t_{n,i}, u_{n,i})$$

Une fois les q pentes $p_{n,1}, \ldots, p_{n,q}$ calculées, la valeur finale est donnée par une quadrature sur tout l'intervalle $[0, 1]$:

$$\boxed{u_{n+1} = u_n + h \sum_{i=1}^{q} b_i p_{n,i}}$$

2.4.3 Tableau de Butcher

On synthétise les coefficients de la méthode (c_i, a_{ij}, b_i) dans un tableau appelé Tableau de Butcher :

$$\begin{array}{c|ccccc} c_1 & 0 & & & & \\ c_2 & a_{2,1} & 0 & & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ c_q & a_{q,1} & a_{q,2} & \ldots & a_{q,q-1} & 0 \\ \hline & b_1 & b_2 & \ldots & b_{q-1} & b_q \end{array}$$

Conditions de consistance : Pour que la méthode ait un sens physique (ordre au moins 1), la somme des poids doit correspondre à la longueur de l'intervalle d'intégration. On impose donc généralement :

$$c_i = \sum_{j=1}^{i-1} a_{ij} \quad \text{et} \quad \sum_{i=1}^{q} b_i = 1.$$

Voici les tableaux correspondants aux méthodes vues précédemment :

- **Point Milieu (RK2)** : On avance d'un demi-pas ($c_2 = \frac{1}{2}$), on évalue la pente, et on l'utilise pour tout l'intervalle ($b_2 = 1$).

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

- **Méthode de Heun (RK2)** : On avance d'un pas complet ($c_2 = 1$), et on fait la moyenne des pentes début/fin ($b_1 = b_2 = \frac{1}{2}$).

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- **Runge-Kutta d'ordre 4 (RK4)** : C'est la méthode la plus célèbre, combinaison de Simpson et du point milieu.

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Utilisation du tableau pour la **Méthode de Runge-Kutta d'ordre 4 (RK4)** :
En utilisant le tableau de Butcher :

$$\begin{cases} p_{n,1} & = f(t_n, u_n) \\ p_{n,2} & = f(t_n + \frac{h}{2},\ u_n + (\frac{h}{2})p_{n,1}) \\ p_{n,3} & = f(t_n + \frac{h}{2},\ u_n + (\frac{h}{2})p_{n,2}) \\ p_{n,4} & = f(t_n + h,\ u_n + hp_{n,3}) \\ u_{n+1} & = u_n + \frac{h}{6}(p_{n,1} + 2p_{n,2} + 2p_{n,3} + p_{n,4}) \end{cases}$$

Chapitre 3

Calcul des valeurs propres

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée. Le problème des valeurs propres consiste à trouver les scalaires $\lambda \in \mathbb{C}$ et les vecteurs non nuls $u \in \mathbb{C}^n$ tels que :

$$Au = \lambda u.$$

Le polynôme caractéristique est défini par $P_A(\lambda) = \det(A - \lambda I_n)$. Les valeurs propres sont les racines de ce polynôme.

Obstacle théorique : D'après le théorème d'Abel-Ruffini, il n'existe pas de formule algébrique générale (exprimable par radicaux) pour résoudre les équations polynomiales de degré $n \geq 5$.

Contrairement à la résolution de systèmes linéaires (ex : Pivot de Gauss) qui est une méthode directe (exacte en un nombre fini d'opérations), le calcul des valeurs propres repose nécessairement sur des méthodes itératives.

3.1 Localisation des valeurs propres

Avant de calculer précisément une valeur propre, il est utile de la localiser dans le plan complexe.

Définition 3.1

Soit $A = (a_{ij})$ une matrice carrée d'ordre n. Pour tout $i \in [\![1, n]\!]$, on définit le i-ème disque de Gershgorin D_i par :

$$D_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq R_i\}, \qquad \text{avec } R_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

L'ensemble D_i n'est jamais vide. En effet, le centre du disque a_{ii} appartient toujours à D_i.

- Calculons la distance du centre à lui-même : $|a_{ii} - a_{ii}| = 0$.
- Le rayon $R_i = \sum_{j\neq i} |a_{ij}|$ est une somme de nombres positifs ou nuls (valeurs absolues), donc $R_i \geq 0$.
- On a donc bien l'inégalité : $|a_{ii} - a_{ii}| = 0 \leq R_i$.

Ainsi, $a_{ii} \in D_i$, ce qui prouve que $D_i \neq \emptyset$.

Théorème 3.1 : - Localisation de Gershgorin

Toutes les valeurs propres de la matrice A sont contenues dans l'union des disques de Gershgorin :

$$\mathrm{Sp}(A) \subset \bigcup_{i=1}^{n} D_i.$$

Démonstration. Soit $\lambda \in \mathrm{Sp}(A)$ une valeur propre et $x = (x_1, \ldots, x_n)^T$ un vecteur propre associé ($x \neq 0$).

