ELEMENTS D’ANALYSE NUMERIQUE
AVEC APPLICATIONS AUX MODELES
BIOMATHEMATIQUES ET ECOLOGIQUES

Abdesslam BOUTAYEB
Abdelaziz CHETOUANI
Mohamed DEROUICH
# Table des matières

## 1 Rappels et définitions

1.1 Notations .................................................. 6
1.2 Quelques rappels sur les matrices .......................... 6
1.3 Valeurs et vecteurs propres ................................. 8
1.4 Normes vectorielles et normes matricielles ............. 11
  1.4.1 Normes vectorielles .................................. 11
  1.4.2 Normes matricielles .................................. 12
  1.4.3 Normes compatibles .................................. 12
1.5 Applications ............................................... 15
1.6 Complément bibliographique sur les matrices positives . 21
1.7 Exercices .................................................. 22

## 2 Méthodes directes de résolution des systèmes linéaires $Ax = b$

2.1 Résolution d’un système par les méthodes de descente ou de remontée 25
2.2 Matrices élémentaires ...................................... 26
  2.2.1 Matrices élémentaires de Gauss ...................... 26
  2.2.2 Matrices élémentaires de Danilevski ................. 26
  2.2.3 Matrices élémentaires de Householder .............. 26
  2.2.4 Matrices élémentaires de permutation .............. 27
  2.2.5 Matrices élémentaires de Perlis ................... 27
  2.2.6 Matrices élémentaires de Givens (ou de rotation) 29
2.3 Méthodes de Gauss ....................................... 29
  2.3.1 Méthode de Gauss sans pivot ....................... 29
  2.3.2 Méthode de Gauss avec pivot partiel ............... 30
  2.3.3 Méthode de Gauss avec pivot total ................. 32
  2.3.4 Méthode de Gauss-Jordan ............................ 32
2.4 Factorisation $LU$ ......................................... 32
2.5 Factorisation de Choleski (matrice symétrique) ........ 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6 Factorisation de Householder (matrice unitaire)</td>
<td>35</td>
</tr>
<tr>
<td>2.7 Conditionnement</td>
<td>36</td>
</tr>
<tr>
<td>2.8 Matrices creuses</td>
<td>37</td>
</tr>
<tr>
<td>2.9 Résultats sur les matrices non carrées</td>
<td>43</td>
</tr>
<tr>
<td>2.10 Résolution des systèmes à matrices non carrées</td>
<td>44</td>
</tr>
<tr>
<td>2.11 Conclusion</td>
<td>48</td>
</tr>
<tr>
<td>2.12 Exercices</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Méthodes indirectes de résolution des systèmes linéaires $Ax=b$</td>
<td>51</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>51</td>
</tr>
<tr>
<td>3.2 Généralités et définitions</td>
<td>52</td>
</tr>
<tr>
<td>3.3 Description des méthodes classiques</td>
<td>54</td>
</tr>
<tr>
<td>3.3.1 Méthode de Jacobi</td>
<td>54</td>
</tr>
<tr>
<td>3.3.2 Méthode de Gauss-Seidel</td>
<td>56</td>
</tr>
<tr>
<td>3.3.3 Méthode de relaxation</td>
<td>58</td>
</tr>
<tr>
<td>3.4 Comparaison des méthodes classiques</td>
<td>61</td>
</tr>
<tr>
<td>3.4.1 Comparaison des méthodes de Jacobi et de Gauss-Seidel</td>
<td>61</td>
</tr>
<tr>
<td>3.4.2 Comparaison des méthodes de Jacobi et de relaxation</td>
<td>62</td>
</tr>
<tr>
<td>3.5 Méthodes semi-itératives</td>
<td>68</td>
</tr>
<tr>
<td>3.6 Décomposition des matrices positives</td>
<td>69</td>
</tr>
<tr>
<td>3.6.1 Décomposition régulière des matrices</td>
<td>71</td>
</tr>
<tr>
<td>3.7 Comparaison des méthodes classiques dans le cas des matrices posi-</td>
<td>72</td>
</tr>
<tr>
<td>tives</td>
<td></td>
</tr>
<tr>
<td>3.8 Complément bibliographique</td>
<td>73</td>
</tr>
<tr>
<td>3.9 Exercices</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Solutions numériques de l’équation $f(x) = 0$</td>
<td>82</td>
</tr>
<tr>
<td>4.1 Cas scalaire</td>
<td>82</td>
</tr>
<tr>
<td>4.1.1 Généralités et définitions</td>
<td>82</td>
</tr>
<tr>
<td>4.1.2 Méthodes du point fixe ou d’approximations successives</td>
<td>83</td>
</tr>
<tr>
<td>4.1.3 Méthode du promoteur de convergence de Wegstein</td>
<td>85</td>
</tr>
<tr>
<td>4.1.4 Méthode de Newton-Raphson</td>
<td>86</td>
</tr>
<tr>
<td>4.1.5 Méthode de Newton modifiée</td>
<td>88</td>
</tr>
<tr>
<td>4.1.6 Méthodes de dichotomie</td>
<td>89</td>
</tr>
<tr>
<td>4.1.7 Regula falsi (fausse position)</td>
<td>89</td>
</tr>
<tr>
<td>4.1.8 Méthode $\Delta^2$ d’Aitken (accélération de la convergence)</td>
<td>90</td>
</tr>
<tr>
<td>4.1.9 Méthode de Steffenson (accélération de la convergence)</td>
<td>91</td>
</tr>
<tr>
<td>4.2 Cas des polynômes: Solutions numériques des équations algébriques</td>
<td>92</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Schéma de Horner</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Méthode de Laguerre</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Méthode de Bairstow</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Méthode de Maehly</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Localisation des racines</td>
</tr>
<tr>
<td>4.2.6</td>
<td>Nombre de racines réelles d’un polynôme</td>
</tr>
<tr>
<td>4.2.7</td>
<td>Cas des racines isolées</td>
</tr>
<tr>
<td>4.3</td>
<td>Applications</td>
</tr>
<tr>
<td>4.4</td>
<td>Complément bibliographique</td>
</tr>
<tr>
<td>4.5</td>
<td>Exercices</td>
</tr>
<tr>
<td>5</td>
<td>Méthodes numériques de résolution des systèmes non linéaires</td>
</tr>
<tr>
<td>5.1</td>
<td>Méthodes itératives</td>
</tr>
<tr>
<td>5.2</td>
<td>Problèmes d’optimisation</td>
</tr>
<tr>
<td>5.3</td>
<td>Applications</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Modèles écologiques</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Modèles à deux espèces</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Modèle proie-prédateur</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Modèle à compétition</td>
</tr>
<tr>
<td>5.3.5</td>
<td>Modèle à coopération</td>
</tr>
<tr>
<td>5.4</td>
<td>Complément bibliographique</td>
</tr>
<tr>
<td>5.5</td>
<td>Exercices</td>
</tr>
<tr>
<td>6</td>
<td>Calcul des valeurs propres et vecteurs propres</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>6.2</td>
<td>Méthodes basées sur le polynôme caractéristique</td>
</tr>
<tr>
<td>6.3</td>
<td>Méthodes itératives</td>
</tr>
<tr>
<td>6.4</td>
<td>Applications</td>
</tr>
<tr>
<td>6.5</td>
<td>Exercices</td>
</tr>
<tr>
<td>7</td>
<td>Analyse numérique des équations différentielles ordinaires (e.d.o)</td>
</tr>
<tr>
<td>7.1</td>
<td>Rappels sur les équations différentielles ordinaires (e.d.o)</td>
</tr>
<tr>
<td>7.2</td>
<td>Systèmes linéaires</td>
</tr>
<tr>
<td>7.3</td>
<td>Notions de stabilité</td>
</tr>
<tr>
<td>7.4</td>
<td>Système d’équations aux différences linéaires avec coefficients constants</td>
</tr>
<tr>
<td>7.5</td>
<td>Méthodes numériques pour les problèmes de condition initiale</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Convergence</td>
</tr>
</tbody>
</table>
Liste des figures

4.1 Convergence($r = 0.9$) .................................................. 105
4.2 2.cycles($r = 2.5$),Chaos($r = 3$) ................................. 105

5.1 Convergence .......................................................... 128
5.2 Stabilité asymptotique .............................................. 128
5.3 Stabilité .............................................................. 129
5.4 2.cycles .............................................................. 129
5.5 4.cycles .............................................................. 129
5.6 Chaos ............................................................... 130
5.7 Extinction d’une espèce ........................................... 131
5.8 Coexistence ........................................................ 131
5.9 Coexistence ........................................................ 132

7.1 Convergence $\lambda = 4.5$,2.cycles $\lambda = 6$ .................. 172
7.2 4.cycles $\lambda = 6.5$,Chaos $\lambda = 7$ ............................ 172
7.3 $\lambda = 3.5,\gamma = 2-\lambda = 5.5,\gamma = 2.5$ ..................... 173
7.4 SIR à deux populations ............................................. 174
7.5 Convergence $\Delta t = 0.01$, Convergence oscillatoire $\Delta t = 0.018$ .... 176
7.6 Oscillation $\Delta t = 0.027$ ........................................... 176
7.7 Convergence-Oscillation ........................................... 177
7.8 Schéma de transmission .......................................... 177
7.9 ................................................................. 180
7.10 Effet de l’effort physique ........................................ 182
Chapitre 1

Rappels et définitions

1.1 Notations

- $A$: matrice carrée d’ordre $n$.
- $a_{ij}$: élément de la $i^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne.
- $A^{-1}$: inverse de $A$.
- $A^\top$: transposée de $A$.
- $\det(A)$: déterminant de $A$.
- $\text{trace}(A)$: trace de $A$.
- $\rho(A)$: rayon spectral de $A$.
- $I$: matrice identité d’ordre $n$.
- $0$: matrice nulle d’ordre $n$.
- $x$: vecteur colonne d’éléments $x_i, i = 1, 2, \ldots, n$.
- $x^\top$: vecteur ligne d’éléments $x_j, j = 1, 2, \ldots, n$.
- $\|A\|$: norme de $A$.
- $\|x\|$: norme de $x$.
- $\langle,\rangle$: produit scalaire.

1.2 Quelques rappels sur les matrices

On note par $A = (a_{ij})$ $i, j = 1, \ldots, n$ une matrice carrée d’ordre $n$ dont les coefficients appartiennent à un corps $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$).

La matrice adjointe de $A$ est notée $A^*$, elle est définie par $\langle Au, v \rangle = \langle u, A^*v \rangle$ pour tous $u, v \in \mathbb{K}^n$, où $\langle x, y \rangle$ désigne le produit scalaire de $x$ et $y$ qu’on note encore

$$\langle x, y \rangle = \sum_i x_i y_i^\top = y^H x.$$

On rappelle alors que
• $A$ est dite symétrique si $A$ est réelle et $A^\top = A$.
• $A$ est dite hermitienne si $A$ est complexe et $A^H = A$.
• $A$ est dite orthogonale si $A$ est réelle et $A^\top A = AA^\top = I$.
• $A$ est dite unitaire si $A$ est complexe et $A^H A = AA^H = I$.
• $A$ est dite normale si $AA^* = A^* A = I$.
• $A$ est dite positive (resp strictement positive) si $a_{ij} \geq 0$ (resp $a_{ij} > 0$) $\forall i, j$.
• $A$ est dite réductible s’il existe une permutation de lignes et de colonnes de telle sorte que $A$ peut s’écrire sous la forme :

$$
\Pi A \Pi^{-1} = \begin{pmatrix}
B & 0 \\
C & D
\end{pmatrix}
$$

où $B$ et $D$ sont des matrices carrées et $\Pi$ est une matrice de permutation. Sinon elle est dite irréductible.
• Soit $A$ une matrice irréductible et positive possédant $l$ valeurs propres de module maximal $|\lambda_1|$ alors
  - Si $l = 1$ $A$ est dite primitive.
  - Si $l > 1$ $A$ est dite imprimitive et $l$ est l’indice d’imprimitivité de $A$.
• Si $A$ est une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$, $A$ est dite M-matrice si $A$ est non singulièr e et $A^{-1} \geq 0$.
• Si $A$ est une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$, $A$ est dite matrice de Stieljes si $A$ est symétrique et définie positive.
• Une matrice $A = (a_{ij})$ est dite matrice de Hessenberg si elle vérifie: $a_{ij} = 0$ pour $i > j + 1$.
• Si $A$ est une matrice complexe, on appelle matrice de comparaison de $A$ la matrice $\tilde{A} = (\alpha_{ij})$ avec $\alpha_{ii} = |a_{ii}|$ et $\alpha_{ij} = -|a_{ij}|$ pour tout $i \neq j$.
• Une matrice complexe $A$ est dite H-matrice si sa matrice de comparaison est une M-matrice.
• $A$ est dite à diagonale strictement dominante en colonnes si elle vérifie la propriété

$$
\sum_{i=1, i\neq j}^n |a_{ij}| < |a_{jj}| \quad \forall j = 1, \ldots, n.
$$
- $A$ est dite à diagonale strictement dominante en lignes si elle vérifie la propriété
  \[ \sum_{j=1}^{n} |a_{ij}| < |a_{ii}| \quad \forall i = 1, \ldots, n. \]

- $A$ est dite à diagonale quasi-dominante s’il existe des entiers positifs $\pi_1, \pi_2, \ldots, \pi_n$ tels que :
  \[ \pi_i |a_{ii}| > \sum_{j \neq i} \pi_j |a_{ij}| \]
  pour tout $i = 1, \ldots, n$.

- $A$ est dite stochastique généralisée si elle vérifie les deux propriétés
  1. $A$ est positive ($a_{ij} \geq 0 \ \forall i, j$).
  2. $\sum_{j=1}^{n} a_{ij} = \sigma > 0$, pour $i = 1, \ldots, n$.
  Si $\sigma = 1$, elle est dite stochastique.

## 1.3 Valeurs et vecteurs propres

**Définition 1.3.1.** Les valeurs propres de $A$ sont les racines de $\det(A - \lambda I) = 0$ ou encore les zéros du polynôme $P_n(\lambda)$ donné par

\[ P_n(\lambda) = \det(A - \lambda I) = (-1)^n \lambda^n + \cdots + \lambda \text{trace}(A) + \det A. \]

On a en particulier

\[ \det A = \prod_i \lambda_i \text{ et } \text{trace}(A) = tr(A) = \sum_i \lambda_i. \]

**Définition 1.3.2.** Un vecteur $x \neq 0$ est dit vecteur propre de $A$ associé à une valeur propre $\lambda$ si $Ax = \lambda x$.

**Définition 1.3.3.** On appelle spectre de $A$ l’ensemble des valeurs propres de $A$; noté $\sigma(A) = \{\lambda_i\}_{i=1}^{n}$. On appelle rayon spectral de $A$, la plus grande valeur propre en module, noté

\[ \rho(A) = \max_i |\lambda_i|. \]

**Définition 1.3.4.** Deux matrices $A$ et $B$ sont dites équivalentes s’il existe deux matrices $P$ et $Q$ telles que $A = PBQ$.

$A$ et $B$ sont dites semblables s’il existe une matrice inversible $P$ telle que $A = PBP^{-1}$. 

\[ 8 \]
Définition 1.3.5. Une matrice $A$ est diagonalisable si elle est semblable à une matrice diagonale.

Définition 1.3.6. Une matrice $A$ est dite définie positive si elle vérifie 
$$\langle Ax, x \rangle > 0$$ pour tout $x \neq 0$.

Remarque 1.3.1. Pour toute matrice $A$, $A^*A$ est une matrice hermitienne définie positive. Il s’ensuit que les valeurs propres de $A^*A$ sont réelles positives. Si ces valeurs propres sont notées $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ les racines carrées positives $\sqrt{\mu_1}, \sqrt{\mu_2}, \cdots, \sqrt{\mu_n}$ sont appelées valeurs singulières de $A$.

Théorème 1.3.1 (Théorème de Schur). Pour toute matrice carrée $A$, il existe une matrice unitaire $U$ telle que 
$$U^H AU = \begin{pmatrix}
\lambda_1 & * & \cdots & * \\
& \lambda_2 & \ddots & \\
& & \ddots & * \\
& & & \lambda_n
\end{pmatrix},$$

où $\lambda_i, i = 1, \cdots, n$, sont les valeurs propres de $A$.

Preuve:
Soit $(e_1, e_2, \cdots, e_n)$ la base canonique de $\mathbb{C}^n$.
On procède par récurrence sur $n$. Pour $n = 1$ le théorème est trivial.
On suppose que le résultat est vrai pour les matrices d’ordre $n - 1$ et soient $A$ une matrice carrée d’ordre $n$ et $\lambda_1$ une valeur propre de $A$ associée à un vecteur propre $x_1 \neq 0$ avec $\|x_1\|_2 = 1$. On peut trouver $n - 1$ vecteurs $x_2, \cdots, x_n$ tels que $(x_1, x_2, \cdots, x_n)$ forment une base de $\mathbb{C}^n$ et la matrice $X = [x_1, \cdots, x_n]$ soit unitaire, $X^HX = I$. Comme

$$X^HAXe_1 = X^HAx_1 = \lambda_1X^Hx_1 = \lambda_1x_1,$$

La matrice $X^HAX$ est de la forme

$$X^HAX = \begin{pmatrix}
\lambda_1 & a \\
0 & A_1
\end{pmatrix},$$

où $A_1$ est une matrice d’ordre $n - 1$ et $a^H \in \mathbb{C}^{n-1}$. Par hypothèse de récurrence, il
existe une matrice unitaire $U_1$ d’ordre $n - 1$ telle que

$$U_1^H A_1 U_1 = \begin{pmatrix}
\lambda_2 & \cdots & * \\
& \ddots & \\
& & \lambda_n
\end{pmatrix}.$$ 

La matrice

$$U = X \begin{pmatrix}
1 & 0 \\
0 & U_1
\end{pmatrix},$$

est une matrice unitaire d’ordre $n$ qui satisfait

$$U^H AU = \begin{pmatrix}
1 & 0 \\
0 & U_1^H
\end{pmatrix} X^H AX \begin{pmatrix}
1 & 0 \\
0 & U_1
\end{pmatrix} = \begin{pmatrix}
\lambda_1 & \cdots & * \\
& \ddots & \\
& & \lambda_n
\end{pmatrix}.$$ 

Ce qui termine la preuve.

Si la matrice $A$ est hermitienne, le théorème de Schur assure que

**Corollaire 1.3.1.**

Pour toute matrice hermitienne $A$, il existe une matrice unitaire $U = [x_1, \cdots, x_n]$ telle que:

$$U^{-1} AU = U^H AU = \begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_n
\end{pmatrix},$$

les valeurs propres de $A$ sont réelles et $A$ est diagonalisable.

**Théorème 1.3.2.** Si les valeurs propres $\lambda_i$, $i = 1, \cdots, n$ d’une matrice hermitienne sont rangées dans un ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n,$$

alors $\lambda_1$ et $\lambda_n$ sont données par

$$\lambda_1 = \max_{x \neq 0} \left( \frac{x^H Ax}{x^H x} \right), \quad \lambda_n = \min_{x \neq 0} \left( \frac{x^H Ax}{x^H x} \right).$$

(1.3.1)
Preuve:
Si \( U^H A U = D = \text{diag}(\lambda_1, \ldots, \lambda_n) \), \( U \) unitaire, alors pour tout \( x \neq 0 \) on a

\[
\frac{x^H A x}{x^H x} = \frac{(x^H U)U^H A U (U^H x)}{(x^H U)(U^H x)} = \frac{y^H D y}{y^H y} = \sum_i \lambda_i |\eta_i|^2 \leq \lambda_1,
\]

où \( y = U^H x = (\eta_1, \ldots, \eta_n)^\top \neq 0 \). Si on prend pour \( x \neq 0 \) le vecteur propre associé à \( \lambda_1 \), \( Ax = \lambda_1 x \), on obtient \( x^H A x / x^H x = \lambda_1 \), donc \( \lambda_1 = \max_{x \neq \theta} \left( \frac{x^H A x}{x^H x} \right) \).

L’autre inégalité s’obtient d’une façon similaire en remplaçant \( A \) par \( -A \).

**Définition 1.3.7.** On appelle coefficient de Rayleigh le réel \( \mu_R \) défini par

\[
\mu_R = \frac{\langle Ax, x \rangle}{\langle x, x \rangle}, \text{ pour tout } x \neq 0.
\]

**Remarque 1.3.2.** Si \( A \) est une matrice hermitienne dont les valeurs propres sont rangées par ordre croissant \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n \), alors d’après le théorème 1.3.2 on a

\[
\mu_1 \leq \mu_R \leq \mu_n.
\]

**Théorème 1.3.3** (Gerschgorin).

Pour toute valeur propre \( \lambda \) de \( A \) on a la relation

\[
|a_{ii} - \lambda| \leq \sum_{j \neq i}^n |a_{ij}|,
\]

pour au moins un indice \( i \).

**Preuve:**
\( \lambda \) étant valeur propre associée à un vecteur propre \( x \), on a

\[
\sum_{j=1, j \neq i}^n a_{ij}x_j - \lambda x_i = -a_{ii}x_i
\]

pour tout \( i = 1, \cdots, n \), il suffit alors de considérer l’indice \( i \) tel que \( |x_i| = \max_j |x_j| \).

### 1.4 Normes vectorielles et normes matricielles

#### 1.4.1 Normes vectorielles

**Définition 1.4.1.** Une norme définie sur un espace vectoriel \( E \) est une application notée \( \| \cdot \| \) de \( E \) dans \( \mathbb{R}_+ \) telle que

i) \( \|x\| = 0 \) si et seulement si \( x = 0 \).
Exemple 1.4.1.
\[ \|x\|_1 = \sum_i |x_i|, \|x\|_2 = \left( \sum_i |x_i|^2 \right)^{1/2}, \|x\|_p = \left( \sum_i |x_i|^p \right)^{1/p}, \|x\|_{\infty} = \max_i |x_i|. \]

1.4.2 Normes matricielles

Définition 1.4.2. Une norme matricielle est une norme qui vérifie, en plus de i), ii) et iii) la condition suivante

iv) \( \|AB\| \leq \|A\| \|B\| \) pour toutes matrices \( A \) et \( B \).

Définition 1.4.3. Étant donné une norme vectorielle \( \| \cdot \| \), l’application
\[ A \rightarrow \|A\| = \sup_{x \neq 0} \left( \frac{\|Ax\|}{\|x\|} \right) \]
définit une norme matricielle dite norme induite par (ou subordonnée à) la norme vectorielle \( \| \cdot \| \).

Les normes vectorielles \( \| \cdot \|_1, \| \cdot \|_{\infty} \) et \( \| \cdot \|_2 \) induisent, respectivement, les normes matricielles suivantes \( \|A\|_1 = \max_j \sum_i |a_{ij}|, \|A\|_{\infty} = \max_i \sum_j |a_{ij}|, \|A\|_2 = \sqrt{\rho(A^\top A)} \).

Remarque 1.4.1. Il existe des normes matricielles qui ne sont subordonnées à aucune norme vectorielle.

Exemple 1.4.2. Norme de Schur \( \|A\| = \left( \sum_i \sum_j a_{ij}^2 \right)^{1/2} \).

1.4.3 Normes compatibles

Définition 1.4.4. Une norme vectorielle \( \| \cdot \| \) et une norme matricielle \( \| \cdot \|_* \) sont compatibles si pour toute matrice \( A \) et tout vecteur \( x \); \( \|Ax\| \leq \|A\|_* \|x\| \).

Remarque 1.4.2. En particulier toute norme matricielle subordonnée à une norme vectorielle est compatible avec celle ci.

Par ailleurs, il existe toujours une norme matricielle compatible avec une norme vectorielle. Réciproquement, étant donnée une norme matricielle \( \| \cdot \|_* \), on peut lui associer une norme \( \| \cdot \| \) vectorielle qui lui soit compatible par \( \|x\| = \|x, 0, \ldots, 0\|_* \).

Théorème 1.4.1. Soit \( \| \cdot \| \) une norme matricielle et \( B \) une matrice complexe vérifiant \( \|B\| < 1 \). Alors la matrice \( I \pm B \) est inversible et on a:

i) \( (I - B)^{-1} = I + B + B^2 + \cdots \)
ii) \[ \|(I \pm B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \]

Inversement, si la série de i) converge alors \(\rho(B) < 1\).

**Preuve:**

Si \(I \pm B\) n’était pas inversible alors elle admettrait \(\lambda = 0\) pour valeur propre et par suite on aurait \(\|B\| \geq 1\). Soit \(G = (I \pm B)^{-1}\) il s’ensuit que \((I \pm B)G = I\) c.à.d \(\|G\| \leq 1 + \|B\|\|G\|\) soit \(\|G\| \leq \frac{1}{1 - \|B\|}\).

En écrivant \(S_p = I + B + B^2 + \cdots + B^p\), on obtient

\[ S_p - BS_p = (I - B)S_p = I - B^{p+1}. \]

Comme \(I - B\) est inversible on a \(S_p = (I - B)^{-1}(I - B^{p+1})\) ou encore

\[ \|(I - B)^{-1} - S^p\| \leq \frac{\|B\|^{p+1}}{1 - \|B\|} \]

d’où \(\lim_{p \to \infty} S_p = (I - B)^{-1}\).

Inversement, si \(\mu\) est une valeur propre de \(B\) associé à un vecteur propre \(v\) alors on a:

\[ (I + B + B^2 + \cdots)v = (1 + \mu + \mu^2 + \cdots)v, \]

la convergence de la série des matrices entraîne la convergence de la série des complexes \(\sum \mu^i\) par conséquent on doit avoir \(|\mu| < 1\) et par suite \(\rho(B) < 1\).

**Remarque 1.4.3.** Sous MATLAB les fonctions matricielles les plus courantes sont: `det(A)`, `eig(A)`, `poly(A)`, `inv(A)`, `rank(A)`, `norm(A)`, `norm(A,1)`, `norm(A,2)`, `norm(A,inf)`, `norm(A,'fro')`, `cond(A,k)`.

**Exemple 1.4.3.**

Considérons la matrice de Hilbert d’ordre 6, \(H = \left(\frac{1}{i + j - 1}\right)_{i,j = 1, \ldots, 6}\)

rang \(H\) = 6.

\[ \det(H) = 5.3673e^{-18}. \]

\[ \text{norm}(H, k) = \left\{ \begin{array}{ll}
\|H\|_1 &= 2.4500 \quad \text{si} \quad k = 1, \\
\|H\|_2 &= 1.6189 \quad \text{si} \quad k = 2, \\
\|H\|_\infty &= 2.4500 \quad \text{si} \quad k = \infty,
\end{array} \right. \]

\[ \text{cond}(H, k) = \left\{ \begin{array}{ll}
\|H\|_1 \|H^{-1}\|_1 &= 2.907e^7 \quad \text{si} \quad k = 1, \\
\|H\|_2 \|H^{-1}\|_2 &= 1.4951e^7 \quad \text{si} \quad k = 2, \\
\|H\|_\infty \|H^{-1}\|_\infty &= 2.907e^7 \quad \text{si} \quad k = \infty,
\end{array} \right. \]

\(\rho(H) = 1.6189\).

Pour les matrices positives on a les résultats suivants
Théorème 1.4.2 (Perron-Frobenius).
Si $A$ est positive, alors $A$ admet une valeur propre $\lambda_0 \geq 0$ égale au rayon spectral de $A$, à $\lambda_0$ correspond un vecteur propre positif.

En ajoutant l’hypothèse d’irréductibilité, on obtient le résultat plus précis.

Théorème 1.4.3 (Perron-Frobenius).
Si $A$ est strictement positive ou positive et irréductible, alors $A$ admet une valeur propre $\lambda_0 > 0$ égale au rayon spectral de $A$, associée à un vecteur propre $\lambda_0$ strictement positif.

De plus $\lambda_0$ est une valeur simple supérieure en module à toute autre valeur propre de $A$, $\lambda_0 > |\lambda_i| \quad \forall \lambda_i$ valeur propre de $A$.

Théorème 1.4.4 (Théorème d’ergodicité).
Soit le système linéaire $n(k+1) =Ln(k)$, où $L$ est une matrice carrée d’ordre $p$ et $n(k) = (n_1(k), n_2(k), \cdots, n_p(k))^\top$.

Si $L$ est strictement positif et primitive alors \( \lim_{k \to +\infty} \left( \frac{L^k}{\lambda_1} \right) = v_1 u_1^\top \) où $v_1$ et $u_1$ sont les vecteurs propres droit et gauche associés à $\lambda_1$ et $u_1^\top v_1 = 1$. Si de plus $\|v_1\| = 1$ alors $\lim_{k \to +\infty} \frac{n(k)}{\|n(k)\|} = v_1$.

Théorème 1.4.5 (Vitesse de convergence).
Si $L$ est strictement positif ou positive et primitive avec valeur propre dominante $\lambda_1$ et $u$ et $v$ sont respectivement les vecteurs propres droit et gauche associés à $\lambda_1$, si on suppose de plus que les valeurs propres distinctes de $L$ sont ordonnées de telle sorte que $\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots > |\lambda_q|$ si de plus $|\lambda_2| = |\lambda_3|$ et l’ordre de multiplicité $m$ de $\lambda_2$ est au moins plus grand que celui de $\lambda_3$ ou de toute autre valeur propre de même module que $\lambda_2$ alors

\[
L^t = \lambda_1uv^\top + O\left(t^{m-1}|\lambda_2|^t\right).
\]

Si $m = 1$

\[
\left\| \frac{n(t)}{\lambda_1} - uv^\top \right\| \leq c \frac{|\lambda_2|^t}{\lambda_1}
\]

$\frac{\lambda_1}{|\lambda_2|}$ est dit rapport d’amortissement.

Remarque 1.4.4. Toute matrice stochastique admet 1 comme valeur propre associée au vecteur propre $I$ dont toutes les composantes sont égales à 1.

Théorème 1.4.6. Soit $A$ une matrice associée à une chaîne de Markov. Alors il existe un vecteur $p > 0$ avec $\|p\| = 1$, $\lim_{k \to +\infty} A^k = \overline{A}$ et $\lim_{k \to +\infty} \frac{n(k)}{\|n(k)\|} = p$, si $n(k+1) = An(k)$.
1.5 Applications

Application 1.5.1 (Modèle de Markov). [41]

Considérons une stratégie de gestion des terres en supposant quatre types: T1=Terres agricoles non irriguées, T2=Terres agricoles irriguées, T3=Terres de construction, T4=Terres non exploitées. Supposons que nous adoptions une stratégie à long terme qui consiste à contrôler chaque année le passage d’un type de terre à un autre pour arriver à un certain équilibre. Soit $M$ une matrice de transition d’un type de terre à un autre, si $n(t)$ désigne l’état du système à l’instant $t$ alors $n(t + 1) = Mn(t)$ où $M = (p_{ij})_{i,j=1,\ldots,4}$, $p_{ij}$ est la probabilité de passage du type de terre $i$ au type $j$ au bout d’une année.

A titre d’exemple, si cette stratégie est appliquée durant une quarantaine d’années avec

$$M = \begin{pmatrix} 0.75 & 0.1 & 0.15 & 0 \\ 0.1 & 0.75 & 0.15 & 0 \\ 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0.05 & 0.25 & 0.7 \end{pmatrix},$$

on arrive à l’équilibre suivant:

$$M^{37} = \begin{pmatrix} 0.1468 & 0.3714 & 0.36 & 0.12 \\ 0.1468 & 0.3714 & 0.36 & 0.12 \\ 0.1468 & 0.3714 & 0.36 & 0.12 \\ 0.1468 & 0.3714 & 0.36 & 0.12 \end{pmatrix}$$

Interprétation: à l’équilibre le domaine des terres non irriguées sera formé de 59% de son étendue initiale, les domaines des terres irriguées et de construction auront augmenté de presque 50% chacun (1.49 et 1.44). Enfin, les terres non exploitées auront diminué de la moitié (0.52%). Ces chiffres ayant été obtenus en prenant pour chaque colonne, les sommes sur les quatre lignes.

Application 1.5.2. (Matrice de Leslie)

On suppose qu’une population de femmes est formée de $m$ classe d’âge et on note $n_i(t)$ la taille de la population d’âge $i$ à l’instant $t$. Si $b_i$ est le nombre moyen de femelles nées par femme d’âge $i$ et $p_i$ la proportion de femmes d’âge $i$ qui survivent à l’âge $i + 1$ alors, à l’instant $t + 1$ on a:

- La première classe d’âge est formée par les nouveaux nés, sa taille est donnée par:

$$n_1(t + 1) = \sum_{i=1}^{m} b_i n_i(t).$$
• Chaque groupe d’âge $i + 1$ est formé par les individus du groupe d’âge $i$ qui survivent

$$n_{i+1}(t + 1) = p_in_i(t) ; \ i \geq 1.$$ 

En notant $n(t)$ le vecteur de composantes $n_i(t); \ i = 1, \cdots , n$ on obtient une formulation matricielle $n(t + 1) = Ln(t)$ où $L$ est la matrice de Leslie donnée par

$$L = \begin{pmatrix}
    b_1 & b_2 & \cdots & b_n \\
    p_1 & 0 & \cdots & 0 \\
    \vdots & \ddots & \ddots & \vdots \\
    0 & \cdots & p_{n-1} & 0
  \end{pmatrix}$$

$L$ est une matrice irréductible si $b_n \neq 0$.

**Exemple 1.5.1. ([41])**

On suppose que les classes d’âge sont d’intervalle de 15 ans, et que seules les femelles des classes 2 et 3 reproduisent et que l’âge maximal est 75 ans.

$$n_1(t + 1) = b_2n_2(t) + b_3n_3(t)$$

$$n_{i+1}(t + 1) = p_in_i(t)$$

l’unité de temps est égal à 15 ans, d’où la forme matricielle suivante

$$n(t + 15) = Ln(t),$$

où

$$L = \begin{pmatrix}
    0 & b_2 & b_3 & 0 & 0 \\
    p_1 & 0 & 0 & 0 & 0 \\
    0 & p_2 & 0 & 0 & 0 \\
    0 & 0 & p_3 & 0 & 0 \\
    0 & 0 & 0 & p_4 & 0
  \end{pmatrix} \quad \text{et} \quad n(0) = \begin{pmatrix}
    1000 \\
    900 \\
    800 \\
    700 \\
    600
  \end{pmatrix}$$

avec $b_1 = 1.5 , b_3 = 1, p_1 = 0.98, p_2 = 0.96, p_3 = 0.93, p_4 = 0.9$, comme

$$\lim_{k \to +\infty} \frac{n_i(k+1)}{n_i(k)} = \lambda_1,$$ 

donc d’après le tableau 1.1 on a $\lambda_1 = 1.4548$. D’autre part

$$\lim_{k \to +\infty} \frac{n(k+1)}{||n(k)||} = v_1,$$ 

donc d’après le tableau 1.2 on a

$v_1 = (0.387, 0.261, 0.172, 0.1102, 0.068)^\top.$
Tableau 1.1: Taux de croissance de chaque classe d’âge

<table>
<thead>
<tr>
<th>période</th>
<th>classe 1</th>
<th>classe 2</th>
<th>classe 3</th>
<th>classe 4</th>
<th>classe 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0889 %</td>
<td>1.0629 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
</tr>
<tr>
<td>2</td>
<td>1.0856 %</td>
<td>1.0629 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
</tr>
<tr>
<td>3</td>
<td>2.15</td>
<td>1.0856 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
</tr>
<tr>
<td>4</td>
<td>1.3297</td>
<td>2.15</td>
<td>1.0856 %</td>
<td>1.0889 %</td>
<td>1.0889 %</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>16</td>
<td>1.4550</td>
<td>1.4554</td>
<td>1.4548</td>
<td>1.4548</td>
<td>1.4557</td>
</tr>
<tr>
<td>17</td>
<td>1.4548</td>
<td>1.4548</td>
<td>1.4554</td>
<td>1.4554</td>
<td>1.4547</td>
</tr>
<tr>
<td>18</td>
<td>1.4549</td>
<td>1.4549</td>
<td>1.4548</td>
<td>1.4548</td>
<td>1.4548</td>
</tr>
<tr>
<td>19</td>
<td>1.4549</td>
<td>1.4549</td>
<td>1.4548</td>
<td>1.4548</td>
<td>1.4548</td>
</tr>
<tr>
<td>20</td>
<td>1.4549</td>
<td>1.4549</td>
<td>1.4548</td>
<td>1.4548</td>
<td>1.4548</td>
</tr>
</tbody>
</table>

Tableau 1.2: Pourcentages de chaque classe d’âge

<table>
<thead>
<tr>
<th>période</th>
<th>classe 1</th>
<th>classe 2</th>
<th>classe 3</th>
<th>classe 4</th>
<th>classe 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40.0522 %</td>
<td>18.2563 %</td>
<td>16.0954 %</td>
<td>13.8599 %</td>
<td>11.7362 %</td>
</tr>
<tr>
<td>2</td>
<td>34.0485 %</td>
<td>30.737 %</td>
<td>13.7244 %</td>
<td>11.7218 %</td>
<td>9.7632 %</td>
</tr>
<tr>
<td>3</td>
<td>40.9743 %</td>
<td>22.8516 %</td>
<td>20.2081 %</td>
<td>8.7412 %</td>
<td>7.2249 %</td>
</tr>
<tr>
<td>4</td>
<td>38.383 %</td>
<td>28.0335 %</td>
<td>15.3154 %</td>
<td>13.1205 %</td>
<td>5.4923 %</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>16</td>
<td>38.7892 %</td>
<td>26.1384 %</td>
<td>17.2402 %</td>
<td>11.0212 %</td>
<td>6.8188 %</td>
</tr>
<tr>
<td>17</td>
<td>38.7910 %</td>
<td>26.1302 %</td>
<td>17.2421 %</td>
<td>11.0208 %</td>
<td>6.8178 %</td>
</tr>
<tr>
<td>18</td>
<td>38.7925 %</td>
<td>26.1248 %</td>
<td>17.2409 %</td>
<td>11.0217 %</td>
<td>6.8176 %</td>
</tr>
<tr>
<td>19</td>
<td>38.7930 %</td>
<td>26.1296 %</td>
<td>17.2416 %</td>
<td>11.0214 %</td>
<td>6.8181 %</td>
</tr>
<tr>
<td>20</td>
<td>38.7930 %</td>
<td>26.1292 %</td>
<td>17.2413 %</td>
<td>11.0214 %</td>
<td>6.8177 %</td>
</tr>
</tbody>
</table>

Exemple 1.5.2. ([31])

Le diabète est une maladie chronique qui touche toutes les catégories de la population à travers le monde.

En des termes simples, cette maladie est causée par un désordre par lequel le corps devient incapable de contrôler la quantité de sucre dans le sang. Le dysfonctionnement du pancréas (plus exactement des cellules bêta) conduit à un manque partiel ou total de l’insuline qui est la clé du mécanisme qui convertit le sucre en énergie. Par conséquent, le système n’est plus auto-régulé et le niveau de sucre dans le sang dépasse les limites seuil.

Théoriquement, un diabétique peut mener une vie tout à fait normale (c. à. d comme un non diabétique) s’il arrive à contrôler quotidiennement sa glycémie en la maintenant aussi proche que possible de la concentration normale qui est environ 1g/ml. Cependant, ce contrôle nécessite la maîtrise de trois paramètres qui sont: l’alimentation, la quantité d’insuline et l’effort physique. La vie du diabétique
dépend de l’équilibre entre les trois paramètres pré-cités et le problème réside à juste titre dans le déséquilibre qui conduit à long terme aux complications diabétiques (cérité, amputations, insuffisance rénale, etc ...). Les études récentes ont montré que l’éducation diabétique et la bonne maîtrise des paramètres permet une réduction significative des complications. Par le modèle donné dans le présent exemple, les auteurs entendent illustrer l’importance du contrôle du passage entre la classe des diabétiques sans complications, $D_k(t)$ et celle des diabétiques avec complications, $C_k(t)$ en tenant en compte de la structure d’âge. Pour les statistiques concernant l’incidence, la prévalence, le coût du diabète et d’autres données, nous renvoyons aux références([22, 30, 31, 145]).

On suppose que les événements tels que naissance, mortalité et guérison, sont enregistrés tous les 15 ans. On néglige l’émigration et l’immigration et on suppose que le nombre de males est égal au nombre de femelles. On suppose qu’il n’y a pas de complication congénitale. On obtient alors le modèle compartimental suivant:

Les hypothèses précédentes conduisent aux équations:

\[
\begin{align*}
D_0(t_{i+1}) &= I_0(t_{i+1}) + \sum_{k=0}^{m} \frac{\theta(t)}{2} \left( 1 - \frac{1}{\gamma_k(t)} \right) b_k(t_i)n_k(t_i) + \frac{\theta'(t_i)}{2} \left( b'_k(t_i)D_k(t_i) + b''_k(t_i)C_k(t_i) \right) \\
C_0(t_{i+1}) &= 0 \\
D_{k+1}(t_{i+1}) &= I_{k+1}(t_{i+1}) + p_k(t)D_k(t_i)(1 - s_k(t_{i+1})) + a_k(t_{i+1})q_k(t_i)C_k(t_i), \\
C_{k+1}(t_{i+1}) &= I'_{k+1}(t_{i+1}) + q_k(t_i)C_k(t_i)(1 - a_k(t_{i+1})) + p_k(t_i)s_k(t_{i+1})D_k(t_i),
\end{align*}
\]

pour $k = 0, 1, \cdots, m - 1$

avec

\[
I_{k+1}(t_{i+1}) = \frac{\epsilon_k(t_i)}{\gamma_k(t_i)} D_k(t_i) = \alpha_{k+1}(t_i)D_k(t_i),
\]

\[
I'_{k+1}(t_{i+1}) = \frac{\epsilon'_k(t_i)}{\gamma'_k(t_i)} C_k(t_i) = \alpha'_{k+1}(t_i)C_k(t_i) \quad \text{et} \quad I_0(t_i) = \alpha_0(t_i)n_0(t_i).
\]

On suppose que $b_k(t_i) = 0$, $b'_k(t_i) = 0$ et $b''_k(t_i) = 0$ pour $k < A_0$ et $k > A_1$. 

18
En Additionnant $C_{k+1}(t_{i+1})$ et $D_{k+1}(t_{i+1})$ On obtient le modèle simplifié avec $n_{k+1}$:

\[
\begin{cases}
n_0(t_{i+1}) = l_0(t_i) + \sum_{k=0}^{m} \frac{\theta(t_i)}{2} \left( \frac{1}{\gamma_k(t_i)} - 1 \right) b_k(t_i) n_k(t_i) + \frac{\theta'(t_i)}{2} \left( b''_k(t_i) D_k(t_i) + b''_k(t_i) C_k(t_i) \right) \\
r_0(t_{i+1}) = 0 \\
n_{k+1}(t_{i+1}) = \left( p_k(t_i) + (q_k(t_i) - p_k(t_i)) r_k(t_i) \right) n_k(t_i) + \alpha_{k+1}(t_i) D_k(t_i) + \alpha'_{k+1}(t_i) C_k(t_i)
\end{cases}
\]

En supposant que les paramètres sont indépendants du temps, le modèle s’écrit:

\[
n(t_{i+1}) = Ln(t_i)
\]

où

\[
L = \begin{pmatrix}
f_0 & f_1 & \cdots & f_m \\
\nu_0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \nu_{m-1} & 0
\end{pmatrix}
\]

avec

\[
f_0 = \alpha_0 + \frac{\theta}{2} \left( \frac{1}{\gamma_0} - 1 \right) b_0 + \frac{\theta'}{2} (b'_0 + (b''_0 - b'_0) r'_0), f_k = \frac{\theta}{2} \left( \frac{1}{\gamma_k} - 1 \right) b_k + \frac{\theta'}{2} (b'_k + (b''_k - b'_k) r'_k).
\]

\[
\nu_k = \alpha_{k+1} + q_k + (\alpha'_{k+1} + q_k - \alpha_{k+1} - p_k) r'_k, n(t_i) = \begin{pmatrix} n_0(t_i), \cdots, n_m(t_i) \end{pmatrix}^T.
\]

**Exemple 1.5.3. ([31])**

On suppose maintenant que les événements tels que naissance, mortalité et guérison, sont enregistrés au début de chaque année. On néglige l’émigration et l’immigration et on suppose que le nombre de mâles est égal au nombre de femelles. On suppose qu’il n’y a pas de complication congénitale. On obtient alors le modèle compartimental suivant:
On obtient

\[ n(t + 1) = L n(t) \]

où

\[
L = \begin{pmatrix}
  f_0 & f_1 & f_2 & \cdots & f_m \\
  \nu_0 & \nu_1' & 0 & \cdots & 0 \\
  \vdots & \ddots & \ddots & \ddots & \vdots \\
  0 & 0 & \cdots & 0 & \nu_{m-1}' & \nu_m'
\end{pmatrix}
\]

avec

\[ f_0 = \mu_0 + \alpha_0 + \frac{\theta}{2} \left( \frac{1}{\gamma_0} - 1 \right) b_0 + \frac{\theta'}{2} (b_0' + (b_0'' - b_0') r_0) . \]

\[ f_k = \frac{\theta}{2} \left( \frac{1}{\gamma_k} - 1 \right) b_k + \frac{\theta'}{2} (b_k' + (b_k'' - b_k') r_k) n_k(t) \text{ pour } k = 1, 2, \ldots m - 1, \]

\[ \nu_k = p_k \omega_k + \alpha_{k+1} + (\alpha_{k+1} + q_k \omega_k' - \alpha_{k+1} - p_k \omega_k) r_k \text{ pour } k = 0, 1, \ldots, m - 1, \]

\[ \nu_k' = p_k \mu_k + (q_k \mu_k' - p_k \mu_k) r_{k+1} \text{ pour } k = 1, \ldots, m. \]
1.6 Complément bibliographique sur les matrices positives

Les matrices positives n’apparaissent pas toujours dans les livres d’analyse numérique. Cependant, leur théorie et leur utilisation en écologie, statistique et numérique de phénomènes réels revêt une grande importance comme en atteste le nombre de publications et de livres dévoués à ce sujet. Il est difficile de donner une liste exhaustive mais on peut retracer brièvement le “chemin” des matrices positives à travers les auteurs suivants:


Signalons enfin qu’un certain nombre de résultats ont été établis sur le lien entre matrices positives et théorie des graphes.
1.7 Exercices

Exercice 1.7.1. Soit $A$ une matrice carrée réelle, montrer que

i) $\|A\|_1 = \max_i \sum |a_{ij}|.$

ii) $\|A\|_\infty = \max_j \sum |a_{ij}|.$

iii) $\|A\|_2^2 = \rho(A^\top A).$

Exercice 1.7.2. Montrer que toute matrice de Leslie $L$ admet une seule valeur propre positive $\lambda_1$ associée au vecteur propre $v_1$ donné par

$$v_1 = \left(1, \frac{p_1}{\lambda_1}, \cdots, \frac{p_1 p_2 \cdots p_{m-1}}{\lambda_1^{m-1}} \right)^\top.$$

Exercice 1.7.3. Soit $A = (a_{ij})$ une matrice carrée.

1. Montrer que $\rho(A) \leq \|A\|$ pour toute norme subordonnée, $\rho(A)$ étant le rayon spectral de $A$.

2. Si $B$ est une matrice non singulière, vérifier que l’application:
   
   $x \longrightarrow \|x\|_\ast = \|Bx\|_\infty$ définit une norme vectorielle.

3. Soient $D$ et $P$ deux matrices non singulières avec $P$ telle que $PAP^{-1} = U$ (triangulaire supérieure), en posant $B = D^{-1}P$, montrer que $\|A\|_\ast \leq \|D^{-1}UD\|_\infty$ et en déduire que, étant donné une matrice carrée $A$ et un réel $\epsilon > 0$, il existe une matrice diagonale $D = \text{diag}(d)$ avec $d > 0$ et telle que: $\|A\|_\ast \leq \rho(A) + \epsilon$

4. Conclure.

5. On suppose que $A$ est triangulaire inférieure avec $a_{ii} = 1 \forall i$ et $a_{ij} = -1 \forall i > j$.
   
   Calculer la $j$ème colonne de $A^{-1}$ et montrer que la matrice $A$ est telle que
   
   $$\|A\|_1 \left\|A^{-1}\right\|_1 = n2^{n-1}.$$

6. On suppose maintenant que la matrice $A$ est telle que

   $$|a_{ii}| - \sum_{j=1}^n |a_{ij}| \geq \delta > 0 \forall i = 1, \cdots, n$$

   montrer que $\|Ax\|_\infty \geq \delta \|x\|_\infty$, en déduire que $A^{-1}$ existe et que $\|A^{-1}\|_\infty \leq \frac{1}{\delta}$.
7. On considère la matrice $A$ donnée par

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

avec $\varepsilon > 0$.

Prouver que la plus grande valeur propre $\lambda_4$ de $A$ vérifie

$$4 + \frac{\varepsilon}{2} \leq \lambda_4 \leq 4 + \varepsilon.$$
Chapitre 2

Méthodes directes de résolution des systèmes linéaires $Ax = b$

Dans ce chapitre, on s’intéresse à la résolution numérique d’un système linéaire

$$Ax = b \quad (2.0.1)$$

où $A$ est une matrice carrée supposée inversible, $b$ un vecteur second membre et $x$ le vecteur des inconnues, $A = (a_{ij})_{i,j=1,...,n}, \ b = (b_1, \cdots, b_n) \top, \ x = (x_1, \cdots, x_n) \top$. Théoriquement, le fait que $A$ soit inversible entraîne que le système (2.0.1) admet une solution unique $x = A^{-1}b$.

Mais cette écriture suppose que l’on dispose de la matrice $A^{-1}$, or l’obtention de $A^{-1}$ est équivalente à la résolution de $n$ systèmes linéaires, $A. (A^{-1})_j = e_j = (0, \cdots, 1, 0, \cdots, 0) \top$ en plus de la multiplication $x = A^{-1}b$.

Une autre méthode consisterait à obtenir les $x_i$ à l’aide des formules de Cramer $x_i = \frac{\det(A_i)}{\det(A)}$ où $\det(A_i)$ désigne le déterminant de la matrice obtenue en remplaçant la $i^{ème}$ colonne de $A$ par le vecteur $b$.

Le calcul de chaque déterminant nécessite $n.n!$ multiplications et $(n! - 1)$ additions.

Soit au total : $(n + 1)!n$ multiplications, $(n + 1)(n! - 1)$ additions et $n$ divisions.

A titre d’exemple, on a besoin de 4319 opérations si $n = 5$ et environ 41000 opérations pour $n = 10$. Comme les problèmes d’analyse numérique donnent lieu à des matrices de grandes tailles ($n$ assez grand), la méthode de Cramer et les méthodes similaires s’avèrent inutilisables.
2.1 Résolution d’un système par les méthodes de descente ou de remontée

C’est le cas où on a à résoudre un système de la forme

\[ Ux = b \] (2.1.1)

ou

\[ Lx = b \] (2.1.2)

avec \( U \) triangulaire supérieure et \( L \) triangulaire inférieure.

Si on prend l’équation (2.1.1) par exemple on obtient

\[ Ux = b \iff \begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n = b_1 \\ u_{22}x_2 + \cdots + u_{2n}x_n = b_2 \\ \vdots \\ u_{nn}x_n = b_n \end{cases} \]

En supposant que les \( u_{kk} \) sont non nuls, on obtient les \( x_i \) de façon évidente en commençant par le bas et en remontant. On obtient ainsi \( x_n = b_n/u_{nn} \) puis

\[ x_i = \left( b_i - \sum_{j=i+1}^n u_{ij}x_j \right) / u_{ii}, \text{ pour } i = n - 1 \text{ à } 1. \]

L’algorithme de résolution est le suivant :

\[ x_n = b_n/u_{nn} \]

Pour \( i = n - 1 \) à 1

\[ x_i = b_i \]

Pour \( j = i + 1 \) à \( n \)

\[ x_i = x_i - u_{ij} \times x_j \]

fin \( j \)

\[ x_i = x_i / u_{ii} \]

fin \( i \)

Le nombre d’opérations nécessaire est: \( \frac{n(n-1)}{2} \) multiplications, \( \frac{n(n-1)}{2} \) additions et \( n \) divisions. Soit au total \( n^2 \) opérations.

**Remarque 2.1.1.** Le cas d’un système avec matrice triangulaire inférieure se traite de façon similaire en obtenant \( x_1 \) d’abord puis en descendant.
2.2 Matrices élémentaires

2.2.1 Matrices élémentaires de Gauss

Soient les matrices

\[ M_1 = \begin{pmatrix} 1 & -m_{21} & 1 & 0 \\ \vdots & 0 & \ddots & \vdots \\ -m_{n1} & 0 & \cdots & 1 \end{pmatrix}, \quad M_k = \begin{pmatrix} 1 & & & \cdots & 0 \\ 0 & \ddots & \cdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & -m_{k+11} & \cdots \\ 0 & \cdots & 0 & -m_{kn} & 1 \end{pmatrix} \]

en posant \( e_k = (0, \cdots, 1, 0, \cdots, 0)^\top \) et \( m_k = (0, \cdots, 0, -m_{k+11}, \cdots, -m_{nk})^\top \), on obtient \( M_k = I - m_k e_k^\top \) et on vérifie facilement que \( M_k \) est inversible et que \( M_k^{-1} = I + m_k e_k^\top \).

2.2.2 Matrices élémentaires de Danilevski

Elles sont de la forme:

\[ M_k = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ m_{k1} & m_{k2} & \cdots & \cdots & m_{kn} \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \]

et permettent de transformer une matrice \( A \) en une matrice de Frobenius \( P \) de la forme

\[ P = \begin{pmatrix} p_1 & p_2 & \cdots & \cdots & p_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}. \]

Ces matrices seront utilisées dans le chapitre 6.

2.2.3 Matrices élémentaires de Householder

Ces matrices sont de la forme

\[ P_k = I_n - 2 \omega_k \omega_k^\top \text{ avec } \omega_k = (0, \cdots, 0, \omega_{k+11}, \cdots, \omega_{nk})^\top \text{ et } \omega_k^\top \omega_k = 1. \]

On vérifie aussi qu'on a \( P_k = P_k^{-1} = P_k^\top \), \( P_k \) est donc une matrice orthogonale symétrique. Elle peut aussi s'écrire sous la forme explicite et en blocs:
\[ P_k = \begin{pmatrix} I_k & \tilde{P}_k \\ \end{pmatrix} \]

avec \( \tilde{P}_k = I_{n-k} - 2\hat{\omega}_k \hat{\omega}_k^\top \) et \( \hat{\omega}_k = (\omega_{k+1}, \ldots, \omega_{nk})^\top \).

### 2.2.4 Matrices élémentaires de permutation

Elles sont de la forme

\[
I_{k,l} = \begin{pmatrix} 1 & 0 & 1 & \cdots & 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 & 1 & \cdots & 0 \\ & & \ddots & \ddots & & & \ddots & \\ & & & \cdots & 0 & 1 & \cdots & 0 \\ & & & & 1 & \cdots & \cdots & \cdots \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & \cdots & \cdots & \cdots \\ & & & & & & & 1 \\ \end{pmatrix} 
\]

**Remarque 2.2.1.** \( I_{k,l}A \) échange les lignes \( k \) et \( l \) de \( A \) alors que \( AI_{k,l} \) échange les colonnes \( k \) et \( l \) de \( A \). On a encore \( I_{k,l} = I_{k,l}^{-1} = I_{l,k}^\top \).

**Exemple 2.2.1.** Soit \( A \) la matrice donnée par:

\[
A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \quad \text{et} \quad I_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}
\]

alors \( I_{13}A = \begin{pmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 3 \end{pmatrix} \) et \( AI_{13} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{pmatrix} \).

**Définition 2.2.1.** Une matrice de permutation est un produit de matrices élémentaires de permutation.

### 2.2.5 Matrices élémentaires de Perlis

Soient les matrices suivantes:

- \( I_{ij} \), une matrice \( I_n \) dont on a permué les \( i^{\text{eme}} \) et \( j^{\text{eme}} \) lignes;
- \( I_i(d) \), une matrice \( I_n \) dont la \( i^{\text{eme}} \) ligne a été multipliée par un scalaire \( d \);
- \( I_{il}(d) \), une matrice \( I_n \) dont la \( i^{\text{eme}} \) ligne est multipliée par: \( e_{ij} + de_{ij} \), où \( I_{ij} = (e_{ij})_{i,j=1,\ldots,n} \).
Les matrices $I_{ij}$, $I_i(d)$, $I_i(d)$ s’appellent matrices élémentaires de Perlis.

