MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU
FACULTE DES SCIENCES
DEPARTEMENT DE MATHEMATIQUES

Polycopié de cours

Filière : MATHEMATIQUES

Deuxième année licence Mathématiques

Rédigé par ATIL Lynda

Analyse Numérique 2

(Cours et exercices)

Préface

Ce polycopié est destiné aux étudiants de deuxième année de Licence en Mathématiques, dans le cadre du module d'Analyse Numérique 2. Il a été conçu pour offrir une introduction claire et progressive aux concepts fondamentaux du calcul numérique, en mettant l'accent sur les méthodes essentielles utilisées pour résoudre des problèmes mathématiques de manière approchée.

L'analyse numérique joue un rôle central en mathématiques appliquées et dans de nombreux domaines scientifiques et techniques. Elle permet de traiter des problèmes complexes qui ne peuvent pas être résolus analytiquement, en proposant des algorithmes efficaces et des estimations contrôlées des erreurs. Ce cours vise à fournir aux étudiants les outils nécessaires pour comprendre, implémenter et critiquer ces méthodes.

Le présent document est structuré en quatre chapitres :
1. Rappels et compléments sur les matrices: Ce chapitre a pour but de rappeler, et de démontrer, un certain nombre de résultats relatifs aux matrices et aux espaces vectoriels de dimension finie, et dont un usage constant sera fait dans toute la suite du polycopié.
2. Résolution numérique des systèmes d'équations linéaires: Nous y étudions les méthodes directes (méthodes de Gauss, Cholesky) qui permettent d'obtenir la solution en un nombre fini d'opérations et les méthodes itératives (méthodes de Jacobi, Gauss-Seidel) qui recherchent la solution de proche en proche en partant d'un vecteur initial arbitraire pour résoudre un système linéaire.
3. Calcul des valeurs et des vecteurs propres: Ce chapitre aborde les méthodes de calcul des approximations de l'ensemble des valeurs propres d'une matrice A et des vecteurs propres associés.
4. Résolution numérique des équations différentielles : Nous abordons différentes méthodes numériques pour la résolutions approchées des équations différentielles ordinaires.

Chaque chapitre est illustré d'exemples et d'exercices pour faciliter l'assimilation des concepts. Ce support pédagogique a pour but de rendre accessible les techniques numériques tout en encourageant une réflexion sur leur précision et leurs limites.

Nous espérons que ce polycopié accompagnera efficacement les étudiants dans leur apprentissage de l'analyse numérique et leur donnera les bases nécessaires pour des études plus avancées en calcul scientifique.

Table des matières

Chapitre I: Rappels et Compléments sur les matrices.

Chapitre II: Résolution numérique des systèmes d'équations linéaires.

Chapitre III: Calcul des valeurs et des vecteurs propres.

Chapitre IV: Résolution numérique des équations différentielles.

CHAPITRE I. Rappels et compléments sur les matrices

1 Introduction.

Ce chapitre a pour but de rappeler, un certain nombre de résultats relatifs aux matrices et aux espaces vectoriels de dimension finie, et dont un usage constant sera fait dans toute la suite du polycopié.
On suppose les lecteurs déjà familiarisés avec les propriétés élémentaires des espaces vectoriels de dimension finie (i.e calcul matriciel notamment), pour lesquelles on revoit au paragraphe l, les principales notations et définitions, ainsi que la notion de décomposition par blocs d'une matrice, qui est à signaler pour son importance en Analyse Numérique Matricielle.
Afin de rendre l'ouvrage aussi "autonome" que possible, tous les résultats importants pour la suite sont énoncés, en particulier la réduction d'une matrice quelconque à la forme triangulaire, la diagonalisation des matrices normales, et l'équivalence d'une matrice à la matrice diagonale de ses' valeurs' singulières. Nous examinons ensuite les caractérisations des valeurs propres des matrices symétriques ou hermitiennes par l'intermédiaire du quotient de Rayleigh, notamment les caractérisations par "min-max" et par "max-min".
On passe ensuite en revue les normes vectorielles les plus couramment utilisées en Analyse Numérique Matricielle, qui sont des cas particuliers des "normes l_p", puis on calcule les normes matricielles subordonnées correspondantes. On rappelle également les conditions d'inversibilité de matrices de la forme $(I + B)$, et on montre que le rayon spectral d'une matrice est la borne inférieure des valeurs de ses normes, ce dernier résultat servant ensuite à démontrer deux résultats relatifs à la suite des puissances successives d'une même matrice, qui jouent un rôle fondamental dans l'étude des méthodes itératives de résolution de systèmes linéaires étudiées dans le polycopié.

2 Principales notations et définitions

Soit V un espace vectoriel de dimension finie n, sur le corps $\mathbb{R}$ des nombres réels, ou sur le corps $\mathbb{C}$ des nombres complexes. S'il n'y a pas lieu dc distinguer, on dit qu'il s'agit du corps $\mathbb{K}$ des scalaires.
Une base de V est un ensemble $\{e_1, e_2, \ldots, e_n\}$ de n vecteurs linéairement indépendants de V, qu'on notera $(e_i)_{i=1}^n$ ou simplement (e_i) si aucune confusion n'est à craindre. Tout vecteur $v \in V$ admet alors une décomposition unique

$$v = \sum_{i=1}^{n} v_i e_i$$

les scalaires v_i, que nous noterons parfois (v_i) étant les composantes du vecteur v sur la base (e_i). Lorsqu'une base est fixée sans ambiguité, on peut ainsi identifier V à $\mathbb{K}^n$, c'est

pourquoi il nous arrivera également de noter $v = (v_i)_{i=1}^n$ ou simplement (v_i), un vecteur de composante (v_i).
En notation matricielle, le vecteur $v = \sum_{i=1}^n v_i e_i$ sera toujours représenté par le vecteur colonne

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

et on désignera par v^T et v^*. les vecteurs lignes suivants: $v^T = (v_1, v_2, \ldots, v_n)$,
$v^* = (\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_n)$, où, en général, $\bar{\alpha}$ désigne le nombre complexe conjugué du nombre α. Le vecteur ligne v^T est le vecteur transposé du vecteur colonne v, et le vecteur ligne v^* est le vecteur adjoint du vecteur colonne v.
L'application $(.\ ,.)$: $V \times V \longrightarrow \mathbb{K}$ définie par
$(u, v) = v^T u = u^T v = \sum_{i=1}^n u_i v_i$ si $\mathbb{K} = \mathbb{R}$,
$(u, v) = v^* u = \overline{u^* v} = \sum_{i=1}^n u_i \bar{v}_i$ si $\mathbb{K} = \mathbb{C}$,
est appelée produit scalaire euclidien si $\mathbb{K} = \mathbb{R}$, hermitien si $\mathbb{K} = \mathbb{C}$, ou canonique si l'on ne précise pas le corps des scalaires. Si l'on souhaite rappeler la dimension de l'espace, on écrira

$$(u, v) = (u, v)_n.$$

Soit V un espace muni de son produit scalaire canonique. Deux vecteurs u et v de V sont orthogonaux si $(u, v) = 0$. Par extension, on dit qu'un vecteur v est orthogonal à une partie U de V, et on note $v \perp U$, lorsque le vecteur v est orthogonal à tous les vecteurs de U. Enfin, un ensemble $\{v_1, \ldots, v_k\}$ de vecteurs de l'espace V est dit orthonormal si

$$(v_i, v_j) = \delta_{ij}, \quad 1 \leq i, j \leq k$$

où δ_{ij} est le symbole de Kronecker; $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ si $i \neq j$.
Soit V et W deux espaces vectoriels sur le même corps, munis de bases $(e_j)_{j=1}^n$, $(f_i)_{i=1}^m$ respectivement. Relativement à ces bases, une application linéaire

$$\mathcal{A} : V \longrightarrow W$$

est représentée par la matrice à m lignes et n colonnes;

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

les éléments a_{ij} de la matrice A étant définis de façon unique par les relations

$$\mathcal{A} e_j = \sum_{i=1}^m a_{ij} f_i, \quad 1 \leq j \leq n$$

Autrement dit, le j-ème vecteur colonne

$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

de la matrice A représente le vecteur $\mathcal{A}e_j$ dans la base $(f_i)_{i=1}^m$. On appelle

$$(a_{i1}, a_{i2}, \ldots, a_{in})$$

le i-ème vecteur ligne de la matrice A.
Une matrice à m lignes et n colonnes est appelée matrice de type (m, n), et on note $\mathcal{A}_{m,n}(\mathbb{K})$, ou simplement $\mathcal{A}_{m,n}$, l'espace vectoriel sur le corps $\mathbb{K}$ formé par les matrices de type (m, n) à éléments dans $\mathbb{K}$. Un vecteur colonne est donc une matrice de type $(m, 1)$ et un vecteur ligne une matrice de type $(1, n)$. Une matrice est dite réelle ou complexe selon que ses éléments sont dans le corps $\mathbb{R}$ ou dans le corps $\mathbb{C}$.
Une matrice A d'éléments a_{ij} est notée

$$A = (a_{ij}),$$

le premier indice i étant toujours celui de la ligne et le second, j, celui de la colonne. Étant donné une matrice A, on désigne par $(A)_{ij}$ l'élément de la i-ème ligne et de la j-ème colonne.
La matrice nulle et le vecteur nul sont désignés par la même lettre O. Étant donné une matrice $A \in \mathcal{A}_{m,n}(\mathbb{C})$, on note $A^* \in \mathcal{A}_{m,n}(\mathbb{C})$ la matrice adjointe de la matrice A, définie de façon unique par les relations

$$(Au, v)_m = (u, A^*v)_n \text{ pour tout } u \in \mathbb{C}^n, v \in \mathbb{C}^m,$$

qui entraînent $(A^*)_{ij} = \bar{a}_{ji}$. De la même façon, étant donné une matrice $A = \mathcal{A}_{m,n}(\mathbb{R})$ on note $A^T \in \mathcal{A}_{m,n}(\mathbb{R})$ la matrice transposée de la matrice A, définie de façon unique par les relations

$$(Au, v)_m = (u, A^Tv)_n \text{ pour tout } u \in \mathbb{R}^n, v \in \mathbb{R}^m,$$

qui entraînent $(A^T)_{ij} = a_{ji}, \forall i, \forall j$.

Remarque 2.1.
(1) Nous pouvons encore définir la matrice transposée d'une matrice complexe, mais c'est une notion d'intérêt moindre, l'application $u, v \longrightarrow \sum_{i=1}^n u_i v_i$ n'étant pas un produit scalaire dans $\mathbb{C}^n$.
(2) Nous préférons la notation A^T à la notation habituelle ${}^tA$, cette dernière étant davan- tage adaptée à la notion de base duale ; la notation A^T rappelle la dépendance de la notion de matrice transposée sur un produit scalaire particulier, le produit scalaire canonique en l'occurrence.

Multiplication de deux matrices.
A la composition des applications linéaires correspond la multiplication des matrices ; Si $A = (a_{ik})$ est une matrice de type (m, l) et $B = (b_{kj})$ de type (l, n), leur produit AB est la matrice de type (m, n) définie par

$$(AB)_{ij} = \sum_{k=1}^{l} a_{ik} b_{kj}.$$

On rappelle que $(AB)^T = B^T A^T$, $(AB)^* = B^* A^*$.

Définition 2.1.
Une matrice de type (n, n) est dite matrice carrée, ou matrice d'ordre n si l'on veut préciser l'entier n ; il est alors commode de dire qu'une matrice est rectangulaire lorsqu'elle n'est pas nécessairement carrée. On note

$$\mathcal{A}_n = \mathcal{A}_{n,n}, \text{ ou} \mathcal{A}_n(\mathbb{K}) = \mathcal{A}_{n,n}(\mathbb{K}),$$

l'anneau des matrices carrées d'ordre n, à éléments dans le corps $\mathbb{K}$.

Définition 2.2.
Soit $A = (a_{ij})$ une matrice carrée, les éléments a_{ii} sont appelés éléments diagonaux, et les éléments $a_{ij}, i \neq j$, sont appelés éléments hors-diagonaux. La matrice unité est la matrice

$$I = (\delta_{ii}).$$

Définition 2.3.
Une matrice A est inversible s'il existe une matrice (unique si elle existe), notée A^{-1} et appelée matrice inverse de la matrice A, telle que $AA^{-1} = A^{-1}A = I$. Dans le cas contraire, on dit que la matrice est singulière. On rappelle que, si A et B sont des matrices inversibles

$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A^T)^{-1} = (A^{-1})^T, \quad (A^*)^{-1} = (A^{-1})^*.$$

Définition 2.4.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est symétrique si A est réelle et $A = A^T$.

Définition 2.5.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est hermitienne si $A = A^*$.

Définition 2.6.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est orthogonale si A est réelle et $AA^T = A^T A = I$.

Définition 2.7.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est idempotente si A est réelle et $A^2 = A$.

Définition 2.8.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est unitaire si $AA^* = A^*A = I$.

Définition 2.9.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est normale si $AA^* = A^*A$.

Définition 2.10.
Soit $A = (a_{ij})$ une matrice carrée, on dit que A est diagonale si $a_{ij} = 0$, pour $i \neq j$.
On la note

$$A = diag(a_{ii}) = diag(a_{11}, a_{22}, \ldots, a_{nn}).$$

La trace d'une matrice $A = (a_{ij})$ est définie par

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

Soit $\mathfrak{S}_n$ le groupe des permutations de l'ensemble $\{1, 2, \ldots, n\}$. A tout élément, $\sigma \in \mathfrak{S}_n$ on associe la matrice de permutation

$$P_\sigma = (\delta_{i\sigma(j)}).$$

On notera qu'une matrice de permutation est orthogonale.

Définition 2.11.
Soit $A = (a_{ij})$ une matrice carrée, le déterminant de la matrice A est défini par

$$det(A) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon_\sigma a_{\sigma(1)1} a_{\sigma(2)2} \ldots a_{\sigma(n)n}$$

où ε_σ désigne la signature de la permutation σ.

Proposition 2.1.
Les valeurs propres $\lambda_i = \lambda_i(A), \quad 1 \leq i \leq n$, d'une matrice A d'ordre n sont les n racines, réelles ou complexes, distinctes ou confondues, du polynôme caractéristique

$$P_A : \lambda \in \mathbb{C} \longrightarrow P_A(\lambda) = det(A - \lambda I)$$

de la matrice A. Le spectre de la matrice A est le sous-ensemble

$$sp(A) = \bigcup_{i=1}^{n} \{\lambda_i(A)\}$$

du plan complexe.

Remarque 2.2.
On rappelle les relations suivantes

$$tr(A) = \sum_{i=1}^{n} \lambda_i(A), \quad det(A) = \prod_{i=1}^{n} \lambda_i(A),$$

$$tr(AB) = tr(BA), \quad tr(A + B) = tr(A) + tr(B),$$

$$det(AB) = det(BA) = det(A)det(B).$$

Définition 2.12.
Le rayon spectral d'une matrice carrée d'ordre n, notée A est le nombre positif ou nul défini par

$$\varrho(A) = max\{|\lambda_i(A)|; 1 \leq i \leq n\}.$$

Définition 2.13.
A toute valeur propre λ d'une matrice A est associé (au moins) un vecteur p tel que $p \neq 0$ et $Ap = \lambda p$, appelé vecteur propre de la matrice A, correspondant à la valeur propre λ. Si $\lambda \in sp(A)$, le sous-espace vectoriel

$$\{\upsilon \in V; A\upsilon = \lambda\upsilon\}$$

(de dimension au moins égale à 1) est appelé sous-espace propre, correspondant à la valeur propre λ.

On conviendra que dans la décomposition par blocs $A = (A_{IJ})$ d'une matrice carrée, les sous-matrices diagonalesA_{II} sont toujours carrées.

Définition 2.14.
Étant donné deux espaces vectoriels de dimensions finies (mais non nécessairement égales) V et W, le rang d'une application linéaire $\mathcal{A} : V \longrightarrow W$ est égal à la dimension du sous-espace vectoriel

$$Im(\mathcal{A}) = \{\mathcal{A}v \in W; v \in V\}.$$

Si les espaces V et W sont munis de bases, vis-à-vis desquelles l'application $\mathcal{A}$ est représentée par une matrice A, le rang de $\mathcal{A}$ est aussi égal au plus grand ordre des sous matrices (carrées) inversibles de la matrice A. C'est pourquoi le rang de $\mathcal{A}$ est aussi appelé rang de la matrice A. On le note $r(A)$.

Remarque 2.3.
Faisons enfin une remarque générale, valable pour toute la suite. Toutes les fois que ce sera "raisonnablement" clair, on ne mentionnera pas les ensembles d'indices. C'est ainsi que si $A = (a_{ij})$ est une matrice de type (m, n) on se contentera d'écrire

$$\max_i\{\min_j a_{ij}\}, \quad \text{au lieu de} \quad \max_{1\leq i\leq n}\{\min_{1\leq j\leq m} a_{ij}\}$$

c'est ainsi qu'on écrira seulement

$$p_i^* p_j = \delta_{ij}, \quad \text{au lieu de} p_i^* p_j = \delta_{ij} \quad 1 \leq i, j \leq n,$$

s'il est clair que les indices i et j décrivent le même ensemble $\{l, 2, \ldots, n\}$, etc.

3 Réduction des matrices

Soit V un espace vectoriel de dimension finie n, et soit $\mathcal{A} : V \longrightarrow V$ une application linéaire, représentée par une matrice (carrée) $A = (a_{ij})$ relativement à une base (e_i). Relativement à une autre base $(\mathfrak{f}_\mathfrak{i})$, la même application est représentée par la matrice

$$B = P^{-1}AP,$$

où P est la matrice inversible dont le j-ème vecteur colonne est formé des composantes du vecteur $(\mathfrak{f}_\mathfrak{j})$ dans la base (e_j). La matrice P est appelée matrice de passage, de la base (e_i) dans la base $(\mathfrak{f}_\mathfrak{i})$.

Une même application linéaire $\mathcal{A}$ étant ainsi représentée par différentes matrices selon la base choisie, le problème se pose de trouver une base vis-à-vis de laquelle la matrice représentant l'application soit "aussi simple que possible". De façon équivalente, étant donné une matrice A, il s'agit de trouver parmi toutes les matrices semblables à la matrice A, c'est-à-dire de la forme $P^{-1}AP$, P étant la matrice inversible, celles qui ont

une forme "aussi simple que possible" : c'est le problème de la réduction d'une matrice. Le cas le plus "favorable" est celui où il existe une matrice inversible P telle que la matrice $P^{-1}AP$ soit diagonale, auquel cas on dit que la matrice A est diagonalisable.

Remarque 3.1.
On notera que, dans ce cas, les éléments diagonaux de la matrice $P^{-1}AP$ sont les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ de la matrice A, et que le j-ème vecteur colonne de la matrice P est formé des composantes (relativement à la même base que pour la matrice A) d'un vecteur propre correspondant à λ_j ; on a en effet l'équivalence

$$P^{-1}AP = diag(\lambda_j) \Leftrightarrow Ap_j = \lambda_j p_j, 1 \leq j \leq n.$$

Autrement dit, une matrice est diagonalisable si et seulement si, il existe une base de vecteurs propres. Il existe des matrices qui ne sont pas diagonalisables. Pour de telles matrices, le théorème de Jordan donne la forme la plus simple des matrices semblables.

Nous rappelons quelques notions utiles à travers la définition suivantes:

Définition 3.1.
Une matrice $A = (a_{ij})$ d'ordre n est triangulaire supérieure si $a_{ij} = 0$ pour $i > j$, et triangulaire inférieure si $a_{ij} = 0$ pour $i < j$. S'il n'y a pas lieu de distinguer, on dit que la matrice est triangulaire.

Théorème 3.1.
(1) *Étant donné une matrice carrée A, il existe une matrice unitaire V telle que la matrice $V^{-1}AV$ soit triangulaire.*
(2) *Étant donné une matrice normale A, il existe une matrice unitaire U telle que la matrice $U^{-1}AU$ soit diagonale.*
(3) *Étant donné une matrice symétrique A, il existe une matrice orthogonale* 0 *telle que la matrice $O^{-1}AO$ soit diagonale.*

Remarque 3.2.
(1) Les matrices de passage vérifiant les conditions de l'énoncé ne sont pas uniques (considérer par exemple $A = I$).
(2) Les éléments diagonaux de la matrice triangulaire $U^{-1}AU$ de (1), ou de la matrice diagonale $U^{-1}AU$ de (2), ou de la matrice diagonale de (3), sont les valeurs propres de la matrice A. En conséquence, ce sont des nombres réels si A est une matrice hermitienne ou symétrique, et des nombres complexes de module 1 si la matrice A est unitaire ou orthogonale.
(3) Il résulte de (2) que toute matrice hermitienne ou unitaire est diagonalisable par une matrice de passage unitaire.
(4) Si O est une matrice orthogonale, le raisonnement précédent montre l'existence d'une matrice unitaire U telle que la matrice $D = U^*OU$ soit diagonale (les éléments diagonaux

de D étant de module 1), mais la matrice U n'est pas en général réelle, c'est-à-dire orthogonale.

Définition 3.2.
On appelle valeurs singulières d'une matrice A carrée les racines carrées positives des valeurs propres de la matrice hermitienne A^*A (ou A^TA si la matrice A est réelle). Ces dernières sont toujours ≤ 0, puisque de la relation $A^*Ap = \lambda p$, $p \neq 0$, on déduit $(Ap)^*Ap = \lambda p^*p$.

Remarque 3.3.
On notera également que les valeurs singulières sont toutes > 0 si et seulement si la matrice A est inversible. En effet

$$Ap = 0 \Rightarrow A^*Ap = 0 \Rightarrow p^*A^*Ap = (Ap)^*Ap = 0 \Rightarrow Ap = 0.$$

Définition 3.3.
Deux matrices A et B de type (m, n) sont dites équivalentes s'il existe une matrice inversible Q d'ordre m et une matrice inversible P d'ordre n telles que

$$B = QAP.$$

Naturellement, il s'agit d'une notion plus générale que celle de la similitude des matrices. On peut d'ailleurs démontrer que toute matrice carrée est équivalente à une matrice diagonale, à travers le théorème suivant,

Théorème 3.2.
Si A est une matrice réelle carrée, il existe deux matrices orthogonales U et V telles que

$$U^TAV = diag(\mu_i),$$

et si A est une matrice complexe carrée, il existe deux matrices unitaires U et V telles que

$$U^*AV = diag(\mu_i).$$

Dans les deux cas, les nombres $\mu_i \geq 0$ sont les valeurs singulières de la matrice A.

4 Propriétés particulières des matrices symétriques et hermitiennes

Nous allons considérer dans ce qui suit le cas des matrices hermitiennes, mais il est entendu que tout le contenu de ce paragraphe s'applique aussi bien au cas des matrices

symétriques, en remplaçant partout "hermitien", "unitaire", "complexe", "matrice adjointe" par "symétrique", "orthogonal", "réel", "matrice transposée", respectivement.
Rappelons que toutes les valeurs propres d'une matrice hermitienne sont réelles, et que toute matrice hermitienne est diagonalisable, la matrice de passage étant unitaire (théorème 3.1). Il existe, de surcroît, diverses caractérisations remarquables des valeurs propres d'une matrice hermitienne, qui font l'objet du théorème ci-dessous. Pour les énoncer, il nous faut tout d'abord une définition.

Définition 4.1.
Soit A une matrice carrée représentant une application linéaire d'un espace V sur le corps $\mathbb{C}$, muni de son produit canonique. Le quotient de Rayleigh de la matrice A est l'application,

$$R_A : V - \{0\} \to \mathbb{C}$$

définie par

$$R_A(v) = \frac{(Av, v)}{(v, v)} = \frac{v^* Av}{v^* v}, \quad v \neq 0$$

On notera que, si la matrice A est hermitienne, le quotient de Rayleigh R_A est à valeurs réelles. Par ailleurs, on remarque aussi que

$$R_A(\alpha v) = R_A(v) \text{ pour tout } \alpha \in \mathbb{C} - \{0\}.$$

En conséquence, toute propriété faisant intervenir l'ensemble des valeurs prises par le quotient de Rayleigh lorsque le vecteur v décrit un sous-espace vectoriel $U \subset V$ peut aussi bien s'étudier sur la sphère unité $\{v \in U; v_* v = 1\}$ de ce même sous-espace ; c'est, en particulier, le cas des propriétés établies dans le résultat qui suit. Pour alléger l'écriture, on omettra dorénavant de préciser que, dans l'écriture $R_A(v)$, l'argument v ne saurait être nul.

Théorème 4.1.
Soit A une matrice hermitienne d'ordre n, de valeurs propres

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n,$$

les vecteurs propres associés $p_1, p_2, \ldots, p_n$ vérifiant

$$p_i p_j^* = \delta_{ij}.$$

Pour $k = 1, \ldots, n$, on note V_k le sous-espace de V engendré par les vecteurs p_i, $1 \leq i \leq k$ et on note $\mathfrak{S}_k$ l'ensemble des sous-espaces de dimension k de V. On pose par ailleurs

$$V_0 = \{0\}, \mathfrak{S}_0 = \{V_0\}.$$

Les valeurs propres admettent alors les caractérisations suivantes, pour $k = 1, 2, \ldots n$:

$$(1) \qquad \lambda_k = R_A(p_k),$$

$$(2) \qquad \lambda_k = \max_{v \in V_k} R_A(v),$$

$$(3) \qquad \lambda_k = \min_{v \perp V_{k-1}} R_A(v),$$

$$(4) \qquad \lambda_k = \min_{W \in \mathfrak{S}_k} \max_{v \in W} R_A(v),$$

$$(5) \qquad \lambda_k = \max_{W \in \mathfrak{S}_{k-1}} \min_{v \perp W} R_A(v),$$

Par ailleurs,

$$(6) \qquad \{R_A(v); v \in V\} = [\lambda_1, \lambda_n] \subset \mathbb{R}.$$

Remarque 4.1.
(1) Comme cas particuliers des caractérisations (2) et (3), on trouve

$$\lambda_1 = \min\{R_A(v); v \in V\},$$

$$\lambda_n = \max\{R_A(v); v \in V\}.$$

(2) Les propriétés (4)-(5) sont dues à Fischer, E. et Courant, R. les a ensuite étendues au cas des opérateurs elliptiques. C'est pourquoi elles sont souvent connues sous le nom de théorème de Courant-Fischer.