1. Choix de la composante dominante : Puisque x n'est pas nul, on peut choisir un indice k tel que le module de la composante x_k soit maximal :

$$|x_k| = \max_{1 \leq j \leq n} |x_j| > 0.$$

2. Isolation du terme diagonal : La k-ième ligne de l'égalité $Ax = \lambda x$ s'écrit :

$$\sum_{j=1}^{n} a_{kj} x_j = \lambda x_k.$$

On isole le terme diagonal $a_{kk}x_k$ dans la somme :

$$a_{kk}x_k + \sum_{j \neq k} a_{kj} x_j = \lambda x_k \quad \Leftrightarrow \quad (\lambda - a_{kk})x_k = \sum_{j \neq k} a_{kj} x_j.$$

3. Majoration : En passant au module et en utilisant l'inégalité triangulaire :

$$|\lambda - a_{kk}| \cdot |x_k| = \left| \sum_{j \neq k} a_{kj} x_j \right| \leq \sum_{j \neq k} |a_{kj}| \cdot |x_j|.$$

Or, par définition de l'indice k, on a pour tout j, $|x_j| \leq |x_k|$. On peut donc majorer chaque $|x_j|$ par $|x_k|$ dans la somme :

$$|\lambda - a_{kk}| \cdot |x_k| \leq \sum_{j \neq k} |a_{kj}| \cdot |x_k| = |x_k| \underbrace{\sum_{j \neq k} |a_{kj}|}_{R_k}.$$

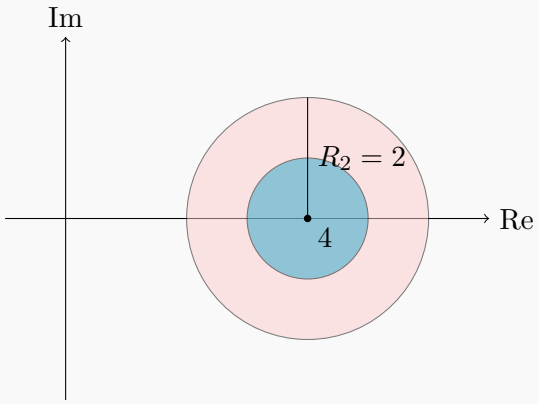
4. Conclusion : Comme $|x_k| > 0$, on peut diviser par $|x_k|$:

$$|\lambda - a_{kk}| \leq R_k.$$

Cela prouve que λ appartient au disque D_k. Par conséquent, λ appartient bien à l'union de tous les disques. □

Soit $A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix}$.

On a $a_{11} = 4$ et $R_1 = |1| + |0| = 1$. De même $R_2 = 2$ et $R_3 = 1$. Les valeurs propres sont donc comprises dans le disque de centre 4 et de rayon 2.



3.2 Stabilité du spectre

Il est crucial de savoir si une petite erreur sur les coefficients de la matrice A (due aux erreurs d'arrondi ou aux incertitudes de mesure) entraîne une petite ou une grande variation des valeurs propres. Dans cette partie, la norme matricielle sera subordonnée à la norme canonique de $\mathbb{R}^n$ notée $|\cdot|$ ($\|x_i\|^2 = \sum_{i=1}^{n} x_i^2$) :

$$\forall A \in \mathcal{M}_n(\mathbb{R}), \|A\| := \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Lemme 3.1 :

Propriétés des normes subordonnées

1. Si $D = \mathrm{diag}(d_1, \ldots, d_n)$ est une matrice diagonale, alors la norme subordonnée est :
$$\|D\| = \max_{1 \leq i \leq n} |d_i|.$$
2. Si U est une matrice unitaire (ou orthogonale réelle), c'est-à-dire $U^*U = I$, alors pour la norme subordonnée euclidienne :
$$\|U\| = 1.$$

Démonstration. 1. Cas de la matrice diagonale.

On a :

— *Majoration :* Pour tout vecteur x, on a $\|Dx\|^2 = \sum_{i=1}^{n} |d_i|^2 \|x_i\|^2 \leq (\max_i |d_i|^2)\|x\|^2$. Par définition de la borne supérieure, on a donc $\|D\| \leq \max_i |d_i|$.

— *Minoration :* Soit k l'indice tel que $|d_k| = \max_i |d_i|$. Considérons le vecteur de la base canonique e_k (qui a 1 en position k et 0 ailleurs).

$$De_k = d_k e_k \implies \|De_k\| = |d_k|\|e_k\|.$$

Le rapport $\dfrac{\|De_k\|}{\|e_k\|}$ vaut exactement $|d_k|$, donc $\sup_{x\neq 0} \dfrac{\|Dx\|}{\|x\|} \geq |d_k|$.

On conclut que $\|D\| = \max_i |d_i|$.

2. Cas de la matrice unitaire (Norme 2).

Soit U telle que $U^*U = I$. Par définition de la norme euclidienne $\|x\|^2 = x^*x$. Pour tout $x \in \mathbb{C}^n$:

$$\|Ux\|^2 = (Ux)^*(Ux) = x^*U^*Ux = x^*Ix = x^*x = \|x\|^2.$$

Ainsi, $\|Ux\|_2 = \|x\|$ pour tout x. L'application linéaire est une isométrie.

$$\|U\|_2 = \sup_{x\neq 0} \frac{\|Ux\|}{\|x\|} = \sup_{x\neq 0} \frac{\|x\|}{\|x\|} = 1.$$

□

On peut maintenant démontrer le Théorème de stabilité spectrale :

Théorème 3.2 : - Théorème de Bauer-Fike

Soit $A \in \mathcal{M}_n(\mathbb{C})$ une matrice diagonalisable, c'est-à-dire qu'il existe une matrice inversible P et une matrice diagonale $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ telles que $A = PDP^{-1}$. Soit $E \in \mathcal{M}_n(\mathbb{C})$ une matrice de perturbation. Pour toute valeur propre μ de la matrice perturbée $A + E$, il existe une valeur propre λ de A telle que :

$$|\mu - \lambda| \leq \kappa(P)\|E\|,$$

où $\kappa(P) := \|P\| \cdot \|P^{-1}\|$ est le conditionnement de P.