**Exemple 2.2.2.** $I_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, $I_2(d) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $I_{12}(d) = \begin{pmatrix} 1 & d & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Ces matrices sont régulières et leurs inverses s’écrivent:

$$I_{ij}^{-1} = I_{ij}$$
$$I_i^{-1}(d) = I_i(1/d)$$
$$I_i^{-1}(d) = I_i(-d)$$

Les transformations élémentaires sur une matrice $A$ peuvent se ramener à la premultiplication de $A$ par l’une des matrices élémentaires précédentes. Ainsi avec $A' = I_{ik}A$, $A'$ est la matrice obtenue après permutation des lignes $i$ et $k$.

$A' = I_{il}(d)A$, $A'$ est la matrice dont la $i^{ème}$ ligne est remplacée par la somme de la $i^{ème}$ ligne de $A$ et $d$ fois la $l^{ème}$ ligne de $A$.

**Remarque 2.2.2.** On peut aussi définir des matrices élémentaires pour les opérations sur les colonnes, on les note $P_{ij}$, $P_j(d)$, $P_{ji}(d)$.

**Exemple 2.2.3.** $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.

Ici la matrice élémentaire $M_1$ de Gauss est donnée par

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ -7 & 0 & 1 \end{pmatrix}.$$ 

Par ailleurs

$$A' = I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 7 & 8 & 9 \end{pmatrix}$$

et

$$A'' = I_{31}(-7)I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{pmatrix}$$

et on a

$$A^{(2)} = M_1A = I_{31}(-7)I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{pmatrix}.$$
2.2.6 Matrices élémentaires de Givens (ou de rotation)

Elles sont données par

\[
R_{k,l} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & k \\
0 & 1 & \cdots & \cdots & l \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\cdots & \cdots & \cdots & 1 & \cdots \\
k & l & \cdots & \cdots & 1 \\
\end{pmatrix}
\]

avec \(c = \cos \theta, \ s = \sin \theta\) et \(k \neq l\)

Remarque 2.2.3. On vérifie facilement que \(R_{k,l}^\top = R_{k,l}^{-1}\).

2.3 Méthodes de Gauss

2.3.1 Méthode de Gauss sans pivot

Elle consiste à transformer le système \(Ax = b\) (1) en un système \(Ux = c\), (2) avec \(U\) triangulaire supérieure puis à résoudre le nouveau système par la méthode de remontée.

Soit \((S_1)\) le système de départ:

\[
(S_1) \begin{cases}
a_{i1}^{(1)}x_1 + a_{i2}^{(1)}x_2 + \cdots + a_{in}^{(1)}x_n = b_1^{(1)} \\
\vdots \\
a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)}
\end{cases}
\]

On pose \(m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}\), en supposant que \(a_{11}^{(1)} \neq 0\), \(i = 2, \cdots, n\).

Ensuite, on remplace la ligne \(L_i\) par \(L_i' = L_i - m_{i1} \times L_1\), ce qui donne :

\[
a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} \times a_{1j}^{(1)} \quad i = 2, \cdots, n \text{ et } j = 2, \cdots, n \\
b_i^{(2)} = b_i^{(1)} - m_{i1} \times b_1^{(1)} \quad i = 2, \cdots, n
\]

On obtient alors le système \((S_2)\) suivant :

\[
(S_2) \begin{cases}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\
0 + a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\
\vdots \\
0 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(2)}x_n = b_n^{(2)}
\end{cases}
\]
On suppose que $a_{22}^{(2)} \neq 0$ et on recommence avec $m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, $i = 3, \ldots, n$.

A l’étape $k$, le système se présente sous la forme :

$$
\begin{cases}
a_{11}^{(1)} x_1 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\
0 \ddots \\
\vdots \\
0 \cdots a_{kk}^{(k)} x_k + \cdots + a_{kn}^{(k)} x_n = b_k^{(k)} \\
\vdots \\
0 \cdots a_{ik}^{(k)} x_k + \cdots + a_{nn}^{(k)} x_n = b_n^{(k)}
\end{cases}
$$

On aboutit alors au système final $(S_n)$ de la forme :

$$
(S_n) \iff UX = c \iff \begin{cases}
a_{11}^{(1)} x_1 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\
0 \ddots \\
\vdots \\
0 \cdots 0 + a_{kk}^{(n)} x_k + \cdots + a_{nn}^{(n)} x_n = b_n^{(n)}
\end{cases}
$$

où $c = \left( b_1^{(1)}, b_2^{(2)}, \ldots, b_n^{(n)} \right) ^\top$.

**Remarque 2.3.1.** Matriciellement, la première étape est équivalente au produit matriciel $A^{(2)} = M_1 A^{(1)}$ où $M_1$ est la matrice élémentaire de Gauss. L’étape finale est alors donnée par : $A^{(n)} = U = M_{n-1} M_{n-2} \cdots M_2 M_1 A^{(1)}$. Evidemment, l’étape finale n’est accessible par ce procédé que si tous les $a_{kk}^{(k)}$ sont non nuls. Si à une étape donnée $a_{kk}^{(k)}$ est nul, et il y a au moins un $a_{ik}^{(k)}$ non nul, avec $i > k$ on permutte les lignes $L_i$ et $L_k$ et on poursuit le procédé. Sinon la matrice $A$ n’est pas inversible et le procédé ne peut continuer.

### 2.3.2 Méthode de Gauss avec pivot partiel

**Exemple 2.3.1.** Soit à résoudre le système

$$
\begin{cases}
10^{-50} x_1 + x_2 = 0 \\
x_1 - x_2 = 0
\end{cases}
$$

La solution exacte est $x_1 = x_2 = 1/(1 + 10^{-50}) \simeq 1$. Cependant, la résolution du système par la méthode de Gauss donne des résultats différents selon qu’on
l’applique avec ou sans pivot.

i) Si on applique la méthode de Gauss sans pivot on obtient

\[ m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{1}{10^{-50}} = 10^{50} \]

et

\[
(S_2) \left\{ \begin{array}{l}
10^{-50}x_1 + x_2 = 0 \\
(-1 - 10^{50})x_2 = 10^{-50}
\end{array} \right.
\]

qui donne pour solution approchée \( x_1 \simeq 1 \) et \( x_2 \simeq 0 \).

ii) Si on adopte la stratégie du pivot partiel qui consiste à mettre en première ligne celle dont le coefficient de \( x_1 \) est le plus grand en module alors on per-mute les lignes pour obtenir le système

\[
(S_1) \left\{ \begin{array}{l}
x_1 - x_2 = 0 \\
10^{-50}x_1 + x_2 = 0
\end{array} \right.
\]

Pour lequel \( m_{21} = \frac{10^{-50}}{1} = 10^{-50} \) et qui conduit à la solution approchée: \( x_2 \simeq 1 \) et \( x_1 = x_2 \).

A travers cet exemple simple, on voit donc le problème que peut poser un pivot trop petit. Pour éviter de diviser par des pivots trop petits pouvant conduire à des solutions absurdes, on peut adopter automatiquement la stratégie du pivot partiel de la manière suivante :

A chaque étape \( k \) : choisir \( a_{kk}^{(k)} \) tel que : \( a_{kk}^{(k)} = \max_{i \geq k} \left| a_{ik}^{(k)} \right| \).

Matriciellement, cette opération revient à multiplier la matrice \( A^{(k)} \) par une matrice de permutation \( I_{kl} \) avant d’appliquer l’élimination de Gauss. La méthode de Gauss avec pivot partiel s’écrit donc :

\[
A^{(2)} = M_1 I_{11} A^{(1)}, \ldots, A^{(n)} = M_{n-1} I_{n-11} \cdots M_1 I_{1i} A^{(1)} = U
\]

où les \( M_i \) sont des matrices élémentaires de Gauss et les \( I_{ki} \) des matrices de permutation pour \( i \geq k \). Si à une étape \( k \) on n’a pas besoin de pivoter, l’écriture reste valable avec \( I_{ki} = I \) où \( I \) désigne la matrice identité.

**Théorème 2.3.1.** Soit \( A \) une matrice carrée, inversible ou non. Il existe (au moins) une matrice inverse \( M \) telle que la matrice \( MA \) soit triangulaire supérieure.

**Preuve:**

Si \( A \) est inversible, le résultat est déjà prouvé en appliquant la méthode de Gauss
sans pivot et en posant $M = M_{n-1} \cdots M_1$. Si $A$ n’est pas inversible cela signifie qu’à une certaine étape $k$ on trouve $a_{ik}^{(k)} = 0$ pour tout $i \geq k$. Mais dans ce cas, il suffit de passer à l’étape suivante.
Matriciellement, cela reviendrait à prendre $I_k = M_k = I$.

### 2.3.3 Méthode de Gauss avec pivot total

On pourrait aussi adopter la stratégie du pivot total qui consiste, à chaque étape $k$, à prendre $a_{kk}^{(k)}$ tel que : $a_{kk}^{(k)} = \max_{i \geq k, j \geq k} |a_{ij}^{(k)}|$. Ce qui reviendrait à multiplier la matrice $A^{(k)}$ par deux matrices de permutation $P$ et $Q$, l’une à droite pour permuter les lignes et l’autre à gauche pour permuter les colonnes. 

$$A^{(2)} = M_1 I_{1i} A^{(1)} I_{1j} \cdots , \quad A^{(n)} = M_{n-1} I_{n-1i} \cdots M_1 I_{1j} A^{(1)} I_{1j} \cdots I_{n-1j}.$$ 

### 2.3.4 Méthode de Gauss-Jordan

C’est une variante qui ressemble à la méthode de Gauss sauf qu’elle aboutit directement à une matrice diagonale. Au lieu des matrices $M_k$ élémentaires on considère les matrices

$$\bar{M}_k = \begin{pmatrix}
1 & & & & & & & \\
0 & & & & & & & -m_{1k}^{(k)} \\
0 & & & & & & \ddots & \\
0 & & & & \ddots & -m_{k-1k}^{(k)} & 0 \\
\vdots & & \ddots & & -m_{k+1k}^{(k)} & \ddots & \\
\vdots & & & & \ddots & \ddots & \\
0 & & & & & -m_{k+1k}^{(k)} & 1
\end{pmatrix}.$$ 

### 2.4 Factorisation $LU$

Si on suppose que la méthode de Gauss sans pivot a été appliquée à toutes les étapes du procédé on aboutit à

$$A^{(n)} = M_{n-1} M_{n-2} \cdots M_2 M_1 A^{(1)} = U \quad \text{avec} \quad A^{(1)} = A,$$
ou encore $MA = U$ avec $M = M_{n-1} \cdots M_1$. Comme les matrices élémentaires de Gauss sont triangulaires inférieures et inversibles. En posant $L = M^{-1}$ on obtient

$A = LU$. 

32
Remarque 2.4.1. Si on utilise des permutations, alors les matrices $M_k I_{ik}$ ne sont plus triangulaires inférieures. On démontre dans ce cas qu’on aboutit à la forme $P A = LU$.

Théorème 2.4.1 (Condition suffisante de la factorisation $LU$).

Soit $A$ une matrice carrée d’ordre $n$ telle que toutes les sous-matrices d’ordre $k$ ($k \leq n$) soient inversibles, alors il existe une matrice triangulaire inférieure $L$ avec $l_{ii} = 1$ et une matrice triangulaire supérieure $U$ telles que $A = LU$. De plus, cette factorisation est unique.

Preuve:

Si $a_{11} \neq 0$, la matrice $a_{11}$ (d’ordre 1) est inversible, donc on peut choisir la matrice de permutation égale à l’identité et appliquer la méthode de Gauss sans pivot à la première étape. Supposons qu’on ait pu choisir toutes les matrices de permutation égales à l’identité jusqu’à l’étape $k$, il s’ensuit que

$$A^{(k)} = M_{k-1} M_{k-2} \cdots M_1 A = \prod_{i=k-1}^{i=1} M_i A.$$  

Avec

$$A_k = \begin{pmatrix} a^{(k)}_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a^{(k)}_{kk} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a^{(k)}_{nk} \end{pmatrix}$$

en écrivant $A$ sous forme de blocs et en effectuant le produit matriciel par blocs, on obtient $a^{(1)}_{11} \times \cdots \times a^{(k)}_{kk} = \det(B_k)$. Comme $\det(B_k) \neq 0$ on a $a^{(k)}_{kk} \neq 0$ et par suite on peut choisir $a^{(k)}_{kk}$ comme pivot et poursuivre le procédé.
Unicité
Supposons qu’il existe $L_1, L_2, U_1$ et $U_2$ telles que $A = L_1 U_1 = L_2 U_2$, comme $L_2$ et $U_1$ sont inversibles alors $L_2^{-1} L_1 = U_2 U_1^{-1}$.
Ce qui impose $L_2^{-1} L_1 = U_2 U_1^{-1} = I$ et donc $L_1 = L_2$ et $U_1 = U_2$.

2.5  Factorisation de Choleski (matrice symétrique)

Théorème 2.5.1. Si $A$ est une matrice symétrique, définie positive, il existe (au moins) une matrice réelle triangulaire inférieure $L$ telle que $A = LL^\top$.
Si de plus on impose aux éléments diagonaux de $L$ d’être strictement positifs, alors la factorisation est unique.

Preuve:
Remarquons d’abord que si $A$ est définie positive, alors toutes les sous-matrices d’ordre $k$ sont inversibles. Le théorème 2.4.1 permet d’affirmer l’existence de deux matrices $L$ et $U$ telles que $A = LU$. Ce que nous cherchons ici c’est de factoriser en utilisant une seule matrice $L$. Raisonnons par récurrence sur $n$.
Si $k = 1$, $A = a_{11} > 0$ donc $a_{11} = \sqrt{a_{11}} \sqrt{a_{11}}$.
Supposons qu’on ait pu factoriser jusqu’à l’ordre $k-1$ et soit $A_k$ une matrice d’ordre $k$ alors $A_k$ peut s’écrire :

$$A_k = \begin{pmatrix} A_{k-1} & v \\ v^\top & a_{kk} \end{pmatrix}$$
avec $A_{k-1} = L_{k-1} L_{k-1}^\top$.

Considérons alors la matrice $L_k$ obtenue à partir de $L_{k-1}$ et telle que :

$$L_k = \begin{pmatrix} L_{k-1} \\ l^\top \end{pmatrix}
\begin{pmatrix} l \\ l_k \end{pmatrix}$$

Le produit matriciel $L_k L_k^\top$ donne :

$$L_k L_k^\top = \begin{pmatrix} L_{k-1} L_{k-1}^\top & L_{k-1} l \\ l^\top L_{k-1} & l^\top l + l_k^2 \end{pmatrix}$$

Par identification on obtient :

$$L_{k-1} l = v$$
$$L_{k-1} L_{k-1}^\top = A_{k-1}$$
$$l^\top l + l_k^2 = a_{kk}$$

i) L’équation (2.5.1) permet alors de résoudre un système et d’obtenir la solution qui est le vecteur $l$. 

34
ii) L’équation (2.5.3) permet d’obtenir la dernière inconnue du problème, à savoir
\[ l_{kk} = \sqrt{a_{kk} - \hat{l}^\top \hat{l}} \text{ et on peut choisir } l_{kk} > 0. \]

**Exemple 2.5.1.** Soit \( A \) la matrice de Hilbert d’ordre 6, la factorisation de Choleski est donnée par \( A = LL^\top \) où

\[
L = \begin{pmatrix}
1 & 0.5 & 0.33 & 0.25 & 0.2 & 0.16 \\
0 & 0.28 & 0.28 & 0.25 & 0.2 & 0.13 \\
0 & 0 & 0.07 & 0.11 & 0.12 & 0.05 \\
0 & 0 & 0 & 0.01 & 0.03 & 0.01 \\
0 & 0 & 0 & 0 & 0.004 & 0.01 \\
0 & 0 & 0 & 0 & 0 & 0.0012
\end{pmatrix}.
\]

### 2.6 Factorisation de Householder (matrice unitaire)

Soit \( P_0 = I - 2\omega_0\omega_0^\top \) une matrice élémentaire de Householder avec

\[
\omega_0^\top \omega_0 = 1. \tag{2.6.1}
\]

On cherche une matrice unitaire \( P_0 \) telle que

\[
P_0a = ke_1, \tag{2.6.2}
\]

pour tout vecteur \( a = (a_1, \ldots, a_n)^\top \), avec \( k \in \mathbb{R} \) et \( e_1 = (1, 0, \ldots, 0)^\top \).

\( P_0 \) est orthogonale c’est à dire \( P_0^\top P_0 = I \) et par suite, on doit avoir

\[
\left( a^\top P_0^\top \right) (P_0a) = k^2 = a^\top a.
\]

Soit \( k = \pm \left( a^\top a \right)^{1/2} \), les équations (2.6.1) et (2.6.2) donnent :

\[
P_0a = a - 2\omega_0\omega_0^\top a = ke_1 \text{ et } 2\omega_0\omega_0^\top a = -ke_1 + a = \nu, \text{ si on pose } \alpha = 2\omega_0^\top a. 
\]

On obtient \( \alpha \omega_0 = \nu \), et comme on cherche \( \omega_0 \) tel que \( \omega_0^\top \omega_0 = 1 \), il vient : \( \alpha^2 = \nu^\top \nu \).

Par suite \( P_0 = I - \frac{2}{\alpha^2\nu^\top \nu} = I - \frac{2\nu\nu^\top}{\nu^\top \nu} \).

**Remarques 2.6.1.**

i) Le choix de \( k \) se fait au signe près, on peut choisir le signe +.

ii) Le même procédé peut être appliqué pour obtenir une matrice

\[
P_k = I_k - 2\omega_k\omega_k^\top \text{ avec } \omega_k = (0, \ldots, 0, \omega_{k+1k}, \ldots, \omega_{nk})^\top.
\]

On a constaté que \( P_k \) peut être décomposée sous la forme

\[
P_k = \begin{pmatrix} I_k & \hat{P}_k \end{pmatrix} \quad \text{ avec } \hat{P}_k = I_{n-k} - 2\hat{\omega}_k\hat{\omega}_k^\top \text{ (voir paragraphe 2.2.3).}
\]
La factorisation de Householder permet d’écrire :

\[ P_{n-2}P_{n-3} \cdots P_1 P_0 A = U, \]

ou encore \( A = Q U \) avec \( Q = P_0 P_1 \cdots P_{n-2} \) une matrice orthogonale.

### 2.7 Conditionnement

**Exemple 2.7.1.** [dû à R.S. Wilson]

Soit à résoudre le système linéaire \( Ax = b \) avec

\[
A = \begin{pmatrix}
10 & 7 & 8 & 7 \\
7 & 5 & 6 & 5 \\
8 & 6 & 10 & 9 \\
7 & 5 & 9 & 10
\end{pmatrix},
\]

\[
b = \begin{pmatrix}
32 \\
23 \\
33 \\
31
\end{pmatrix}.
\]

La solution exacte est donnée par \( x = (32, 23, 33, 31)^\top \), si on perturbe le second membre d’un \( \delta b \), quel est l’effet de cette perturbation sur la solution?

Soit \( b + \delta b = (32.1, 22.9, 33.1, 30.9)^\top \), alors il en résulte une solution

\( \tilde{x} = x + \delta x = (9.2, -12.6, 4.5, -1.1) \).

Soit une erreur de l’ordre de \( \frac{1}{200} \) sur les données et un rapport d’amplification de l’erreur relative de l’ordre 2000 ce qui montre que le système n’est pas stable, il reste vulnérable à toute perturbation, on caractérise la stabilité intrinsèque d’un tel système indépendamment de la méthode de résolution en disant qu’il est mal conditionné. De même si on perturbe les éléments \( a_{ij} \) de la matrice \( A \) de \( \delta A \), on aboutit à une solution approchée \( \tilde{x} \) qui est très différente de la solution exacte.

**Définition 2.7.1.** Soit \( \| . \| \) une norme subordonnée et \( A \) une matrice inversible. On appelle conditionnement de \( A \) relativement à la norme \( \| . \| \), le nombre \( \| A \| \| A^{-1} \| \) noté \( C(A) \) ou \( \text{cond}(A) \).

**Propriétés du conditionnement**

On vérifie facilement les propriétés suivantes

i) \( C(A) \geq 1 \).

ii) \( C(A) = C(A^{-1}) \).

iii) \( C(\alpha A) = |\alpha|C(A), \quad \alpha \neq 0 \).

iv) \( C_2(A) = \| A \|_2 \| A^{-1} \|_2 = \frac{\mu_n(A)}{\mu_1(A)}, \mu_i \) valeur singulière de \( A \).
v) $C(QA) = C(A)$ pour toute matrice orthogonale $Q$.

**Théorème 2.7.1.** Soit $A$ une matrice inversible, $x$ et $x + \delta x$ les solutions respectives de $Ax = b$ et $A\tilde{x} = b + \delta b$ où $\tilde{x} = x + \delta x$, alors on a

$$\frac{1}{C(A)} \left( \frac{\|\delta b\|}{\|b\|} \right) \leq \frac{\|\delta x\|}{\|x\|} \leq C(A) \left( \frac{\|\delta b\|}{\|b\|} \right).$$

(2.7.1)

De même la solution obtenue après perturbation de $A$ par $\delta A$ vérifie :

$$\frac{\|\delta x\|}{\|x\|} \leq \left( \frac{C(A)}{1 - C(A)} \right) \left( \frac{\|\delta A\|}{\|A\|} \right).$$

Si on perturbe en même temps $A$ et $b$ alors on obtient :

$$\frac{\|\delta x\|}{\|x\|} \leq \left( \frac{C(A)}{1 - C(A)} \right) \left( \frac{\|\delta A\|}{\|A\|} \right) \left( \frac{\|\delta b\|}{\|b\|} \right).$$

**Preuve:**

On a $Ax = b$ et $Ax + A\delta x = b + \delta b$ d'où $A\delta x = \delta b$ ou encore $\delta x = A^{-1}\delta b$, comme $\|b\| \leq \|A\|\|x\|$, on déduit que $\frac{\|\delta x\|}{\|x\|} \leq C(A) \frac{\|\delta b\|}{\|b\|}$.

Les autres inégalités s'obtiennent de façons similaires.

**Remarques 2.7.1.**

i) L’équation (2.7.1) donne une estimation de l’erreur relative de la solution en fonction de l’erreur relative connue $\frac{\|\delta b\|}{\|b\|}$.

ii) Tous les calculs sont effectués sur un ordinateur, des erreurs d’arrondi sont accumulées à chaque étape de calcul. Si $\varepsilon$ désigne la précision numérique relative (dépend de la machine), l’erreur relative de la solution explode si $C(A) \times \varepsilon \approx 1$.

iii) La matrice de Hilbert d’ordre $n$, $H = \left( \frac{1}{i + j - 1} \right)_{i,j=1}^n$ présente un exemple classique de matrices mal conditionnées.

### 2.8 Matrices creuses

On appelle matrice creuse une matrice dont la plupart des éléments sont égaux à zéro. Si de plus la structure des éléments non nuls est simple, il n’est pas nécessaire de réserver une quantité de mémoire égale à celle de la matrice complète. Des algorithmes spécifiques permettent de réduire le temps de calcul de manière considérable. Parmi les cas simples des matrices creuses, citons les matrices :
- tridiagonales: les éléments non nuls sont sur la diagonale et de part et d’autre de celle-ci sur les deux lignes adjacentes. On a $a_{ij} = 0$ pour $|i - j| > 1$.

- diagonale par bande de largeur $m$: les éléments de matrices tels que $|i - j| > m$ sont nuls,

- simplement ou doublement bordées : par rapport à la définition précédente, des éléments non nuls supplémentaires existent le long des lignes ou colonnes du bord de la matrice.

**Matrices tridiagonales**

Soit le système

$$
\begin{pmatrix}
  d_1 & e_1 & & & & \\
  c_2 & d_2 & e_2 & & & \\
  & \ddots & \ddots & \ddots & & \\
  & & c_{n-1} & d_{n-1} & e_{n-1} & \\
  & & & c_n & d_n & \\
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  \vdots \\
  \vdots \\
  \vdots \\
  x_n \\
\end{pmatrix}
=
\begin{pmatrix}
  b_1 \\
  \vdots \\
  \vdots \\
  \vdots \\
  b_n \\
\end{pmatrix}
$$

(2.8.1)

Il y’a un vaste choix de bibliothèques disponibles pour calculer les solutions qui prennent un temps de calcul proportionnel à $n$.

De manière générale, on peut obtenir un algorithme proportionnel à $n$ pour une matrice à bande. Le préfacteur de l’algorithme est proportionnel à $m$.

**Formule de Sherman-Morison**

Supposons que l’on ait une matrice $A$ dont on a facilement calculé l’inverse (cas d’une matrice triangulaire). Si on fait un petit changement dans $A$ en modifiant par exemple un ou quelques éléments de l’ensemble de la matrice, peut on calculer facilement l’inverse de cette nouvelle matrice?

En effet soit $B$ une matrice telle que

$$
B = A + uv^\top
$$

(2.8.2)
La formule de Sherman-Morison donne l’inverse de $B$.

\[
B^{-1} = \left(I + A^{-1}uv^\top\right)^{-1} A^{-1} \\
= \left(I - A^{-1}uv^\top + A^{-1}uv^\top A^{-1}uv^\top + \cdots \right) A^{-1} \\
= A^{-1} - A^{-1}uv^\top A^{-1} \left(1 - \lambda + \lambda^2 + \cdots \right) \\
= A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + \lambda}
\]

où $\lambda = v^\top A^{-1}u$. Posons $z = A^{-1}u$ et $w = (A^{-1})^\top v$ on a $\lambda = v^\top z$ et $B^{-1} = A^{-1} - \frac{zw^\top}{1 + \lambda}$.

**Remarque 2.8.1.** Dans le cas où les vecteurs $u$ et $v$ sont des matrices $U$ et $V$ d’ordre respectif $n \times k$ et $k \times n$ et si de plus la matrice $I + VA^{-1}U$ est inversible, alors

\[
(A + UV)^{-1} = A^{-1} - A^{-1}(I + VA^{-1}U)^{-1}VA^{-1}.
\]

Supposons que les éléments de la matrice $A$ vérifient

\[
\begin{align*}
i) & \quad |d_1| > |e_1| \\
ii) & \quad |d_i| > |e_i| + |c_i| \quad i = 1, \ldots, n - 1 \\
iii) & \quad |d_n| > |e_n|
\end{align*}
\]

et que

\[
c_i e_{i-1} \neq 0, \quad i = 2, \ldots, n
\]

Sous ces conditions la matrice $A$ est irréductible à diagonale dominante donc $A$ est inversible. Wendroff a montré qu’il existe une factorisation de la matrice $A$ sous la forme suivante

\[
A = \hat{L}\hat{R}
\]

où $\hat{R}$ est une matrice triangulaire supérieure avec éléments diagonaux égaux à 1 et $\hat{L}$ est une matrice triangulaire inférieure. Si $A$ est tridiagonale, les matrices $\hat{R}$ et $\hat{L}$ sont données par

\[
\hat{L} = \begin{pmatrix}
\alpha_1 \\
c_2 & \alpha_2 \\
& \ddots & \ddots \\
c_n & & \cdots & \alpha_n
\end{pmatrix}
\quad \text{et} \quad
\hat{R} = \begin{pmatrix}
1 & \gamma_1 & & \\
& \ddots & \ddots & \\
& & \ddots & \gamma_{n-1} \\
& & & 1
\end{pmatrix},
\]

où

\[
\begin{align*}
i) & \quad \alpha_1 = d_1, \quad \gamma_1 = \frac{e_1}{d_1} \\
ii) & \quad \alpha_i = d_i - c_i \gamma_{i-1}, \quad i = 2, \ldots, n \\
iii) & \quad \gamma_i = \frac{e_i}{\alpha_i}, \quad i = 2, \ldots, n - 1
\end{align*}
\]


Puisque $A$ est inversible et $\det(A) = \prod_{i=1}^{n} \alpha_i$, alors tous les $\alpha_i$ sont non nuls et le schéma récursif (2.8.6) est bien défini.

**Remarque 2.8.2.** On note que la factorisation (2.8.5) est obtenue par application de la méthode d'élimination de Gauss usuelle à $A^\top$.

La résolution du système linéaire est équivalente à

$$\hat{L}v = b, \quad \hat{R}x = v,$$

qui est obtenue de la manière suivante

$v_1 = b_1/\alpha_1$

*Pour $i = 2$ à $n$ :

$$v_i = (b_i - c_i v_{i-1})/\alpha_i$$

*fin

$x_n = v_n$

*Pour $i = 1$ à $n-1$ :

$$x_{n-i} = v_{n-i} - \gamma_{n-i} v_{n-i+1}$$

*fin

Le théorème suivant donne une majoration des quantités $\gamma_i$ et $\alpha_i$ indépendamment de l'ordre de la matrice $A$.

**Théorème 2.8.1.** Si les éléments de $A$ vérifient les conditions (2.8.3) alors

\begin{align*}
\text{i)} \quad & |\gamma| < 1 \\
\text{ii)} \quad & 0 < |d_i| - |c_i| < |d_i| + |c_i|, \quad i = 2, \ldots, n
\end{align*}

(2.8.7)

**Preuve:**

On a $\gamma_i = e_i\alpha_i = e_i/d_i$, la condition i) de (2.8.3) donne $|\gamma_i| < |e_i|/|d_i| < 1$.

Supposons que $\gamma_j < 1$ pour $j = 1, \ldots, i-1$, en remplacant (2.8.6 (iii)) dans (2.8.6 (ii)) on obtient

$$\gamma_i = \frac{e_i}{d_i - c_i \gamma_{i-1}}$$

et

$$|\gamma_i| < \frac{|e_i|}{|d_i| - |c_i| |\gamma_{i-1}|} < \frac{|e_i|}{|d_i| - |c_i|}$$

par hypothèse. Ensuite en considérant (2.8.3 (iii)) il s'ensuit que $|\gamma_i| < 1$.

Pour montrer (2.8.7 (ii)) il suffit de considérer (2.8.6 (ii)) et (2.8.3 (ii)).
L’algorithme présenté précédemment peut être simplifié en éliminant les $\alpha_i$, c’est le cas où la factorisation $\tilde{L}\tilde{R}$ n’est pas nécessaire. L’algorithme se présente comme suit

$$v_1 = b_1/d_1$$
$$t = d_1$$

Pour $i = 2$ à $n$

$$\gamma_{i-1} = e_{i-1}/t$$
$$t = d_i - c_i\gamma_{i-1}$$
$$v_i = (b_i - c_iv_{i-1})/t$$

fin $x_n = v_n$

Pour $i = 1$ à $n-1$

$$x_{n-i} = v_{n-i} - \gamma_{n-i}x_{n-i+1}$$

fin

Si $A$ est définie positive, elle possède une factorisation de la forme $L^\top DL$ qui est obtenue facilement du schéma (2.8.6). Si

$$L = \begin{pmatrix} 1 \\ l_1 \\ \vdots \\ \vdots \\ l_{n-1} \\ 1 \end{pmatrix}$$

(2.8.8)

et $D = diag(\delta_1, \cdots, \delta_n)$, par identification des écritures et en posant $c_i = e_{i-1}$ on obtient

$$\delta_1 = d_1$$

Pour $i = 1$ à $n-1$

$$l_i = e_i/\delta_i$$
$$\delta_{i+1} = d_{i+1} - l_ie_i$$

fin

$A$ est définie positive donc $u^\top L^\top DLu > 0$, posons $v = L^\top u$ ($v \neq 0$ car $L$ est inversible), alors $v^\top Dv > 0$, ainsi $D$ est une matrice diagonale définie positive, dont les éléments diagonaux sont tous non nuls, et par suite le schéma récursif est bien défini. Si on pose $Lv = b$, l’algorithme de résolution s’écrit

$$v_1 = b_1$$

Pour $i = 1$ à $n-1$

$$v_{i+1} = b_{i+1} - l_iv_i$$

fin

$$x_n = v_n/\delta_n$$
Pour $i = 1$ à $n-1$

\[ x_{n-i} = \frac{(v_{n-i} - e_{n-i}x_{n-i+1})}{\delta_{n-i}} \]

fin

Si la factorisation n’est pas nécessaire on obtient l’algorithme simplifié suivant

\[ v_1 = b_1 \]
\[ \delta_1 = d_1 \]

Pour $i = 1$ à $n-1$

\[ t = e_i/\delta_i \]
\[ \delta_{i+1} = d_{i+1} - te_i \]
\[ v_{i+1} = b_{i+1} - tv_i \]

fin

\[ x_n = v_n/\delta_n \]

Pour $i = 1$ à $n-1$

\[ x_{n-i} = \frac{(v_{n-i} - e_{n-i}x_{n-i+1})}{\delta_{n-i}} \]

fin

**Exemple 2.8.1.** Soit le système linéaire $Ax = b$, où $A$ est une matrice carrée d’ordre $n$ de la forme

\[
A = \begin{pmatrix}
a_1 & \cdots & \cdots & \cdots & \cdots & a_n \\
1 & \lambda & 1 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
1 & \lambda & 1 \\
\end{pmatrix},
\]

Ce type de matrices est issu de la discrétisation de certains problèmes paraboliques voir [32, 130].

Considérons la matrice

\[
A_0 = \begin{pmatrix}
\alpha & 1 \\
1 & \alpha & 1 \\
\vdots & \ddots & \ddots & \ddots \\
& \ddots & \ddots & \ddots & \ddots \\
1 & \alpha & 1 \\
1 & 1 & \alpha \\
\end{pmatrix},
\]

avec $\alpha = \frac{\lambda - \sqrt{\lambda^2 - 4}}{2}$ et $\beta = \frac{\lambda + \sqrt{\lambda^2 - 4}}{2}$. Il s’ensuit que

\[ A_0 = LU \]
avec
\[ L = \begin{pmatrix} \alpha & \cdots & \cdots & \cdots & \cdots \alpha \\ 1 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \ddots & \ddots & \cdots \\ 1 & \cdots & \cdots & \cdots & \beta \end{pmatrix}, \quad U = \begin{pmatrix} 1 & \beta & \cdots & \cdots & \cdots \beta \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \cdots & \ddots & \cdots \end{pmatrix}. \]

La solution \( y \) de \( A_0 y = \omega \), est donnée par
\[
y_n = v_n \\
y_m = v_m - \beta y_{m+1}, \quad m = 0, \ldots, n - 1,
\]
où \( v \) est donnée par
\[
v_n = \beta \omega_0 \\
v_m = \beta (\omega_m - \omega_{m-1}), \quad m = 1, \ldots, n.
\]

Soit \( z = x - y \), on obtient
\[ Az = \begin{pmatrix} \omega_0 - \sum_{k=0}^{\infty} a_k y_{k} & \cdots & \omega_n - \sum_{k=0}^{\infty} b_k y_{k} \end{pmatrix}^\top.
\]
Comme
\[ z_{m-1} + \lambda z_m + z_{m+1} = 0, \quad m = 1, \ldots, n - 1.
\]
Il vient
\[ z_m = c_0 \gamma^{n-m} + c_1 \gamma^m, \quad m = 0, \ldots, n,
\]
où \( \gamma = \frac{-\lambda - \sqrt{\lambda^2 - 4}}{2} \) et \( c_0, c_1 \) sont des constantes à déterminer en résolvant le système
\[
c_0 \sum_{k=0}^{\infty} a_k \gamma^{n-k} + c_1 \sum_{k=0}^{\infty} a_k \gamma^k = \omega_0 - \sum_{k=0}^{\infty} a_k y_{k} \\
c_0 \sum_{k=0}^{\infty} b_k \gamma^{n-k} + c_1 \sum_{k=0}^{\infty} b_k \gamma^k = \omega_n - \sum_{k=0}^{\infty} b_k y_{k}
\]
Finalement on obtient la solution du système initial en posant \( x = y + z \).

### 2.9 Résultats sur les matrises non carrées

**Théorème 2.9.1.** soit \( A \) une matrice rectangulaire \( (m, n) \) alors il existe deux matrices carrées unitaires \( U \) et \( V \) d’ordre respectivement \( m \) et \( n \) tel que : \( U^* AV = \Sigma \), où \( \Sigma \) est une
matrice rectangulaire \((m,n)\) donnée par :

\[
\Sigma = \begin{pmatrix}
\mu_1 & \cdots & \mu_r \\
0 & \ddots & 0 \\
\vdots & \ddots & \vdots \\
m & & 0
\end{pmatrix}
\]

\(m > n\)

où les \(\mu_i\) sont les valeurs singulières de \(A\).

**Corollaire 2.9.1.**

- Le rang de \(A\) est égal au nombre de valeurs singulières non nulles.

- La forme \(A = U\Sigma V^*\) est appelée décomposition en valeurs singulières de \(A\) et on a
  \(A = \sum_{i=1}^{r} \mu_i u_i v_i^*\) et \(A^* A = \sum_{i=1}^{r} \mu_i^2 u_i v_i^*\) où \(u_i\) et \(v_i\) désignent, respectivement, les \(i\)ème colonnes de \(U\) et de \(V\).

**Preuve:**

Soient \(I_n = \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix}\) et \(I_m = \begin{pmatrix} e_1 & \cdots & e_m \end{pmatrix}\)

\(\Sigma V^* = \sum_{i=1}^{n} \mu_{i} e_i v_i^*\),  \(V^* = \sum_{i=1}^{n} e_i v_i^*\),  \(\Sigma e_i = \mu_{i} e_i\) pour  \(i = 1 \cdots n\),

\(\Sigma V^* = \sum_{i=1}^{n} \mu_{i} e_i v_i^*,\) \(U\Sigma V^* = \sum_{i=1}^{r} \mu_{i} u_i v_i^*\) et \(A^* A = \sum_{i=1}^{r} \mu_{i}^2 u_i v_i^*\).

**2.10 Résolution des systèmes à matrices non carrées**

**Exemple 2.10.1.**

\[
\begin{pmatrix}
1 & 2 \\
-1 & 1 \\
-1 & -3 \\
-1 & -2
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= \begin{pmatrix}
2 \\
1 \\
1 \\
2
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 & 2 & 2 \\
1 & 2 & 1 \\
2 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix}
= \begin{pmatrix}
1 \\
2
\end{pmatrix}
\]
Cas où \( m \geq n \) et \( \text{rang } A = n \)

Considérons le système \( Ax = b \), on définit le \( i \)ème résidu

\[
r_i = \sum_{j=1}^{n} a_{ij}x_j - b_i \quad i = 1, \ldots, m.
\]

(Vectoriellement \( r = Ax - b \)).

La méthode des moindres carrés consiste à minimiser \( \|r\|_2^2 \).

Posons

\[
I(x) = r^\top r = (Ax - b)^\top (Ax - b).
\]

Donc minimiser \( I(x) \) revient à chercher des solutions de l’équation

\[
\frac{\partial I(x)}{\partial x_i} = 0.
\]

Posons \( e_i = (0, \ldots, 1, \ldots, 0)^\top \), on a d’une part

\[
r^\top r = x^\top A^\top Ax - b^\top Ax - x^\top A^\top b + b^\top b
\]

\[
= x^\top A^\top Ax - 2x^\top A^\top b + b^\top b.
\]

D’autre part

\[
\frac{\partial I(x)}{\partial x_i} = e_i^\top A^\top Ax + xA^\top Ae_i - 2e_i^\top A^\top b
\]

\[
= 2e_i^\top (A^\top Ax - A^\top b).
\]

Donc trouver un minimum de \( I(x) \) revient à résoudre le système linéaire

\( A^\top Ax = A^\top b \).

Définition 2.10.1. Soit \( \Sigma \) la matrice rectangulaire \((m, n)\) donnée par

\[
\Sigma = \begin{pmatrix}
\mu_1 & \cdots & \mu_n \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{pmatrix}
\]

\[
m > n
\]
on appelle **pseudo-inverse** de $\Sigma$ la matrice $\Sigma^+$ de la forme $(n, m)$ définie par:

$$\Sigma^+ = \begin{pmatrix}
\mu_1^{-1} & \cdots & 0 \\
0 & \ddots & \vdots \\
0 & \cdots & \mu_r^{-1}
\end{pmatrix}
$$

$m > n$

**Définition 2.10.2.** Soit $A$ une matrice de forme $(m, n)$ dont la décomposition en valeurs singulières est $A = U\Sigma V^*$. On appelle **pseudo-inverse** (ou inverse généralisé) de la matrice $A$, la matrice $A^+$ de la forme $(n, m)$ donnée par $A^+ = V\Sigma^+ U^*$.

**Remarque 2.10.1.**

1. $A^+A = V\Sigma^+ U^* U \Sigma V^* = V \Sigma^+ \Sigma V^*$.

2. Sous Matlab la commande `svd(A)` donne la décomposition en valeurs singulières de $A$.

**Exemple 2.10.2.** Soit la matrice $A$ donnée

$$A = \begin{pmatrix}
0.1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.5 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.6 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.7 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0.8 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}.$$ 

La décomposition en valeurs singulières de $A$ est donnée par

$$A = UwV^T,$$

où

$$U = \begin{pmatrix}
0.34 & 0.54 & -0.32 & -0.003 & -0.30 & -0.48 & 0.09 & -0.37 \\
0.34 & 0.39 & 0.55 & -0.23 & -0.3 & 0.12 & 0.02 & 0.5 \\
0.35 & 0.23 & -0.12 & 0.33 & 0.66 & 0.09 & 0.44 & 0.22 \\
0.35 & 0.08 & -0.1 & 0.08 & -0.13 & 0.78 & -0.16 & -0.43 \\
0.35 & -0.07 & -0.23 & 0.19 & 0.14 & -0.16 & -0.78 & 0.32 \\
0.35 & -0.22 & 0.02 & -0.78 & 0.38 & -0.09 & -0.01 & -0.2 \\
0.35 & -0.38 & 0.59 & 0.41 & -0.01 & -0.27 & 0.03 & -0.35 \\
0.36 & -0.53 & -0.39 & -0.001 & -0.43 & 0.02 & 0.37 & 0.31
\end{pmatrix}.$$
Théorème 2.10.1. Si $A$ est une matrice rectangulaire $(m,n)$ alors on a

1. $A^+ = \sum_{i=1}^{n} \mu_i^{-1} v_i u_i^*$.

2. $AA^+ = \sum_{i=1}^{n} u_i u_i^*$ matrice de projection orthogonale sur $\text{Im}A$.

3. $A^+ A = \sum_{i=1}^{n} v_i v_i^*$ matrice de projection orthogonale sur $\text{Im}A^*$.

Remarques 2.10.2. 1. Le nombre $\mathcal{K}(A) = \|A\|_2\|A^+\|_2$ est appelé conditionnement généralisé de $A$.

2. Si $\text{rang}(A) = n$ alors $\mathcal{K}^2(A) = \mathcal{K}(AA^\top) = \text{Cond}(AA^\top)$.

3. La commande MATLAB \ (backslash) est la commande générique pour résoudre un système linéaire. L'algorithme mis en œuvre dépend de la structure de la matrice $A$ du système. MATLAB utilise dans l’ordre les méthodes suivantes:

   i) Si $A$ est une matrice triangulaire, le système est résolu par simple substitution.
ii) Si $A$ est une matrice symétrique ou hermitienne, définie positive, la résolution est effectuée par la méthode de Choleski.

iii) Si $A$ est une matrice carrée, mais n’entrant pas dans les deux cas précédents, une factorisation $LU$ est réalisée en utilisant la méthode d’élaboration de Gauss avec stratégie de pivot partiel.

iv) Si $A$ n’est pas une matrice carrée, la méthode QR est utilisée.

4. La matrice $A$ peut être creuse, elle comporte une forte proportion de coefficients nuls (de nombreux problèmes issus de la physique conduisent à l’analyse de systèmes linéaires à matrices creuses), l’intérêt de telles matrices résulte non seulement de la réduction de la place mémoire (on ne stocke pas les zéros) mais aussi de la réduction des nombres d’opérations. Dans le cas de ces matrices des algorithmes particuliers sont mis en œuvre.

5. Chacune des méthodes précédentes peut être utilisée de manière spécifique grâce aux commandes chol, lu, qr.

2.11 Conclusion

Avant d’entamer la résolution des systèmes linéaires de grandes tailles, il est impératif de commencer par une analyse des propriétés de la matrice afin de déterminer la méthode la plus adaptée afin d’obtenir une solution avec une précision correcte et pour un temps de calcul qui sera minimal. Les différentes méthodes dans ce chapitre ne sont qu’une introduction à ce très vaste sujet.
2.12 Exercices

Exercice 2.12.1. Une matrice $A = (a_{ij})$ carrée d’ordre $n$ à diagonal strictement dominante en colonnes.


2. Montrer que la matrice $B$ d’ordre $(n - 1)$ obtenue à partir de $A^{(2)}$ en enlevant la première ligne et la première colonne est à diagonal strictement dominante en colonnes.

3. Quel est l’intérêt de cette propriété?

Exercice 2.12.2. Soit $a$ et $b$ deux vecteurs ayant le même nombre de composantes et $\alpha$ un scalaire.

1. Montrer que la matrice $I - \alpha ab^\top$ admet une matrice inverse de la même forme en imposant une condition sur $\alpha$. En déduire l’expression de $(A + ab^\top)^{-1}$ en fonction de $A^{-1}$ et des vecteurs $a$ et $b$.

2. Appliquer ce résultat aux matrices de Gauss: $I + me_k^\top$

3. Montrer que:

   $$(I + m_1 e_1^\top)(I + m_2 e_2^\top)\cdots(I + m_{n-1} e_{n-1}^\top) = I + m_1 e_1^\top + \cdots + m_{n-1} e_{n-1}^\top$$

Exercice 2.12.3. Soit $A$ une matrice rectangulaire $(m, n)$ avec $m > n$.

1. Montrer que il existe deux matrices carrées unitaires $U$ et $V$ d’ordre respectivement $m$ et $n$ telles que: $U^*AV = \widetilde{D}$ où $\widetilde{D}$ est une matrice rectangulaire $(m, n)$ donné par:

   $$\widetilde{D} = \begin{pmatrix}
   \mu_1 &  & \\
   & \ddots & \\
   0 & \cdots & 0 \\
   \vdots & \cdots & \vdots \\
   0 & \cdots & 0
   \end{pmatrix}$$

   avec $\mu_i$ valeurs singulières de $A$

2. Déduire le rang de $A$
Exercice 2.12.4. Soit \( A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 5 & 3 & 4 \\ 0 & 3 & -20 & 15 \\ 0 & 4 & 15 & -5 \end{pmatrix} \)

1. Utiliser les transformations de Householder pour obtenir une matrice tridiagonale.

2. Utiliser les transformations de Givens pour obtenir une matrice tridiagonale.

Exercice 2.12.5. Soit \( B = (b_{ij}) \) une matrice tridiagonale donnée par: \( b_{ij} = 1 \) et \( b_{i-1,i} = b_{i+1,i} = \frac{1}{4} \) et \( E \) une matrice carrée vérifiant: \( \|E\|_1 = \epsilon < \frac{1}{2} \).
Montrer que les systèmes \( Bx = b \) et \( (B + E)y = b \) ont des solutions uniques \( x \) et \( y \) et prouver que: \( \|y - x\|_1 \leq \frac{4\epsilon}{1-2\epsilon} \|b\|_1 \).
Chapitre 3

Méthodes indirectes de résolution des systèmes linéaires $Ax = b$

3.1 Introduction

Pour résoudre le système

$$Ax = b,$$  \hspace{1cm} (3.1.1)

on utilise des méthodes, dites indirectes, du type

$$x^{(k+1)} = Tx^{(k)} + C$$ \hspace{1cm} (3.1.2)

où $T$ est une matrice obtenue à partir de $A$ et $C$ un vecteur dépendant de $A$ et de $b$. Pour passer de l’équation (3.1.1) à l’équation (3.1.2) on écrit $A$ sous la forme : $A = M - N$ avec $M$ inversible. En remplaçant dans (3.1.2) on obtient $Mx = Nx + b$ et par la suite

$$x = M^{-1}Nx + M^{-1}b,$$ \hspace{1cm} (3.1.3)

qui suggère le procédé itératif suivant

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$$ \hspace{1cm} (3.1.4)

qui est de la forme (3.1.2) avec $T = M^{-1}N$ et $C = M^{-1}b$.

Remarque 3.1.1. En général il y a une multitude de façons d’écrire $A$ sous la forme $A = M - N$. Dans le cadre de ce chapitre, nous nous limiterons à deux familles de décomposition. La première comprend les méthodes classiques $A = M - N = D - L - U$, en précisant:

- Méthode de Jacobi: $M = D, N = L + U$. 

51
• Méthode de Gauss-Seidel: \( M = D - L, N = U \).
• Méthode de relaxation : \( A = A(\omega) = M(\omega) - N(\omega) \), avec \( M(\omega) = \frac{1}{\omega}D - L \),
\( N(\omega) = \frac{1-\omega}{\omega}D - U \) où \( \omega \) est un scalaire.

La deuxième sera consacrée aux matrices positives.

### 3.2 Généralités et définitions

**Définition 3.2.1.** Une méthode de type (3.1.2) est dite convergente si pour toute valeur initiale \( x(0) \) on a \( \lim_{n \to +\infty} x(n) = x \). Si une telle limite \( x \) existe alors elle vérifie \( Tx + C = x \).

**Définition 3.2.2.** On appelle erreur de la méthode (à la \( k \)ème itération) la quantité \( e(k) = x(k) - x \). Avec \( e(0) = x(0) - x \) on obtient \( e(k) = T^k e(0) \) pour tout \( k \).

La méthode est convergente si \( \lim_{k \to +\infty} T^k = 0 \).

**Définition 3.2.3.** Une matrice carrée \( B \) est dite convergente si \( \lim_{k \to +\infty} B^k = 0 \).

**Théorème 3.2.1.** Soit \( A \) une matrice carrée. Les assertions suivantes sont équivalentes :

i) \( \lim_{k \to +\infty} A^k = 0 \).

ii) \( \lim_{k \to +\infty} A^k v = 0 \) pour tout \( v \).

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

iv) \( \|A\| < 1 \), pour au moins une norme matricielle induite.

**Preuve:** On utilisera les résultats de l'exercice 1.5.3

i) \( \Rightarrow \) ii)

On a \( \|A^k v\| \leq \|A^k\| \|v\| \); l'assertion i) et la continuité de la norme implique que \( \lim_{k \to +\infty} \|A^k\| = 0 \) et par suite \( \lim_{k \to +\infty} \|A^k v\| = 0 \) et \( \lim_{k \to +\infty} A^k v = 0 \) pour tout \( v \).

ii) \( \Rightarrow \) iii)

Ceci revient à montrer que : Non iii) \( \Rightarrow \) Non ii)

Supposons que \( \rho(A) \geq 1 \) et soit \( v \) le vecteur propre associé à la valeur propre \( \lambda \) qui vérifie \( |\lambda| = \rho(A) \), donc \( Av = \lambda v \) et on en tire \( A^k v = \lambda^k v \) d'où, si \( |\lambda| \geq 1 \) finalement \( A^k v \) ne converge pas vers 0 pour ce \( v \) et donc Non ii) est vraie.

iii) \( \Rightarrow \) iv)

D’après l'exercice 1.5.3, pour tout \( \varepsilon > 0 \) il existe une norme induite telle que \( \|A\| \leq \rho(A) + \varepsilon \) il suffit de considérer un \( \varepsilon \) tel que \( \rho(A) + \varepsilon < 1 \), pour ce \( \varepsilon \),
l’exercice 1.5.3 assure que \( \|A\| \leq \rho(A) + \varepsilon \) et par suite \( \|A\| < 1 \).

iv) \( i \)\\
Elle est évidente car si \( \|A\| < 1 \) alors \( \lim_{k \to +\infty} \|A\|^k = 0 \) et par ailleurs \( \|A^k\| \leq \|A\|^k \)
d’où \( \lim_{k \to +\infty} \|A^k\| = 0 \) et \( \lim_{k \to +\infty} A^k = 0 \)

**Théorème 3.2.2.** Soit \( A \) une matrice carrée et \( \|.\| \) une norme quelconque,
alors \( \lim_{k \to +\infty} \left\| A^k \right\|^{1/k} = \rho(A) \).

**Preuve:**

i) On a \( \rho(A) \leq \|A\| \) et comme \( \rho(A^k) = (\rho(A))^k \)
il s’ensuit que \( \rho(A^k) = (\rho(A))^k \leq \|A\|^k \) et par suite \( \rho(A) \leq \|A^k\|^{1/k} \)
donc \( \rho(A) \leq \lim_{k \to +\infty} \left\| A^k \right\|^{1/k} \).

ii) Pour montrer que \( \lim_{k \to +\infty} \left\| A^k \right\|^{1/k} \leq \rho(A) \), introduisons pour tout \( \varepsilon > 0 \) la matrice \( A_\varepsilon = \frac{1}{\rho(A) + \varepsilon} A \), pour cette matrice on a \( \rho(A_\varepsilon) = \frac{\rho(A)}{\rho(A) + \varepsilon} < 1 \) et d’après le théorème précédent on obtient : \( \|A_\varepsilon\| < 1 \) pour au moins une norme matricielle induite, donc \( \lim_{k \to +\infty} A_\varepsilon^k = 0 \) et \( \lim_{k \to +\infty} \|A_\varepsilon\|^k = 0 \).

Donc \( \forall \varepsilon > 0, \exists N(\varepsilon) \) tels que pour tout \( k \geq N(\varepsilon) \) on ait : \( \|A_\varepsilon^k\| < 1 \)
or encore \( \|A^k\| \leq (\rho(A) + \varepsilon)^k \) soit encore \( \|A^k\|^{1/k} \leq (\rho(A) + \varepsilon) \) pour tout \( \varepsilon > 0 \)
finalement \( \lim_{k \to +\infty} \left\| A^k \right\|^{1/k} \leq \rho(A) \).

**Remarque 3.2.1.** \( \|A\|_2 = \rho(A^* A)^{1/2} \) et si \( A \) est hermitienne alors \( \|A\|_2 = \rho(A) \).

**Théorème 3.2.3.** Considérons deux méthodes itératives \( \tilde{x}^{(k+1)} = \tilde{T} \tilde{x}^{(k)} + \tilde{C} \) et \( x^{(k+1)} = T x^{(k)} + C \), avec \( \rho(T) < \rho(\tilde{T}) \) et \( x^{(0)} = \tilde{x}^{(0)} \), alors \( \forall \varepsilon > 0 \) \( \exists k_0 > 0 \) tel que \( \forall k \geq k_0 \) on a

\[
\sup_{\varepsilon \geq \varepsilon(0)} \frac{\tilde{e}^{(k)}}{e^{(k)}} \geq \left( \frac{\rho(\tilde{T})}{\rho(T) + \varepsilon} \right)^k.
\]

Donc la méthode itérative de matrice \( T \) converge plus rapidement que celle de matrice \( \tilde{T} \), en résumé, l’étude des méthodes itératives consiste à étudier les deux problèmes suivants:

1. Etant donné une méthode itérative de matrice \( T \), déterminer si la méthode converge, ie si \( \rho(T) < 1 \) ou s’il existe une norme telle que \( \|T\| < 1 \).

2. Etant donné deux méthodes itératives convergentes de matrice \( T \) et \( \tilde{T} \), les comparer, la méthode la plus rapide est celle ayant le plus petit rayon spectral.
Définition 3.2.4. On appelle taux moyen de convergence sur \( k \) itérations le nombre
\[
\bar{R}(k, T) = -\log \left\| T^k \right\|^{1/k}
\]
et taux asymptotique de convergence le nombre
\[
R(T) = \lim_{k \to +\infty} \bar{R}(k, T) = -\log(\rho(T)).
\]

\( R(T) \) joue le rôle de vitesse de convergence, plus \( R(T) \) est grand plus rapide est la convergence.

3.3 Description des méthodes classiques

3.3.1 Méthode de Jacobi

Elle consiste à choisir \( M = D = \text{diag}(a_{ii}) \) inverse et \( N = (-a_{ij})_{i \neq j} \) le schéma itératif est comme suit
\[
x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b.
\]

La matrice \( T_j = D^{-1}(L + U) \) est dite matrice de Jacobi associée à la matrice \( A \). Si \( x^{(0)} \) est le vecteur initial (donné), l’algorithme de Jacobi est de la forme
\[
x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{i \neq j} a_{ij}x_j^{(k)} + \frac{b_i}{a_{ii}} \quad i = 1, 2, \ldots, n.
\]

Cet algorithme nécessite \( a_{ii} \neq 0 \) pour \( i = 1, \ldots, n \). Explicitement, on obtient :
\[
\begin{align*}
a_{11}x_1^{(k+1)} &= -a_{12}x_2^{(k)} - \cdots - a_{1n}x_n^{(k)} + b_1 \\
& \vdots \\
a_{nn}x_n^{(k+1)} &= -a_{n1}x_1^{(k)} - \cdots - a_{nn-1}x_{n-1}^{(k)} + b_n
\end{align*}
\]

On a besoin de stocker les \( n \) composantes de \( x^{(k)} \) et les \( n \) composantes de \( x^{(k+1)} \).