Pour terminer, rappelons quelques définitions,

Définition 4.2.
Une matrice hermitienne A est définie positive si

$$v^*Av > 0 \quad \text{pour tout} \quad v \in V - \{0\},$$

et positive si

$$v^*Av \geq 0 \quad \text{pour tout} \quad v \in V,$$

Remarque 4.2.
On établit facilement qu'une matrice hermitienne est définie positive, ou positive, si et seulement si toutes ses valeurs propres sont > 0 ou ≥ 0, respectivement.

Remarque 4.3.
La terminologie "matrice positive" désigne aussi une matrice dont tous les éléments sont positifs ou nuls.

5 Normes vectorielles et normes matricielles

Dans ce paragraphe, nous allons introduire la notion de norme dans le cas des vecteurs et des matrices.

Définition 5.1.
Soit V un espace vectoriel sur le corps $\mathbb{K}$ des scalaires. Une norme sur V est une application $||.|| : V \to \mathbb{R}$ qui vérifie les propriétés suivantes;

$$||v|| = 0 <=> v = 0, \quad \text{et} \quad ||v|| \geq 0 \quad \text{pour tout} \quad v \in V,$$

$$||\alpha v|| = |\alpha|||v|| \quad \text{pour tout} \quad \alpha \in \mathbb{K} \quad \text{et} \quad v \in V,$$

$$||u + v|| \leq ||u|| + ||v|| \quad \text{pour tout} \quad u, v \in V,$$

la dernière propriété étant connue sous le nom d'inégalité triangulaire. Une norme sur V sera également appelée norme vectorielle. Lorsque plusieurs espaces sont en cause, la notation $||.||_V$ sera parfois utilisée pour rappeler l'espace V considéré. Enfin, on appelle espace vectoriel normé un espace vectoriel muni d'une norme.

Soit V un espace de dimension finie. Les trois normes suivantes sont les plus couramment utilisées en pratique:

$$||v||_1 = \sum_i |v_i|,$$

$$||v||_2 = \left(\sum_i |v_i|^2 \right)^{1/2} = (v, v)^{1/2},$$

$$||v||_\infty = \max_i |v_i|,$$

la norme $||.||_2$ étant appelée norme euclidienne. Il est facile de vérifier directement que les applications $||.||_1$ et $||.||_\infty$ sont effectivement des normes. Pour l'application $||.||_2$, c'est un cas particulier du résultat général suivant:

Théorème 5.1.
Soit V un espace de dimension finie. Pour tout nombre réel $p \geq 1$, l application $||.||_p$ définie par

$$||v||_p = \left(\sum_i |v_i|^p \right)^{1/p}$$

est une norme

Définition 5.2.
Pour $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, l'inégalité

$$\sum_i |u_i v_i| \leq \left(\sum_i |u_i|^p \right)^{1/p} \left(\sum_i |v_i|^p \right)^{1/p}$$

s'appelle inégalité de Hölder. L'inégalité de Hölder pour $p = 2$;

$$\sum_i |u_i v_i| \leq \left(\sum_i |u_i|^2 \right)^{1/2} \left(\sum_i |v_i|^2 \right)^{1/2}$$

s'appelle inégalité de Cauchy-Schwarz, ou encore inégalité de Bouniakovski. L'inégalité triangulaire pour la norme $||.||_p$;

$$\left(\sum_i |u_i + v_i|^p \right) \frac{1}{p} \leq \left(\sum_i |u_i|^p \right)^{1/p} + \left(\sum_i |v_i|^p \right)^{1/p}$$

s'appelle l'inégalité de Minkowski.

Les normes définies ci-dessus sont équivalentes, cette propriété étant un cas particulier de l'équivalence des normes sur un espace vectoriel de dimension finie. On rappelle que deux normes $||.||$ et $||.||'$, définies sur un même espace vectoriel V, sont équivalentes s'il existe deux constantes C et C' telles que

$$||v||' \leq C||v|| \quad \text{et} \quad ||v|| \leq C'||v||' \quad \text{pour tout} \quad v \in V.$$

Définition 5.3.
Soit $\mathcal{A}_n$ l'anneau des matrices d'ordre n, à éléments dans le corps $\mathbb{K}$. Une norme matricielle est une application $||.|| : \mathfrak{A}_n \to \mathbb{R}$ qui vérifie les propriétés suivantes:

$$||A|| = 0 <=> A = 0, \quad \text{et} \quad ||A|| \geq 0 \quad \text{pour tout} \quad A \in \mathcal{A}_n,$$

$$||\alpha A|| = |\alpha|.||A|| \quad \text{pour tout} \quad \alpha \in \mathbb{K} \quad \text{et} \quad A \in \mathcal{A}_n,$$

$$||A + B|| \leq ||A|| + ||B|| \quad \text{pour tout} \quad A, B \in \mathcal{A}_n,$$

$$||AB|| \leq ||A||.||B|| \quad \text{pour tout} \quad A, B \in \mathcal{A}_n,$$

L'anneau $\mathcal{A}_n$ étant aussi un espace vectoriel de dimension n^2 , les trois premières propriétés ci-dessus ne sont autres que celles d'une norme vectorielle, une matrice étant alors considérée comme un vecteur à n^2 composantes. La dernière propriété est évidemment particulière aux matrices carrées.
Le résultat qui suit donne un moyen très simple de construire des normes matricielles:

Proposition 5.1.
Etant donné une norme vectorielle $||.||$ sur C^n, l'application $||.|| : \mathcal{A}_n(C) \to \mathbb{R}$ définie par

$$||A|| = \sup_{v \in C^n, v \neq 0} \frac{||Av||}{||v||} = \sup_{v \in C^n, ||v|| \leq 1} ||Av|| = \sup_{v \in C^n, ||v|| = 1} ||Av||$$

est une norme matricielle, appelée norme matricielle subordonnée (à la norme vectorielle donnée). C'est évidemment un cas particulier de la définition usuelle de la norme d'une application linéaire, mais attention! il existe des normes matricielles qui ne sont subordonnées à aucune norme vectorielle.

Calculons maintenant chacune des normes subordonnées aux normes vectorielles $||.||_1$, $||.||_2$, $||.||_\infty$. Pour alléger l'écriture, on omettra dorénavant l'indication que les bornes supérieures sont à évaluer sur l'ensemble des vecteurs non nuls de C^n.

Théorème 5.2.
Soit $A = (a_{ij})$ une matrice carrée. Alors

$$||A||_1 = \sup \frac{||Av||_1}{||v||_1} = \max_j \sum_i |a_{ij}|,$$

$$||A||_2 = \sup \frac{||Av||_2}{||v||_2} = \sqrt{\varrho(A^*A)} = \sqrt{\varrho(AA^*)} = ||A^*||_2,$$

$$||A||_\infty = \sup \frac{||Av||_\infty}{||v||_\infty} = \max_i \sum_j |a_{ij}|.$$

La norme $||.||_2$ est invariante par transformation unitaire:

$$UU^* = I \Rightarrow ||A||_2 = ||AU||_2 = ||UA||_2 = ||U^*AU||_2$$

Par ailleurs, si la matrice A est normale :

$$AA^* = A^*A \Rightarrow ||A||_2 = \varrho(A).$$

Remarque 5.1.
(1) La norme $||A||_2$ n'est autre que la plus grande valeur singulière de la matrice A.
(2) Si une matrice A est hermitienne, ou symétrique (donc normale), on a $||A||_2 = \varrho(A)$.
(3) Si une matrice U est unitaire, ou orthogonale (donc normale), on a $||A||_2 = \sqrt{\varrho(A^*A)} = \sqrt{\varrho(I)} = 1$.
(4) Du point de vue pratique, on observera que, si les normes $||A||_1$ et $||A||_\infty$ se calculent facilement à partir de la seule connaissance des éléments de la matrice A, il n'en va pas de même pour Ia norme $||A||_2$.

Théorème 5.3.
***(1)** Soit A une matrice carrée quelconque et $||.||$ une norme matricielle, subordonnée ou non, quelconque. Alors*

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

***(2)** Étant donné une matrice A et un nombre $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée telle que*

$$||A|| \leq \varrho(A) + \varepsilon$$

Un exemple important de norme matricielle non subordonnée est donné dans le théorème ci-dessous.

Théorème 5.4.
L'application $||.||_E : \mathcal{A}_n \to \mathbb{R}$ *définie par*

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

pour toute matrice $A = (a_{ij})$ *d'ordre* n*, est une norme matricielle non subordonnée (pour* $n \geq 2$*), invariante par transformation unitaire:*

$$UU^* = I \Rightarrow ||A||_E = ||AU||_E = ||UA||_E = ||U^*AU||_E,$$

et qui vérifie

$$||A||_2 \leq ||A||_E \leq \sqrt{n}||A||_2, \quad \textit{pour tout} \quad A \in \mathcal{A}_n.$$

Remarque 5.2.
Contrairement à la norme matricielle subordonnée $||.||_2$, la norme $||.||_E$ se prête facilement à un calcul effectif. C'est là un de ses principaux intérêts, puis- qu'elle fournit en particulier un majorant de la norme $||.||_2$

Terminons par un théorème qui rassemble quelques propriétés utiles concernant les matrices de la forme $(I + B)$.

Théorème 5.5.
(1) *Soit* $||.||$ *une norme matricielle subordonnée, et* B *une matrice vérifiant*

$$||B|| < 1$$

Alors la matrice $(1 + B)$ *est inversible, et*

$$||(1 + B)^{-1}|| \leq \frac{1}{1 - ||B||}$$

(2) *Si une matrice de la forme* $||(I + B)||$ *est singulière, alors nécessairement*

$$||B|| \geq 1$$

pour toute norme matricielle, subordonnée ou non.

6 Suites de vecteurs et de matrices

Une suite (infinie) d'éléments $x_0, x_1, \ldots$, d'un ensemble X sera notée $(x_k)_{k\geq 0}$, où même simplement (x_k) .

Définition 6.1.
Dans un espace vectoriel V, muni d'une norme $||.||$, on dit qu'une suite (v_k) d'éléments de V converge vers un élément $v \in V$, ou encore que v est la limite de la suite (v_k), si

$$\lim_{k\to\infty} ||v_k - v|| = 0$$

et on écrit

$$v = \lim_{k\to\infty} v_k$$

Si l'espace est de dimension finie, l'équivalence des normes montre que la convergence d'une suite est indépendante de la norme choisie. Le choix particulier de la norme $||.||_\infty$ montre que la convergence d'une suite de vecteurs équivaut à la convergence des n suites (n = dimension de l'espace) de scalaires formées par les composantes des vecteurs.

En considérant l'ensemble $\mathcal{A}_{m,n}(\mathbb{K})$ des matrices de type (m, n) comme un espace vectoriel à mn dimensions, on voit de la même façon que la convergence d'une suite de matrices de type (m, n) est indépendante de la norme choisie, et qu'elle équivaut à la convergence des mn suites de scalaires formées par les éléments des matrices.

Le résultat qui suit donne des conditions nécessaires et suffisantes pour que la suite particulière formée par les puissances successives d'une matrice donnée (carrée...) converge vers la matrice nulle. De ces conditions découlera le critère fondamental de convergence des méthodes itératives de résolution de systèmes linéaires.

Théorème 6.1.
Soit B une matrice carrée. Les conditions suivantes sont équivalentes

$$(1) \quad \lim_{k\to\infty} B^k = 0,$$

$$(2) \quad \lim_{k\to\infty} B^k v = 0, \quad \textit{pour tout vecteur } v$$

$$(3) \quad \varrho(B) < 1,$$

$$(4) \quad ||B|| < 1 \quad \textit{pour au moins une norme matricielle subordonnée } ||.||.$$

Le résultat qui suit sert également à l'étude des méthodes itératives, en ce qui concerne la rapidité de leur convergence. Ce n'est d'ailleurs qu'un cas particulier (celui de la dimension finie) d'un résultat d'analyse fonctionnelle vrai dans les espaces de Banach.

Théorème 6.2.
Soit B une matrice carrée, et $||.||$ une norme matricielle quelconque. Alors

$$\lim_{k\to\infty} ||B^k||^{1/k} = \varrho(B).$$

avec $\varrho(B)$ représentant le rayon spectral de B.

CHAPITRE II. Résolution numérique des systèmes d'équations linéaires

1 Introduction.

L'analyse matricielle étudie deux problèmes fondamentaux : l'inversion de matrices ou la résolution de systèmes linéaires qui fait l'objet du présent chapitre et le calcul des valeurs et des vecteurs propres d'une matrice qui sera traité dans le chapitre suivant. On montre comment la méthode bien connue de Cramer a des limites en ce sens que le nombre d'opérations à effectuer est énorme. Cependant, il existe des algorithmes de résolution des systèmes linéaires, celles ci se classent en deux grandes catégories : les méthodes directes (méthodes de Gauss, Cholesky) qui permettent d'obtenir la solution en nombre fini d'opérations et les méthodes itératives (méthodes de Jacobi, Gauss-Seidel) qui recherchent la solution de proche en proche en partant d'un vecteur initial arbitraire. Les algorithmes et leurs implantations en machine mettent en jeu des techniques spéciales lorsque les matrices ont des formes particulières (matrices bandes, tridiagonales, creuses, diagonales par blocs, etc...).

2 Conditionnement d'un système

Résoudre le système linéaire $Ax = b$ serait idéalement de déterminer le vecteur x vérifiant l'équation donnée. Cependant, à cause des erreurs d'arrondi, la solution calculée est x^*. Considérons le vecteur $r = Ax - Ax^* = A.(x - x^*)$ appelé le vecteur résidu. On serait tenté de penser que si r est petit (au sens d'une norme choisie) alors x^* représentera une bonne approximation de x, mais, ce n'est pas toujours le cas. En effet, soit l'exemple suivant

Exemple 2.1.

$$A = \begin{pmatrix} 3.02 & -1.05 & 2.53 \\ 4.33 & 0.56 & -1.78 \\ -0.23 & -0.54 & 1.47 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} -1.61 \\ 7.23 \\ -3.38 \end{pmatrix}$$

La solution exacte est égale à $x = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$

Soit $x^* = \begin{pmatrix} 0.88 \\ -2.35 \\ -2.66 \end{pmatrix}$. Alors, $A.x^* = \begin{pmatrix} -1.61 \\ 7.23 \\ -3.37 \end{pmatrix}$ et par suite,

$$r = Ax - Ax^* = \begin{pmatrix} 0.00 \\ 0.00 \\ -0.01 \end{pmatrix}$$

Ainsi, $\|r\|$ est petit alors que x et x^* sont complètement différentes.

Cet exemple est important car il signifie que l'on ne peut pas vérifier la précision de la solution d'un système linéaire en remplaçant simplement la solution dans l'équation et en calculant les résidus. Ceci peut être déduit de la relation suivante :

$$r = A.(x - x^*) \Longrightarrow x - x^* = A^{-1}r$$

Cette relation implique que si des éléments de A^{-1} sont "grands" alors même si r est petit, l'erreur sur la solution peut être grande.

Exemple 2.2.
Considérons les deux systèmes suivants :

$$\left\{ \begin{array}{rcrcl} 2x & + & 6y & = & 8 \\ 2x & + & 6.00001\ y & = & 8.00001 \end{array} \right. \qquad (I)$$

et

$$\left\{ \begin{array}{rcrcl} 2x & + & 6y & = & 8 \\ 2x & + & 5.99999\ y & = & 8.00002 \end{array} \right. \qquad (II)$$

La solution de (I) est: $x = y = 1$.
La solution de (II) est: $x = 10$ et $y = -2$.
Les deux systèmes sont quasiment identiques mais ils admettent des solutions "complètement" différentes. Posons

$$A = \begin{pmatrix} 2 & 6 \\ 2 & 6.00001 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 8 \\ 8.00001 \end{pmatrix}$$

Ainsi, un changement de 2×10^{-5} de l'élément a_{22} de A et de 10^{-5} de la deuxième composante de b a causé un changement important de la solution.
Supposons que (II) est le système contenant des données expérimentales de (I). Si les données n'ont pas une précision meilleure que 10^{-5} (ce qui est déjà remarquable) alors qu'elle que soit la méthode de résolution utilisée, on obtiendra une solution qui pourra être fort erronée. On dit que la matrice A (ou le système $Ax = b$) est mal conditionnée.

Définition 2.1.
On appelle conditionnement d'une matrice A, le nombre

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

Remarque 2.1.

$$\|A\|.\|A^{-1}\| \geq \|A.A^{-1}\| = \|I\| = 1$$

Théorème 2.1.
On pose $e = x - \bar{x}$ *et* $r = Ax - A\bar{x} = b - A\bar{x}$.
On a alors la relation suivante :

$$\frac{1}{Cond(A)}.\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq Cond(A).\frac{\|r\|}{\|b\|}$$

où $\frac{\|e\|}{\|x\|}$ *représente l'erreur relative*

Remarque 2.2.
On dit que le système $Ax = b$ est mal conditionné lorsque $Cond(A)$ est "grand".
Pour fixer les idées, supposons que $\frac{\|r\|}{\|b\|} = 10^{-4}$. Ainsi,
Si $Cond(A) \simeq 1$ alors, $\frac{\|e\|}{\|x\|} \simeq 10^{-4}$
Si $Cond(A) \simeq 10$ alors, $10^{-5} \leq \frac{\|e\|}{\|x\|} \simeq 10^{-3}$
Si $Cond(A) \simeq 10^5$ alors, $10^{-9} \leq \frac{\|e\|}{\|x\|} \simeq 10$

Lorsque $Cond(A)$ est "grand", on n'a quasiment aucune information sur la valeur de $\frac{\|e\|}{\|x\|}$ alors que l'on a une bonne estimation de $\frac{\|e\|}{\|x\|}$ par rapport à $\frac{\|r\|}{\|b\|}$ dans le cas où $Cond(A)$ est "petit".

Remarque 2.3.
(1). La matrice identité est la matrice la mieux conditionnée. En d'autres termes, une petite perturbation des éléments de I ou de b n'aura pas beaucoup d'effets sur la solution de $Ix = b$.
(2). Autant $Cond(A)$ s'éloigne de 1, autant la matrice A est mal conditionnée pour des valeurs de $\frac{\|r\|}{\|x\|}$ presque identiques.
(3). Le calcul d'un conditionnement ne permet pas d'atténuer les erreurs mais il ne fait que réveler que, lorsque $Cond(A)$ est grand, le risque d'erreurs sur la solution l'est aussi.
(4). Comme exemples de matrices mal conditionnées, on peut citer les matrices H_n de

Hilbert qui sont du type

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n+1} \end{pmatrix}$$

On a

$$\begin{aligned} Cond(H_3) &\simeq 524.06 \\ Cond(H_4) &\simeq 15514 \\ Cond(H_5) &\simeq 4.7\text{x}10^5 \\ Cond(H_8) &\simeq 1.5\text{x}10^{10} \end{aligned}$$

Maintenant, nous nous proposons de majorer l'erreur relative sur la solution d'un système en fonction de son conditionnement. Pour cela, nous considérons deux cas. Soit un système linéaire $Ax = b$

1er cas. Seul b est perturbé

Le système $Ax = b$ devient alors $A.(x + \delta x) = B + \delta b$. Soit une norme vectorielle quelconque et une norme matricielle subordonnée. On écrit

$$A.(x + \delta x) = b + \delta b \Longrightarrow \delta x = A^{-1}.\delta b$$

$$\delta x = A^{-1}.\delta b \Longrightarrow \|\delta x\| \leq \|A^{-1}\|.\|\delta b\|$$

$$b = Ax \Longrightarrow \|b\| \leq \|A\|.\|x\|$$

Ce qui donne

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|.\|A^{-1}\|.\frac{\|\delta b\|}{\|b\|} \Longrightarrow \frac{\|\delta x\|}{\|x\|} \leq Cond(A).\frac{\|\delta b\|}{\|b\|}$$

2 ème cas. Seule A est perturbée

le système $Ax = b$ s'écrit alors $(A + \delta A).(x + \delta x) = b$. En procédant de la même manière que le premier cas, on obtient

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{Cond(A).\frac{\|\delta A\|}{\|A\|}}{1 - Cond(A).\frac{\|\delta A\|}{\|A\|}}$$

3 Méthode de Cramer et système triangulaire

3.1 Méthode de Cramer

Soit à résoudre un système linéaire $Ax = b$ d'ordre n avec A inversible. La méthode de Cramer permet de dire que si le déterminant de la matrice A est non nul ($\det(A) \neq 0$) , alors les solutions $x_i,\ i = 1, \ldots, n$ sont obtenues par la formule suivante

$$x_i = \frac{\det(A_i)}{\det(A)} \quad i = 1, \ldots, n$$

où A_i est la matrice obtenue en remplaçant la $i^{\text{ème}}$ colonne de A par le vecteur b.

Exemple 3.1.
Soit le système $Ax = b$ avec

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \\ 2 & 3 & -5 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 10 \\ 2 \\ 7 \end{pmatrix}$$

On a alors, $det(A) = 31$, $det(A_1) = 93$, $det(A_2) = 62$, $det(A_3) = 31$. La solution est alors,

$$x_1 = 3 \quad , \quad x_2 = 2 \quad , \quad x_3 = 1$$

Remarque 3.1.
L'usage de la méthode de Cramer nécessite le calcul de $(n+1)$ déterminants et d'effectuer n divisions.
Le nombre d'opérations nécessaires est environ égal à
$(n + 1)!$ additions (une soustraction est considérée comme une addition)
$(n + 2)!$ multiplications
et n divisions.

Par conséquent, cela montre qu'elle reste une méthode limitée en ce sens que le nombre d'opérations nécessaires rend l'usage de la méthode impossible. En pratique, c'est le cas quand $n > 10$. Nous avons alors besoin de faire appel à des méthodes qui peuvent contourner ce problème. C'est le cas, notamment des méthodes numériques dont quelques unes sont présentées dans cet ouvrage sous forme de deux types :
• Les méthodes directes qui permettent théoriquement de calculer la solution en un nombre fini d'opérations connu d'avance. C'est le cas des méthodes de Gauss, de Gauss-Jordan et de Cholesky.
• Les méthodes indirectes ou itératives qui donnent la solution obtenue après convergence d'une suite de vecteurs. C'est le cas des méthodes de Jacobi et de Gauss-Seidel.

3.2 Système triangulaire

On dit qu'un système linéaire $Ax = b$ est triangulaire supérieur si la matrice A est triangulaire supérieure, c'est-à-dire,

$$a_{ij} = 0 \quad \forall i > j \quad ; \; i, j = 1, \ldots, n$$

Il a donc la forme suivante

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots\ldots + a_{1n}x_n = b_1 \\ \quad + a_{22}x_2 + a_{23}x_3 + \ldots\ldots + a_{2n}x_3 = b_2 \\ \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots \\ \qquad\qquad\qquad\quad a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ \qquad\qquad\qquad\qquad\quad a_{n,n}x_n = b_n \end{cases}$$

La résolution d'un tel système se fait par remontée triangulaire que l'on résume comme suit:

$$\begin{cases} x_n = \dfrac{b_n}{a_{nn}} \\ x_i = \dfrac{1}{a_{ii}}.\left(b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij}x_j\right) \qquad i = n-1, n-2, \ldots, 1 \end{cases}$$

Exemple 3.2.
Soit à résoudre le système suivant:

$$\begin{pmatrix} 2 & 3 & 1 \\ 0 & -5 & 1 \\ 0 & 0 & 6/5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ -7 \\ -12/5 \end{pmatrix}$$

On a alors,
$6/5x_3 = -12/5 \Longrightarrow x_3 = -2$
$-5x_2 + x_3 = -7 \Longrightarrow x_2 = 1$
$2x_1 + 3x_2 + x_3 = 5 \Longrightarrow x_1 = 2$
d'où la solution

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}$$

4 Méthode de Gauss

La méthode de Gauss est une méthode directe qui repose sur le résultat suivant

Théorème 4.1.
Soit une matrice carrée A quelconque. Il existe une matrice inversible M telle que $D = MA$ soit triangulaire supérieure.

4.1 Principe de la méthode de Gauss

La résolution d'un système linéaire $Ax = b$ par la méthode de Gauss consiste à trouver une matrice M inversible telle que MA soit une matrice triangulaire supérieure (On dit que l'on a triangularisé A) et de remplacer $Ax = b$ par le système triangulaire $MAx = Mb$ qui possède les mêmes solutions. Ensuite, nous appliquons la remontée triangulaire à ce nouveau système.
Sur le plan numérique, la connaissance explicite de la matrice M n'est pas importante car il est possible d'obtenir directement MA et Mb sans passer par le calcul de M. Cette opération se fait en n étapes,
On pose $A^{(1)} = A$ et $b^{(1)} = b$. On se propose de trouver les matrices $A^{(2)}$, $A^{(3)}$, ... et $A^{(n)}$ ainsi que les vecteurs $b^{(2)}$, $b^{(3)}$, ... et $b^{(n)}$ de telle manière que $A^{(i)}x = b^{(i)} \quad \forall i = 1, \ldots, n$ et $A^{(n)}$ soit triangulaire supérieure. Chaque matrice $A^{(k)}$ $k = 2, \ldots, n$ a la forme suivante

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1k}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2k}^{(2)} & \ldots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3k}^{(3)} & \ldots & a_{3n}^{(3)} \\ \vdots & \vdots & 0 & \ddots & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & 0 & a_{k,k}^{(k)} & \ldots & a_{k,n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a_{n,k}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix}$$

avec $a_{ij}^{(k)}$ qui représente le terme géneral de la matrice $A^{(k)}$. L'élément $a_{kk}^{(k)}$ s'appelle le pivot.