Démonstration. Soit μ une valeur propre de $A + E$ et x un vecteur propre associé ($\|x\| = 1$).

$$(A + E)x = \mu x \quad \Leftrightarrow \quad (\mu I - A)x = Ex.$$

Si μ est une valeur propre de A, l'inégalité est trivialement vérifiée (avec la distance nulle). Supposons donc que $\mu \notin \mathrm{Sp}(A)$. La matrice $(\mu I - A)$ est alors inversible. En utilisant la diagonalisation $A = PDP^{-1}$, on a :

$$\mu I - A = P(\mu I - D)P^{-1}.$$

L'équation devient :

$$P(\mu I - D)P^{-1}x = Ex \quad \Leftrightarrow \quad x = P(\mu I - D)^{-1}P^{-1}Ex.$$

Passons à la norme :

$$\|x\| \leq \|P\| \cdot \|(\mu I - D)^{-1}\| \cdot \|P^{-1}\| \cdot \|E\| \cdot \|x\|.$$

Puisque $\|x\| = 1$, on divise par $\|x\|$ et on regroupe $\kappa(P) = \|P\|\|P^{-1}\|$:

$$1 \leq \kappa(P)\|E\|\|(\mu I - D)^{-1}\|.$$

Or, $(\mu I - D)^{-1}$ est une matrice diagonale dont les éléments sont $\dfrac{1}{\mu - \lambda_i}$. La norme d'une matrice diagonale (pour les normes usuelles subordonnées aux normes vectorielles $1, 2, \infty$) est le maximum des modules diagonaux :

$$\|(\mu I - D)^{-1}\| = \max_i \frac{1}{|\mu - \lambda_i|} = \frac{1}{\min_i |\mu - \lambda_i|}.$$

En injectant ceci dans l'inégalité :

$$1 \leq \kappa(P)\|E\| \frac{1}{\min_i |\mu - \lambda_i|} \quad \Leftrightarrow \quad \min_i |\mu - \lambda_i| \leq \kappa(P)\|E\|.$$

Il existe donc bien un λ_i (celui qui réalise le minimum) tel que $|\mu - \lambda_i| \leq \kappa(P)\|E\|$. □

Conséquence importante :

— Si A est symétrique, la matrice de passage P peut être choisie orthogonale ($P^{-1} = P^T$). Dans ce cas, $\|P\|_2 = 1$, donc $\kappa_2(P) = 1$.

— Le problème des valeurs propres des matrices symétriques est donc **très bien conditionné** : une perturbation de taille ε sur la matrice entraîne une perturbation d'au plus ε sur les valeurs propres.

— À l'inverse, si les vecteurs propres sont presque colinéaires (matrice non symétrique proche d'une matrice non diagonalisable), $\kappa(P)$ est grand et les valeurs propres deviennent très instables.

3.3 La Méthode de la Puissance

Cette méthode permet de déterminer la valeur propre de plus grand module (dite dominante) et un vecteur propre associé.

Définition 3.2

Soit A une matrice et $x_0 \in \mathbb{R}^n$ tel que $\|x_0\| = 1$. La méthode de la puissance construit une suite de vecteurs (x_k) et de scalaires (μ_k) définie par :

$$\begin{cases} x_{k+1} &= Ax_k \\ y_k &= \dfrac{x_k}{\|x_k\|} \\ \mu_k &= \langle y_k, Ay_k \rangle = y_k^T A y_k \quad \text{(Quotient de Rayleigh)} \\ z_k = & \dfrac{\overline{\lambda_1}^k}{|\lambda_1|^k} y_k. \end{cases}$$

Théorème 3.3 : - Convergence de la méthode de la puissance

Supposons que A soit diagonalisable et admette une valeur propre dominante unique λ_1, c'est-à-dire :

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Si le vecteur initial x_0 n'est pas orthogonal au sous-espace propre associé à λ_1, alors :

$$\lim_{k \to +\infty} \mu_k = \lambda_1 \qquad \text{et} \qquad \lim_{k \to +\infty} \|z_k - (\pm v_1)\| = 0,$$

où v_1 est un vecteur propre unitaire associé à λ_1.

Démonstration. **1. Décomposition spectrale** : Comme A est diagonalisable, il existe une base de vecteurs propres $(u_1, \ldots, u_n)$ associés aux valeurs propres $(\lambda_1, \ldots, \lambda_n)$. Le vecteur initial x_0 se décompose dans cette base :

$$x_0 = \sum_{i=1}^{n} \alpha_i u_i = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n.$$

Par hypothèse, x_0 n'est pas orthogonal à l'espace propre dominant, ce qui se traduit ici par le coefficient $\alpha_1 \neq 0$.

2. Itération de la puissance : Calculons $A^k x_0$. Par linéarité :

$$A^k x_0 = \sum_{i=1}^{n} \alpha_i \lambda_i^k u_i = \alpha_1 \lambda_1^k u_1 + \sum_{i=2}^{n} \alpha_i \lambda_i^k u_i.$$

3. Factorisation par le terme dominant : On met en facteur le terme λ_1^k :

$$A^k x_0 = \lambda_1^k \left(\alpha_1 u_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k u_i \right).$$

Posons le vecteur reste $\varepsilon_k = \sum_{i=2}^{n} \alpha_i (\frac{\lambda_i}{\lambda_1})^k u_i$. Comme $|\lambda_1| > |\lambda_i|$ pour tout $i \geq 2$, on a $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$, et donc $\lim_{k \to +\infty} \|\varepsilon_k\| = 0$.

Ainsi, pour k grand, le vecteur $A^k x_0$ s'aligne sur la direction de u_1 :

$$A^k x_0 = \lambda_1^k \alpha_1 u_1 + o(1).$$

4. Convergence du vecteur normalisé : Le vecteur y_k est défini par $y_k = \frac{x_k}{\|x_k\|}$. En utilisant l'estimée précédente :

$$y_k = \frac{\lambda_1^k \alpha_1 u_1 + o(1)}{|\lambda_1^k \alpha_1| \|u_1\| + o(1)}.$$

Ainsi $z_k = \frac{\overline{\lambda_1}^k}{|\lambda_1|^k} y_k = \operatorname{sign}(\alpha_1) \frac{u_1}{\|u_1\|} + o(1)$.