Matriciellement, le schéma itératif est du même type que le schéma (3.1.2) avec \( T_j = D^{-1}(L + U) \) et \( C = D^{-1}b \).

D’après les théorèmes précédents, une condition suffisante pour que la méthode de Jacobi converge est \( \rho(T_j) < 1 \) ou \( \|T_j\|_{\infty} < 1 \).

Théorème 3.3.1. Si \( A \) est une matrice carrée à diagonale strictement dominante en lignes alors la méthode de Jacobi converge.

Preuve:

On a
\[
\sum_{j=1}^{n} |a_{ij}| < |a_{ii}| \quad i = 1, \ldots, n
\]  
(3.3.1)
d’autre part on a : 
\[ t_{ij} = -\frac{a_{ij}}{a_{ii}} \] pour \( i \neq j \) et \( t_{ii} = 0 \) d’où
\[ \|T_j\|_\infty = \max_i \sum_j |t_{ij}| = \max_i \left\{ \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| \right\} \] et d’après (3.3.1) on a \( \|T_j\|_\infty < 1 \).

**Corollaire 3.3.1.** Si \( A \) est une matrice à diagonale strictement dominante en colonnes, alors la méthode de Jacobi converge.

**Preuve :** Identique à celle du théorème 3.3.1.

Pour le cas des matrices irréductibles la stricte dominance en lignes ou en colonnes peut être affaiblie, pour de telles matrices le théorème qui va suivre assure que

**Théorème 3.3.2.** Si \( A \) est une matrice irréductible et vérifie
\[ (|a_{ii}| \geq \sum_{k \neq i} |a_{ik}|, \ i = 1, 2, \cdots, n) \], avec inégalité stricte pour au moins un indice \( i_0 \), alors la méthode de Jacobi converge.

**Preuve:**
On procède d’une manière analogue à celle de la preuve du théorème 3.3.1 pour montrer que \( \|T_j\|_\infty \leq 1 \).

Donc
\[ |T_j|e \leq e, \quad |T_j|e \neq e, \quad e = (1, 1, \cdots, 1)^\top. \] (3.3.2)

Puisque \( A \) est irréductible, \( T_j \) est aussi irréductible.

Pour montrer le théorème, il suffit de montrer que
\[ |T_j|^n e \leq e, \]

car si c’était le cas alors
\[ (\rho(T_j))^n = \rho(T_j^n) \leq \|(T_j^n)\|_1 < 1. \]

D’après (3.3.2) et le fait que \( |T_j| \geq 0 \), on a
\[ |T_j|^2 e \leq |T_j|e < e, \]
et par suite
\[ |T_j|^{i+1} e \leq |T_j|^i e \leq \cdots < e, \]
donc le vecteur \( t^{(i)} = e - |T_j|^i e \) satisfait
\[ 0 < t^{(1)} \leq t^{(2)} \leq \cdots \] (3.3.3)
Montrons que le nombre de composantes non nulles $\tau_i$ de $t^{(i)}$ croît avec $i$.
Si ce n’était pas le cas, (3.3.3) impliquerait qu’il existe $i \geq 1$ tel que $\tau_i = \tau_{i+1}$. $t^{(i)}$ peut s’écrire sous la forme

$$
\begin{pmatrix}
a \\
0
\end{pmatrix}, \quad a > 0, \quad a \in \mathbb{R}^p.
$$

(3.3.3) et $\tau_i = \tau_{i+1}$ impliquent que $t^{(i+1)}$ est aussi de la forme

$$
\begin{pmatrix}
b \\
0
\end{pmatrix}, \quad b > 0, b \in \mathbb{R}^p.
$$

Si $|T_j|$ est écrite sous la forme

$$
|T_j| = \begin{pmatrix}
|T_{11}| & |T_{12}| \\
|T_{21}| & |T_{22}|
\end{pmatrix}, \quad |T_{11}| \text{ matrice } p \times p.
$$

Il s’ensuit que

$$
\begin{pmatrix}
b \\
0
\end{pmatrix} = t^{(i+1)} = e - |T_j|^{i+1} e \geq |T_j| e - |T_j|^{i+1} e
$$

$$
= |T_j| t^{(i)} = \begin{pmatrix}
|T_{11}| & |T_{12}| \\
|T_{21}| & |T_{22}|
\end{pmatrix} \begin{pmatrix}
a \\
0
\end{pmatrix}
$$

Puisque $a > 0$, ceci n’est possible que si $T_{21} = 0$ ie $T_j$ est réductible. Ceci contredit les hypothèses du théorème.
Donc $0 < \tau_1 < \tau_2 < \cdots$, et $t^{(n)} > 0$. La preuve est ainsi achevée.
Pour remédier au problème du stockage et dans l’espoir d’améliorer les résultats en accélérant la convergence, on cherche une méthode qui utilise les composantes de $x^{(k+1)}$ au fur et à mesure qu’elles sont calculées. C’est ce que réalise la méthode de Gauss-Seidel.

### 3.3.2 Méthode de Gauss-Seidel

Pour cette méthode, les matrices $M$ et $N$ sont données par : $M = D - L$ inversible et $N = U$ où $D$, $L$ et $U$ proviennent de l’écriture $A = D - L - U$, le schéma itératif est comme suit :

$$
(D - L)x^{(k+1)} = Ux^{(k)} + b \quad \text{(3.3.4)}
$$

ou encore

$$
x^{(k+1)} = (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b \quad \text{(3.3.5)}
$$
en supposant que $D - L$ est inversible.

Les équations (3.3.4) et (3.3.5) peuvent aussi être présentées sous les formes:

$$D x^{(k+1)} = L x^{(k+1)} + U x^{(k)} + b$$

(3.3.6)

et (si $D$ est inversible)

$$x^{(k+1)} = D^{-1} L x^{(k+1)} + D^{-1} U x^{(k)} + D^{-1} b$$

(3.3.7)

en explicitant (3.3.6) on obtient:

$$a_{11} x_1^{(k+1)} = -a_{12} x_2^{(k)} - \cdots - a_{1n} x_n^{(k)} + b_1$$
$$a_{22} x_2^{(k+1)} = -a_{21} x_1^{(k+1)} - a_{23} x_3^{(k)} - \cdots - a_{2n} x_n^{(k)} + b_2$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{ii} x_i^{(k+1)} = -a_{i1} x_1^{(k+1)} - \cdots - a_{i-1} x_{i-1}^{(k+1)} - a_{i+1} x_{i+1}^{(k)} - \cdots - a_{in} x_n^{(k)} + b_i$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{nn} x_n^{(k+1)} = -a_{n1} x_1^{(k+1)} - \cdots - a_{n-1} x_{n-1}^{(k+1)} + b_n$$

La matrice $T_{GS} = (D - L)^{-1} U$ est dite matrice de Gauss-Seidel associée à la matrice $A$.

**Remarque 3.3.1.** Si $D$ est inversible, la matrice de Gauss-Seidel s’écrit

$$T_{GS} = (I - D^{-1} L)^{-1} D^{-1} U.$$  

**Théorème 3.3.3.** Si $A$ est une matrice carrée à diagonale strictement dominante en lignes alors la méthode de Gauss-Seidel converge.

**Preuve:**

Posons $T = (D - L)^{-1} U$ et montrons que $\|T\|_\infty < 1$ où $\|T\|_\infty = \max_{x \neq 0} \frac{\|Tx\|_\infty}{\|x\|_\infty}$.

Soit $y = Tx = (D - L)^{-1} U x$ on a alors $(D - L) y = U x$ ou encore $D y = Ly + U x$ et $y = D^{-1} Ly + D^{-1} U x$. Considérons l’indice $i_0$ tel que

$$|y_{i_0}| = \max_i |y_i| = \|y\|_\infty = \|Tx\|_\infty.$$

Il vient :

$$y_{i_0} = \sum_{j=1}^{i_0-1} (D^{-1} L)_{i_0 j} y_j + \sum_{j=i_0+1}^n (D^{-1} U)_{i_0 j} x_j.$$  

Par suite

$$|y_{i_0}| = \|y\|_\infty \leq \sum_{j=1}^{i_0-1} \left| \frac{a_{i_0 j}}{a_{i_0 i_0}} \right| \|y\|_\infty + \sum_{j=i_0+1}^n \left| \frac{a_{i_0 j}}{a_{i_0 i_0}} \right| \|x\|_\infty.$$  

57
En regroupant les termes

\[
\left( 1 - \sum_{j=1}^{i_0-1} \frac{|a_{i0j}|}{a_{i0i_0}} \right) \frac{\|y\|_\infty}{\|x\|_\infty} \leq \sum_{j=i_0+1}^{n} \frac{|a_{i0j}|}{a_{i0i_0}}
\]

Par hypothèse, le terme \(1 - \sum_{j=1}^{i_0-1} \frac{|a_{i0j}|}{a_{i0i_0}}\) est strictement positif d'où on en tire :

\[
\frac{\|Tx\|_\infty}{\|x\|_\infty} = \frac{\|y\|_\infty}{\|x\|_\infty} \leq \left( \sum_{j=i_0+1}^{n} \frac{|a_{i0j}|}{a_{i0i_0}} \right) \left( 1 - \sum_{j=1}^{i_0-1} \frac{|a_{i0j}|}{a_{i0i_0}} \right)^{-1},
\]

finalement

\[
\max_{x \neq 0} \frac{\|Tx\|_\infty}{\|x\|_\infty} < 1.
\]

**Remarques 3.3.2.**

1. Un résultat de convergence similaire a lieu si \(A\) est à diagonale dominante en colonnes.

2. Si on se place dans les conditions du théorème 3.3.2, la méthode de Gauss-Seidel est convergente.

### 3.3.3 Méthode de relaxation

Si on considère des matrices \(M\) et \(N\) dépendantes d’un paramètre \(\omega\)
on obtient : \(A = M(\omega) - N(\omega)\).

Prenons \(M(\omega) = \frac{1}{\omega}D - L\) et \(N(\omega) = \frac{1}{\omega}D + U\), en supposant \(M(\omega)\) inversible.

Le schéma itératif qui en résulte est le suivant :

\[
x^{(k+1)} = \left( \frac{1}{\omega}D - L \right)^{-1} \left( \frac{1}{\omega}D + U \right) x^{(k)} + \left( \frac{1}{\omega}D - L \right)^{-1} b
\]

l’équation (3.3.8) peut être remplacée par :

\[
x^{(k+1)} = \left( \frac{1}{\omega}D \right)^{-1} Lx^{(k+1)} + \left( (1 - \omega)I + \frac{1}{\omega}D^{-1}U \right) x^{(k)} + \left( \frac{1}{\omega}D^{-1} \right) b
\]

La matrice de relaxation est donnée par \(T_\omega = \left( \frac{1}{\omega}D - L \right)^{-1} \left( \frac{1}{\omega}D + U \right)\).

**Remarques 3.3.3.**

- Si \(D\) est inversible, \(T_\omega = (I - \omega D^{-1}L)^{-1} \left( (1 - \omega)I + \omega D^{-1}U \right)\).

- Si \(\omega = 1\), on retrouve la méthode de Gauss-Seidel.

- Si \(\omega > 1\), on parle de sur-relaxation.
Si $\omega < 1$, on parle de sous-relaxation.

Ici la condition de convergence $\|T_\omega\| < 1$ dépendra du paramètre $\omega$ et par conséquent, on est amené à chercher tous les $\omega$ pour lesquels il y a convergence et ensuite choisir la valeur optimale $\omega_0$ de telle sorte que la vitesse de convergence soit la meilleure possible.

**Théorème 3.3.4.** Soit $A$ une matrice hermitienne et inversible définie positive ($A = M - N$) telle que $M$ soit inversible, et la matrice $M^* + N$ soit définie positive, alors le schéma (3.1.4) converge si et seulement si la matrice $A$ est définie positive.

**Preuve:**

i) Supposons que $A$ est définie positive

$$A = A^* \implies (M - N)^* = M^* - N^* = M - N.$$ 

D’où $M^* + N = (M^* + M - A) = M + N = (M^* + N)^*$. Par conséquent si $A$ est hermitienne alors $M^* + N$ est aussi hermitienne. Comme $A$ est définie positive, l’application $v \rightarrow (v^* A v)^{1/2}$ définit une norme : $v \rightarrow \|v\|_A = (v^* A v)^{1/2}$. Considérons alors la norme matricielle induite par la norme $\|\cdot\|_A$ on a :

$$\left\|M^{-1}N\right\|_A = \left\|I - M^{-1}A\right\|_A = \sup_{v/\|v\|_A = 1} \left\|(I - M^{-1}A)v\right\|_A.$$ 

En posant $\omega = M^{-1}A v$ il vient que $M \omega = A v$ et on est amené à travailler sur $\|v - \omega\|_A$ avec $\|v\|_A = 1$ on a :

$$\|v - \omega\|_A^2 = (v - \omega)^* A (v - \omega) = \|v\|_A^2 - v^* A \omega - \omega^* A v + \|\omega\|_A^2 = 1 - \omega^* M \omega - \omega^* M^* \omega + \omega^* A \omega = 1 - \omega^* M \omega - \omega^* N \omega = 1 - \omega^* (M^* + N) \omega$$

On a $v \neq 0$ donc $\omega \neq 0$ et $M^* + N$ est définie positive donc $\omega^* (M^* + N) \omega > 0$ par conséquent $\|v - \omega\|_A^2 < 1$ et $\|M^{-1}N\|_A < 1$.

ii) Réciproquement, posons $T = I - M^{-1}A$ et $R = AM^{-1}(M^* + N)M^{-1}A$.

$$\langle Rx, x \rangle = \langle (M + M^* - A) y, y \rangle$$ avec $y = M^{-1}A x$.

Or la matrice $M + M^* - A = M + N$ est définie positive et $A$ est inversible, on a donc $\langle Rx, x \rangle > 0 \ \forall x \neq 0$. 

59
Par suite $R$ est hermitienne définie positive.

En remarquant que $A = R + T^*AT$, on obtient

$$A = R + T^*(R + T^*AT)T = R + T^*RT + T^2(R + T^*AT)T^2$$

$$= \sum_{i=0}^{k-1} T^{*i}RT^i + T^{*k}RT^k.$$

Par hypothèse le schéma (3.1.4) est convergent, donc

$$\rho \left(M^{-1}N\right) < 1,$$

or $M^{-1}N = I - M^{-1}A = T$ donc $\rho(T) < 1$ et on a

$$\lim_{k \to \infty} T^k = \lim_{k \to \infty} T^{*k} = 0.$$

Donc

$$A = \sum_{i=0}^{\infty} T^{*i}RT^i = R + \sum_{i=0}^{\infty} T^{*i}RT^i.$$

Puisque $R$ est définie positive, et que $\langle T^{*k}RT^kx, x \rangle \geq 0, \forall x$, il en résulte que $A$ est définie positive.

**Théorème 3.3.5** (Condition nécessaire de convergence).

*Si $A$ est une matrice hermitienne définie positive alors la méthode de relaxation converge si $\omega \in ]0,2[.*

**Preuve:**

D’après le théorème précédent, si $A$ est hermitienne définie positive et $M^* + N$ définie positive alors la méthode converge. Il suffit donc que $M^* + N$ soit définie positive.

Or $M^* + N = \frac{2 - \omega}{\omega}D$ et par suite $M^* + N$ est définie positive si $\frac{2 - \omega}{\omega} > 0$ c.à.d si $0 < \omega < 2$.

**Théorème 3.3.6** (Kahan).

*Le rayon spectral de la matrice de relaxation vérifie toujours l’inégalité $ho(T_\omega) \geq |\omega - 1|$, $\omega \neq 0$ il s’ensuit que la méthode de relaxation ne peut converger que si $\omega \in ]0,2[.*

**Preuve:**

$$T_\omega = \left(\frac{1}{\omega}D - L\right)^{-1} \left(\frac{1 - \omega}{\omega}D + U\right)$$
Si les valeurs propres de $T_\omega$ sont notées $\lambda_i(\omega)$ on a :

$$\det T_\omega = \prod_{i=1}^{n} \lambda_i(\omega) = \frac{\det \left( \frac{1 - \omega}{\omega} D + U \right)}{\det \left( \frac{1}{\omega} D - L \right)} = (1 - \omega)^n.$$ 

D'où $\rho(T_\omega) \geq \left| \frac{1}{1 - \omega} \right|^n n = |1 - \omega|$.

Pour que la méthode converge, il est nécessaire d’avoir $\rho(T_\omega) < 1$ et par conséquent $|1 - \omega| < 1$ d’où $\omega \in [0, 2]$.

**Remarque 3.3.4.** Le résultat reste valable si $A$ est tridiagonale par blocs (voir Ciarlet(1984) ou Sibony(1986)).

### 3.4 Comparaison des méthodes classiques

#### 3.4.1 Comparaison des méthodes de Jacobi et de Gauss-Seidel

**Théorème 3.4.1.** Soit $A$ une matrice tridiagonale. Alors les méthodes de Jacobi et de Gauss-Seidel convergent ou divergent simultanément; lorsqu’elles convergent, la méthode de Gauss-Seidel est plus rapide que celle de Jacobi, plus précisément, on a $\rho(T_{GS}) = \left( \rho(T_J) \right)^2$ où $T_{GS}$ et $T_J$ sont les matrices de Gauss-Seidel et Jacobi (respectivement).

**Preuve:**

$$T_{GS} = (D - L)^{-1}U \text{ et } T_J = D^{-1}(L + U)$$

On est donc amené à chercher les valeurs propres de ces deux matrices.

1) Soit $\lambda$ une valeur propre de $T_J$. Alors $\lambda$ est racine du polynôme caractéristique $P_J(\lambda)$

$$P_J(\lambda) = \det \left( D^{-1}(L + U) - \lambda I \right) = \det \left( -D^{-1} \right) \det(\lambda D - (L + U)) = K_1 \det(\lambda D - (L + U))$$

2) Soit $\alpha$ une valeur propre de $T_{GS}$. Alors $\alpha$ est racine du polynôme caractéristique $P_{GS}(\alpha)$.

$$P_{GS}(\alpha) = \det \left( (D - L)^{-1}(U) - \alpha I \right) = \det \left( -(D - L)^{-1} \right) \det(\alpha D - \alpha L - U) = K_2 \det(\alpha D - \alpha L - U).$$

61
Comme il s’agit de matrices tridiagonales, on a le résultat suivant (voir exercice (3.7.2))

\[ P_{GS}(\alpha) = K_2 \det \left( \alpha D - \alpha \mu L - \mu^{-1} U \right) \text{ pour tout } \mu \neq 0. \]

En particulier pour \( \mu = \alpha^{-1/2} \) on obtient

\[ P_{GS}(\alpha) = K_2 \alpha^{n/2} \det \left( \alpha^{1/2} D - (L + U) \right). \]

On voit bien donc que \( P_{GS}(\alpha) = K_1 2^{n/2} P_J \left( \alpha^{1/2} \right) \). Par suite, si \( \beta \) est une valeur propre de \( T_J \) alors \( \beta^2 \) est une valeur propre de \( T_{GS} \).

Réciproquement, si \( \beta^2 \) est une valeur propre non nulle de \( T_{GS} \) alors \( \beta \) et \(-\beta\) sont des valeurs propres de \( T_J \). En effet, en prenant \( \mu = -1 \), il vient que \( \det(\beta D - L - U) = \det(\beta D + L + U) = 0 \) donc \( P_J(\beta) = 0 \) et \( P_J(-\beta) = 0 \).

3.4.2 Comparaison des méthodes de Jacobi et de relaxation

**Théorème 3.4.2.** Soit \( A \) une matrice tridiagonale, telle que toutes les valeurs propres de la matrice de Jacobi associée soient réelles. Alors les méthodes de Jacobi et de relaxation, pour \( \omega \in ]0, 2[ \), convergent ou divergent simultanément. Lorsqu’elles convergent, on peut déterminer une valeur optimale \( \omega_0 \) du paramètre \( \omega \) telle que

\[ \omega_0 = \frac{2}{1 + \sqrt{1 - \rho(T_J)^2}}, \quad \rho(T_{\omega_0}) = \inf_{\omega \in [0, 2]} \rho(T_{\omega}) = \omega_0 - 1. \]

Plus précisément, si on considère la fonction qui à \( \omega \) fait correspondre \( \rho(T_{\omega}) \) alors cette fonction a l’allure suivante

![Graphique de la fonction ρ(T_ω)](image)
Remarques 3.4.1. 1. \( \rho(T_{GS}) = \rho(T_j) \) est obtenue comme cas particulier avec \( \omega = 1 \) on sait d’après le théorème précédent que \( \rho(T_{GS}) = (\rho(T_j))^2 \).

2. On voit que pour la valeur \( \omega_0 \) de \( \omega \) on obtient

\[
\rho(T_{\omega_0}) < \rho(T_{GS}) < \rho(T_j)
\]

qui indique qu’avec un choix approprié de \( \omega \), la méthode de relaxation converge plus rapidement que celles de Gauss-Seidel et de Jacobi.

Preuve:
Notons \( \alpha \) une racine de \( P_j \) et notons \( \lambda \) une racine de \( P_\omega \). On obtient alors

\[
P_j(\alpha) = K_1 \det(\alpha D - L - U)\]

et

\[
P_\omega(\lambda) = K_\omega \det \left( \frac{1 - \omega - \lambda}{\omega} D + \lambda L + U \right)
\]

\[
= (-1)^n K_\omega \det \left( \frac{\lambda + \omega - 1}{\omega} D - \lambda L - U \right)
\]

\[
P_\omega(\lambda^2) = \tilde{K}_\omega \det \left( \frac{\lambda^2 + \omega - 1}{\omega} D - \lambda^2 L - U \right)
\]

Soit encore :

\[
P_\omega(\lambda^2) = \tilde{K}_\omega \lambda^n \det \left( \frac{\lambda + (\omega - 1)}{\omega} D - L - U \right) \quad \text{pour tout } \lambda \neq 0
\]

Par identification, on voit que :

\[
P_\omega(\lambda^2) = \tilde{K}_\omega, P_j \left( \frac{\lambda + (\omega - 1)}{\omega} \right)
\]

\[
= \tilde{K}_\omega, P_j \left( \frac{\lambda^2 + \omega - 1}{\lambda \omega} \right)
\]

Il s’ensuit qu’on peut déterminer une relation liant les valeurs propres non nulles de \( T_\omega \) et celles de \( T_j \).

Soit \( \lambda^2 \neq 0 \) une valeur propre de \( T_\omega \) donc \( \pm \frac{\lambda^2 + \omega - 1}{\lambda \omega} \) est une valeur propre de \( T_j \). Inversement, si \( \alpha \) est valeur propre de \( T_j \), on sait que \( -\alpha \) est aussi valeur propre de \( T_j \) et ceci implique que \( \lambda_1^2 \) et \( \lambda_2^2 \) sont valeurs propres de \( T_\omega \).
par :

\[
\lambda_1^2 = \frac{1}{2} \left( \alpha^2 \omega^2 - 2 \omega + 2 \right) - \frac{\alpha \omega}{2} \sqrt{\left( \alpha^2 \omega^2 + 4 - 4 \omega \right)}
\]

\[
\lambda_2^2 = \frac{1}{2} \left( \alpha^2 \omega^2 - 2 \omega + 2 \right) + \frac{\alpha \omega}{2} \sqrt{\left( \alpha^2 \omega^2 + 4 - 4 \omega \right)}
\]

\(\lambda_1\) et \(\lambda_2\) proviennent de la résolution de l’équation du second degré en \(\lambda\), à savoir

\[
\alpha = \frac{\lambda^2 + \omega - 1}{\lambda \omega}
\]

ou encore

\[
\lambda^2 - \alpha \omega \lambda + \omega - 1 = 0 \quad (\ast)
\]

on a

\[
\Delta(\alpha) = \alpha^2 \omega^2 - 4 \omega + 4 = \alpha^2 \omega^2 - 4(\omega - 1) = \alpha^2 \omega^2 + 4(1 - \omega)
\]

1. Si \(\alpha \geq 1\), on est dans le cas où la méthode de Jacobi diverge.

Considérons alors \(\Delta(\alpha)\) comme équation du second degré en \(\omega\) : \(\Delta(\alpha) = \alpha^2 \omega^2 - 4 \omega + 4\), son discriminant est \(\Delta' = 4 - 4\alpha^2 < 0\) d’où \(\Delta(\alpha)\) est toujours du signe de \(\alpha^2\) c’est à dire positif.

\(\Delta(\alpha) \geq 0\) implique que \((\ast)\) admet deux solutions réelles

\[
\lambda_1 = \frac{\alpha \omega - \sqrt{\Delta(\alpha)}}{2} \text{ et } \lambda_2 = \frac{\alpha \omega + \sqrt{\Delta(\alpha)}}{2}
\]

on en tire

\[
\lambda_1^2 = \frac{1}{2} \left( \alpha^2 \omega^2 - 2 \omega + 2 \right) - \frac{\alpha \omega}{2} \sqrt{\left( \alpha^2 \omega^2 + 4 - 4 \omega \right)}
\]

et \(\lambda_2^2 = \frac{1}{2} \left( \alpha^2 \omega^2 - 2 \omega + 2 \right) + \frac{\alpha \omega}{2} \sqrt{\left( \alpha^2 \omega^2 + 4 - 4 \omega \right)}
\]

on a

\[
\lambda_2^2 \geq \frac{1}{2} \left( \omega^2 - 2 \omega + 2 \right) + \frac{\omega}{2} |\omega - 2| \text{ si } \alpha \geq 1
\]

\[
\geq \frac{1}{2} \left( \omega^2 - 2 \omega + 2 + 2 \omega - \omega^2 \right) = 1
\]

Donc si la méthode de Jacobi diverge \((\alpha \geq 1)\) alors on a aussi \(\rho(T_\omega) \geq 1\), par conséquent la méthode de relaxation diverge aussi.

2. Si \(|\alpha| < 1\)

On est amené à étudier

\[
\rho(T_\omega) = \max_{\alpha \in \mathbb{S}(T)} \left\{ \max \left( |\lambda_1^2(\alpha, \omega)|, |\lambda_2^2(\alpha, \omega)| \right) \right\}
\]
pour cela, considérons la fonction $f$ définie par :

$$f : \mathbb{R}_+ \times [0, 2] \rightarrow \mathbb{R}_+
\begin{array}{c}
(\alpha, \omega) \rightarrow \max \left\{ |\lambda_1^2|, |\lambda_2^2| \right\}
\end{array}$$

**Remarque 3.4.2.** on n'a pas besoin de considérer $\alpha \in \mathbb{R}_-$ car $f(\alpha, \omega) = f(-\alpha, \omega)$.

i) D’abord si $\alpha = 0$ alors $f(0, \omega) = |1 - \omega|$.

ii) Si $0 < \alpha < 1$ alors le trinôme $\omega \rightarrow \alpha^2 \omega^2 - 4\omega + 4 = \Delta(\alpha)$ considéré

Comme équation de second degré en $\omega$ admet pour discriminant,

$$\Delta' = 4 - 4\alpha^2 > 0 \text{ d'où deux racines réelles :}
$$

$$\omega_0(\alpha) = \frac{2 - \sqrt{4 - 4\alpha^2}}{\alpha^2} = \frac{4\alpha^2}{2\alpha^2 \left( 1 + \sqrt{1 - \alpha^2} \right)} = \frac{2}{1 + \sqrt{1 - \alpha^2}}$$

$$\omega_1(\alpha) = \frac{2 + \sqrt{4 - 4\alpha^2}}{\alpha^2} = \frac{4\alpha^2}{2\alpha^2 \left( 1 - \sqrt{1 - \alpha^2} \right)} = \frac{2}{1 - \sqrt{1 - \alpha^2}}$$

$\omega_0(\alpha)$ et $\omega_1(\alpha)$ satisfont les inégalités

$$1 < \omega_0(\alpha) = \frac{2}{1 + \sqrt{1 - \alpha^2}} < 2 < \omega_1(\alpha)$$

Donc pour $\alpha$ telle que $0 < \alpha < 1$ et pour $\omega$ compris entre $\omega_0$ et $\omega_1$ (en particulier $\omega_0 < \omega < 2$) $\Delta(\alpha, \omega)$ est négatif. Ce qui veut dire que pour ces valeurs, $\lambda_1^2$ et $\lambda_2^2$ sont des complexes conjugués.

Comme $\lambda_1$ et $\lambda_2$ vérifient $\lambda_1 + \lambda_2 = \alpha \omega$ et $\lambda_1 \lambda_2 = \omega - 1$.

Or $\lambda_1 \lambda_2 = |\lambda_1|^2 = |\lambda_2|^2 = |\lambda_2^2| = |\lambda_1^2| = \omega - 1$ et par suite $f(\alpha, \omega) = \omega - 1$.

On vérifie que $f(\alpha, \omega) = |\lambda_2^2| = \lambda_2^2$ car $\Delta(\alpha) > 0$ et on a deux racines réelles

$\lambda_1$ et $\lambda_2$ et on est amené à étudier l’allure des courbes $f(\alpha, \omega)$ on obtient

$$\frac{\partial f}{\partial \omega} < 0 \text{ pour } 0 < \alpha < 1 \text{ et } 0 < \omega < \omega_0(\alpha) \text{ donc la fonction

$\omega \rightarrow f(\alpha, \omega)$ est décroissante.}

(i) Si $0 < \omega < \omega_0(\alpha)$.

Puisque $0 < \omega < \omega_0(\alpha)$ alors $f(\alpha, \omega) = |\lambda_2^2| = \lambda_2^2$.

$\Delta(\alpha) > 0$ donc on a deux racines réelles

$$\frac{\partial f}{\partial \omega} = \left( \frac{\alpha}{2} + \frac{\omega \alpha^2 - 2}{2\sqrt{\Delta(\alpha)}} \right) 2\lambda_2$$

65
en utilisant $2\lambda_2 = \alpha \omega + \sqrt{\Delta(\alpha)}$ donc $\sqrt{\Delta(\alpha)} = 2\lambda_2 - \alpha \omega$ d’où

$$\frac{\partial f}{\partial \omega} = \left(\frac{\alpha}{2} + \frac{\omega \alpha^2 - 2}{2(2\lambda - \alpha \omega)}\right) 2\lambda_2$$

$$= \left(\frac{2\alpha \lambda - \omega \alpha^2 + \omega \alpha^2 - 2}{2(2\lambda - \alpha \omega)}\right) 2\lambda_2$$

$$= \left(\frac{\alpha \lambda - 1}{2\lambda - \alpha \omega}\right) 2\lambda_2$$

$$= 2\lambda_2 \left(\frac{\lambda_2 \alpha - 1}{2\lambda_2 - \alpha \omega}\right)$$

Donc $\frac{\partial f}{\partial \omega} < 0$ car $\alpha \lambda_2 < \lambda_2 < 1$ pour $0 < \alpha < 1$ en effet

$$\lambda_2^2 = \frac{\alpha^2 \omega^2 - 2\omega + 2}{2} + \frac{\alpha \omega}{2} \left(\alpha^2 \omega^2 - 4\omega + 4\right)^{\frac{1}{2}}$$

$$< \frac{\omega^2 - 2\omega + 2}{2} + \frac{\omega}{2} \left(\omega^2 - 4\omega + 2\right)^{\frac{1}{2}}$$

$$< \frac{\omega^2 - 2\omega + 2}{2} + \frac{\omega}{2} \left(2 - \omega\right)$$

$$< 1$$

Par ailleurs

1- $\frac{\partial f}{\partial \omega}(0) = 0$ donc $\Delta(\alpha)(\omega = 0) = 4$

et $\lambda_2(\omega = 0) = \sqrt{\Delta(\alpha)(\omega = 0)} = \frac{\sqrt{4}}{2} = \frac{1}{2}$ par conséquent

$$\frac{\partial f}{\partial \omega}(\omega = 0) = 2\lambda_2 \left(\lambda_2 + \frac{\omega \lambda^2 - 2}{\sqrt{\Delta(\alpha)(\omega = 0)}}\right)$$

$$= 2\lambda_2(0) \left(\lambda_2(0) - \frac{2}{\sqrt{\Delta(\alpha)(\omega = 0)}}\right)$$

$$= 2(1 - \frac{2}{2})$$

$$= 0$$

2- $\frac{\partial f}{\partial \omega}(\omega_0) = \infty$ car $\omega_0$ est racine de $\Delta(\alpha)$ d’où $\frac{1}{\Delta(\alpha)}$ tend vers l’infini quand $\omega$ tend vers $\omega_0$.

En conclusion, pour $0 < \alpha < 1$ et $0 < \omega < \omega_0$ la courbe de $f$ comme fonction de $\omega$ a l’allure suivante
Enfin, remarquons que la fonction :

\[ \alpha \rightarrow \frac{2}{1 + \sqrt{1 - \alpha^2}} \]

est croissante en fonction de \( \alpha \)
donc le maximum est atteint pour \( \rho(T_f) \)
3.5 Méthodes semi-itératives

Dans la première partie de ce chapitre, nous avons considéré les méthodes itératives du type \( x^{(k+1)} = Tx^{(k)} + C \), impliquant deux itérés successifs \( x^{(k)} \) et \( x^{(k+1)} \).

A présent, nous considérons des méthodes semi-itératives permettant d’exprimer \( y^{(k)} \) comme combinaison algébrique des \( k \) itérés précédents \( x^{(0)}, x^{(1)}, \ldots, x^{(k-1)} \).

Nous obtenons l’expression:

\[
y^{(k)} = \sum_{j=0}^{k} \theta_j(k) x^{(j)}, \quad k \geq 0,
\]

avec \( \sum_{j=0}^{k} \theta_j(k) = 1, \quad k \geq 0 \).

Si \( e^{(j)} = x^{(j)} - x \) est l’erreur de la \( j \)ème itération et \( e^{(j)} = y^{(j)} - x \) est l’erreur de la méthode semi-itérative alors on a:

\[
e^{(k)} = \sum_{j=0}^{k} \theta_j(k) e^{(j)},
\]

et comme \( e^{(k)} = T^k e^{(0)} \), il vient:

\[
e^{(k)} = \left( \sum_{j=0}^{k} \theta_j(k) T^j \right) e^{(0)},
\]

ou encore

\[
e^{(k)} = Q_k(T)e^{(0)}, \quad (3.5.1)
\]

où \( Q_k(x) \) est le polynôme \( Q_k(x) = \sum_{j=0}^{k} \theta_j(k) x^j \).

**Remarques 3.5.1.**

1. Si on prend \( \theta_j(k) = \frac{1}{k+1} \) pour \( j = 0, 1, \ldots, k \), on retrouve \( y^{(k)} \) comme moyenne arithmétique des \( x^{(k)} \).

2. La méthode de Richardson donnée à l’exercice 3.7.6 peut être considérée comme un cas particulier des méthodes semi-itératives.

**Théorème 3.5.1.** Si les coefficients des polynômes \( Q_k(x) \) sont réels et si la matrice \( T \) est hermitienne et ses valeurs propres vérifient \(-1 \leq a \leq \lambda_1 \leq \cdots \leq \lambda_n \leq b \leq 1\), alors

\[
\|Q_k(T)\|_2 = \max_{\lambda_i \in \text{Sp}(T)} |Q_k(\lambda_i)| \leq \max_{a \leq \lambda \leq b} |Q_k(\lambda)|.
\]

68
En revenant à l’équation (3.5.1), nous sommes donc amenés au problème de minimisation suivant:

$$\min_{Q_k(1)=1} |Q_k(x)|.$$ 

Ceci conduit au problème bien connu de minmax de Chebyshev:

$$\min_{Q_k(1)=1} \max_{-1 \leq a \leq x \leq b \leq 1} |Q_k(x)|,$$

dont la solution est donnée par les polynômes de Chebyshev (voir Varga(2000), Golub(1989))

$$C_k(x) = \begin{cases} 
\cos(k \cos^{-1}(x)), & -1 \leq x \leq 1, \quad k \geq 0 \\
\cosh(k \cosh^{-1}(x)), & x \geq 1, \quad k \geq 0
\end{cases}$$

### 3.6 Décomposition des matrices positives

#### Théorème 3.6.1.

Soit $B$ une matrice carrée; si $B \geq 0$ alors les assertions suivantes sont équivalentes:

i) $\beta > \rho(B)$.

ii) $\beta I - B$ est inversible et on a:

$$(\beta I - B)^{-1} \succeq 0$$

**Preuve:**

Supposons que $\beta > \rho(B)$.

En posant $M = (\beta I - B) = \beta(I - B_\beta)$ où $B_\beta = \frac{1}{\beta} B$, il est évident que $\rho(B_\beta) < 1$ et il découle du théorème 1.4.1 que $B$ est inversible et $(\beta I - B)^{-1} \succeq 0$.

Réciproquement, en supposant ii) et en utilisant le théorème 1.4.1 avec $B \geq 0$, si $v \succeq 0$ est un vecteur propre de $B$ associé à la valeur propre $\rho(B)$ alors $v$ est aussi vecteur propre de $(\beta I - B)$ associé à la valeur propre $(\beta - \rho(B))$.

De l’inversibilité de $(\beta I - B)$ il s’ensuit que $(\beta - \rho(B)) \neq 0$ et par suite

$$(\beta I - B)^{-1} v = \frac{1}{(\beta - \rho(B))} v.$$ 

Enfin, $v$ étant un vecteur propre $\succeq 0$ (non identiquement nul) et $(\beta I - B)^{-1} \succeq 0$ entraîne que $(\beta - \rho(B)) > 0$.

#### Théorème 3.6.2.

Soit $A$ une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $D = (a_{ii})$, on a équivalence entre:
i) $A$ est inversible et $A^{-1} \geq 0$

et

ii) $D > 0$, et en posant $B = I - D^{-1}A$ on a $B \geq 0$ et $B$ est convergente.

**Preuve:**
Supposons que $A$ est inversible et $A^{-1} = (\alpha_{ij})$ avec $\alpha_{ij} \geq 0 \ \forall \ i, j = 1, n$, en explicitant $A^{-1}A = I$, il vient

$$\alpha_{ii}a_{ii} + \sum_{j \neq i} \alpha_{ij}a_{ji} = 1 \ \text{pour tout} \ i = 1, \cdots, n$$

et par suite $\alpha_{ii}a_{ii} = 1 - \sum_{j \neq i} \alpha_{ij}a_{ji} \geq 1 \ \forall \ i = 1, \cdots, n$ du fait que $a_{ij} \leq 0$ et $\alpha_{ij} \geq 0$

pour tout $i \neq j$.

On a donc $a_{ii} > 0 \ \forall \ i = 1, \cdots, n$ et la matrice $D$ est inversible avec $D^{-1} > 0$, d’où $B = I - D^{-1}A \geq 0$ et $I - B = D^{-1}A$ est inversible comme produit de deux matrices inversibles.

Enfin, la convergence de $B$ ($\rho(B) < 1$) s’obtient en remarquant que $(I - B)^{-1} = A^{-1}D \geq 0$ puis en appliquant le théorème 3.6.1 avec $\beta = 1$.

Réciproquement, toujours d’après le théorème 3.6.1, on a $I - B$ inversible et $(I - B)^{-1} \geq 0$ ce qui entraîne que $A^{-1}D \geq 0$ et $A^{-1} \geq 0$.

**Autres résultats**

Avec des hypothèses supplémentaires, notamment d’irréductibilité, on obtient les résultats suivants (voir Varga [169]).

**Théorème 3.6.3.** Si $A \geq 0$ alors les assertions suivantes sont équivalentes:

i) $\alpha > \rho(A)$ et $A$ est irréductible.

ii) $\alpha I - A$ est inversible et on a:

$$(\alpha I - A)^{-1} > 0$$

**Théorème 3.6.4.** Soit $A$ une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $D = (a_{ii})$, on a équivalence entre:

i) $A$ est inversible et $A^{-1} > 0$ et

ii) $D > 0$ et en posant $B = I - D^{-1}A$ on a $B \geq 0$ et $B$ est irréductible et convergente.

**Théorème 3.6.5.** Soit $A$ une matrice réelle irréductible à diagonale dominante avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $a_{ii} > 0 \ \forall i = 1, \cdots, n$ alors $A^{-1} > 0$.

**Théorème 3.6.6.** Soit $A$ une matrice réelle symétrique inversible et irréductible avec $a_{ij} \leq 0$ pour tout $i \neq j$, alors $A^{-1} > 0$ si et seulement si $A$ est définie positive.
3.6.1 Décomposition régulière des matrices

Définition 3.6.1. Soient $A, M$ et $N$ des matrices carrées d’ordre $n$. Une décomposition $A = M - N$ est dite régulière si: $M$ est inversible avec $M^{-1} \geq 0$ et $N \geq 0$.
La décomposition est dite régulière faible si: $M$ est inversible avec $M^{-1} \geq 0$ et $M^{-1}N \geq 0$.

Théorème 3.6.7. Soit $A = M - N$ une décomposition régulière de $A$. Alors $A$ est inversible avec $A^{-1} \geq 0$ si et seulement si $\rho(M^{-1}N) < 1$ où
\[ \rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1 \]
et par conséquent, si $A$ est inversible avec $A^{-1} \geq 0$ alors la méthode $x^{(k+1)} = M^{-1}Nx^{(k)} + c$ est convergente pour n’importe quel choix initial.

Preuve:
Supposons $A$ monotone.
En posant $B = A^{-1}N$, on peut aisément voir que
- $M^{-1}N = (I + B)^{-1}B$,
- si $M^{-1}Nv = \lambda v$ alors $Bv = \mu v$ avec $\mu = \frac{\lambda}{1 - \lambda}$ et $\lambda = \frac{\mu}{1 + \mu}$,
- $M^{-1}N \geq 0$ et $B \geq 0$,
- la fonction $\mu \rightarrow \frac{\mu}{1 + \mu}$ est strictement croissante pour $\mu \geq 0$ d’où $\rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1$ et par conséquent la méthode converge.

Réciproquement, si $\rho(M^{-1}N) < 1$ alors $(I - M^{-1}N)$ est inversible et on a successivement
\[ M^{-1}A = I - M^{-1}N \text{ et } A^{-1} = (I - M^{-1}N)^{-1}M^{-1}, \]
comme $M^{-1} \geq 0$ et que le théorème 1.4.1 donne $(I - M^{-1}N)^{-1} \geq 0$ on en déduit que $A^{-1} \geq 0$.

Théorème 3.6.8. Soient $A = M_1 - N_1 = M_2 - N_2$ deux décompositions régulières de $A$ avec $A^{-1} \geq 0$.
Si $0 \leq N_1 \leq N_2$ alors $0 \leq \rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) \leq 1$.
Si de plus $A^{-1} > 0$ et si $0 < N_1 < N_2$ alors $0 < \rho(M_1^{-1}N_1) < \rho(M_2^{-1}N_2) < 1$.

Preuve:
On a $A^{-1}N_1 \leq A^{-1}N_2$ et par suite $\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) \leq 1$. 

71
3.7 Comparaison des méthodes classiques dans le cas des matrices positives

Soit la décomposition $A = D - E - F$ où $D = (a_{ii})_{i=1,...,n}$ et $E$ et $F$ sont respectivement triangulaire inférieure et supérieure. En supposant $D$ et $(D - E)$ inversibles et en posant $L = D^{-1}E$ et $U = D^{-1}F$, on retrouve les matrices de Jacobi $T_J = L + U$ et de Gauss-Seidel $T_{GS} = (I - L)^{-1}U$ qu’on peut comparer dans certains cas particuliers.

**Théorème 3.7.1** (Stein & Rosenberg).

Soit $T_J = L + U$ la matrice de Jacobi supposée positive avec diagonale nulle et $T_{GS} = (I - L)^{-1}U$ la matrice de Gauss-Seidel. Alors une et une seule des relations suivantes a lieu

1. $\rho(T_J) = \rho(T_{GS}) = 0$,
2. $0 < \rho(T_{GS}) < \rho(T_J) < 1$,
3. $\rho(T_J) = \rho(T_{GS}) = 1$,
4. $1 < \rho(T_J) < \rho(T_{GS})$.

**Corollaire 3.7.1.** Si la matrice de Jacobi est positive avec $\rho(T_J) < 1$, alors

$R(T_J) < R(T_{GS})$.

**Remarques 3.7.1.**
1. Le théorème de Stein & Rosenberg affirme donc que les matrices de Jacobi et de Gauss-Seidel convergent ou divergent simultanément. Lorsqu’elles convergent, la matrice de Gauss-Seidel converge asymptotiquement plus rapidement.

2. En prenant $L + U$ positive et en posant

$$M_1 = I; N_1 = L + U,$$
$$M_2 = I - L; N_2 = U.$$

Le théorème 3.6.8 permet d’obtenir directement le point 2 du théorème de Stein & Rosenberg.

**Théorème 3.7.2.** Soit $A = I - B$ où $B = L + U$ est positive, irreductible et convergente avec $L$ et $U$, respectivement, triangulaire strictement inférieure et triangulaire strictement supérieure alors la matrice de relaxation définie par

$$M_\omega = (I - \omega L)^{-1}(\omega U + (1 - \omega)I)$$
est convergente pour \(0 < \omega \leq 1\). De plus, si \(0 < \omega_1 < \omega_2 \leq 1\), alors

\[0 < \rho(M_{\omega_2}) < \rho(M_{\omega_1}) < 1\]

et par conséquent \(R(M_{\omega_1}) < R(M_{\omega_2})\).

### 3.8 Complément bibliographique

Selon le type de matrice à savoir, positive, définie positive, symétrique, hermitienne, à diagonale dominante, tridiagonale ou autre, plusieurs auteurs se sont intéressés aux méthodes itératives, à leur convergence et leurs comparaisons. Là encore, il serait vain de vouloir faire la review de tous les papiers traitant ces sujets. À titre indicatif, nous donnons une chronologie de quelques travaux tout en renvoyant à Ostrowski(1956), Collatz(1950), Bonnet & Meurant(1979), Stoer & Bulirsch(1983), Golub(1989),, Berman & Plemmons(1994) et Varga(2000) pour plus de détails.


3.9 Exercices

**Exercice 3.9.1.** Soit $B$ une approximation de l’inverse $A^{-1}$ d’une matrice carrée $A$ et $\hat{x}$ une solution approchée du système $Ax = b$.

En posant $R = I - BA$ et $r = b - A\hat{x}$, montrer que si $\|R\| < 1$ alors on a:

i) $\|A^{-1}\| \leq \frac{\|B\|}{1 - \|R\|}$

ii) $\|A^{-1} - B\| \leq \frac{\|B\|\|R\|}{1 - \|R\|}$

iii) $\|x - \hat{x}\| \leq \frac{\|B\|\|r\|}{1 - \|R\|}$

**Exercice 3.9.2.** Soit $\theta$ un scalaire non nul et $A(\theta)$ la matrice tridiagonale suivante:

$$A(\theta) = \begin{pmatrix}
a_1 & b_1\theta^{-1} & & \\
c_2\theta & a_2 & b_2\theta^{-1} & \\
& c_3\theta & \ddots & \ddots \\
& & \ddots & b_{n-1}\theta^{-1} \\
& & & c_n\theta & a_n
\end{pmatrix}$$

montrer que $\det A(\theta) = \det A(1)$.

**Exercice 3.9.3.** Soit $E$ un espace vectoriel normé complet et $A$ un opérateur de $E \rightarrow E$ vérifiant:

$$\|Ax - Ay\| \leq \alpha\|x - y\| \quad \forall x \in E, \forall y \in E;$$

avec $0 \leq \alpha < 1$.

1. Soit $x_0 \in E$ donné, montrer que l’itération $x_{n+1} = Ax_n$ définit une suite de Cauchy dans $E$. On notera $x^*$ la limite de cette suite.

2. On considère une itération approchée donnée par: $y_{n+1} = Ay_n + \sigma_n, \ y_0 \in E$ donné et $\|\sigma_n\| \leq \varepsilon$ pour tout $n \geq 0$, prouver que

$$\|x^* - y_n\| \leq \frac{\alpha}{1 - \alpha}\|y_n - y_{n-1}\| + \frac{\varepsilon}{1 - \alpha}$$

3. Soit $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $M = \frac{1}{3} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ et $b(x) = \frac{1}{4} \begin{pmatrix} \cos x_1 \\ \sin x_2 \end{pmatrix}$
a) Montrer que l’équation $x = Mx + b(x)$ admet une solution unique $x^*$

b) Montrer que si on utilise l’itération $x_{n+1} = Mx_n + b(x_n)$ alors on obtient

$$\|x^* - x_{n+1}\|_\infty \leq k\|x_{n+1} - x_n\|_\infty,$$

où $k$ est une constante à déterminer.

c) Comment peut-on améliorer la vitesse de convergence de l’itération proposée ci-dessus?

Exercice 3.9.4. On considère les deux matrices suivantes:

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

Chercher les matrices de Jacobi et de Gauss-Seidel associées aux matrices $A$ et $B$ et comparer les rayons spectraux.

Exercice 3.9.5. Soit $A$ une matrice symétrique définie positive et $T$ la matrice définie par: $T = 2D^{-1} - D^{-1}AD^{-1}$. On suppose que $2D - A$ est définie positive.

1. Montrer que la méthode de Jacobi appliquée au système $Ax = b$ converge .

2. Soit la méthode itérative suivante:

$$\begin{cases} x^{(0)} \text{ donné} \\ x^{(n+1)} = x^{(n)} + T \left( b - Ax^{(n)} \right) \end{cases}$$

Montrer que cette méthode converge et comparer sa vitesse de convergence à celle de la la méthode de Jacobi.

Exercice 3.9.6. Soit $A$ une matrice carrée d’ordre $n$ à coefficients dans $\mathbb{C}^n$. Pour chercher la solution du système $Ax = b$ on considère le schéma itératif suivant:

$$\begin{cases} x^{(0)} \in \mathbb{C}^n \text{ donné} \\ x^{(n+1)} = (I - \alpha A)x^{(n)} + \alpha b \end{cases}$$

avec $\alpha > 0$.

I- On suppose que $A$ est à diagonale strictement dominante en lignes.

Montrer que si $0 < \alpha \leq \frac{1}{\max_{i,j} |a_{ij}|}$ alors (*) converge.

II- On suppose que $A$ est hermitienne définie positive dont les valeurs propres sont rangées par ordre décroissant: $0 \leq \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$. 

75
1. Quelle condition doit vérifier $\alpha$ pour que la méthode converge?

2. Déterminer la valeur optimale de $\alpha$.

**Exercice 3.9.7.** Soit $A(h)$ la matrice tridiagonale donnée par:

$$
A(h) = \frac{1}{h^2} \begin{pmatrix}
2 + h^2 a_1 & -1 & & & \\
-1 & 2 + h^2 a_2 & -1 & & \\
& -1 & \ddots & \ddots & \\
& & \ddots & -1 & \\
& & & -1 & 2 + h^2 a_n
\end{pmatrix}
$$

avec $a_i > 0, i = 1, \cdots, n$.

1. Montrer que $A(h)$ est définie positive

2. Donner un algorithme de décomposition $A = LU$ où $A$ est donnée par:

$$
A = \begin{pmatrix}
1 & d_2 & & & \\
c_1 & a_2 & d_3 & & \\
& c_2 & \ddots & \ddots & \\
& & \ddots & d_n & \\
& & & c_{n-1} & a_n
\end{pmatrix}
\begin{pmatrix}
1 & l_1 & 1 & & \\
l_1 & l_2 & & & \\
& l_{n-1} & 1 & & \\
& & & \ddots & \\
& & & & u_n
\end{pmatrix}
\begin{pmatrix}
u_1 & u_2 & d_3 & & \\
d_2 & u & & & \\
& & \ddots & \ddots & \\
& & & u & \\
& & & & u
\end{pmatrix}
$$

en posant: $c_{n+1} = d_1 = l_1 = 0$ et en supposant que $|a_i| > |c_{i+1}| + |d_i|$

Montrer que: $|l_i| < 1$ et $|u_i| > |c_{i+1}|$

**Exercice 3.9.8.** Soit $A$ la matrice donnée par $A = \begin{pmatrix}a & b & b \\ b & a & b \\ b & b & a\end{pmatrix}$ avec $a > b > 0$

1. Ecrire les matrices de Jacobi et de relaxation associées à la matrice $A$.

2. Donner une condition sur $a$ et $b$ pour que la méthode de Jacobi soit convergente.

3. On suppose que $a > 2b$, montrer que $A$ est inversible que le conditionnement de $A$ pour la norme $\| \|_\infty$ vérifie $C(A) \leq \frac{\alpha a + \beta b}{\alpha a - \beta b}$ où $\alpha$ et $\beta$ sont des constantes à déterminer.

**Exercice 3.9.9.** Soit $A$ une matrice carrée hermitienne et définie positive, pour résoudre le système $Ax = b$, on pose $A = D - E - F, L = D^{-1}E$ et $U = D^{-1}F$ et on considère le procédé itératif $B(\omega)x^{(k+1)} = (B(\omega) - A)x^{(k)} + b$ avec $B(\omega) = \frac{1}{\omega}D(I - \omega L)$.  

76
1. Écrire l’itération sous la forme $x^{(k+1)} = T(\omega)x^{(k)} + c \ (*)$ et montrer que $T(\omega) = (I - \omega L)^{-1}((1 - \omega)I + \omega U)$.

2. Montrer que la matrice $B(\omega) + B^H(\omega) - A$ est définie positive pour $0 < \omega < 2$.

3. Montrer que si $\lambda$ est valeur propre de $Q(\omega) = A^{-1}(2B(\omega) - A)$ alors $Re\lambda > 0$.

4. Vérifier que $Q(\omega) + I$ est inversible et que $(Q(\omega) - I)(Q(\omega) + I)^{-1} = T(\omega)$.

5. Trouver une relation entre les valeurs propres $\mu$ de $T(\omega)$ et les $\lambda$.

6. En écrivant $\mu$ en fonction de $\lambda$, prouver que le carré du module de $\mu$ peut s’écrire $|\mu|^2 = |\lambda|^2 + \delta - 2Re\lambda$ où $\delta$ est un nombre strictement positif à exprimer.

7. En déduire que l’itération $\ (*)$ converge pour $0 < \omega < 2$.

**Exercice 3.9.10.** Examen d’Analyse Numérique Faculté des sciences Oujda

**Exercice 3.9.1.** Soit $A = (a_{ij})$ une matrice carrée inversible dont les éléments diagonaux sont non nuls. $A$ est écrite sous la forme $A = D - L - U$ où $D$ est une matrice diagonale et $L$ (respectivement $U$) est triangulaire inférieure (respectivement supérieure). Pour résoudre le système $Ax = b$, on utilise la méthode itétabte suivante:

$$a_{ii}x_i^{(k+1)} = a_{ii}x_i^{(k)} - \theta \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - (1 - \theta)\omega \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} + (1 - \omega)\theta \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \omega \sum_{j=i}^{n} a_{ij}x_j^{(k)} + \omega b_i$$

où $\theta$ et $\omega$ sont des réels fixés ($\omega$ non nul) et $k = 0, 1, \ldots$

1. (a) Montrer que la méthode proposée peut s’écrire sous la forme matricielle:

$$x^{(k+1)} = M(\theta, \omega)x^{(k)} + c(\theta, \omega) \ (*)$$

(b) Pour quelles valeurs de $\theta$ et $\omega$ obtient-on les méthodes de Jacobi et Gauss-Seidel ?

2. On prend $\theta = 1$

(a) Trouver une relation entre les valeurs propres de $M(\theta, \omega)$ et celles de la matrice de Gauss-Seidel associée à $A$.  

77
(b) Peut-on comparer la vitesse de convergence de la méthode $(\ast)$ à celle de Gauss-Seidel ?

3. On prend $U = L^t$ et $\theta = 0$

(a) Trouver une relation entre les valeurs propres de $M(\theta, \omega)$ et celles de la matrice de Jacobi associée à $A$.

(b) En supposant que la méthode de Jacobi converge, montrer que la méthode $(\ast)$ avec $\theta = 0$ converge pour $0 < \omega < \frac{2}{1 - \mu_1}$ où $\mu_1$ est la plus petite valeur propre de la matrice de Jacobi associée à $A$.

4. En supposant que $A$ est une matrice tridiagonale, trouver une relation entre les valeurs propres $M(\theta, \omega)$ et celles de la matrice de Jacobi associée à $A$.

5. En supposant que $A$ est définie positive et symétrique peut-on comparer les vitesses de convergences des méthodes de Jacobi, Gauss-Seidel, $M(0, \omega)$ et $M(1, \omega)$.