Algorithme

$$\left.\begin{array}{l} m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \\ b_i^{(k+1)} = b_i^{(k)} - m_{ik}.b_k^{(k)} \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}.a_{kj}^{(k)} \end{array}\right\} j = k+1, \ldots, n \Bigg\} i = k+1, \ldots, n \Bigg\} k = 1, \ldots, n-1$$

L'écriture matricielle de cet algorithme est comme suit

$$\left.\begin{array}{l} A_{i.}^{(k+1)} = A_{i.}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.A_{k.}^{(k)} \\ b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.b_k^{(k)} \end{array}\right\} i = k+1, \ldots, n$$

où $A_{i.}^{(k+1)}$désigne la $i^{\text{ème}}$ ligne de la matrice $A^{(k+1)}$.

Remarque 4.1.
Le nombre d'opérations est

$$\begin{aligned} \frac{n.(n+1)}{2} &\text{ divisions} \\ \frac{2n^3+3n^2-5n}{6} &\text{ multiplications} \\ \frac{2n^3+3n^2-5n}{6} &\text{ additions} \end{aligned}$$

Ce qui donne le nombre total d'opérations de l'ordre de $n^3/3 + O(n^2)$.

Exemple 4.1.
Soit à résoudre le système suivant:

$$\underbrace{\begin{pmatrix} 2 & 3 & 1 \\ 4 & 1 & 3 \\ -2 & 3 & -1 \end{pmatrix}}_{\text{A}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{\text{x}} = \underbrace{\begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix}}_{\text{b}}$$

Pour faire apparaître en même temps la matrice du système et le vecteur correspondant, utilisons la notation suivante

$$\underbrace{\left(\begin{array}{ccc} 2 & 3 & 1 \\ 4 & 1 & 3 \\ -2 & 3 & -1 \end{array}\right)}_{A^{(1)}} \underbrace{\begin{array}{c} 5 \\ 3 \\ 1 \end{array}}_{b^{(1)}}$$

On obtient:

$$\text{Première étape : } \underbrace{\left(\begin{array}{ccc} 2 & 3 & 1 \\ 0 & -5 & 1 \\ 0 & 6 & 0 \end{array}\right)}_{A^{(2)}} \underbrace{\begin{array}{c} 5 \\ -7 \\ 6 \end{array}}_{b^{(2)}}$$

$$\text{Deuxième étape : } \underbrace{\left(\begin{array}{ccc} 2 & 3 & 1 \\ 0 & -5 & 1 \\ 0 & 0 & 6/5 \end{array}\right)}_{A^{(3)}} \underbrace{\begin{array}{c} 5 \\ -7 \\ -12/5 \end{array}}_{b^{(3)}}$$

La résolution par remontée triangulaire donne, comme le montre l'exemple 3.2,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}$$

4.2 Méthode de Gauss avec pivot

4.2.1 Exemple introductif.

On considère le système

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{1}$$

Supposons que l'on veuille résoudre ce système avec une calculatrice à 4 chiffres. L'application de la méthode de Gauss à ce système comporte une seule étape.

$$\begin{pmatrix} 10^{-4} & 1 \\ 0 & -9999 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -9998 \end{pmatrix} \tag{2}$$

La résolution de (2) donne

$$x_1 = 0 \text{ et } x_2 \simeq 1$$

Ce qui est une solution fausse.
Le système (1) est équivalent au suivant:

$$\begin{pmatrix} 1 & 1 \\ 10^{-4} & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \tag{3}$$

La méthode de Gauss conduit au système suivant :

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.9999 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9998 \end{pmatrix} \tag{4}$$

La résolution de (4) donne

$$x_1 \simeq 1 \quad \text{et} \quad x_2 \simeq 1$$

Quoique sur le plan strictement mathématique, les deux méthodes sont applicables, l'une des deux a donné une réponse fausse. La raison est que la division par des pivots très petits gonfle l'erreur d'arrondi et induit une erreur trop importante. C'est pourquoi nous avons besoin de faire appel à des stratégies qui nous amènent à diviser par des pivots toujours plus grands. Pour ce faire, il existe deux stratégies décrites ci-après

4.2.2 Stratégie du pivot partiel

La stratégie du pivot partiel consiste à choisir comme pivot, à chaque étape l ($l = 1, \ldots, n$), l'un des éléments $a_{il}^{(l)}, \quad i = l, \ldots, n$ tel que

$$|a_{il}^{(l)}| = \max_{l \leq p \leq n} |a_{pl}^{(l)}|$$

Exemple 4.2.
Soit à résoudre le système

$$\begin{pmatrix} 3 & 1 & 2 \\ 6 & 4 & 1 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \\ 7 \end{pmatrix}$$

en utilisant les notations décrites précédemment, on aura

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & 4 \\ 6 & 4 & 1 & 10 \\ 3 & 4 & 1 & 7 \end{array}\right)$$

Les différentes étapes sont :

$$\underbrace{\left(\begin{array}{ccc|c} 6 & 4 & 1 & 10 \\ 3 & 1 & 2 & 4 \\ 3 & 4 & 1 & 7 \end{array}\right)}_{\text{Stratégie de pivot}} \longrightarrow \underbrace{\left(\begin{array}{ccc|c} 6 & 4 & 1 & 10 \\ 0 & -1 & 3/2 & -1 \\ 0 & 2 & 1/2 & 2 \end{array}\right)}_{\text{Méthode de Gauss}}$$

$$\underbrace{\left(\begin{array}{ccc|c} 6 & 4 & 1 & 10 \\ 0 & 2 & 1/2 & 2 \\ 0 & -1 & 3/2 & -1 \end{array}\right)}_{\text{Stratégie de pivot}} \longrightarrow \underbrace{\left(\begin{array}{ccc|c} 6 & 4 & 1 & 10 \\ 0 & 2 & 1/2 & 2 \\ 0 & 0 & 7/4 & 0 \end{array}\right)}_{\text{Méthode de Gauss}}$$

La résolution par retour arrière du système

$$\left(\begin{array}{ccc} 6 & 4 & 1 \\ 0 & 2 & 1/2 \\ 0 & 0 & 7/4 \end{array}\right)\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 10 \\ 2 \\ 0 \end{array}\right)$$

donne la solution

$$\left(\begin{array}{c} 1 \\ 1 \\ 0 \end{array}\right)$$

4.2.3 Stratégie du pivot total

La stratégie du pivot total consiste à choisir comme pivot, à chaque étape l, $(l = 1, \ldots, n)$, l'un des éléments $a_{ij}^{(l)}, \quad i, j = l, \ldots, n$ tel que

$$|a_{ij}^{(l)}| = \max_{l \leq p,q \leq n} |a_{pq}^{(l)}|$$

Exemple 4.3.
Soit à résoudre le système

$$\left(\begin{array}{ccc} 3 & 1 & 2 \\ 6 & 4 & 1 \\ 3 & 4 & 1 \end{array}\right)\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 4 \\ 10 \\ 7 \end{array}\right)$$

Comme une permutation éventuelle de colonnes entrainerait une permutation dans le vecteur solution, nous écrirons celui-ci à chaque étape en utilisant la notation suivante :

$$\left(\begin{array}{ccc|c|c} 3 & 1 & 2 & 4 & \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right] \\ 6 & 4 & 1 & 10 & \\ 3 & 4 & 1 & 7 & \end{array}\right)$$

les différentes étapes sont :

$$\underbrace{\left(\begin{array}{ccc|c|c} 2 & 1 & 3 & 4 & x_3 \\ 1 & 4 & 6 & 10 & x_2 \\ 1 & 4 & 3 & 7 & x_1 \end{array}\right)}_{\text{Stratégie de pivot}} \longrightarrow \underbrace{\left(\begin{array}{ccc|c|c} 10 & 1 & 3 & 4 & x_3 \\ 0 & 39/10 & 57/10 & 96/10 & x_2 \\ 0 & 39/10 & 27/10 & 66/10 & x_1 \end{array}\right)}_{\text{Méthode de Gauss}}$$

$$\underbrace{\left(\begin{array}{ccc|c|c} 10 & 3 & 1 & 4 & x_3 \\ 0 & 57/10 & 39/10 & 96/10 & x_1 \\ 0 & 27/10 & 39/10 & 66/10 & x_2 \end{array}\right)}_{\text{Stratégie de Pivot}} \longrightarrow \underbrace{\left(\begin{array}{ccc|c|c} 10 & 1 & 3 & 4 & x_3 \\ 0 & 57/10 & 39/10 & 96/10 & x_1 \\ 0 & 0 & 39/19 & 39/19 & x_2 \end{array}\right)}_{\text{Méthode de Gauss}}$$

La résolution du système

$$\left(\begin{array}{ccc} 10 & 1 & 3 \\ 0 & 57/10 & 39/10 \\ 0 & 0 & 39/19 \end{array}\right)\left(\begin{array}{c} x_3 \\ x_1 \\ x_2 \end{array}\right) = \left(\begin{array}{c} 4 \\ 96/10 \\ 39/19 \end{array}\right)$$

donne

$$\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 1 \\ 1 \\ 0 \end{array}\right)$$

5 Méthode de Gauss-Jordan

La méthode de Gauss-Jordan est une méthode directe qui repose sur le résultat suivant dû à Jordan

Théorème 5.1.
Soit une matrice A quelconque d'ordre n. Il existe une matrice M telle que $SA = D$ diagonale.

5.1 Principe de la méthode

Elle consiste à transformer le système Cramérien $Ax = b$ en un système $A'x = b'$ tel que A' est une matrice diagonale. L'objectif est donc de trouver les matrices $A^{(1)}, A^{(2)}, \ldots, A^{(n)}$

telles que la matrice $A^{(k)}$ $(k = 1, \ldots, n)$ ait la forme suivante :

$$A^{(k)} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & a_{1k}^{(k)} & \ldots & a_{1n}^{(k)} \\ 0 & 1 & 0 & \ldots & 0 & a_{2k}^{(k)} & \ldots & a_{2n}^{(k)} \\ 0 & 0 & 1 & \ldots & 0 & a_{3k}^{(k)} & \ldots & a_{3n}^{(k)} \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & 0 & 1 & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & a_{k,k}^{(k)} & \ldots & a_{k,n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & 0 & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & a_{n,k}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix}$$

où $a_{ij}^{(k)}$ représente le terme général de la matrice $A^{(k)}$.
On utilise le même procédé que pour la méthode de Gauss à la différence qu'à chaque étape, on transforme toutes les lignes sauf celle contenant le pivot. Rappelons que pour la méthode de Gauss, seules les lignes en-dessous du pivot sont transformées.

Remarque 5.1.
La méthode de Gauss-Jordan nécessite

$$n(n^2-1)/2 \quad \text{additions}$$
$$n(n^2-1)/2 \quad \text{multiplications}$$
$$n(n+1)/2 \quad \text{divisions}$$

L'algorithme de la méthode s'écrit :

$$\left.\begin{array}{l} a_{ij}^{(k+1)} = a_{ij}^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.a_{kj}^{(k)} \\ b_i^{(k+1)} = b_i^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.b_k^{(k)} \end{array}\right\} j = k, \ldots, n \Bigg\} i = 1, \ldots, n \quad (i \neq k) \Bigg\} k = 1, \ldots, n$$

La solution du système s'écrit alors

$$x_i = \frac{b_i^{(n)}}{a_{ii}^{(n)}} \qquad i = 1, \ldots, n$$

L'écriture matricielle de cet algorithme est comme suit

$$\left.\begin{array}{l} A_{i.}^{(k+1)} = A_{i.}^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.A_{k.}^{(k)} \\ b_i^{(k+1)} = b_i^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.b_k^{(k)} \end{array}\right\} k = 1, \ldots, n \quad \text{et} \quad i = 1, \ldots, k-1, k+1, \ldots, n$$

où $A_{i.}^{(k+1)}$désigne la $i^{\text{ème}}$ ligne de la matrice $A^{(k+1)}$.

Remarque 5.2.
Si, à chaque étape, on normalise le système en rendant le pivot égal à un, la solution du système sera égale au vecteur $b^{(n)}$.

Exemple 5.1.
Soit à résoudre le système suivant:

$$\begin{pmatrix} 2 & 3 & 1 \\ 4 & 1 & 3 \\ -2 & 3 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix}$$

En utilisant les notations de la méthode de Gauss, on aura les étapes suivantes :

$$\begin{pmatrix} 2 & 3 & 1 \\ 4 & 1 & 3 \\ -2 & 3 & -1 \end{pmatrix} \begin{matrix} 5 \\ 3 \\ 1 \end{matrix}$$

Première étape

$$\text{Normalisation} \longrightarrow \begin{pmatrix} 1 & 3/2 & 1/2 \\ 4 & 1 & 3 \\ -2 & 3 & -1 \end{pmatrix} \begin{matrix} 5/2 \\ 3 \\ 1 \end{matrix}$$

$$\text{Gauss-Jordan} \longrightarrow \begin{pmatrix} 1 & 3/2 & 1/2 \\ 0 & -5 & 1 \\ 0 & 6 & 0 \end{pmatrix} \begin{matrix} 5/2 \\ -7 \\ 6 \end{matrix}$$

Deuxième étape

$$\text{Normalisation} \longrightarrow \begin{pmatrix} 1 & 3/2 & 1/2 \\ 0 & 1 & -1/5 \\ 0 & 6 & 0 \end{pmatrix} \begin{matrix} 5/2 \\ 7/5 \\ 6 \end{matrix}$$

$$\text{Gauss-Jordan} \longrightarrow \begin{pmatrix} 1 & 0 & 4/5 \\ 0 & 1 & -1/5 \\ 0 & 0 & 6/5 \end{pmatrix} \begin{matrix} 2/5 \\ 7/5 \\ -12/5 \end{matrix}$$

Troisième étape

$$\text{Normalisation} \longrightarrow \begin{pmatrix} 1 & 0 & 4/5 \\ 0 & 1 & -1/5 \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} 2/5 \\ 7/5 \\ -2 \end{matrix}$$

$$\text{Gauss-Jordan} \longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} 2 \\ 1 \\ -2 \end{matrix}$$

La solution est donc

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}$$

5.2 Application : Calcul de la matrice inverse A^{-1}

Pour calculer l'inverse d'une matrice A que nous noterons B, il suffit de rappeler que

$$A.A^{-1} = A.B = I$$

où I signifie la matrice identité. On applique alors la même méthode à la matrice

$$\left(\begin{array}{cccc|cccc} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1 \end{array}\right)$$

comme le montre l'exemple suivant :

Exemple 5.2.
Supposons que l'on veuille inverser la matrice suivante

$$A = \begin{pmatrix} 1 & 3 & 3 \\ 2 & 2 & 0 \\ 3 & 2 & 6 \end{pmatrix}$$

Les différentes étapes sont

$$\text{Normalisation} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 3 & 2 & 6 & 0 & 0 & 1 \end{array}\right)$$

$$\text{Gauss-Jordan} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 0 & -4 & -6 & -2 & 1 & 0 \\ 0 & -7 & -3 & -3 & 0 & 1 \end{array}\right)$$

$$\text{Normalisation} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 0 & 1 & 3/2 & 1/2 & -1/4 & 0 \\ 0 & -7 & -3 & -3 & 0 & 1 \end{array}\right)$$

$$\text{Gauss-Jordan} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & -3/2 & -1/2 & 3/4 & 0 \\ 0 & 1 & 3/2 & 1/2 & -1/4 & 0 \\ 0 & 0 & 15/2 & 1/2 & -7/4 & 1 \end{array}\right)$$

$$\text{Normalisation} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & -3/2 & -1/2 & 3/4 & 0 \\ 0 & 1 & 3/2 & 1/2 & -1/4 & 0 \\ 0 & 0 & 1 & 1/15 & -7/30 & 2/15 \end{array}\right)$$

$$\text{Gauss-Jordan} \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -2/5 & 2/5 & 1/5 \\ 0 & 1 & 0 & 2/5 & 1/10 & -1/5 \\ 0 & 0 & 1 & 1/15 & -7/30 & 2/15 \end{array}\right)$$

La matrice inverse est donc

$$B = A^{-1} = \begin{pmatrix} -2/5 & 2/5 & 1/5 \\ 2/5 & 1/10 & -1/5 \\ 1/15 & -7/30 & 2/15 \end{pmatrix}$$

6 Décomposition LU

6.1 Introduction

Soit à résoudre le système linéaire $Ax = b$ où A est une matrice inversible de $\mathbb{M}_n(\mathbb{R})$. Si A peut s'écrire sous la forme $A = L.U$ où L est une matrice triangulaire inférieure (L provient de *Lower* en anglais) et U une matrice triangulaire supérieure (U provient de *Upper*), alors il suffit de résoudre successivement les deux systèmes triangulaires

$$Ly = b \qquad \text{et} \qquad Ux = y$$

Remarque 6.1.
a). Une matrice A n'est pas toujours décomposable sous la forme $L.U$ (on dit aussi qu'elle n'admet pas de décomposition LU). Soit par exemple la matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. En effet, en supposant

$$A = L.U = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix}.\begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}$$

On obtient

$$\begin{cases} l_{11}u_{11} = 0 \\ l_{11}u_{12} = 1 \\ l_{21}u_{11} = 1 \\ l_{21}l_{22} + l_{22}l_{22} = 0 \end{cases}$$

Ce qui est impossible.
b). Si A admet une décomposition LU alors elle admet une infinité de décompositions. En effet,

$$A = LU \Longrightarrow A = \left(\frac{1}{k}L\right).(kU) \qquad \forall k \in \mathbb{R}^\star$$

Exemple 6.1.

$$\begin{aligned} A = \begin{pmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 0 \end{pmatrix} &= \begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 6 & 0 & 3 \end{pmatrix}.\begin{pmatrix} 1 & -1/2 & -1/2 \\ 0 & 1 & -1/2 \\ 0 & 0 & 1 \end{pmatrix} = L_1U_1 \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}.\begin{pmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 0 & 0 & 3 \end{pmatrix} = L_2U_2 \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 0 & 1 \end{pmatrix}.\begin{pmatrix} 2 & -1 & -1 \\ 0 & -2 & 1 \\ 0 & 0 & 3 \end{pmatrix} = L_3U_3 \end{aligned}$$

6.2 Méthode de Crout

Parmi les différents choix de U, on prend celui où la matrice U ne comporte que des "1" sur la diagonale comme c'est le cas pour U_1 dans l'exemple précédent.

Détermination des l_{ij} et u_{ij}
Prenons le cas d'une matrice 4×4.

$$\begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \cdot \begin{pmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

Par identification, on obtient

$$l_{11} = a_{11} \quad , \quad l_{21} = a_{21} \quad , \quad l_{31} = a_{31} \quad , \quad l_{41} = a_{41}$$

(La première colonne de L est la même que celle de A).

$$u_{12} = \frac{a_{12}}{l_{11}} \quad , \quad u_{13} = \frac{a_{13}}{l_{11}} \quad , \quad u_{14} = \frac{a_{14}}{l_{11}} \quad ,$$

Dans cette méthode, on alterne entre une colonne de L et une ligne de U.
L'étape suivante consiste à trouver la deuxième colonne de L et la deuxième ligne de U.
Soit

$$\begin{aligned} l_{22} &= a_{22} - l_{21}u_{12} \\ l_{32} &= a_{32} - l_{31}u_{12} \\ l_{42} &= a_{42} - l_{41}u_{12} \end{aligned}$$

Remarque 6.2.
Si $l_{ii} \neq 0$, alors l_{ij} et u_{ij} sont determinés de manière unique.

Exemple 6.2.
Si on considère la matrice $A = \begin{pmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 0 \end{pmatrix}$, on obtient

$$L = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 6 & 0 & 0 \end{pmatrix} \qquad \text{et} \qquad U = \begin{pmatrix} 1 & -1/2 & -1/2 \\ 0 & 1 & -1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

6.3 Méthode de Doolittle

On procède comme pour la méthode de Crout sauf que la matrice A se décompose comme suit (pour une matrice 4×4):

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}$$

6.4 Méthode de Cholesky

6.4.1 Matrice définie positive

La forme quadratique $Q = \phi(x) = {}^t xAx$ est dite définie positive si :
$\phi(x) \geq 0, \quad \forall x$
$\phi(x) = 0 \Longleftrightarrow x = 0$
Si Q est définie positive, la matrice A est dite aussi définie positive.
La méthode de Cholesky s'appuie sur le résultat suivant appelé théorème de Cholesky:

Théorème 6.1.
Soit A une matrice symétrique définie positive. Alors, elle peut se mettre sous la forme $A = L.{}^tL$ où L est une matrice réelle triangulaire inférieure.

6.4.2 Factorisation

La matrice L a la forme suivante

$$ {}^tL = \begin{pmatrix} l_{11} & l_{12} & l_{13} & \dots & l_{1n} \\ 0 & l_{22} & l_{23} & \dots & l_{2n} \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & l_{n-1,n} \\ 0 & \dots & \dots & 0 & l_{nn} \end{pmatrix} $$

Si on multiplie L par ${}^tL$, puis on égalise avec A on obtient l'algorithme suivant :

$$ \left.\begin{array}{ll} l_{rr} = \left[a_{rr} - \sum_{k=1}^{r-1} l_{rk}^2\right] & \\ l_{jr} = \left[a_{rj} - \sum_{k=1}^{r-1} l_{rk}.l_{jk}\right] / l_{rr} & j = r+1, \dots, n \end{array}\right\} r = 1, \dots, n $$

Remarque 6.3.
Si $l_{ii} > 0 \qquad \forall i = 1, \dots, n$, alors la factorisation ci-dessus est unique.

6.5 Résolution de systèmes

Pour résoudre un système linéaire $Ax = b$ avec A symétrique définie et positive moyennant cette méthode, on procède comme suit :

- On factorise A sous la forme $L{}^tL$. Ce qui donne $Ax = L.({}^tLx) = b$.
- On pose $y = {}^tLx$ et on résoud $Ly = b$
- On résoud ${}^tLx = y$

Remarque 6.4.
Le nombre d'opérations est de

$$\frac{n^3+9n^2+2n}{6} \quad \text{Multiplications/divisions}$$

$$\frac{n^3+6n^2-7n}{6} \quad \text{additions/soustractions}$$

2. Si $A = L^tL$ avec L inversible, alors A est symétrique définie positive. En effet,

$$\forall x \in \mathbb{R}^n \text{ tel que } x \neq 0, {}^t x.A.x = {}^t x.{}^tL.L.x = {}^t (Lx).(Lx) = \|Lx\|^2 > 0$$

Exemple 6.3.
Soit à résoudre le système linéaire $Ax = b$ avec

$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 1 \\ 5 \\ 5 \end{pmatrix}$$

La factorisation de A donne $A = {}^tL.L$ avec

$$L = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

La résolution du système se fait alors en deux étapes:

Première étape:

$${}^tL.y = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix} . \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 5 \end{pmatrix}$$

La solution de ce premier système est :

$$y = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$$

Deuxième étape:

$$L.x = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = y = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$$

La solution finale est donc

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

7 Méthodes itératives

On considère le système $Ax = b$ où A est une matrice inversible de $\mathbb{M}_n(\mathbb{R})$. Le principe général des méthodes itératives consiste à construire une suite de vecteurs $(x_n)_n$ convergeant vers x. Pour cela, on choisit un vecteur initial $x^{(0)}$ et on modifie les composantes de $x^{(k)}$ dans un certain ordre pour obtenir $x^{(k+1)}$.

7.1 Méthode de Jacobi

7.1.1 Algorithme.

Soit $A = (a_{ij})_{i,j}$ tel que $a_{ii} \neq 0$, $i = 1, \ldots, n$. On considère la $i^{\text{ème}}$ équation de $Ax = b$ à savoir

$$a_{ii}x_i + \sum_{j=1, j\neq i}^{n} a_{ij}x_j = b_i$$

On définit $x^{(k+1)}$ à partir de $x^{(k)}$ par

$$a_{ii}x_i^{(k+1)} + \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)} = b_i \tag{5}$$

Autrement dit,

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

7.1.2 Interprétation matricielle

On définit les matrices D, E et F telles que $A = D - E - F$ avec

$$D = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & \ddots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{nn} \end{pmatrix}$$

$$-E = \begin{pmatrix} 0 & 0 & \ldots & \ldots & 0 \\ a_{21} & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \ddots & \ldots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{n,n-1} & 0 \end{pmatrix} \quad \text{et} \quad -F = \begin{pmatrix} 0 & a_{12} & \ldots & \ldots & a_{1n} \\ 0 & 0 & \ddots & \ldots & a_{2n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix}$$

On a alors,

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b \tag{6}$$

7.2 Méthode de Gauss-Seidel

7.2.1 Algorithme

On définit $x^{(k+1)}$ par

$$a_{ii}x_i^{(k+1)} + \sum_{j<i} a_{ij}x_j^{(k+1)} + \sum_{j>i} a_{ij}x_j^{(k)} = b_i \tag{7}$$

C'est-à-dire,

$$x^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right)$$

7.2.2 Interprétation matricielle

$$Dx^{(k+1)} - Ex^{(k+1)} - Fx^{(k)} = b$$

Par suite,

$$x^{(k+1)}(D-E) = Fx^{(k)} + b$$

Soit,

$$x^{(k+1)} = (D-E)^{-1}.Fx^{(k)} + (D-E)^{-1}b \tag{8}$$

7.3 Convergence

Soit l'équation

$$x^{(k+1)} = Mx^{(k)} + v \tag{11}$$

On veut savoir sous quelle condition la suite $(x^{(k)})_k$ converge-t-elle vers la solution de $Ax = b$. Il est à noter que les relations (6) et (8) sont de la forme (9) avec $M = D^{-1}(E+F)$, $v = D^{-1}b$ pour la méthode de Jacobi et $M = (D-E)^{-1}F$, $v = (D-E)^{-1}b$ pour celle de Gauss-Seidel.