5. Convergence du quotient de Rayleigh : Par continuité du produit scalaire :

$$\mu_k = \langle y_k, A y_k \rangle = \langle \frac{\lambda_1^k \alpha_1 u_1 + o(1)}{|\lambda_1^k \alpha_1|}, \frac{\lambda_1^{k+1} \alpha_1 u_1 + o(1)}{|\lambda_1^k \alpha_1|} \rangle = \lambda_1 + o(1)$$

□

Vitesse de convergence : La convergence est dite linéaire. L'erreur à l'étape k est proportionnelle à $\left|\frac{\lambda_2}{\lambda_1}\right|^k$. Plus l'écart spectral est grand, plus la méthode est rapide.

3.4 Calcul des autres valeurs propres : La Déflation

La méthode de la puissance ne fournit que le couple (λ_1, v_1) correspondant à la valeur propre dominante. Pour obtenir les valeurs propres suivantes, on utilise la technique de la déflation. **Principe** : L'idée est de construire une nouvelle matrice A_1 dont le spectre est modifié de sorte que λ_1 soit remplacée par 0, tandis que les autres valeurs propres $\lambda_2, \ldots, \lambda_n$ restent inchangées. Ainsi, la valeur propre dominante de B deviendra λ_2 (supposée distincte), et on pourra lui appliquer de nouveau la méthode de la puissance.

3.4.1 Déflation de Wielandt (Cas Général)

Proposition 3.1 :

Soit (λ_1, v_1) un couple propre de $A \in \mathbb{R}^{n\times n}$. Soit $u \in \mathbb{R}^n$ un vecteur arbitraire tel que $u^T v_1 = 1$. Considérons la matrice déflatée :

$$A_1 = A - \lambda_1 v_1 u^T.$$

Alors le spectre de A_1 est exactement $\{0, \lambda_2, \ldots, \lambda_n\}$, où $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A comptées avec leur multiplicité.

Démonstration. **Il s'agit ici d'une démonstration pédagogique simple mais incomplète démontrant que $\{0, \lambda_2, \cdots \lambda_n\} \subset \mathbf{Sp}(A_1)$ dans le cas où $\lambda_i \neq 0$ pour $2 \leq j \leq n$.**

1. La valeur propre dominante devient nulle : Calculons l'image de v_1 par A_1 :

$$A_1 v_1 = (A - \lambda_1 v_1 u^T)v_1 = Av_1 - \lambda_1 v_1 \underbrace{(u^T v_1)}_{1} = \lambda_1 v_1 - \lambda_1 v_1 = 0.$$

Donc 0 est valeur propre de A_1 (associée au vecteur v_1).

2. Conservation des autres valeurs propres : Soit $i \in \{2, \ldots, n\}$. On cherche un vecteur propre w_i de A_1 associé à la valeur propre λ_i (supposée non nulle).

Cherchons w_i sous la forme $w_i = v_i - \alpha v_1$, où α est une constante à déterminer. Calculons $A_1 w_i$:

$$A_1(v_i - \alpha v_1) = A_1 v_i - \alpha \underbrace{A_1 v_1}_{0} = A_1 v_i.$$

Développons $A_1 v_i$:

$$A_1 v_i = (A - \lambda_1 v_1 u^T)v_i = \underbrace{Av_i}_{\lambda_i v_i} - \lambda_1 v_1 (u^T v_i) = \lambda_i v_i - \lambda_1 (u^T v_i) v_1.$$

On souhaite que $A_1 w_i = \lambda_i w_i$. Identifions les termes :

$$\lambda_i v_i - \lambda_1 (u^T v_i) v_1 = \lambda_i (v_i - \alpha v_1) = \lambda_i v_i - \lambda_i \alpha v_1.$$

En simplifiant par $\lambda_i v_i$, il reste :

$$-\lambda_1 (u^T v_i) v_1 = -\lambda_i \alpha v_1.$$

Cette égalité est vérifiée si l'on choisit le scalaire $\alpha = \dfrac{\lambda_1}{\lambda_i}(u^T v_i)$.

Conclusion : Pour tout $i \geq 2$, le vecteur $w_i = v_i - \dfrac{\lambda_1 (u^T v_i)}{\lambda_i} v_1$ est un vecteur propre de A_1 pour la valeur propre λ_i. Le spectre est bien conservé. □

Ici on donne la démonstration complète et rigoureuse :

Démonstration. La preuve repose sur la relation entre les polynômes caractéristiques de A et A_1. Soit $P_A(\lambda) = \det(A - \lambda I)$. Calculons $P_{A_1}(\lambda) = \det(A_1 - \lambda I)$.

En remplaçant A_1 par son expression :

$$P_{A_1}(\lambda) = \det(A - \lambda_1 v_1 u^T - \lambda I) = \det((A - \lambda I) - \lambda_1 v_1 u^T).$$

Utilisons le lemme du déterminant (formule de mise à jour de rang 1) : $\det(M + xy^T) = \det(M) + y^T \text{adj}(M)x$. Ici, posons $M = A - \lambda I$, $x = -\lambda_1 v_1$ et $y = u$. On sait que pour tout λ n'appartenant pas au spectre de A, M est inversible et $\text{adj}(M) = \det(M)M^{-1}$. Ainsi :

$$\begin{aligned} P_{A_1}(\lambda) &= \det(A - \lambda I) + u^T \left[P_A(\lambda)(A - \lambda I)^{-1} \right] (-\lambda_1 v_1) \\ &= P_A(\lambda) \left[1 - \lambda_1 u^T (A - \lambda I)^{-1} v_1 \right]. \end{aligned}$$

Puisque v_1 est vecteur propre de A pour λ_1, il est vecteur propre de $(A - \lambda I)^{-1}$ pour la valeur propre $\dfrac{1}{\lambda_1 - \lambda}$. D'où $(A - \lambda I)^{-1} v_1 = \dfrac{1}{\lambda_1 - \lambda} v_1$.