**Exercice 3.9.11.** $\|\cdot\|$ une norme matricielle subordonnée.

1. On considère le système linéaire: $(I + E)x = b$ et le système perturbé

   $$(I + E + F)(x + \delta x) = b$$

   où $I$ est la matrice carrée identité d’ordre $n$, $E$ et $F$ deux matrices carrées d’ordre $n$; $b$ et $x$ sont des vecteurs à $n$ composantes. On suppose que

   $\|E\| = \frac{1}{2}$ et $\|F\| = \varepsilon < \frac{1}{2}$. Montrer que $\|\delta x\| \leq \frac{a + b\varepsilon}{c + d\varepsilon} \|b\|$ où $a, b, c$ et $d$ sont des constantes à déterminer.

2. Soit $A$ une matrice diagonalisable admettant $\lambda_1, \lambda_2, \cdots, \lambda_n$ pour valeurs propres, Montrer que si $\mu$ est une valeur propre de la matrice perturbée $A + \delta A$ alors $\min_{i=1}^{i=n} |\lambda_i - \mu| \leq C(P) \|\delta A\|$ où $C(P)$ désigne le conditionnement de la matrice $P$ telle que $P^{-1}AP = \text{diag}(\lambda_i)$.

**Exercice 3.9.12.** Soit $A = (a_{ij})$ une matrice carrée inversible dont les éléments diagonaux sont non nuls. $A$ est écrite sous la forme $A = D - L - U$ où $D$ est une matrice diagonale et $L$ (respectivement $U$) est triangulaire inférieure (respectivement supérieure). Pour résoudre le système $Ax = b$, on utilise la méthode itérative suivante: $a_{ii}x_i^{(k+1)} = a_{ii}x_i^{(k)} + \omega \left( b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)} \right) + r \sum_{j=1}^{i-1} a_{ij} \left( x_j^{(k)} - x_j^{(k+1)} \right)$, où $r$ et $\omega$ sont des réels fixés ($\omega$ non nul) et $k = 0, 1, \cdots$.
1. Montrer que la méthode proposée peut s’écrire sous la forme matricielle:
\[ x^{(k+1)} = M(r, \omega) x^{(k)} + c \]
avec: \( M(r, \omega) = (D - rL)^{-1}(aD + bL + eU) \)
où \( a, b \) sont des réels qu’on exprimera en fonction de \( r \) et/ou de \( \omega \).

2. Vérifier que cette méthode permet d’obtenir les méthodes de Jacobi, Gauss-Seidel et de relaxation pour des choix appropriés de \( r \) et \( \omega \).

3. Montrer que les valeurs propres de \( M(r, \omega) \) sont les racines de l’équation:
\[ \det(\alpha D - \beta L - \omega U) \]
avec \( \alpha = \lambda + \omega - 1 \) et \( \beta = (\lambda - 1)r + \omega \).

4. En supposant que \( A \) est une matrice tridiagonale, montrer que les valeurs propres \( \mu \) de \( M(0,1) \) sont liées aux valeurs propres \( \lambda \) de la matrice générale \( M(r, \omega) \) par la relation \( (\lambda + \omega - 1)^2 = \omega \mu^2((\lambda - 1)r + \omega) \).

5. Comparer la vitesse de convergence des méthodes classiques en discutant selon les valeurs des paramètres.

**Exercice 3.9.13.**

1. Soit \( H \) une matrice hermitienne définie positive.
   
a- Montrer que pour tout \( r > 0 \), la matrice \( rI + H \) est inversible.
   
b- Montrer que la matrice \( (rI - H)(rI + H)^{-1} \) est hermitienne et en déduire que:
     \[ \| (rI - H)(rI + H)^{-1} \|_2 = \max \left| \frac{r - \lambda_j}{r + \lambda_j} \right| \]    
     où les \( \lambda_j \) sont les valeurs propres de \( H \).

2. Soient \( H_1 \) et \( H_2 \) deux matrices Hermitiennes définies positives.
   
   Etant donné un vecteur initial \( x^{(0)} \) on définit la suite des vecteurs \( x^{(k)} \) par:
   
   \[
   \begin{pmatrix}
   (rI + H_1)x^{(k + \frac{1}{2})} = (rI - H_2)x^{(k)} + b \\
   (rI + H_2)x^{(k+1)} = (rI - H_1)x^{(k + \frac{1}{2})} + b
   \end{pmatrix}
   \]

   où \( b \) est un vecteur et \( r \) un scalaire strictement positif.

   a- Ecrire le vecteur \( x^{(k+1)} \) sous la forme: \( x^{(k+1)} = Tx^{(k)} + c \)
   
   b- Montrer que \( \rho(T) < 1 \) (on pourra considérer la matrice \( T = (rI - H_2)(rI + H_2)^{-1} \))
   
   c- Montrer que la suite \( \left( x^{(k)} \right) \) converge vers \( x^* \), solution d’un système linéaire que l’on déterminera.
   
   c- On suppose que les valeurs propres de \( H_1 \) et de \( H_2 \) sont dans l’intervalle \([a, b], a > 0\)
   
   trouver la valeur de \( r \) qui rend minimum la quantité:
   
   \[ \| (rI - H_1)(rI + H_1)^{-1} \|_2 \| (rI - H_2)(rI + H_2)^{-1} \|_2 \]
(On pourra montrer que \( \min_{r \geq 0} \max_{0 < a \leq x \leq b} \frac{|r - x|}{r + x}^2 = \left( \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \right)^2 \) et qu’il est atteint pour \( r = \sqrt{ab} \))

**Exercice 3.9.14.** Soit \( A \) une matrice carrée autoadjointe. On suppose que \( A \) est symétrique.

1. En décomposant \( A \) sous la forme \( A = M - N \), établir les relations suivantes:

   (a) \( AM^{-1}N = N - NM^{-1}N \)

   (b) \( (M^{-1}N)\top AM^{-1}N = AM^{-1}N - (I - (M^{-1}N)\top)N(I - M^{-1}N) \)

   (c) \( A = (I - (M^{-1}N)\top)M\top \)

   (d) \( A - (M^{-1}N)\top A(M^{-1}N) = (I - M^{-1}N)\top(M\top + N)(I - M^{-1}N) \)

   (e) En déduire que si \( x \) est un vecteur propre de \( M^{-1}N \) associé à la valeur propre \( \lambda \) alors

   \[
   \left(1 - |\lambda|^2\right) \langle Ax, x \rangle = |1 - \lambda|^2 \left(\langle (M\top + N) x, x \rangle\right)
   \]

   et que si \( |\lambda| < 1 \) alors \( \langle Ax, x \rangle \) et \( \langle (M\top + N) x, x \rangle \) sont de même signe.

2. Théorème d’Ostrowski.

   En supposant \( A \) définie positive et en la décomposant sous la forme \( A = D - L - L\top \)

   (a) En notant \( T \) la matrice de Gauss-Seidel associée à \( A \) et en posant \( T_1 = D^{1/2}TD^{-1/2} \) et \( L_1 = D^{-1/2}LD^{-1/2} \) montrer que \( T_1 = (I - L_1)^{-1}L\top_1 \)

   (b) Montrer que \( T_1 \) et \( T \) ont mêmes valeurs propres.

   (c) Soit \( \lambda \) une valeur propre de \( T_1 \) associée au vecteur propre \( x \) tel que

   \( x\top x = 1 \) en posant: \( x\top L_1 x = a + ib \), montrer que:

   \[
   |\lambda|^2 = \frac{a^2 + b^2}{1 - 2a + a^2 + b^2}.
   \]

   (d) Montrer que \( 1 - 2a > 0 \) et déduire que \( \rho(T) < 1 \).

   (e) Conclure.

**Problème 3.9.1.** \( A \) désigne toujours une matrice carrée d’ordre \( n \).

1. Montrer que si la matrice \( A = M - N \) est singulière alors on ne peut pas avoir \( \rho(M^{-1}N) < 1 \) même si \( M \) est régulière.

2. Soit \( A \) une matrice hermitienne mise sous la forme \( A = M - N \) où \( M \) est inversible. on note \( B = I - (M^{-1}A) \) la matrice de l’itération associée à \( Ax = \) 80
b. On suppose que \( M + M^* - A \) est définie positive.
Montrer que si \( x \) est un vecteur et \( y = Bx \) alors:
\[
\langle x, Ax \rangle - \langle y, Ay \rangle = \langle x - y, (M + M^* - A)(x - y) \rangle.
\]
En déduire que \( \rho(B) < 1 \) si et seulement si \( A \) est définie positive.

3. Soit \( a \in \mathbb{R} \); la matrice \( A \) est définie de la manière suivante:
\[
A = \begin{pmatrix}
1 & a & a \\
a & 1 & a \\
a & a & 1
\end{pmatrix}
\]
(a) Pour quelles valeurs de \( a \) de la matrice \( A \) est-elle définie positive. Peut-on en déduire les valeurs de \( a \) pour lesquelles la méthode de Gauss-Seidel est convergente.
(b) Ecrire la matrice \( J \) de l’itération de Jacobi. Pour quelles valeurs de \( a \) la matrice de Jacobi converge-t-elle.
(c) Ecrire la matrice \( G \) de l’itération de Gauss-Seidel. Calculer \( \rho(G) \). Pour quelles valeurs de \( a \) la matrice de gauss-Seidel converge-t-elle plus vite que la méthode de Jacobi.

4. On donne deux matrices réelles régulières \( A \) et \( B \) et \( a, b \in \mathbb{R}^n \).
(a) On construit les deux itérations suivantes:
\[
\begin{cases}
    x_0, y_0 \in \mathbb{R}^n \\
x_{k+1} = Bx_k + a \\
y_{k+1} = Ay_k + b
\end{cases}
\tag{3.9.1}
\]
Donner une condition nécessaire et suffisante de convergence des deux suites \( x_k \) et \( y_k \).
(b) On pose \( z_k = (x_k, y_k)^\top \). Expliciter les matrices \( C \) et \( c \) telles que \( z_{k+1} = Cz_k + c \) et comparer \( \rho(C) \) et \( \rho(AB) \).
On considère les deux itérations
\[
\begin{cases}
    x_{k+1} = By_k + a \\
y_{k+1} = Ax_{k+1} + b
\end{cases}
\tag{3.9.2}
\]
(c) Donner une condition nécessaire et suffisante de convergence de (3.9.2).
(d) Mettre (3.9.2) sous la forme \( z_{k+1} = Dz_k + d \). Expliciter \( D \) et \( d \) et comparer \( \rho(D) \) et \( \rho(AB) \).
(e) Comparer les taux de convergence des algorithmes (3.9.1) et (3.9.2).
Chapitre 4

Solutions numériques de l’équation \( f(x) = 0 \)

4.1 Cas scalaire

4.1.1 Généralités et définitions

Dans ce chapitre, on s’intéresse aux solutions numériques des équations du type \( f(x) = 0 \) où \( f \) est une fonction de \( \mathbb{R} \) dans \( \mathbb{R} \). On traitera, en particulier, une partie de recherche des zéros d’un polynôme \( P \) de degré \( n \).

Nous supposons que le lecteur est bien familiarisé avec des notions d’analyse élémentaire notamment: théorèmes de Taylor, des Accroissements finis, de Rolle, des valeurs intermédiaires ainsi que les propriétés fondamentales des suites.

Définition 4.1.1. Soit \( (x_n) \) une suite réelle convergeant vers \( x^* \) et \( e_i = x_i - x^* \) l’erreur à la \( i^{\text{ème}} \) itération. On dira que la convergence est d’ordre \( p \) si :

\[
\lim_{i \to \infty} \frac{|e_{i+1}|}{|e_i|^p} = c
\]

\( c \) (constante).

(pour \( p = 1 \) on imposera la condition \( c < 1 \)), \( c \) est appelé constante de l’erreur asymptotique.

\( p = 1 \) : convergence linéaire.
\( p = 2 \) : convergence quadratique.
\( p = 3 \) : convergence cubique.

Remarque 4.1.1. L’ordre de convergence indique la vitesse de convergence de la suite. Plus l’ordre est élevé, plus rapide est la convergence.
4.1.2 Méthodes du point fixe ou d’approximations successives

Supposons que l’équation \( f(x) = 0 \) (1) est transformée en l’équation \( x = \Phi(x) \) (2).

Alors on peut définir un procédé itératif comme suit :

\[
\begin{cases}
x_0 \text{ donné} \\
x_{n+1} = \Phi(x_n) \quad \text{pour } n = 1, 2, \cdots
\end{cases}
\]

Si la suite récurrente \((x_n)\) est convergente et si \(\Phi\) est continue alors on obtient nécessairement \(x^* = \Phi(x^*)\) et par suite \(f(x^*) = 0\). La limite de la suite \((x_n)\) est donc la solution de l’équation (1). En général, il y a plusieurs façons de passer de l’équation (1) à l’équation (2) comme le montre l’exemple suivant

**Exemple 4.1.1.** \( f(x) = x^3 - e^x + 1, \) on peut prendre
\[
x = \Phi_1(x) \quad \text{avec } \Phi_1(x) = x^3 - e^x + x + 1.
\]
\[
x = \Phi_2(x) \quad \text{avec } \Phi_2(x) = (e^x - 1)^{1/3}.
\]
\[
x = \Phi_3(x) \quad \text{avec } \Phi_3(x) = \log(x^3 + 1).
\]

Il est donc important de choisir le procédé itératif de telle manière que la suite obtenue soit convergente et qu’en plus la convergence soit la ”meilleure possible”.

**Théorème 4.1.1** (Convergence globale).

Soit \(\Phi\) une fonction définie et continue sur un intervalle fermé \(I = [a, b]\) vériant :

i) \(\Phi(I) \subset I\).

ii) Il existe une constante \(0 < L < 1\) telle que \(|\Phi(x) - \Phi(y)| \leq L|x - y|\) pour tous \(x, y\) de \(I\).

Alors le procédé itératif \(x_{n+1} = \Phi(x_n)\) converge vers une solution \(x^*\) qui est la solution unique du problème du point fixe \(x_{n+1} = \Phi(x_n)\). Cette convergence a lieu pour n’importe quel point de départ \(x_0 \in I\). De plus on a une majoration de l’erreur :

\[
|x_{n+1} - x_n| \leq L^n|x_1 - x_0|, \quad |e_n| \leq \frac{L^n}{1-L}|x_1 - x_0|.
\]

**Preuve:**

On montre que la suite \((x_n)\) est de cauchy :

\[
|x_{n+1} - x_n| = |\Phi(x_n) - \Phi(x_{n-1})| \leq L|x_n - x_{n-1}| \leq \cdots \leq L^n|x_1 - x_0|
\]

et pour tout \(p > 0\) on obtient :

\[
|x_{n+p} - x_n| \leq (L^n + L^{n+1} + \cdots + L^{n+p})|x_1 - x_0|.
\]
Soit encore 
\[ |x_{n+p} - x_n| \leq L^n \left( \frac{1 - L^p}{1 - L} \right) |x_1 - x_0| \] (4.1.1)
pour \( n \) assez grand \( L^n \) tend vers zéro et la suite \( x_n \) est de Cauchy.
Soit \( x^* \) une limite de \( (x_n) \) alors par la continuité de \( \Phi \) on a \( x^* = \Phi(x^*) \).

Unicité :
Supposons que \( (x_n) \) admet deux limites \( x^* \) et \( y^* \) alors : \( x^* = \Phi(x^*) \) et \( y^* = \Phi(y^*) \) et par suite \( |x^* - y^*| = |\Phi(x^*) - \Phi(y^*)| \leq L|x^* - y^*| \) qui contredit l'hypothèse \( L < 1 \), si on suppose que \( x^* \neq y^* \).

Remarques 4.1.2. 1. En général, il est difficile de connaître la constante \( L \). On préfère la remplacer par une condition plus forte mais plus usuelle qui est : 
\[ |\Phi'(x)| < 1. \]
2. L'inégalité (4.1.1) permet d'estimer à l'avance le nombre d’itérations pour approximer \( x^* \) avec une précision donnée \( \varepsilon \).
3. Pratiquement on arrête le processus itératif dès que la différence, en valeur absolue, de deux approximations successives \( x_n \) et \( x_{n-1} \) ne dépasse pas une certaine précision imposée à l’avance.

Exemple 4.1.2. Soit à chercher les solutions de \( f(x) = 0 \) avec \( f(x) = x - \tan(x) \) dans \( [\pi, \frac{3\pi}{2}] \).

i) On peut prendre \( x = \Phi_1(x) \) avec \( \Phi_1(x) = \tan(x) \) ou \( x = \Phi_2(x) \) avec 
\( \Phi_2(x) = \arctan(x) + \pi. \)

ii) On vérifie facilement que \( |\Phi_1'(x)| > 1 \) et \( |\Phi_2'(x)| < 1. \)

Théorème 4.1.2. Si l’itération \( x_{n+1} = \Phi(x_n) \) produit une suite qui converge linéairement vers \( x^* \) et si \( \Phi \) est de classe \( C^1 \) alors
\[ C = \lim_{i \to \infty} \frac{|e_{i+1}|}{|e_i|} = |\Phi'(x^*)|. \]

Preuve:
Il suffit d’appliquer le théorème des accroissements finis
\[ |e_{i+1}| = |x_{i+1} - x^*| = |\Phi(x_i) - \Phi(x^*)| = |(x_i - x^*) \Phi'(\xi)| \text{ avec } x^* < \xi < x_i, \text{ d’où} \]
\[ \lim_{i \to \infty} \frac{|e_{i+1}|}{|e_i|} = \lim_{i \to \infty} |\Phi'(\xi)| = |\Phi'(x^*)|. \]

Remarque 4.1.3. Le théorème 4.1.1 donne une convergence globale c.à.d pour tout choix \( x_0 \in I \). On peut avoir une condition analogue à celle du théorème 4.1.1 mais qui n’est pas vérifiée sur \( I \) tout entier.
Corollaire 4.1.1 (Convergence locale).

Si $\Phi$ est de classe $C^1$ dans un ouvert contenant $x^*$ et si $|\Phi'(x^*)| < 1$, alors :

il existe $\varepsilon > 0$ tel que l’itération $x_{n+1} = \Phi(x_n)$ converge vers $x^*$ si $x_0$ satisfait :

$|x_0 - x^*| \leq \varepsilon$. En d’autres termes, on obtient convergence pour des choix de $x_0$ assez proche de la solution $x^*$.

Preuve:

Puisque $\Phi'$ est continue dans un voisinage de $x^*$ et que $|\Phi'(x^*)| < 1$ alors pour tout $K$ tel que $|\Phi'(x)| \leq K < 1$, il existe $\varepsilon > 0$ tel que $|\Phi'(x)| \leq K$ pour tout $x \in [x^* - \varepsilon, x^* + \varepsilon]$.

Soit $K$ fixé et $\varepsilon(K)$ donné alors $I = [x^* - \varepsilon, x^* + \varepsilon]$ répond aux conditions du théorème 4.1.1

i) $\Phi(I) \subset I$ : en effet si $x \in I$ alors $|x - x^*| < \varepsilon$ et $\Phi(x) - x^* = \Phi(x) - \Phi(x^*)$

$= \Phi'(\eta)(x - x^*)$, d’où $|\Phi(x) - x^*| = |\Phi'(\eta)(x - x^*)| \leq K|x - x^*| \leq K\varepsilon < \varepsilon$

et la conclusion s’ensuit d’après le théorème 4.1.1.

Exemple 4.1.3. Considérons $f(x) = 0$ avec $f(x) = x^2 - x - 2$ les solutions sont $x^*_1 = 2$ et $x^*_2 = -1$.

On peut prendre $\Phi_1(x) = x^2 - 2$ et $\Phi'_1(x) = 2x$ ou $\Phi_2(x) = \sqrt{x + 2}$ et $\Phi'_2(x) = \frac{1}{2\sqrt{x + 2}}, x > -2$. Si on cherchait à approcher $x^*_2$ par $\Phi_1$, on aurait

$\Phi'_1(x) > 1$ pour $x > 1/2$ donc 2 n’est pas un point d’attraction.

4.1.3 Méthode du promoteur de convergence de Wegstein

La méthode du promoteur de convergence de Wegstein consiste à modifier la méthode du point fixe vue au paragraphe précédent de façon à accélérer (ou forcer) systématiquement sa convergence.

Dans la méthode précédente, le schéma de calcul était le suivant:

$$x_{n+1} = x_n + \Delta x_n$$  \hspace{2cm} (4.1.2)

Wegstein suggère de modifier l’équation (4.1.2) comme suit

$$\begin{cases} x_{n+1} = x_n + \alpha \Delta x_n \\ \Delta x_n = \Phi(x_n) - x_n \end{cases}$$  \hspace{2cm} (4.1.3)

L’introduction du facteur de relaxation $\alpha$ approprié force la convergence dans le cas où la méthode du point fixe divergerait, $\alpha$ est bien donc un promoteur de convergence. Dans le cas où la méthode du point fixe convergerait, l’introduction du facteur $\alpha$ permet d’accélérer la convergence.
Le meilleur choix à pour α qui optimiserait la convergence du processus itératif (4.1.3) serait tel que \( x_{n+1} = x^\star \). Or ce point est a priori inconnu, donc à l’est également.

Toutefois, on peut obtenir un estimé de à de la manière suivante:

La distance entre \( \Phi(x_n) \) et \( x^\star \) est égale à \( \Phi(x_n) - x^\star = (\alpha - 1) \Delta x_n \).

Si \( \theta \) désigne l’angle de l’intersection entre la droite passant par \((x_n, \Phi(x_n))\) et parallèle à l’axe des abscisses et la droite passant par \((x_n, \Phi(x_n))\) et \((x^\star, \Phi(x^\star))\) alors

\[
\tan(\theta) = \frac{(\alpha - 1) \Delta x_n}{\alpha \Delta x_n} = \frac{\alpha - 1}{\alpha} \tag{4.1.4}
\]

d’autre part \( \tan(\theta) = \frac{\Phi(x^\star) - \Phi(x_n)}{x^\star - x_n} \).

Le théorème des accroissements finis donne:

\[
\tan(\theta) = \Phi'(e) \quad \text{où} \quad e \in [x_n, x^\star] \tag{4.1.5}
\]

Des équations (4.1.4) et (4.1.5) on tire

\[
\alpha = \frac{1}{1 - \Phi'(e)}.
\]

La valeur de \( \Phi'(e) \) est inconnue mais on peut l’approcher par

\[
\Phi'(e) = \frac{\Phi(x_n) - \Phi(x_{n-1})}{x_n - x_{n-1}} = \frac{\Phi(x_n) - x_n}{x_n - x_{n-1}}.
\]

La méthode de Wegstein consiste à calculer à partir de \( x_{n-1} \) la valeur de \( x_n \), \( \tan(\theta) \), \( \alpha \) et évaluer \( x_{n+1} \) selon le schéma (4.1.3).

**Remarque 4.1.4.** Si \( \alpha = 1 \), on retrouve la méthode du point fixe.

A présent, on va s’intéresser aux schémas itératifs avec \( \Phi \) du type \( \Phi(x) = x - \frac{f(x)}{h(x)} \), de telle sorte que le schéma itératif associé converge plus rapidement que le schéma des approximations successives.

### 4.1.4 Méthode de Newton-Raphson

Elle consiste à prendre \( h(x) = f'(x) \). Le procédé itératif de Newton est alors donné par:

\[
x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \Phi(x_n)
\]

on voit aisément que \( f(x^\star) = 0 \) à condition que \( f'(x^\star) \neq 0 \) ce qui nous assure au moins une convergence locale (\(|\Phi'(x^\star)| < 1\)).
En effet $x_{n+1} - x^* = x_n - x^* - \frac{f(x_n) - f(x^*)}{f'(x_n)} = \frac{1}{2} (x_n - x^*)^2 \frac{f''(\eta)}{f'(x_n)}$.

donc $\frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2} \frac{|f''(\eta)|}{|f'(x_n)|}$ tend vers $\frac{1}{2} \frac{|f''(x^*)|}{|f'(x^*)|}$ lorsque $n$ tend vers l’infini.

**Exemple 4.1.4.** Soit à calculer la racine carrée d’un réel positif $A$ par le schéma itératif

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{A}{x_n} \right)$$

- $A = 2$, $x_0 = 2$, $x_1 = 1,5$, $x_2 = 1,416666$, $x_3 = 1,414215$, $x_4 = 1,41421356$, $x_5 = 1,41421356$
- $A = 2$, $x_0 = 20$, $x_1 = 10,05$, $x_2 = 5.1245$, $x_3 = 2.7574$, $x_4 = 1.7414$, $x_5 = 1.4449$
- $A = 10$, $x_0 = -10$, $x_1 = -5,1$, $x_2 = -2,7461$, $x_3 = -1.4442$, $x_4 = -1.4142$, $x_5 = -1.4142$

**Interprétation géométrique de la méthode de Newton**

Le schéma itératif $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ revient à écrire l’équation de la tangente de $f$ en $x_n$, soit $y(x) = f'(x_n) (x - x_n) + f(x_n)$ et prendre $x_{n+1}$ tel que $y(x_{n+1}) = 0$.

**Remarques 4.1.5.**

1. D’après le corollaire 4.1.1, la méthode de Newton converge pour un bon choix de $x_0$. En fait, pour $x_0$ pas assez proche de $x^*$, cette méthode diverge ou converge vers un point autre que $x^*$, ceci se produit lorsque $f''$ change de signe au voisinage de $x^*$.

2. On peut généraliser l’algorithme de Newton au cas où la racine est multiple (d’ordre $p$), il suffit de l’appliquer à la fonction $g(x) = (f(x))^{1/p}$.

**Théorème 4.1.3 (Convergence globale).** Soit $f$ une fonction de classe $C^2 [a, b]$, satisfaisant les conditions suivantes:

- $i)$ $f(a)f(b) < 0$,
- $ii)$ $f'(x) \neq 0 \forall x \in [a, b]$. (Monotone)
- $iii)$ $f''(x) \geq 0$ (ou $\leq 0$) (convexité ou concavité)
- $iv)$ $\left| \frac{f(a)}{f'(a)} \right| < b - a$ et $\left| \frac{f(b)}{f'(b)} \right| < b - a$.

Alors la méthode de Newton converge vers l’unique solution $x^*$ de $f(x) = 0$ dans $[a, b]$ pour n’importe quel choix de $x_0 \in [a, b]$.
**Preuve:**
Considérons le cas \( f(a) < 0, f(b) > 0, f'(x) > 0 \) et \( f''(x) \geq 0 \) (les autres cas se traitent de façon similaire).
D’après i) et ii), il existe une solution unique \( x^* \in [a, b] \) qui vérifie:
\[
x_{n+1} - x^* = x_n - x^* + \frac{f(x^*) - f(x_n)}{f'(x_n)} = \frac{1}{2}(x_n - x^*)^2 \frac{f''(\xi)}{f'(x_n)} \quad \text{avec} \quad x_n \leq \xi \leq x^*
\]
Par conséquent, le signe de \( x_{n+1} - x^* \) est celui de \( \frac{f''(\xi)}{f'(x_n)} \).
Si \( f' > 0 \) et \( f'' \geq 0 \) alors \( x_{n+1} \geq x^* \quad \forall n, (x_n) \) est donc minorée par \( x^* \) à partir de \( x_1 \).
Par ailleurs on a deux possibilités de choix de \( x_0 \):
- ou bien \( x_0 > x^* \) et alors \( x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0 \) (puisque \( f(x_0) > f(x^*) = 0 \))
et \( x_n > x^* \) avec \( f(x_n) > f(x^*) = 0 \) \( \forall n \in \mathbb{N} \).
Dans ce cas, la suite \((x_n)\) est décroissante minorée donc elle converge.
- ou bien \( x_0 < x^* \) et alors \( f(x_0) < 0 \) d’où \( x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} > x_0 \),
mais on aussi \( x_1 > x^* \) et \( f(x_1) > 0 \) et par suite \( x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} < x_1 \)
Donc mis à part le point \( x_0 \), on aura encore une suite décroissante minorée donc convergente.

**Remarques 4.1.6.**
1) les conditions i) et ii) assurent que \( f(x) = 0 \) admet une solution unique dans \([a, b]\),
2) les conditions ii) et iii) assurent que \( f' \) est monotone sur \([a, b]\),
3) la condition iv) assure que le point \( x_1 \) restera dans \([a, b]\) si le point de départ \( x_0 \) est \( a \) ou \( b \).

**4.1.5 Méthode de Newton modifiée**
La méthode de Newton est d’ordre 2, mais elle présente un inconvénient dans la mesure où elle nécessite à chaque étape l’estimation de \( f(x_n) \) et de \( f'(x_n) \).
Une modification consiste à utiliser \( f'(x_0) \) durant tout le procédé itératif. On obtient alors \( x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \) qui est une méthode d’ordre 1 avec
\[
C = \left| \frac{f'(x^*) - f(x_0)}{f'(x_0)} \right|.
\]
Cette méthode est particulièrement intéressante dans le cas où \( f' \) ne varie pas trop.

Entre la méthode de Newton qui nécessite le calcul de \( f'(x_n) \) à chaque étape et la méthode de Newton modifiée qui utilise seulement la valeur \( f'(x_0) \), on peut aussi utiliser une méthode qui estime \( f'(x_{\phi(n)}) \) pour des valeurs intermédiaires de \( n \).

En général, on obtient des méthodes de Newton modifiées (encore dites discrètes) en considérant le schéma:

\[
x_{n+1} = x_n - \frac{f(x_n)}{a_n},
\]

où \( a_n \) est une approximation de \( f'(x_n) \).

En prenant \( a_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \) on obtient la méthode de la sécante (voir exercice 4.4.9).

### 4.1.6 Méthodes de dichotomie

En général, la méthode de Newton converge uniquement si le choix de \( x_0 \) est "assez proche" de \( x^* \). Une alternative qui garantit la convergence consiste à utiliser une dichotomie de l’intervalle \([a, b]\) en localisant la solution à chaque itération. Si \( f(a)f(b) < 0 \) alors il existe au moins une solution \( x^* \in [a, b] \) telle que \( f(x^*) = 0 \).

Soit \( c = \frac{a + b}{2} \)

- Si \( f(a)f(c) < 0 \) on prend \( b = c \)
- Si \( f(b)f(c) < 0 \) on prend \( a = c \),

et on reitère le procédé, après \( n \) itérations l’intervalle \([a, b]\) est réduit à \( \frac{a - b}{2^n} \).

### 4.1.7 Regula falsi (fausse position)

Au lieu de prendre automatiquement le milieu de \([a, b]\), on prend le point d’intersection de l’axe des abscisses avec la droite passant par \((a, f(a))\) et \((b, f(b))\) dont l’équation est donnée par

\[
\frac{x-a}{x-b} = \frac{f(x) - f(a)}{f(x) - f(b)}
\]

et coupe l’axe des abscisses au point \((a + f(a)\frac{a-b}{f(a) - f(b)}, 0)\).

- Si \( f(a)f(c) < 0 \) on remplace \( b \) par \( c \)
- Si \( f(b)f(c) < 0 \) on remplace \( a \) par \( c \).

Inconvénient si \( f \) est convexe ou concave.
4.1.8 Méthode $\Delta^2$ d’Aitken (accélération de la convergence)

Soit $x_n$ une suite générée par un procédé itératif, on pose $\Delta x_n = x_{n+1} - x_n$ ($\Delta$ opérateur de différence), la méthode d’Aitken consiste à générer une suite $y_n$ à partir de la suite $x_n$ par la formule:

$$y_n = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$  

On a alors le résultat suivant:

**Théorème 4.1.4.** Supposons qu’il existe un réel $k$ tel que $|k| < 1$ pour lequel la suite $x_n$ vérifie:

$$x_n \neq x^* et x_{n+1} - x^* = (k + \delta_n)(x_n - x^*) avec \lim_{n \to \infty} \delta_n = 0.$$  

alors la suite $y_n$ est bien définie pour $n$ assez grand et on a

$$\lim_{n \to \infty} \frac{y_n - y^*}{x_n - x^*} = 0.$$  

Autrement dit, la suite $y_n$ converge plus rapidement que la suite $x_n$ vers la limite commune $x^*$.

**Remarque 4.1.7.** On a $\lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|} = \lim_{n \to \infty} (k + \delta_n) = k$, donc il y’a convergence linéaire avec $k < 1$, encore dite géométrique.

**Preuve:**

$$x_{n+1} - x_n = (x_{n+1} - x^*) - (x_n - x^*) = e_{n+1} - e_n = (k + \delta_n)e_n - e_n = (k - 1 + \delta_n)e_n$$

$$e_{n+2} = (k + \delta_{n+1})e_{n+1} = (k + \delta_n)(k + \delta_n)e_n$$

$$\Delta^2 x_n = x_{n+2} - 2x_{n+1} + x_n = e_{n+2} - 2e_{n+1} + e_n = (k^2 + k(\delta_{n+1} + \delta_n) + \delta_{n+1}\delta_n)e_n$$

$$= (k - 1)^2 + \mu_n)e_n avec \mu_n = \delta_{n+1}\delta_n - 2k - 2\delta_n + 1, \lim_{n \to \infty} \mu_n = 0.$$  

Par conséquent, pour $n$ assez grand, $\Delta^2 x_n$ est différent de 0 car $k < 1$, $\mu_n$ tend vers 0 et par suite, $y_n$ est bien définie.

Par ailleurs on a

$$(\Delta x_n)^2 = ((k - 1 + \delta_n)^2 e_n^2$$

$$(\Delta x_n)^2 = ((k - 1)^2 + 2\delta_n(k - 1) + \delta_n^2 e_n^2$$

$$(\Delta x_n)^2 = ((k - 1)^2 + \sigma_n^2 e_n^2 avec \sigma_n \to 0$$
d'où

\[ y_n - x^* = x_n - x^* - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = e_n - \frac{((k-1)^2 + \sigma_n^2)e_n^2}{((k-1)^2 + \mu_n^2)e_n} \]

ou encore

\[ y_n - x^* = \left( 1 - \frac{((k-1)^2 + \sigma_n^2)}{((k-1)^2 + \mu_n^2)} \right) e_n \]

i.e.

\[ \frac{y_n - x^*}{x_n - x^*} = 1 - \frac{((k-1)^2 + \sigma_n^2)}{((k-1)^2 + \mu_n^2)} \]

et par passage à la limite on obtient le résultat annoncé \( \lim_{n \to \infty} \frac{y_n - x^*}{x_n - x^*} = 0 \).

### 4.1.9 Méthode de Steffenson (accélération de la convergence)

Considérons une méthode itérative \( x_{n+1} = \varphi(x_n) \) pour approcher la solution \( x^* \) de \( f(x) = 0 \).

Posons

\[
\begin{align*}
y_n & = \varphi(x_n) \\
z_n & = \varphi(y_n) = \varphi \circ \varphi(x_n) .
\end{align*}
\]

La méthode de Steffenson est définie par

\[ x_{n+1} = x_n - \frac{(y_n - x_n)^2}{z_n - 2y_n + x_n} \]

ceci conduit à un procédé itératif avec la nouvelle fonction \( \Psi \), soit \( \Psi(x_n) = x_{n+1} \) avec \( \Psi \) définie par

\[ \Psi(x) = \frac{x \varphi \circ \varphi(x) - (\varphi(x))^2}{\varphi \circ \varphi(x) - 2 \varphi(x) + x} . \]

On obtient alors les deux résultats:

#### Théorème 4.1.5.

i) Si \( \Psi(x^*) = x^* \) alors \( \varphi(x^*) = x^* \),

ii) et si \( \varphi'(x^*) \neq 1 \) alors \( \varphi(x^*) = x^* \) implique que \( \Psi(x^*) = x^* \).

**Preuve:** Voir Exercice 4.4.3.

#### Théorème 4.1.6.

Soit \( \varphi \) une fonction de classe \( C^{p+1} \) définissant un procédé itératif \( x_{n+1} = \varphi(x_n) \) dont la convergence est d’ordre \( p \). Alors la fonction \( \Psi \) donne un procédé itératif dont la convergence est d’ordre \( 2p - 1 \) si \( p > 1 \) et d’ordre au moins égal à \( 2 \) si \( p = 1 \) et \( \varphi'(x^*) \neq 1 \).

**Preuve:** Voir Exercice 4.4.3.
4.2 Cas des polynômes: Solutions numériques des équations algébriques

On s’intéresse à l’équation \( P(x) = 0 \), avec
\[
P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n, \quad a_i \in \mathbb{R} \text{ et } a_0 \neq 0.
\] (4.2.1)

Toutes les méthodes qu’on a vues pour résoudre \( f(x) = 0 \) restent applicables. Cependant, les polynômes ont des propriétés spéciales et nécessitent des fois des méthodes appropriées. Dans le cadre de ce chapitre, on se limitera à quelques résultats, l’étudiant intéressé pourra trouver des compléments dans la littérature qui abonde sur ce sujet.

4.2.1 Schéma de Horner

Chaque fois que nous aurons à évaluer \( P(x) \) en un point \( x = a \), nous utiliserons l’algorithme suivant
\[
\begin{align*}
    b_0 &= a_0 \\
    b_i &= b_{i-1} x + a_i, & i = 1, 2, \cdots, n
\end{align*}
\] (4.2.2)

La valeur du polynôme \( P(x) \) est alors \( P(a) = b_n \). Le calcul se fait donc de la manière suivante :
\[
P(x) = (\cdots (((a_0 x + a_1)x + a_2)x + a_3) \cdots )x + a_n
\] (4.2.3)

L’algorithme permettant d’obtenir \( P(x) \) à l’aide des \( b_i \) est connu sous le nom de schéma de Horner. Ce schéma nécessite \( n \) multiplications et \( n \) additions pour l’évaluation de \( P(x) \) alors que la méthode directe demande \( \frac{n(n+1)}{2} \) multiplications.

Si on écrit \( P(x) \) pour \( x = \alpha \) à l’aide du schéma de Horner alors les quantités \( b_i \) sont les coefficients du polynôme \( P_1(x) \) obtenu par division euclidienne de \( P(x) \) par \( x - \alpha \). Soit \( P(x) = (x - \alpha) P_1(x) + b_n \) et \( P_1(x) = b_0 x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-1} \).

On obtient aussi \( P'(\alpha) = P_1(\alpha) \). Ce qui implique que la valeur de la dérivée \( P'(\alpha) \) peut être calculée à l’aide du schéma de Horner en utilisant les \( b_i \) cette fois-ci comme coefficients, \( P'(\alpha) = (\cdots (b_0 \alpha + b_1) \alpha + b_2) \cdots ) \alpha + b_{n-1} \).

4.2.2 Méthode de Laguerre

La recherche de racines d’un polynôme se construit de la manière suivante: soit \( P(x) \) un polynôme de degré \( n \). Si on obtient une première racine, on peut écrire
\[
P(x) = (x - r_1) Q(x)
\] (4.2.4)
où \( Q(x) \) est un polynôme de degré \( n - 1 \). Ainsi théoriquement, une fois obtenue une première racine, on peut commencer la recherche d’une autre racine par un polynôme de degré strictement inférieur. Successivement, on poursuit cette procédure jusqu’à l’obtention de l’ensemble des \( n \) racines du polynôme \( P(x) \). Rappelons que les polynômes à coefficients complexes se factorisent en un produit de monômes de degré 1. Cette propriété exprime le fait que les polynômes à coefficients complexes ont l’ensemble de leurs racines dans le plan complexe.

\[
P(x) = \prod_{i=1}^{n} (x - r_i)
\]

(4.2.5)

La méthode de Laguerre utilise le fait que les dérivées logarithmiques successives d’un polynôme divergent au voisinage d’une racine. En prenant le logarithme de l’équation (4.2.5) on obtient

\[
\ln(|P(x)|) = \sum_{i=1}^{n} \ln(|x - r_i|)
\]

(4.2.6)

En dérivant l’équation (4.2.6), on obtient

\[
\frac{d \ln(|P(x)|)}{dx} = \sum_{i=1}^{n} \frac{1}{x - r_i}
\]

(4.2.7)

\[
= \frac{P'(x)}{P(x)}
\]

(4.2.8)

En dérivant l’équation (4.2.7), il vient

\[
-\frac{d^2 \ln(|P(x)|)}{dx^2} = \sum_{i=1}^{n} \frac{1}{(x - r_i)^2}
\]

(4.2.9)

\[
= \left( \frac{P'(x)}{P(x)} \right)^2 - \frac{P''(x)}{P(x)}
\]

\[
= H
\]

(4.2.10)

Soit la racine \( r_1 \) à déterminer, on suppose que la valeur de départ \( r_0 \) est située à une distance \( a \) de \( r_1 \) et que l’ensemble des autres racines sont situées à une distance supposée identique et qui vaut \( b \)

\[
x - r_1 = a
\]

(4.2.10)

\[
x - r_i = b, \quad i = 2, \cdots, n
\]

(4.2.11)
en insérant les équations (4.2.10) et (4.2.11) dans les équations (4.2.7) et (4.2.9), on en déduit respectivement les relations suivantes

\[
\frac{1}{a} + \frac{n - 1}{b} = G \\
\frac{1}{a^2} + \frac{n - 1}{b^2} = H
\]  

(4.2.12)

Après élimination de \( b \), la valeur de \( a \) est

\[
a = \frac{n}{G \pm \sqrt{(n - 1)(nH - G^2)}},
\]  

(4.2.13)

Le signe placé devant la racine du dénominateur est choisi tel que le dénominateur soit le plus grand possible, \( x - a \) devient alors la nouvelle valeur de départ et on itère le processus.

4.2.3 Méthode de Bairstow

Soit \( P \) un polynôme de degré \( n \):

\[
P(x) = x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n
\]
on on considère sa division par le polynôme du second degré \( x^2 + bx + c \):

\[
P(x) = (x^2 + bx + c)Q(x) + rx + s
\]

On cherche à déterminer \( b \) et \( c \) pour que \( r(b, c) = s(b, c) = 0 \).

Pour ceci on calcule \( b \) et \( c \) par la méthode de Newton-Raphson: à partir d’une estimation de \( b \) et \( c \) on cherche \( \Delta b \) et \( \Delta c \) tels que

\[
r(b + \Delta b, c + \Delta c) = s(b + \Delta b, c + \Delta c) = 0
\]

soit au premier ordre:

\[
r(b, c) + \frac{\partial r}{\partial b} \Delta b + \frac{\partial r}{\partial c} \Delta c = 0 \\
s(b, c) + \frac{\partial s}{\partial b} \Delta b + \frac{\partial s}{\partial c} \Delta c = 0
\]
on on calcule \( \Delta b \) et \( \Delta c \) en résolvant ce système de deux équations puis on remplace \( b \) par \( b + \Delta b \) et \( c \) par \( c + \Delta c \); la suite des estimations obtenues converge.

En dérivant la relation entre \( P \) et \( Q \) par rapport à \( b \) et \( c \) on calcule les dérivées partielles \( \frac{\partial r}{\partial b}, \frac{\partial r}{\partial c}, \frac{\partial s}{\partial b}, \frac{\partial s}{\partial c} \), qui s’obtiennent par division de \( Q(x) \) et de \( xQ(x) \) par \( x^2 + bx + c \).
4.2.4 Méthode de Maehly

Soit \( P \) un polynôme de degré \( n \):

\[
P(x) = x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n
\]

On calcule ses racines par la méthode de Newton : \( x_0 \) donné on forme

\[
x_{k+1} = x_k - \frac{P(x_k)}{P'(x_k)}
\]

Soit \( r_1 \) la première racine trouvée et \( P_1(x) = \frac{P(x)}{x - r_1} \), on a

\[
P_1'(x) = \frac{P'(x)}{x - r_1} - \frac{P(x)}{(x - r_1)^2}
\]

La méthode appliquée à \( P_1 \) donne la relation de récurrence:

\[
x_{k+1} = x_k - \frac{P(x_k)}{P'(x_k) - \frac{P(x_k)}{x_k - r_1}}
\]

Si on a déjà trouvé \( j \) racines réelles de \( P \): \( r_1, \cdots, r_j \) et si on note

\[
P_j(x) = \frac{P(x)}{(x - r_1) \cdots (x - r_j)}
\]

alors

\[
P_j'(x) = \frac{P'(x)}{(x - r_1) \cdots (x - r_j)} - \frac{P(x)}{(x - r_1) \cdots (x - r_j)} \sum_{i=1}^{j} \frac{1}{x - r_i}
\]

et la méthode de Newton appliquée à \( P_j \) donne la relation de récurrence:

\[
x_{k+1} = x_k - \frac{P(x_k)}{P'(x_k) - \frac{P(x_k)}{\sum_{i=1}^{j} \frac{1}{x_k - r_i}}}
\]

4.2.5 Localisation des racines

**Théorème 4.2.1.** Soit \( A = \max \{|a_1|, |a_2|, \cdots, |a_n| \} \) où les \( a_i \) sont les coefficients du polynôme \( P \) avec \( a_0 \neq 0 \). Alors si \( \alpha \) est une racine de \( P \), \( \alpha \) vérifie :

\[
|\alpha| < R \quad \text{avec} \quad R = 1 + \frac{A}{|a_0|}
\] (4.2.14)
Preuve :
Soit $\alpha$ une racine de $P$.
Si $|\alpha| \leq 1$ alors $\alpha < R$.
Si $|\alpha| > 1$ la formule (4.2.1) entraîne :

$$|P(\alpha)| = |a_0\alpha^n + \cdots + a_n|$$
$$\geq |a_0||\alpha|^n - |a_1||\alpha|^{n-1} - \cdots - |a_n|$$
$$\geq |a_0||\alpha|^n - A \sum_{k=1}^{n} |\alpha|^k$$
$$\geq |a_0||\alpha|^n - A \left( \frac{|\alpha|^n - 1}{|\alpha| - 1} \right)$$
$$\geq \left[ |a_0| - \frac{A}{|\alpha| - 1} \right] |\alpha|^n$$

Il s’ensuit que $|P(\alpha)| > 0$ si $|a_0| - \frac{A}{|\alpha| - 1} \geq 0$, c’est à dire $|P(\alpha)| > 0$ si $|\alpha| \geq R$.
Donc si $|\alpha| \geq R$ on obtient $P(\alpha) \neq 0$ ce qui contredit notre hypothèse. Par conséquent, toute racine $\alpha$ de (4.2.1) vérifie (4.2.14).

Corollaire 4.2.1. Soit $a_n \neq 0$ et $B = \max\{|a_0|, |a_1|, \cdots, |a_{n-1}|\}$. Alors toute racine $\alpha$ de $P$ satisfait :

$$|\alpha| > r \text{ avec } r = \frac{1}{1 + \frac{B}{|a_n|}}.$$  

Preuve :
Il suffit de reprendre la démonstration du théorème 4.2.1 en faisant le changement de variable $\beta = \frac{1}{\alpha}$. En effet

$$P(\alpha) = \alpha^n Q(\beta) \text{ avec } Q(\beta) = a_n \beta^n + \cdots + a_0$$

le théorème 4.2.1 appliqué à $Q$ donne

$$|\beta| < R = 1 + \frac{B}{|a_n|}, \text{ d'où } |\alpha| = \frac{1}{|\beta|} > \frac{1}{1 + \frac{B}{|a_n|}} = r.$$  

Remarque 4.2.1. Le Théorème 4.2.1 et le corollaire 4.2.1 entraînent que, si $a_0$ et $a_n$ sont non nuls alors toute racine en module de (4.2.1) est encadrée par $r$ et $R$ c.a.d

$$r < |\alpha| < R.$$  

En particulier, $r$ et $R$ sont les limites inférieures et supérieures des racines réelles positives et $-r$ et $-R$ sont les limites correspondantes pour les réelles négatives.
**Théorème 4.2.2** (Théorème de Lagrange).

Si $a_0 > 0$ et $a_k$ est le premier des coefficients négatifs du polynôme $P(x)$, on peut prendre pour limite supérieure des racines réelles positives le nombre $R = 1 + \sqrt[1]{\frac{B}{a_0}}$ où $B$ est la plus grande des valeurs absolues des coefficients négatifs de $P(x)$.

**Preuve :**

Elle est analogue à celle du théorème 4.2.1. En effet, soit $\alpha$ une racine de $P$, supposée réelle positive.

Si $\alpha < 1$ on a bien $\alpha < R$.

Si $\alpha > 1$, en écrivant l’équation (4.2.1) dans laquelle on remplace tous les coefficients positifs $a_1, \cdots, a_{k-1}$ par zéro et tous les coefficients $a_k, a_{k+1}, \cdots, a_n$ par $-B$, on obtient la majoration :

$$P(\alpha) \geq a_0 \alpha^n - B (\alpha^{n-k} + \cdots + \alpha^0) = a_0 \alpha^n - B \frac{\alpha^{n-k+1}}{\alpha - 1}.$$  

De telle sorte qu’avec $\alpha > 1$ on obtient

$$P(\alpha) > a_0 \frac{\alpha^{n-k+1}}{\alpha - 1} \left[ a_0 (\alpha - 1)^k - B \right]$$

et par suite, pour $\alpha \geq R = 1 + \sqrt[1]{\frac{B}{a_0}}$, on a $P(\alpha) > 0$. Donc toute racine positive $\alpha$ de (4.2.1) vérifie $\alpha < R$.

**Théorème 4.2.3** (Méthode de Newton).

Si pour $x = c > 0$ le polynôme $P(x)$ et toutes ses dérivées sont non négatives : $P^{(k)}(c) \geq 0$, $k = 0, 1, \cdots, n$ et $P^{(n)}(c) = n! a_0 > 0$, alors on peut prendre pour limite supérieure des racines positives la valeur $R = c$.

**Preuve :**

En fait, on peut construire une suite croissante $0 < c_1 \leq c_2 \leq \cdots \leq c_n$ qui vérifie

$$P^{(n-1)}(c_1) \geq 0, \ P^{(n-2)}(c_2) \geq 0, \cdots, P^{(n-1)}(c_{n-1}) \geq 0, P(c_n) > 0.$$  

Il suffit d’écrire la formule de Taylor appliquée à $P$.

$$P(x) = P(c) + (x - c)P'(c) + \cdots + \frac{(x - c)^n}{n!} P^{(n)}(c)$$  

et par suite, pour toute racine positive $\alpha$ on a $\alpha \leq c$, sinon on aboutit à une contradiction avec $P(\alpha) > 0$.  

97
**Théorème 4.2.4** (Théorème de Cauchy).

Soient

\[ Q_1(x) = |a_0|x^n - |a_1|x^{n-1} - \cdots - |a_n| \] et \[ Q_2(x) = |a_0|x^n + |a_1|x^{n-1} + \cdots + |a_{n-1}|x - |a_n|. \]

Alors \( Q_1 \) et \( Q_2 \) admettent chacun une seule racine positive, notons les \( R \) et \( r \), alors toute racine \( \alpha \) de \( P(x) \) vérifie \( r < |\alpha| < R \).

### 4.2.6 Nombre de racines réelles d’un polynôme

**Définition 4.2.1.** Soit \( x_1, x_2, \cdots, x_n \) une suite de réels avec \( x_1 \neq 0 \) et \( x_n \neq 0 \). On appelle nombre inférieur de changements de signes de la suite, le nombre \( N \) de changements de signes obtenu en omettant les termes nuls.

De même, le nombre supérieur de changements de signes de la suite, qu’on note \( \overline{N} \), est obtenu en remplaçant chaque terme nul par un terme ayant le signe opposé du terme qui le suit.

**Exemple 4.2.1.** Pour la suite \((x_1, x_2, x_3, x_4, x_5, x_6) = (5, 0, 0, -1, 0, 7)\). On a : \( N = 2 \) et \( \overline{N} = 4 \) (signe de \((5, -\varepsilon, \varepsilon, -1, -\varepsilon, 7))\).

**Remarque 4.2.2.** Si la suite ne comporte aucun terme nul, alors \( N = \overline{N} = N \), où \( N \) est le nombre de changements de signes.

**Théorème 4.2.5** (Théorème de Budan-Fourier).

Étant donnés deux réels \( a \) et \( b \) (\( a < b \)) tels que \( P(a) \neq 0, P(b) \neq 0 \), le nombre \( N(a, b) \) de racines réelles de \( P(x) = 0 \), comprises entre \( a \) et \( b \) est égal à \( \Delta N - 2k \), \( k \in \mathbb{N} \) où \( \Delta N \) est le nombre donné par : \( \Delta N = N(a) - N(b) \) pour la suite des dérivées \( \left(P(x), P'(x), \cdots, P^{(n)}(x)\right) \) prise en \( x = a \) et \( x = b \).

**Corollaire 4.2.2.** \( N(a, b) = \Delta N - 2k = N(a) - N(b) - 2k \) si les dérivées ne s’annulent pas pour \( x = a \) ni pour \( x = b \).

**Corollaire 4.2.3.** Si \( \Delta N = 0 \) alors pas de racines réelles entre \( a \) et \( b \).

**Corollaire 4.2.4.** Si \( \Delta N = 1 \) alors une et une seule racine réelle dans l’intervalle \([a, b]\).

**Corollaire 4.2.5** (Théorème de Descartes).

Le nombre de racines positives d’une équation algébrique où chaque racine est prise avec son ordre de multiplicité, est égal au nombre de changements de signes dans la suite des coefficients \( a_1, a_2, \cdots, a_n \) de \( P(x) \), les coefficients nuls étant omis, ou inférieur à ce nombre d’un nombre pair.
Preuve :
On applique le théorème 4.2.5 à $P^{(k)}$ sur $[a, b] = [0, \infty]$ en remarquant que $a_{n-k} = 0, k = 0, 1, \cdots, n$. Les $a_{n-k}$ ont même signe que $P^{(k)}(0)$ d’où les changements de signe dans $(a_0, a_1, \cdots, a_n)$ sont les mêmes que ceux de la suite $(P(0), P'(0), \cdots, P^{(k)}(0))$ ils sont égaux à $N(0)$. Par ailleurs, $P^{(k)}(+\infty)$ ont toutes même signe d’où $N(+\infty) = 0$. On a donc

$$N(a, b) = N(0, +\infty) = N(0) - c(+\infty) - 2k = N(0) - 2k.$$ 

Corollaire 4.2.6. Le nombre de racines réelles négatives est égal au nombre de changements de signes $-2k$.

Preuve : On considère $Q(x) = P(-x)$.

Exemple 4.2.2. $P(x) = x^4 - x^3 - x^2 + x - 1$
Ici $N = \overline{N} = N = 3$. Le nombre de racines positives est égal à $3 - 2k$ soit 3 ou 1.
Posons $Q(x) = P(-x) = x^4 + x^3 - x^2 - x - 1$.
Le nombre de racines négatives est égal à $1 - 2k$ d'où une seule possibilité : 1.
En résumé $P(x)$ admet soit 3 racines positives et une négative soit une positive, une négative et 2 racines complexes conjuguées.

4.2.7 Cas des racines isolées

a- Fonctions symétriques élémentaires des racines

Soient $x_1, x_2, \cdots, x_n$ les $n$ racines de $P(x) = 0$, alors par identification des écritures on obtient les relations :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n = -\frac{a_1}{a_0}$$
$$\sum_{i<j}^{n} x_i x_j = x_1 x_2 + x_1 x_3 + \cdots + x_1 x_n + \cdots + x_{n-1} x_n = \frac{a_2}{a_0}$$
$$\sum_{i<j<k}^{n} x_i x_j x_k = x_1 x_2 x_3 + x_1 x_3 \cdots \cdots = -\frac{a_3}{a_0}$$
$$\vdots$$
$$\prod_{i=1}^{n} x_i = (-1)^n \frac{a_n}{a_0}$$

Exemple 4.2.3. Si $x_1, x_2, x_3$ sont des racines de $x^3 + px^2 + qx + r = 0$ on obtient

$$\begin{cases} 
  x_1 + x_2 + x_3 = -p \\
  x_1 x_2 + x_1 x_3 + x_2 x_3 = q \\
  x_1 x_2 x_3 = r
\end{cases}$$
b- Relation de Newton

Si $S_p$ désigne la somme $S_p = x_1^p + x_2^p + \cdots + x_n^p = -p$ on a

\[
\begin{cases}
a_0S_1 + a_1 = 0 \\
a_0S_2 + a_1S_1 + 2a_2 = 0 \\
\vdots \\
a_0S_n + a_1S_{n-1} + \cdots + a_{n-1}S_1 + na_n = 0
\end{cases}
\]

c- Méthode de Graeffe :

On suppose ici que les racines sont nettement séparées c’est à dire :

\[
|x_1| >> |x_2| >> \cdots >> |x_n|.
\]

Alors, d’après (paragraphe a-), on obtient pour $x_1$

\[
x_1 \left[ 1 + \frac{x_2}{x_1} + \cdots + \frac{x_n}{x_1} \right] = -\frac{a_1}{a_0}
\]

On peut donc prendre $-\frac{a_1}{a_0}$ comme approximation de $x_1$, la plus grande racine en module. (En général, la racine $x_k$ peut être approchée par $-\frac{a_k}{a_{k-1}}$). De même $x_1x_2 \simeq \frac{a_2}{a_0}$.