Théorème 7.1.
Soit M une matrice de $\mathbb{M}_n(\mathbb{R})$, alors

$$\lim_{k\longrightarrow\infty} M^k = 0 \Longrightarrow \rho(M) < 1$$

où $\rho(M)$ désigne le rayon spectral de M.

Preuve
On considère le cas où M est diagonalisable. Ainsi, il existe une matrice P inversible telle que $M = PDP^{-1}$ où $D = Diag\{\lambda_i\}$ où λ_i est une valeur propre de M.

$$M = PDP^{-1} \Longleftrightarrow M^k = PD^kP^{-1}$$

et par suite,

$$\lim_{k\longrightarrow\infty} M^k = 0 \Longleftrightarrow \lim_{k\longrightarrow\infty} D^k = 0$$

soit

$$\begin{aligned}\lim_{k\longrightarrow\infty} D^k = 0 &\iff |\lambda_i| < 1 \quad \forall i \\ &\iff \rho(M) < 1\end{aligned}$$

Dans le cas où M est non diagonalisable, alors $M = PJP^{-1}$ où J est la matrice de Jordan. ■

7.3.1 Condition nécessaire et suffisante de convergence

Théorème 7.2.
Soit l'itération linéaire $x^{(k+1)} = Mx^{(k)} + v$. Alors,

$$\lim_{k\longrightarrow\infty} x^{(k)} = x, \forall x^{(0)} \iff \rho(M) < 1$$

où x représente la solution de $x = Mx + v$.

Preuve
1). On suppose que $\rho(M) < 1$.

$$x = Mx + v \iff (I - M)x = v$$

Soit λ_i une valeur propre de M :$Mu_i = \lambda_i u_i$. Ainsi,

$$(I - M)u_i = (1 - \lambda_i)u_i$$

et les valeurs propres de $(I - M)$ sont de la forme $1 - \lambda_i$.
Comme $\rho(M) < 1$, donc $|\lambda_i| < 1$, il s'ensuit que $1 - \lambda_i \neq 0\ \forall i$ et par suite, le système $(I - M)x = v$ admet une solution unique.
Soit $x^{(k+1)} = Mx^{(k)} + v.$, montrons que $\lim_{k\longrightarrow\infty} x^{(k)} = x$.
$x^{(k+1)} = Mx^{(k)} + v$ et $x = Mx + v$ donnent

$$\begin{aligned}x^{(k+1)} - x &= (Mx^{(k)} + v) - (Mx + v) \\ &= M(Mx^{(k-1)} + v)) = M^2(x^{(k-1)} - x) \\ &= \dots \\ &= M^{(k+1)}(x^{(0)} - x)\end{aligned}$$

On a alors,

$$\rho(M) < 1 \Longrightarrow \lim_{k\longrightarrow\infty} = 0$$

et donc, $\lim_{k\longrightarrow\infty} x^{(k+1)} = x$.
On a donc la convergence quelle que soit la valeur du vecteur initial $x^{(0)}$.
2). On suppose que $\lim_{k\longrightarrow\infty} x^{(k+1)} = x$ où $x^{(k+1)} = Mx^{(k)} + v$. On veut montrer que $\rho(M) < 1$.

$$\lim_{k\longrightarrow\infty} x^{(k+1)} = x \Longrightarrow \lim_{k\longrightarrow\infty} x^{(k+1)} = \lim_{k\longrightarrow\infty} Mx^{(k)} + v$$

Soit $x = Mx + v$.

$$\begin{aligned}\lim_{k\longrightarrow\infty}(x^{(k+1)} - x) &= \lim_{k\longrightarrow\infty} M(x^{(k)} - x)\\ &= \lim_{k\longrightarrow\infty} M^{k+1}(x^{(0)} - x)\\ &= 0\end{aligned}$$

Donc, $\lim_{k\longrightarrow\infty} M^{k+1} = 0$ et par conséquent, d'après le théorème précédent, on a $\rho(M) < 1$.

■

Remarque 7.1.
La condition $\rho(M) < 1$ est difficilement utilisable en pratique sauf si on utilise un logiciel approprié (exemple : Matlab). C'est la raison pour laquelle on essaye de trouver des conditions beaucoup plus faciles à appliquer.

7.3.2 Condition suffisante de convergence

Définition 7.1.
On dit qu'une matrice A de $\mathbb{M}_n(\mathbb{R})$ est à diagonale strictement dominante si

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \qquad \forall i = 1, \ldots, n$$

Exemple 7.1.
La matrice

$$A = \begin{pmatrix} -5 & 1 & -3 \\ 2 & 4 & 0 \\ -1 & -2 & 5 \end{pmatrix}$$

est à diagonale strictement dominante.

Théorème 7.3.
Soit $A = (a_{ij})_{i,j}$ une matrice inversible de $\mathbb{M}_n(\mathbb{R})$. Si A est à diagonale strictement dominante, alors les méthodes de Jacobi et de Gauss-Seidel convergent pour tout vecteur initial vers la solution du système de matrice A.

Preuve
(i). Méthode de Jacobi.

On considère l'equation $Jx = \lambda x$ où $J = D^{-1}.(E + F)$.

$$\begin{aligned}Jx = \lambda x &\iff D^{-1}.(E + F)x = \lambda x\\ &\iff (E + F)x = \lambda D x\\ &\iff \sum_{\substack{j=1\\ j\neq i}} -(a_{ij}x_j) = \lambda a_{ii}x_i \forall i = 1, \ldots, n \qquad (1)\end{aligned}$$

Soit k tel que $|x_k| = \max |x_i|$ ($|x_k|$ est strictement positif car x est un vecteur propre). On considère alors la $k^{\text{ème}}$ itération de (1) et on aura

$$\sum_{\substack{j=1\\ j\neq i}} -(a_{ij}x_j) = \lambda a_{kk}x_k \Longrightarrow |\lambda|.|a_{kk}| = \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|.\frac{|x_j|}{|x_k|}$$

Donc,

$$|\lambda|.|a_{kk}| \leq \sum_{\substack{j=1\\ j\neq k}} n|a_{kj}| \Longrightarrow |\lambda| \leq \frac{\sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|}{|a_{kk}|} < 1 \Longrightarrow \rho(J) < 1$$

D'où la convergence de la méthode de Jacobi.

(ii). Méthode de Gauss-Seidel.

On considère l'équation $Gx = \lambda x$ avec $G = (D-E)^{-1}.F$. On obtient

$$\lambda.\left(a_{kk} + \sum_{j=1}^{k-1} a_{kj}.\frac{x_j}{x_k}\right) = \sum_{j=k+1}^{n} a_{kj}.\frac{x_j}{x_k}$$

et donc,

$$|\lambda| \leq \frac{\displaystyle\sum_{j=k+1}^{n} |a_{kj}|}{\left|a_{kk} + \displaystyle\sum_{j=1}^{k-1} a_{kj}.\frac{x_j}{x_k}\right|} \leq \frac{\displaystyle\sum_{j=k+1}^{n} |a_{kj}|}{|a_{kk}| - \displaystyle\sum_{j=1}^{k-1} |a_{kj}|} < 1$$

Ceci implique que $\rho(G) < 1$ et donc la convergence de la méthode de Gauss-Seidel . ■

7.3.3 Majoration d'erreur pour la méthode de Jacobi

Pour résoudre un système $Ax = b$, considérons la suite itérative du type $x^{(k+1)} = J.x^{(k)} + v$ avec J étant la matrice de Jacobi ($J = D^{-1}(E+F)$ et $v = D^{-1}b$). Alors,

$$\|x - x^{(n)}\| \leq \frac{\|J\|^n}{1-\|J\|}.\|x^{(1)} - x^{(0)}\|$$

avec $\|J\| < 1$. En effet,

Soit $X^{(k+1)} = J.x^{(k)} + c$. Alors,

$$x^{(2)} - x^{(1)} = J.(x^{(1)} - x^{(0)}) \Longrightarrow \|x^{(2)} - x^{(1)}\| = \|J\|.\|x^{(1)} - x^{(0)}\|$$

De même,

$$x^{(3)} - x^{(2)} = J.(x^{(2)} - x^{(1)}) \Longrightarrow \|x^{(3)} - x^{(2)}\| = \|J\|^2.\|x^{(1)} - x^{(0)}\|$$

Ainsi de suite, on aura

$$\|x^{(n+1)} - x^{(n)}\| = \|J\|^n.\|x^{(1)} - x^{(0)}\|$$

De plus,

$$x^{(n+p)} - x^{(n)} = \sum_{i=1}^{p} x^{(n+i)} - x^{(n+i-1)}$$

$$\Longrightarrow \|x^{(n+p)} - x^{(n)}\| = \sum_{i=1}^{p} \|x^{(n+i)} - x^{(n+i-1)}\|$$

$$\Longrightarrow \quad \|x^{(n+p)} - x^{(n)}\| = \|J\|^n.\|x^{(1)} - x^{(0)}\|.\sum_{i=1}^{p} \|J\|^{i-1}$$

$$\Longrightarrow \lim_{p\longrightarrow\infty} \|x^{(n+p)} - x^{(n)}\| = \|x - x^{(n)}\| = \frac{\|J\|^n}{1-\|J\|}.\|x^{(1)} - x^{(0)}\| \qquad \text{car } \|J\| < 1$$

Exemple 7.2.
On considère le système linéaire $Ax = b$ où

$$A = \begin{pmatrix} 5 & 2 & -1 \\ 1 & 5 & 2 \\ 3 & 0 & 5 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 3 \\ 4 \\ 4 \end{pmatrix}$$

La méthode de jacobi converge car la matrice A du système est à diagonale strictement dominante. On a

$$J = \frac{1}{5}\begin{pmatrix} 0 & -2 & 1 \\ -1 & 0 & -2 \\ -3 & 0 & 0 \end{pmatrix}$$

D'où,

$$\|J\|_1 = 3/5 < 1$$

Prenons $x^{(0)} = \begin{pmatrix} 3/5 \\ 4/5 \\ 4/5 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.8 \\ 0.8 \end{pmatrix}$. Alors, on aura $x^{(1)} = \begin{pmatrix} -0.16 \\ -0.44 \\ -0.36 \end{pmatrix}$ et $x^{(2)} = \begin{pmatrix} 0.104 \\ 0.176 \\ 0.096 \end{pmatrix}$

Remarque 7.2.
La stricte dominance est une condition suffisante pour garantir la convergence. Cependant, si la matrice est seulement à diagonale dominante, alors la condition n'est plus suffisante pour assurer la convergence comme le montre l'exemple suivant : supposons que l'on veuille résoudre les systèmes $Ax = b$ avec

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}$$

la matrice A est dominante seulement. En appliquant la méthode de Jacobi, avec $x^{(0)} =^t (0,0,0)$, on obtient $x^{(1)} =^t (2,2,2)$, $x^{(2)} =^t (0,0,0)$,... soit un cycle d'ordre 2.

Comparaison des méthodes itératives.

Dans ce qui suit, on donne des exemples retraçant les différents cas de figures.

a. Divergence simultanée des deux méthodes

Soit à résoudre le système linéaire $Ax = b$ avec

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

Soient

$$J_A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad \text{et} \qquad G_A = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$$

les matrices de correction M correspondant aux méthodes de Jacobi et de Gauss-Seidel respectivement. Dans ce cas, $\rho(J_A) = \rho(G_A) = 1$.

b. Convergence simultanée des deux méthodes

Soit à résoudre le système $Ax = b$ avec

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix} \text{ et } b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

(i). Méthode de Jacobi

La matrice est à diagonale strictement dominante. Donc, la méthode de Jacobi converge quelque soit le vecteur initial $x^{(0)}$ choisi. Le calcul donne

$$J_A = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 \end{pmatrix} \text{ et } v = \begin{pmatrix} 0 \\ 1/3 \\ 0 \end{pmatrix}$$

Dans ce qui suit, on donne quelques itérés obtenus avec $x^{(0)}$ égal au vecteur nul

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(5)}$	$x^{(9)}$	$x^{(10)}$	$x^{(15)}$	$x^{(16)}$	$x^{(17)}$	$x^{(18)}$
0	0	0.1667	0.2222	0.2469	0.2490	0.2499	0.2500	0.2500	0.2500
0	0.3333	0.3333	0.4815	0.4979	0.4979	0.4999	0.4999	0.5000	0.5000
0	0	0.1667	0.2222	0.2469	0.2490	0.2499	0.2500	0.2500	0.2500

La solution est atteinte après 17 itérations au sens de précision maximale du format numérique utilisé ($x^{(17)} = x^{(18)} = ...$).

(ii).**Méthode de Gauss-Seidel**

Si on applique la méthode de Gauss-Seidel à l'exemple ci-dessus, on peut dire aussi que la méthode converge car la matrice A est à diagonale strictement dominante. En prenant $x^{(0)}$ égal au vecteur nul, on obtient

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$	$x^{(9)}$
0	0.1667	0.2222	0.2469	0.2490	0.2497	0.2499	0.2500
0	0.4444	0.4815	0.4979	0.4993	0.4998	0.4999	0.5000
0	0.2222	0.2407	0.2490	0.2497	0.2499	0.2500	0.2500

Avec le même vecteur initial que la méthode de Jacobi, la solution est atteinte après seulement 9 itérations.

c. Convergence de la méthode de Jacobi et divergence de la méthode de Gauss-Seidel

Soit le système $Ax = b$ avec

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

Dans ce cas, on trouve $\rho(J_A) = 0 < 1$ et $\rho(G_A) = 2 > 1$.
En prenant $x^{(0)} =^t (1, 0, 0)$, on trouve $x^{(1)} =^t (1, 2, 3)$, $x^{(2)} =^t (3, -1, -1)$, $x^{(3)} =^t (1, 1, 1)$, $x^{(4)} =^t (1, 1, 1)$,etc.

d. Convergence de la méthode de Gauss-Seidel et divergence de la méthode de Jacobi

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 2 \\ 6 \\ 0 \end{pmatrix}$$

Dans ce cas, on trouve $\rho(J_A) = \sqrt{5}/2 > 1$ et $\rho(G_A) = 0.5 < 1$

8 Exercices résolus

8.1

On considère le système linéaire $Ax = b$ où A est une matrice inversible de $\mathbb{M}_n(\mathbb{R})$ et b un vecteur de $\mathbb{R}^n$.

1. Supposons que la matrice A est connue exactement mais que le vecteur b l'est avec une certaine incertitude. On considère alors le système $Ax^* = b + \Delta b$ où Δb est un vecteur dont les composantes sont relativement petites devant celles de b. Posons $\Delta x = x^* - x$. Montrer qu'il existe une constante $C(A) > 0$ ne dépendant pas de A, telle que
$$\frac{\|\Delta x\|}{\|x\|} \leq C(A).\frac{\|\Delta b\|}{\|b\|} \tag{1}$$
2. Soit le système linéaire
$$\begin{cases} 11x + 3.3y = 1 \\ 3.3x + y = 2 \end{cases} \tag{2}$$
On suppose que l'on a obtenu une approximation x^* de la solution x de (2) vérifiant
$$\|Ax^* - Ax\| \leq 10^{-3}$$
Peut-on dire que x^* est une "bonne" approximation de x ? Justifiez votre réponse.

Solution.
1. On a
$$Ax = b \qquad \text{et} \qquad Ax^* = b + \delta b$$
$$Ax^* = A.(x + \delta x) = b + \delta b \Longrightarrow Ax + A\delta x = b + \delta b$$
Comme $Ax = b$, alors, $A\delta x = \delta b$. Ce qui donne
$$\delta x = A^{-1}.\delta b \Longrightarrow \|\delta x\| \leq \|A^{-1}\|.\|\delta b\| \tag{i}$$
De plus, $b = Ax \Longrightarrow \|b\| \leq \|A\|.\|x\|$ d'où,
$$\|x\| \geq \frac{\|b\|}{\|A\|} \tag{ii}$$
(i) et (ii) donnent
$$\frac{\|\delta x\|}{\|x\|} \leq Cond(A).\frac{\|\delta b\|}{\|b\|}$$
2. Soit x^* la solution approchée du système. On a
$$r = Ax - Ax^* = A.(x - x^*) \Longrightarrow x - x^* = A^{-1}.r$$
$$\|x - x^*\| = \|A^{-1}.r\| \leq \|A^{-1}\|.\|r\| \Longrightarrow \|x - x^*\| \leq \|A^{-1}\|.10^{-3}$$
On ne peut pas dire que x^* est ou n'est pas une bonne approximation de x car si $\|A^{-1}\|$ est grande, la majoration devient insignifiante.

8.2

On considère le système $Ax = b$ où A est une matrice inversible de $\mathbb{M}_n(\mathbb{R})$. Soit x^* une solution approchée de x. Posons $r = Ax - Ax^*$ et $e = x - x^*$.

1. Démontrer que
$$\frac{\|r\|}{\|A\|} \leq \|e\| \leq \|A^{-1}\|.\|r\|$$
et que
$$\frac{\|b\|}{\|A\|} \leq \|x\| \leq \|A^{-1}\|.\|b\|$$
2. En déduire que
$$\frac{1}{Cond(A)}.\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq Cond(A).\frac{\|r\|}{\|b\|}$$

Solution.
1. $r = Ae \Longrightarrow \|r\| \leq \|A\|.\|e\|$. D'où,
$$\frac{\|r\|}{\|A\|} \leq \|e\|$$
De plus,
$$e = A^{-1}.r \Longrightarrow \|e\| \leq \|A^{-1}.\|r\|$$
d'où,
$$\frac{\|r\|}{\|A\|} \leq \|e\| \leq \|A^{-1}\|.\|r\| \quad (1)$$
Même raisonnement pour montrer que
$$\frac{b\|}{\|A\|} \leq \|x\| \leq \|A^{-1}\|.\|b\| \quad (2)$$

2. (1) et (2) impliquent
$$\frac{\|r\|}{\|A\|.\|A^{-1}\|.\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \frac{\|A^{-1}\|.\|r\|}{\|b\|/\|A\|}$$
Soit
$$\frac{1}{Cond(A)}\frac{\|e\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq Cond(A).\frac{\|r\|}{\|b\|}$$

8.3

Résoudre les deux systèmes linéaires suivants:
$$\begin{cases} 2x + 1.4y = 1.4 \\ 1.4x + y = 1 \end{cases} \text{ et } \begin{cases} 2x + 1.4y = 1.44 \\ 1.4x + y = 1 \end{cases}$$
Commentez vos résultats.

Solution.
La premier système a pour solution $x = 0$ et $y = 1$. Quant au deuxième, sa solution est $x = 1$ et $y = -0.4$.
Commentaire. Les deux systèmes ne diffèrent que par le second membre de la première équation. Cette différence est égale à 0.04. Cependant, le système est très sensible à ce léger changement. Ce qui le rend mal conditionné.

8.4

Soit

$$H = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix}$$

et

$$b = \begin{pmatrix} 25/12 \\ 77/60 \\ 57/60 \\ 319/420 \end{pmatrix}$$

1. En utilisant 3 chiffres significatifs, résoudre le système $Hx = b$.
2. Trouver la solution de $Hx = b$ en utilisant 5 chiffres significatifs.
3. Que peut-on conclure sur le conditionnement de la matrice H ?
4. Sachant que la solution exacte est $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, peut-on avoir un ordre de grandeur de $\det(H)$?

Remarque: H s'appelle matrice de Hilbert d'ordre 4.

Solution.
1. La solution du système (avec 3 chiffres significatifs) est

$$x = \begin{pmatrix} 0.99 \\ 1.42 \\ -0.43 \\ 2.10 \end{pmatrix}$$

2. Avec 5 chiffres significatifs, la solution est

$$x = \begin{pmatrix} 1.0000 \\ 0.9995 \\ 1.0017 \\ 0.9990 \end{pmatrix}$$

3. Vu que les solutions ci-dessus sont très différentes, ce qui montre l'influence considérable de l'erreur d'arrondi, la matrice H doit sûrement être mal conditionnée.
4. La solution exacte est $x = {}^t(1,1,1,1)$. Comme la solution obtenue avec 3 chiffres significatifs est loin de x, on peut dire que $\det(A) < 10^{-3}$. D'ailleurs, le déterminant de A vaut 1.65×10^{-5}.

8.5

Résoudre le système $Ax = b$ avec

$$A = \begin{pmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{pmatrix} \text{ et } b = \begin{pmatrix} 12 \\ 11 \\ 2 \end{pmatrix}$$

sachant que A peut se mettre sous la forme $A = L.U$ avec

$$L = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 7/3 & 0 \\ 2 & -4/3 & -1 \end{pmatrix} \text{ et } U = \begin{pmatrix} 1 & -1/3 & 2/3 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

Solution.

$$Ax = b \Longleftrightarrow L.U.x = b$$

Comme L est inversible, on a

$$L^{-1}.(L.Ux) = L^{-1}.b \longleftrightarrow Ux = L^{-1}b$$

Posons $y = L^{-1}b$. Alors,

$$Ly = b \Longrightarrow y = \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix}$$

Pour trouver x, il suffit de résoudre $Ux = y$ duquel on obtient

$$x = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

8.6

En utilisant la méthode de Gauss, trouver l'inverse de la matrice

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$$

Solution.
On cherche A^{-1} telle que $A.A^{-1} = I$. On pose

$$A^{-1} = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix} ; \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad ; \quad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad ; \quad z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Le calcul de A^{-1} se ramène à la résolution des trois systèmes suivants

$$Ax = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad (1) \quad ; \quad Ay = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad (2) \quad ; \quad Az = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (3) \quad ;$$

L'application de la méthode de Gauss aux trois systèmes (1), (2) et (3) conduit respectivement à la résolution des trois systèmes triangulaires suivants

$$A'x = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} \quad ; \quad A'y = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \quad ; \quad A'z = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

avec A' la matrice triangulaire supérieure suivante

$$A' = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

Les solution sont

$$x = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \quad ; \quad y = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \quad ; \quad z = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

Ce qui donne

$$A^{-1} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 1 & 1 & -1 \end{pmatrix}$$

8.7

Soit un système linéaire $Ax = b$ de matrice

$$A = \begin{pmatrix} 2 & \alpha & 0 \\ \alpha & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

Peut-on trouver les valeurs de α pour lesquelles la méthode de Jacobi diverge ? Même question pour la méthode de Gauss Seidel.

Solution.
Avec la méthode de Jacobi, la matrice de correction est

$$B = \begin{pmatrix} 0 & -\alpha/2 & 0 \\ -\alpha & 0 & -1 \\ 0 & -1/2 & 1 \end{pmatrix}$$

Les valeurs propres de B sont $\lambda_1 = 0$ et $\lambda_2 = -\lambda_3 = \pm\sqrt{\frac{\alpha+1}{2}}$. La convergence a lieu si $-1 < \alpha < 1$.
Avec la méthode de Gauss-Seidel, la matrice de correction est

$$B = \begin{pmatrix} 0 & -2\alpha & 0 \\ 0 & 2\alpha^2 & -4 \\ 0 & -\alpha^2 & 2 \end{pmatrix}$$

Les valeurs propres de B sont $\lambda_1 = 0$ et $\lambda_2 = -\lambda_3 = \pm\sqrt{\frac{\alpha+1}{2}}$. La convergence a lieu si $-1 < \alpha < 1$.
En conclusion, les deux méthodes convergent ou divergent en même temps.

8.8

On considère le système linéaire $Ax = b$ où

$$A = \begin{pmatrix} 5 & 2 & -1 \\ 1 & 5 & 2 \\ 3 & 0 & 5 \end{pmatrix}$$

et

$$b = \begin{pmatrix} 3 \\ 4 \\ 4 \end{pmatrix}$$

1. Montrer que l'algorithme de Jacobi converge.

2. Soit $B = (b_{ij})$ une matrice carrée d'ordre n $(1 \leq n, 1 \leq j \leq n)$. On définit par $\|B\|$ par

$$\|B\| = \max_i \sum_{j=1}^{n} |b_{ij}|$$

Soit J la matrice de Jacobi relative au problème donné.

a. Montrer que $\|J\| < 1$.

b. En considérant $\|x^{(n+p)} - x^{(n)}\|$. Montrer que

$$\|x - x^{(n)}\| \leq \frac{\|J\|}{1 - \|J\|}.\|x^{(1)} - x^{(0)}\|.$$

c. Soit $x^{(0)} = D^{-1}.b$ où

$$D = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Montrer que

$$\|x - x^{(n)}\| \leq \frac{\|J\|^{n+1}}{1 - \|J\|}.\|D^{-1}.b\|$$

Solution.