L'expression devient :

$$P_{A_1}(\lambda) = P_A(\lambda) \left[1 - \frac{\lambda_1}{\lambda_1 - \lambda} (u^T v_1) \right].$$

Comme $u^T v_1 = 1$, on simplifie le crochet :

$$1 - \frac{\lambda_1}{\lambda_1 - \lambda} = \frac{(\lambda_1 - \lambda) - \lambda_1}{\lambda_1 - \lambda} = \frac{-\lambda}{\lambda_1 - \lambda}.$$

On obtient la relation finale (valable pour tout λ par continuité polynomiale) :

$$P_{A_1}(\lambda) = P_A(\lambda) \cdot \frac{-\lambda}{\lambda_1 - \lambda}.$$

Conclusion : Le facteur $(\lambda_1 - \lambda)$ au dénominateur simplifie la racine λ_1 présente dans $P_A(\lambda)$, et le facteur $-\lambda$ introduit une racine 0. Les autres racines de $P_A(\lambda)$ restent inchangées. Le spectre est donc bien modifié de $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ en $\{0, \lambda_2, \ldots, \lambda_n\}$. □

3.4.2 Déflation de Hotelling (Cas Symétrique)

Dans le cas où A est symétrique réelle, on sait que les vecteurs propres associés à des valeurs propres distinctes sont orthogonaux. C'est le cadre idéal pour la déflation.

On choisit $u = v_1$ (car $\|v_1\| = 1 \implies \langle v_1, v_1 \rangle = 1$). La formule devient :

$$A_1 = A - \lambda_1 v_1 v_1^T$$

Proposition 3.2 : A14

Si A est symétrique et $(v_1, \ldots, v_n)$ est une base orthonormée de vecteurs propres associés à $(\lambda_1, \ldots, \lambda_n)$, alors :

1. $A_1 v_1 = 0$;
2. Pour tout $j \neq 1$, $A_1 v_j = \lambda_j v_j$.

Démonstration. **Point 1** : Déjà démontré plus haut ($A_1 v_1 = 0$).

Point 2 : Soit v_j un autre vecteur propre ($j \neq 1$). Comme A est symétrique, v_j est orthogonal à v_1, donc $\langle v_1, v_j \rangle = v_1^T v_j = 0$.

$$A_1 v_j = (A - \lambda_1 v_1 v_1^T) v_j = \underbrace{A v_j}_{\lambda_j v_j} - \lambda_1 v_1 \underbrace{(v_1^T v_j)}_{0} = \lambda_j v_j.$$

Conclusion : Les vecteurs propres $v_2, \ldots, v_n$ de A sont **aussi** des vecteurs propres de A_1, avec les mêmes valeurs propres. Comme la valeur propre λ_1 a été remplacée par 0, la valeur propre de plus grand module de A_1 est maintenant λ_2. □

3.4.3 Algorithme itératif

Si $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ et A symétrique, on peut donc itérer la méthode de la puissance pour trouver successivement les valeurs propres.

- **Initialisation** : Poser $A_0 = A$.
- **Itération** : Pour $k = 1$ à $n - 1$:
  1. Appliquer la **méthode de la puissance** sur la matrice A_k.
  2. On obtient une approximation de la valeur propre dominante λ_k et de son vecteur propre v_k.
  3. On construit la matrice suivante par déflation :
  $$A_{k+1} = A_k - \lambda_k v_k v_k^T$$
  4. Grâce à la proposition précédente, la valeur propre dominante de A_k sera λ_{k+1}. Elle est de plus symétrique donc diagonalisable avec chacune des valeurs propres non nulles de module distinct. On pourra donc appliquer la méthode de la puissance à l'étape suivante.

Soit A symétrique de spectre $\{10, 5, 2\}$.
- Puissance sur $A_1 = A \implies$ converge vers $\lambda = 10, v_1$.
- On calcule $A_2 = A - 10 v_1 v_1^T$. Son spectre est $\{0, 5, 2\}$.
- Puissance sur $A_2 \implies$ converge vers $\lambda = 5, v_2$ (car 5 est le max de $\{0, 5, 2\}$).
- On calcule $A_3 = A_2 - 5 v_2 v_2^T$. Son spectre est $\{0, 0, 2\}$.
- Puissance sur $A_3 \implies$ converge vers $\lambda = 2, v_3$.

Remarque (Stabilité) : En pratique, comme λ_k et v_k sont approchés, la matrice A_{k+1} contient une petite erreur. Cette erreur se propage et s'amplifie à chaque étape. Cette méthode n'est précise que pour les premières valeurs propres.

Peut-on appliquer la méthode de Hotelling lorsque A n'est pas symétrique ?

Dans le cas symétrique, la matrice déflatée A_1 reste symétrique, donc elle est toujours diagonalisable. Dans le cas non symétrique, la symétrie est perdue. Cependant, si les valeurs propres sont distinctes, la structure diagonalisable est conservée à chaque étape.

Proposition 3.3 :

Supposons que la matrice A possède n valeurs propres de module **distincts non nuls** :

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

Alors les matrices déflatées par la méthode de Hotelling A_k sont **diagonalisables** de valeurs propres $(0, \lambda_{k+1}, \cdots, \lambda_n)$ et $\dim(\mathrm{Ker}(A_k)) = k$.