Remarque 4.2.3. Si les racines $x_1, x_2, \cdots, x_n$ ne sont pas nettement séparées, on transforme l’équation $P(x) = 0$ de la manière suivante :

Si $P(x) = a_0(x - x_1)(x - x_2)\cdots(x - x_n)$ on pose $f_0(x) = P(x)$ et $P(-x) = (-1)^n a_0(x + x_1)(x + x_2)\cdots(x + x_n)$.

Alors $(-1)^n P(x) = a_0^2(x^2 - x_1^2)(x^2 - x_2^2)\cdots(x^2 - x_n^2) = f_2(x^2) = 0$.

On peut continuer ainsi pour obtenir :

$f_m(x^m) = a_0^m(x^m - x_1^m)\cdots(x^m - x_n^m)$ avec $m = 2^k = 2, 4, 8, \cdots$, si maintenant on écrit $f_m(x^m) = f_m(z) = 0$, avec $f_m(z) = A_0z^n + A_1z^{n-1} + \cdots + A_n$, les racines $z_k$ seront nettement séparées et pourront être approchées par $z_k \simeq \frac{A_k}{A_{k-1}}$ et par suite $x_k = \sqrt[n]{z_k}$.

e- Suites de Sturm et les racines réelles.

Définition 4.2.2. soit $f_0, f_1, \cdots, f_n$ une suite de fonctions définies sur $[a, b]$, supposées continues (avec $f_0$ de classe $C^1$). On dit que cette suite de fonctions est une suite de Sturm si
1. $f_0$ n’admet que des racines simples dans $[a, b]$. 

2. $f_n$ garde un signe constant sur $[a, b]$. 

3. Pour tout $k, 0 < k < n$ tel que $f_k(\alpha) = 0$ on a $f_{k-1}(\alpha)f_{k+1}(\alpha) < 0$. 

4. Si $\alpha$ est une racine de $f_0(\alpha) = 0$ alors $f_0'(\alpha)f_1(\alpha) > 0$.

**Exemple 4.2.4.** Soit $P$ un polynôme de degré $n$ qui n’admet que des racines simples et réelles $\alpha_1, \alpha_2, \ldots, \alpha_n$, on obtient une suite de Sturm de la façon suivante :

\[
\begin{align*}
    f_0(x) &= P(x) \\
    f_1(x) &= P'(x) \\
    \vdots \\
    f_{n-2}(x) &= f_{n-1}(x)q_{n-1}(x) - f_n(x) \quad \text{avec } d^n f_n = 0
\end{align*}
\]

autrement dit, on prend pour $f_1$ l’opposé du reste de la division euclidienne de $f_{k-2}$ par $f_{k-1}$. On vérifie facilement que :

- $f_{k+1}(\alpha) = 0 \Rightarrow f_k(\alpha)f_{k+2}(\alpha) < 0$. 

- Deux polynômes consécutifs ne peuvent pas s’annuler pour la même racine. 

- $f_n$ est constante.

**Théorème 4.2.6 (Théorème de Sturm).**

Soit $\{f_0, f_1, \ldots, f_n\}$ une suite de Sturm dans $[a, b]$ et $C(x)$ le nombre de changement de signe de la suite $(f_0(x), \ldots, f_n(x))$. Alors le nombre de racines réelles distinctes de $f_0$ dans $[a, b]$, en supposant que ni $a$ ni $b$ ne sont racines est égal à la différence entre le nombre de changements de signes de la suite $(f_0(a), f_1(a), \ldots, f_n(a))$ et celui de la suite $(f_0(b), f_1(b), \ldots, f_n(b))$. Soit $N(a, b) = C(a) - C(b)$. 

**Preuve:**

Pour tout $x \in [a, b]$, posons $N(x) = C(a) - C(x)$, $N(x)$ est un entier qui vériifie :

1. $N(x)$ reste constant tant que $f_i(x)$ ne s’annule pour aucun $i$.

2. Si pour $i \neq 0$, $f_i(\alpha) = 0$ alors d’après la définition, $f_{i-1}(\alpha)f_{i+1}(\alpha) < 0$, ce qui entraîne que $f_{i-1}(\alpha)$ et $f_{i+1}(\alpha)$ sont non nuls, et par continuité, $f_{i-1}$ et $f_{i+1}$ garderont le même signe dans un voisinage de $\alpha$ ; c’est-à-dire pour $\varepsilon$ assez petit, $f_{i-1}$ et $f_{i+1}$ gardent chacune un signe constant dans $[\alpha - \varepsilon, \alpha + \varepsilon] = I_\varepsilon$. Si on considère les trois fonctions $f_{i-1}, f_i$ et $f_{i+1}$ sur $I_\varepsilon$ alors les quatre possibilités suivantes peuvent se présenter (en supposant que $f_i$ ne change pas de
Le nombre de changements de signe de la suite \((f_0(\alpha - \epsilon), f_1(\alpha - \epsilon), \cdots, f_n(\alpha - \epsilon))\)
est même que celui de la suite \((f_0(\alpha + \epsilon), f_1(\alpha + \epsilon), \cdots, f_n(\alpha + \epsilon))\).
Par conséquent, \(N(x)\) est constant dans \([\alpha - \epsilon, \alpha + \epsilon]\).

3. Si \(f_0(\alpha) = 0\) (\(\alpha \neq 0, \alpha \neq b\)) on obtient les deux cas suivants :

Dans les deux cas, on remarque que la suite \((f_0(\alpha - \epsilon), f_1(\alpha - \epsilon), \cdots, f_n(\alpha - \epsilon))\)
presents un changement de signe de plus que la suite \((f_0(\alpha + \epsilon), f_1(\alpha + \epsilon), \cdots, f_n(\alpha + \epsilon))\)
dauss dans \(I_\epsilon\). D'où \(N(\alpha + \epsilon) = N(\alpha - \epsilon) + 1\) et \(N(b) = C(a) - C(b) = n\).
La fonction \(N(x)\) se présente comme une fonction en escalier.

Remarque 4.2.4. 1. Si on suppose que \(f_i\) change de signe en \(\alpha\), on arrive à la même conclusion.

2. Sous MATLAB le polynôme de degré \(n\), \(p(x) = a_n x^n + \cdots + a_0\) est défini par un vecteur \(p\) de dimension \(n + 1\) contenant les coefficients \(\{a_i\}_{i=0}^n\) rangées dans l’ordre décroissant des indices. c’est à dire que l’on a \(p(1) = a_n, \cdots, p(n+1) = a_0\).
La commande \texttt{polyval} permet d’évaluer \(p\) (la fonction polynomiale) en des points donnés. La syntaxe est \texttt{polyval(p,x)} où \(x\) est une valeur numérique ou un vecteur. Dans le second cas on obtient un vecteur contenant les valeurs de la fonction polynomiale aux différents points spécifiés dans le vecteur \(x\).
Utilisé avec la commande \texttt{fplot}, la commande \texttt{polyval} permet de tracer le graphe de la fonction polynomiale sur un intervalle \([x_{\text{min}}, x_{\text{max}}]\) donné. La syntaxe de l’instruction \(\texttt{fplot('polyval([a_n, \cdots, a_0],x'),[x_{\text{min}}, x_{\text{max}}])}\).
On obtient les racines du polynôme grâce à l’instruction \texttt{roots(p)}.
4.3 Applications

Application 4.3.1. Dynamique élémentaire de population

Considérons une population de taille \( P(t) \) à un certain temps \( t \) et supposons que nous nous intéressons à son évolution dans le temps. Quatre paramètres peuvent contribuer au changement de \( P(t) \):

- \( B \): naissances (birth),
- \( D \): mortalités (death),
- \( E \): émigration,
- \( I \): immigration.

Si on suppose que les paramètres \( B, D, E \) et \( I \) ne dépendent pas du temps et si on note \( P_n \) la population à l’instant \( t_n \) et \( P_{n+1} \) celle de l’instant \( t_{n+1} \) on peut exprimer la variation de la population par la formule:

\[
P_{n+1} - P_n = R
\]  

où \( R = B + I - D - E \).

On voit facilement que la population croît ou décroît avec le temps selon que \( R \) est positif ou négatif.

Application 4.3.2. Malthus et la croissance géométrique (Malthus, 1798)

Supposons maintenant que \( P_0 \) désigne la population initiale et qu’à chaque instant \( t_n \) le paramètre \( R \) soit donné par \( R = rP_n \) l’équation (4.3.1) devient alors:

\[
P_{n+1} = (1 + r)P_n
\]  

A partir de l’équation (4.3.2), il est facile de voir que:

- si \( r = 0 \) alors \( P_n = P_0 \) et la taille de la population reste constante dans le temps (c’est le cas des populations pour lesquelles les naissances compensent juste les mortalités et l’immigration compense l’émigration),
- si \( r < 0 \) alors \( P_{n+1} < P_n \) et la population décroît jusqu’à extinction \( (P_n \to 0 \text{ quand } n \to \infty) \),
- si \( r > 0 \) alors \( P_{n+1} > P_n \) et la population croît jusqu’à explosion \( (P_n \to \infty \text{ quand } n \to \infty) \).

Si \( r \neq 0 \) la croissance ou la décroissance de la population est géométrique et nous retrouvons ainsi la théorie de Malthus qui peut être résumée par l’équation suivante:
une croissance géométrique de la population + une croissance arithmétique des moyens de subsistance = grande misère humaine.

**Remarque 4.3.1.** La théorie de Malthus a été beaucoup critiquée. Elle est notamment mise en cause par les économies des pays industrialisés mais aussi et surtout par la répartition inégale et injuste des richesses dans d’autres pays.

**Application 4.3.3. Equation logistique et ressources limitées** (Verlhust, 1845)

Si maintenant on suppose que les ressources sont limitées et que toute population ne peut continuer à croître au delà de la capacité permise par l’environnement dans lequel vit cette population, on est amené à l’équation

\[ P_{n+1} = (1 + r)P_n - \frac{r}{K}P_n^2 \]  

(4.3.3)

En écrivant l’équation (4.3.3) sous la forme:

\[ P_{n+1} = g(P_n) \]

on voit aisément que l’équation logistique admet deux points d’équilibre (les points fixes de \( g \)) qui sont:

l’équilibre trivial \( P^* = 0 \) et l’équilibre non trivial atteint pour la capacité d’environnement \( P^{**} = K \).

En étudiant la dérivée de \( g \) aux points d’équilibre, on déduit que si \( r > 0 \) alors le point d’équilibre trivial est instable (\( g'(0) > 1 \)) tandis que pour le point d’équilibre non trivial (\( g'(K) = r - 1 \)), la zone de stabilité est donnée par \( 0 < r < 2 \).

L’équation logistique constitue un exemple pédagogique intéressant. La variation de \( r \) exhibe différentes situations allant de l’équilibre stable (\( 0 < r < 1 \)) voir figure 4.1, au chaos (\( r = 3 \)) voir figure 4.2, en passant par des oscillations cycliques (\( 2 < r < 3 \)) voir figure 4.2.
Figure 4.1: Convergence($r = 0.9$)

Figure 4.2: 2_cycles($r = 2.5$), Chaos($r = 3$)
Application 4.3.4. **Equation logistique et captivité (harvest)** (Schaefer, 1954)

Si on considère une population de poissons vivant dans un environnement aux ressources limitées et assujettie à une captivité (harvest) on obtient l’équation

\[ P_{n+1} = (1 + r)P_n - \frac{r}{K}P_n^2 - qEP_n \]  \hspace{1cm} (4.3.4)

L’équation (4.3.4) admet deux points d’équilibre qui sont l’équilibre trivial \( P^* = 0 \), et l’équilibre non trivial \( P^{**} = K\left(1 - \frac{qE}{r}\right) \).

L’équilibre trivial correspond à l’épuisement du stock de poissons et l’équilibre non trivial permet d’élaborer des stratégies optimales en fonction du stock \( P_n \), de l’effort de pêche \( E \) et de la constante \( q \) mesurant la captivité.

Application 4.3.5. **Equation logistique avec retard** (Nicholson, 1954)

Certains phénomènes n’apparaissent pas instantanément dans l’évolution des populations, la dynamique exige alors de considérer un effet à retard. L’effet d’une guerre ou d’une saturation d’environnement peut être vécu par des populations à posteriori. L’équation logistique avec retard (ou retardée) est donnée par

\[ P_n - P_{n-\tau} = r\tau P_{n-\tau}(1 - P_{n-\tau}/K). \]
4.4 Complément bibliographique

- Population de rats et suites de Fibonacci (Fibonacci, 1202)

Fibonacci a proposé un modèle de croissance d’une population de rats en partant d’un couple de rats (une femelle et un male) qui atteignent maturité pour donner un nouveau couple, qui, après maturité, donne un autre couple qui s’ajoute au couple que donne le premier couple s’il survit, etc... Les suites de Fibonacci vérifient

\[ B_0 = 1, B_1 = 1, \]
\[ B_{n+1} = B_n + B_{n-1}, \quad n = 1, 2, \ldots \]

- Population avec structure d’âge (Euler, 1760)

Si on considère une population humaine recensée à des intervalles de temps réguliers en tenant compte uniquement des femelles parmi les nouveaux nés. On note \( B_n \) le nombre de femelles nées durant le \( k^{\text{eme}} \) recensement, la population est divisée en classe d’âge de longueur égale à la longueur du recensement (10 ans par exemple). On note \( \lambda_k \) la proportion de ceux qui sont nés \( k \) recensements auparavant et qui survivent, autrement dit, \( \lambda_k \) représente la probabilité qu’un nouveau né survive au \( k^{\text{eme}} \) recensement. De façon similaire, on note \( b_k \) la fertilité des femmes qui sont dans leur \( k^{\text{eme}} \) intervalle de recensement. Par conséquent, les femmes nées \( k \) recensements auparavant devraient produire \( \lambda_k b_k B_{n-k} \) filles durant le \( k^{\text{eme}} \) recensement. On obtient ainsi ce qu’on appelle l’équation de renouvellement donnée par

\[ B_n = h_n + \lambda_1 b_1 B_{n-1} + \lambda_2 b_2 B_{n-2} + \cdots + \lambda_n b_n B_0. \]

Cette équation peut être vue comme une généralisation des suites de Fibonacci (en prenant \( h_n = 0, \lambda_1 = \lambda_2 = 1 \) et \( \lambda_j = 0 \) pour \( j \geq 3 \)).

- Dépendance de la densité (Verhulst, 1845)

Le modèle de Malthus suppose un taux de croissance \( r \) constant alors qu’en réalité, à partir d’un certain niveau, la conjugaison des facteurs d’environnement et de la densité d’une population peut empêcher la croissance de celle-ci. On obtient alors une généralisation des modèles de Malthus et de l’équation logistique en prenant

\[ B_{n+1} = R(B_n)B_n, \]

où \( R(B_n) \) est le taux de croissance intrinsèque de la population à l’instant \( t_n \). Les exemples ci-dessous peuvent en constituer des cas particuliers.
a- Equation de Ricker (1954): $B_{n+1} = \exp \left[ r \left( 1 - \frac{B_n}{K} \right) \right] B_n$.

b- Equation de Beverton-Holt (1957): $B_{n+1} = \frac{R_0}{1 + \left( \frac{R_0 - 1}{K} \right) B_n} B_n$.

c- Equation généralisée de Beverton-Holt (1957):

$$B_{n+1} = \frac{R_0^\beta}{\left( 1 + \frac{R_0 - 1}{K} B_n \right)^\beta} B_n.$$ 

d- Equation de Hassel-Lawton-May (1976): $B_{n+1} = (1 + r) \left( 1 + aB_n \right)^{-\beta} B_n$.

Voir aussi exercices 4.4.3, 4.4.4, 4.4.5.
4.5 Exercices

Exercice 4.5.1. Etudier la convergence de l’itération définie par

\[ x_{n+1} = \Phi(x_n) \quad \text{avec} \quad \Phi(x) = \frac{x^3 + \alpha x}{sx^2 - 1 + \alpha}. \]

Exercice 4.5.2. Soit \( f \) une fonction réelle admettant un minimum \( m \) en un point \( x^* \) de l’intervalle \([a, b]\).

On cherche à encadrer ce minimum par une méthode de dichotomie. La méthode du nombre d’or consiste à prendre \( L_1 = b - a \) puis à localiser \( x^* \) dans un intervalle de longueur \( L_i \) en calculant \( L_i \) par la relation de récurrence: \( L_i = L_{i-2} - L_{i-1} \) et en lui imposant que le rapport de réduction soit constant c’est à dire

\[ \frac{L_i}{L_{i-1}} = \frac{L_{i-1}}{L_{i-2}} = \text{constante.} \]

a) Trouver le nombre d’or \( \alpha \).

b) Proposer un algorithme permettant d’obtenir \( x^* = \min_{x \in [a, b]} f(x) \).

Exercice 4.5.3. Soit \( g \) une fonction de classe \( C^p \) avec \( p \geq 1 \) et \( g(x^*) \neq 1 \) on suppose que le procédé itératif \( x_{n+1} = g(x_n) \) vérifie \( x_{n+1} - x^* = (k + \delta_n)(x_n - x^*) \) avec \(|k| < 1 \) et \( \lim_{n \to \infty} \delta_n = 0 \). Soit

\[ z_n = x_n + \frac{(x_{n+1} - x_n)^2}{2x_{n+1} + x_n} \]

1. Chercher \( \lim_{n \to \infty} z_n - x^* \) et comparer la vitesse de convergence des suites \((x_n)\) et \((z_n)\).

2. On généralise le procédé précédent de la manière suivante: on définit la fonction \( h \) par \( h(x) = \frac{g(z(x)) - g(x)^2}{g(z(x)) - 2g(x) + x} \) puis on pose \( y_n = g(x_n), z_n = g(y_n) \) et \( x_{n+1} = h(x_n) \) pour \( n = 0, 1, 2, \ldots \).

(a) Prouver que \( g(x^*) = x^* \) si et seulement si \( h(x^*) = x^* \).

(b) On suppose de plus que \( g'(x^*) = \cdots = g^{(p-1)}(x^*) = 0 \) et que

\[ g^{(p)}(x^*) = p!A \neq 0 \]

en prenant \( x^* = 0 \) et \( g(x) = Ax^p + \frac{x^{p+1}}{(p+1)!}g'(\theta x) \), \( 0 < \theta < 1 \), montrer que:

\[ h(x) = -A^2x^{2p-1} + O(x^{2p}) \] si \( p > 1 \) et \( h(x) = O(x^2) \) si \( p = 1 \)

Exercice 4.5.4. Soit \( f \) une fonction réelle de classe \( C^2 \) sur un intervalle \([a, b]\) et telle que \( f''(x) = m \ \forall x \in [a, b] \).

109
1. Donner l’ordre de convergence de la méthode de fausse position avec

\[ x_0 \text{ fixé } (x_0 = a), \quad x_{n+1} = x_n - \frac{x_n - x_0}{f(x_n) - f(x_0)} f(x_n), \]

quelle est la constante d’erreur asymptotique.

2. On considère l’algorithme consistant à effectuer alternativement une itération de Newton-Raphson et une itération de la méthode de la sécante

\[ z_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x_{n+1} = z_n - \frac{x_n - z_n}{f(x_n) - f(z_n)} f(z_n). \]

Trouver l’ordre de cette méthode et la constante d’erreur asymptotique (on pourra poser: \( \varepsilon_n = x_n - x^* \) et \( e_n = z_n - x^* \)).

**Exercice 4.5.5.** (May-Conway-Hassell-Southwood, 1974)

Chercher les points d’équilibre et donner les conditions de stabilité pour les équations suivantes:

\[ P_{n+1} = \lambda P_n^{-(1+b)} \]
\[ P_{n+1} = \lambda P_n \exp \left[-\alpha P_n\right] \]
\[ P_{n+1} = P_n \left\{1 + r(1 - P_n/K)\right\} \]

**Exercice 4.5.6.** (Ludwing-Jones-Holling, 1978)

En l’absence de prédation, la dynamique d’une population de vers est régie par l’équation logistique. Soumise à une prédation par des oiseaux, l’équation devient:

\[ P_{n+1} - P_n = f(P_n) - g(P_n) \]

avec

\[ f(P_n) = rP_n - \frac{r}{K} P_n^2, \quad g(P_n) = \frac{\beta P_n^2}{\alpha^2 + P_n^2}. \]

En traçant les graphes de \( f \) et \( g \) en fonction de \( P_n \), montrer que le système admet un point non trivial d’équilibre stable ou 3 points d’équilibre dont 2 stables et un instable.

**Exercice 4.5.7.** (Brauer & Sanchez, 1975)

Supposons qu’une population régie par l’équation logistique subit une moisson à taux constant donnant lieu à l’équation:

\[ \frac{dP(t)}{dt} = rP(t) - \frac{r}{K} P^2(t) - E \]

a) Montrer que si \( E < \frac{1}{4} rK \); il ya deux possibilités d’équilibre, l’un stable, l’autre instable.

b) Montrer que si \( E > \frac{1}{4} rK \) alors la population disparaîtra en un temps fini.
c) Que deviennent les résultats a) et b) si $E$ est remplacé par $EP$.

d)Comparer avec le cas discret en considérant l’équation:

$$P_{n+1} = rP_n - \frac{r}{K}P_n^2 - E$$

Exercice 4.5.8. Soit $f(x) = x^3 - x^2 - x - 1 = 0$

1. Montrer que cette équation admet une racine dans l’intervalle $[1, 2]$.

2. Montrer que: l’itération $x_{n+1} = x_n^3 - x_n^2 - 1$ ne converge pour aucune valeur initiale $x_0 \in [1, 2]$ mais que la méthode $x_{n+1} = 1 + \frac{1}{x_n} + \frac{1}{x_n^2}$ converge pour toute valeur initiale $x_0 \in [\frac{7}{4}, 2]$.

Exercice 4.5.9. On suppose que l’équation $x = g(x)$ admet une seule racine $\alpha$ et que $g$ est de classe $C^p$ dans l’intervalle $I = \{x \text{ tels que: } |x - \alpha| \leq r \text{ avec } r > 0\}$.

Pour chercher $\alpha$, on utilise le procédé itératif $x_{n+1} = g(x_n)$.

1. Montrer que l’ordre de convergence est $p$ si et seulement si: $\alpha = g(\alpha)$ ; $g^{(i)}(\alpha) = 0$ pour $0 \leq i \leq p - 1$ et $g^{(p)}(\alpha) \neq 0$.

2. Déterminer la constante d’erreur asymptotique.

3. Pour approcher $\sqrt{A}, A > 0$, on utilise le procédé ci-dessus avec

$$g(x) = \frac{3A}{8x} + 3x^3 - \frac{x^3}{8A}$$

donner l’ordre de convergence et la constante d’erreur asymptotique.

Exercice 4.5.10. On s’intéresse à la solution numérique de l’équation $f(x) = 0$ où $f$ est une fonction scalaire régulière de $\mathbb{R}$ dans $\mathbb{R}$.

1. On considère la méthode de Newton modifiée

$$x_{n+1} = x_n - \frac{f(x_n)}{a_n},$$

où $a_n$ est une approximation de $f'(x_n)$.

Montrer que l’ordre de convergence est supérieur à 1 si $a_n \longrightarrow f'(x^*)$ quand $n \longrightarrow \infty$.

2. La méthode de la sécante est un cas particulier avec $a_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$.

Prouver que si cette méthode converge vers $x^*$ alors

$$\lim_{n \longrightarrow \infty} \frac{e_{n+1}}{e_ne_{n-1}} = \frac{f''(x^*)}{2f'(x^*)}, \quad e_n = x_n - x^*.$$  

En déduire l’ordre de convergence de cette méthode.
3. Etant donné deux réels $a$ et $b$ tels que $f(a)f(b) < 0$, de quelle façon peut-on adapter la méthode de la sécante pour garantir la convergence vers $x^* \in [a, b]$.

**Exercice 4.5.11.** Soit $p$ un entier $\geq 2$ et soit $\Phi$ la fonction définie sur $\mathbb{R}_+^*$ par $\phi(x) = Ax^{1-p}$.

On voudrait calculer une valeur approchée de la racine $n^{\text{ème}}$ de $A > 0$.

1. Le schéma: $x_{n+1} = \Phi(x_n)$ converge-t-il vers la solution $A^{1/p}$.

2. On considère une autre fonction définie par:

$$f(x) = x + \lambda(\Phi(x) - x).$$

(a) Pour quelles valeurs de $\lambda$ a-t-on convergence de l’itération

$$x_{n+1} = f(x_n).$$

(b) Donner la valeur optimale de $\lambda$.

(c)Comparer avec la méthode de Newton appliquée à la fonction $g(x) = x^p - A$.

3. Soit à résoudre $f(z) = 0$, $z \in \mathbb{C}$.

En écrivant $z = x + iy$ et $f(z) = u(x, y) + iv(x, y)$ et sachant que:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{et} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

Montrer que le schéma de Newton dans $\mathbb{C}$ est équivalent à deux schémas de Newton dans $\mathbb{R}$ et écrire ces deux schémas.

**Exercice 4.5.12 (Conditions de Routh-Hurwitz).**

Soit $A$ une matrice carrée d’ordre $n$, dont le polynôme caractéristique est donné par

$$P(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n,$$

où $a_i, i = 1, 2, \cdots, n$ sont réels. Montrer que les conditions nécessaires et suffisantes pour que les racines de $P$ soient de partie réelle strictement negative sont:

$$a_n > 0, D_1 = a_1 > 0, D_2 = \begin{vmatrix} a_1 & a_3 \\ 1 & a_2 \end{vmatrix} > 0$$

$$D_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ 1 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix} > 0$$
Exercice 4.5.13. Soit $P$ un polynôme de degré $n \geq 2$ à coefficients réels. On suppose que les zéros de $P(x)$ sont tous réels et vérifient:

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$  

On considère le schéma itératif de Newton appliqué à $P(x)$ avec:

$$x^{(0)}$$ donné tel que $x^{(0)}$ et $P(x^{(0)}) > 0$.

1. Vérifier que $P(x) > 0$ et $P'(x) > 0$ pour tout $x > \alpha_1$.

2. Montrer que la suite $(x^{(k)})$ est strictement décroissante.

3. Montrer que $x^{(k)} > \alpha_1$ et $x^{(k+1)} > \alpha_1$.

4. Conclure.

Exercice 4.5.14. Soit $f$ une fonction $f : \mathbb{R} \rightarrow \mathbb{R}$. On suppose que $f$ est de classe $C^2$ et qu'elle admet une fonction réciproque $g$. On cherche à résoudre numériquement l’équation $f(x) = 0$. Si $f(\alpha) = 0$ alors $\alpha = g(0)$ d’où l’idée de chercher $\alpha$ en passant par $g(y)$ évaluée au point $y = 0$. Soit $P$ un polynôme vérifiant l’une des 3 conditions suivantes:

$C_1 : P(y_n) = g(y_n) \text{ et } P(y_{n-1}) = g(y_{n-1})$

$C_2 : P(y_n) = g(y_n) \text{ et } P'(y_n) = g'(y_n)$

$C_3 : P(y_n) = g(y_n), P'(y_n) = g'(y_n) \text{ et } P''(y_n) = g''(y_n)$

1. On suppose que $P$ est un polynôme de degré 1 vérifiant la condition $C_2$

(a) Donner l’expression de $P(y)$ en fonction de $y, y_n, g(y_n)$ et $g'(y_n)$.

(b) Donner l’expression de $P(y)$ en fonction de $y, x_n, f(x_n)$ et $f'(x_n)$.

(c) Quel procédé obtient-on si on pose $x_{n+1} = P(0)$ (avec $y_n = f(x_n)$) ?
2. On suppose que $P$ est un polynôme de degré 1 vérifiant la condition $C_1$, quel procédé d’itération obtient-on en évaluant $P(y)$ et en prenant $x_{n+1} = P(0)$?

3. On suppose que $P$ est un polynôme de degré 2 vérifiant la condition $C_3$
   
   (a) Donner l’expression de $P(y)$ en fonction de $y$, $y_n$, $g(y_n)$, $g'(y_n)$ et $g''(y_n)$.
   
   (b) En exprimant les dérivées de $g$ en fonction de celles de $f$ et en prenant $x_{n+1} = P(0)$ montrer qu’on obtient le procédé de Tchebychev:
   $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f^2(x_n)f''(x_n)}{2(f'(x_n))^3}$$
   avec $f'(x_n) \neq 0$.
   
   (c) Montrer que l’ordre de cette méthode est 3 puis donner la constante d’erreur asymptotique.

**Exercice 4.5.15.** Soit $P(x)$ le polynôme donné par
$$P(x) = 2x^5 - 100x^2 + 2x - 1.$$ 

On cherche un encadrement des zéros $x_1, \ldots, x_5$ de l’équation $P(x) = 0$ (*).

a) Déterminer les limites inférieure et supérieure des modules des racines.

b) Préciser les limites des racines positives

c) En posant $z = -x$, montrer que l’équation (*) n’admet pas de racines négatives

d) En utilisant le théorème de Newton, montrer qu’on peut trouver des nombres $c_1, c_2, \ldots, c_5$ tels que $c_1 \leq c_2 \leq \cdots \leq c_5$ avec $P^{(4)}(c_1) \geq 0$, $P^{(3)}(c_2) \geq 0$, $P'(c_3) \geq 0$, $P''(c_4) \geq 0$ et $P(c_5) \geq 0$.

**Exercice 4.5.16.** Soit $A_k$ la matrice tridiagonale donnée par
$$A_k = \begin{pmatrix}
\alpha_1 & \beta_2 & & & \\
\beta_2 & \alpha_2 & \beta_3 & & \\
& \beta_3 & \ddots & \ddots \\
& & \ddots & \ddots & \beta_k \\
& & & \beta_k & \alpha_k
\end{pmatrix}.$$ 

En posant $P_0 = 1$, $P_1 = \alpha_1 - \lambda$ et $P_k = \det(A_k - \lambda I_k)$.

1. Montrer que $P_k = (\alpha_k - \lambda)P_{k-1} - (\beta_k)^2P_{k-2}$ pour $k \geq 2$.

2. Montrer que les zéros de $P_{n+1}$ sont distincts et strictement séparés par les zéros de $P_n$. 

114
3. En déduire que \((P_n, P_{n-1}, \cdots, P_0)\) est une suite de Sturm.

4. Application: soit A la matrice donnée par

\[
\begin{pmatrix}
1 & 2 & 0 & 0 \\
2 & 3 & -1 & 0 \\
0 & -1 & 4 & 1 \\
0 & 0 & 1 & 1
\end{pmatrix}
\]

(a) Montrer que les valeurs propres de A sont comprises entre \(-1\) et \(6\).

(b) En appliquant le théorème de Sturm, prouver que A possède une seule valeur propre négative.

(c) Obtenir d’autres encadrements des valeurs propres positives (sans les calculer).
Chapitre 5

Méthodes numériques de résolution des systèmes non linéaires

5.1 Méthodes itératives

Soit \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) une fonction vectorielle.

On cherche à résoudre numériquement l’équation \( F(x) = 0, \ x \in \mathbb{R}^n, \ 0 \in \mathbb{R}^n \). Comme au chapitre précédent, on peut se ramener aux méthodes du point fixe \( \Phi(x) = x \).

Théorème 5.1.1. Soit \( M \) un sous-espace fermé borné de \( \mathbb{R}^n \) tel que

\begin{enumerate}
  \item \( \Phi : M \rightarrow M \) continue.
  \item \( \Phi \) est contractante sur \( M \) c.à.d \( \exists \ 0 < L < 1 \) tel que:
    \[ \| \Phi(x) - \Phi(y) \| \leq L \| x - y \| \ \ \forall (x, y) \in M \times M \]

pour une certaine norme \( \| . \| \).
\end{enumerate}

Alors \( \Phi \) possède un point fixe unique \( x^* \in M \) et l’itération \( x^{(k+1)} = \Phi \left( x^{(k)} \right) \) converge vers \( x^* \) pour n’importe quel vecteur initial \( x^{(0)} \in M \), de plus

\[ \| x^{(k)} - x^* \| \leq \left( \frac{L^k}{1-L} \right) \| x^{(1)} - x^{(0)} \|. \]

Preuve :

Analogue à celle du cas scalaire en remplaçant \( |.| \) par \( \| . \| \).
Résoudre si on choisit. Par suite, soit Cependant, on évite de calculer l'inverse de. Elle est définie par le procédé itératif :

Méthode de Newton :

\[ x^{(k)} = \Phi \left( x^{(k-1)} \right) \] vériﬁe \( x^{(k)} \in V \) et \( x^{(k)} \) converge vers \( x \).

\[ x_{k+1} = x_k - J_F^{-1} \left( x_k \right) F \left( x_k \right) . \] (5.1.1)

Cependant, on évite de calculer l'inverse de \( J_F \) en solvant à chaque étape un système linéaire. L'équation (5.1.1) est équivalente à :

\[ J_F \left( x^{(k)} \right) \left( x^{(k+1)} - x^{(k)} \right) = - F \left( x^{(k)} \right) \]

soit \( J_F \left( x^{(k)} \right) C^{(k)} = - F \left( x^{(k)} \right) \) et \( x^{(k+1)} = x^{(k)} + C^{(k)} \), les \( C^{(k)} \) sont appelés des corrections. A chaque étape, on obtient \( x^{(k+1)} \) en corrigant la valeur de l'étape précédente en lui ajoutant \( C^{(k)} \) obtenue comme solution d'un système linéaire.

Exemple 5.1.1. Résoudre \( F(x) = 0 \) avec : \( F(x) = (x_1^2 + x_2^2 - 4, x_1 + x_2 - 2)^\top \).

Si on choisit \( x^{(0)} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \) alors \( J_F(x) = \begin{pmatrix} 2x_1 & 2x_2 \\ 1 & 1 \end{pmatrix}, \]

\( J_F \left( x^{(0)} \right) = \begin{pmatrix} 2 & 6 \\ 1 & 1 \end{pmatrix} \).
et \( F(x^{(0)}) = (6, 2)^\top \), on résoud donc : 
\[ I_F \left( x^{(0)} \right) C^{(0)} = -F \left( x^{(0)} \right) \]
ce qui donne : 
\[ C^{(0)} = (-3/2, -1/2)^\top, \quad x^{(1)} = x^{(0)} + C^{(0)} = (-1/2, 5/2)^\top \]
\[ x^{(2)} = x^{(1)} + C^{(1)} = (-1/12, 25/12)^\top \simeq (-0.08, 2.08)^\top. \]
\( \left( x^{(k)} \right) \) converge vers la solution \((0, 2)^\top \).

**Théorème 5.1.3 (Convergence quadratique de la méthode de Newton).**

Si \( x^* \) est une solution de \( F(x^*) = 0 \) et \( F \) est de classe \( C^1 \) dans \( \overline{B}(x^*, d) \) et \( \alpha \) et \( \gamma \) sont des constantes telles que

i) \( I_F(x) \) est inversible et \( \| I_F^{-1}(x) \| < \alpha \) \( \forall x \in \overline{B}(x^*, d) \).

ii) \( \| I_F(x) - I_F(y) \| < \gamma \| x - y \| \) \( \forall x, y \in \overline{B}(x^*, d) \).

iii) \( d < \frac{2}{\alpha \gamma} \).

Alors la méthode de Newton converge vers \( x^* \) pour tout choix de \( x^{(0)} \in \overline{B}(x^*, d) \) et la convergence est quadratique.

**Preuve :**

Posons \( \Phi(x) = x - I_F^{-1}(x)F(x) \)

Si \( x \in B(x^*, d) \) donc
\[
\| \Phi(x) - x^* \| = \left\| x - I_F^{-1}(x)F(x) - x^* \right\|
\]
\[
= \left\| x - x^* - I_F^{-1}(x) \left[ F(x) - F(x^*) \right] \right\|
\]
\[
= \left\| I_F^{-1}(x) \left[ F(x^*) - F(x) - I_F(x)(x^* - x) \right] \right\|
\]
\[
\leq \left\| I_F^{-1}(x) \right\| \left\| \int_0^1 (I_F(\xi(t)) - I_F(x))(x^* - x)dt \right\|
\]
\[
\leq \left\| I_F^{-1}(x) \right\| \int_0^1 \left\| I_F(\xi(t)) - I_F(x) \right\| \| x - x^* \| dt
\]
\[
\leq \alpha \gamma \| x - x^* \| \int_0^1 \| \xi(t) - x \| dt
\]
\[
\leq \alpha \gamma \| x - x^* \| \int_0^1 t dt \| x - x^* \|
\]
\[
\leq \alpha \gamma \| x - x^* \|^2
\]
\[
\leq \frac{\alpha \gamma}{2} d \| x - x^* \|
\]
d’où \( \| \Phi(x) - x^* \| < \| x - x^* \| \leq d \) si la condition iii) est vérifiée et comme on a
\[
\| x^{(k+1)} - x^* \| \leq \frac{\alpha \gamma}{2} \| x^{(k)} - x^* \|^2
\] alors la suite \( \left( x^{(k)} \right) \) converge vers \( x^* \) et la convergence est quadratique.
**Remarques 5.1.1.** 1. D’après le théorème 5.1.3, la méthode de Newton converge avec ordre de convergence quadratique mais à condition que le choix de départ soit assez proche de $x^*$, $x^{(0)} \in B(x^*, d)$. Il s’agit donc d’une convergence locale. Comme dans le cas scalaire, on peut obtenir une convergence globale mais au prix d’hypothèses supplémentaires sur $F$.

2. Pour réduire le nombre de calculs nécessaires, on peut remplacer l’équation (5.1.1) par :

$$x^{(k+1)} = x^{(k)} - J_F(x^{(0)}) F(x^{(k)})$$  

(5.1.2)

on obtient ainsi une méthode de Newton modifiée qui est évidemment moins précise que la précédente mais moins coûteuse.

3. Si au lieu de l’équation (5.1.2) on considère un Schéma itératif où $J_F(x^{(0)})$ est remplacé par une matrice $\Theta$, on obtient avec $\Theta$ diagonale, la méthode de Whittaker :

$$x^{(k+1)} = x^{(k)} - \Theta F(x^{(k)})$$

qui est une méthode de relaxation semblable à ce qu’on obtient en dimension un.

**Exemple 5.1.2.** Soit à résoudre $F(x)$ avec $F(x) = (x_1^2 + x_2^2 - x_3 - 2, x_1 + 5x_2 + 2, x_1x_3 - 2x_1)^\top$ alors en choisissant $\Theta = \begin{pmatrix} -1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & -1/2 \end{pmatrix}$, on obtient le schéma

$$x^{(k+1)} = \Phi(x^{(k)}) \quad \text{où} \quad \Phi(x) = \begin{pmatrix} x_1 + \frac{1}{4}x_1^2 + \frac{1}{4}x_2^2 - \frac{1}{4}x_3 - \frac{1}{2}, \frac{1}{5}x_1 - \frac{2}{5}, x_3 + \frac{1}{2}x_1x_3 - x_1 \end{pmatrix}^\top$$

$J_\Phi(x) = \begin{pmatrix} \frac{1}{2}x_1 + 1 & \frac{1}{2}x_2 & -\frac{1}{4} \\ -1/5 & 0 & 0 \\ \frac{1}{2}x_3 - 1 & 0 & \frac{1}{2}x_1 + 1 \end{pmatrix}$

$x^* = (-2, 0, 2)^\top$ est solution de $F(x) = 0$ et on vérifie que

$J_\Phi(x^*) = \begin{pmatrix} 0 & 0 & -1/4 \\ -1/5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

et $\|J_\Phi(x^*)\| < 1$, impliquant convergence locale.

**Théorème 5.1.4** (Convergence globale quadratique de la méthode de Newton). 
Soient $C$ un ouvert de $\mathbb{R}^n$ et $C_0$ un convexe tel que $C_0 \subset C$, et soit $F : C \rightarrow \mathbb{R}^n$ une fonction différentiable sur $C_0$ et continue sur $C$.

Pour $x_0 \in C_0$ on définit les constantes positives $\alpha$, $\gamma$, $\beta$, $h$ telles que

$$B(x_0, r) = \{x/\|x - x_0\| < r\} \subset C_0,$$
\[
    h = \frac{\alpha \beta \gamma}{2} < 1,
    \\
    r = \frac{\alpha}{1 - h}.
\]

si \( F(x) \) vériﬁe les conditions suivantes

i) \( J_F(x) \) est inversible et \( \|J_F^{-1}(x)\| \leq \beta \ \forall x \in C_0 \)

ii) \( \|J_F(x) - J_F(y)\| \leq \gamma \|x - y\| \ \forall x, y \in C_0. \)

iii) \( \|J_F^{-1}(x_0)F(x_0)\| \leq \alpha. \)

Alors la méthode de Newton converge vers \( x^* \) et la convergence est quadratique.

**Preuve:** Voir Stoer et Bulirsch [165].

Avec des conditions plus fortes on obtient unicité de la solution

**Théorème 5.1.5** (Newton - Kantorovich).

Soit \( F : C \subset \mathbb{R}^n \rightarrow \mathbb{R}^n \) une fonction continûment differentiable sur un convexe \( C_0 \subset C \) vériﬁant les conditions suivantes

i) \( \|J_F^{-1}(x_0)\| \leq \beta, \)

ii) \( \|J_F(x) - J_F(y)\| \leq \gamma \|x - y\| \ \forall x, y \in C_0. \)

iii) \( \|J_F^{-1}(x_0)F(x_0)\| \leq \alpha, \)

pour \( x_0 \in C_0 \). Soient \( h \) et \( r_{12} \) des constantes telles que

\[
    h = \alpha \beta \gamma,
    \\
    r_{12} = \frac{1 \pm \sqrt{1 - 2h}}{h} \alpha.
\]

Si \( h \leq \frac{1}{2} \) et \( \overline{B}(x_0, r_{12}) \subset C_0 \) alors la méthode de Newton converge vers la solution unique \( x^* \).

**Preuve:** Voir Stoer et Bulirsch [165].

### 5.2 Problèmes d’optimisation

Considérons les deux problèmes suivants:

1. Trouver \( x^* \in \mathbb{R}^n \) solution de \( F(x) = 0 \) avec \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \).
2. Trouver \( x^* \in \mathbb{R}^n \) minimisant \( f(x) \) avec \( f : \mathbb{R}^n \rightarrow \mathbb{R} \).
Ces deux problèmes sont liés de la manière suivante

i) En posant \( f(x) = (F(x))^\top F(x) \), \( f \) admet un minimum \( x^* \) qui est solution de 
\( F(x) = 0 \).

ii) Réciproquement, en prenant 
\( F(x) = \nabla f = \text{grad} f = \left( \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)^\top \), alors 
\( F(x^*) = 0 \) en \( x^* \) solution du problème 2.

A- Méthodes de descente :
Nous cherchons des solutions du problème 2 en utilisant des méthodes itératives donnant une suite de points \( x^{(1)}, x^{(2)}, \ldots \) à partir d’un point initial \( x^{(0)} \) et qui converge vers un minimum local \( x^* \). On voudrait obtenir une convergence monotone c’est à dire 
\( f(x^{(0)}) > f(x^{(1)}) > \cdots > f(x^{(i)}) > f(x^{(i+1)}) > \cdots > f(x^*) \).

Pour cela, on a besoin de deux paramètres pour passer d’un itéré \( x^{(i)} \) à l’itéré suivant \( x^{(i+1)} \).

a) La direction dans laquelle on se dirigera (direction de recherche ).

b) La longueur avec la quelle on avance à chaque étape.

On obtient ainsi un algorithme de descente qu’on peut présenter de la manière suivante :

1. \( i = 0 \) ; évaluer \( f(x^{(0)}) \)
2. chercher \( \rho^{(i)} \), direction de recherche
3. chercher \( \alpha_i \), longueur de l’étape
4. calculer le nouveau point \( x^{(i+1)} = x^{(i)} + \alpha_i \rho^{(i)} \)
5. évaluer \( f(x^{(i+1)}) \)
6. faire \( i = i + 1 \) et tester pour s’arrêter ou repartir de 2.

1- Quelle direction choisir ?

Nous cherchons une direction de manière à aller en descendant, en d’autres termes:
\( f(x^{(i)} + \varepsilon \rho^{(i)}) < f(x^{(i)}) \) pour tout \( \varepsilon \) assez petit.

En appliquant la formule de Taylor à \( f(x^{(i)} + \varepsilon \rho^{(i)}) \) on obtient :
\[
f(x^{(i)} + \varepsilon \rho^{(i)}) = f(x^{(i)}) + \varepsilon \left( \nabla f(x^{(i)}) \right)^\top \rho^{(i)} + O(\varepsilon^2)
\]
et par suite nous exigeons que \( (\nabla f(x^{(i)}))^\top \rho^{(i)} < 0 \) \( (*) \)
En posant \( g(x) = \nabla f(x) \) pour alléger les notations, l’inégalité \( (*) \) s’écrit :
\[
\left( g^{(i)} \right)^\top \rho^{(i)} < 0,
\]
c’est donc la condition qui permet de choisir à chaque étape \( i \) la direction \( \rho^{(i)} \).

2- Longueur de l’étape (le pas)

Une fois la direction \( \rho^{(i)} \) choisie, nous sommes ramenés à un problème de minimisation à une dimension de \( f \) dans la direction \( \rho^{(i)} \). Le problème se pose donc comme suit :

Trouver \( \alpha^* \) tel que \( f \left( x^{(i)} + \alpha^* \rho^{(i)} \right) < f \left( x^{(i)} + \alpha \rho^{(i)} \right) \) pour tout \( \alpha \) dans un voisinage de \( \alpha^* \).

Ce procédé s’appelle une recherche en ligne. On peut utiliser des méthodes numériques pour encadrer \( \alpha^* \). Si on utilise une recherche en ligne exacte de telle sorte que \( f \left( x^{(i)} + \alpha^* \rho^{(i)} \right) \) soit un minimum local de \( f \) dans la direction \( \rho^{(i)} \), alors on doit avoir
\[
\frac{\partial f}{\partial \alpha} = 0 \text{ pour } \alpha = \alpha^*,
\]
ou encore \( \left( \frac{\partial f}{\partial x} \right)^\top \rho^i = 0 \) au point \( x = x^{(i)} + \alpha^* \rho^{(i)} \).

Ce qui donne la condition d’orthogonalité suivante : \( \left( g^{(i+1)} \right)^\top \rho^i = 0 \). En d’autres termes, le gradient de \( f \) au point \( x^{(i+1)} \) doit être orthogonal à la direction de recherche si on utilise la recherche en ligne exacte.

Choix des directions:

a- Un moyen simple et naturel de choisir des directions de recherche est de minimiser par rapport à chaque variable à part. Partant de \( x^{(0)}_1 \), on garde \( x^{(0)}_2, \ldots, x^{(0)}_n \) fixes et on minimise par rapport à \( x^{(0)}_1 \) ensuite on refait le même chemin avec \( x^{(1)}_1, x^{(0)}_3, \ldots, x^{(0)}_n \) fixes etc... En dimension deux on obtient une allure en escalier.

**Exemple 5.2.1.** \( f(x_1, x_2) = 2x_1^2 - 6x_1x_2 + 5x_2^2 - 2x_1 + 2x_2 \).

Prenons \( x^{(0)} = (0, 0)^\top, f \left( x^{(0)} \right) = 0, \)

1. On garde \( x^{(0)}_2 = 0 \) et en cherche \( x^{(1)}_1 \) tel que \( \left( \frac{\partial f}{\partial x_1} \right) \left( x^{(1)}_1, 0 \right) = 0 \) on obtient
On garde

\[ \frac{\partial f}{\partial x_1} \left(x_1^{(1)}, 0\right) = 4x_1 - 6x_2 - 2 \text{ d'où } x_1^{(1)} = \frac{1}{2} \text{ et } x^{(1)} = (1/2, 0)^T, \quad f(x^{(2)}) = -\frac{1}{2} \]

2. On garde \( x_1^{(1)} = \frac{1}{2} \) et on cherche \( x^{(2)} \) tel que \( \left( \frac{\partial f}{\partial x_2} \right) \left(x_1^{(1)}, x^{(2)} \right) = 0 \)

Ce qui donne \( \frac{\partial f}{\partial x_2} \left(x_1^{(1)}, x^{(2)} \right) = -6x_1 + 10x_2 + 2 \text{ d'où } x_2^{(2)} = \frac{1}{10} \text{ d'où } x^{(2)} = (1/2, 1/10)^T \text{ et } f \left( x^{(2)} \right) = -\frac{11}{20} \)

3. On garde \( x_2^{(2)} = \frac{1}{10} \) et on cherche \( x^{(3)} \) tel que \( \left( \frac{\partial f}{\partial x_1} \right) \left(x_1^{(3)}, x_2^{(2)} \right) = 0 \ldots \)

b- Méthode de la meilleure descente (de la plus grande pente).

En écrivant \( f \left( x^{(i)} + \rho \right) = f \left( x^{(i)} \right) + \left( g^{(i)} \right)^T \rho + \cdots \)

On choisit \( \rho \) de telle sorte à minimiser \( \left( g^{(i)} \right)^T \rho \) avec une condition de minimisation \( ||\rho|| = C \) pour une certaine norme \( ||.|| \). On obtient alors (avec la norme euclidienne)

\[ -\left( g^{(i)} \right)^T \rho < \left\| g^{(i)} \right\|_2 ||\rho||_2 \text{ pour tout } \rho, \text{ d'où } \left( g^{(i)} \right)^T \rho > -\left\| g^{(i)} \right\|_2 ||\rho||_2 \text{ pour tout } \rho. \]

Soit \( \tilde{\rho} = -\frac{C}{2} g^{(i)} \) alors \( \left( g^{(i)} \right)^T \tilde{\rho} = -C \left\| g^{(i)} \right\|_2 \), donc \( \left( g^{(i)} \right)^T \rho > -C \left\| g^{(i)} \right\|_2 \)

Dans la méthode classique on prend \( \rho^{(i)} = -g^{(i)} \).

**Exemple 5.2.2.** \( F(x_1, x_2) = 2x_1^2 - 6x_1x_2 + 5x_2^2 - 2x_1 + 2x_2 \) qu'on peut écrire

\[ f(x_1, x_2) = \frac{1}{2} x^T G x + b^T \text{ avec } G = \begin{pmatrix} 4 & -6 \\ -6 & 10 \end{pmatrix}, \quad b = \begin{pmatrix} -2 \\ 2 \end{pmatrix}. \]

On a \( g(x) = G x + b = \begin{pmatrix} -4x_1 - 6x_2 - 2, -6x_1 + 10x_2 + 2 \end{pmatrix}^T \). Prenons \( x^0 = (0, 0)^T \) donc \( g^{(0)} = (-2, 2)^T, \rho^{(0)} = -g^{(0)} = (2, -2)^T \) et \( x^{(i)} = x^{(0)} + \alpha \rho^{(0)} = (2\alpha, -2\alpha) \)

\[ f \left( x^{(1)} + \alpha \rho^{(0)} \right) = 52\alpha^2 - 8\alpha \text{ qui atteint son minimum pour } \alpha = \frac{8}{2 \times 52} = \frac{1}{13} \]

on obtient donc \( x^{(1)} = x^{(0)} + \frac{1}{13} \rho^{(0)} = (2/13, -2/13)^T \) et \( f(x^{(1)}) = -\frac{4}{13} \)

ensuite \( \rho^{(0)} = -g^{(1)} = (-6/13, -6/13)^T \) etc \ldots

**Remarque 5.2.1.** Si on considère la méthode de Newton pour la résolution de \( F(x) = 0 \) c.à.d.

\( x^{(k+1)} = x^{(k)} - I_F \left( x^{(k)} \right)^{-1} F \left( x^{(k)} \right) = x^{(k)} + \alpha_k \rho^{(k)} \) avec \( \rho^{(k)} = I_F \left( x^{(k)} \right)^{-1} F \left( x^{(k)} \right) \)

et \( \alpha_k = 1. \) Alors il s’agit d’une méthode de descente pour minimiser la fonction

\[ f(x) = 1/2F(x)^T F(x). \]

123
En effet, $\nabla f = f^T F$ et on vérifie aussi que $\left(g^{(k)}\right)^T \rho^{(k)} < 0$ puisque

$$
\left(g^{(k)}\right)^T \rho^{(k)} = (J_k^T F_k) \left(-I_k^{-1} F_k\right) = -F_k^T J_k I_k^{-1} F_k = -F_k^T F_k.
$$

**B- Méthodes des directions conjuguées**

Ces méthodes sont aussi basées sur les algorithmes de descente : $x^{(k+1)} = x^{(k)} + \alpha_k \rho^{(k)}$ mais elles choisissent des directions conjuguées.

**Définition 5.2.1.** Soit $\rho^{(1)}, \ldots, \rho^{(n)}$ un ensemble de directions et $M$ une matrice définie positive. Alors on dira que les directions $\rho^{(i)}$ sont conjuguées par rapport à $M$ ou $M-$ conjuguées si :

$$
\left(\rho^{(i)}\right)^T M \rho^{(j)} = 0 \quad \forall \ i \neq j.
$$

On sait que pour les méthodes de descente avec direction en ligne exacte on obtient :

$$
\left(g^{(k+1)}\right)^T \rho^{(k)} = 0 \quad k = 0, 1, \ldots
$$

posons $s^{(k)} = x^{(k+1)} - x^{(k)}$ et $t^{(k)} = g^{(k+1)} - g^{(k)}$ et supposons que $f$ est une fonction quadratique $f(x) = \frac{1}{2}x^T Gx + b^T x + c$ de telle sorte que $g(x) = \nabla f(x) = Gx + b$ et $\nabla^2 f(x) = G$ Hessien. Il s’ensuit alors que :

$$
t^{(k)} = G \left(s^{(k)}\right) = \alpha_k G \rho^{(k)}
$$

par suite, pour tous $i$ et $j$ tel que $1 \leq j \leq i \leq n$ on a :

$$
\left(g^{(i+1)}\right)^T \rho^{(j)} = \left(g^{(j+1)} + t^{(j+1)} + \cdots + t^{(i)}\right)^T \rho^{(j)}
= \left(g^{(j+1)} + \alpha_{j+1} G \rho^{(j+1)} + \cdots + \alpha_i G \rho^{(i)}\right)^T \rho^{(j)}
= 0
$$

en particulier, on obtient :

$$
\left(g^{(n+1)}\right)^T \rho^{(j)} = 0 \quad j = 1, 2, \cdots, N
$$

et comme les $\rho^{(j)}$ sont linéairement indépendantes et donc engendrent $\mathbb{R}^n$, on conclue que $g \left(x^{(n+1)}\right) = 0$, nous venons donc de prouver le théorème suivant

**Théorème 5.2.1.** Si $f$ est une fonction quadratique avec Hessien définie positive alors les directions de recherche sont conjuguées par rapport à $G$.

**C- Méthode du gradient conjugué**

Elle consiste à choisir la direction $\rho^{(k)}$ sous la forme:

$$
\rho^{(k)} = -g^{(k)} + \beta_k \rho^{(k-1)} \quad \text{avec} \quad \rho^0 = -g^0
$$
L’algorithme est donc le suivant:

\[
\begin{cases}
  x^{(i+1)} & = x^{(i)} + \alpha_i \rho^{(i)} \\
  \rho^{(i)} & = -g^{(i)} + \beta_i \rho^{(i-1)} \\
  \rho^{(0)} & = -g^{(0)}
\end{cases}
\]

On démontre alors que pour les fonctions quadratiques les \(\alpha_i\) et les \(\beta_i\) sont données par

\[
\alpha_i = \left(\frac{g^{(i)}}{\rho^{(i)}}\right) \top g^{(i)}\left(\frac{\rho^{(i)}}{\rho^{(i)}}\right), \quad \text{où } G \text{ est le Hessien et } \beta_i = \left(\frac{g^{(i)}}{g^{(i-1)}}\right) \top g^{(i)}.
\]

**Remarques 5.2.2.**

1. La méthode du gradient conjugué appliquée à la résolution du système \(Ax = b\) se présente sous une forme similaire :
   - *on choisit* \(x^{(0)} \in \mathbb{R}^n\)
   - *on prend* \(\rho^{(0)} = r^{(0)} = b - Ax^{(0)}\)
   - *pour* \(k = 1, 2, \ldots\)
     - si \(\rho^{(k)} = 0\) stop, \(x^{(k)}\) est la solution de \(Ax = b\)
     - sinon calculer \(a_k = \left(\frac{r^{(k)}}{\rho^{(k)}}\right) \top \rho^{(k)}\), \(x^{(k+1)} = x^{(k)} + a_k \rho^{(k)}\),
       \(r^{(k+1)} = r^{(k)} - a_k A \rho^{(k)}\), \(b_k = \frac{\left(\frac{r^{(k+1)}}{r^{(k)}}\right) \top r^{(k+1)}}{\left(\frac{r^{(k)}}{r^{(k)}}\right) \top r^{(k)}}\) et
       \(\rho^{(k+1)} = r^{(k+1)} + b_k \rho^{(k)}\).