1. La méthode de Jacobi converge car la matrice du système est à diagonale strictement dominante.

2.

a. On a

$$J = \frac{1}{5}\begin{pmatrix} 0 & -2 & 1 \\ -1 & 0 & -2 \\ -3 & 0 & 0 \end{pmatrix}$$

D'où,

$$\|J\| = 3/5 < 1$$

b. Soit $X^{(k+1)} = J.x^{(k)} + c$. Alors,

$$x^{(2)} - x^{(1)} = J.(x^{(1)} - x^{(0)}) \Longrightarrow \|x^{(2)} - x^{(1)}\| = \|J\|.\|x^{(1)} - x^{(0)}\|$$

De même,

$$x^{(3)} - x^{(2)} = J.(x^{(2)} - x^{(1)}) \Longrightarrow \|x^{(3)} - x^{(2)}\| = \|J\|^2.\|x^{(1)} - x^{(0)}\|$$

Ainsi de suite, on aura

$$\|x^{(n+1)} - x^{(n)}\| = \|J\|^n.\|x^{(1)} - x^{(0)}\|$$

De plus,

$$x^{(n+p)} - x^{(n)} = \sum_{i=1}^{p} x^{(n+i)} - x^{(n+i-1)}$$

$$\Longrightarrow \|x^{(n+p)} - x^{(n)}\| = \sum_{i=1}^{p} \|x^{(n+i)} - x^{(n+i-1)}\|$$

$$\Longrightarrow \|x^{(n+p)} - x^{(n)}\| = \|J\|^n.\|x^{(1)} - x^{(0)}\|.\sum_{i=1}^{p} \|J\|^{i-1}$$

$$\Longrightarrow \|x^{(n+p)} - x^{(n)}\| = \frac{\|J\|^n}{1-\|J\|}.\|x^{(1)} - x^{(0)}\| \qquad \text{car } \|J\| < 1$$

c. $x^{(0)} = D^{-1}.b \Longrightarrow x^{(1)} = Jx^{(0)} + D^{-1}b \Longrightarrow x^{(1)} - x^{(0)} = J.D^{-1}b.$
$\|x^{(1)} - x^{(0)}\| \leq \|J\|.\|D^{-1}b\|$.D'où

$$\|x - x^{(n)}\| = \frac{\|J\|^{n+1}}{1-\|J\|}.\|D^{-1}b\|$$

8.9

Utiliser la méthode de Gauss pour résoudre le système suivant :

$$\begin{cases} 3x + 4y - 7z = -7 \\ x - 2y + z = 1 \end{cases}$$

Solution.
Le système à résoudre est

$$\begin{pmatrix} 3 & 4 & -7 \\ 1 & -2 & 1 \end{pmatrix} . \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -7 \\ 1 \end{pmatrix} \tag{1}$$

$$(1) \Longleftrightarrow \begin{pmatrix} 3 & 4 & -7 \\ 0 & 10 & -10 \end{pmatrix} . \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -7 \\ -10 \end{pmatrix}$$

Soit $10y - 10z = -10$ et donc, $z = y + 1$. La première équation donne $x = y$. D'où la solution du problème qui est : $u = (x, x, x+1)$ avec x quelconque dans $\mathbb{R}$.
Une autre possibilité est de transformer la matrice rectangulaire en une matrice carrée (ajouter une ligne obtenue par une combinaison linéaire des deux autres lignes). Par exemple,

$$(1) \Longleftrightarrow \begin{pmatrix} 3 & 4 & -7 \\ 1 & -2 & 1 \\ 4 & 2 & -6 \end{pmatrix} . \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -7 \\ 1 \\ -6 \end{pmatrix}$$

$$(1) \Longleftrightarrow \begin{pmatrix} 3 & 4 & -7 \\ 0 & 10 & -10 \\ 0 & 5/2 & -5/2 \end{pmatrix} . \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -7 \\ -10 \\ -5/2 \end{pmatrix}$$

$$(1) \Longleftrightarrow \begin{pmatrix} 3 & 4 & -7 \\ 0 & 10 & -10 \\ 0 & 0 & 0 \end{pmatrix} . \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -7 \\ -100 \end{pmatrix}$$

8.10

Soient les matrices

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ et } B = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

1. Calculer $\rho(J_A)$, $\rho(J_B)$, $\rho(G_A)$ et $\rho(G_B)$ où la notation $\rho(J_A)$ désigne le rayon spectral de la matrice de Jacobi J associée à la matrice A et $\rho(G_B)$ représente le rayon spectral de la matrice de Gauss -Seidel G associée à la matrice B.

2. Discuter les résultats précédents.

Solution.
1. Calcul des rayons spectraux

$$J_A = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} \quad \text{et} \quad G_A = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}$$

Ce qui nous donne $\rho(J_A) = 0$ et $\rho(G_A) = 2$.
De même, on a

$$J_B = \frac{1}{2}\begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{pmatrix} \quad \text{et} \quad G_B = \begin{pmatrix} 0 & 1/2 & -1/2 \\ 0 & -1/2 & -1/2 \\ 0 & 0 & -1/2 \end{pmatrix}$$

Ce qui donne $\rho(J_B) = \sqrt{5}/2$ et $\rho(G_B) = 1/2$.

2. D'après les résultats ci-dessus, on peut dire que

$$\rho(J_A) < 1 < \rho(G_A) \tag{1}$$

et

$$\rho(G_B) < 1 < \rho(J_B) \tag{2}$$

(1) signifie que la méthode de Jacobi appliquée à A converge tandis que celle de Gauss-Seidel diverge.
(2) signifie que la méthode de Jacobi appliquée à B diverge tandis que celle de Gauss-Seidel converge.
Ces deux exemples montrent que l'on ne peut rien dire de la comparaison des deux méthodes.

8.11

Soit le système $Ax = b$ avec :

$$A = \begin{pmatrix} 10 & 1 & 1 \\ 1 & 10 & 1 \\ 1 & 1 & 10 \end{pmatrix} \text{ et } b = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

Trouver les premières valeurs de $x^{(k)}$ par la méthode de Jacobi, en prenant pour vecteur initial $x^{(0)} = {}^t(1, 1, 1)$.

Solution.
L'algorithme de Jacobi est convergent car la matrice A est à diagonale strictement dominante. On a

$$J = D^{-1}(E+F) = \begin{pmatrix} 0 & -0.1 & -0.1 \\ -0.1 & 0 & -0.1 \\ -0.1 & -0.1 & 0 \end{pmatrix} \qquad \text{et} \qquad D^{-1}b = \begin{pmatrix} 0.6 \\ 0.6 \\ 0.6 \end{pmatrix}$$

Si on prend $x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, on aura les résultats suivants

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$
1	0.4	0.52	0.496	0.5008	0.49984	0.50032
1	0.4	0.52	0.496	0.5008	0.49984	0.50032
1	0.4	0.52	0.496	0.5008	0.49984	0.50032

8.12

Soit le système

$$\begin{cases} x + y = 2 \\ x - y = 0 \end{cases} \tag{1}$$

Peut-on appliquer les algorithmes de Jacobi et de Gauss-Seidel pour obtenir la solution de (1) ?

Solution.
Posons $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. On a

$$D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad ; \quad -E = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad ; \quad -E = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

La matrice de Jacobi est

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

Les valeurs propres sont $\lambda_{1,2} = \pm i$. Ainsi, $\rho(J) = 1$. Par conséquent, le procédé de Jacobi diverge.
La matrice de Gauss-Seidel est

$$G = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$$

Les valeurs propres sont $\lambda - 1 = 0$ et $\lambda_2 = -1$. Par conséquent, le procédé de Gauss-Seidel diverge aussi.

CHAPITRE III. Calcul des valeurs et des vecteurs propres

1 Introduction.

Pour calculer des approximations de l'ensemble des valeurs propres d'une matrice A, une idée couramment exploitée consiste à construire une suite de matrices $(P_k)_{k\geq 1}$ telle que les matrices $P_k^{-1}AP_k$ "convergent" (dans un sens à préciser: en effet, il ne s'agit pas toujours d'une véritable convergence) vers une matrice de valeurs propres connues, c'est-à-dire diagonale ou triangulaire.
Cette idée est à la base de la méthode de Jacobi pour les matrices symétriques, où les matrices P_k sont des produits de matrices orthogonales "élémentaires" très simples à construire. On peut alors montrer que

$$\lim_{k\to\infty} P_k^{-1}AP_k = diag(\lambda_i)$$

où les nombres λ_i sont les valeurs propres de la matrice A (à une permutation près). Lorsque ces dernières sont toutes distinctes, on établit que les vecteurs colonnes des matrices P_k constituent de surcroît des approximations des vecteurs propres de la matrice A.
Pour des matrices quelconques, la remarquable méthode QR, dont le principe sera décrit dans ce chapitre, relève de la même idée. Utilisant à chaque itération de la méthode la factorisation QR des matrices, on construit une suite de matrices (P_k) telle que, moyennant certaines hypothèses assez restrictives,

$$\lim_{k\to\infty} (P_k^{-1}AP_k)_{ij} = 0 \quad \text{pour} \quad j < i,$$

$$\lim_{k\to\infty} (P_k^{-1}AP_k)_{ii} = \lambda_i,$$

les scalaires λ_i étant les valeurs propres de la matrice A.
On note généralement l'absence de méthodes de calculs de l'ensemble des valeurs propres des matrices quelconques à partir du polynôme caractéristique. C'est d'ailleurs exactement l'inverse qui se produit: pour calculer l'ensemble des racines d'un polynôme de degré élevé, il est en effet courant d'appliquer la méthode QR.
D'autres méthodes permettent de calculer seulement certaines valeurs propres sélectionnées. C'est le cas notamment de la méthode de Givens-Householder que nous étudierons, et qui s'applique aux matrices symétriques: on commence par réduire une telle matrice à la forme tridiagonale à l'aide de matrices de Householder, puis un ingénieux procédé de Givens permet le calcul approché, avec une précision arbitraire, d'une valeur propre de rang donné d'une telle matrice.

Soit A une matrice de $\mathbb{M}_n(\mathbb{R})$. Ce chapitre s'intéresse donc au calcul approché des valeurs propres et des vecteurs propres de A.

D'un point de vue pratique, l'information nécessaire sur les éléments propres d'une matrice varie selon le cas. Ainsi, par exemple
• La résolution de certains problèmes de mécanique, de chimie, .. exige la connaissance de toutes les valeurs propres et parfois de tous les vecteurs propres.
• Dans certains problèmes, on demande de déterminer la distance séparant un nombre donné du spectre de la matrice.
La diversité des problèmes et de l'information nécessaire sur les éléments propres a pour conséquence la création d'une multitude de modes opératoires (méthodes numériques).
On peut dire, de manière générale, que les différentes méthodes numériques de calcul de valeurs propres et de vecteurs propres répondent surtout aux grandes difficultés d'utilisation pratique du polynôme caractéristique de la matrice particulièrement pour de grandes matrices.

2 Théorèmes de localisation de valeurs propres

2.1 Théorème de Gershgorin

Théorème 2.1.
Soit A une matrice d'ordre $n \times n$ (avec des éléments dans $\mathbb{R}$ ou dans $\mathbb{C}$). Si λ est une valeur propre de A, alors il existe un indice i tel que

$$|\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

c'est à dire que toutes les valeurs propres de A se trouvent dans l'union des disques

$$D_i = \left\{ \lambda, |\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}$$

Preuve
Soit $v \neq 0$ un vecteur propre et choisissons l'indice i tel que $|v_i| \geq |v_j|$ pour tout j. La ligne i de l'équation $Av = \lambda v$ donne

$$\sum_{j\neq i}^{n} |a_{ij} v_j| = (\lambda - a_{ii})_{v_i}$$

En divisant par v_i et en utilisant l'inégalité de triangle, on obtient

$$|\lambda - a_{ii}| = \left| \sum_{j\neq i} a_{ij} \frac{v_j}{v_i} \right| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

■

Remarque 2.1.
On voit encore, par un argument de continuité (en faisant tendre les éléments en dehors de la diagonale vers zéro), que si les disques de Gershgorin sont tous disjoints, alors chacun contient exactement une valeur propre.

Interprétation
λ appartient au disque de centre a_{kk} et de rayon

$$r = (\sum_{i=1}^{n} |a_{ki}|) - |a_{kk}|$$

Ainsi, ce théorème nous permet de localiser toutes les valeurs propres de A dans la réunion de n disques fermés, appelés disques de Gershgorin.

Exemple 2.1.
Considérons la matrice carrée A suivante

$$A = \begin{pmatrix} -10 & 1 & -0.5 & 11 \\ -0.6 & 10 & 11.3 & 11.4 \\ 0.3 & -1.1 & 1.1 & 2 \\ 1.2 & 17.1 & 0.1 & 17 \end{pmatrix}$$

L'application du théorème de Gershgorin permet de déduire la localisation des valeurs propres λ_1, λ_2, λ_3, λ_4 de A dans la réunion des disques fermés. $D_1 = D(-10, 12.5)$, $D_2 = D(10, 13.3)$, $D_3 = D(1.1, 3.4)$ et $D_4 = D(17, 18.4)$.
On ne peut pas conclure que chaque disque de Gershgorin contient une valeur propre.
D'ailleurs, le spectre de A à 10^{-4} près est

$$sp(A) = \{10.6980, 3.9192 - 4.7928i, 3.9192 + 4.7928i, 20.9596\}$$

et le disque D_3 ne contient aucune des quatre valeurs propres de la matrice A.

2.2 Théorème de Schur

Théorème 2.2.
Soit A une matrice carrée réelle,alors il existe une matrice réelle et orthogonale U et une matrice triangulaire supérieure par blocs T tels que ${}^tUAU = T$. Sur la diagonale de T se trouvent des blocs 1×1 correspondant aux valeurs propres réelles de A et des blocs 2×2 correspondant aux valeurs propres complexes conjuguées de A.

Proposition 2.1.
Soit A une matrice d'ordre n et $\lambda_1, \lambda_2, \ldots, \lambda_n$, ses valeurs propres. Alors,

$$\sum_{i=1}^{n} |\lambda_i|^2 \leq \sum_{j=1}^{n}\sum_{k=1}^{n} |a_{jk}|^2$$

Cette inégalité s'appelles inégalité de Schur.

Corollaire 2.1.
On obtient une borne supérieure du module de toute valeur propre de la matrice A par la formule suivante

$$\rho(A) \leq \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} |a_{jk}|^2}$$

Exemple 2.2.
Soit la matrice carrée A

$$A = \begin{pmatrix} -1 & 1 & 2 & 1 \\ -0.5 & 3 & 1 & 0.4 \\ 0.3 & -1.1 & 1.1 & 2 \\ 1.2 & 0.7 & 0.1 & 1.7 \end{pmatrix}$$

On a alors,

$$\sum_{i=1}^{4} |\lambda_i|^2 \leq 28.75 \quad \text{et} \quad \max_{1\leq i\leq 4} |\lambda_i| \leq 5.37$$

3 Méthodes itératives

Le premier type de méthodes que nous présentons dans ce paragraphe sont dites "itératives" . Ce sont des algorithmes qui approchent progressivement les valeurs propres par une suite de calculs répétés.
Ces méthodes sont idéales pour les matrices de grande dimension où les méthodes directes sont trop coûteuses.
Cependant, l'inconvénient majeur est que la convergence n'est pas toujours garantie pour toutes les matrices et dépend des propriétes des la matrice et du choix de l'algorithme. Elles nécessitent un critère d'arrêt et peuvent demander un grand nombre d'itérations pour atteindre la précision désirée. Voici quelques méthodes itératives.

3.1 Méthode de la puissance

La méthode des puissances, encore appelée méthode de la puissance itérée, repose sur l'idée qu'en appliquant un grand nombre de fois la matrice sur un vecteur quelconque,

les vecteurs successifs obtenus prennent une direction qui se rapproche de la direction du vecteur propre associé à la plus grande valeur propre en valeur absolue.
La méthode de la puissance itérée consiste donc à trouver par une méthode itérative, la valeur propre λ_1 de plus grand module d'une matrice A, ainsi que d'un vecteur propre associé.

3.1.1 Procédé

a. Valeurs Propres

Soit A une matrice de $\mathbb{M}_n(\mathbb{R})$ ayant n valeurs propres distinctes $\lambda_1, \lambda_2, \ldots, \lambda_n$ avec
$|\lambda_1| > |\lambda_2| > \ldots |\lambda_n|$ et pour vecteurs propres $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Soit y un vecteur de $\mathbb{R}^n$, on a

$$y = \sum_{j=1}^{n} C_j.x^{(j)}$$

L'écriture est possible car $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$=base.
On pose

$$y^{(1)} = Ay = \sum_{j=1}^{n} C_j A x^{(j)} = \sum_{j=1}^{n} C_j \lambda_j x^{(j)}$$

et

$$y^{(m)} = A^m y = \sum_{j=1}^{n} C_j \lambda_j^m x^{(j)} \tag{1}$$

Soit $\{e_1, e_2, \ldots, e_n\}$ une base quelconque de $\mathbb{R}^n$. On pose

$$y^{(m)} = \begin{pmatrix} y_1^{(m)} \\ \vdots \\ y_n^{(m)} \end{pmatrix}$$

Donc, $y_i^{(m)}$ représente la $i^{\text{ème}}$ coordonnée du vecteur $y^{(m)}$ dans la base $\{e_1, e_2, \ldots, e_m\}$.
Le developpement des vecteurs propres $x^{(j)}$ par rapport à la base $\{e_1, e_2, \ldots, e_m\}$ donne

$$x^{(j)} = \sum_{i=1}^{n} x_{ij} e_i \tag{2}$$

En portant (2) dans (1), on obtient

$$y^{(m)} = \sum_{i=1}^{n} e_i \sum_{j=1}^{n} C_j x_{ij} \lambda_j^m \tag{3}$$

et donc,

$$y_i(m) = \sum_{j=1}^{n} C_j x_{ij} \lambda_j^m \tag{4}$$

et de même

$$y_i(m+1) = \sum_{j=1}^{n} C_j x_{ij} \lambda_j^{(} m+1) \tag{5}$$

Le rapport de (5) par (4) donne

$$\frac{y_i(m+1)}{y_i(m)} = \frac{C_1 x_{i1} \lambda_1^{m+1} + \ldots + C_n x_{in} \lambda_n^{m+1}}{C_1 x_{i1} \lambda_1^{m} + \ldots + C_n x_{in} \lambda_n^{m}} \tag{6}$$

Supposons que $C_1 \neq 0$ et $x_{i1} \neq 0$ (Ceci est possible grâce au choix approprié du vecteur initial y). Alors, (6) devient

$$\frac{y_i(m+1)}{y_i(m)} = \lambda_1 . \frac{1 + \frac{C_2 x_{i2}}{C_1 x_{i1}} . \left(\frac{\lambda_2}{\lambda_1}\right)^{m+1} + \ldots}{1 + \frac{C_2 x_{i2}}{C_1 x_{i1}} . \left(\frac{\lambda_2}{\lambda_1}\right)^{m} + \ldots}$$

Par suite,

$$\lim_{m \longrightarrow \infty} \frac{y_i(m+1)}{y_i(m)} = \lambda_1 \tag{7}$$

car $\lim_{m \longrightarrow \infty} \frac{\lambda_2}{\lambda_1} = 0$ pour $j > 1$.

Remarque 3.1.
Lorsque le choix du vecteur initial y est mauvais (cas extrêmement rare), il se peut que la forme (7) ne donne pas la valeur recherchée. On s'en aperçoit facilement en pratique d'après les valeurs "sautantes" du rapport $\frac{y_i(m+1)}{y_i(m)}$ quand m varie.

b. Vecteurs propres

On a

$$y = \sum_{j=1}^{n} C_j x^{(j)}$$

et

$$y^{(m)} = \lambda_1^m \left[C_1 x^{(1)} + \sum_{j=2}^{n} C_j \left(\frac{\lambda_j}{\lambda_1}\right)^m x^{(j)} \right] \tag{8}$$

D'où,

$$\lim_{m \longrightarrow \infty} \frac{1}{\lambda_1^m} . y^{(m)} = C_1 x^{(1)} \tag{9}$$

(9) signifie que la direction définie par $y^{(m)}$ tend vers celle définie par le vecteur propre $x^{(1)}$ associé à λ_1. En d'autres termes, $\lim_{m \longrightarrow \infty} y^{(m)}$ est un vecteur propre associé à λ_1.

Remarque 3.2.
D'après (8), lorsque m augmente, les composantes de $y^{(m)}$ tendent vers 0 si $|\lambda_1| < 1$ ou vers l'infini si $|\lambda_n| > 1$. En pratique, pour éviter de travailler avec des nombres extrêmement grands ou petits, on impose au vecteur $y^{(m)}$ une condition de normalisation, par exemple, la première composante égale à 1. On construit ainsi une suite de vecteurs $Y^{(1)}, \ldots, Y^{(n)}$ définie par

$$\begin{aligned} Y^{(1)} &= \frac{1}{\alpha} Ay \\ Y^{(m)} &= \frac{1}{\alpha_m} AY^{(m-1)} \end{aligned}$$

α_m est le premier élément de la matrice $AY^{(m-1)}$. D'après (10), on a

$$Y^{(m)} = \frac{1}{\alpha_1 . \ldots . \alpha_m} A^m y = \frac{1}{\alpha_1 . \ldots , \alpha_m} y^{(m)}$$

Il s'ensuit alors que $Y^{(m)}$ *tend vers un vecteur propre associé à* λ_1 *et que*

$$\lambda_1 = \lim_{m \longrightarrow \infty} \alpha_m$$

En pratique, on arrête les calculs lorsque tous les rapports des composantes de $Y^{(m)}$ et de $Y^{(m-1)}$ sont compris entre $1 - \epsilon$ et $1 + \epsilon$ pour ϵ fixé à l'avance.

Exemple 3.1.
Soit

$$A = \begin{pmatrix} 5 & 2 & -2 \\ 2 & 2 & -1 \\ -2 & -1 & 2 \end{pmatrix}$$

On prend $y = {}^t(1, 2, 3)$. Cela donne

$$A.y = \frac{1}{\alpha_1} y^{(1)} = {}^t(3, 3, 2) = \frac{1}{3} {}^t(1, 1, 2/3)$$

$$A.y^{(1)} = \frac{1}{\alpha_2} y^{(2)} = {}^t(17/3, 10/3, -5/3) = \frac{17}{3} {}^t(1, 10/17, -5/17)$$

Les résultats sont donnés par le tableau suivant:

$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$	$y^{(7)}$	$y^{(8)}$
1	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	0.5882	0.5130	0.5019	0.5003	0.5000	0.5000	0.5000
0.6667	-0.2941	-0.4696	-0.4956	-0.4994	-0.5000	-0.5000	-0.5000
α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
3	5.6667	6.7647	6.9652	6.9950	6.9993	6.9999	7.0000

Nous présentons maintenant l'algorithme de la méthode de la puissance.

Algorithme 1.

$x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right) \in \mathbb{R}^n$, `un vecteur arbitraire donné,` k $= 0$

`répéter`

$y^{(k+1)} = Ax^{(k)}$

$x^{(k+1)} = \frac{y^{(k+1)}}{\|y^{(k+1)}\|}$

$\tilde{\lambda}^{(k+1)} = (x^{(k)})^t Ax^{(k)}$

$k = k + 1$

`jusqu'à` [test d'arrêt].

Le test d'arrêt est souvent basé sur $|\tilde{\lambda}^{(k+1)} - \tilde{\lambda}^{(k)}| \leq \varepsilon$, c'est-à-dire que l'itération s'arrête dès que la différence entre deux estimations de la valeur propre est suffisamment petite.

3.2 Méthode de la puissance inverse

Cette méthode permet d'approcher la valeur propre de la matrice A la plus proche d'un nombre μ donné n'appartenant pas à son spectre. On applique la méthode de la puissance à la matrice $(A - \mu I_n)^{-1}$ dont les vecteurs propres sont $\{\upsilon_1, \ldots, \upsilon_n\}$ ceux de la matrice A et les valeurs propres sont $(\lambda_i - \mu)^{-1}, 1 \leq i \leq n$.
Il suffit d'adapter l'algorithme 1 pour obtenir un nouvel algorithme.

Algorithme 2.

Méthode de la puissance inverse

$x^{(0)} \in \mathbb{R}^n$ donné k $= 0$

`répéter`

$y^{(k+1)} = (A - \mu I_n)^{-1} x^{(k)}$

$x^{(k+1)} = \frac{y^{(k+1)}}{\|y^{(k+1)}\|}$

$\tilde{\lambda}^{(k+1)} = (x^{(k)})^t (A - \mu I_n)^{-1} x^{(k)}$

$k = k + 1$

`jusqu'à` [test d'arrêt].

Remarque 3.3.
On calcule une décomposition LU de la matrice $A - \mu I_n$ afin de résoudre le système

$$(A - \mu I_n) y^{(k+1)} = x^{(k)}$$

Exemple 3.2.
Considérons la matrice carrée

$$A = \begin{pmatrix} -3 & 0 & 0 \\ 17 & 13 & -7 \\ 16 & 14 & -8 \end{pmatrix}$$

avec le spectre de A, $spec(A) = \{\lambda_1 = 6, \lambda_2 = -3, \lambda_3 = -1\}$.
En appliquant la méthode de la puissance avec

$$x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ on obtient, } x^{(1)} = \begin{pmatrix} -0.0938417 \\ 0.7194529 \\ 0.6881724 \end{pmatrix}, x^{(10)} = \begin{pmatrix} 0.0002303 \\ 0.7070492 \\ 0.7071643 \end{pmatrix}$$

et $\tilde{\lambda}^{(1)} = 4.1986301$, $\tilde{\lambda}^{(10)} = 6.0036631$.

Si on applique la méthode de la puissance inverse, avec $\mu = 0$ et en partant de

$$x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ on obtient, } x^{(1)} = \begin{pmatrix} -0.13938466 \\ 0.6270597 \\ 0.7664063 \end{pmatrix}, x^{(10)} = \begin{pmatrix} 0.0000151 \\ 0.4472257 \\ -0.8944211 \end{pmatrix}$$

et $\tilde{\lambda}^{(1)} = 0.5242718$, $\tilde{\lambda}^{(10)} = -1.0000478$.

3.3 Méthode de la déflation

Soit A une matrice réelle symétrique possédant n valeurs propres distinctes $\lambda_1, \lambda_2, \ldots, \lambda_n$ avec
$|\lambda_1| > |\lambda_2| > \ldots |\lambda_n|$. Soient $v_1, v_2, \ldots, v_n$ les vecteurs propres associés.
La méthode de déflation consiste, en appliquant n fois la méthode de la puissance itérée, à calculer toutes les valeurs propres et tous les vecteurs propres de la matrice A.