Démonstration. Nous avons établi dans la **Proposition 3.1** (cas général de la déflation) que le spectre de A_1 est exactement :

$$\mathrm{Sp}(A_1) = \{0, \lambda_2, \lambda_3, \ldots, \lambda_n\}.$$

Montrons que A_1 admet une base de vecteurs propres (c'est-à-dire qu'elle est diagonalisable) :

1. **Les valeurs propres non nulles :** Les scalaires $\lambda_2, \ldots, \lambda_n$ sont distincts deux à deux et non nuls (par hypothèse sur A). D'après la théorie spectrale, à des valeurs propres distinctes correspondent des vecteurs propres linéairement indépendants. Il existe donc une famille libre de $n-1$ vecteurs propres $\{w_2, \ldots, w_n\}$ associés respectivement à $\lambda_2, \ldots, \lambda_n$.
2. **La valeur propre nulle :** Nous savons que $A_1 v_1 = 0$. Comme 0 est une valeur propre distincte de toutes les autres ($\lambda_i \neq 0$), le vecteur propre v_1 est linéairement indépendant de la famille $\{w_2, \ldots, w_n\}$.

Conclusion : La famille $(v_1, w_2, \ldots, w_n)$ constitue une famille libre de n vecteurs dans un espace de dimension n. C'est donc une base de vecteurs propres. La matrice A_1 est bien diagonalisable. □

Justification pour les itérations suivantes : Ce raisonnement s'applique par récurrence. À l'étape k, la matrice A_k possède le spectre $\{0, \ldots, 0, \lambda_{k+1}, \ldots, \lambda_n\}$. Les valeurs propres non nulles restant distinctes, elles fournissent $n-k$ vecteurs propres indépendants. La déflation étant construite pour annuler la contribution des directions précédentes, on peut montrer que la dimension du noyau (l'espace propre associé à 0) augmente exactement de 1 à chaque étape. La somme des dimensions des sous-espaces propres reste égale à n, garantissant la diagonalisabilité nécessaire à la convergence de la méthode de la puissance pour l'étape suivante.

3.5 Pour aller plus loin : L'algorithme QR

Pour calculer simultanément toutes les valeurs propres de manière stable, on utilise l'algorithme QR (aussi appelé méthode de Francis). Avant de décrire l'algorithme, définissons l'outil central : la décomposition QR.

3.5.1 La décomposition QR

Théorème 3.4 :

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Il existe un unique couple de matrices (Q, R) tel que :

$$A = QR$$

avec :

- Q une matrice orthogonale ($Q^T Q = I_n$) ;
- R une matrice triangulaire supérieure à coefficients diagonaux strictement positifs.

Démonstration. L'existence découle directement de l'orthogonalisation de Gram-Schmidt appliquée aux colonnes de A.

Si on note $a_1, \ldots, a_n$ les colonnes de A, l'algorithme de Gram-Schmidt produit une base orthonormée $q_1, \ldots, q_n$ telle que pour tout k, $\text{Vect}(a_1, \ldots, a_k) = \text{Vect}(q_1, \ldots, q_k)$.

On peut alors écrire chaque a_j comme combinaison linéaire des q_i (pour $i \leq j$) :

$$a_j = \sum_{i=1}^{j} r_{ij} q_i.$$

En écriture matricielle, cela donne exactement $A = QR$, où R est triangulaire supérieure ($r_{ij} = 0$ si $i > j$). □

3.5.2 L'algorithme de Francis (QR)

Définition 3.3

On génère une suite de matrices $(A_k)_{k\in\mathbb{N}}$ définie par $A_0 = A$ et pour tout $k \geq 0$:

$$\begin{cases} A_k = Q_k R_k & \text{(1. Décomposition QR de } A_k\text{)} \\ A_{k+1} = R_k Q_k & \text{(2. Recombinaison inversée)} \end{cases}$$

Proposition 3.4 :

Toutes les matrices A_k sont semblables à la matrice originale A. En particulier, elles ont toutes les mêmes valeurs propres.

Démonstration. D'après l'étape de décomposition, on a $A_k = Q_k R_k$. Comme Q_k est orthogonale, elle est inversible et $Q_k^{-1} = Q_k^T$. On peut donc écrire :

$$R_k = Q_k^T A_k.$$

Injectons ceci dans l'étape de recombinaison $A_{k+1} = R_k Q_k$:

$$A_{k+1} = (Q_k^T A_k) Q_k = Q_k^{-1} A_k Q_k.$$

La matrice A_{k+1} est donc semblable à A_k. Par récurrence, A_k est semblable à $A_0 = A$. □

3.5.3 Convergence

La clé de la preuve réside dans le fait que l'algorithme QR effectue implicitement une orthogonalisation de Gram-Schmidt sur les puissances de la matrice A. Définissons les matrices cumulées :

$$\underline{Q}_k = Q_0 Q_1 \dots Q_{k-1} \quad \text{et} \quad \underline{R}_k = R_{k-1} \dots R_1 R_0.$$

Lemme 3.2 :

Pour tout $k \geq 1$, on a les deux relations fondamentales :

1. **Similitude** : $A_k = \underline{Q}_k^T A \underline{Q}_k$.
2. **QR des puissances** : $A^k = \underline{Q}_k \underline{R}_k$.