2. Résultat analogue pour \(f\) quelconque
5.3 Applications

5.3.1 Modèles écologiques

À la fin du chapitre quatre, nous avons considéré la dynamique des populations en nous limitant à l’étude d’une seule espèce vivant dans un certain environnement qui caractérise son comportement (croissance, décroissance, extinction, explosion, stabilité, etc...).

Dans ce chapitre, nous retenons toutes les considérations intraclasses liées au comportement d’une espèce en l’absence des autres espèces mais nous introduisons en plus l’effet d’interactions interspécies.

5.3.2 Modèles à deux espèces

Une population ne peut vivre isolée dans un environnement, elle entre en contact avec d’autres populations d’une manière ou d’une autre, et la nature de ces interactions contribue à son évolution. Un ensemble de deux ou plusieurs espèces est dit communauté. Une classification traditionnelle d’un système de populations est basée sur la nature des interactions interspécies et le comportement intraespèce.

Une grande et importante classe de modèles écologiques consacrés à la modélisation de la dynamique des communautés de populations, notamment avec interaction, peut être obtenue à partir du modèle général de Lotka-Volterra

\[
\frac{dN_i}{dt} = N_i(r_i - \sum_{j=1}^{n} a_{ij}N_j) \quad i = 1, \ldots, n
\]

où \(N_i(t)\) représente la taille ou la densité de l’espèce \(i\), \(r_i\) est le taux intrinsèque de croissance ou décroissance de l’espèce \(i\) en l’absence des autres espèces, \(A = (a_{ij})\) est appelée matrice des communautés, \(|a_{ij}|\) indique l’intensité d’interaction entre l’espèce \(i\) et l’espèce \(j\). Le signe de \(a_{ij}\) indique la nature d’interaction entre l’espèce \(i\) et l’espèce \(j\).

Dans le cas de deux espèces \(N_i\) et \(N_j\), les formes d’interaction les plus fréquentes dans la littérature sont données par le tableau suivant:

Volterra est considéré comme le pionnier de la modélisation écologique mais depuis ses premiers travaux en 1926, une littérature abondante est consacrée à ce sujet.

Dans le cadre de ce chapitre, nous donnons succintement quelques exemples mais sans rentrer dans les détails d’étude de stabilité (voir exercices). Le lecteur
Coopération ou mutualisme ou symbiose

Proie-Prédateur ou ressource-consommateur ou hôte-parasite

Commensalisme

compétition

Amensalisme

Neutralisme

Tableau 5.1: Natures des interactions

interessé trouvera dans la littérature les études complètes des comportements des différents systèmes avec les outils mathématiques et écologiques nécessaires.

5.3.3 Modèle proie-prédateur

-Cas continu :

\[
\begin{aligned}
& x' = r_1 x - b_1 x y \\
& y' = -r_2 x + b_2 x y
\end{aligned}
\]

\(x\): densité de populations des proies.
\(y\): densité de populations des prédateurs.
\(r_1\): taux de croissance des proies.
\(r_2\): taux de mortalité des prédateurs.
\(b_1\) et \(b_2\): taux de prédation.

L’approximation de

\[
\frac{\Delta x}{\Delta t} = \frac{x(t + \Delta t) - x(t)}{\Delta t}
\]

avec \(\Delta t = 1\) donne:

\[
\begin{aligned}
& x_{n+1} = (1 + r_1 - b_1 y_n) x_n \\
& y_{n+1} = (1 - r_2 + b_2 x_n) y_n
\end{aligned}
\]

\(r_2 = 1\) le système devient alors:

\[
\begin{aligned}
& x_{n+1} = (1 + r_1 - b_1 y_n) x_n \\
& y_{n+1} = b_2 x_n y_n
\end{aligned}
\]

En l’absence des prédateurs les proies croissent infiniment mais en réalité cette croissance est soumise à la capacité d’accueil de l’environnement, on introduit alors un terme \(-c x_n^2\) qui caractérise la compétition intraespèce au sein de la communauté des proies, les nouvelles équations sont :

\[
\begin{aligned}
& x_{n+1} = (1 + r_1 - c x_n - b_1 y_n) x_n \\
& y_{n+1} = b_2 x_n y_n
\end{aligned}
\]
Exemple 5.3.1. ([41])

\[ r_1 = 1, \ b_1 = 2, \ b_2 = 2, c = 3 \implies \text{convergence vers } (0.333,0) \text{ voir figure 5.1); } \]
\[ r_1 = 1, \ b_1 = 1, \ b_2 = 2, c = 2 \implies \text{convergence asymptotique vers } (0.333,0.33) \text{ voir figure 5.2); } \]
\[ r_1 = 3, \ b_1 = 4.5, \ b_2 = 3, c = 2 \implies \text{stabilité du point } (0.333,0.777) \text{ voir figure(5.3);} \]
\[ r_1 = 2.3, \ b_1 = 3.3, \ b_2 = 3.3, c = 3.3 \implies 2 \text{-points cycles voir figure 5.4);} \]
\[ r_1 = 2.4, \ b_1 = 3.4, \ b_2 = 3.4, c = 3.4 \implies 4 \text{-points cycles voir figure 5.5);} \]
\[ r_1 = 2.5, \ b_1 = 3.5, \ b_2 = 3.5, c = 3.5 \implies \text{comportement chaotique voir figure 5.6).} \]

Figure 5.1: Convergence

Figure 5.2: Stabilité asymptotique

128
Figure 5.3: Stabilité

Figure 5.4: 2 cycles

Figure 5.5: 4 cycles
5.3.4 Modèle à compétition

L’interaction entre les deux espèces a un effet négatif sur les deux densités; le modèle est régi par le système d’équations suivantes:

\[
\begin{align*}
x_{n+1} &= (1 + r_1 - a_{11}x_n - a_{12}y_n)x_n \\
y_{n+1} &= (1 + r_2 - a_{22}y_n - a_{21}x_n)y_n
\end{align*}
\]

- \(r_i\) : taux de croissance de chaque espèce.
- \(a_{ii}\) : compétition intra-espèce.
- \(a_{ij}\) : compétition inter-espèce.

**Exemple 5.3.2.** ([41])

\(r_1 = 0.5\), \(a_{11} = 1\), \(a_{12} = 3\), \(a_{21} = 4\), \(r_2 = 0.2\), \(a_{22} = 1\) \(\implies\) extinction de la première espèce voir figure(5.7);  

\(r_1 = 1.6\), \(a_{11} = 1\), \(a_{12} = 4\), \(a_{21} = 3\), \(r_2 = 1.5\), \(a_{22} = 2\) \(\implies\) extinction de la deuxième espèce voir figure(5.7);  

\(r_1 = 0.9\), \(a_{11} = 0.5\), \(a_{12} = 0.4\), \(a_{21} = 0.3\), \(r_2 = 0.8\), \(a_{22} = 0.6\) \(\implies\) coexistence des deux espèces voir figure(5.8);
Figure 5.7: Extinction d’une espèce

Figure 5.8: Coexistence

131
5.3.5 Modèle à coopération

Il est donné par le système d’équations suivantes:

\[
\begin{align*}
x_{n+1} &= (1 + r_1 - a_{11}x_n + a_{12}y_n)x_n \\
y_{n+1} &= (1 + r_2 - a_{22}y_n + a_{21}x_n)y_n
\end{align*}
\]

**Exemple 5.3.3.** ([41])

\(r_1 = 0.1, a_{11} = 0.4, a_{12} = 0.2, a_{21} = 0.15, r_2 = 0.15, a_{22} = 0.5\) génèrent une coexistence des deux espèces voir figure (5.9).

![Figure 5.9: Coexistence](image)

5.4 Complément bibliographique

En 1926, sur proposition d’un ensemble de statsiques concernant la vente de poissons dans les marchés de Trieste par le biologiste et marin d’Ancona, Volterra proposa le modèle classique proie-prédateur

\[
N' = rN - cNP
\]

\[
P' = bNP - mP.
\]

Pour les modèles à compétition, la formulation classique due à Lotka(1932) et Volterra(1926) suppose une compétition par interférence dans laquelle chaque espèce empêche la croissance de l'autre. Le modèle se présente comme suit:

\[
\begin{align*}
N'_1 &= r_1 N_1 (1 - N_1/K) \\
N'_2 &= r_2 N_2 (1 - N_2/K)
\end{align*}
\]

May & Leonard(1975) ont considéré un modèle de compétition non transitive entre trois espèces, d'autres auteurs ont étudié la possibilité de coexistence de deux prédateurs en compétition par rapport à une troisième espèce. Le lecteur intéressé pourra consulter la review des recherches récentes sur ce sujet donnée par Grover(1997).

Contrairement aux cas compétition et prédateur-proie, dans le modèle de Mutualisme la présence mutuelle de deux espèces favorise la croissance de chacune d'elles. Une telle symbiose a été notamment décrite en 1874 par le naturaliste Thomas Belt qui a observé un mutualisme formidable entre les fourmis et les arbres acacias. Une formulation du modèle est donnée par le système suivant:

\[
\begin{align*}
N'_1 &= r_1 N_1 (1 - \frac{N_1 - \alpha_{12} N_2}{K_1}) \\
N'_2 &= r_2 N_2 (1 - \frac{N_2 - \alpha_{21} N_1}{K_2})
\end{align*}
\]

5.5 Exercices

Exercice 5.5.1. Supposons que $f$ est une fonction quadratique montrer pour une recherche en ligne exacte on obtient

$$\alpha_k = \left( g^{(k)} \right)^\top g^{(k)}; \alpha_k = -\left( g^{(k)} \right)^\top \rho^{(k)}.$$

Exercice 5.5.2. Démontrer que le système suivant admet une solution unique

$$\begin{cases} \frac{2}{x} e^x + \frac{\sin(x)}{2} + ye^y = 0 \\ \frac{3}{y} e^y - \frac{\cos(y)}{2} = 0 \end{cases}$$

Exercice 5.5.3. Résoudre par la méthode de Newton-Raphson le système

$$\begin{cases} -2x_1 + 4x_2 - 2x_1x_2 = 0 \\ -4x_1 - 5x_2 + 5x_2^2 + x_1x_2 - 39 = 0 \end{cases}$$

Exercice 5.5.4. soit $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ une fonction de classe $C^p$ ayant un point fixe $\alpha$ on note $g(x) = (g_1(x), \cdots, g_n(x))^t$ $op$ et $g_{ij} = \frac{\partial g_j}{\partial x_i}$. Soit $\rho > 0$ et on suppose que pour la boule

$$B = \{x \in \mathbb{R}^n/\|x - \alpha\|_\infty \leq \rho\}$$

on a $|1 - g_{ij}(x)| > \sum_{j \neq i} |g_{ij}(x)|$, pour $1 \leq i \leq n$ et on considère le schéma itératif suivant

$$\begin{cases} x^0 \in B \\ x^{k+1} = DG(x^k) + (I - D)x^k \end{cases}$$

où $D$ est une matrice diagonale inversible.

1. On pose $e^k = x^k - \alpha$, montrer que $e^{k+1} = (I - D + DG_k)e^k$ où $G_k$ est une matrice qui ne dépend que des $g_{ij}$. (On pourra utiliser la formule de Taylor).

2. Trouver une matrice $D$ telle que $x^k \rightarrow \alpha$ (on pourra utiliser la norme $\|\cdot\|_\infty$).

3. Trouver la matrice $D$ qui minimise $\|I - D + DG\|$.

4. Quel est l’intérêt du schéma $(S)$.

Exercice 5.5.5. 1. Soit $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, on considère le schéma de Newton

$$x_{k+1} = x_k - J^{-1}(x_k)F(x_k)$$
(a) En posant $s_k = x_{k+1} - x_k$, écrire le système en $s_k$ qui permet d’éviter d’inverser $J$.

(b) Rappeler l’expression de l’inverse $P^{-1}$ d’une matrice de la forme $P = I - \alpha uv^\top$.

(c) Pour éviter le calcul de $J^{-1}$ à chaque itération, on considère une matrice $B_k$ qui approche $J$ et on utilise l’itération: $x_{k+1} = x_k - B_k^{-1}F(x_k)$.

Montrer que si $B_{k+1} = B_k + uv^\top$ alors $B_{k+1}^{-1} = B_k^{-1} - \frac{B_k^{-1}uv^\top B_k^{-1}}{1 + v^\top B_k^{-1}u}$.

(d) On considère l’algorithme de Broyden qui consiste à prendre :

$$u = y_k - B_k s_k \text{ et } v = \frac{s_k}{s_k^\top s_k}$$

Exprimer $B_{k+1}^{-1}$ en fonction de $B_k^{-1}, s_k$ et $y_k$ et vérifier que : $B_{k+1}^{-1} y_k = s_k$.

2. On suppose maintenant que $f : \mathbb{R}^n \rightarrow \mathbb{R}$ est définie par :

$$f(x) = 1/2 x^\top G x + b^\top x$$

où $x$ et $b$ sont des vecteurs de $\mathbb{R}^n$ et $G$ est une matrice symétrique définie positive.

On note $g(x) = \nabla f(x)$ le gradient de $f$ et pour simplifier on adopte les notations suivantes : $f_k = f(x_k), g_k = g(x_k), y_k = g_{k+1} - g_k$ et $s_k = x_{k+1} - x_k$, on considère alors le schéma suivant:

$$x_{k+1} = x_k + \alpha_k p_k; \alpha_k \in \mathbb{R} \text{ et } p_k \in \mathbb{R}^n$$

tel que

$$p_i^\top G p_j = 0 \text{ pour tout } i \neq j; i, j = 1, \cdots, n$$

$$g_{i+1}^\top p_i = 0 \text{ pour } i = 1, \cdots, n$$

(a) Montrer que : $g_{i+1}^\top p_j = 0 \text{ pour } 1 \leq j \leq i \leq n$.

(b) En déduire que $g(x_{n+1}) = 0$.

(c) Conclure.

**Exercice 5.5.6. (Modèle prédateur-proie)**

On suppose qu’une population de proies $(x_n)$ et une population de prédateurs $(y_n)$ sont amenées à vivre dans le même environnement. En l’absence de prédateurs, la population des proies croît avec le taux $1 + r_1$ alors que l’absence des proies entraîne la décroissance des prédateurs avec un taux $1 - r_2$. 

135
Si les deux populations sont présentes on obtient le système suivant:

\[ X_{n+1} = g(X_n) \]

donné par:

\[ x_{n+1} = (1 + r_1 - b_1 y_n) x_n \]
\[ y_{n+1} = (1 - r_2 + b_2 x_n) y_n \]

1- Chercher les points d'équilibre (points fixes de \( g \))

2- Etudier la convergence de la méthode itérative \( X_{n+1} = g(X_n) \)

3- Donner les conditions de coexistence des deux populations ou de leur extinction.

4- Que signifie le remplacement de la première équation par: \( x_{n+1} = (1 + r_1 - cx_n - b_1 y_n) x_n \)

**Exercice 5.5.7. (Modèle à compétition)**

Dans ce modèle on suppose que l’absence d’une espèce entraine la croissance de l’autre espèce mais la présence des deux espèces entraîne la décroissance de chacune d’elles à cause de la compétition pour les ressources. Le modèle peut-être formulé comme suit

\[ x_{n+1} = (1 + r_1 - a_{11} x_n - a_{12} y_n) x_n \]
\[ y_{n+1} = (1 + r_2 - a_{22} y_n - a_{21} x_n) y_n \]

1- Chercher les points d’équilibre.

2- Etudier la convergence.

3- Donner les conditions qui favorisent la première espèce par rapport à la deuxième.

**Exercice 5.5.8. (Modèle à coopération)**

Dans le modèle de coopération, l’interaction entre les deux espèces joue en leur faveur.

On obtient alors le système d’équations suivantes

\[ x_{n+1} = (1 + r_1 - a_{11} x_n + a_{12} y_n) x_n \]
\[ y_{n+1} = (1 + r_2 - a_{22} y_n + a_{21} x_n) y_n \]

Etudier le comportement écologique des deux espèces en fonction des paramètres du modèle.
Exercice 5.5.9. (Modèle Maynard Smith)

\[ \begin{align*}
N_{n+1} &= N_n + rN_n(1 - N_n/K) - cN_nP_n \\
P_{n+1} &= bN_nP_n
\end{align*} \]

1- En posant \( x_n = \frac{N_n}{K}, y_n = \frac{cP_n}{bK} \) et \( a = bK \), on obtient le système

\[ \begin{align*}
x_{n+1} &= (1 + r)x_n - rx_n^2 - ax_ny_n \\
y_{n+1} &= ax_ny_n
\end{align*} \]

2- Montrer que le nouveau système admet 3 points d’équilibre \( E_1 \) (extinction des deux espèces), \( E_2 \) (extinction de la deuxième espèce) et \( E_3 \) (coexistence des deux espèces).

3- Etudier la convergence locale autour de chaque point d’équilibre.

Problème 5.5.1. 1. Soit \( A \) une matrice carrée d’ordre \( n \) symétrique, définie positive. Deux vecteurs \( a \) et \( v \) sont dits \( A \)-conjugés si \( \langle Av, a \rangle = 0 \).

(a) Montrer que si les vecteurs propres \( v_0, \ldots, v_{n-1} \) sont \( A \)-conjugés, ils forment une base de \( \mathbb{R}^n \).

(b) On définit les deux suites de matrices suivantes:

\[ C_k = \sum_{i=0}^{k-1} \frac{v_i a_i^\top}{\langle Av_i, v_i \rangle}, \quad D_k = I - C_k A. \]

Montrer que

\[ C_k Av_j = v_j \]

\[ D_k v_j = 0 \]

\[ D_k^\top Av_j = 0 \]

Que valent \( D_n \) et \( C_n \).

(c) Supposons que \( v_0, \ldots, v_{k-1} \) soient connues.

Si \( D_k = 0 \) que peut-on conclure.

Sinon, soit \( v \in \mathbb{R}^n \) tel que \( D_k v \neq 0 \), montrer que \( v_k = D_k v \) est \( A \)-conjugé par rapport à \( v_0, \ldots, v_{k-1} \).

(d) Écrire un algorithme qui à partir de \( v_0 \), donné construit la suite \( v_0, \ldots, v_{n-1} \).

On peut considérer un vecteur de la forme \( D_k e_i \) où \( e_i \) est le \( i^{\text{ème}} \) vecteur de la base canonique.

En déduire un algorithme de calcul de \( A^{-1} \).
Chapitre 6

Calcul des valeurs propres et vecteurs propres

6.1 Introduction

Comme on l’a vu dans les chapitres précédents, l’analyse spectrale (calcul des valeurs et vecteurs propres, rayon spectral, · · ·) joue un rôle important quant aux décisions concernant la convergence, la stabilité etc · · · Par ailleurs, de nombreux problèmes (statistique, modélisation, · · ·) se ramènent à l’étude spectrale des matrices ou d’opérateurs en général.

Il est donc nécessaire de savoir calculer numériquement les valeurs et les vecteurs propres dans des situations concrètes impliquant des grands systèmes où les méthodes théoriques sont pratiquement inutilisables.

La stratégie générale pour déterminer les valeurs propres et les vecteurs propres consiste à construire une suite de transformations de similarité jusqu’à l’obtention d’une matrice diagonale, ou plus simplement jusqu’à l’obtention d’une matrice triangulaire à partir de laquelle il est possible de déterminer assez facilement les vecteurs propres.

Pour réaliser ces transformations, deux grandes classes de méthodes sont disponibles : les méthodes directes et les méthodes itératives.

Les premières consistent à effectuer une suite de transformations similaires et ne s’appliquant que sur des matrices de taille modeste, car le temps de calcul croît comme le cube de la dimension linéaire de la matrice.

Pour les matrices de grande taille et généralement creuse, les méthodes itératives sont plus adaptées. Dans la mesure où la plupart du temps, le spectre complet d’une très grande matrice n’est pas recherché mais seulement une partie, les métho-
des itératives peuvent réaliser aisément cette tâche.

Nous allons voir dans la suite de ce chapitre quelques algorithmes de base, tout en ayant à l’esprit qu’il existe une vaste littérature sur ce sujet, et pour un problème particulier il est nécessaire de commencer par analyser précisément le type de matrice dont on souhaite obtenir le spectre, afin de choisir la méthode la plus adaptée pour résoudre ce problème.

6.2 Méthodes basées sur le polynôme caractéristique

Soit \( P_n(\lambda) = \det(A - \lambda I) \)

et \( P(\lambda) = (-1)^n P_n(\lambda) \)

\[ = \lambda^n - a_1 \lambda^{n-1} - \cdots - a_{n-1} \lambda - a_n \]

\[ = \lambda^n - \sum_{k=1}^{n} a_k \lambda^{n-k} \]

les racines de \( P(\lambda) \) sont les valeurs propres de \( A \)

**Exemple 6.2.1.** Méthode de Krylov

Elle consiste à calculer les coefficients du polynôme \( P \) dont les racines sont approchées par les méthodes du chapitre 3.

D’après le théorème de Cayley-Hamilton on a \( P(A) = 0 \) donc \( A^n = \sum_{k=1}^{n} a_k A^{n-k} \).

Soit alors \( x^{(0)} \) un vecteur non nul quelconque et \( a \) le vecteur

\[ a = (a_1, \cdots, a_n)^T, \quad x^{(0)} = \left( x_1^{(0)}, \cdots, x_n^{(0)} \right)^T \]

On a : \( A^n x^{(0)} = \sum_{k=1}^{n} a_k A^{n-k} x^{(0)} \)

Alors, en posant \( x^{(1)} = A x^{(0)} \) et \( x^{(k)} = A x^{(k-1)} = A^k x^{(0)} \), on obtient : \( x^{(n)} = a_1 x^{(n-1)} + \cdots + a_n x^{(0)} \), ou encore

\[
\begin{pmatrix}
  x_1^{(n)} \\
  \vdots \\
  x_i^{(n)} \\
  \vdots \\
  x_n^{(n)}
\end{pmatrix}
= \begin{pmatrix}
  x_1^{(n-1)} & x_1^{(n-2)} & \cdots & x_1^{(1)} & x_1^{(0)} \\
  \vdots & \vdots & \cdots & \vdots & \vdots \\
  x_i^{(n-1)} & x_i^{(n-2)} & \cdots & x_i^{(1)} & x_i^{(0)} \\
  \vdots & \vdots & \cdots & \vdots & \vdots \\
  x_n^{(n-1)} & x_n^{(n-2)} & \cdots & x_n^{(1)} & x_n^{(0)}
\end{pmatrix}
\begin{pmatrix}
  a_1 \\
  \vdots \\
  a_i \\
  \vdots \\
  a_n
\end{pmatrix}
\]

(6.2.1)
Soit en condensant

\[ Ba = A^n x^{(0)}. \]

La détermination des coefficients \( a_i \) du polynôme \( P \) revient donc à la résolution d’un système linéaire. Si la matrice \( B \) est inversible on obtient une solution unique, sinon on change le \( x^{(0)} \). Cette méthode permet aussi de calculer les vecteurs propres en écrivant \( x^{(0)} \) sur la base de ces vecteurs propres \( x^{(0)} = \sum \alpha_i v_i \).

**Exemple 6.2.2. Méthode de Danilevski**

Soient \( A \) et \( P \) les matrices réelles carrées d’ordre \( n \).

\[
A = \begin{pmatrix}
a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
a_{n1} & \cdots & \cdots & a_{nn}
\end{pmatrix}
\]

et

\[
P = \begin{pmatrix}
p_1 & p_2 & \cdots & \cdots & p_n \\
1 & 0 & \cdots & \cdots & 0 \\
0 & 1 & 0 & \cdots & \cdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 1 & 0
\end{pmatrix}.
\]

La méthode de Danilevski consiste à transformer la matrice carrée quelconque \( A \) en une matrice de Frobenius de type \( P \) en cherchant une matrice \( S \) telle que

\[ P = S^{-1} AS. \]

On effectue \( n-1 \) réductions transformant successivement les lignes de la matrice \( A \) en lignes respectives de \( P \) en commençant par la dernière ligne.

**étape 1** On cherche à transformer la dernière ligne de \( A, (a_{n1}, \cdots, a_{nn}) \) pour obtenir la dernière ligne de \( P, \left( 0, \cdots, 1, 0 \right) \).

Ceci peut-être réalisé si on suppose que \( a_{nn-1} \neq 0 \) puis en divisant tous les termes de la \((n-1)\)ème colonne de \( A \) par \( a_{nn-1} \) et en retranchant la \((n-1)\)ème colonne transformée multipliée respectivement par les nombres \( a_{n1}, a_{n2}, \cdots, a_{nn} \) de toutes les colonnes de \( A \).

Matriciellement, cette étape est équivalente au produit matriciel \( AM_{n-1} \) où la matrice élémentaire \( M_{n-1} \) est donnée par

\[
M_{n-1} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & 0 \\
0 & 1 & \cdots & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
m_{n-11} & m_{n-12} & \cdots & \cdots & m_{n-1n} \\
0 & \cdots & \cdots & 0 & 1
\end{pmatrix},
\]

où \( m_{n-1n-1} = \frac{1}{a_{nn-1}} \) et \( m_{n-1i} = -\frac{a_{ni}}{a_{nn-1}} \) pour \( i \neq n-1 \).
On vérifie que la matrice inverse $M_{n-1}^{-1}$ a la même forme que $M_{n-1}$ et qu'elle est donnée par

\[
M_{n-1}^{-1} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & 0 \\
0 & 1 & \cdots & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
a_{n-11} & a_{n-12} & \cdots & a_{n-1n} & 0 \\
0 & \cdots & \cdots & 0 & 1
\end{pmatrix}
\]

On aboutit donc à une matrice semblable à $A$ et dont la dernière ligne est celle de la matrice de Frobenius soit $C = M_{n-1}^{-1}AM_{n-1}$ et comme $C$ est obtenu à partir de $B$ par le produit $M_{n-1}^{-1}B$, il est facile de voir que tous éléments de $C$ sont égaux aux éléments de $B$ sauf ceux de la $(n-1)^{\text{ème}}$ ligne qui sont donnés par

\[
c_{n-1j} = \sum_{k=1}^{n} a_{nk} b_{kj} \quad \text{pour} \quad j = 1, \cdots, n.
\]

étape 2 Si $c_{n-1n-2} \neq 0$ on reprend les mêmes opérations de l’étape 1 avec la matrice $C$ au lieu de $A$.

Si toutes les étapes sont exécutées on aboutit à la matrice de Frobenius

\[
P = M_1^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} AM_{n-1} M_{n-2} \cdots M_1.
\]

La méthode de Danilevski permet aussi de calculer les vecteurs propres d’une matrice $A$ si ses valeurs propres sont connues.

En effet si $\lambda$ est une valeur propre de $A$, elle est aussi valeur propre de la matrice de Frobenius semblable à $A$ Soit donc $v = (v_1, v_2, \cdots, v_n)^\top$ un vecteur propre de $P$ associé à la valeur propre $\lambda$, alors en écrivant $Pv = \lambda v$ et en identifiant composante à composante, on obtient le système

\[
p_1 v_1 + p_2 v_2 + \cdots + p_n v_n = \lambda v_1 \\
v_1 = \lambda v_2 \\
v_2 = \lambda v_3 \\
\vdots & \vdots & \vdots \\
v_{n-1} = \lambda v_n.
\]

141
6.3 Méthodes itératives

Méthode de la puissance itérée

Soient \( \lambda_1, \lambda_2, \ldots, \lambda_n \) les valeurs propres de \( A \). On suppose que \(|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|\). Tout vecteur \( x^{(0)} \) peut s'écrire sur la base des vecteurs propres \( x^{(0)} = \sum_{i=1}^{n} \alpha_i v_i \).

On considère alors les deux suites \( (x^{(k)}) \) et \( m_k \): \( x^{(k)} = \frac{A x^{(k-1)}}{m_k} \) où \( m_k \) est l'élément de \( A x^{(k-1)} \) le plus grand en valeur absolue, \( m_k = \|A x^{(k-1)}\|_\infty \) de telle sorte que 1 est la plus grande composante de \( x^{(k)} \).

**Théorème 6.3.1.** Dans ces conditions, les suites \( m_k \) et \( (x^{(k)}) \) convergent vers \( \lambda_1 \) et \( v_1 \), respectivement, où \( v_1 \) est le vecteur propre associé à \( \lambda_1 \).

**Preuve :**

soit \( x^{(0)} = \sum_{i=1}^{n} \alpha_i v_i \), \( A^k x^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k v_i = \lambda_1^k \left[ \alpha_1 v_1 + \sum_{i=1}^{n} \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k v_i \right] \)

d'où

\[
x^{(k)} = \frac{A x^{(k-1)}}{m_k} = \frac{1}{m_k} \frac{1}{m_{k-1}} A^2 x^{(k-2)} = \frac{A^k x^{(0)}}{\prod_{i=1}^{k} m_i} = \frac{1}{C_k} A^k x^{(0)} \quad \text{avec} \quad C_k = \prod_{i=1}^{k} m_i
\]

Par suite \( x^{(k)} = \frac{\lambda_1^k}{C_k} \left[ \alpha_1 v_1 + \sum_{i=1}^{n} \alpha_i \left( \frac{\lambda_i}{\lambda_1} \right)^k v_i \right] \) donc lorsque \( k \) tend vers l'infini \( x^{(k)} \) tend vers \( \frac{\lambda_1^k}{C_k} \alpha_1 v_1 \). Par identification, on obtient \( \lim_{k \to \infty} \frac{\alpha_1 \lambda_1^k}{C_k} = 1 \) et donc \( \lim_{k \to \infty} x^{(k)} = v_1 \).

Comme \( \lim_{k \to \infty} A x^{(k)} = A v_1 = \lambda_1 v_1 \) on a \( \lim_{k \to \infty} x^{(k)} = \lim_{k \to \infty} \frac{A x^{(k-1)}}{m_k} = \frac{\lambda_1 v_1}{\lim_{k \to \infty} m_k} \) d'où \( \lim_{k \to \infty} m_k = \lambda_1 \).

**Exemple 6.3.1.** \( A = \begin{pmatrix} 4 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 4 \end{pmatrix} \) \( x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \), \( A x^{(0)} = \begin{pmatrix} 6 \\ 0 \\ 4 \end{pmatrix} \) \( m_1 = 6 \) donc.
\[ x^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 2/3 \end{pmatrix}, \quad \text{et} \quad Ax^{(1)} = \begin{pmatrix} 14/3 \\ 1/3 \\ 11/3 \end{pmatrix}, \quad m_2 = 14/3, \quad x^{(2)} = \begin{pmatrix} 1 \\ 1/14 \\ 11/14 \end{pmatrix}, \ldots \]

**Remarque 6.3.1.**

- Si \( \lambda_1 \) est de multiplicité \( r \) alors on écrit

\[
x^{(0)} = \sum_{i=1}^{r} \alpha_i v_i + \sum_{i=r+1}^{n} \alpha_i v_i \quad \text{et} \quad A^k x^{(0)} = \sum_{i=1}^{r} \alpha_i \lambda_1 v_i + \cdots = \lambda_1 u + \cdots \quad \text{avec} \quad u = \sum_{i=1}^{r} \alpha_i v_i.
\]

- Si les valeurs propres ne sont pas bien séparées, i.e \( \frac{\lambda_1}{\lambda_2} \simeq 1 \), la convergence sera très lente.

**Méthode LR**

Soit \( A \) une matrice carrée d’ordre \( n \), la méthode LR est basée sur le procédé suivant: \( A_1 = A = LR \) où \( L \) est triangulaire inférieure avec \( l_{ii} = 1 \) et \( U \) triangulaire supérieure, \( A_i = L_i R_i, A_{i+1} = R_i L_i \). On suppose que toutes les décompositions \( L_i R_i \) existent. Alors on a:

**Théorème 6.3.2.**

1. \( A_{i+1} \) est semblable à \( A_i \).
2. \( A_{i+1} = T_i^{-1} A_i T_i \) avec \( T_i = L_1 L_2 \cdots L_i \).
3. \( A^i = T_i U_i \) avec \( U_i = R_i R_{i+1} \cdots R_1 \).

**Preuve:**

1. \( L_i \) étant inversible, il suffit de multiplier \( A_i \) par \( L_i^{-1} \) à gauche et par \( L_i \) à droite pour obtenir \( L_i^{-1} A_i L_i = R_i L_i = A_{i+1} \).

2. \( A_{i+1} = L_i^{-1} A_i L_i = R_i L_i \) est une conséquence directe de 1. Puisqu’on a

\[
A_2 = L_1^{-1} A_1 L_1, \quad A_3 = L_2^{-1} (L_1^{-1} A_1 L_1) L_2 \quad \text{etc} \cdots
\]

3. on en tire \( T_i A_{i+1} = A_i T_i \) et en écrivant \( T_i U_i \), il vient

\[
T_i U_i = L_1 \cdots L_{i-1} (L_i R_i) R_{i-1} \cdots R_1
= A_1 L_1 \cdots (L_{i-1} R_{i-1}) \cdots R_1
= \cdots
= A_i^{i-1} T_i U_1
= A_i^{i-1} T_i R_1 = A_i = A^i
\]

143
Théorème 6.3.3. *On suppose que les valeurs propres de $A$ vérifient: $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ et que la matrice de passage $P (P^{-1}AP = D)$ ainsi que $P^{-1}$ admettent les décompositions $P = L_P R_P$ et $P^{-1} = Q = L_Q R_Q$, $(L_P)_{ii} = 1$ et $(L_Q)_{ii} = 1$. Alors les suites de matrices $(A_i)$, $(L_i)$ et $(R_i)$ convergent et on a:

$$
\lim_{i \to \infty} A_i = \lim_{i \to \infty} R_i = \begin{pmatrix}
\lambda_1 & * & * \\
0 & \ddots & * \\
0 & 0 & \lambda_n
\end{pmatrix}
\quad \text{et} \quad
\lim_{i \to \infty} L_i = I.
$$

**Preuve:**
Soient $A = PDQ$ et $A^i = PD^iQ = L_P R_P D^i L_Q R_Q$ qui s’écrit $A = L_P R_P (D^i L_Q D^{-i}) D^i R_Q$.

Comme $D^i L_Q D^{-i} = (l^i_{jk}) = \left(\frac{\lambda_j}{\lambda_k}\right)^i l_{jk}$ avec $|\lambda_j| < |\lambda_k|$ pour $j \geq k$, il s’ensuit que

$$
\lim_{i \to \infty} l^i_{jk} = 0 \text{ pour } j \neq k.
$$

On peut donc écrire $D^i L_Q D^{-i} = I + E_i$ avec $\lim_{i \to \infty} E_i = 0$.

Maintenant on remplace $D^i L_Q D^{-i}$ par $I + E_i$ dans l’expression de $A^i$, on obtient:

$$
A^i = L_P R_P (I + E_i) D^i R_Q
\quad =
L_P \left( I + R_P E_i R_P^{-1} \right) R_P D^i R_Q
\quad =
L_P (I + F_i) R_P D^i R_Q \text{ en posant } F_i = R_P E^i R_P^{-1}
$$

Comme $\lim_{i \to \infty} F_i = 0$, la matrice $I + F_i$ admet bien une décomposition $I + F_i = \tilde{L}_i \tilde{R}_i$ pour $i$ assez grand, avec $\tilde{L}_i = \begin{pmatrix} 1 & * & \cdots & * \\
& \ddots & \ddots & \ddots \\
& & 1 & \end{pmatrix}$ et $\tilde{R}_i$ triangulaire supérieure. De plus on a

$$
\lim_{i \to \infty} \tilde{L}_i = \lim_{i \to \infty} \tilde{R}_i = I.
$$

D’après le théorème 6.3.2 on a: $A^i = T_i U_i = L_P \tilde{L}_i \tilde{R}_i R_P D^i R_Q$ et par unicité de décomposition on tire : $T_i = L_P \tilde{L}_i$ et $U_i = \tilde{R}_i R_P D^i R_Q$.

Finalement, $\tilde{L}_i$ converge vers $I$ ce qui implique que $T_i$ converge vers $L_P$ et $\tilde{R}_i$ converge vers $I$ donc $U_i$ converge vers $\lim_i R_P D^i R_Q$, d’où

$$
\lim_i L_i = \lim_i T_{i-1}^{-1} T_i = I
\quad \text{et}
\lim_i R_i = \lim_i U_i U_i^{-1} = R_P D R_P^{-1}
$$

$$
\lim_{i \to \infty} A_i = \lim_{i \to \infty} L_i R_i = R_P D R_P^{-1} = \begin{pmatrix}
\lambda_1 & \ast & \cdots \\
& \ddots & \ast \cdots \\
& & \lambda_n
\end{pmatrix}
$$

144
Méthode QR

Considérons une matrice carrée d’ordre \( n \) \((A = A_1)\), des matrices \( Q_i \) unitaires et des matrices \( R_i \) triangulaires et posons:

\[
A_i = Q_i R_i
\]

et \( A_{i+1} = R_i Q_i \) (\( Q_i R_i \) est la décomposition \( Q_i R_i \) de \( A_i \) avec \( Q_i \) unitaire). On a alors :

**Lemme 6.3.1.**

i) \( A_{i+1} = (Q_1 \cdots Q_i)^H A_1 (Q_1 \cdots Q_i) \),

ii) \( Q_1 \cdots Q_i R_i \cdots R_3 = A_1^i \),

iii) la décomposition QR de \( A_1^i \) est unique si \( R \) est triangulaire supérieure avec diagonale strictement positive.

**Preuve:** voir exercice(6.5.1).

**Théorème 6.3.4.**

\[
\lim_{i \to \infty} A_i = T
\]

où \( T \) est triangulaire supérieure contenant en diagonale les valeurs propres \( \lambda_j \) rangées par ordre décroissant en module.

**Preuve:** Supposons que \(|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|\).

Alors:

1. il existe une matrice \( P \) telle que:

\[
P^{-1} A_1 P = D = diag(\lambda_j)
\]

2. En posant \((Q_1 \cdots Q_i) = Q_i\), on a d’après 1- et le lemme 6.3.1:

\[
A_{i+1} = Q_i^H P D P^{-1} Q_i = Q_i^H Q_i R_i P D R_i^P P^{-1} Q_i^H Q_i
\]

où \( P = Q_i R_i \) est la décomposition QR de \( P \).

Le reste de la démonstration est similaire à celle du théorème 6.3.3.
Méthode de Jacobi

Considérons la matrice élémentaire

\[ \Omega = \begin{pmatrix} 1 & 0 & 1 & \cdots & c & s & \cdots & p \\ 0 & 1 & \cdots & s & c & \cdots & q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \cdots & 1 & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \]

avec \( c = \cos \theta \), \( s = \sin \theta \) et \( k \neq l \)

\( \Omega \) est une matrice orthogonale \( \Omega^\top \Omega = I \).

**Théorème 6.3.5.** Soit \( \theta \) un nombre réel et \( p \) et \( q \) tels que \( 1 \leq p < q \leq n \) deux entiers aux quels on associe une matrice élémentaire \( \Omega \)

1. Si \( A \) est une matrice symétrique alors la matrice \( B = \Omega^\top A \Omega \) est également symétrique et vérifie \( \sum_{i,j} b_{ij}^2 = \sum_{i,j} a_{ij}^2 \) soit encore \( \|B\|_E^2 = \|A\|_E^2 \).

2. Si \( a_{pq} \neq 0 \), il existe une seule valeur de \( \theta \) dans \( I = ] -\frac{\pi}{4}, 0[ \cup ] 0, \frac{\pi}{4} [ \) telle que \( b_{pq} = 0 \), c'est la solution dans 1 de l'équation:

\[ \cotg(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}}. \]

\( \theta \) ainsi choisi, on obtient

\[ \sum_{i=1}^n b_{ii}^2 = \sum_{i=1}^n a_{ii}^2 + 2a_{pq}^2. \]

**Preuve:**

1. \( \|B\|_E^2 = \|\Omega^\top A \Omega\|_E^2 = \|A\|_E^2 = tr(A^\top A) = tr(A^2). \)

2. De ce qui précède et de fait que le produit \( \Omega^\top A \Omega \) touche uniquement les lignes et les colonnes \( p \) et \( q \) on obtient \( a_{pp}^2 + 2a_{pq}^2 + a_{qq}^2 = b_{pp}^2 + 2b_{pq}^2 + b_{qq}^2 \) et \( b_{pq} = a_{pq} \cos(2\theta) + \frac{a_{qq} - a_{pp}}{2} \sin(2\theta) \), d'où \( b_{qp} = 0 \) si \( \theta \) est tel que \( \cotg(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}} \), pour cette valeur de \( \theta \) il est évident que \( \sum_{i=1}^n b_{ii}^2 = \sum_{i=1}^n a_{ii}^2 + 2a_{pq}^2. \)

Dans la méthode de Jacobi classique, on choisit \( a_{pq}^{(k)} \) tel que \( \left| a_{pq}^{(k)} \right| = \max_{i,j} \left| a_{ij}^{(k)} \right| \).
Méthode de Jacobi classique

On note maintenant $\Omega_k$ la matrice $\Omega$ de l’étape $k$ (pour $p$ et $q$ choisi tel que $|a_{pq}^{(k)}| = \max_{i,j} |a_{ij}^{(k)}|$. La matrice $A_{k+1}$ est donnée par $A_{k+1} = \Omega^\top A_k \Omega$ et on a

**Théorème 6.3.6.** La suite $(A_{k+1})$ des matrices obtenues par la méthode de Jacobi classique converge vers diag$(\lambda_{\sigma(i)})$ où $\sigma$ est une permutation convenable.

Pour la preuve de ce théorème, on se sert de trois lemmes.

**Lemme 6.3.2.** Soit $A_k$ décomposée sous la forme $A_k = D_k + E_k$ avec $D_k = \text{diag} \left( a_{ii}^{(k)} \right)$, alors $\lim_{k \to \infty} E_k = 0$.

**Preuve:**
Posons $\varepsilon_k = \sum_{i \neq j} \left( a_{ij}^{(k)} \right)^2 = \|E_k\|^2_E$ d’après le théorème 6.3.5 on a: $\varepsilon_{k+1} = \varepsilon_k - 2 \left( a_{pq}^{(k)} \right)^2 \sum_{i \neq j} \left( a_{ij}^{(k+1)} \right)^2 = \varepsilon_{k+1} + \sum_{i = 1}^{n} \left( a_{ii}^{(k+1)} \right)^2 = \varepsilon_k + \sum_{i = 1}^{n} \left( a_{ii}^{(k)} \right)^2$.
Comme on a supposé que $\left| a_{pq}^{(k)} \right| = \max_{i,j} \left| a_{ij}^{(k)} \right|$ on a :

$\varepsilon_k = \sum_{i \neq j} \left( a_{ij}^{(k)} \right)^2 \leq \sum_{i \neq j} \left( a_{pq}^{(k)} \right)^2 = \left( n^2 - n \right) \left( a_{pq}^{(k)} \right)^2$ d’où $\varepsilon_{k+1} \leq \left( 1 - \frac{2}{n(n-1)} \right) \varepsilon_k$
i.e $\varepsilon_{k+1} \leq \alpha \varepsilon_k$ avec $0 < \alpha < 1$ ce qui entraîne que $\lim_{k \to \infty} \varepsilon_k = 0$ et donc $\lim_{k \to \infty} \|E_k\| = 0$.

**Lemme 6.3.3.** Soit $E$ un espace vectoriel de dimension finie et $(x_k)$ une suite bornée de $E$, admettant un nombre fini de valeurs d’adhérence et telle que: $\lim_{k \to \infty} \|x_{k+1} - x_k\| = 0$.
Alors la suite $(x_k)$ converge vers l’une des valeurs d’adhérence.

**Lemme 6.3.4.** La suite $(D_k)$ vérifie les hypothèses du lemme 6.3.3, par conséquent $(D_k)$ converge vers une des valeurs d’adhérence, qui est de la forme diag$(\lambda_{\sigma(i)})$.

**Preuve:**
Puisque $D_{k+1} - D_k = \text{diag} \left( a_{ii}^{(k+1)} - a_{ii}^{(k)} \right)$ alors on a

$a_{ii}^{(k+1)} - a_{ii}^{(k)} = \begin{cases} 0 & \text{si } i \neq p \land i \neq q \\ \pm \tan(\theta_k) a_{pq}^{(k)} & \text{si } i = p \text{ ou } i = q. \end{cases}$

avec $|\theta_k| < \frac{\pi}{4}$,

$\left| a_{pq}^{(k)} \right| \leq \|E_k\|$ d’où $\lim_{k} \left( a_{ii}^{(k+1)} - a_{ii}^{(k)} \right) = 0$ puisque $\|D_k\|_E \leq \|A_k\|_E = \|A_k\|_E$ est bornée. $D_k$ n’a qu’un nombre fini des valeurs d’adhérence, qui sont nécessairement de la forme diag$(\lambda_{\sigma(i)})$.

$\lambda_1 \cdot \cdot \cdot \lambda_n$ supposées rangées dans un certain ordre.
en effet:
si $(D_{k'})$ est une suite extraite de $(D_k)$ et telle que $\lim_{k'} D_{k'} = D$, alors
\[ \lim_{k'} A_{k'} = \lim_{k'} (D_{k'} + B_{k'}) = D \]
\[ \det (D - \lambda I) = \lim_{k' \to \infty} \det (A_{k'} - \lambda I) = \det (A - \lambda I). \]

Donc $A$ et $D$ ont mêmes valeurs propres. Comme $D$ est diagonale, elle contient les $\lambda_i$ en diagonale.
Finalement les lemmes 6.3.2, 6.3.3 et 6.3.4 entraînent que
\[ \lim_k A_k = \lim_k D_k = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \]

**Théorème 6.3.7.** En posant $A_{k+1} = \Omega_k^T A_k \Omega_k = \Omega_k^T \Omega_{k-1} \cdots \Omega_1^T A \Omega_2 \cdots \Omega_k = H_k^T A H_k$
et en supposant que toutes les valeurs propres de $A$ sont distinctes, alors la suite des matrices $(H_k)$ construite par la méthode de Jacobi classique, converge vers une matrice $H$
orthogonale dont les colonnes constituent une base orthogonale formée des vecteurs propres de $A$.

**Preuve:** Voir Ciarlet(1984).
6.4 Applications

Notions de persistance et d’extinction

Définition 6.4.1. Une population sera dite persistante si \( \lim_{t \to +\infty} n(t) > 0 \) pour toute valeur initiale \( n(0) \) et elle tendra vers l’extinction si \( \lim_{t \to +\infty} n(t) = 0 \) pour un certain \( n(0) \).

Soit un modèle du type \( n(t + 1) = L(n(t))n(t) \).

Théorème 6.4.1. Soit \( L \) une matrice à coefficients constants non négatifs et régulière avec valeur propre dominante \( \lambda \)

- Si \( \lim_{t \to +\infty} L(n) = L \) et \( \lambda > 1 \) alors la population est persistante.
- Si \( L(n) \leq L \) pour \( t \geq t_1 > 0 \) et \( \lambda < 1 \) alors la population tend vers l’extinction.

Preuve: Voir ([41]).

Remarques 6.4.1. 1. Les résultats du théorème sont évidents puisque le taux de croissance mesuré par la valeur propre dominante de \( L \) croît pour la persistance et décroît pour l’extinction.

2. Comme conséquence du théorème de persistance et extinction on définit un paramètre critique de persistance et extinction \( N_c \).

Supposons que \( \lambda(n, p) \) est une fonction continue et monotone pour \( n \in \mathbb{N} \) où \( n \) désigne par exemple le nombre de régions et \( p \) un vecteur de paramètres du modèle, \( p \in \mathbb{R}_+^l \).

Le nombre critique de régions est défini comme étant le plus petit entier \( N_c \) s’il existe satisfaisant

i) \( \lambda(n, p) > 1 \) et \( \lambda(n + 1, p) < 1 \),

ou

ii) \( \lambda(n, p) > 1 \) et \( \lambda(n - 1, p) < 1 \) et \( n > 1 \).

\( N_c \) a deux interprétations différentes:

Dans le cas i) il est le nombre maximal pour la persistance donc la population est persistante dans \( n \) régions telle que \( n \leq N_c \), dans le cas ii) la population est persistante si \( n \geq N_c \).

Application 6.4.1. ([41])

On suppose que \( \lim L(n) = L \times D_k \), où \( L \) est une matrice de Leslie \( m \times m \) et \( D \) est une matrice de diffusion \( n \times n \).
Exemple 6.4.1. (Modèle “Island”)

\[
D = \begin{pmatrix}
1 - (n-1)d_1 & d_2 & \cdots & d_2 \\
d_2 & 1 - (n-1)d_1 & \cdots & d_2 \\
\vdots & \vdots & \ddots & \vdots \\
d_2 & \cdots & \cdots & 1 - (n-1)d_1
\end{pmatrix}
\]

La valeur propre dominante \(\lambda\) de \(L \times D^k\) est \(\lambda_1 \times \lambda_D^k\) où \(\lambda_D\) et \(\lambda_1\) sont les valeurs propres dominantes de \(D\) et \(L\) respectivement.

La matrice \(D\) est une matrice circulaire et \(\lambda_D = 1 + (n-1)(d_2 - d_1)\).

- Si \(d_2 > d_1\) alors \(\lambda > 1\), la population est persistante pour tout \(n\).
- Si \(d_2 < d_1\), \(\lambda\) est décroissante en fonction de \(n\) en résolvant l’équation \(\lambda(N_c, p) = 1\)

on obtient

\[N_c = E \left(1 + \frac{1-d}{d_1-d_2}\right)\]

avec

\[d = \left(\frac{1}{\lambda_1}\right)^{1/k}\]

\(N_c\) est le nombre maximal des régions qui assure la persistance puisque \(\lambda(n, p)\) est décroissante.

- Si \(n < N_c\), la population persiste.
- Si \(n > N_c\), la population tend vers l’extinction.

Exemple 6.4.2. (Modèle ”stepping Stone”)
Dans ce type de modèle le mouvement est seulement entre les régions voisines, \(D\) est la de forme

\[
\begin{pmatrix}
a & d_2 & \cdots & 0 \\
d_2 & a & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & d_2 & a
\end{pmatrix}
\]

avec \(a = 1 - 2d_1\) et \(0 < d_1, d_2 < \frac{1}{2}\), \(\lambda_D = 1 - 2d_1 - 2d_2 \cos\left(\frac{\pi N}{N+1}\right)\).

\[\lambda(n, p) = \lambda_1 \left(1 - 2d_1 - 2d_2 \cos\left(\frac{\pi N}{N+1}\right)\right)^k\]
La valeur critique $N_c$ vérifie l’équation
\[
\lambda_1 \left( 1 - 2d_1 - 2d_2 \cos \left( \frac{\pi N}{N+1} \right) \right)^k = 1.
\]
Cette équation n’admet de solution que si
\[
1 - b \leq 2(d_1 + d_2)
\]
avec
\[
b = \left( \frac{1}{\lambda_1} \right)^{1/k}.
\]
\[
N_c = E \left( \frac{\arccos \left( 1 - 2d_1 - \frac{b}{2d_2} \right)}{\pi - \arccos \left( 1 - 2d_1 - \frac{b}{2d_2} \right)} \right) + 1.
\]

$N_c$ est la valeur minimale qui assure la persistance.

- Si $n > N_c$ la population est persistante.
- Si $n < N_c$ la population tend vers l’extinction.
6.5 Exercices

Exercice 6.5.1. 1. Rappelez la méthode de Householder permettant d’obtenir une matrice réelle orthogonale $P$ telle que:

$$Px = ke_1$$

où $e_1$ représente la première colonne de la matrice identité, $x$ est un vecteur à $n$ composantes et $k$ un scalaire convenablement choisi.

2. On définit la méthode QR pour une matrice symétrique $A$ comme suit:

$$Q_i R_i = A_i - p_i I;$$

$$A_{i+1} = R_i Q_i + p_i I, \quad i = 1, 2, \cdots.$$ 

où $Q_i$ est une matrice réelle orthogonale, $R_i$ une matrice triangulaire supérieure et $p_i$ un paramètre de translation (shift).

(a) Montrer que $A_{i+1} = (Q_1 \cdots Q_i)^T A_1 (Q_1 \cdots Q_i)$.

(b) déduire que: $Q_1 \cdots Q_i R_i \cdots R_1 = \prod_{k=i}^{k=n} (A_1 - p_k I)$.

Exercice 6.5.2. Soit $A$ une matrice symétrique ayant les valeurs propres $(\lambda_i)_{i=1}^n$ associées aux vecteurs propres $(u_i)_{i=1}^n$.

On pose $\rho(\mu, x) = Ax - \mu x \quad \forall x \in \mathbb{R}^n$ ($\rho = 0$ si $x$ est vecteur propre associé à $\mu$)

1. Montrer que le spectre de $A$ contient au moins une valeur propre $\lambda_i$ telle que

$$|\lambda_i - \mu| \leq ||\rho(\mu, x)||.$$ 

2. Soit $\lambda_i$ une valeur propre vérifiant les deux conditions

i) $\min_{j \neq i} |\lambda_i - \lambda_j| = d > 0,$

ii) $|\lambda_i - \mu| \leq ||\rho(\mu, x)||.$

Montrer qu’il existe une constante $\alpha$ telle que $||x - \alpha u_i|| \leq \frac{||\rho(\mu, x)||}{d - ||\rho(\mu, x)||}.$

Exercice 6.5.3. Soit $A$ une matrice dont les vecteurs propres $v_1, \cdots, v_n$ sont associés aux valeurs propres $\lambda_1, \cdots, \lambda_n$ vérifiant

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n$$
1. On suppose que la valeur propre dominante $\lambda_1$ et le vecteur propre associé $v_1$ aient été déterminés et soit

$$A_1 = A - \lambda_1 \frac{v_1 v_1^\top}{v_1^\top v_1}$$

(a) Montrer que $\lambda_2$ est la valeur propre dominante de $A_1$.

(b) Appliquer la méthode des puissances itérées à $A_1$ pour calculer $\lambda_2$ et le vecteur propre associé $v_2$.

(c) Conclure.

2. On suppose toujours que $\lambda_1$ et $v_1$ aient été déterminés.

(a) Soit $x$ un vecteur tel que $x^\top v_1 = 1$ et $A_1 = A - \lambda_1 v_1 x^\top$ déterminer les valeurs propres de $A_1$.

(b) Montrer que les vecteurs propres de $A_1$ sont donnés par

$$w_1 = v_1, \quad w_i = v_i - \frac{\lambda_1}{\lambda_i} v_1 x^\top v_i \quad i \neq 1.$$  

(c) Montrer que

$$v_i = w_i + \frac{\lambda_1}{\lambda_i - \lambda_1} x^\top w_i v_1 \quad i \neq 1.$$  

(d) On prend $x^\top = \left( \frac{a_{i1}}{\lambda_1 v_{i1}}, \frac{a_{i2}}{\lambda_1 v_{i1}}, \cdots, \frac{a_{in}}{\lambda_1 v_{i1}} \right)^\top$, $(a_{i1}, a_{i2}, \cdots, a_{in})^\top$ est la $i_{\text{ème}}$ ligne de $A$ et $v_{i1}$ est la $i_{\text{ème}}$ composante de $v_1$. Montrer que la $i_{\text{ème}}$ composante de $w_j$, $j = 2, \cdots, n$ est nulle.

(e) Soit $A_1'$ la matrice obtenue de $A_1$ en supprimant la $i_{\text{ème}}$ ligne et la $i_{\text{ème}}$ colonne de $A$, donner la valeur propre dominante de $A_1'$. Peut on appliquer la méthode des puissances itérées pour déterminer $\lambda_2$ et le vecteur propre associé $w_2'$.

(f) Déduire le vecteur propre $w_2$ de la matrice $A_1$ associé à $\lambda_2$.