Principe.

On pose

$$A^1 = A - \lambda_1 . \frac{v_1 . {}^t v_1}{{}^t v_1 . v_1}$$

A^1 a pour éléments propres $(0, v_1), (\lambda_2, v_2), \ldots, (\lambda_n, v_n)$
En effet, on a

$$A^1 v_1 = A v_1 - \lambda_1 . \frac{v_1 . ({}^t v_1 . v_1)}{{}^t v_1 . v_1} = \lambda_1 v_1 - \lambda_1 v_1 = 0 . v_1$$

$$A^1 v_j = \lambda_j v_j - \lambda_1 . \frac{v_1 . ({}^t v_1 . v_j)}{{}^t v_1 . v_1}$$

Comme A est symétrique alors le système de vecteurs propres $(v_1, v_2, \ldots, v_n)$ est orthogonal. Donc, ${}^t v_1 . v_j = 0 \ \forall j \in \{2, \ldots, n\}$. Pour calculer (λ_2, v_2) on applique la méthode de la puissance itérée à la matrice A^1.
D'une manière générale, pour obtenir (λ_j, v_j), on considère la matrice

$$A^{j-1} = A - \lambda_{j-1} . \frac{v_{j-1} . {}^t v_{j-1}}{{}^t v_{j-1} . v_{j-1}}$$

qui a pour éléments propres

$$(0, v_1), \ldots, (0, v_{j-1}), (\lambda_j, v_j), \ldots, (\lambda_n, v_n)$$

Remarque 3.4.
Les matrices A^i sont symétriques mais non régulières puisque 0 n'est pas valeur propre multiple.

Exemple 3.3.
Soit la matrice carrée

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 1 & 2 & 4 \\ 4 & 1 & 3 & 2 \end{pmatrix}$$

$Sp(A) = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. Par application directe de la méthode de la puissance itérée, on obtient la plus grande valeur propre de A en module, $\lambda_1 = 10.0000$ et son vecteur propre associé

$$\upsilon_1 = \begin{pmatrix} 0.5000 \\ 0.5000 \\ 0.5000 \\ 0.5000 \end{pmatrix}$$

Le spectre de la matrice

$$B = A - \lambda_1 \frac{\upsilon_1^t \upsilon_1}{{}^t\upsilon_1 \upsilon_1} = \begin{pmatrix} -3/2 & -1/2 & 1/2 & 3/2 \\ 3/2 & -3/2 & -1/2 & 1/2 \\ 1/2 & -3/2 & -1/2 & 3/2 \\ 3/2 & -3/2 & 1/2 & -1/2 \end{pmatrix}$$

contient les trois autres valeurs propres de A et zéro. En appliquant la méthode de la puissance itérée à B, on obtient

$$\lambda_2 = -2.8019 \quad \text{et} \quad \upsilon_2 = \begin{pmatrix} -0.6032 \\ 0.4224 \\ -0.0340 \\ 0.6757 \end{pmatrix}$$

3.4 Méthode de Krylov (1931)

La méthode de Krylov consiste à determiner le polynôme caractéristique à l'aide de la résolution d'un système linéaire.

3.4.1 Calcul des valeurs propres

Soit A une matrice de $\mathbb{M}_n(\mathbb{R})$ et $P(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n$ son polynôme caractéristique. D'après le théorème d'Hamilton-Cayley, $P(A) = 0$.

$$P(A) = 0 \Longleftrightarrow A^n + a_1.A^{n-1} + \ldots + a_n I = 0 \tag{1}$$

Soit $x^{(0)}$ un vecteur non nul de $\mathbb{R}^n$. Posons

$$x^{(k)} = A^k.x^{(0)} \tag{2}$$

En multipliant (1) par $x^{(0)}$, on obtient

$$A^n x^{(0)} + a_1.A^{(n-1)}x^{(0)} + \ldots + a_n x^{(0)} \tag{3}$$

En introduisant (2) dans (3), on a

$$x^{(n)} + a_1 x^{(n-1)} + \ldots + a_n x^{(0)} \tag{4}$$

(4) équivaut à

$$a_1 x^{(n-1)} + \ldots + a_n x^{(0)} \tag{5}$$

(5) peut se mettre sous la forme suivante

$$\begin{pmatrix} x_1^{(n-1)} & x_1^{(n-2)} & \ldots & x_1^{(0)} \\ x_2^{(n-1)} & x_2^{(n-2)} & \ldots & x_2^{(0)} \\ \vdots & \vdots & \vdots & \vdots \\ x_n^{(n-1)} & x_n^{(n-2)} & \ldots & x_n^{(0)} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = - \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_n^{(n)} \end{pmatrix} \tag{6}$$

On obtient ainsi un système linéaire de n équations à n inconnues dont la résolution permet de trouver $a_1, a_2, \ldots, a_n$.

Remarque 3.5.
Après avoir obtenu $P(\lambda)$, on peut appliquer une méthode numérique pour avoir les racines de $P(\lambda) = 0$.

Exemple 3.4.
On veut calculer les valeurs propres de la matrice

$$A = \begin{pmatrix} 4 & 2 \\ 1 & 2 \end{pmatrix}$$

On prend $x^{(0)} =^t (1, 0)$. Alors $x^{(1)} =^t (4, 1)$ et $x^{(2)} =^t (18, 7)$. D'où,

$$\begin{pmatrix} 4 & 2 \\ 1 & 2 \end{pmatrix} . \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = - \begin{pmatrix} 18 \\ 7 \end{pmatrix}$$

soit $a = -7$ et $a_2 = 10$. Donc, $P(\lambda) = \lambda^2 - 7\lambda + 10$ et par suite, $\lambda_1 = 5$ et $\lambda_2 = 2$.

3.4.2 Calcul des vecteurs propres

On considère le cas où les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont distinctes. On cherche à trouver lee vecteurs propres $v_1, v_2, \ldots, v_n$ associés à $\lambda_1, \lambda_2, \ldots, \lambda_n$. Posons

$$x^{(0)} = \sum_{i=1}^{n} c_i v_i \tag{7}$$

(7) implique que

$$x^{(k)} = \sum_{i=1}^{n} c_i \lambda_i^k v_i \tag{8}$$

Soit

$$\phi_i(\lambda) = \lambda^{n-1} + q_{1i}\lambda^{n-2} + \ldots + q_{n-1,i} \qquad i = 1, 2, \ldots, n \tag{9}$$

Un système de polynômes arbitraires. On considère la combinaison linéaire suivante :

$$x^{(n-1)} + q_{1i}x^{(n-2)} + \ldots + q_{n-1,i}x^{(0)} \tag{10}$$

De (7), (8) et (9) on déduit que

$$\begin{aligned} (10) &= c_1(q_{n-1,i} + \ldots + q_{1i}\lambda_1^{n-2} + \lambda_1^{n-1})v_1 + \ldots + c_n(q_{n-2,i} + \ldots + q_{1i}\lambda_n^{n-2} + \lambda_n^{n-1})v_n \\ &= c_1\phi_i(\lambda_1)v_1 + \ldots + c_n\phi_i(\lambda_n)v_n \qquad (11) \end{aligned}$$

Si on prend $\phi_i(\lambda) = \frac{P(\lambda)}{\lambda - \lambda_i} \quad i = 1, \ldots, n$ alors, $\phi(\lambda_j) = 0$ pour $i \neq j$
et $\phi(\lambda_i) = P'(\lambda_i) \neq 0$ car λ_i est une racine simple de $P(\lambda) = 0$. Dans ce cas, (11) devient

$$x^{(n-1)} + q_{1i}x^{(n-2)} + \ldots + q_{n-1,i}x^{(0)} = c_i\phi_i(\lambda_i)v_i \qquad i = 1, 2, \ldots, n \tag{12}$$

(12) montre que (10) est un vecteur propre associé à λ_i (on a supposé $c_i \neq 0$, ce qui est toujours possible par un choix approprié du vecteur initial $x^{(0)}$). Les coefficients $q_{ji} \quad j = 1, \ldots, n$ peuvent être determinés par le schéma de Hörner

$$\begin{cases} q_{0i} = 1 \\ q_{ji} = \lambda_i q_{j_1,i} + a_j \end{cases}$$

Exemple 3.5.

$$\begin{pmatrix} 4 & 2 \\ 1 & 2 \end{pmatrix}$$

Avec $x^{(0)} =^t (1, 0)$, on a obtenu $x^{(1)} =^t (4, 1)$ et $x^{(2)} =^t (18, 7)$. Ce qui donne $\lambda_1 = 5$ et $\lambda_2 = 2$.
$v_1 = x^{[1)} + q_{11}x^{(0)}$ où $q_{11} = \lambda_1 q_{01} + q_1$.
soit $v_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

4 Méthode de tridiagonalisation de matrices

4.1 Méthode de Householder

Soit $A = (a_{ij})_{i,j}$ une matrice symétrique de $\mathbb{M}_n(\mathbb{R})$.

4.1.1 Principe

On pose $A^{(1)} = A$ et on génère $(n-2)$ matrices symétriques $A^{(k)}$, $2 \leq k \leq n-1$, semblables à la matrice et ce à l'aide des transformations de Householder (matrices hermitiennes et élémentaires). La matrice $A^{(k)}$ est de la forme

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(2)} & 0 & 0 & & & & \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(3)} & 0 & & 0 & & \\ 0 & a_{32}^{(3)} & a_{33}^{(3)} & a_{34}^{(4)} & & & & \\ 0 & 0 & a_{43}^{(4)} & a_{44}^{(4)} & & & & \\ & & & & \ddots & & & \\ & & & & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \dots & a_{kn}^{(k)} \\ & & & & a_{k+1,k}^{(k)} & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ & 0 & & & \vdots & \dots & \dots & \vdots \\ & & & & \vdots & \dots & \dots & \vdots \\ & & & & a_{nk}^{(k)} & a_{n,k+1}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}$$

Le passage à la matrice $A^{(k+1)}$ se fait comme suit :
Pour passer de $A^{(k)}$ à $A^{(k+1)}$, on cherche une matrice orthogonale élémentaire $H^{(k)}$ telle que $H^{(k)} = I - 2u^{(k)}.{}^t u^{(k)}$ avec $u_i^{(k)} = 0$ pour $1 \leq i \leq k$

$$A^{(k+1)} = H^{(k)}.A^{(k)}.H^{(k)}$$

On peut vérifier que dans la transformation $A^{(k)} \longrightarrow A^{(k+1)}$, seule la sous-matrice $\tilde{A}^{(k)} = (a_{ij})^{(k)})_{k \leq i,j \leq n}$ est modifiée, à l'exception de $a_{kk}^{(k)}$.

4.1.2 Algorithme

$a_{ij}^{(1)} = a_{ij} \quad 1 \leq i, j \leq n$
Pour k allant de 2 à $n-1$ faire,

$$\begin{aligned}
a_{k,k+1}^{(k+1)} &= -\Big(\text{ sgn } a_{k,k+1}^{(k)}\Big)\left(\sum_{i=k+1}^{n} a_{ik}^2\right)^{1/2} = a_{k+1,k}^{(k+1)} \\
\nu^{(k)} &= a_{k,k+1}^{(k+1)}\left(a_{k,k+1}^{(k+1)} - a_{k,k+1}^{(k)}\right) \\
v_{k+1}^{(k)} &= a_{k,k+1}^{(k)} - a_{k,k+1}^{(k+1)} \\
v_i^{(k)} &= a_{ik}^{(k)} \quad \text{pour} \quad k+2 \leq i \leq n \\
\delta_i^{(k)} &= \frac{1}{\nu^{(k)}} \sum_{j=k+1}^{n} a_{ij}^{(k)}.v_j^{(k)} \quad \text{pour} \quad k+1 \leq i \leq n \\
\beta^{(k)} &= \frac{1}{2\nu^{(k)}} \sum_{i=k+1}^{n} \delta_i^{(k)}.v_i^{(k)} \quad \text{pour} \quad k+1 \leq i \leq n \\
q_i(k) &= \delta_i(k) - \beta^{(k)}.v_i^{(k)} \quad \text{pour} \quad k+1 \leq i \leq n \\
a_{ij}^{(k+1)} &= a_{ij}^{(k)} - v_i^{(k)}.q_j^{(k)} - q_i^{(k)}.v_j^{(k)} \quad \text{pour} k+1 \leq i, j \leq n \\
a_{kj}^{(k+1)} &= 0 \quad \text{pour} \quad j = k+2, \ldots, n
\end{aligned}$$

La matrice $A^{(n-1)}$ est symétrique et tridiagonale. De plus, si on pose

$$Q = H^{(n-2)}.H^{(n-3)}.\ldots.H^{(1)}$$

alors on aura $A^{(n-1)} = Q.A.{}^tQ$.
La matrice Q étant orthogonale, le procédé ainsi défini permet de tridiagonaliser la matrice A par une transformation orthogonale semblable.

Remarque 4.1.
(i). Le coût de la méthode est environ égal à $2n^3/3$.
(ii). La méthode est très stable.

4.2 Méthode de Givens

Soit A une matrice symétrique de $\mathbb{M}_n(\mathbb{R})$. La méthode de Givens conciste à obtenir une matrice tridiagonale J semblable à A par une suite de transformations unitaires (les mêmes que celles utilisées dans l'algorithme de Jacobi). Soit T_{pq} la matrice dite de Jacobi définie par

$$\begin{aligned}
t_{pp}^{(pq)} &= \cos\varphi \quad , \quad t_{qq}^{(pq)} = -\cos\varphi \\
t_{pq}^{(pq)} &= t_{qp}^{(pq)} = \sin\varphi \quad \forall i, j \neq p, q \quad t_{ii}^{(pq)} = \delta_{ij}
\end{aligned}$$

On pose $A^{(1)} = A$ et $A^{(2)} = T_{pq}.A.T_{pq}^{-1}$. Givens choisit φ de façon à annuler $a_{p-1,q}^{(2)}$. Or,

$$a_{iq}^{(2)} = a_{ip}\sin\varphi - a_{iq}\cos\varphi$$

On prend donc, φ tel que

$$\tan\varphi = \frac{a_{p-1,q}}{a_{p-1,p}}$$

Dans la transformation $A^{(1)} \longrightarrow A^{(2)}$, seules les lignes et les colonnes d'indice p et q sont modifiées. Givens procède ainsi,
Il fixe $p = 2$ et fait varier $q = 3, 4, \ldots, n$.
Puis, $p = 3$ et $q = 4, 5, \ldots, n$.
Et ainsi de suite.

$$\begin{pmatrix} \times & \times & \times & & \times & & \times & & \times \\ \times & \times & (1) & \longrightarrow & (2) & \longrightarrow & (3) & \longrightarrow & (4) \\ \times & \times & \times & & (5) & \longrightarrow & (6) & \longrightarrow & (7) \\ \times & \times & \times & & \times & & (8) & \longrightarrow & (9) \\ \times & \times & \times & & \times & & \times & & (10) \\ \times & \times & \times & & \times & \times & \times & & \end{pmatrix}$$

La réduction complète à la forme tridiagonale nécessite environ $4n^3/3$ opérations, soit quasiment le double de la méthode de Householder.

5 Méthodes de réduction de matrices

Dans qui suit, on présente de manière assez succinte quelques méthodes qui ont pour principe la réduction de la matrice A à une forme particulière.

5.1 Méthode de Jacobi (1846)

La méthode de Jacobi est une méthode itérative applicable à une matrice A symétrique. Elle consiste à faire opérer le groupe des rotations planes sur A c'est-à-dire à multiplier A par des transformations orthogonales afin de la mettre sous forme diagonale, les éléments diagonaux étant les valeurs propres de la matrice A. Partant de la matrice $A_1 = A$ symétrique, la méthode de Jacobi a pour principe de construire une suite $(T_k)_k$ de matrices orthogonales telles que la suite de matrices symétriques $A_{k+1} = {}^t T_k.A_k.T_k$ converge vers la matrice orthogonale D. De plus, $Q_k = T_1.T_2.\ldots.T_k$ converge vers une matrice orthogonale dont les colonnes sont les vecteurs propres de A.

5.2 Méthode LU (Rutishauser, 1958)

Cette méthode est basée sur la décomposition d'une matrice en un produit de deux matrices L_1 et U_1. On forme les matrices
$A_1 = A = U_1.L_1 = L_2.U_2$

$A_2 = U_2.L_2 = L_3.U_3$
$A_k = U_k.L_k = L_{k+1}.U_{k+1}$
On démontre que si la suite $\{A_k\}_k$ converge, elle tend vers une matrice triangulaire inférieure dont les éléments de la diagonale sont les valeurs propres de A.

Remarque 5.1.
• La méthode ne peut s'appliquer que si l'une des matrices A_i n'admet pas de décomposition LU.
• Le calcul de L_i ou U_i peut être mal conditionné.

Remarque 5.2.
Dans la décomposition initiale de A, au lieu de choisir U_1 comme matrice à diagonale unitaire, on peut fixer L_1 comme matrice triangulaire inférieure à diagonale unitaire (décomposition de Doolittle) et alors la suite converge vers une matrice triangulaire supérieure

Exemple 5.1.
Soit la matrice carrée

$$A = \begin{pmatrix} 5 & 5 \\ 5 & 3 \end{pmatrix}$$

alors la matrice A peut s'écrire

$$A = \begin{pmatrix} 5 & 5 \\ 5 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 5 & 5 \\ 0 & -2 \end{pmatrix} = L_1 U_1$$

$$A_1 = U_1 L_1 = \begin{pmatrix} 5 & 5 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 5 \\ -2 & -2 \end{pmatrix}$$

$$\begin{pmatrix} 10 & 5 \\ -0.2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} 10 & 5 \\ 0 & -1 \end{pmatrix} = L_2 U_2$$

Et donc

$$A_2 = U_2 L_2 = \begin{pmatrix} 9.0000 & 5.0000 \\ 0.2000 & -1.0000 \end{pmatrix}$$

par suite, on obtient successivement

$$A_3 = \begin{pmatrix} 9.1111 & 5.0000 \\ 0.0247 & -1.1111 \end{pmatrix}, A_4 = \begin{pmatrix} 9.0976 & 5.0000 \\ 0.0030 & -1.0976 \end{pmatrix}, A_5 = \begin{pmatrix} 9.0992 & 5.0000 \\ -0.0004 & -1.0992 \end{pmatrix}$$

$$A_6 = \begin{pmatrix} 9.0990 & 5.0000 \\ 0.0000 & -1.0990 \end{pmatrix}, A_7 = \begin{pmatrix} 9.0990 & 5.0000 \\ 0.0000 & -1.0990 \end{pmatrix}$$

les valeurs propres de A sont : $\lambda_1 = 9.0990$ et $\lambda_2 = -1.0990$.

5.3 Méthode QR (Francis, 1961)

La méthode de Francis est identique à la méthode de Rutishauser à ceci près qu'elle utilise la décomposition QR (au lieu de la décomposition LU). Cette méthode consiste donc à factoriser A en un produit d'une matrice orthogonale Q et d'une matrice triangulaire R. On construit une suite de matrices A_k unitairement semblables à A par
$A_1 = A = Q_1.R_1$
$A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1} \quad k \geq 1$
Si $\{A_k\}_k$ converge, elle tend vers R dont les éléments diagonaux sont les valeurs propres de A.

Remarque 5.3.
• La factorisation QR existe toujours.
• Même si elle est stable, cette méthode est trop coûteuse quand elle est appliquée directement. Cependant, elle possède l'avantage de conserver la structure de départ.
• Si A est une matrice de $\mathbb{M}_n(\mathbb{R})$ qui possède des valeurs propres réelles, distinctes en valeurs absolues, alors la suite des matrices $(A_k)_k$ converge vers une matrice triangulaire supérieure.

Exemple 5.2.
Soit la matrice carrée

$$A = \begin{pmatrix} 5 & 5 \\ 5 & 3 \end{pmatrix}$$

alors la matrice A peut s'écrire

$$A = \begin{pmatrix} 5 & 5 \\ 5 & 3 \end{pmatrix} = \begin{pmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{pmatrix} \begin{pmatrix} -7.0711 & -5.6569 \\ 0 & -1.4142 \end{pmatrix} = Q_1 R_1$$

D'où

$$A_1 = R_1 Q_1 = \begin{pmatrix} 9 & 1 \\ 1 & -1 \end{pmatrix}$$

par suite, on obtient successivement

$$A_2 = \begin{pmatrix} 9.0976 & 0.1220 \\ 0.1220 & -1.0976 \end{pmatrix}, A_3 = \begin{pmatrix} 9.0990 & 0.0147 \\ 0.0147 & -1.0990 \end{pmatrix}, A_4 = \begin{pmatrix} 9.0990 & 0.0018 \\ 0.0018 & -1.0990 \end{pmatrix}$$

$$A_5 = \begin{pmatrix} 9.0990 & 0.0002 \\ 0.0002 & -1.0990 \end{pmatrix}, A_6 = \begin{pmatrix} 9.0990 & 0.0000 \\ 0.0000 & -1.0990 \end{pmatrix}$$

les valeurs propres de A sont : $\lambda_1 = 9.0990$ et $\lambda_2 = -1.0990$.

6 Méthodes directes

Les méthodes directes représentent les méthodes qui nous permettent d'obtenir les valeurs propres directement à partir du polynôme caractéristique seulement. Et c'est pour cela que ces méthodes ne seront pas décrites car elles sont équivalentes à la résolution du polynôme caractéristique.

7 Exercices résolus

7.1

Soit les matrices

1. Calculer les spectres des matrices suivantes
$$A = \begin{pmatrix} 1 & 0 \\ 1000 & 1 \end{pmatrix} \text{ et } B = \begin{pmatrix} 1 & 0.0001 \\ 1000 & 1 \end{pmatrix}$$
2. On considère la matrice
$$A(\alpha) = \begin{pmatrix} 1 & \alpha + 2 \\ \alpha - 1 & 4 \end{pmatrix}$$
Calculer les valeurs propres de $A(\alpha)$ en fonction de α.
3. Analyser les résultats obtenus.

7.2

Soit la matrice
$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$
Par la méthode de la puissance itérée, déterminer ses valeurs et vecteurs propres.

Solution.
Résumons un peu la méthode,
En pratique, on se donne un vecteur initial $x^{(0)}$ (de norme égale à 1 ou quelconque), et on calcule les $x^{(k)}$ par la relation $x^{(k+1)} = Ax^{(k)}$.
De plus, on préfère ramener à 1 la valeur de l'une des composantes de ces vecteurs (par exemple la 1ère).
La composante correspondante du vecteur itéré suivant donne le rapport à la valeur précédente et tend vers la valeur propre de plus grand module.
A présent, on veut trouver les valeurs propres suivantes, on applique la méthode de déflation qui consiste à créer la matrice
$$A_1 = A - \lambda_1 \frac{\upsilon_1 (x^{(1)})^T}{(x^{(1)})^T \upsilon_1}$$
où le vecteur $x^{(1)}$ est le vecteur propre de la matrice transposée de A, correspondant à la valeur propre λ_1. On se donne un vecteur initial $x^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ à partir duquel on construit la suite de vecteurs,
$x^{(1)} = Ax^{(0)}$, $x^{(2)} = Ax^{(1)}, \ldots$, reporté dans le tableau ci-dessous

λ		1	5	$\frac{13}{5}$	$\frac{14}{13}$	$\frac{121}{41}$	$\frac{365}{121}$	...
$x^{(k)}$	1	1	1	1		1	1	
	0	2	$\frac{4}{5}$	$\frac{13}{5}$	$\frac{14}{13}$	$\frac{121}{41}$	$\frac{365}{121}$	...
	0	0	0	0	0	0	0	

Les composantes de chaque vecteur de ce tableau ont été divisées par la première composante correspondant à chaque vecteur.

La suite des vecteurs $x^{(k)}$ tend vers le vecteur $\upsilon^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ et que la suite des valeurs de λ converge vers la valeur $\lambda_1 = 3$.

La matrice A considérée étant symétrique, le vecteur propre de sa transposée, correspondant à la valeur propre λ_1 sera $x^{(1)} = \upsilon^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.

Construisons la matrice A_1 qui est telle que

$$A_1 = A - \lambda_1 \frac{\upsilon_1 (x^{(1)})^T}{(x^{(1)})^T \upsilon_1} = \begin{pmatrix} -1/2 & 1/2 & 0 \\ 1/2 & -1/2 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

On se donne à nouveau un vecteur $x^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ et on procède de la même manière qu'avec la matrice A,

λ		$\frac{-1}{2}$	-1	-1	...
$x^{(k)}$	1	1	1	1	
	0	-1	-1	-1	...
	0	0	0	0	

On voit que cette suite de vecteurs converge vers le vecteur propre: $\upsilon_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ et que la valeur propre qui lui correspond est $\lambda_2 = -1$.

On recommence le processus en définissant une matrice A_2 à partir de A, $x^{(2)} = \upsilon_2$:

$$A_2 = A_1 - \lambda_2 \frac{\upsilon_2 (x^{(2)})^T}{(x^{(2)})^T \upsilon_2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Il est clair que la valeur propre de cette matrice est $\lambda_3 = -1$ et que le vecteur propre correspondant est $\upsilon_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

7.3

Soit la matrice

$$\begin{pmatrix} 3 & -1 & 0 \\ -1 & 8 & 2 \\ 0 & 2 & 15 \end{pmatrix}$$

1. Montrer que A admet 3 valeurs propres distinctes λ_1, λ_2 et $\lambda_3 1$ telles que

$$|\lambda_1| > |\lambda_1| > |\lambda_1|$$

2. Appliquer la méthode de la puissance itérée à A en prenant $x^{(0)} =^t (0, 0, 1)$.

7.4

Soit la matrice carrée

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

Déterminer sa plus grande valeurs propre ainsi que le vecteur propre associé par la méthode de la puissance itérée.