Démonstration. **1. Similitude** : Par définition, $A_k = R_{k-1} Q_{k-1}$. Or $A_{k-1} = Q_{k-1} R_{k-1}$, donc $R_{k-1} = Q_{k-1}^T A_{k-1}$. En remplaçant : $A_k = Q_{k-1}^T A_{k-1} Q_{k-1}$. Par récurrence immédiate :

$$A_k = Q_{k-1}^T (Q_{k-2}^T \dots A_0 \dots Q_{k-2}) Q_{k-1} = \underline{Q}_k^T A \underline{Q}_k.$$

2. Puissances : Montrons par récurrence que $A^k = \underline{Q}_k \underline{R}_k$.

— Pour $k = 1$: $A^1 = A_0 = Q_0 R_0 = \underline{Q}_1 \underline{R}_1$. C'est vrai.

— Supposons $A^k = \underline{Q}_k \underline{R}_k$. Alors :

$$A^{k+1} = A \cdot A^k = A \underline{Q}_k \underline{R}_k.$$

D'après la relation de similitude (point 1), on a $A = \underline{Q}_k A_k \underline{Q}_k^T$, d'où $A\underline{Q}_k = \underline{Q}_k A_k$. Ainsi :

$$A^{k+1} = \underline{Q}_k A_k \underline{R}_k.$$

On utilise la définition de l'étape QR : $A_k = Q_k R_k$.

$$A^{k+1} = \underline{Q}_k (Q_k R_k) \underline{R}_k = (\underline{Q}_k Q_k)(R_k \underline{R}_k) = \underline{Q}_{k+1} \underline{R}_{k+1}.$$

□

Ce lemme prouve que la matrice $\underline{Q}_k$ n'est rien d'autre que la matrice unitaire issue de la décomposition QR de la puissance A^k.

3.5.4 Convergence (Théorème et Preuve détaillée)

Théorème 3.5 : - Convergence de l'algorithme QR

Soit A une matrice diagonalisable dont les valeurs propres satisfont :

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

Alors la suite (A_k) converge vers une matrice triangulaire supérieure dont la diagonale contient les λ_i.

Démonstration. Définissons la matrice orthogonale cumulée $\underline{Q}_k = Q_1 Q_2 \dots Q_k$ et la matrice triangulaire cumulée $\underline{R}_k = R_k R_{k-1} \dots R_1$.

Montrons par récurrence que $A^k = \underline{Q}_k \underline{R}_k$.

— **Initialisation ($k = 1$) :** $A = Q_1 R_1 = \underline{Q}_1 \underline{R}_1$. C'est la définition.

— **Hérédité :** Supposons $A^{k-1} = \underline{Q}_{k-1} \underline{R}_{k-1}$. On sait que $A_{k-1} = \underline{Q}_{k-1}^T A \underline{Q}_{k-1}$. De plus, par définition de l'étape k, $A_{k-1} = Q_k R_k$. Donc :

$$Q_k R_k = \underline{Q}_{k-1}^T A \underline{Q}_{k-1} \iff \underline{Q}_{k-1} Q_k R_k = A \underline{Q}_{k-1}.$$

Multiplions à droite par $\underline{R}_{k-1}$:

$$\underline{Q}_{k-1} Q_k R_k \underline{R}_{k-1} = A(\underline{Q}_{k-1} \underline{R}_{k-1}).$$

En utilisant l'hypothèse de récurrence ($A^{k-1} = \underline{Q}_{k-1} \underline{R}_{k-1}$) et les définitions cumulées :

$$\underline{Q}_k \underline{R}_k = A(A^{k-1}) = A^k.$$

Conclusion : $\underline{Q}_k \underline{R}_k$ est la décomposition QR unique (si on impose les coefficients diagonaux de R positifs) de la matrice A^k.

Étape 2 : Convergence des sous-espaces (Itération de sous-espaces)

Soit $\mathcal{E}_j = \text{vect}(e_1, \dots, e_j)$ le sous-espace engendré par les j premiers vecteurs de la base canonique. Soit $\mathcal{V}_j = \text{vect}(u_1, \dots, u_j)$ le sous-espace invariant dominant de dimension j (engendré par les j premiers vecteurs propres).

L'égalité $A^k = \underline{Q}_k \underline{R}_k$ implique une relation cruciale sur les images des sous-espaces. Comme $\underline{R}_k$ est triangulaire supérieure, elle conserve les drapeaux (elle ne mélange les vecteurs $e_1 \dots e_j$ qu'entre eux). Ainsi, l'espace engendré par les j premières colonnes de A^k est identique à l'espace engendré par les j premières colonnes de $\underline{Q}_k$:

$$\text{Im}(A^k|_{\mathcal{E}_j}) = \text{vect}(A^k e_1, \dots, A^k e_j) = \text{vect}(q_1^{(k)}, \dots, q_j^{(k)}).$$

D'après la théorie de la puissance itérée appliquée aux sous-espaces, comme $|\lambda_j| > |\lambda_{j+1}|$, l'espace image $A^k \mathcal{E}_j$ converge vers l'espace invariant dominant $\mathcal{V}_j$ lorsque $k \to \infty$ (sous réserve que la projection de $\mathcal{E}_j$ sur $\mathcal{V}_j$ soit non singulière, condition générique).

Nous avons donc établi que :

$$\text{vect}(q_1^{(k)}, \dots, q_j^{(k)}) \xrightarrow{k \to \infty} \text{vect}(u_1, \dots, u_j).$$

Étape 3 : Analyse des coefficients de A_k La matrice à l'étape k est donnée par $A_k = \underline{Q}_k^T A \underline{Q}_k$. Le coefficient (i,j) de A_k est le produit scalaire :

$$(A_k)_{ij} = \langle q_i^{(k)}, Aq_j^{(k)} \rangle.$$

Analysons le cas sous-diagonal, c'est-à-dire $i > j$.