(g) Montrer que

$$v_2 = \frac{(\lambda_2 - \lambda_1)w_2 + \lambda_1 x^\top w_2 v_1}{\lambda_2 - \lambda_1}.$$  

(h) Conclure.
Chapitre 7

Analyse numérique des équations différentielles ordinaires (e.d.o)

7.1 Rappels sur les équations différentielles ordinaires (e.d.o)

On considère l'e.d.o du premier ordre
\[ y'(x) = f(x, y), \quad f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m, x \in \mathbb{R}, \quad y \in \mathbb{R}^m \quad \text{où} \quad y = (y_1, y_2, \ldots, y_m)^\top \]

(7.1.1)

L'équation (7.1.1) peut encore s'écrire sous la forme d'un système d'e.d.o :
\[
\begin{align*}
    y'_1 &= f_1(x, y_1, \ldots, y_m) \\
    y'_2 &= f_2(x, y_1, \ldots, y_m) \\
    &\vdots \\
    y'_m &= f_m(x, y_1, \ldots, y_m)
\end{align*}
\]

(7.1.2)

Lorsque les conditions initiales sont précisées, on obtient un problème de condition initiale (p.c.i) encore appelé problème de Cauchy :
\[
\begin{align*}
    y'(x) &= f(x, y) \\
    y(a) &= \alpha \quad \text{avec} \quad \alpha = (\alpha_1, \ldots, \alpha_m)^\top \quad \text{donné}
\end{align*}
\]

(7.1.3)

L'existence et l'unicité de la solution du p.c.i (7.1.3) sont données par le théorème suivant

Théorème 7.1.1. Soit \( f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m \) une fonction définie et continue pour tout couple \((x, y) \in D\) où \( D = \{(x, y); a \leq x \leq b, -\infty < y_i < \infty\}\), avec \(a\) et \(b\) finis. On suppose qu'il existe une constante \( L \) telle que :
\[
\|f(x, y) - f(x, y^*)\| \leq L\|y - y^*\| \quad \text{pour tous} \quad (x, y) \quad \text{et} \quad (x, y^*) \quad \text{appartenant à} \quad D
\]

(7.1.4)
Alors pour tout $\alpha \in \mathbb{R}^m$, il existe une solution unique $y(x)$ du problème (7.1.3), où $y$ est continue differentiable pour tout couple $(x, y) \in D$.

**Remarque 7.1.1.** Un système différentiel d'ordre $q$ peut toujours être ramené à un système du premier ordre du type (7.1.3)

$$y^{(q)} = \varphi \left(x, y^{(0)}, \cdots, y^{(q-1)}\right); \varphi : \mathbb{R} \times \mathbb{R}^m \times \cdots \times \mathbb{R}^m \rightarrow \mathbb{R}^m$$

En posant: $Y_1 = y, Y_2 = Y'_1, \cdots, Y_q = Y'_{q-1} = y^{(q-1)}$ on obtient

$Y' = F(x, Y)$ avec $Y = \left(Y^T_1, Y^T_2, \cdots, Y^T_q\right)^T \in \mathbb{R}^{qm}$

$$F = \left(Y^T_2, Y^T_3, \cdots, Y^T_q, \varphi^T\right)^T \in \mathbb{R}^{qm}$$

où $y^{(r)}(a) = \alpha_{r+1}, r = 0, 1, \cdots, q - 1$.

On obtient

$$Y' = F(x, Y); Y(a) = \alpha$$

**Exemple 7.1.1.**

$$y^{(iv)} = f(x, y), \quad f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$$

$Y_1 = y, Y_2 = y', Y_3 = y'', Y_4 = y'''$ si on pose $Y = (Y_1, Y_2, \cdots, Y_4)^T$ on obtient $Y = F(x, Y)$.

### 7.2 Systèmes linéaires

Le système

$$y'(x) = f(x, y), \quad f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$$

est dit linéaire si $f$ est de la forme

$$f(x, y) = A(x)y + \psi(x) \quad \text{(7.2.1)}$$

où $A(x)$ est une matrice d’ordre $m$ et $\psi \in \mathbb{R}^m$.

Si de plus $A(x) = A$ est indépendante de $x$, on obtient un système linéaire à coefficients constants de la forme:

$$y'(x) = Ay + \psi(x) \quad \text{(7.2.2)}$$

Si $\psi(x) = 0$, le système est dit homogène.

$$y' = Ay \quad \text{(7.2.3)}$$
Si les valeurs propres de A sont distinctes, la solution du système homogène est de la forme

\[ y(x) = \sum_{j=1}^{k} C_j \exp(\lambda_j x) V_j + \varphi(x) \]  

(7.2.4)

où \( \lambda_j \) est valeur propre de A associée au vecteur propre \( V_j \) et les \( C_j \) sont des constantes arbitraires et \( \varphi(x) \) est une solution particulière de l’équation (7.2.2).

**Exemple 7.2.1.**

\[ y' = Ay + \psi(x), \quad y(0) = \alpha \]

avec \( A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \), \( \psi(x) = (2x - 1, x + 1)^\top \), \( \alpha = (0, 0)^\top \), les valeurs propres de A sont \( \lambda_1 = 3 \) et \( \lambda_2 = -1 \).

Les vecteurs propres de A sont \( V_1 = (1, 1)^\top \) et \( V_2 = (1, -1)^\top \).

En cherchant une solution particulière de la forme: \( \varphi(x) = \begin{pmatrix} ax + b \\ cx + d \end{pmatrix} \), on obtient

\[ y(x) = C_1 \exp(3x) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2 \exp(-x) \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -5/3 \\ -x + 4/3 \end{pmatrix}. \]

Enfin, en considérant la condition initiale on obtient \( C_1 = 1/6 \) et \( C_2 = 3/2 \), d’où

\[ y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix} = \begin{pmatrix} 1/6 \exp(3x) + 3/2 \exp(-x) - 5/3 \\ 1/6 \exp(3x) - 3/2 \exp(-x) - x + 4/3 \end{pmatrix}. \]

**7.3 Notions de stabilité**

Considérons le système nonlinéaire suivant

\[ \frac{dX(t)}{dt} = F(X(t)) \]  

(**)

où \( X(t) = (X_1, \ldots, X_n)^\top \) est un vecteur de \( \mathbb{R}^n \) et \( F = (f_1, \ldots, f_n)^\top \) une fonction de \( \mathbb{R}^n \) dans \( \mathbb{R}^n \) suffisamment régulière.

Si F est linéaire le système est dit **linéaire**

**Définition 7.3.1** (Point d’équilibre). Un vecteur \( \bar{X} \) est un point équilibre du système (**)) si à l’instant \( t_0 \) l’état du système est égal à \( \bar{X} \) et il restera égal à \( \bar{X} \) dans le futur c.a.d:

\[ X(t_0) = \bar{X} \quad \text{et} \quad \forall \ t > t_0 \ X(t) = \bar{X}. \]
On parle aussi de point stationnaire, état stationnaire et solution stationnaire qui désignent la même notion.

Définition 7.3.2. 1. Un point d’équilibre $\bar{X}$ est dit stable si :

$\exists R_0 > 0$ et $\forall R < R_0 \exists r \ 0 < r < R$ tel que si $X(t_0) \in B(\bar{X}, r)$ alors

$X(t) \in B(\bar{X}, R) \ \forall \ t > t_0$, $B(\bar{X}, r)$ : boule de centre $\bar{X}$ et de rayon $r$;

2. Un point d’équilibre est dit asymptotiquement stable s’il est stable et en plus $\exists R_0$ tel que pour tout $X(t_0) \in B(\bar{X}, R_0)$ on a

$\lim_{t \to +\infty} X(t) = \bar{X}$;

3. Un point d’équilibre est dit marginalement stable s’il est stable et non asymptotiquement stable;

4. Un point est instable si il n’est pas stable;

5. Un point est globalement asymptotiquement stable si pour tout $X(t_0)$ on a

$\lim_{t \to +\infty} X(t) = \bar{X}$.

Théorème 7.3.1. Si $F$ est linéaire, une condition nécessaire et suffisante pour que le système $(\ast \ast)$ admette 0 comme point d’équilibre asymptotiquement stable est que $Re(\lambda) < 0$, pour toute valeur propre $\lambda$ de $F$.

Si au moins une des valeurs propres de $F$ vérifie $Re(\lambda) > 0$ alors le point d’équilibre 0 est instable.

Définition 7.3.3. Si le système non linéaire $(\ast \ast)$ admet un point d’équilibre $\bar{X}$, on appelle matrice de linéarisation du système la matrice

$$A = \begin{pmatrix}
\frac{\partial f_1}{\partial X_1} & \cdots & \frac{\partial f_1}{\partial X_n} \\
\frac{\partial f_2}{\partial X_1} & \cdots & \frac{\partial f_2}{\partial X_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_n}{\partial X_1} & \cdots & \frac{\partial f_n}{\partial X_n}
\end{pmatrix}_{\bar{X}},$$

le système $\frac{dY(t)}{dt} = AY(t)$ est dit système linearisé obtenu à partir du système $(\ast \ast)$.

Théorème 7.3.2. Si le système non linéaire $(\ast \ast)$ admet l’origine comme point fixe unique alors dans un voisinage de l’origine, les comportements du système non linéaire et du système linearisé sont équivalents à condition que le système n’admette pas de point centre (valeurs propres imaginaires purses).
Remarque 7.3.1. Le théorème peut s’appliquer aux points d’équilibre qui sont distincts de l’origine, il suffit d’introduire des coordonnées locales.

Théorème 7.3.3 (Théorème de Liapunov). Soient $\bar{X}$ un point d’équilibre de $(\star \star)$ et $V$ une fonction définie de $U$ dans $\mathbb{R}$ de classe $C^1$, $U$ un voisinage de $\bar{X}$ tel que:

i) $V(\bar{X}) = 0$ et $V(X) > 0$ si $X \neq \bar{X}$,

ii) $\dot{V}(X) \leq 0$ $\forall X \in U \setminus \{\bar{X}\}$.

Alors $\bar{X}$ est stable. Si de plus

iii) $\dot{V}(X) < 0$ $\forall X \in U \setminus \{\bar{X}\}$

Alors $\bar{X}$ est asymptotiquement stable. $V(X)$ est dite fonction de Liapunov.

Pour les preuves des théorèmes 7.3.1, 7.3.2 et 7.3.3, voir ([10, 88, 103]).

7.4 Système d’équations aux différences linéaires avec coefficients constants

Soit $\{y_n, n = N_0, N_0 + 1, \cdots, \}$ une suite de vecteurs de $\mathbb{R}^m$ et $(\alpha_n)$ une suite de réels. On appelle système d’équations aux différences linéaires d’ordre $k$ le système d’équations suivantes

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \psi_n, n = N_0, N_0 + 1, \cdots \text{ avec } \psi_n \in \mathbb{R}^m \quad (7.4.1)$$

Si de plus $\psi_n = 0$, le système est dit homogène:

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0, n = N_0, N_0 + 1, \cdots \quad (7.4.2)$$

La solution générale du système (7.4.1) s’obtient de façon similaire à celle qui donne la solution du système d’équations différentielles linéaires (7.2.1). Elle est donnée par la somme d’une solution $(Y_n)$ du système homogène (7.4.2) et d’une solution particulière $\varphi_n$ du système (7.4.1) $(y_n = Y_n + \varphi_n)$.

Définition 7.4.1. On appelle polynôme caractéristique de l’équation aux différences linéaires le polynôme défini par $\pi(r) = \sum_{j=0}^{k} \alpha_j r^j$. Si les racines $r_j$ de $\pi(r)$ sont simples alors la solution générale du système (7.4.1) est donnée sous la forme

$$y_n = \sum_{j=1}^{k} \theta_j r_j^n + \varphi_n,$$

où $\varphi_n$ est une solution particulière du système (7.4.1).
Si \( r_1 \) est une racine de \( \pi(r) \) de multiplicité \( m \), la solution générale du système (7.4.1) est donnée sous la forme: 
\[
y_n = \sum_{j=1}^{m} n^{j-1} \theta_j r_1^n + \sum_{j=m+1}^{k} \theta_j r_j^n + \varphi_n.
\]

**Exemple 7.4.1.** 
\( y_{n+2} - \frac{1}{2} y_{n+1} + \frac{1}{2} y_n = 1 \), \( K = 2 \) est une solution particulière

\[
\pi(r) = r^2 - r + \frac{1}{2}, \quad \Delta = 1 - 2 = i^2.
\]

Les racines sont complexes conjuguées \( r_1 = (1 + i)/2 \) et \( r_2 = (1 - i)/2 \). La solution générale est donnée par 
\[
y_n = \theta_1 (1 + i)^n / 2^n + \theta_2 (1 - i)^n / 2^n + 2.
\]

### 7.5 Méthodes numériques pour les problèmes de condition initiale

Considérons le problème de condition initiale
\[
y'(x) = f(x, y), \quad y(a) = \alpha. \tag{7.5.1}
\]

On suppose que ce problème admet une solution unique dans l’intervalle \([a, b]\).

Les méthodes numériques utilisent la discretisation de l’intervalle \([a, b]\) en posant

\[
x_i = a + ih, \quad i = 0, 1, \cdots, N
\]

où \( h = \frac{b - a}{N} \) est le pas de discrétisation (ici le pas est supposé constant mais on peut envisager des pas \( h_i \) variables).

La solution exacte au point \( x_i \) est notée \( y(x_i) \), la solution approchée est notée \( y_i \) (\( y(x_i) \simeq y_i \)), une méthode numérique est un système d’équations aux différences impliquant un certain nombre d’approximations successives \( y_n, y_{n+1}, \cdots, y_{n+k} \) où \( k \) désigne le nombre de pas de la méthode.

Si \( k = 1 \) on parle de méthode à un pas.

Si \( k > 1 \) la méthode est dite à pas multiples ou multi-pas.

**Exemple 7.5.1.**

\[
y_{n+2} - y_{n+1} = hf_n
\]

\[
y_{n+2} + y_{n+1} - 2y_n = \frac{h}{4} (f_{n+2} + 8f_{n+1} + 3f_n)
\]

\[
y_{n+2} - y_{n+1} = \frac{h}{3} (3f_{n+1} - 2f_n)
\]

\[
y_{n+1} - y_n = \frac{h}{4} (k_1 + k_2)
\]
avec 
\[ k_1 = f_n = f(x_n, y_n) \quad \text{et} \quad k_2 = f \left( x_n + h, y_n + \frac{1}{2}hk_1 + \frac{1}{2}hk_2 \right). \]

Ces exemples peuvent être donnés sous la forme générale
\[ \sum_{j=0}^{k} \alpha_j y_{n+j} = \phi_f(y_{n+k}, \cdots y_n, x_n; h). \quad (7.5.2) \]

### 7.5.1 Convergence

Considérons une méthode numérique de type (7.5.2) avec des valeurs initiales appropriées:
\[
\begin{align*}
\sum_{j=0}^{k} \alpha_j y_{n+j} &= \phi_f(y_{n+k}, \cdots y_n, x_n; h) \\
y_{\kappa} &= \gamma_{\kappa}(h), \quad \kappa = 0, 1, \cdots, k - 1
\end{align*}
\]

**Définition 7.5.1.** La méthode (7.5.3) est dite convergente si pour tout problème de condition initiale vérifiant les hypothèses du théorème 7.1.1 (existence et unicité) on a
\[ \max_{0 \leq n \leq N} \| y(x_n) - y_n \| = 0 \quad \text{quand} \quad h \rightarrow 0 \]

**Définition 7.5.2.** On appelle erreur de troncature locale de la méthode numérique le résidu \( R_{n+k}(h) \) défini par
\[ R_{n+k}(h) = \sum_{j=0}^{k} \alpha_j y_{n+j} - h\phi_f(y(x_n+k), y(x_{n+k-1}), \cdots, y(x_n), x_n, h) \quad (7.5.4) \]

**Remarque 7.5.1.** Parfois on utilise aussi \( \tau_n(h) = \frac{1}{h} R_{n+k}(h) \) comme définition d’erreur de troncature locale, mais dans le cadre de ce chapitre c’est la première définition qui est adoptée.

**Définition 7.5.3.** La méthode (7.5.3) est dite d’ordre \( p \) si \( R_{n+k} = O(h^{p+1}) \).

### 7.5.2 Consistance

La méthode numérique est dite consistante si son ordre est au moins 1 ou encore, pour tout p.c.i vérifiant les hypothèses du théorème 7.1.1 d’existence et d’unicité on a
\[ \lim_{h \rightarrow 0} \frac{1}{h} R_{n+k}(h) = 0, \quad x = a + nh. \]

**Lemme 7.5.1.** La méthode numérique (7.5.2) est consistante si et seulement si
i) $\sum_{j=0}^{k} \alpha_j = 0$.

ii) $(\phi_f(y(x_n), y(x_n) \cdots y(x_n), x_n, 0)) / \sum_{j=0}^{k} j\alpha_j = f(x_n, y(x_n))$.

7.5.3 Stabilité

Considérons le problème de condition initiale

$$z' = f(x, z), \quad z(a) = \alpha, \quad x \in [a, b]$$

Avant de définir la stabilité de la méthode numérique, on pourrait se poser la question de savoir comment réagirait la solution $z(x)$ de ce problème à une perturbation des conditions initiales et/ou de $f$ ? Soit donc $z^*$ la solution du problème perturbé:

$$z' = f(x, z) + \delta(x), \quad z(a) = \alpha + \delta, \quad x \in [a, b].$$

Définition 7.5.4 (Hahn, Stetter). Si $(\delta_n, n = 0, 1, \cdots N)$ et $(\delta^*_n, n = 0, 1, \cdots N)$ sont deux perturbations et $z(x)$ et $z^*(x)$ les solutions qui en résultent et s’il existe une constante $S$ telle que

$$\forall x \in [a, b], \|z(x) - z^*(x)\| < \delta_0 \varepsilon \text{ si } \|\delta(x) - \delta^*(x)\| < \varepsilon \text{ et } \|\delta - \delta^*\| < \varepsilon$$

Alors le p.c.i est dit totalement stable.

Remarque 7.5.2. Dans cette définition de stabilité, on exige simplement l’existence d’une constante $S$ finie (mais pas nécessairement petite) et on montre que les hypothèses du théorème 7.1.1 sont suffisantes pour que le p.c.i soit totalement stable. On dit aussi que le problème est bien posé.

Si maintenant on considère la méthode numérique, on peut se demander quel effet aurait une perturbation de l’équations aux différences sur la solution numérique $y_n$. On a alors la définition suivante:

Définition 7.5.5 (Lambert). Soient $(\delta_n, n = 0, 1, \cdots N)$ et $(\delta^*_n, n = 0, 1, \cdots N)$ deux perturbations de l’équations aux différences (7.1.1)et $(z_n, n = 0, 1, \cdots N)$ et $(z^*_n, n = 0, 1, \cdots N)$ les solutions qui en résultent. On dira que la méthode numérique est zéro-stable si il existe deux constantes $S$ et $h_0$ telles que pour tout $h \in [0, h_0]$, si $\|\delta_n - \delta^*_n\| < \varepsilon$ $0 \leq n \leq N$ alors $\|z_n - z^*_n\| < S\varepsilon, \quad 0 \leq n \leq N$.

Remarque 7.5.3. La zéro-stabilité est encore appelée stabilité au sens de Dahlquist.

Définition 7.5.6. On appelle 1er polynôme caractéristique de la méthode numérique le polynôme $\rho(t)$ défini par $\rho(t) = \sum_{j=0}^{k} \alpha_j t^j$.  

161
Définition 7.5.7. On dit que la méthode numérique satisfait les conditions aux racines si tous les zéros du 1er polynôme caractéristique sont de module inférieur ou égal à 1 et ceux ayant un module égal à 1 sont des zéros simples.

Théorème 7.5.1. Une condition nécessaire et suffisante pour que la méthode soit zéro-stable et que la méthode vérifie la condition aux racines.

Théorème 7.5.2. Une condition suffisante et nécessaire pour que la méthode (7.5.2) soit convergente est qu’elle soit zéro-stable et consistante.

Pour les preuves des théorèmes 7.5.1 et 7.5.2, voir ([48, 103]).

7.5.4 Méthode d’Euler

La plus simple et la plus connue des méthodes d’approximation des solutions des e.d.o est la méthode d’Euler donnée par:

\[ y_{n+1} - y_n = hf_n, \quad y_0 = \alpha \] (7.5.5)

En considérant

\[ y(x_n + h) - y(x_n) - hf(x_n, y(x_n)) = \frac{1}{2} h^2 y''(\theta_n) \text{ avec } x_n \leq \theta_n \leq x_{n+1} \] (7.5.6)

On voit que la méthode d’Euler est une méthode explicite d’ordre 1 (donc consistante).

Lemme 7.5.2. i) Pour tout \( x \geq -1 \) et toute constante positive \( m \) on a

\[ 0 \leq (1 + x)^m \leq \exp(mx) \]

ii) Si \( s \) et \( t \) sont des réels positifs et \( (z_n)_{n=0}^{n=k} \) une suite vérifiant

\[ z_0 \geq -\frac{t}{s} \text{ et } z_{n+1} \leq (1 + s)z_n + t \quad \forall n = 1, \ldots, k \]

Alors on a

\[ z_{n+1} \leq \exp((n + 1)s) \left( \frac{t}{s} + z_0 \right) - \frac{t}{s} \]

Théorème 7.5.3. Soit \( f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R} \) une fonction continue et vérifiant la condition de Lipschitz pour \( y \) sur \( D = \{(x, y); a \leq x \leq b, -\infty < y < \infty\} \) avec \( a \) et \( b \) finis. On suppose qu’il existe une constante \( M \) telle que \( |y'(x)| \leq M \) pour tout \( x \in [a, b] \). Soit \( y(x) \) la solution unique du p.c.i \( y'(x) = f(x, y) \), \( y(a) = \alpha \) et \( (y_n)_{n=0}^{n=N} \) la suite des approximations générée par la méthode d’Euler. Alors on a:

\[ |y(x_n) - y_n| \leq \frac{hM}{2L} (\exp(L(x_n - a)) - 1) \text{ pour chaque } n = 0, 1, \ldots, N \]
Preuve.
Pour $n = 0$, l’inégalité est vérifiée puisque $y(x_0) = y_0 = \alpha$.

Les équations (7.5.5) et (7.5.6) donnent:

$$|y(x_{n+1}) - y_{n+1}| \leq |y(x_n) - y_n| + hf(x_n, y_n) + \frac{1}{2} h^2 |y''(\theta_n)|$$  (7.5.7)

Les hypothèses du théorème conduisent alors à la majoration

$$|y(x_{n+1}) - y_{n+1}| \leq |y(x_n) - y_n|(1 + hL) + \frac{h^2 M}{2}$$

En appliquant le lemme avec $z_n = |y(x_n) - y_n|$, $s = hL$ et $t = \frac{h^2 M}{2}$ il vient:

$$|y(x_{n+1}) - y_{n+1}| \leq \exp(((n + 1)hL) \left(|y(x_0) - y_0| + \frac{h^2 M}{2hL} \right) - \frac{h^2 M}{2hL}$$

et comme $|y(x_0) - y_0| = 0$ et $(n + 1)h = x_{n+1} - a$, on obtient:

$$|y(x_{n+1}) - y_{n+1}| \leq \frac{hM}{2L} (\exp(L(x_{n+1} - a)) - 1)$$

D’après le théorème 7.5.3, l’erreur de la méthode d’Euler est majorée par une fonction linéaire en $h$. Ceci laisse comprendre que plus on diminue $h$ plus on réduit l’erreur. Cependant, le corollaire qui va suivre indique autre chose.

**Corollaire 7.5.1.** Soit $y(x)$ la solution unique du p.c.i $y'(x) = f(x, y)$, $a \leq x \leq b$, $y(a) = \alpha$ et $w_0, w_1, \ldots, w_N$ les approximations vérifiant $w_0 = \alpha + \delta_0$ et $w_{n+1} = w_n + hf(x_n, w_n) + \delta_{n+1}$; $n = 0, \ldots, N - 1$. Si $|\delta_n| < \delta \forall n = 0, \ldots, N$ et les hypothèses du théorème précédent sont vérifiées, alors

$$|y(x_n) - w_n| \leq \frac{1}{L} \left( \frac{hM}{2} + \frac{\delta}{h} \right) \exp(L(x_n - a)) - 1) + |\delta_0| \exp L(x_n - a)$$

**Preuve.** Identique à celle du théorème avec $y(x_0) - w_0 = \delta_0$ et $t = \frac{h^2 M}{2} + |\delta_n|$.

**Remarque 7.5.4.** La simplicité de la méthode d’Euler en fait un exemple pédagogique d’introduction aux autres méthodes plus complexes. Cependant, un des critères principaux de l’analyse numérique consiste à chercher des méthodes ayant l’ordre de précision le plus élevé possible et comme la méthode d’Euler est d’ordre 1, son utilisation se trouve limitée en pratique et on est amené à considérer des méthodes plus précises. Trois directions principales permettent d’obtenir des méthodes d’ordres élevés.

La première direction consiste à travailler avec des méthodes à un pas mais en cherchant à atteindre des ordres élevés en utilisant un développement de Taylor
et en négligeant le terme d’erreur mais ces méthodes ont un handicap à cause des dérivées sucsessives de \( f \).

Une deuxième possibilité est donnée par des choix appropriés de \( \phi_f(y_{n+k}, \cdots y_n, x_n; h) \) dans l’équation (7.5.1), les méthodes de Runge-Kutta sont la meilleure illustration de cette direction.

Enfin, une troisième direction est offerte par les Méthodes Linéaires à Pas Multiples (MLPM).

En se basant sur le critère de précision, on voit qu’on est obligé de chercher des méthodes dont la performance est supérieure à celle d’Euler. On fait donc appel à d’autres méthodes plus précises.

### 7.5.5 Méthodes de Taylor dans le cas scalaire

Supposons que la solution \( y(x) \) du p.c.i

\[
y'(x) = f(x, y), \quad a \leq x \leq b, \quad y(a) = \alpha
\]  

est de classe \( C^{(n+1)} \). Alors en écrivant le développement de Taylor au point \( x_{n+1} = x_n + h \)
on obtient

\[
y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2!} y''(x_n) + \cdots + \frac{h^n}{n!} y^{(n)}(x_n) + \frac{h^{n+1}}{(n+1)!} y^{(n+1)}(\zeta_n),
\]

\[x_n < \zeta_n < x_{n+1}\]

En remplaçant \( y'(x_n) \) par \( f(x_n, y_n) \) ainsi que les dérivées supérieures de \( y' \) par celles de \( f \) puis en laissant tomber le terme \( \frac{h^{n+1}}{(n+1)!} y^{(n+1)}(\zeta_n) \), on obtient la méthode numérique:

\[
\begin{align*}
y_0 &= \alpha \\
y_{n+1} &= y_n + h T_f^n(x_n, y_n, h), \\
T_f^n(x_n, y_n, h) &= f(x_n, y_n) + \frac{h}{2!} f'(x_n, y_n) + \cdots + \frac{h^{n-1}}{n!} f^{(n-1)}(x_n, y_n) \\
n &= 0, 1, \cdots, N-1
\end{align*}
\]

**Exemple 7.5.2.** La méthode d’Euler fait partie des méthodes de Taylor.

**Exercice 7.5.1.** En faisant un développement de Taylor de \( y(x) \) à l’ordre 3 puis en remplaçant \( y''(x_n) \) par \( \frac{1}{h}(y'(x_{n+1}) - y'(x_n)) + O(h) \), montrer qu’on obtient la méthode d’Euler modifiée \( y_{n+1} - y_n = \frac{h}{2} (f(x_n, y_n) + f(x_{n+1}, y_{n+1})) \).
Remarque 7.5.5. Au vu du critère de précision et bien que les méthodes de Taylor paraissent faciles dans leur écriture, elles sont rarement utilisées dans la pratique à cause des difficultés engendrées par le calcul des dérivées successives de \( f \) comme fonction de deux variables. C’est pour cette raison qu’on cherche des méthodes permettant d’atteindre un ordre élevé tout en évitant le calcul des dérivées successives de \( f \). Au critère de précision s’ajoute le critère de coût.

7.5.6 Méthodes de Runge-Kutta (R.K) dans le cas scalaire

Revenons au p.c.i (7.1.3), les méthodes de R.K se présentent sous la forme:

\[
y_{n+1} - y_n = h \sum_{i=1}^{l} b_i k_i
\]

avec \( k_i = f \left( x_n + c_i h, y_n + h \sum_{j=1}^{l} a_{ij} k_j \right) \), \( i, = 1, 2, \cdots , l \) et on suppose que \( c_i = \sum_{j=1}^{l} a_{ij}, i, = 1, 2, \cdots , l \)

Si \( a_{ij} = 0 \) pour \( j > i \) alors:

\[
k_i = f \left( x_n + c_i h, y_n + h \sum_{j=1}^{i} a_{ij} k_j \right) \quad i, = 1, 2, \cdots , l
\]

on obtient ainsi:

\[
k_1 = f(x_n, y_n)
\]

\[
k_2 = f(x_n + c_2 h, y_n + c_2 h k_1)
\]

\[
k_3 = f(x_n + c_3 h, y_n + (c_3 - a_{32}) h k_1 + a_{32} h k_2)
\]

Remarques 7.5.6. 1. Les méthodes de R.K sont des méthodes à un pas, elles peuvent-être écrite sous la forme générale \( y_{n+1} - y_n = h \phi_f (x_n, y_n, h) \), où

\[
\phi_f (x, y, h) = \sum_{i=1}^{l} b_i k_i.
\]

2. Les méthodes de R.K satisfont la condition aux racines, elles sont donc zérostables. Par conséquent, pour étudier la convergence il suffit d’étudier la consistance.

7.5.7 Méthodes de Runge-Kutta explicites

Les méthodes explicites de R-K d’ordre 1,2,3 et 4 sont obtenues en considérant

\[
y_{n+1} = y_n + h (b_1 k_1 + b_2 k_2 + b_3 k_3 + b_4 k_4)
\]

165
en déterminant pour chaque ordre les valeurs possibles des paramètres.

\[ k_1 = f(x_n, y_n) \]
\[ k_2 = f(x_n + c_2 h, y_n + c_2 h k_1) \]
\[ k_3 = f(x_n + c_3 h, y_n + (c_3 - a_{32}) h k_1 + a_{32} h k_2) \]
\[ \vdots \]

Les formes explicites des méthodes d’ordre 1, 2 et 3 sont obtenues après différentiation et identification.

**Méthode d’ordre 1**

En considérant \( y_{n+1} = y_n + h b_1 k_1 \) avec \( k_1 = f(x_n, y_n) \) on voit que le l’ordre maximal est 1 obtenu avec le paramètre \( b_1 \) égal à 1 et qui donne la méthode d’Euler:

\[ y_{n+1} - y_n = h f(x_n, y_n) \]

**Méthodes d’ordre 2**

En partant de \( y_{n+1} = y_n + h b_1 k_1 + b_2 k_2 \) avec \( k_1 = f(x_n, y_n) \) et \( k_2 = f(x_n + c_2 h, y_n + c_2 k_1) \), \( p = 2 \) est l’ordre maximal possible qu’on peut atteindre si les paramètres \( b_1, b_2 \) et \( c_2 \) vérifient les équations \( b_1 + b_2 = 1 \) et \( b_2 c_2 = \frac{1}{2} \). Comme on a deux équations pour trois inconnues, il en résulte une infinité de méthodes explicites d’ordre 2. On en donne quelques unes:

- **Méthode d’Euler modifiée**: (\( b_1 = 0, b_2 = 1 \) et \( c_2 = \frac{1}{2} \))
  \[ y_{n+1} - y_n = h f \left( x_n + \frac{1}{2} h, y_n + \frac{1}{2} k_1 \right) \]

- **Méthode d’Euler améliorée**: (\( b_1 = b_2 = \frac{1}{2} \) et \( c_2 = 1 \))
  \[ y_{n+1} - y_n = \frac{h}{2} \left( f(x_n, y_n) + f(x_n + h, y_n + k_1) \right) \]

- **Méthode de Heun d’ordre 2** (\( b_1 = \frac{1}{4} \), \( b_2 = \frac{3}{4} \) et \( c_2 = \frac{2}{3} \))
  \[ y_{n+1} = y_n + \frac{h}{4} \left( f(x_n, y_n) + 3 f(x_n + \frac{2}{3} h, y_n + \frac{2}{3} k_1) \right) \]
  \[ k_1 = h f(x_n, y_n) \]

**Méthodes d’ordre 3**

En considérant \( y_{n+1} = y_n + b_1 k_1 + b_2 k_2 + b_3 k_3 \) avec \( k_1 = h f(x_n, y_n) \), \( k_2 = h f(x_n + c_2 h, y_n + c_2 k_1) \) et \( k_3 = h f(x_n + c_3 h, y_n + (c_3 - a_{32}) k_1 + a_{32} k_2) \).
On obtient une famille de méthode d’ordre 3 si les paramètres vérifient les équations

\[ \begin{align*}
    b_1 + b_2 + b_3 &= 1 \\
    b_2c_2 + b_3c_3 &= \frac{1}{2} \\
    b_2c_2^2 + b_3c_3^2 &= \frac{1}{3} \\
    b_3c_2a_{32} &= \frac{1}{6}
\end{align*} \]

Deux représentants de cette famille de méthode d’ordre 3 sont :

Méthode de Heun d’ordre 3 \((b_1 = \frac{1}{4}, b_2 = 0, b_3 = \frac{3}{4} \text{ et } c_2 = \frac{1}{3} \text{ et } c_3 = \frac{2}{3}, a_{32} = \frac{2}{3})\)

\[ \begin{align*}
    y_{n+1} &= y_n + \frac{h}{4}(k_1 + 3k_3) \\
    k_1 &= hf(x_n, y_n) \\
    k_2 &= hf \left( x_n + \frac{1}{3}h, y_n + \frac{1}{3}k_1 \right) \\
    k_3 &= hf \left( x_n + \frac{2}{3}h, y_n + \frac{2}{3}k_2 \right)
\end{align*} \]

Méthode de Kutta d’ordre 3 \((b_1 = \frac{1}{6}, b_2 = \frac{2}{3}, b_3 = \frac{1}{6} \text{ et } c_2 = \frac{1}{2} \text{ et } c_3 = 1, a_{32} = 2)\)

\[ \begin{align*}
    y_{n+1} &= y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \\
    k_1 &= hf(x_n, y_n) \\
    k_2 &= hf \left( x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1 \right) \\
    k_3 &= hf(x_n + h, y_n - k_1 + 2k_2)
\end{align*} \]

Méthode d’ordre 4

Méthode de Runge-Kutta d’ordre 4

\[ \begin{align*}
    y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
    k_1 &= hf(x_n, y_n) \\
    k_2 &= hf \left( x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1 \right) \\
    k_3 &= hf \left( x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2 \right) \\
    k_4 &= hf(x_n + h, y_n + k_3)
\end{align*} \]
Méthode de Runge-Kutta d’ordre 4
\[
y_{n+1} = y_n + \frac{h}{8} (k_1 + 3k_2 + 3k_3 + k_4)
k_1 = hf(x_n, y_n)
k_2 = hf\left(x_n + \frac{1}{3}h, y_n + \frac{1}{3}k_1\right)
k_3 = hf\left(x_n + \frac{2}{3}h, y_n - \frac{1}{3}k_1 + k_2\right)
k_4 = hf(x_n + h, y_n + k_1 - k_2 + k_3)
\]

Méthodes de Runge-Kutta d’ordre supérieur et contrôle de l’erreur

Des méthodes de R-K d’ordres successifs peuvent être combinées pour le contrôle de l’erreur. Deux types de cette utilisation sont illustrées [37] par:

1. Méthode de Runge-Kutta-Fehlberg. Elle consiste à utiliser une méthode de R-K d’ordre 5
\[
\tilde{y}_{n+1} = y_n + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6
\]

pour estimer l’erreur de troncature locale de la méthode de R-K d’ordre 4
\[
y_{n+1} = y_n + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5
\]

avec
\[
k_1 = hf(x_n, y_n)
k_2 = hf\left(x_n + \frac{1}{4}h, y_n + \frac{1}{4}k_1\right)
k_3 = hf\left(x_n + \frac{3}{8}h, y_n + \frac{3}{32}k_1 + \frac{9}{32}k_2\right)
k_4 = hf\left(x_n + \frac{12}{13}h, y_n + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right)
k_5 = hf\left(x_n + h, y_n + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right)
k_6 = hf\left(x_n + \frac{1}{2}h, y_n - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right)
\]

2. Méthode de Runge-Kutta-Verner. Elle consiste à utiliser une méthode de R-K d’ordre 6 :
\[
\tilde{y}_{n+1} = y_n + \frac{3}{40}k_1 + \frac{875}{2244}k_3 + \frac{23}{72}k_4 + \frac{264}{1955}k_5 + \frac{125}{11592}k_7 + \frac{43}{616}k_8
\]
pour estimer l’erreur de troncature locale de la méthode de R-K d’ordre 5 :

\[ y_{n+1} = y_n + \frac{13}{160}k_1 + \frac{2375}{5984}k_3 + \frac{5}{16}k_4 + \frac{12}{85}k_5 + \frac{3}{44}k_6 \]

avec

\[
\begin{align*}
  k_1 &= hf(x_n, y_n) \\
  k_2 &= hf \left( x_n + \frac{1}{6}h, y_n + \frac{1}{6}k_1 \right) \\
  k_3 &= hf \left( x_n + \frac{4}{15}h, y_n + \frac{4}{75}k_1 + \frac{16}{75}k_2 \right) \\
  k_4 &= hf \left( x_n + \frac{2}{3}h, y_n + \frac{5}{6}k_1 - \frac{8}{3}k_2 + \frac{5}{2}k_3 \right) \\
  k_5 &= hf \left( x_n + \frac{5}{6}h, y_n - \frac{165}{64}k_1 + \frac{55}{6}k_2 - \frac{425}{64}k_3 + \frac{85}{96}k_4 \right) \\
  k_6 &= hf \left( x_n + h, y_n - \frac{12}{5}k_1 - 8k_2 + \frac{4015}{612}k_3 - \frac{11}{36}k_4 + \frac{88}{255}k_5 \right) \\
  k_7 &= hf \left( x_n + \frac{1}{15}h, y_n - \frac{8263}{15000}k_1 + \frac{124}{75}k_2 - \frac{643}{680}k_3 - \frac{81}{250}k_4 + \frac{2484}{10625}k_5 \right) \\
  k_8 &= hf \left( x_n + h, y_n + \frac{3501}{1720}k_1 - \frac{300}{43}k_2 + \frac{297275}{52632}k_3 - \frac{319}{2322}k_4 + \frac{24068}{8405}k_5 + \frac{3850}{26703}k_7 \right)
\end{align*}
\]

Méthodes Linéaires à Pas Multiples (MLPM)

En posant \( f_n = f(x_n, y_n) \), on définit une MLPM par

\[ \sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}, \]

avec \( \alpha_j \) et \( \beta_j \) des constantes vérifiant les conditions \( \alpha_k = 1 \) et \( |\alpha_0| + |\beta_0| \neq 0 \).

Exemple 7.5.3. \( y_{n+2} - y_{n+1} = hf_{n+1} \) (ou de façon équivalente \( y_{n+1} - y_n = hf_n \)) est la méthode d’Euler à un pas.

Définition 7.5.8. On appelle \( 2^{ème} \) polynôme caractéristique de la méthode, le polynôme défini par \( \sigma(t) = \sum_{j=0}^{k} \beta_j t^j \).

L’opérateur de différence linéaire est défini par

\[ L(z(x); h) := \sum_{j=0}^{k} (\alpha_j z(x + jh) - h\beta_j z'(x + jh)) \]

Si on suppose que \( z \) est une fonction qu’on peut dériver autant de fois qu’on veut, on obtient:

\[ L(z(x); h) = C_0 z(x) + C_1 h z'(x) + \cdots + C_q h^q z^{(q)}(x) + \cdots \]
Définition 7.5.9. La MLPM est dite d’ordre \( p \) si

\[
C_0 = C_1 = \cdots = C_p = 0 \text{ et } C_{p+1} \neq 0.
\]

\( C_{p+1} \) est appelée constante d’erreur de la MLPM

### 7.6 Applications

La modélisation et la simulation par ordinateur ont fait de l’outil mathématique un moyen important pour la compréhension, l’analyse et le contrôle de certaines maladies. La formulation d’un modèle permet de tester des hypothèses, d’estimer des paramètres, de construire des théories, de discuter des conjonctures, de formuler des scénarios prédictifs, de visualiser certaines sensibilités et de remédier à l’impossibilité de certaines expériences pour coût élevé ou danger opérationnel. La modélisation mathématique permet de mieux saisir les concepts d’épidémie, de seuil et des nombres de contacts, de replacements ou de reproduction. Le modèle mathématique fournit également le support nécessaire à la réalisation d’appareils et d’équipements d’analyse et de contrôle dans le domaine médical. À ce propos, nous renvoyons à une review récente de Hethcote [85] et aux 200 références citées par l’auteur.

Dans ce chapitre, nous considérons trois types d’applications, la première traite les maladies d’infection transmise de façon directe, la deuxième est consacrée à la transmission par vecteur et la troisième traite l’effet de l’effort physique sur les dynamiques d’insuline et de glucose.

Pour les modèles à transmission directe, nous nous limiterons aux cas discrets.

**Application 7.6.1. Modèles épidémiologiques**

Pour les maladies infectieuses, les modèles mathématiques utilisés sont en général des modèles à compartiments \( c \) à \( d \) que la population est subdivisée en sous classes:

- \( M(t) \) : nombre d’individus avec immunité passive à l’instant \( t \);
- \( S(t) \) : nombre des susceptibles à l’instant \( t \) (on désigne par susceptibles, les individus qui peuvent avoir la maladie);
- \( E(t) \) : nombre des exposés à l’instant \( t \) (on désigne par exposés, les individus qui sont infectés mais ne sont pas infectieux);
- \( I(t) \) : nombre des infectieux à l’instant \( t \) (les personnes qui sont déjà touchées par la maladie et peuvent transmettre la maladie);
• $R(t)$ : nombre des résistants à l’instant $t$ (les personnes qui sont guéries avec immunisation)

Chacune des classes précédentes est dite compartiment, et suivant les caractéristiques de chaque maladie on a les types de modèles suivants : SI, SIS, SIR, SEIR, SEIRS, · · ·

**Remarque 7.6.1.**
- En général les classes $M$ et $E$ sont souvent omises.
- une population est dite fermée si on néglige l’émigration et l’immigration,
- une population est dite homogène si la population se mixe de façon homogène c à d si on néglige, les détails associés à l’âge, la location géographique (ville, village), les facteurs socio-culturels et l’hétérogénéité génétique,
- une population est de taille fixe si le taux de naissance est égal au taux de mortalité.

**Exemple 7.6.1. Modèle SIS**
C’est un modèle à deux compartiments utilisé essentiellement pour décrire les maladies sexuellement transmissibles. Les guéris ne développent pas une immunité et redeviennent susceptibles à la maladie: deux classes de sous populations susceptibles et infectieux. La dynamique des sous populations est régie par le système d’équations suivant:

$$
\begin{align*}
S_{n+1} &= S_n(1 - \lambda.\Delta t.\! I_n) + \gamma.\Delta t.\! I_n + \mu.\Delta t(1 - S_n) \\
I_{n+1} &= I_n(1 - \gamma.\Delta t - \mu.\Delta t + \lambda.\Delta t.S_n) \\
I_0 + S_0 &= 1 \text{ avec } S_0 \geq 0 \text{ et } I_0 > 0
\end{align*}
$$

où $S_n$ et $I_n$ sont les proportions d’individus susceptibles et infectieux, respectivement.

Comme $S_n + I_n = 1$ alors $I_{n+1} = I_n(1 - (\gamma + \mu).\Delta t + \lambda.\Delta t - \lambda.\Delta t.I_n)$.

- si $(\lambda - \gamma)\Delta t < 2$ alors on a convergence vers le point fixe;
- si $2 < (\lambda - \gamma)\Delta t < 2.449$ alors on a 2-points cycles;
- si $(\lambda - \gamma)\Delta t$ est assez grand alors on a comportement chaotique.

$N = 150, I_0 = 2, \gamma = 1.5, \Delta t = .5$

$\lambda = 4.5 \implies$ convergence voir figure 7.1;

$\lambda = 6.0 \implies$ 2-points cycles voir figure 7.1;

$\lambda = 6.5 \implies$ 4-points cycles voir figure 7.2;

$\lambda = 7.0 \implies$ comportement chaotique voir figure 7.2.
Figure 7.1: Convergence $\lambda = 4.5$, cycles $\lambda = 6$

Figure 7.2: 4 cycles $\lambda = 6.5$, Chaos $\lambda = 7$
Exemple 7.6.2. Modèle SIR
Le modèle SIR est un modèle simple de transmission des maladies par contact entre les individus, ces derniers deviennent immunisés après une seule infection, une troisième classe de sous-populations est considérée: la classe R des isolés ou enlevés “removed” ce sont les personnes qui, une fois infectées soit deviennent immunisées soit ils meurent. Le modèle SIR est plus adapté pour les maladies d’enfants comme la rougeole. Les équations qui régissent ce modèle sont :

\[
\begin{align*}
S_{n+1} &= S_n(1 - \lambda \Delta t I_n) + \mu \Delta t (1 - S_n) \\
I_{n+1} &= I_n(1 - \gamma \Delta t - \mu \Delta t + \lambda \Delta t S_n) \\
R_{n+1} &= R_n(1 - \mu \Delta t) + \gamma \Delta t I_n \\
S_0 + I_0 + R_0 &= 1 \text{ avec } S_0 > 0, I_0 > 0 \text{ et } R_0 \geq 0
\end{align*}
\]

\( N = 200, I_0 = 0.115, S_0 = 0.885 \) et \( \Delta t = 0.2 \):

1. \( \lambda = 3.5, \gamma = 2 \) voir figure 7.3 ;

2. \( \lambda = 5.5, \gamma = 2.5 \) voir figure 7.3.

Figure 7.3: \( \lambda = 3.5, \gamma = 2 \cdot \lambda = 5.5, \gamma = 2.5 \)
(SIR à deux populations)

i) $\Delta t = 0.25, \alpha_{11} = 2, \alpha_{12} = 0.5, \alpha_{21} = 4, \alpha_{22} = 2, \gamma_1 = 2, \gamma_2 = 1, N^1 = 100, N^2 = 200, I_0^1 = 10, I_0^2 = 50.$

ii) $\Delta t = 0.25, \alpha_{11} = 2, \alpha_{12} = 0.5, \alpha_{21} = 4, \alpha_{22} = 2, \gamma_1 = 2, \gamma_2 = 1, N^1 = 100, N^2 = 200, I_0^1 = 10, I_0^2 = 150.$

Figure 7.4: SIR à deux populations
Exemple 7.6.3. Modèle SEIR

Modèle SEIR avec $\lambda$ indépendante du temps

Le modèle SEIR est une variante du modèle SIR où la période de latence est prise en compte c.à.d une classe de la population est intermédiaire entre la classe des susceptibles et les infectieux: la classe des exposés qu’un susceptible traverse pour passer à l’état d’infection. Un autre paramètre est introduit qu’on note $\beta$ et on parle de période de latence $\frac{1}{\beta}$. Les équations régissant ce modèle sont:

\[
\begin{align*}
S_{n+1} &= S_n(1 - \lambda I_n \Delta t) + \mu \Delta t (1 - S_n) \\
E_{n+1} &= E_n + S_n \lambda I_n \Delta t - (\beta + \mu) \Delta t E_n \\
I_{n+1} &= I_n(1 - \gamma \Delta t - \mu \Delta t) + \beta \Delta t E_n \\
R_{n+1} &= R_n(1 - \mu \Delta t) + \gamma \Delta t I_n \\
S_0 + I_0 + R_0 + E_0 &= 1
\end{align*}
\]

On suppose toujours que la taille de la population est constante.

Pour $\lambda = 10^{-6}$, $\beta = 45.6$, $\gamma = 73$, $\mu = 0.02$, $I_0 = 0.0006$, $E_0 = 0.001$ et $S_0 = 0.25$, le modèle SEIR a le comportement suivant:

i) $\Delta t = 0.01$ —> convergence voir figure 7.5.

ii) $\Delta t = 0.018$ —> oscillation puis convergence voir figure 7.5.

iii) $\Delta t = 0.027$ —> oscillation voir figure 7.6.
Exemple 7.6.4. Modèle SEIR saisonnier

Le modèle saisonnier SEIR tient compte de la variation du taux d’infections selon les saisons donc on a le système suivant:

\[
\begin{align*}
S_{n+1} &= S_n(1 - \lambda_n \Delta t) + \mu \Delta t (1 - S_n) \\
E_{n+1} &= E_n + S_n \lambda_n I_n \Delta t - (\beta + \mu) \Delta t E_n \\
I_{n+1} &= I_n(1 - \gamma \Delta t - \mu \Delta t) + \beta \Delta t E_n \\
R_{n+1} &= R_n(1 - \mu \Delta t) + \gamma \Delta t I_n \\
S_0 + I_0 + R_0 + E_0 &= 1
\end{align*}
\]

avec \( \lambda_n = \lambda(n \Delta t) = c_0 + c_1 (1 + \cos(2\pi n \Delta t)) \) (dû à Bolker et Grenfell)

i) Pour \( \beta = 45, \gamma = 78, \mu = 0.01, \Delta t = 0.001, I_0 = 0.006, E_0 = 0.03 \) et \( S_0 = 0.25 \) il y’a une épidémie voir figure 7.7.
ii) Pour $\beta = 45$, $\gamma = 73$, $\mu = 0.02$, $\Delta t = 0.027$, $I_0 = 0.001$, $E_0 = 0.01$ et $S_0 = 0.25$ il y’a oscillation voir figure 7.7.

Application 7.6.2. Transmission indirecte par vecteur
Nous considérons le cas de la maladie infectueuse DENGUE transmise à l’homme par le mosquito (Aedes) et qui est actuellement endémique dans plus de cent pays d’Afrique, d’Amérique et d’Asie [53].

Le modèle comporte un ensemble d’équations pour l’homme et un autre pour le vecteur (mosquito) exprimés à partir du diagramme de la figure 7.8.

On suppose qu’on dispose d’une population humaine (respectivement de vecteurs mosquitos) de taille $N_h$ (resp. $N_v$) composée de Susceptibles $S_h$, d’infectieux $I_h$ et de guéris $R_h$ (resp. $S_v$ et $I_v$).
Le modèle suppose un mixage homogène d’humain et de vecteur de telle sorte que chaque piqure a la même probabilité d’être appliquée à un humain particulier. En notant \( b_s \) le taux moyen de piqures d’un vecteur susceptible, \( p_{hv} \) la probabilité moyenne de transmission d’un infectieux humain à vecteur susceptible, le taux d’exposition pour les vecteurs est donné par: \( (p_{hv}I_h b_s)/N_h \).

En notant \( p_{vh} \) la probabilité moyenne de transmission d’un infectieux vecteur à un humain et \( I_v \) le nombre de vecteurs infectieux, le taux d’exposition pour les humains est donné par: \( (p_{vh}I_v b_i)/N_h \) par conséquent:

- Le taux de contact adéquat d’humain à vecteurs est donné par: \( C_{hv} = p_{hv} b_s \)
- Le taux de contact adéquat de vecteurs à humain est donné par:

\[
C_{vh} = p_{vh} b_i.
\]

La durée de vie humaine est prise égale à 25 000 jours (68.5 ans), et celle des vecteurs est de: 4 jours. Les autres paramètres sont donnés dans le tableau 7.1 d’après [141].

<table>
<thead>
<tr>
<th>Nom du paramètre</th>
<th>Notation</th>
<th>valeur de base</th>
</tr>
</thead>
<tbody>
<tr>
<td>probabilité de transmission de vecteur à humain</td>
<td>( p_{hv} )</td>
<td>0.75</td>
</tr>
<tr>
<td>probabilité de transmission d’humain à vecteur</td>
<td>( p_{vh} )</td>
<td>0.75</td>
</tr>
<tr>
<td>piqure par susceptible mosquito par jour</td>
<td>( b_s )</td>
<td>0.5</td>
</tr>
<tr>
<td>piqure par infectieux mosquito par jour</td>
<td>( b_i )</td>
<td>1.0</td>
</tr>
<tr>
<td>taux de contact effectif humain à vecteur</td>
<td>( C_{hv} )</td>
<td>0.375</td>
</tr>
<tr>
<td>taux de contact effectif, vecteur à humain</td>
<td>( C_{vh} )</td>
<td>0.75</td>
</tr>
<tr>
<td>Durée de vie des humains</td>
<td>( \mu_h )</td>
<td>25 000 jours</td>
</tr>
<tr>
<td>Durée de vie des vecteurs</td>
<td>( \mu_v )</td>
<td>4 jours</td>
</tr>
<tr>
<td>durée de l’infection</td>
<td>( \mu_h + \gamma_h )</td>
<td>3 jours</td>
</tr>
</tbody>
</table>

Tableau 7.1: définitions et valeurs des paramètres

Les équations qui régissent le modèle sont données par

**Population humaine**

\[
\begin{align*}
\frac{dS_h}{dt} &= \mu_h N_h - (\mu_h + p + C_{vh} I_v / N_h) S_h \\
\frac{dI_h}{dt} &= (C_{vh} I_v / N_h) S_h - (\mu_h + \gamma_h) I_h \\
\frac{dR_h}{dt} &= p S_h + \gamma_h I_h - \mu_h R_h \\
\end{align*}
\]
Population vecteur

\[
\begin{align*}
\frac{dS_v}{dt} &= \mu_v N_v - (\mu_v + \frac{C_{hv}I_h}{N_h})S_v \\
\frac{dI_v}{dt} &= (\frac{C_{hv}I_h}{N_h})S_v - \mu_v I_v
\end{align*}
\]

Avec les conditions \( S_h + I_h + R_h = N_h,\ S_v + I_v = N_v,\ R_h = N_h - S_h - I_h \) et \( S_v = N_v - I_v \)

Le système s’écrit

\[
\begin{align*}
\frac{dS_h}{dt} &= \mu_h N_h - (\mu_h + p + \frac{C_{vh}I_v}{N_h})S_h \\
\frac{dI_h}{dt} &= (\frac{C_{vh}I_v}{N_h})S_h - (\mu_h + \gamma_h)I_h \\
\frac{dI_v}{dt} &= \frac{C_{hv}I_h}{N_h}(N_v - I_v) - \mu_v I_v
\end{align*}
\]

Pour l’étude de stabilité et d’autres détails concernant les paramètres du modèle, le lecture pourra consulter le papier cité en références. Dans le cadre restreint de ce chapitre, nous discutons très brièvement l’output du modèle.

La difficulté principale de la dengue provient du fait qu’elle est causée par quatre virus différents et que l’immunité provisoire acquise contre un virus ne protège pas contre les autres virus, au contraire, un individu attaqué par un deuxième virus court le danger d’évoluer vers le stade de Dengue Hemorragique.

La recherche de vaccin se complique par le fait que ce vaccin doit couvrir le spectre des quatre virus et ceci constitue un problème.

Le modèle montre que la réduction du nombre de mosquitos (par insecticides ou autre) n’est pas suffisante pour éradiquer l’épidémie de la Dengue. Elle peut tout au plus retarder l’apparition de l’épidémie. Comme la découverte d’un vaccin global n’est pas attendue dans le court terme, les auteurs suggèrent la combinaison du contrôle des facteurs environnementaux et de vaccins partiels contre chaque virus.
Application 7.6.3. (Diabète et effort physique[54])

Pour un diabétique, l’effort physique figure au même niveau que le traitement medical et la diététique. Il peut l’aider à contrôler la quantité de sucre dans le sang de façon directe ou indirecte en améliorant la sensibilité de l’insuline et la réponse des muscles et des cellules ou encore en combattant l’excès de poids qui constitue un facteur à risque. S’il est bien connu que les spécialistes du diabètes insistent toujours sur le rôle de l’effort physique en prenant en compte les capacités et les données de chaque individu, il est aussi intéressant de voir comment les modèles mathématiques sont utilisés dans ce domaine.

En 1939, Himsworth et Ker ont introduit la première approche de mesure de sensibilité d’insuline in vivo. Les modèles mathématiques ont été utilisés pour la dynamique de glucose et d’insuline. Le pionnier dans ce domaine est Bolie(1961) qui a proposé un modèle simple supposant que la disparition du glucose est une fonction linéaire du glucose et d’insuline, la secretion d’insuline est proportionnelle au glucose et elle disparaît proportionnellement à la concentration d’insuline dans le plasma. Ce modèle qui sera utilisé avec quelques modifications par d’autres auteurs ( Akerman et al (1965), Della et al (1970), Serge et al. (1973)) peut-être formulé par le système différentiel suivant:

\[
\begin{align*}
\frac{dG(t)}{dt} &= -a_1 G - a_2 I + p \\
\frac{dI(t)}{dt} &= -a_3 G - a_4 I
\end{align*}
\]

où \( G = G(t) \) représente la concentration de glucose et \( I = I(t) \) représente l’insuline, \( p, a_1, a_2, a_3, a_4 \) sont des paramètres.
Durant les dernières décennies, une littérature abondante a été consacrée à ce sujet, le lecteur intéressé pourra consulter les deux reviews récentes par Bellazi et al (2001) et Parker et al. (2001).