Solution.
En partant du vecteur initial $x^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$,
l'algorithme de la puissance $x^{(k)} = Ax^{(k-1)}$ fournit les valeurs suivantes

k	0	1	2	3	4	5	6	7
$x^{(k)}$	1	2	5	14	41	122	365	1094
	0	-1	-4	-13	-40	-121	-364	-1093
					2.928	2.975	2.991	2.9973

Le rapport de 2 composantes homologues successives $\left(\frac{x_i^{(k)}}{x_i^{(k-1)}}\right)$ est donné dans la dernière ligne, dans la dernière colonne, il converge vers la valeur 3, qui serait la plus grande des valeurs propres.
Le vecteur $x^{(k)}$ semble tendre vers une limite proportionnelle au vecteur $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ qui serait le vecteur propre υ_1 associé à la valeur propre $\lambda_1 = 3$.

Remarque 7.1.
Comme il est expliqué dans l'algorithme, il vaut mieux normaliser $x^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}$ dés qu'il est calculé avec l'inconvénient que le calcul est très lourd fait à la main.

Voici les résultats calculés à l'aide du logiciel Matlab.

k	0	1	2	...	5	6
$v^{(k)}$	1	0.8944272	0.7808688	...	0.7100107	0.7080761
	0	-0.4472136	-0.6246950	...	-0.7041909	-0.7061361

Ainsi la valeur approchée de la valeur propre est obtenue par la formule
$\lambda \simeq v^{(6)T} A v^{(6)} = 2.9999962$. Cette approximation est plus précise que celle calculée précedemment.

7.5

Calculer les valeurs propres de la matrice suivante

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

avec la méthode de Krylov.

CHAPITRE IV. Résolution numérique des équations différentielles

1 Introduction.

Les équations différentielles, souvent obtenues lors de la modélisation des phénomènes physiques, peuvent être diviser en deux grandes familles: Les équations différentielles ordinaires et les équations aux dérivées partielles. Dans ce dernier cas les équations sont caractériséés par le fait que la variable dépendante (ou les variables dépendantes) est fonction de plusieurs variables indépendantes. L'autre catégorie d'équations différentielles sont les équations différentielles ordinaires qui sont caractérisées par le fait que la variable dépendante (ou les variables dépendantes) ne dépend que d'une seule variable indépendante et que sa dérivée soit alors, une dérivée totale. Un exemple de ce genre d'équations est l'équation qui gouverne l'évolution, en fonction du temps t, de la vitesse v, d'une masse m dans le champ de gravité terrestre g en chute amortie par un coeffcient d'amortissement k :

$$\frac{dv}{dt} = g - \frac{k}{m} \times v$$

ou bien encore, l'équation de la position angulaire θ d'un pendule, de longueur l, qui oscille dans un plan vertical:

$$\frac{d^2\theta}{dt^2} + \frac{g}{l} \times \sin\theta = 0.$$

Il existe d'autres modèles basés sur des équations différentielles ordinaires qui sont extrêmement courants tel que:
En cinétique chimique, dynamique des populations et en météorologie. Quand ces équations sont linéaires ou qu'elles peuvent raisonnablement être rendues linéaires, souvent des solutions analytiques peuvent être obtenues pour ces équations.
Toutefois, quand ces équations sont complexes ou non linéaires, les méthodes analytiques échouent et une approche numérique s'avère être une solution. L'objectif de ce chapitre est de présenter quelques méthodes numériques pour la résolution des équations différentielles ordinaires.

2 Position du problème

On appelle équation différentielle une équation reliant une fonction et ses dérivées successives. Si l'équation ne fait intervenir que la fontion et sa dérivée, on parle d'équation du premier ordre.
Soit Ω un ouvert de $\mathbb{R} \times \mathbb{R}^N$. Etant donnés $f : \Omega \subset \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ qui à un couple $(t, u) \in \Omega$ associe $f(t, u)$ continue, un point quelconque $t_0 \in \mathbb{R}$, un point quelconque $U_0 \in \mathbb{R}^N$, nous cherchons une fontion $u : I \to \mathbb{R}^N$, où I est un voisinage de t_0 dans $\mathbb{R}$,

qui à $t \in I$ associe $u(t)$, continûment dérivable, telle que u vérifie le problème de Cauchy suivant:

$$\begin{cases} u'(t) &= f(t,u(t)) \\ u(t_0) &= U_0 \qquad \text{condition initiale ou condition de Cauchy} \end{cases} \tag{1}$$

On dit qu'il y a unicité au problème de Cauchy en (t_0, U_0) s'il existe au moins une solution à ce problème et si pour toutes solutions $\varphi : I \to \mathbb{R}^N$ et $\psi : J \to \mathbb{R}^N$, les fonctions φ et ψ coincident sur $I \cap J$.
L'étude mathématique de l'existence et de l'unicité d'une solution u est délicate et constitue une branche entière des mathématiques. Nous nous contenterons de donner une condition suffisante à l'existence et l'unicité d'une solution au problème de Cauchy. Nous ne nous intéresserons ensuite qu'à la résolution numérique de ces équations différentielles.

Théorème 2.1. *Théorème de Cauchy-Lipschitz*
Si f est continue dans Ω et si f est localement lipschitzienne par rapport à sa deuxième variable, ie si pour tout $(t_0, u_0) \in \Omega$, il existe un voisinage V de (t_0, u_0) et $L > 0$ tels que :

$$\forall (t,u),(t,v) \in V, |f(t,u) - f(t,v)| \leq L|u - v|,$$

alors pour tout $(t_0, u_0) \in \Omega$, il existe I voisinage de t dans $\mathbb{R}$ et une application u, continûment dérivable, de I dans $\mathbb{R}^N$ solution unique de :

$$\begin{cases} u'(t) &= f(t,u(t)) \\ u(t_0) &= U_0 \end{cases} \tag{2.1}$$

Remarque 2.1.
Si f est de classe C^1 alors f est localement lipschitzienne par rapport à sa deuxième variable, et le théorème de Cauchy-Lipschitz s'applique.

Exemple 2.1.

- $$\begin{cases} u'(t) &= u(t) - t \\ u(0) &= 0 \end{cases} \tag{2.2}$$

La fonction $u(t) = -\exp(t) + t + 1,\ \forall t \geq 0$ est solution.

- $$\begin{cases} u'(t) &= \frac{-u(t)}{t \ln t} + \frac{1}{\ln t}, \quad t \in [e,5] \\ u(e) &= e \end{cases} \tag{2.3}$$

La fonction $u(t) = \frac{t}{\ln t}$, est solution unique car

$$|f(t,u) - f(t,v)| = \left|\frac{v-u}{t \ln t}\right| \leq \left|\frac{v-u}{e}\right| = \frac{1}{e}|v-u|$$

3 Stabilité d'une équation différentielle ordinaire

Une question importante lors de la résolution numérique d'une équation différentielle est la stabilité de l'équation. La stabilité détermine si une méthode numérique peut être appliquée, ou le cas échéant, impose une borne sur la taille du pas afin d'obtenir une solution fiable. Afin de comprendre ce qu'est la stabilité d'une équation différentielle, nous allons tout d'abord étudier une équation différentielle instable.

Exemple 3.1.
Equation différentielle ordinaire instable: Soit le problème aux valeurs initiales,

$$\begin{cases} x'(t) &= x(t) \\ x(0) &= C \end{cases} \tag{3.1}$$

Pour le problème (3.1), la solution analytique peut être trouvée. En effet,

$$x(t) = Ce^t$$

est la solution unique à (3.1). Considérons à présent une technique de résolution numérique de (3.1). Le principe de toutes les méthodes numériques est d'approximer $x(t)$ pour des temps discrétisés $t_0, t_0 + h, t_0 + 2h, t_0 + 3h, \ldots$. Inévitablement une erreur par rapport à la solution analytique sera introduite à chaque itération. Pour voir l'effet qu'a une erreur sur la suite du calcul, nous allons modéliser l'erreur ϵ faite lors du calcul, par une différence ϵ sur la condition initiale. Si l'on résout donc un problème perturbé $x_0(t) = x(t)$, $x(0) = C - \epsilon$, la solution obtenue est $x_\epsilon = (C - \epsilon)e^t$. Comparant $x(t)$ à $x_\epsilon(t)$, on obtient

$$e(t) = x(t) - x_\epsilon(t) = \epsilon e^t.$$

On voit que l'erreur introduite grandit exponentiellement lorsque t augmente. En d'autres termes, une erreur introduite au départ du calcul va être amplifiée exponentiellement en cours de calcul. Cela explique pourquoi on parle d'une équation différentielle instable. On comprend dès lors, pourquoi il sera plus difficile de résoudre une telle équation.

Nous passons tout naturellement à l'exemple d'une équation différentielle stable.

Exemple 3.2.
Equation différentielle ordinaire stable: Similairement à l'exemple précédent, nous considérons à présent le problème

$$\begin{cases} x'(t) &= -x(t) \\ x(0) &= C \end{cases} \tag{3.2}$$

dont la solution analytique est $x(t) = Ce^{-t}$. Une petite perturbation dans la condition initiale $x(0) = C - \epsilon$ nous donne cette fois comme solution $x_\epsilon(t) = (C - \epsilon)e^{-t}$ et comme erreur $e(t) = x(t) - x_\epsilon(t) = \epsilon e^{-t}$. Cette fois, l'erreur décroît exponentiellement. En d'autres termes, une erreur commise en début de calcul ne se répercutera pratiquement pas sur la suite du calcul.

Définition 3.1.
Soit une équation différentielle,

$$\begin{cases} x'(t) &= f(x(t),t) \\ x(t_0) &= x_0 \end{cases} \tag{3.3}$$

est dite stable en $(x(t),t)$ si son Jacobien

$$J(x(t),t) = \frac{df(x(t),t)}{dx}(x(t),t) < 0$$

et instable si son Jacobien

$$J(x(t),t) = \frac{df(x(t),t)}{dx}(x(t),t) > 0$$

La définition précédente se justifie car l'erreur commise en modifiant légèrement la condition initiale s'amplifie dans le cas instable et s'amenuise dans le cas stable comme la proposition suivante nous l'indique.

Proposition 3.1. *Soient les deux problèmes aux valeurs initiales*

$$\begin{cases} x'(t) &= f(x(t),t) \\ x(t_0) &= x_0 \end{cases} \tag{*}$$

$$\begin{cases} x'(t) &= f(x(t),t) \\ x(t_0) &= x_0 - \epsilon \end{cases} \tag{**}$$

dont les solutions sont $x^*(t)$ et $x^{**}(t)$ respectivement. Si on définit $e(t) = x^{**}(t) - x^*(t)$, on a

$$e'(t) \approx J(x(t),t)e(t),$$

et dès lors,

$$e(t) \approx \epsilon exp\left(\int_{t_0}^{t} J(x(s),s)ds\right). \tag{3.4}$$

Preuve
On a

$$e'(t) = x'^{**}(t) - x'^{*}(t) = f(x^{**}(t),t) - f(x^*(t),t) \approx f(x^*(t),t) + J(x^*(t),t)(x^{**}(t) - x^*(t)) - f(x^*(t),t)$$

$$e'(t) \approx J(x^*(t),t)e(t)$$

où la première approximation est obtenue grâce à un développement de Taylor tronqué à l'ordre 1. Finalement, l'expression (3.4) est obtenue en résolvant l'équation différentielle. ■

On déduit de la proposition précédente que l'erreur va croître exponentiellement pour une équation différentielle instable et décroître exponentiellement pour une équation différentielle stable.
Il est, par exemple, assez intuitif qu'une équation instable soit très difficile à résoudre numériquement. Il faut utiliser des méthodes très particulières pour ce faire. La situation des équations stables n'est pas aussi simple pour autant. Nous verrons que dans le cas de certains systèmes très stables, la situation peut également être problématique.

4 Méthodes de Taylor

Nous abordons différentes méthodes numériques pour résoudre les équations différentielles ordinaires en commençant par les plus intuitives de toutes. Tout d'abord, il est utile de remarquer, et c'est le cas pour toutes les méthodes que nous exposons dans ce cours, que pour trouver la fonction $x(t)$ recherchée, nous allons en réalité approximer $x(t)$ en $t_0, t_0 + h, t_0 + 2h, t_0 + 3h, \ldots$. La notation que nous adoptons dans tout le reste de ce chapitre est présentée ci-dessous.
Notations:
• Les temps pour lesquels une approximation de $x(t)$ est calculée sont notés par $t_0, t_1, t_2, \ldots$.
• Les approximations de $x(t)$ calculées aux temps $t_0, t_1, t_2, \ldots$ sont notées respectivement $\bar{x}_0, \bar{x}_1, \bar{x}_2, \ldots$
La première méthode que nous considérons ici consiste à écrire le développement de Taylor de la fonction recherchée $x(t)$ autour de t afin d'approximer au mieux la valeur de x en $t + h$. La longueur du développement choisie indique le degré de la méthode considérée.

4.1 Méthode d'Euler explicite

La méthode d'Euler explicite considère un développement de Taylor tronqué à l'ordre 1. Cela implique que l'on doit connaître la première dérivée de x en t. Cette dérivée est toutefois connue, puisqu'elle nous est donnée par l'intermédiaire de f. Rappelons-nous en effet que $x'(t) = f(x(t), t)$. Si on procède de la sorte, on obtient

$$x(t + h) \approx x(t) + hx'(t) = x(t) + hf(x(t), t).$$

Le processus $x(t+h) = x(t)+hf(x(t), t)$ est mieux connu sous le nom de méthode d'Euler explicite.
Méthode 1 (Méthode d'Euler explicite)
Les itérés de la méthode d'Euler explicite sont calculés successivement comme suit

$$\bar{x}_{i+1} = \bar{x}_i + hf(\bar{x}_i, t_i).$$

Exemple 4.1.
Soit le problème aux valeurs initiales,

$$\begin{cases} x'(t) &= -x^2(t) + t \\ x(0) &= 2 \end{cases}$$

Nous allons appliquer la méthode d'Euler explicite. Les valeurs obtenues par la méthode sont notées $\bar{x}_i$. Prenons un pas de $h = 0.3$. On a $\bar{x}_0 = 2$, et $\bar{x}_1 = 2 + hf(2, 0) = 2 + 0.3(-4) = 0.8$. Ensuite $\bar{x}_2 = 0 + hf(0.8, 0.3) = 0 + 0.3(-0.64 + 0.3) = 0.698$

Nous allons à présent analyser l'erreur commise lorsque l'on résout un problème aux valeurs initiales en utilisant la méthode d'Euler explicite. On se doute qu'à chaque pas dit d'intégration, une erreur va s'introduire dans le calcul. Mais il faut se rendre compte que l'erreur introduite va impliquer que l'approximation calculée à l'étape suivante le sera pour un problème légèrement différent résultant en une nouvelle erreur. On va donc voir dans la proposition suivante que l'erreur peut être décomposée en une erreur commise localement et une erreur globale résultant de l'accumulation des différentes erreurs locales et menant à la résolution d'un problème légèrement modifié.

Proposition 4.1. *Soit le problème au valeurs initiales*

$$\begin{cases} x'(t) &= f(x(t), t) \\ x(t_0) &= x_0 \end{cases} \tag{4.1}$$

et sa solution $x^*(t)$. On définit également par $\bar{x}_0, \bar{x}_1, \bar{x}_2, \ldots$ les différentes approximations de $x^*(t)$ obtenues en utilisant la méthode d'Euler explicite avec un pas h. L'erreur EG_i commise en $t_i = t_0 + ih$ peut s'exprimer comme

$$EG_i = (1 + hJ_i)EG_{i-1} + EL_i,$$

où EG signifie erreur globale et EL erreur locale et où $J_i = \frac{df}{dx}(\zeta_i, t_i)$ et $EL_i = -\frac{h^2}{2}x''(\xi_i)$, où ζ_i est compris entre $\bar{x}_{i-1}$ et $x^*(t_{i-1})$ et où $\xi_i \in [t_{i-1}, t_i]$.

Preuve
Ecrivons l'erreur commise au pas i. On a

$$EG_i = \bar{x}_i - x^*(t_i) = \bar{x}_{i-1} + hf(\bar{x}_{i-1}, t_{i-1}) - x^*(t_i) \tag{4.2}$$

$$EG_i = \bar{x}_{i-1} + hf(\bar{x}_{i-1}, t_{i-1}) - (x^*(t_{i-1}) + hf(x^*(t_{i-1}), t_{i-1}) + \frac{h^2}{2}(x^*)''(\xi_i)) \tag{4.3}$$

$$EG_i = EG_{i-1} + hf(\bar{x}_{i-1}, t_{i-1}) - f(x^*(t_{i-1}), t_{i-1}) - \frac{h^2}{2}(x^*)''(\xi_i)$$

$$EG_i = EG_{i-1} + h(\bar{x}_{i-1} - x^*(t_{i-1}))\frac{df}{dx}(\zeta_i, t_{i-1}) + EL_i \tag{4.4}$$

$$EG_i = EG_{i-1}(1 + hJ_i) + EL_i$$

où (4.2) est obtenue en exprimant comment $\bar{x}_i$ est obtenu en utilisant la méthode d'Euler explicite, (4.3) est obtenue en développant $x^*(t)$ en série de Taylor autour de t_{i-1}, et (4.4) est obtenue en appliquant le théorème des accroissements finis à la première composante de f. ■

Il est important de comprendre ce que signifie exactement la proposition précédente. L'exemple suivant tente de clarifier la situation.

Exemple 4.2.
Soit le problème suivant

$$\begin{cases} x'(t) &= -x(t)^2 + t \\ x(0) &= 2 \end{cases}$$

que nous avons déjà considéré dans l'exemple précédent. Lors de la première itération, on obtient $\bar{x}_1 = 2 + 0.3f(2, 0) = 0.8$. L'erreur obtenue ici est uniquement locale, c'est-à-dire qu'elle peut être exclusivement interprétée par l'intermédiaire du développement de Taylor. Dans ce cas, l'erreur peut être approximée par 0.48. A la deuxième itération, la méthode d'Euler calcule

$$\bar{x}_{2,0.3} = 0.8 + 0.3f(0.8, 0.3) = 0.698.$$

Dans ce cas-ci, une erreur s'est introduite par rapport à la résolution du problème $x'(t) = -x^2(t) + t$, $x(0.3) = 0.8$. En effet, la solution en 0.6 de ce problème est 0.76, ce qui implique qu'une erreur de 0.04 a été introduite en plus à la deuxième étape. Mais le point important à remarquer est que nous avons résolu l'équation différentielle pour la condition initiale $x(0.3) = 0.8$ au lieu de $x(0.3) = 1.28$. En d'autres termes, nous avons utilisé, dans l'approximation de Taylor, une pente de $f(0.8, 0.3)$ au lieu de $f(1.28, 0.3)$, ce qui fait une erreur approximative de 1 dans la pente utilisée. Cette erreur implique une accumulation d'erreurs venant des itérations précédentes.

Nous venons de voir dans l'exemple précédent qu'une partie importante de l'erreur commise en utilisant la méthode d'Euler explicite provient des erreurs commises lors des itérations précédentes. On en vient à présent au choix du pas de la méthode. Dans ce cas-ci, on choisira donc un pas de façon à ce que l'erreur qui se propage d'une itération à l'autre s'amenuise (comme dans le cas de l'exemple) au lieu de croître. Dans le cas où les erreurs propagées d'une itération à l'autre restent sous contrôle, on dit que la méthode est stable. Si, au contraire, les erreurs provenant des itérations précédentes croissent, on dit que la méthode est instable.

Proposition 4.2. *La méthode d'Euler explicite est stable si on a*

$$-2 < hJ_i < 0 \text{ pour tout } i.$$

Preuve
L'erreur globale à l'itération i est EG_i et est donnée par $EG_i = (1 + hJ_i)EG_{i-1} + EL_i$. On aura en particulier

$$EG_i = (1 + hJ_i)(1 + hJ_{i-1}) \dots (1 + hJ_2)EL_1 + \dots + (1 + hJ_i)EL_{i-1} + EL_i.$$

Pour que tous les termes tendent vers 0, il faut donc $|1 + hJ_i| < 1$ ce qui est équivalent au résultat annoncé. ■

On voit donc que la méthode d'Euler explicite n'est jamais stable lorsque l'équation différentielle n'est elle-même pas stable. Par contre, et c'est plus surprenant, il faut choisir un pas extrêmement petit lorsque l'équation différentielle est fortement stable, c'est-à-dire lorsque $J_i \ll 0$. Les équations très stables sont donc également des problèmes particulièrement ardus pour les méthodes numériques traditionnelles.

4.2 Méthodes d'ordre supérieur

Rien n'empêche de construire un développement de Taylor comportant plus de termes afin d"obtenir une approximation plus précise de l'itéré $\bar{x}_{i+1}$ en fonction de $\bar{x}_i$. Nous allons voir que ceci implique une connaissance approfondie de la fonction f et que ce n'est pas toujours très praticable. En effet, si on écrit le développement de Taylor de $x(t+h)$ autour du point t, on obtient

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \frac{h^3}{3!}x'''(t) + \dots$$

$$x(t+h) = x(t) + hf(x(t),t) + \frac{h^2}{2}\frac{df}{dt}(x(t),t) + \frac{h^3}{3!}\frac{d^2f}{dt^2}(x(t),t) + \dots$$

$$x(t+h) = x(t) + hf(x(t),t) + \frac{h^2}{2}(\frac{df}{dx}(x(t),t)f(x(t),t) + \frac{df}{dt}(x(t),t)) + \dots \tag{4.5}$$

$$x(t+h) = x(t) + hf(x(t),t) + \frac{h^2}{2}(\frac{df}{dx}(x(t),t)f(x(t),t) + \frac{df}{dt}(x(t),t)) +$$
$$\frac{h^3}{3!}\left(\frac{d^2f}{dx^2}f^2 + 2\frac{d^2f}{dxdt}f + \frac{d^2f}{dt^2} + \frac{df}{dx}\frac{df}{dt} + (\frac{df}{dx})^2 f\right)(x(t),t) + \dots \tag{4.6}$$

L'expression (4.5) donne la méthode de Taylor d'ordre 2. L'expression (4.6) donne l'expression d'ordre 3.

Méthode 2 (Méthode de Taylor d'ordre 2)
Après calcul préalable de $\frac{df}{dx}$ et de $\frac{df}{dt}$, on calcule successivement les itérés

$$\bar{x}_{i+1} = \bar{x}_i + hf(\bar{x}_i,t_i) + \frac{h^2}{2}(\frac{df}{dx}(\bar{x}_i,t_i)f(\bar{x}_i,t_i) + \frac{df}{dt}(\bar{x}_i,t_i))$$

Méthode 3 (Méthode de Taylor d'ordre 3)
Après calcul préalable des différentes dérivées partielles premières et secondes, on calcule successivement les itérés

$$\bar{x}_{i+1} = \bar{x}_i + hf(\bar{x}_i,t_i) + \frac{h^2}{2}(\frac{df}{dx}(\bar{x}_i,t_i)f(\bar{x}_i,t_i) + \frac{df}{dt}(\bar{x}_i,t_i))$$

$$\frac{h^3}{3!}\left(\frac{d^2f}{dx^2}f^2 + 2\frac{d^2f}{dxdt}f + \frac{d^2f}{dt^2} + \frac{df}{dx}\frac{df}{dt} + (\frac{df}{dx})^2 f\right)(\bar{x}_i,t_i) + \dots$$

On le voit, la complexité de ces formules croît très rapidement. Pour pouvoir appliquer ces méthodes, il faudra donc passer au préalable par une étape de dérivation symbolique.

Dans la pratique, les méthodes de Taylor d'ordre supérieur à 1 sont très peu utilisées. Ceci dit, avec la venue de logiciels de calcul symbolique, il n'est pas inintéressant de considérer ces méthodes dans certaines applications où la dérivation symbolique est possible. Finalement, il est également possible d'analyser les conditions de stabilité de telles méthodes. Au fur et à mesure que l'ordre du développement de Taylor considéré augmente, la région de stabilité augmente également. Il n'y a pas, ceci dit, de différence drastique avec la méthode d'Euler explicite.

4.3 La méthode d'Euler implicite

On peut adapter la méthode d'Euler dite explicite de façon à ce qu'elle adopte un comportement beaucoup plus stable. Cependant, il y a malheureusement un lourd coût à payer au niveau du temps de calcul à effectuer à chaque itération. Souvenons-nous que pour déterminer la méthode d'Euler explicite, nous avons simplement écrit un développement de Taylor autout du point t, pour en déduire une expression de $x(t+h)$. L'idée de la méthode d'Euler implicite est d'écrire le développement en $t+h$ plutôt qu'en t. On a donc $x(t) = x(t+h) - hx'(t+h) + \ldots$ et dès lors

$$x(t+h) \approx x(t) + hf(x(t+h), t+h). \tag{4.7}$$

Le problème dans (4.7) est évidemment que l'on ne connaît pas $f(x(t+h), t+h)$ si on ne connaît pas encore $x(t+h)$. C'est la raison pour laquelle la méthode est qualifiée d'implicite puisqu'il faudra résoudre une équation non linéaire à chaque pas de temps.