1. Le vecteur $q_j^{(k)}$ appartient (à la limite) à l'espace $\mathcal{V}_j = \text{vect}(u_1, \ldots, u_j)$.
2. L'espace $\mathcal{V}_j$ est stable par A (car engendré par des vecteurs propres). Donc $Aq_j^{(k)}$ appartient aussi, asymptotiquement, à $\mathcal{V}_j$.
3. Par construction de la décomposition QR, le vecteur $q_i^{(k)}$ est orthogonal à tous les vecteurs précédents $q_1^{(k)}, \ldots, q_{i-1}^{(k)}$.
4. Puisque $i > j$, l'espace $\mathcal{V}_j$ est inclus (à la limite) dans l'espace $\text{vect}(q_1^{(k)}, \ldots, q_{i-1}^{(k)})$.

Conséquence : $q_i^{(k)}$ devient orthogonal à l'espace contenant $Aq_j^{(k)}$.

$$\lim_{k \to \infty} (A_k)_{ij} = 0 \quad \text{pour tout } i > j.$$

Conclusion Puisque tous les termes sous-diagonaux tendent vers 0, la matrice A_k converge vers une matrice triangulaire supérieure T.

Comme A_k est semblable à A (donc même spectre) et que les valeurs propres d'une matrice triangulaire sont ses coefficients diagonaux, les termes diagonaux $(A_k)_{ii}$ convergent vers les valeurs propres λ_i (dans l'ordre décroissant des modules à cause de l'ordre imposé par la convergence des sous-espaces $\mathcal{E}_j$ vers $\mathcal{V}_j$). □

3.5.5 Accélération de l'algorithme : Formes de Hessenberg et Tridiagonales

Bien que l'algorithme QR soit l'outil de choix pour calculer simultanément toutes les valeurs propres de manière stable, le calcul de la décomposition QR d'une matrice dense A_k à chaque itération coûte $\mathcal{O}(n^3)$ opérations. Pour optimiser ce processus, on transforme préalablement la matrice pour lui donner une structure comportant un maximum de zéros.

Définition 3.4

Une matrice $H \in \mathcal{M}_n(\mathbb{R})$ est dite sous forme de Hessenberg supérieure si tous ses coefficients situés strictement sous la première sous-diagonale sont nuls :

$$\forall i > j+1, \qquad h_{ij} = 0.$$

L'intérêt majeur de cette forme réside dans sa conservation tout au long de l'algorithme de Francis générant la suite $A_{k+1} = R_k Q_k$.

Proposition 3.5 :

Si la matrice A_k est inversible et sous forme de Hessenberg supérieure, alors A_{k+1} l'est également.

Démonstration. Par définition de l'étape de décomposition, nous avons $A_k = Q_k R_k$. Puisque A_k est inversible, la matrice triangulaire supérieure R_k l'est aussi.

Posons $U = R_k^{-1}$. Cette matrice est également triangulaire supérieure. *(En effet, la j-ème colonne de U est solution du système $R_k x = e_j$; par résolution par remontée sur une matrice triangulaire supérieure, les composantes x_i sont nécessairement nulles pour tout $i > j$).*

1. Structure de Q_k : On écrit $Q_k = A_k R_k^{-1} = A_k U$. Le coefficient générique (i, j) de Q_k est donné par le produit matriciel :

$$(Q_k)_{ij} = \sum_{p=1}^{n} (A_k)_{ip} U_{pj}.$$

Supposons que $i > j + 1$ (position sous la première sous-diagonale).

— Comme U est triangulaire supérieure, $U_{pj} \neq 0 \implies p \leq j$.

— La somme se restreint donc aux indices $p \in [\![1, j]\!]$.

— Or, pour tout $p \leq j$, on a $p < i - 1$ (car $j < i - 1$). Donc $i > p + 1$.

— Comme A_k est de Hessenberg supérieure, $(A_k)_{ip} = 0$ pour tout $i > p + 1$.

Tous les termes de la somme sont nuls. Donc $(Q_k)_{ij} = 0$ pour $i > j + 1$. Q_k est bien une matrice de Hessenberg supérieure.

2. Structure de A_{k+1} : L'étape de recombinaison s'écrit $A_{k+1} = R_k Q_k$. Le coefficient (i, j) est :

$$(A_{k+1})_{ij} = \sum_{p=1}^{n} (R_k)_{ip} (Q_k)_{pj}.$$

Supposons encore $i > j + 1$.

— Comme R_k est triangulaire supérieure, $(R_k)_{ip} \neq 0 \implies p \geq i$.

— La somme se restreint donc aux indices $p \in [\![i, n]\!]$.

— Or, pour tout $p \geq i$, on a $p > j + 1$ (car $i > j + 1$).

— Comme Q_k vient d'être montrée Hessenberg, $(Q_k)_{pj} = 0$ pour tout $p > j + 1$.

La somme est nulle. $(A_{k+1})_{ij} = 0$ pour $i > j+1$. Ainsi, A_{k+1} conserve bien la forme de Hessenberg supérieure. □

Conséquences algorithmiques :

- **Gain de complexité (Cas général)** : Grâce à la structure de Hessenberg, la décomposition QR peut être réalisée avec des rotations de Givens (voir exercice). Le coût d'une itération chute de $\mathcal{O}(n^3)$ à $\mathcal{O}(n^2)$ opérations.
- **Matrice Tridiagonale (Cas symétrique)** : Si la matrice initiale A est **symétrique**, toutes les matrices A_k le restent, puisqu'elles sont de la forme $Q_k^T A_k Q_k$. Or, une matrice qui est à la fois symétrique et de Hessenberg supérieure est nécessairement tridiagonale (seules la diagonale principale et ses deux diagonales adjacentes sont non nulles). Dans ce cas optimal, le coût de l'itération QR s'effondre pour atteindre seulement $\mathcal{O}(n)$ opérations.