Incorporant l’effet de l’effort physique sur la dynamique du glucose et de l’insuline, Derouich et Boutayeb (2002) ont proposé le modèle suivant

\[
\begin{align*}
\frac{dG(t)}{dt} &= -(1 + q_2)X(t)G(t) + (p_1 + q_1)(G_b - G(t)) \\
\frac{dI(t)}{dt} &= (p_3 + q_3)(I(t) - I_b) + p_2X(t)
\end{align*}
\]

où \((I(t) - I_b)\) représente la différence entre l’insuline dans le plasma et l’insuline de base \((G_b - G(t))\) représente la différence entre la concentration de glucose dans le plasma et le glucose de base \(X(t)\) est l’insuline interstitiale, \(q_1, q_2, q_3\) sont des paramètres liés à l’effort physique.

Les auteurs ont discuté les résultats de ce modèle dans trois cas différents:

- cas normal (non diabétique),
- cas diabétique insulino-dépendant,
- cas diabétique non insulino-dépendant.

Dans les trois cas, le modèle met en évidence l’intérêt de l’effort physique à améliorer la sensibilité de l’insuline mais le dernier cas reste le plus illustratif comme l’indique la figure 7.10 où on voit qu’un diabétique non insulino-dépendant pourrait être amené à vivre continuellement avec 2g/ml de glucose dans le sang sans se rendre compte des conséquences à long terme de cette overdose. Avec un effort physique, le diabétique peut ramener la courbe de glucose au voisinage de la concentration normale de 1g/ml.
7.7 Complément bibliographique

La modélisation épidémologique remonte au moins à 1760 lorsque Bernouilli proposa un modèle mathématique pour la variole.

En 1906 un modèle discret pour l’épidémie de la rougeole a été considéré par Haner.

Ross utilisa les équations différentielles pour des modèles hôte-vector de Malaria en 1911.

D’autres modèles déterministes ont été proposés par Ross, Hudson et Lotka (1922).

A partir de 1926 la contribution de Kermack et McKendrick restera marquée par le seuil endémique que la densité des susceptibles doit dépasser pour que l’épidémie se propage.

Le livre de Bailey publié en 1957 sur la théorie mathématique des maladies infectieuses marquera la lancée d’une nouvelle vague de la deuxième moitié du 20ème siècle.

Bartlett (1960) est parmi les precurseurs des modèles épidémiologiques stochastiques même si ceux-ci ont été utilisés auparavant par Yule(1924) et d’autres.

Les années 60 verront aussi le débat s’animer entre stochastique et déterminisme. Certains écologistes vont soutenir que les modèles stochastiques sont les mieux adaptés à décrire la nature avec des facteurs imprédictibles et des événements

Hoppensteadt (1975) est considéré par certains auteurs comme le premier à avoir proposé une analyse mathématique des modèles avec structure d’âge.

7.8 Exercices

**Exercice 7.8.1.** On considère le modèle discret suivant:

\[
S_{n+1} = \exp(-aI_n)S_n \\
I_{n+1} = bI_n + (1 - \exp(-aI_n))S_n \\
R_{n+1} = (1 - b)I_n + R_n
\]

On suppose que \( R_0 = 0 \), montrer que

1. La population des susceptibles tend vers une limite \( S \) quand \( n \) tend vers l’infini.
2. Si \( F = S/S_0 \) alors
   \[
   F = \exp(-aS_0/(1-b))(1 + I_0/S_0 - F).
   \]
3. Expliquer comment \( F \) peut jouer un rôle de seuil.

**Exercice 7.8.2.** Chercher l’ordre et la constante d’erreur \( C_p \) pour la méthode suivante:

\[
y_{n+2} - (1 + \alpha)y_{n+1} + \alpha y_n = h((1 + \beta)f_{n+2} - (\alpha + \beta + \alpha\beta)f_{n+1} + \alpha\beta f_n)
\]

**Exercice 7.8.3.** Chercher les solutions des systèmes différentiels et aux différences suivants: \( y' = Ay + \psi(x) \), \( y(0) = \alpha \) avec

\[
A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \psi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}, \alpha = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix}
\]

\[
y_{n+4} - 6y_{n+3} + 14y_{n+2} - 16y_{n+1} + 8y_n = \phi_n \text{ avec } y_n \in \mathbb{R}^2 \text{ et } \phi_n = \begin{pmatrix} n \\ 1 \end{pmatrix}.
\]

**Exercice 7.8.4.** Considérons le système aux différences :

\[
y_{n+2} - 2\mu y_{n+1} + \mu y_n = c, \quad n = 0, 1, \ldots
\]

avec \( y_n, c \in \mathbb{R}^m \) et \( 0 \leq \mu \leq 1 \).

Montrer que \( y_n \) converge vers \( c/(1 - \mu) \) quand \( n \to \infty \).

**Exercice 7.8.5.** A- Montrer que:

i) Pour tout \( x \geq -1 \) et toute constante positive \( m \) on a:

\[
0 \leq (1 + x)^m \leq \exp(mx)
\]

184
ii) Si $s$ et $t$ sont des réels positifs et $(z_n)_{n=0}^{n=k}$ une suite vérifiant:

$$z_0 \geq -\frac{t}{s} \text{ et } z_{n+1} \leq (1+s)z_n + t \ \forall \ n = 1, \cdots, k$$

Alors on a:

$$z_{n+1} \leq \exp((n+1)(1+s))\left(\frac{t}{s} + z_0\right) - \frac{t}{s}.$$ 

B- Soit $y(x)$ la solution unique du p.c.i $y'(x) = f(x, y), a \leq x \leq b$, $y(a) = \alpha$ et $w_0, w_1, \cdots, w_N$, les approximations vérifiant: $w_0 = \alpha + \delta_0$ et $w_{n+1} = w_n + hf(x_n, w_n) + \delta_{n+1}; n = 0, ..., N - 1$

1. On suppose que: $D = \{(x, y); a \leq x \leq b, -\infty < y < \infty\}$ avec $a$ et $b$ finis

i) il existe une constante $L$ positive telle que:

$$|f(x, y) - f(x, y^*)| \leq L|y - y^*| \forall y, y^* \in D$$

ii) il existe une constante $M$ telle que: $|y''(x)| \leq M$ pour tout $x \in [a, b]$.

iii) les perturbations vérifient: $|\delta_n| < \delta$ $\forall n = 0, \cdots, N$

Montrer que:

$$|y(x_n) - w_n| \leq \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{L}\right)(\exp(L(x_n - a)) - 1) + |\delta_0|\exp(L(x_n - a))$$

(Preuve: identique à celle du théorème du cours avec $y(x_0) - w_0 = \delta_0$ et $t = \frac{h^2M}{2} + |\delta_n|$)

2. En posant $\varepsilon(h) = \frac{hM}{2} + \frac{\delta}{h}$ et en remarquant que $\lim \varepsilon(h) = \infty$, montrer qu’on peut déterminer une limite inférieure $h_0$ de $h$ qui rend minimum $\varepsilon(h)$.

**Exercice 7.8.6.** Soit $A$ une matrice diagonalisable admettant les vecteurs propres $V_j$ associés aux valeurs propres $\lambda_j$ avec $Re(\lambda_j) < 0$ pour tout $j$.

On considère les deux schémas d’Euler donnés par:

\begin{align*}
E_1 & : \quad u_{n+1} = u_n + hAu_n, \quad u_0 = \eta \\
E_2 & : \quad u_{n+1} = u_n + hAu_{n+1}, \quad u_0 = \eta
\end{align*}
Montrer que les solutions de ces deux schémas sont données par:

\[ S_1 : u_n = \sum_{j=1}^{k} c_j (1 + h\lambda_j)^n V_j \]

\[ S_2 : u_n = \sum_{j=1}^{k} c_j (1 - h\lambda_j)^{-n} V_j \]

Quelles conditions doit-on imposer à \( h \) pour que les solutions \( S_1 \) et \( S_2 \) tendent vers zéro ou restent bornées quand \( n \to \infty \)

**Exercice 7.8.7.** En posant \( f_n = f(x_n, y_n) \), on définit une MLPM par:

\[ \sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j} \]

avec \( \alpha_j \) et \( \beta_j \) des constantes vérifiant les conditions:

\[ \alpha_k = 1 \text{ et } |\alpha_0| + |\beta_0| \neq 0 \]

On appelle 1er (resp. 2ème) polynôme caractéristique de la MLPM le polynôme \( \rho(t) \) (resp. \( \sigma(t) \)):

\[ \rho(t) = \sum_{j=0}^{k} \alpha_j t^j, \quad \sigma(t) = \sum_{j=0}^{k} \beta_j t^j. \]

Si l’opérateur de différence linéaire est donné par:

\[ L(z(x); h) := \sum_{j=0}^{k} (\alpha_j z(x + jh) - h\beta_j z'(x + jh)) = C_0 z(x) + C_1 h z'(x) + \cdots + C_q h^q z^{(q)}(x) + \cdots \]

1. Montrer que:

\[ C_0 = \sum_{j=0}^{k} \alpha_j = \rho(1) \]

\[ C_1 = \sum_{j=0}^{k} (j\alpha_j - \beta_j) = \rho'(1) - \sigma(1) \]

\[ C_q = \sum_{j=0}^{k} \left( \frac{1}{q} j^q \alpha_j - \frac{1}{(q-1)} j^{q-1} \beta_j \right), \quad q = 2, 3, \ldots \]

2. La MLPM est dite consistante si son ordre est supérieur ou égal à 1 (\( p \geq 1 \))

Montrer que la MLPM est consistante si et seulement si:

\[ \rho(1) = 0 \text{ et } \rho'(1) = \sigma(1) \]

186
Exercice 7.8.8. Chercher l’ordre des méthodes numériques suivantes:

1) \( y_{n+2} = y_n + \frac{h}{6}(f_n + 4f_{n+1} + f_{n+2}) \)

2) \( y_{n+4} = y_n + \frac{4h}{3}(2f_{n+1} - f_{n+2} + 2f_{n+3}) \)

3) \( y_{n+1} = y_n + \frac{h}{4}\left( f(x_n, y_n) + 3f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}k_1) \right), \quad k_1 = hf(x_n, y_n) \)

4) \( y_{n+1} = y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \)

\[
\begin{align*}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1) \\
k_3 &= hf(x_n + h, y_n - k_1 + 2k_2)
\end{align*}
\]

Exercice 7.8.9. Considérons la méthode numérique

\( y_{n+2} - (1 + \alpha)y_{n+1} + \alpha y_n = \frac{h}{2}\left( (3 - \alpha)f_{n+1} - (1 + \alpha)f_n \right) \)

où \( f_n = f(x_n, y_n) \) et \(-1 \leq \alpha \leq 1\).

1. Donner l’erreur de troncature locale de la méthode et l’ordre de la méthode en fonction de \( \alpha \).

2. Cette méthode est utilisée numériquement pour résoudre l’équation scalaire

\[
\begin{cases}
y'(x) = y(x) \\
y(0) = 1
\end{cases}
\]

en supposant que \( y_0 = 1 + \omega h^3 \)

\( y_1 = \exp(h) + \theta h^3 \)

Montrer que la solution approchée \( y_n \) peut-être donnée sous la forme

\[
y_n = \frac{\Omega(r_2)r_1^n - \Omega(r_2)r_1^n}{r_1 - r_2}
\]

où \( r_1 \) et \( r_2 \) sont les racines de l’équation caractéristique associées à l’équation aux différences. et \( \Omega(r) = \exp(h) - r + (\theta - r\omega)h^3 \).
3. On suppose que $r_1$ est de la forme

$$r_1 = 1 + h + \frac{h^2}{2} + O(h^3)$$

(a) Donner une expression analogue pour $r_2$.
(b) Etudier la convergence de $y_n$ dans le cas $\alpha = -1$.

**Exercice 7.8.10.** On considère la méthode numérique suivante

$$y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h(\beta_1 f_{n+1} + \beta_0 f_n) \quad (M)$$

où $f_n = f(x_n, y_n)$ et $f(x_n, y(x_n)) = y'(x_n)$.

1. Déterminer les constantes $\alpha_0, \alpha_1, \beta_0$ et $\beta_1$ pour que la méthode soit d'ordre maximum.

2. La méthode est-elle zéro-stable pour les valeurs des constantes trouvées?

3. On utilise la méthode $(M)$ avec $y_0 = 1$ et $y_1 = \exp(-h)$ pour approcher la solution du problème de condition initiale suivant

$$y'(x) = -y(x), y(0) = 1 \quad (P)$$

Montrer qu'on obtient l’équation aux différences suivante

$$y_{n+2} + 4(1 + h)y_{n+1} + (-5 + 2h)y_n = 0; \ n = 0, 1, \ldots \quad (D)$$

4. Montrer que les racines $r_1$ et $r_2$ de l’équation caractéristique associée à $(D)$ sont

$$r_1 = -2 - 2h + 3 \left(1 + \frac{2}{3}h + \frac{4}{9}h^2\right)^{1/2} \quad \text{et} \quad r_2 = -2 - 2h - 3 \left(1 + \frac{2}{3}h + \frac{4}{9}h^2\right)^{1/2}.$$ 

5. Montrer que la solution de $(D)$ est de la forme

$$y_n = C_1(r_1)^n + C_2(r_2)^n,$$

avec $C_1 = \frac{r_2 - \exp(-h)}{r_2 - r_1}$ et $C_2 = \frac{\exp(-h) - r_1}{r_2 - r_1}$.

6. On donne $\left(1 + \frac{2}{3}h + \frac{4}{9}h^2\right)^{1/2} = 1 + \frac{1}{3}h + \frac{1}{6}h^2 - \frac{1}{18}h^3 + \frac{1}{216}h^4 + O(h^5)$. En déduire que $r_1 = 1 - h + \frac{1}{2}h^2 - \frac{1}{6}h^3 + \frac{1}{72}h^4 + O(h^5), \ r_2 = -5 - 3h + O(h^2)$. 

188
7. Montrer alors que $C_1 = 1 + O(h^2)$ et $C_2 = -\frac{1}{216}h^4 + O(h^5)$. 

8. En considérant pour $x$ fixé, $nh = x$, montrer que $C_1(r_1)^n \to \exp(-x)$ quand $n \to \infty$. 

Cette annexe est un guide d’initiation à MATLAB. MATLAB est un programme interactif de calcul scientifique utilisable pour la résolution numérique de nombreux problèmes mathématiques ou appliqués. En outre, MATLAB dispose de potentialités graphiques importantes.

L’objectif est de permettre au débutant de se familiariser avec MATLAB qui est avant tout un programme de calcul matriciel.

Toutefois, nous nous sommes limités aux méthodes les plus courantes décrites dans les chapitres précédents, il est recommandé d’exécuter les exemples dans l’ordre où ils apparaissent.

```matlab
function [n]=mat_square(A)
[n,m] = size(A);
if n ~= m
    disp(' Error: the matrix should be squared');
    stop
end
return
```

**Broyden**

```matlab
function [x,it]=broyden(x,Q,nmax,toll,f)
[n,m]=size(f); it=0; err=1;
fk=zeros(n,1); fk1=fk;
for i=1:n, fk(i)=eval(f(i,:)); end
while it < nmax & err > toll
    s=-Q\fk; x=s+x; err=norm(s,inf);
    if err > toll
        for i=1:n, fk1(i)=eval(f(i,:)); end
        Q=Q+1/(s'*s)*fk1*s';
    end
end

190
it=it+1; fk=fk1;
end

Choleski

function [A] = cholesky (A)
[n,n]=size(A); for k=1:n-1
    A(k,k)=sqrt(A(k,k));
    A(k+1:n,k)=A(k+1:n,k)/A(k,k);
    for j=k+1:n
        A(j:n,j)=A(j:n,j)-A(j:n,k)*A(j,k);
    end
end A(n,n)=sqrt(A(n,n)); return

Jacobi

function [xvect,err,nit]=jacobi(x0,n_max,tol,A,b)
P=diag(diag(A));
N=P-A; B=eye(size(A))-inv(P)*A; bb=P\b; r=b-A*x0; erreur=norm(r);
nit=1; xvect(:,1)=x0; err(1)=erreur;
while (nit <= n_max) & (erreur > tol)
    xvect(:,nit+1) = B*xvect(:,nit) +bb;
    r=b-A*xvect(:,nit+1);
    erreur = norm(r);
    err(nit+1) =erreur;
    nit = nit + 1;
end
return

Gauss-Seidel

function [xvect,err,nit]=gauss_seidel(x0,n_max,tol,A,b)
P=tril(A); F=-triu(A,1); B=inv(P)*F; bb=P\b; r=b-A*x0;
erreur=norm(r); nit=1; xvect(:,1)=x0; err(1)=erreur;
while (nit <= n_max) & (erreur > tol)
    xvect(:,nit+1) = B*xvect(:,nit) +bb;
    r=b-A*xvect(:,nit+1);
end
return
erreur = norm(r);
err(nit+1) = erreur;
nit = nit + 1;
end

return

Householder

function [H,Q]=houshess(A) n=max(size(A));
Q=eye(n); H=A; for
k=1:(n-2),
    [v,beta]=vhouse(H(k+1:n,k));
    I=eye(k);
    N=zeros(k,n-k);
    m=length(v);
    R=eye(m)-beta*v*v';
    H(k+1:n,k:n)=R*H(k+1:n,k:n);
    H(1:n,k+1:n)=H(1:n,k+1:n)*R;
    P=[I, N; N', R];
    Q=Q*P;
end return

Horner

function [pnz,b] = horner(a,n,z)
b(1)=a(1);
for j=2:n+1,
b(j)=a(j)+b(j-1)*z;
end; pnz=b(n+1);

LU

function [L,U,P,Q] = LUpivtot(A,n)
P=eye(n); Q=P; Minv=P; for
k=1:n-1
    [Pk,Qk]=pivot(A,k,n);
    A=Pk*A*Qk;
    [Mk,Mkinv]=MGauss(A,k,n);
    A=Mk*A;
    P=Pk*P;

192
Gradient

function \([x, \text{error}, \text{niter}, \text{flag}] = \text{gradient}(A, x, b, M, \text{maxit},\text{tol})\)

\[\text{flag} = 0;\]
\[\text{niter} = 0;\]
\[\text{bnrm2} = \text{norm}(b);\]
\[\text{if } (\text{bnrm2} == 0.0),\]
\[\text{bnrm2} = 1.0;\]
\[\text{end}\]
\[r = b - A*x;\]
\[\text{error} = \text{norm}(r) / \text{bnrm2};\]
\[\text{if } (\text{error} < \text{tol})\]
\[\text{return},\]
\[\text{end}\]
\[\text{for } \text{niter} = 1: \text{maxit}\]
\[z = M \backslash r;\]
\[\text{rho} = (r'*z);\]
\[q = A*z;\]
\[\alpha = \text{rho} / (z'*q);\]
\[x = x + \alpha * z;\]
\[r = r - \alpha*q;\]
\[\text{error} = \text{norm}(r) / \text{bnrm2};\]
\[\text{if } (\text{error} <= \text{tol})\]
\[\text{break},\]
\[\text{end}\]
\[\text{end}\]
\[\text{if } (\text{error} > \text{tol})\]
\[\text{flag} = 1;\]
\[\text{end}\]
\[\text{return}\]

Substitution

function \([b]=\text{backward}_\text{col}(U,b)\)
\[[n]=\text{mat}_\text{square}(U);\]
l = min(diag(abs(U))); if l == 0 disp('Error: the matrix is singular'); b = []; break end for j = n:-1:2, b(j) = b(j)/U(j,j); b(1:j-1) = b(1:j-1) - b(j)*U(1:j-1,j); end; b(1) = b(1)/U(1,1); return

Relaxation

function [x, iter]= sor ( a, b, x0, nmax, toll, omega)
[n,n]=size(a); iter = 0; r = b - a * x0; r0 = norm (r); err = norm (r); xold = x0; while err > toll & iter < nmax
iter = iter + 1;
for i=1:n
s = 0;
for j = 1:i-1
s = s + a (i,j) * x (j);
end
for j = i+1:n
s = s + a (i,j) * xold (j);
end
x (i) = omega * ( b(i) - s) / a(i,i) + (1 - omega) * xold (i);
end
x = x(:); xold = x;
r = b - a * x;
err = norm (r) / r0;
end
return

Sturm

function [p]=sturm(dd,bb,x)
n=length(dd);
p(1)=1;
p(2)=d(1)-x;
for i=2:n
    p(i+1)=(dd(i)-x)*p(i)-bb(i-1)^2*p(i-1);
end
return

Point fixe

function [xvect,xdif,fx,nit]=fixpoint(x0,nmax,toll,fun,phi)
err=toll+1;
nit=0;
xvect=x0;
x=x0;
fx=eval(fun);
xdif=[];
while (nit < nmax & err > toll),
    nit=nit+1;
    x=xvect(nit);
    xn=eval(phi);
    err=abs(xn-x);
    xdiff=[xdiff; err];
    x=xn;
    xvect=[xvect;x];
    fx=[fx;eval(fun)];
end;

Gradient conjugué

function [x, error, niter, flag] = conjgrad(A, x, b, P, maxit, tol)
flag = 0;
niter = 0;
bnrm2 = norm( b );
if ( bnrm2 == 0.0 ),
    bnrm2 = 1.0;
end
r = b - A*x;
error = norm( r ) / bnrm2;
if ( error < tol )
    return,
for niter = 1:maxit
    z = P \ r;
    rho = (r'*z);
    if niter > 1
        beta = rho / rho1;
        p = z + beta*p;
    else
        p = z;
    end
    q = A*p;
    alpha = rho / (p'*q);
    x = x + alpha * p;
    r = r - alpha*q;
    error = norm( r ) / bnrm2;
    if ( error <= tol ),
        break,
    end
    rho1 = rho;
end
if ( error > tol )
    flag = 1;
end
return
décomposition LU

function [A] = lu_band (A,p,q)
%& decomposition LU de A
[n,n]=size(A);
for k = 1:n-1
    for i = k+1:min(k+p,n)
        A(i,k)=A(i,k)/A(k,k);
    end
    for j = k+1:min(k+q,n)
        for i = k+1:min(k+p,n)
            A(i,j)=A(i,j)-A(i,k)*A(k,j);
        end
    end
end
Méthode de Newton

function [xvect,xdif,fx,nit]=newton(x0,nmax,toll,fun,dfun)
err=toll+1;
nit=0;
xvect=x0;
x=x0;
fx=eval(fun);
xdif=[];
while (nit < nmax & err > toll),
nit=nit+1;
x=xvect(nit);
dfx=eval(dfun);
if (dfx == 0), err=toll*1.e-10;
disp(' Stop ');
else,
xn=x-fx(nit)/dfx;
err=abs(xn-x);
xdif=[xdif; err];
x=xn;
xvect=[xvect;x];
fx=[fx;eval(fun)];
end;
end;

Méthode de Newton cas vectoriel

function [x, nit] = newtonsys(F, J, x0, toll, nmax, p)
[n,m]=size(F); nit=0; Fxn=zeros(n,1); x=x0; err=toll+1; for i=1:n,
    for j=1:n, Jxn(i,j)=eval(J((i-1)*n+j,:)); end; end
[L,U,P]=lu(Jxn); step=0; while err > toll
    if step == p
        step = 0;
        for i=1:n;
            Fxn(i)=eval(F(i,:));
            for j=1:n; Jxn(i,j)=eval(J((i-1)*n+j,:)); end
            end
        end
    end
    end
end;
[L,U,P] = lu(Jxn);
else
    for i=1:n, Fxn(i) = eval(F(i,:)); end
end
nit = nit + 1; step = step + 1; Fxn = -P*Fxn; y = forward_col(L,Fxn);
deltax = backward_col(U,y); x = x + deltax; err = norm(deltax);
if nit > nmax
    disp('pas de convergence');
    break
end
end
Bibliographie


201


202


[111] Leslie P. H. (1948), Some further notes on matrices in population mathematics, Biometrika, 35: 213-245


209


Notes de solutions des exercices

Chapitre 1

exercice 1.6.1

i) 
\[ \| A \|_1 = \sup_{x \neq 0} \frac{\sum_j |a_{ij}x_j|}{\sum_j x_j} \leq \sup_{x \neq 0} \frac{\sum_i \sum_j |a_{ij}x_j|}{\sum_j |x_j|} \leq \max_i \sum_j |a_{ij}|. \]

Si on considère \( e_i = (0, 0, \cdots, 1, \cdots, 0)^\top \) alors
\[ \| Ae_i \|_1 = \sum_j |a_{ij}|, \]

et par suite \( \max_i \sum_j |a_{ij}| \leq \| A \|_1. \)

ii) On procède comme dans i).

iii) \( A^\top A \) est une matrice symétrique définie positive, considérons une base orthonormée \( (e_1, \cdots, e_n) \) de vecteurs propres de \( A^\top A \).
\[ \| A \|_2 = \sup_{x \neq 0} \frac{\langle A^\top Ax, x \rangle}{\langle x, x \rangle} \]
\[ = \sup_{x \neq 0} \frac{\sum \lambda_i |x_i|^2}{\sum |x_i|^2} \]
\[ \leq \lambda_n = \rho(A^\top A). \]

On prend ensuite \( x = e_n \) vecteur propre associé à \( \lambda_n \).

exercice 1.6.2. En effet \( L v = \lambda_1 v \).

exercice 1.6.3

1. Il suffit de considérer un vecteur propre \( x \) associé à une valeur propre \( \lambda \).
2. évident.

3. 
\[ \|A\|_* = \sup_{x \neq 0} \frac{\|Ax\|_*}{\|x\|_*} = \sup_{x \neq 0} \frac{\|BAx\|_\infty}{\|Bx\|_\infty}. \]

\[ B = D^{-1} P \implies \|A\|_* = \sup_{x \neq 0} \frac{\|D^{-1}PAx\|_\infty}{\|D^{-1}Px\|_\infty} = \sup_{x \neq 0} \frac{\|D^{-1}UDD^{-1}Px\|_\infty}{\|D^{-1}Px\|_\infty} = \sup_{x \neq 0} \frac{\|D^{-1}Uy\|_\infty}{\|y\|_\infty} \leq \|D^{-1}U\|_\infty. \]

4. Il existe une norme matricielle \( \|\cdot\|_* \) telle que \( \rho(A) = \|A\|_* \).

5. la \( j^{\text{ème}} \) colonne de \( A^{-1} \) est \( (2^{j-2}, 2^{j-3}, \cdots, 1, \cdots, 0)^\top \) 1 est situé à la \( j^{\text{ème}} \) position.

\[ \|A\|_1 = n \text{ et } \|A^{-1}\|_1 = 2^{n-1}. \]

6. 
\[ \|Ax\|_\infty = \max_j \left| \sum_i a_{ij}x_i \right| \geq |a_{ii}| - \sum_{i \neq j} |a_{ij}| \max_j |x_j| \geq \delta \|x\|_\infty. \]

On prend \( y = A^{-1}x \).

7. \( \|A\|_\infty = 4 + \varepsilon \), le 1 donne \( \lambda_4 \leq 4 + \varepsilon. \)

On prend dans le 2, \( B = I \).

Chapitre 2

exercice 2.12.1

1. 
\[ a_{ij}^{(2)} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}, \quad i, j = 2, \cdots, n. \]
\[ a_{ii}^{(2)} = a_{ii} - \frac{a_{i1}}{a_{11}}a_{1j}, \quad i = 2, \cdots, n. \]

2. 
\[ |a_{ij}^{(2)}| > |a_{jj}| - \frac{a_{i1}}{a_{11}}|a_{1j}| > \sum_{i \neq j} |a_{jj}| - \frac{a_{i1}}{a_{11}}|a_{1j}| \]
\[ > \sum_{i \neq j} a_{ij}^{(2)} + \frac{a_{i1}}{a_{11}}|a_{1j}| + |a_{1j}| - \frac{a_{i1}}{a_{11}}|a_{1j}| \]
\[ > \sum_{i \neq j} a_{ij}^{(2)} - \sum_{i \neq j} |a_{i1}| \left| \frac{a_{1j}}{a_{11}} \right| + |a_{1j}| \]
\[ > \sum_{i \neq j} a_{ij}^{(2)}. \]
3. Il n’est pas nécessaire de pivoter

exercice 2.12.2
1. Si $\alpha \neq \frac{1}{b^\top a}$ alors $I - \alpha ab^\top$ est inversible.

$$(A + ab^\top)^{-1} = A^{-1} - \frac{A^{-1}ab^\top A^{-1}}{1 + \lambda}$$

où $\lambda = b^\top A^{-1}a$.

2. $(I + m_k e_k^\top)^{-1} = I - m_k e_k^\top$.

3. $(I + m_1 e_1^\top)(I + m_2 e_2^\top) = I + m_1 e_1^\top + m_2 e_2^\top$ car $e_i^\top m_k = 0$.

On procède ensuite par récurrence.

exercice 2.12.3 Appliquer les résultats du cours.
exercice 2.12.4 Appliquer les résultats du cours.

exercice 2.12.5 On montre que $\|B^{-1}\|_1 = 2$, les propriétés du conditionnement donnent

$$\|y - x\|_1 \leq \frac{\|B^{-1}\|_1 \|E\|_1 \|b\|_1}{1 - \|B^{-1}\|_1 \|E\|_1 \|b\|_1}.$$  

**Chapitre 3**

exercice 3.8.1

i) $A^{-1} = (I - R)^{-1}B$.

ii) $A^{-1} - B = A^{-1}R$.

iii) $\hat{x} - x = A^{-1}r$.

exercice 3.8.2 Procéder par récurrence.

exercice 3.8.3

1. $A$ est contractante donc admet un point fixe.

2. Calcul facile.

3. (a) $(I - M)^{-1}b(x)$ est contractante car $\|(I - M)^{-1}\|_\infty \leq 3$ et $\|b\|_\infty \leq \frac{1}{4}$.

(b) $k = \frac{24}{33}$.

(c) On peut améliorer la vitesse de convergence en prenant $\varepsilon b(x)$ avec $\varepsilon << 1$.

exercice 3.8.4 Exemples de matrices dont les méthodes de Jacobi et Gauss-Seidel ne convergent pas simultanément.

exercice 3.8.5
1. Utiliser le théorème 3.3.4.

2. Posons $T_S = T_J^2$ où $x^{(n+1)} = T_Sx^{(n-1)} - Tb$.
   Si $\lambda$ est valeur propre de $T_J$ alors $\lambda^2$ est valeur propre de $T_S$ donc $\rho(T_S) < 1$ ainsi convergence. La méthode converge rapidement que la méthode de Jacobi.

exercice 3.8.6

I il suffit qu’il existe une norme matricielle telle que $\|I - \alpha A\| < 1$, ceci est assuré si on prend $\|\cdot\|_\infty$.

II 1) $1 + \lambda_i \alpha$ est valeur propre de $I - \alpha A$ par suite pour que la méthode converge, il faut et il suffit que $\alpha < \frac{2}{\lambda_1}$.

2) $\alpha_s = \frac{2}{\lambda_1 + \lambda_n}$.

exercice 3.8.7

1. $M(r, \omega) = (D - rL)^{-1}((1 - \omega)D + (\omega - r)L + \omega U)$.

2. $\theta = 0$ et $\omega = 1$ donnent Jacobi.
   $r = \omega = 1$ donnent Gauss-Seidel.
   $r = \omega$ donne relaxation.

3. il suffit d’écrire.

4. Comme il s’agit de matrices tridiagonales, on utilise la propriété
   $$\det(\alpha D - \beta L - U) = \det(\alpha D - \beta \mu L - \mu^{-1} U).$$

exercice 3.8.8

1. On suppose ici que les $a_i$ sont positifs. Vous pouvez le montrer directement en regroupant dans la somme $\langle A(h)x, x \rangle$ des termes positifs, ou utiliser un théorème qui assure qu’une matrice est définie positive si et seulement si ces éléments diagonaux sont positifs.

2. Voir chapitre 3.

exercice 3.8.9

1. (a) $rI + H$ est définie positive et hermitienne donc inversible.

(b) $\|(rI - H)(rI + H)\|_2 = \rho((rI - H)(rI + H)) = \max_j \left| \frac{r - \lambda_j}{r + \lambda_j} \right|$. 

214
2. (a) $T = (rI + H_2)^{-1}(rI - H_1)(rI + H_1)^{-1}(rI - H_2)\), $ c = (rI + H_2)^{-1}(rI - H_1)(rI + H_1)^{-1}(rI - H_2)b + (rI + H_2)^{-1}b$. 

(b) $\rho(T) \leq \max_j \left| \frac{r - \lambda_j}{r + \lambda_j} \right| \max_j \left| \frac{r - \mu_j}{r + \mu_j} \right|$, où $\lambda_i$ et $\mu_i$ désignent les valeurs pro-

(c) $x_k$ converge vers $x^*$ solution de $(I + T)^{-1}(I - T)x^* = b$.

(d) $\min_{r \geq 0} \max_{a \leq x \leq b} \left| \frac{(r-x)^2}{(r+x)^2} \right| = \min_{r \geq 0} \max_{a \leq x \leq b} \left( \frac{(r-a)^2}{(r+a)^2}, \frac{(r-b)^2}{(r+b)^2} \right)$, et $r_{opt} = \sqrt{ab}$. 

exercice 3.8.10

1. (a), (b), (c), (d) calculs simples.

2. On utilise le (d) et le fait que $M^{-1}Nx = \lambda x$.

3. (a) $T = (D - L)^{-1}L^\top$, $T_1 = T = T = D^{-1/2}TD^{-1/2} = D^{-1/2}(D - L)^{-1}D^{-1/2} = (I - D^{-1/2}LD^{-1/2})^{-1}L^\top D^{-1/2} = (I - L_1)^{-1}L_1^\top$. 

(b) $\lambda = \frac{a - ib}{1 - a + ib}$.

(c) On a $x^HL_1x + x^HL_1^\top x = x^HD^{-1/2}(L_1 + L_1^\top)D^{-1/2}x = 2a$, or $A = D - L - L^\top$ est définie positive, donc $x^H D^{-1/2} Ax^HD^{-1/2} > 0$ ainsi $1 - 2a > 0$ et par suite $|\lambda| < 1$ et $\rho(T) < 1$.

(d) La méthode de Gauss-Seidel est convergente.

Chapitre 4

exercice 4.4.1 les points d’équilibre sont 0 et $\pm \sqrt{\alpha}$ si $\alpha > 0$.

Si $\alpha > 0$ alors 0 n’est pas un point d’attraction.

Si $\alpha = 0$ alors 0 est un point d’attraction si $s > 1$.

Le point $x^* = \pm \sqrt{\alpha}$ est un point d’attraction si $s < 1$.

exercice 4.4.2

1. $\alpha = \frac{\sqrt{5} - 1}{2}$, $\beta = 1 - \alpha$.

2. Posons $L_1 = b - a$, $x_1 = a$, $x_4 = b$, $x_2 = x_0 + L_1\beta$, $x_3 = x_4 - L_1\beta$.

Si $f(x_3) < f(x_2)$ on remplace $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ par $\begin{pmatrix} x_2 \\ x_3 \\ x_4 - \alpha\beta L_1 \\ x_4 \end{pmatrix}$.
Si $f(x_3) > f(x_2)$ on remplace \[
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
\quad \text{par} \quad \begin{pmatrix}
x_1 \\
x_1 + \alpha \beta L_1 \\
x_2 \\
x_3
\end{pmatrix}
\]
\[
\min_{x \in [a,b]} = \frac{x_1 + x_2}{2}.
\]

**exercice 4.4.6**

1. $\Phi(x) = x^3 - x^2 - 1, |\Phi'(x)| \geq 1$.

2. $\Phi(x) = 1 + \frac{1}{x} + \frac{1}{x^2}, |\Phi'(x)| < 1$.

**exercice 4.4.3**

1. Soit $\lambda > 0$.
   Si $1 + b > 0$ alors il existe un seul point d’équilibre $P^* = \frac{2}{\sqrt{\lambda}}$, il est stable si $\lambda < \frac{1}{1 + b}$.
   Si $b = -1$, le système est stationnaire.
   Si $1 + b < 0$ alors il existe deux points d’équilibre $P^* = 0$ et $P^{**} = \frac{1}{\sqrt{-r}}$, où $r = -(1 + b)$, $0$ est stable si $\lambda < \frac{1}{-(1 + b)}$, $P^{**}$ est stable si $-2 < b < -1$.

2. Les points d’équilibre sont $P^* = 0$ et $P^{**} = \frac{1}{\alpha} \log \lambda$.
   $0$ est stable si $0 < \lambda < 1$.
   $P^{**}$ est stable si $e^{-2} < \lambda < 1$.

3. Les points d’équilibre sont $P^* = 0$ et $P^{**} = K$.
   $0$ est instable si $r > 0$.
   $P^{**}$ est stable si $0 < r < 2$.

**exercice 4.4.9**

1. Utiliser la formule de Taylor.

2. Utiliser la formule de Taylor en considérant $f(x_n)$ et $f(x_{n-1})$.
   La méthode est d’ordre supérieur à 1.

3. On peut adapter la méthode de convergence la sécante de la manière suivante: $f(a)f(b) < 0$ et on définit la suite
   \[
x_{n+1} = x_n - \frac{f(x_n)(x_n - x^*)}{f(x_n) - f(x_{n-1})},
   \]
   où $x^*$ est le dernier $x_j, j = n - 1, \cdots, 1$ calculé tel que $f(x_n)f(x^*) < 0$.
exercice 4.4.10

1. Affirmatif.

2. (a), (b) voir (paragraphe 4.1.3, page 85).
   (c) Si $\lambda = \frac{1}{p}$ on retrouve la méthode de Newton.

3. 

$$\begin{align*}
  x_{n+1} &= x_n - \frac{u(x_n, y_n) \frac{\partial u}{\partial x}(x_n, y_n) + v(x_n, y_n) \frac{\partial v}{\partial x}(x_n, y_n)}{\left(\frac{\partial u}{\partial x}(x_n, y_n)\right)^2 - \left(\frac{\partial v}{\partial x}(x_n, y_n)\right)^2} \\
  y_{n+1} &= y_n - \frac{v(x_n, y_n) \frac{\partial u}{\partial x}(x_n, y_n) - u(x_n, y_n) \frac{\partial v}{\partial x}(x_n, y_n)}{\left(\frac{\partial u}{\partial x}(x_n, y_n)\right)^2 - \left(\frac{\partial v}{\partial x}(x_n, y_n)\right)^2}
\end{align*}$$

exercice 4.4.14

a) Les points d’équilibre sont $P^* = \frac{K\left(r + \sqrt{r^2 - \frac{4rE}{K}}\right)}{2r}$ et $P^{**} = \frac{K\left(r - \sqrt{r^2 - \frac{4rE}{K}}\right)}{2r}$.
   Le point $P^*$ est stable et $P^{**}$ est instable.

b) Si $E > \frac{rK}{4}$, le système n’admet pas de points d’équilibre et $P(t)$ est strictement décroissante.

c) Les points d’équilibre sont $P^* = 0$ et $P^{**} = \frac{K(r - E)}{r}$, si $r > E$.
   Le point $P^*$ est instable et $P^{**}$ est stable.

c) Les points d’équilibre sont $P^* = \frac{K\left((r - 1) + \sqrt{(r - 1)^2 - \frac{4rE}{K}}\right)}{2r}$ et

$$P^{**} = \frac{K\left((r - 1) - \sqrt{(r - 1)^2 - \frac{4rE}{K}}\right)}{2r},$$
   si $(r - 1)^2 > \frac{4rE}{K}$.
   Le point $P^*$ est stable si $0 < (r - 1)^2 - \frac{4rE}{r} < 4$ et $P^{**}$ est instable.

exercice 4.4.7

a) $\frac{1}{101} < \alpha < 51$.

b) $R = 51, r = \frac{1}{101}$.

217
c) Le nombre de racines négatives $= 0 - 2k$, donc une seule possibilité: 0.

d) Appliquer le théorème de Newton.

exercice 4.4.4. Il suffit de tracer les graphes des fonctions $f$ et $g$ (les points d’équilibre sont l’intersection des 2 graphes) et de dériver les conditions de stabilité en les imposant à la fonction $\Phi(x) = x + rx - \frac{\beta x^2}{\alpha + x^2}$.

exercice 4.4.2

1. Utiliser la Formule de Taylor.
2. $C = \frac{g''(x)}{p!}$.

exercice 4.4.4

1. Utiliser la Formule de Taylor.
2.

exercice 4.4.13

1. (a) $P(y) = g'(y_n)y + (g(y_n) - g'(y_n)y_n)$. 
   (b) $P(y) = \frac{y}{f''(f^{-1}(x_n))} + (x_n - \frac{f^{-1}(x_n)}{f''(f^{-1}(x_n))})$. 
   (c) Méthode de Newton.
2. Méthode de la sécante.
3. (a) $P(y) = \frac{g''(y_n)}{2}y^2 + (g(y_n) - g''(y_n)y_n)y + g(y_n) + \frac{g''(y_n)}{2}y_n^2 - g'(y_n)y_n$. 
   (b) $g'' = \frac{f''(f^{-1})^3}{f''(f^{-1})^3}g' = \frac{1}{f''(f^{-1})}$. 
   (c) $C = \frac{-f''(\alpha)}{6f'''(\alpha)} + \frac{f''(\alpha)}{2(f'(\alpha))^2}$.

Chapitre 5

1. a) $f(x_k)s_k = F(x_k), x_{k+1} = x_k + s_k$. 
   b) $p^{-1} = I + \frac{\alpha uv^\top}{1 - \alpha v^\top u}$. 
   c) $B_{k+1}^{-1} = B_k^{-1} - \frac{B_k^{-1} \alpha uv^\top B_k^{-1}}{1 + \alpha v^\top B_k^{-1} u}$. 

218
d) $B_{k+1}^{-1} y_k = s_k$.

2. a), b), c) définissent la méthode des directions conjuguées décrite dans le chapitre cinq.

Chapitre 6

exercice 6.5.2 Il suffit de considérer une base orthonormée de vecteurs propres.

exercice 6.4.3

(a) Les valeurs propres de $A_1$ sont $0, \lambda_2, \ldots, \lambda_n$.

(b) Considérer les suites $x^{(k)} = \frac{A_1 x^{(k-1)}}{m_k}$, $m_k = \|A_1 x^{(k-1)}\|_\infty$.

(c) Cette méthode peut être utilisée pour calculer d’autres valeurs propres de $A$.

(a) Les valeurs propres de $A_1$ sont $0, \lambda_2, \ldots, \lambda_n$.

(b) En effet $A w_i = \lambda_i w_i$, $i = 1, \ldots, n$.

(c) La $i^{\text{ème}}$ ligne de $A'_1$ est nulle et ses valeurs propres sont $\lambda_2, \ldots, \lambda_n$. La méthode des puissances peut être utilisée à condition que $|\lambda_2| > |\lambda_1|$.

(d) Si $w'_2 = (w'_1, \ldots, w'_{n-1})^\top$ alors $w_2 = (w'_1, \ldots, w'_{i-1}, 0, w'_i, \ldots, w'_{n-1})^\top$

(e) utiliser le 2) c).

(f) On peut commencer recommencer ce processus sur $A'_1$ et obtenir ainsi de proche en proche une matrice d’ordre 1.

Chapitre 7

exercice 7.7.1

si $\alpha \neq 1$ et $\beta \neq -\frac{1}{2}$ l’ordre $p = 1$ et $C_2 = (\alpha - 1) \left(\beta + \frac{1}{2}\right)$

si $\alpha \neq 1$ et $\beta = -\frac{1}{2}$ l’ordre $p = 2$ et $C_3 = \frac{1}{12} (\alpha - 1)$

si $\alpha = 1$ et $\beta \neq -\frac{1}{2}$ l’ordre $p = 2$ et $C_3 = - \left(\beta + \frac{1}{2}\right)$

si $\alpha = 1$ et $\beta = -\frac{1}{2}$ l’ordre $p = 3$ et $C_4 = - \frac{1}{12}$
exercice 7.7.2 On cherchera une solution particulière du système différentiel (resp. du système aux différences) de la forme:

\[ \phi(x) = \left( \begin{array}{c} ax + b \\ cx + d \end{array} \right) \quad \text{(resp. } \psi_n = \left( \begin{array}{c} an + b \\ cn + d \end{array} \right) \text{)} \]

où \( a, b, c \) et \( d \) sont des constantes réelles.

exercice 7.7.9

1. \( \alpha_0 = -5, \alpha_1 = 4, \beta_0 = 2, \beta_1 = 4 \), ordre 3.
2. non zéro stable car \(|\alpha_0| = 5\).
3. 4,5,6,7,8 il suffit d’écrire.
4. diverge car \( \alpha_0 = -5 \).

exercice 7.7.4
Par passage à la limite \( y_n \to \frac{c}{1 - \mu} \).

exercice 7.7.5

A)  
1. Considérer \( f(x) = m \log(1 + x) - mx \).
2. Procéder par récurrence.

B)  
1. Identique à celle du théorème du cours avec \( y(x_0) - w_0 = \delta_0 \) et \( t = \frac{h^2 M}{2} + |\delta| \).
2. \( h_0 = \sqrt{\frac{2\delta}{M}} \).

exercice 7.7.6 On considère une base de vecteurs propres de \( A \).

Pour \( S_1 \) : \( h < \frac{2 \Re(\lambda)}{|\lambda|^2} \).

Pour \( S_1 \) : inconditionnellement stable.

exercice 7.7.7

1. Il suffit de développer en utilisant la formule de Taylor.
2. La condition est nécessaire et suffisante (\( c_0 = 0 \) et \( c_1 = 0 \)).

exercice 7.7.8

1. Méthode d’ordre 1.
2. 
3.
Chapitre 8

problème 8.0.1


2. Voir (théorème 3.3.4,59).

3. (a) $a > 1$.

   (b) $J = \begin{pmatrix} 0 & -a & -a \\ a & 0 & -a \\ a & a & 0 \end{pmatrix}$, $J$ converge si $a < 1/2$.

   (c) $G = \begin{pmatrix} 0 & 0 & 0 \\ -a & 0 & 0 \\ -a & -a & 0 \end{pmatrix}$, $\rho(G) = 0$, si $a \neq 0$ la méthode de Gauss Seidel converge plus vite que la méthode de Jacobi.

4. (a) Si $\exists$ une norme telle que $\|A\| < 1$ et $\|B\| < 1$.

   (b) $C = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$, $c = (a, b)^\top$, $\rho(AB) \leq \rho(C)$.

   (c) $\rho(AB) < 1$ ou $\|AB\| < 1$.

   (d) $D = AB, d = Aa + b, \rho(AB) = \rho(D)$.

   (e) Le deuxième schéma converge plus rapidement que le premier.

Problème 8.0.2

1. On montre qu’elles sont linéairement indépendentes.

2. Il suffit d’écritre, $D_n = 0, C_n = A^{-1}$.

3. Si $D_k = 0$ alors $C_k = A^{-1}$.

4. Utiliser l’indication, la suite étant ainsi construite alors $C_n = A^{-1}$.

Problème 8.0.3

exercice 1

1. $M_J = \begin{pmatrix} 0 & -b & -b \\ -b & a & a \\ -b & 0 & -a \\ -b & -a & 0 \end{pmatrix}$
2. Méthode convergente si \( \|M_j\|_\infty = \frac{2b}{a} < 1 \) donc \( a > 2b \).

3. \( A \) est inversible car elle est à diagonale dominante, \( C(A) \leq \frac{a + 2b}{a - 2b} \).

exercice 2

1. Méthode d’ordre 1.


exercice 3. Il s’agit de simples calculs, il suffit d’écrire.

Examen exercice 1

1. \( \|\delta x\| \leq \frac{\|F\|\|b\|}{(1 - \|E\| - \|F\|)(1 - \|E\|)} = \frac{4\varepsilon}{1 - 2\varepsilon}\|b\| \).

2. Méthode convergente si \( \|M_j\|_\infty = \frac{2b}{a} < 1 \) donc \( a > 2b \).

3. On distingue deux cas.
   
   Si \( \exists i \) tel que \( \lambda_i = \mu \) c’est fini.

   Sinon soit \( v \) un vecteur propre de \( A \) associé à \( \mu \) on montre alors que

   \[
   v = -P^{-1}\text{diag} \left( \frac{1}{\lambda_i - \mu} \right) P\delta A v \text{ et parsuite } \min_i |\lambda_i - \mu| \leq C(P)\|\delta A\|.
   \]

exercice 2

1. (a) \( M(\theta, \omega) = (D - \theta L)^{-1} ((1 - \omega) D + (\omega - \theta) L + \omega U) \).
   
   (b) \( \theta = 0 \) et \( \omega = 1 \) donnent Jacobi.

   \( \theta = \omega = 1 \) donnent Gauss-Seidel.

2. (a) \( \theta = 1 \), si \( \mu \) est valeur propre de \( M(1, \omega) \) et \( \mu \) est valeur propre de \( M(1, 1) \) alors

   \( \mu = 1 - \omega + \omega \lambda \).

   (b) Affirmatif.

3. (a) Si \( \mu \) est valeur propre de \( M(0, 1) \) et \( \lambda \) est valeur propre de \( M(0, \omega) \) alors

   \( \lambda = 1 - \omega + \omega \mu \).

   (b) Si la méthode de Jacobi converge alors \( |\mu_i| < 1 \) pour tout \( i \), \( A \) étant symétrique on peut ranger ses valeurs propres par ordre croissant \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n \), donc

   \[ \frac{1}{1 - \mu_1} \leq \frac{1}{1 - \mu_i} \quad \forall i. \]

   La méthode \( (*) \) converge si \( -1 < 1 - \omega + \omega \mu_i < 1 \) et ceci est vérifié si

   \[ 0 < \omega < \frac{2}{1 - \mu_1} \].

222
Index

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acacias</td>
<td>Bailey</td>
</tr>
<tr>
<td>Age</td>
<td>Bartlett</td>
</tr>
<tr>
<td>Akerman</td>
<td>Becker</td>
</tr>
<tr>
<td>Algorithme ...</td>
<td>Bellazi</td>
</tr>
<tr>
<td>Anderson</td>
<td>Belt</td>
</tr>
<tr>
<td>Applications</td>
<td>Berman</td>
</tr>
<tr>
<td>Approximation</td>
<td>Bernardelli</td>
</tr>
<tr>
<td>Axelsson</td>
<td>Bernouilli</td>
</tr>
<tr>
<td></td>
<td>Beverton-Holt</td>
</tr>
<tr>
<td></td>
<td>Bolle</td>
</tr>
<tr>
<td></td>
<td>Bolker</td>
</tr>
<tr>
<td></td>
<td>Bonnet</td>
</tr>
<tr>
<td></td>
<td>Boutayeb</td>
</tr>
<tr>
<td></td>
<td>Brauer</td>
</tr>
<tr>
<td></td>
<td>Broyden</td>
</tr>
<tr>
<td></td>
<td>broyden</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrée . 6, 22, 24, 31, 33, 47–49, 51–53, 74–75, 77, 86, 142</td>
<td>Décomposition ... 44, 46, 75, 142, 143, 195</td>
</tr>
<tr>
<td>Castillo-Chavez</td>
<td>Définie positive</td>
</tr>
<tr>
<td>Caswell</td>
<td>Définie positive 9, 34, 41, 46, 58, 59, 74, 75, 78, 123</td>
</tr>
<tr>
<td>Cauchy</td>
<td>Danilevski</td>
</tr>
<tr>
<td>Cayley-Hamilton</td>
<td></td>
</tr>
<tr>
<td>Chebyshev</td>
<td></td>
</tr>
<tr>
<td>Choleski</td>
<td></td>
</tr>
<tr>
<td>Ciarlet</td>
<td></td>
</tr>
<tr>
<td>Classique</td>
<td></td>
</tr>
<tr>
<td>Collatz</td>
<td></td>
</tr>
<tr>
<td>Comparaison</td>
<td></td>
</tr>
<tr>
<td>Conditionnement</td>
<td></td>
</tr>
<tr>
<td>Conolly</td>
<td></td>
</tr>
<tr>
<td>Consistance</td>
<td></td>
</tr>
<tr>
<td>Convergence</td>
<td></td>
</tr>
<tr>
<td>Convergente</td>
<td></td>
</tr>
<tr>
<td>Conway</td>
<td></td>
</tr>
<tr>
<td>Cooke</td>
<td></td>
</tr>
<tr>
<td>Cramer</td>
<td></td>
</tr>
<tr>
<td>Creuse</td>
<td></td>
</tr>
</tbody>
</table>

223
Norme . 6, 11, 12, 22, 36, 51, 52, 58, 75, 77, 115, 122
Numérique . . . . . . . . 24, 37, 81, 91, 101, 121

O
Okubo ........................................ 182
Oliveira ...................................... 131
Opération ................................. 24, 25, 47
Optimisation ............................. 119
Orthogonale . . . . . . 7, 26, 35–37, 46, 145, 147
Ostrowski ................................. 72

P
Parker ........................................ 180
Perlis ........................................ 27
Permutation ................. 7, 27, 31–33, 146
Perron ................................. 14, 21
Perturbation .......................... 36, 37
Pielou .......................................... 182
Pivot ........................................ 29–33, 47
Plemmons .................................. 21, 72
Point fixe ............... 82, 115, 116, 194
Polynôme8, 60, 81, 91, 94, 96, 97, 100, 101, 138, 139
Primitive .................................. 7, 14
Produit scalaire ................. 6
Puissance .............................. 141

Q
QR ............................................ 47, 143, 144
Quasi-dominante ...................... 8

R
Réductible ................................. 7
Résolution 24, 25, 30, 36, 40, 41, 44, 46, 50, 63, 122, 124, 139
Racine . 8, 9, 60, 62, 64, 65, 78, 86, 94–101, 110, 113, 138
Rayon spectral . . . . . . 6, 8, 22, 52, 59, 74, 137
Rectangulaire ......................... 43–46, 48
Regula falsi ............................... 88
Relaxation . . . . . . 51, 57–59, 61–63, 75, 78, 118, 193
Ricker ........................................ 107
Rosenberg ............................... 71, 72
Ross ......................................... 181
Rotation ................................. 29
Runge-Kutta . . . . . . . . . . . . . . .163, 164, 166, 167

S
Schaefer ..................................... 105
Schur ......................................... 9, 10, 12
Scott ......................................... 182
Scuda ......................................... 131
SEIR ......................................... 170, 174
SEIRS ......................................... 170
Semblable ................................. 8
Serje ......................................... 179
Sherman ..................................... 38
SI ............................................. 170
Singulière . . . . . . 9, 22, 36, 44, 46, 48
SIR ........................................... 170, 172
SIS ........................................... 170
Smilatova .................................. 132
Smith ......................................... 136, 182
Sous-matrice ......................... 33, 34
Southwood ............................... 109
Spectre ...................................... 8
Stable ......................................... 36, 137
Stationnaire ............................. 156
Steffenson ............................... 90
Stein ......................................... 71, 72
Stetter ....................................... 160
Stieljes ................................. 7, 72
Stochastique ............................ 8
Structure .................................. 37, 46
Sturm ....................................... 99, 100, 114, 193

226
Subordonnée ................................. 12, 22, 36
Sujan ........................................... 132
Supérieure 14, 22, 25, 31, 33, 39, 77, 96, 142, 143
Symétrique ............... 7, 26, 34, 46, 74, 98, 145
Système . . 14, 24, 25, 29–31, 34, 36, 38, 40, 44–47, 49, 50, 73, 74, 77, 78, 116, 124, 137, 139
Taylor ................. 81, 96, 120, 162–164
Triangulaire .. 22, 25, 29–34, 38, 39, 46, 76, 77, 137, 142, 143
Tridiagonale . . 39, 49, 60, 61, 73, 75, 77, 78, 113
Troncature ...................... 159, 167, 168
Unitaire ......................... 7, 35, 43, 48
Valeur propre ...................... 7–9, 11, 14, 22, 23, 60–62, 74, 76–79, 114, 137, 138, 141–143, 147
Van Loan ......................... 72
Varga .................. 21, 72
Vecteur propre . . . 8, 14, 137, 139, 141, 147
Verhulst ......................... 106
Verner ........................ 167
Vitesse ...... 14, 53, 58, 74, 77, 78, 81, 108
Volterra ......................... 131, 132
von Mises ......................... 72
Wegstein ......................... 84, 85
White ......................... 72
Whittaker ......................... 118
Wickwire ......................... 182
Young .............................. 72
Yule .............................. 181
Ziegler ......................... 131
Verhulst ........................ 106
Verner ........................ 167
Vitesse ...... 14, 53, 58, 74, 77, 78, 81, 108
Volterra ......................... 131, 132
von Mises ......................... 72