Méthode 4 (Méthode d'Euler implicite)
L'itéré $\bar{x}_{i+1}$ est obtenu comme étant une solution de l'équation

$$\bar{x}_{i+1} = \bar{x}_i + hf(\bar{x}_{i+1}, t_{i+1}).$$

Exemple 4.3.
Soit à nouveau le problème suivant

$$\begin{cases} x'(t) &= -x(t)^2 + t \\ x(0) &= 2 \end{cases}$$

On considère une itération de l'algorithme d'Euler implicite. On part de $\bar{x}_0 = 2$ et on recherche $\bar{x}_1$ tel que

$$\bar{x}_0 = \bar{x}_1 - 0.3(-\bar{x}_1^2 + 0.3).$$

Dans ce cas-ci, on voit qu'il suffit de résoudre l'équation non linéaire $0.3\bar{x}_1^2 + \bar{x}_1 - 2.09 = 0$. Deux solutions sont possibles, $\bar{x}_1 = 1.254$ ou $\bar{x}_1 = -2.254$. En choisissant la solution la plus proche de $\bar{x}_0$, on obtient donc $\bar{x}_1 = 1.254$. Cette fois, l'erreur n'est plus que de 0.03.

La méthode semble donc être une bonne alternative. Malheureusement, la résolution d'une équation non linéaire à chaque pas rend son utilisation impraticable. On peut malgré tout analyser la stabilité de la méthode. Une analyse similaire au cas de la

méthode d'Euler explicite nous mène à la proposition suivante que nous énonçons sans démonstration.

Proposition 4.3. *La méthode d'Euler implicite est stable si*

$$\left|\frac{1}{1-hJ_i}\right| < 1 \text{ pour tout } i.$$

On voit, en particulier, que lorsque l'équation est stable ($J_i < 0$), la méthode est stable pour tout choix de pas h. La méthode d'Euler implicite admet donc des conditions de stabilité très robustes. Pour des valeurs positives très grandes de J_i, c'est-à-dire pour un problème très instable, il semblerait que la méthode d'Euler implicite soit également stable. Ceci n'a évidemment aucune valeur car la stabilité apparente de la méthode numérique n'aura rien à voir avec la solution analytique.

5 Méthodes de Runge-Kutta

Les méthodes de Runge-Kutta sont aux méthodes de Taylor ce que la méthode de la sécante est à la méthode de Newton dans le cadre de la résolution d'équations non linéaires. On se souvient que la méthode de la sécante approxime numériquement la dérivée nécessaire à la méthode de Newton. Dans le cadre d'équations différentielles ordinaires, nous avons vu que les méthodes de Taylor requièrent une lourde phase de différentiation analytique. Les méthodes de Runge-Kutta vont remplacer cette partie par une approximation numérique des différentes dérivées partielles.

Méthode 5 (Runge-Kutta d'ordre 2)
Les différents itérés de la méthode de Runge-Kutta d'ordre 2 sont obtenus par le processus

$$x_{i+1} = \bar{x}_i + \frac{h}{2}f(\bar{x}_i, t_i) + \frac{h}{2}f(\bar{x}_i + hf(\bar{x}_i, t_i), t_{i+1}) \tag{5.1}$$

Explication:
L'idée de la méthode est que l'on va copier le plus possible de termes du développement de Taylor de $x(t+h)$ en utilisant le calcul de f en deux points seulement, à savoir $F_1 = f(x(t), t)$ et $F_2 = f(x(t) + \beta h f(x(t), t), t + \alpha h)$ où α et β sont inconnus. Ces deux points sont utilisés en écrivant

$$x(t+h) \approx x(t) + w_1 h F_1 + w_2 h F_2 \tag{5.2}$$

où w_1 et w_2 sont également inconnus. La suite de cette "explication" est donc de montrer qu'on peut déterminer $\alpha, \beta, , w_1, w_2$ de manière à ce que la formule (5.2) se rapproche le plus possible du développement de Taylor de $x(t+h)$. Pour ce faire, nous allons tout d'abord faire un développement de Taylor tronqué à l'ordre 1 de F_2. On a

$$F_2 = f(x(t) + \beta h f(x(t), t), t + \alpha h) \tag{5.3}$$

$$F_2 \approx f(x(t),t) + \beta h f(x(t),t)\frac{df}{dx}(x(t),t) + \alpha h \frac{df}{dt}(x(t),t). \tag{5.4}$$

Si on utilise (5.4) dans (5.1), on trouve l'approximation

$$x(t+h) \approx x(t)+(w_1+w_2)hf(x(t),t)+\alpha w_2 h^2 \frac{df}{dt}(x(t),t)+\beta h^2 w_2 f(x(t),t)\frac{df}{dx}(x(t),t). \tag{5.5}$$

Par ailleurs, nous avons vu lors de l'exposition des méthodes de Taylor, qu'une expression d'ordre 2 de $x(t+h)$ est

$$x(t+h) \approx x(t) + hf(x(t),t) + \frac{h^2}{2}(f(x(t),t)\frac{df}{dx}(x(t),t) + \frac{df}{dt}(x(t),t)). \tag{5.6}$$

Si on compare (5.5) à (5.6), on voit qu'il faut avoir

$$w_1 + w_2 = 1, \quad \alpha w_2 = \frac{1}{2}, \quad \beta w_2 = \frac{1}{2}. \tag{5.7}$$

Une solution possible et pratique à (5.7) est de choisir $\alpha = \beta = 1$ et $w_1 = w_2 = \frac{1}{2}$. Remarquons que la méthode proposée n'est pas la seule qui pourrait donner un ordre 2. On pourrait par exemple choisir pour satisfaire (5.7) $\alpha = \beta$ et $w_1 = 1 - \frac{1}{2\alpha}$ et $w_2 = \frac{1}{2\alpha}$. Dans la pratique, les méthodes de Runge-Kutta d'ordre 2, bien que très simples à mettre en oeuvre sont assez peu utilisées car leur erreur n'est que de $\mathcal{O}(h^3)$. La méthode de Runge-Kutta la plus utilisée est celle d'ordre 4. Déterminer une telle formule est un travail très fastidieux que nous ne détaillerons pas ici. Nous présentons la formule de la méthode sans l'expliquer.

Méthode 6 (Runge-Kutta d'ordre 4)
Les différents itérés de la méthode de Runge-Kutta d'ordre 4 sont obtenus par le processus

$$\bar{x}_{i+1} = \bar{x}_i + \frac{1}{6}(K_1 + K_2 + K_3 + K_4)$$

où

$$K_1 = hf(\bar{x}_i, t_i)$$
$$K_2 = hf(\bar{x}_i + \frac{1}{2}K_1, t_i + \frac{1}{2}h)$$
$$K_3 = hf(\bar{x}_i + \frac{1}{2}K_2, t_i + \frac{1}{2}h)$$
$$K_4 = hf(\bar{x}_i + K_3, t_i + h)$$

Comme son nom l'indique, la méthode de Runge-Kutta d'ordre 4 copie le développement de Taylor jusqu'aux termes d'ordre 4. Le terme d'erreur est donc en $\mathcal{O}(h^5)$.

6 Méthodes adaptatives de Runge-Kutta-Fehlberg

Comme on l'a vu précédemment, il est souvent difficile de déterminer le pas à choisir pour assurer une stabilité de la méthode numérique tout en conservant une quantité limitée de calculs. En général, on aimerait que l'utilisateur puisse déterminer une tolérance dans laquelle la solution doit se trouver. Mais même en ayant accès à l'erreur locale commise par une méthode, il est souvent difficile de déterminer le pas à utiliser. Il se pourrait qu'il soit nécessaire de choisir un pas très petit sur certaines portions du problème alors que l'on pourrait se contenter de pas plus grands sur d'autres portions. Pour cette raison, plusieurs méthodes de choix automatiques du pas ont été imaginées.

Pour comprendre le principe des méthodes de Runge-Kutta-Fehlberg, imaginons tout d'abord le principe suivant. On considère la méthode de Runge-Kutta d'ordre 4 avec un pas h. On peut aussi considérer la même méthode avec un double pas de $h/2$. Si le pas h est satisfaisant, la différence entre l'approximation obtenue avec un pas h ou deux pas de $h/2$ sera très faible. Dans ce cas, le pas h est suffisant. Dans le cas contraire, il faudra réduire le pas. Le problème de cette méthode est qu'elle nécessite quatre appels à la fonction f pour le pas h et 7 autres appels pour le double pas de $h/2$. Cela fait un total de 11 appels à la fonction f par itération, ce qui peut s'avérer coûteux en temps de calcul dans certaines applications. Or, nous avons vu dans la section précédente qu'il y a une certaine flexibilité dans le choix des coefficients des méthodes de Runge-Kutta. L'idée est de chosir une méthode de Runge-Kutta d'ordre 5 et une méthode d'ordre 4 qui partagent le plus possible d'évaluations communes de f de façon à minimiser la quantité de travail à chaque itération. L'avantage est de disposer de deux évaluations de $x(t+h)$. En comparant les deux évaluations, nous pouvons ainsi décider si le pas h est adapté ou pas. La méthode suivante est l'algorithme implémenté dans la fonction ode45 de matlab.

Méthode 7 (Runge-Kutta-Fehlberg d'ordres 4 et 5)

$$K_1 = hf(\bar{x}_i, t_i)$$

$$K_2 = hf(\bar{x}_i + \frac{1}{4}K_1, t_i + \frac{1}{4}h)$$

$$K3 = hf(\bar{x}_i + \frac{3}{32}K_1 + \frac{9}{32}K_2, t_i + \frac{3}{8}h)$$

$$K_4 = hf(\bar{x}_i + \frac{1932}{2197}K_1 - \frac{7200}{2197}K_2 + \frac{7296}{2197}K_3, t_i + \frac{12}{13}h)$$

$$K_5 = hf(\bar{x}_i + \frac{439}{216}K_1 - 8K_2 + \frac{3680}{513}K_3 - \frac{845}{4104}K_4, t_i + h)$$

$$K_6 = hf(\bar{x}_i - \frac{8}{27}K_1 + 2K_2 - \frac{3544}{2565}K_3 + \frac{1859}{4104}K_4 - \frac{11}{40}K_5, t_i + \frac{1}{2}h)$$

On obtient deux approximations de $x(t+h)$, à savoir

$$\bar{x}_{i+1}^{[4]} = x(t) + \frac{25}{216}K_1 + \frac{1408}{2565}K_3 + \frac{2197}{4104}K_4 - \frac{1}{5}K_5$$

$$\bar{x}_{i+1}^{[5]} = x(t) + \frac{16}{135}K_1 + \frac{6656}{12825}K_3 + \frac{28561}{56430}K_4 - \frac{9}{50}K_5 + \frac{2}{55}K_6$$

qui sont respectivement une approximation d'ordre 4 et d'ordre 5 obtenues à l'aide de 6 évaluations de fonction. La différence $|\bar{x}_{i+1}^{[5]} - \bar{x}_{i+1}^{[4]}|$ est une estimation de l'erreur en t_{i+1}.

7 Méthodes à pas liés

Jusqu'à présent, nous avons uniquement analysé des méthodes à pas séparés. Une méthode est à pas séparés lorsque l'on se sert uniquement de l'intervalle $[t, t+h]$ et de l'expression de f dans celui-ci pour calculer la nouvelle valeur $x(t+h)$. L'idée d'une méthode à pas liés est que l'on peut se servir de la connaissance des points précédemment calculés afin d'avoir une meilleure perception de la manière dont f se comporte et ceci sans devoir procéder à une différentiation analytique qui s'avérerait trop lourde. La forme générique d'une méthode à pas liés peut être formulée comme suit.

Méthode 8 (Méthode à pas liés)
Si on dénote par $\bar{x}_i$ les différentes approximations obtenues par la méthode aux points t_i, on calcule successivement

$$\bar{x}_{i+1} = \bar{x}_i + h \sum_{j=-1}^{n} \beta_j f(\bar{x}_{i-j}, t_{i-j}). \tag{7.1}$$

Dans (7.1), on remarque que j peut prendre la valeur -1 ce qui correspond à considérer que pour obtenir la valeur $\bar{x}_{i+1}$, on se sert de la valeur $\bar{x}_{i+1}$. On reconnaît là le principe d'une méthode implicite. Si $\beta_{-1} = 0$, on ne se sert que de points connus pour calculer la nouvelle valeur $\bar{x}_{i+1}$, il s'agit alors d'une méthode explicite.
Pour calculer les coefficients d'une formule de type (6.1), on doit évaluer l'intégrale

$$\bar{x}_{i+1} = \bar{x}_i + \int_{t_i}^{t_{i+1}} f(x(s), s)ds.$$

En particulier, pour obtenir les coefficients de (7.1) dans le cas d'une méthode explicite, on peut écrire le polynôme qui interpole les $n+1$ points obtenus lors des itérations précédentes $(\bar{x}_{i-n}, f(\bar{x}_{i-n}, t_{i-n})), \ldots, (\bar{x}_i, f(\bar{x}_i, t_i))$ et l'intégrer sur l'intervalle $[t_i, t_{i+1}]$. Nous donnons les méthodes explicites et implicites d'ordre 2 et 3 à titre informatif. Remarquons que les méthodes explicites à pas liés sont appelées méthodes d'Adams-Bashforth et les méthodes implicites méthodes d'Adams-Moulton.

Méthode 9 (Adams-Bashforth d'ordre 2)

$$\bar{x}_{i+1} = \bar{x}_i + \frac{h}{2}(-f(\bar{x}_{i-1}, t_{i-1}) + 3f(\bar{x}_i, t_i))$$

Méthode 10 (Adams-Bashforth d'ordre 3)

$$\bar{x}_{i+1} = \bar{x}_i + \frac{h}{12}(5f(\bar{x}_{i-2}, t_{i-2}) - 16f(\bar{x}_{i-1}, t_{i-1}i) + 23f(\bar{x}_i, t_i))$$

Méthode 11 (Adams-Moulton d'ordre 2)

$$\bar{x}_{i+1} = \bar{x}_i + \frac{h}{2}(f(\bar{x}_i, t_i) + f(\bar{x}_{i+1}, t_{i+1}))$$

Méthode 12 (Adams-Moulton d'ordre 3)

$$\bar{x}_{i+1} = \bar{x}_i + \frac{h}{2}(-f(\bar{x}_{i-1}, t_{i-1}) + 8f(\bar{x}_i, t_i) + 5f(\bar{x}_{i+1}, t_{i+1}))$$

L'intérêt des méthodes à pas liés est qu'elles n'utilisent qu'une seule évaluation de la fonction f à chaque pas d'intégration. Cela peut s'avérer un gain de temps conséquent par rapport à une méthode de Runge-Kutta d'ordre élevé qui requiert un grand nombre d'évaluations à chaque pas, et ce, surtout lorsque l'évaluation de la fonction est assez coûteuse. Dans le cadre des méthodes de Runge-Kutta, nous avons vu l'amélioration adaptative proposée par Fehlberg. Les méthodes à pas liés se prêtent également très bien à une version adaptative ou prédicteur-correcteur.
La méthode prédicteur-correcteur consiste à utiliser une méthode explicite et implicite conjointement. On va ainsi se servir de l'approximation donnée par la méthode explicite comme $\bar{x}_{i+1}$ dans la formule implicite. On évitera ainsi la coûteuse phase de résolution d'une équation non linéaire. La version adaptative consiste à utiliser la différence entre la sortie de la formule explicite et de la formule implicite pour savoir s'il faut considérer un changement de la taille du pas.

Exemple 7.1.
Soit à nouveau le problème suivant

$$\begin{cases} x'(t) &= -x(t)^2 + t \\ x(0) &= 2 \end{cases}$$

On peut remarquer que les méthodes d'Euler implicite et explicite sont en réalité les méthodes à pas liés d'ordre 1. Nous allons ici uniquement montrer comment on peut, pour l'ordre 1, mettre en pratique la méthode prédicteur-correcteur. Rappelons les formules d'Euler explicite $\bar{x}_{i+1} = \bar{x}_i + hf(x_i, t_i)$ et implicite $\bar{x}_{i+1} = \bar{x}_i + hf(\bar{x}_i, t_{i+1})$. Dans notre cas, et pour un pas de 0.3, on obtient le prédicteur donné par Euler explicite $\bar{x}_{i+1} = 2 - 1.2 = 0.8$. Le correcteur est ensuite donné en utilisant la première approximation comme $\bar{x}_{i+1} = 2 + 0.3f(0.8, 0.3) = 2 - 0.102 = 1.898$. Remarquons qu'ici, vu la différence entre les deux approximations obtenues, il serait judicieux de réduire le pas.

Remarquons également qu'il est possible d'itérer plusieurs fois le processus prédicteur-correcteur afin d'obtenir une approximation plus précise. La pratique montre cependant qu'une seule itération suffit à donner de très bonnes approximations de la valeur réelle.

8 Exercices résolus

8.1

Considérons le problème de Cauchy

$$\begin{cases} y'(t) &= y(t) + t \\ y(0) &= 1, \end{cases}$$

Approcher à 10^{-3} la solution du problème en $t = 1$ à l'aide de la méthode d'Euler en subdivisant l'intervalle [0, 1] en 10 parties égales, sachant que la solution exacte du problème ci dessus est $y(t) = 2\exp(t) - t - 1$.

Solution.
Par l'algorithme d'Euler $0 \leq n \leq 9$ et $h = 0.1$: on trouve $y(1) \simeq 3.187$ et la solution exacte de l'équation (1) est donnée par l'équation $y(t) = 2\exp(t) - t - 1$. Ce qui donne $y(1) \simeq 3.437$. L'approximation calculée est donc très grossière.

8.2

Stabilité de la méthode d'Euler explicite en fonction du pas
On considère le problème de Cauchy

$$\begin{cases} y'(t) &= -y(t) \\ y(0) &= 1, \end{cases}$$

sur l'intervalle [0, 10].

1. Calculer la solution exacte du problème de Cauchy.
2. Soit Δt le pas temporel. Écrire la méthode d'Euler explicite pour cette équation différentielle ordinaire (EDO).
3. En déduire une forme du type

   $$y_{k+1} = g(\Delta t, k)$$

   avec $g(\Delta t, k)$ à préciser (autrement dit, l'itérée en t_k ne dépend que de Δt et k et ne dépend pas de y_k).
4. Utiliser la formulation ainsi obtenue pour tracer les solutions
   - exacte,
   - obtenue avec la méthode d'Euler avec $\Delta t = 2.5$,

- obtenue avec la méthode d'Euler avec $\Delta t = 1.5$,
- obtenue avec la méthode d'Euler avec $\Delta t = 0.5$.

5. Que peut-on en déduire sur la stabilité de la méthode ?

Solution.
1. Il s'agit d'une EDO à variables séparables. L'unique solution constante de l'EDO est la fonction $y(t) \equiv 0$, toutes les autres solutions sont du type $y(t) = Ce^{-t}$. Donc l'unique solution du problème de Cauchy est la fonction $y(t) = e^{-t}$ définie pour tout $t \in \mathbb{R}$.

2. La méthode d'Euler est une méthode d'intégration numérique d'EDO du premier ordre de la forme $y'(t) = F(t, y(t))$ C'est une méthode itérative : la valeur y à l'instant $t + \Delta t$ se déduisant de la valeur de y à l'instant t par l'approximation linéaire

$$y(t + \Delta t) \approx y(t) + y'(t)\Delta t = y(t) + F(t, y(t))\Delta t.$$

En choisissant un pas de discrétisation Δt , nous obtenons une suite de valeurs (t_k, y_k) qui peuvent être une excellente approximation de la fonction $y(t)$ avec

$$\begin{cases} t_k &= t_0 + k\Delta t \\ y_k &= y_{k-1} + F(t_{k-1}, y_{k-1})\Delta t \end{cases}$$

La méthode d'Euler explicite pour cette EDO s'écrit donc

$$y_{k+1} = (1 - \Delta t)y_k.$$

3. En procédant par récurrence sur k, on obtient

$$y_{k+1} = (1 - \Delta t)^{k+1}.$$

4. On a donc

- si $\Delta t = 2.5$ alors $y_k = \left(-\frac{3}{2}\right)^k$
- si $\Delta t = 1.5$ alors $y_k = \left(-\frac{1}{2}\right)^k$
- si $\Delta t = 0.5$ alors $y_k = \left(\frac{1}{2}\right)^k$

5. De la formule $y_{k+1} = (1 - \Delta t)^{k+1}$ on déduit que

- si $0 < \Delta t < 1$ alors la solution numérique est stable et convergente,
- si $1 < \Delta t < 2$ alors la solution numérique oscille mais est encore convergente,
- si $\Delta t > 2$ alors la solution numérique oscille et divergente.

En effet, on sait que la méthode est absolument stable si et seulement si $|1 - \Delta t| < 1$.

Remarque 8.1.
la suite obtenue est une suite géométrique de raison $q = 1 - \Delta t$. On sait qu'une telle suite

- diverge si $|q| > 1$ ou $q = -1$,
- est stationnaire si $q = 1$,
- converge vers 0 si $|q| < 1$.

8.3

Méthode de Taylor
La méthode de Taylor est basée sur la relation

$$y(x+h) \approx y(x) + y'(x)h + \frac{1}{2!}y''(x)h^2 + \frac{1}{3!}y'''(x)h^3 + \ldots + \frac{1}{m!}y^{(m)}(x)h^m$$

Cette relation prédit $y(x+h)$ à partir de $y(x)$, ainsi elle permet d'écrire une formule d'intégration numérique. Le dernier terme indique l'ordre de la méthode et l'erreur de troncature, due aux termes omis, est

$$E = \frac{1}{(m+1)!}y(m+1)(\xi)h^{m+1} \text{ pour } x < \xi < x+h,$$

que l'on peut approcher par

$$E \approx \frac{h^m}{(m+1)!}(y^{(m)}(x+h) - y^{(m)}(x)).$$

Considérons le problème de Cauchy

$$\begin{cases} y'(x) + 4y(x) &= x^2 \\ y(0)=1 \end{cases}$$

Estimer $y(0.1)$ par la méthode de Taylor d'ordre 4 avec un seul pas d'intégration.

Solution
Le développement de Taylor en 0 jusqu'à l'ordre 4 est

$$y(h) \approx y(0) + y'(0)h + \frac{1}{2!}y''(0)h^2 + \frac{1}{3!}y'''(0)h^3 + \frac{1}{4!}y^{(4)}(0)h^4$$

En dérivant l'EDO on trouve $y(0) = 1$,
$y'(x) = -4y(x) + x^2, \quad y'(0) = -4$
$y''(x) = -4y'(x) + 2x, \quad y''(0) = 16$
$y'''(x) = -4y''(x) + 2, \quad y'''(0) = -62$
$y^{(4)}(x) = -4y'''(x), \quad y^{(4)}(0) = 248$
donc, pour $x = 0$ et $h = 0.1$, on obtient

$$y(0.1) \approx 1 + \frac{-4}{10} + \frac{16}{200} + \frac{-62}{6000} + \frac{248}{240000} = 0.6707$$

et comme

$$y^{(4)}(x+h) = -4y'''(x) = -4(-4y''(x)+2) = (-4(-4y'(x)+2x)+2) = (-4(-4(-4y(x)+x^2)+2x)+2)$$

alors $y^{(4)}(0.1) \approx (-4(-4(-4\times 0.6707+0.1^2)+0.2)+2) = 166.259$ et on obtient l'estimation de l'erreur

$$E \approx \frac{248}{960000}(y^{(4)}(0.1) - y^{(4)}(0)) = \frac{248}{960000}(166.259 - 248) = -0.000068.$$

8.4

Soit le problème de Cauchy

$$\begin{cases} y'(t) &= y(t) - \frac{-2t}{y(t)} \\ y(0) &= 1 \end{cases}$$

dont la solution exacte est $y = \sqrt{2x+1}$: Approcher la solution du système précédent en $t = 0.2$ en exécutant les calculs avec 6 décimales à l'aide des méthodes d'Euler modifié et de Runge-Kutta d'ordre 2.

Solution
La solution exacte étant $y = \sqrt{2x+1}$, on considère l'intervalle $[t_0, t_1]$ avec $t_0 = 0$, $t_1 = t_0 + h = 0.2$ et $h = 0.2$.
Méthode de Runge-Kutta d'ordre 2:
$\hat{y} = y_0 + hf(t_0, y_0)$, avec $f(t_0, y_0) = y_0 - \frac{2t_0}{y_0} = 1$ donc $\hat{y} = 1.2$.
$f(t_1, y_0 + hf(t_0, y_0)) = f(0.2, 1.2) = 1.2 - \frac{2\times 0.2}{1.2} = 0.866667$.
$y_1 = y_0 + \frac{h}{2}(f(t_0, y_0) + f(t_1, \hat{y})) = 1 + \frac{0.2}{2}(1 + 0.866667) = 1.1866667$. Ainsi $y(0.2) \approx y_1 = 1.1866667$. La valeur exacte étant $y(0.2) = \sqrt{1.4} = 1.183216$.
L'erreur commise est $|y(0.2) - y_1| = 0.003451 < 0.510^{-}2$. D'où $y(0.2) = 1.19 \pm 0.01$, ainsi y_1 approche $y(0.2)$ avec 3 chiffres significatifs exacts.

Bibliographie

• M. Atteia, M., Pradel, M. , Eléments d'analyse numérique, Ceradues-Editions.

• Baranger, J. , Introduction à l'analyse numérique, Ed. Hermann 1977.

• Boumahrat, M. , Bourdin, A. Méthodes numériques appliquées. Ed. OPU 1983.

• Démodovitch, B. Maron, I. , Eléments de calcul numérique, Ed. Mir Mosco.

• Ciarlet, Ph. G. Introduction à l'analyse numérique matricielle et à l'optimisation, Dunod, Paris 1998.

• Curtis F. Gerald, P. O. Wheatdey : Applied Numerical Analysis, Addison-Wesley Pub. Compagny.

• Lascaux, P. Theodor, R. Analyse numérique matricielle appliquée à l'art d'ingénieur, Tomes I et II, Masson, Paris.

• Meurant, G. Résolution numérique des grands systèmes, Ed. Stanford University.

• Lascaux,P. , Theodor, R. Analyse numérique matricielle appliquée à l'art d'ingénieur Tomes I et II, Masson, Paris.