

Année 2010

---

# Méthodes numériques pour les écoulements incompressibles

---

Patrick Le Quéré et Bérengère Podvin



*Température dans une cavité circulaire différentiellement chauffée*  
*S. Xin et P. Le Quéré*

## Résumé

Ces notes de cours présentent les fondements mathématiques et physiques de la résolution numérique des équations de Navier-Stokes incompressibles. Une attention particulière est donnée aux méthodes spectrales. Nous commençons par rappeler les équations fondamentales et les conditions dans lesquelles un écoulement peut être considéré comme incompressible. Nous nous intéressons ensuite à la classification des équations aux dérivées partielles et nous montrons que les équations de Navier-Stokes discrétisées sont de type elliptiques en espace. Dans une deuxième partie, nous abordons la discrétisation des EDP en temps et en espace. Nous introduisons les idées générales des approches par différences finies, volumes finis et éléments finis. La recherche de la solution dans un espace approprié nous conduit à la présentation des méthodes spectrales et à leur application pour la résolution des équations elliptiques. Dans une troisième partie, nous nous intéressons à l'erreur de discrétisation et aux méthodes de résolution itératives. Enfin, dans la dernière partie, nous définissons le problème de Stokes et abordons les méthodes de résolution de la pression.

---

## Quelques ouvrages de références

- Mécanique des Fluides (physique):  
**An introduction to Fluid Dynamics**, Batchelor , Cambridge University Press 2000 (réédition).
- Mécanique des Fluides (numérique):
  - **Computation Fluid Mechanics and Heat Transfer**, Tannehill, Andresen and Pletcher, Hemisphere 1984
  - **Numerical computation of Internal and External Fluid Dynamics**, C. Hirsch, Butterworth-Heinemann, 2007 (réédition)
  - **Spectral Methods in Fluid Dynamics**, Canuto, Hussaini, Quarteroni, Zang, Springer-Verlag 1991.
- Méthodes numériques (général) :
  - **A multigrid tutorial**, Briggs, Henson, McCormick, SIAM Monographs
  - **Numerical Recipes: The Art of Scientific Computing**, Press, Teutolsky, Vetterling, Flannery , Cambridge University Press, 1992

# Table des matières

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Le caractère elliptique des équations de Navier-Stokes incompressible</b>   | <b>4</b>  |
| 1.1      | Equations de Navier-Stokes . . . . .   | 4         |
| 1.2      | Ecoulement incompressible . . . . .  | 5         |
| 1.2.1    | Définition . . . . .   | 5         |
| 1.2.2    | Conditions d'incompressibilité . . . . .                                       | 5         |
| 1.2.3    | Rôle de la pression . . . . .  | 7         |
| 1.3      | Généralités sur les équations aux dérivées partielles . . . . .                | 8         |
| 1.3.1    | Classification des EDP . . . . .   | 8         |
| 1.3.2    | Schéma conservatif - Equations en forme conservative . . . . .                 | 10        |
| <b>2</b> | <b>Discrétisation des EDP</b>  | <b>12</b> |
| 2.1      | Discrétisation compacte en espace . . . . .                                    | 12        |
| 2.1.1    | Discrétisation par différences finies . . . . .                                | 12        |
| 2.1.2    | Approche par éléments finis . . . . .  | 14        |
| 2.1.3    | Approche par volumes finis . . . . .   | 17        |
| 2.2      | Discrétisation temporelle . . . . .  | 19        |
| 2.2.1    | Les schémas d'Euler . . . . .  | 20        |
| 2.2.2    | Les schémas de Runge-Kutta . . . . .   | 20        |
| 2.2.3    | Discrétisation du terme source . . . . .                                       | 21        |
| 2.3      | Discrétisation non compacte en espace . . . . .                                | 22        |
| 2.3.1    | La recherche d'un espace où définir la solution . . . . .                      | 22        |
| 2.3.2    | Séries de Fourier . . . . .  | 23        |
| 2.3.3    | Polynômes orthogonaux . . . . .  | 28        |
| 2.4      | Résolution spectrale d'équations elliptiques . . . . .                         | 36        |
| 2.4.1    | Généralités . . . . .  | 36        |
| 2.4.2    | Méthodes spectrales . . . . .  | 37        |
| 2.4.3    | Méthodes de collocation . . . . .  | 40        |
| 2.4.4    | Résolution de problèmes bidimensionnels . . . . .                              | 42        |
| <b>3</b> | <b>L'erreur de discrétisation</b>  | <b>44</b> |
| 3.1      | Schémas numériques: notions de stabilité, convergence et consistance . . . . . | 44        |

|          |  |
|----------|--|
|          | 3  |
| 3.2      | Analyse de stabilité . . . . . 45                        |
| 3.2.1    | Analyse de Von Neumann . . . . . 45                      |
| 3.2.2    | Stabilité matricielle . . . . . 49                       |
| 3.3      | Résolution itérative d'un problème linéaire . . . . . 50 |
| 3.3.1    | Conditionnement de l'opérateur . . . . . 50              |
| 3.3.2    | Méthodes itératives . . . . . 52                         |
| 3.3.3    | Méthodes multi-grille . . . . . 55                       |
| <b>4</b> | <b>Résolution du problème de Stokes 57</b>               |
| 4.1      | Formulation . . . . . 57                                 |
| 4.2      | Equation de Poisson pour la Pression . . . . . 58        |
| 4.2.1    | Principe . . . . . 58                                    |
| 4.2.2    | Mise en œuvre . . . . . 60                               |
| 4.3      | Opérateur d'Uzawa . . . . . 63                           |
| 4.3.1    | Principe . . . . . 63                                    |
| 4.3.2    | Choix des espaces d'approximation . . . . . 64           |
| 4.3.3    | Mise en œuvre . . . . . 67                               |

# Chapitre 1

## Le caractère elliptique des équations de Navier-Stokes incompressible

### 1.1 Equations de Navier-Stokes

On considère un écoulement de fluide de densité  $\rho$ , et on notera  $u_i$  les composantes du champ de vitesse et  $p$  la pression du fluide.

Les équations de Navier-Stokes pour un fluide Newtonien expriment pour un volume de contrôle fixe dans l'écoulement

- la conservation de la masse

$$\frac{\partial \rho}{\partial t} + \frac{(\rho \partial u_j)}{\partial x_j} = 0 \quad (1.1)$$

- la conservation de la quantité de mouvement

$$\frac{\partial(\rho u_i)}{\partial t} + u_j \frac{\partial(\rho u_i)}{\partial x_j} = \frac{\partial p}{\partial x_j} + \nu \frac{\partial u_i}{\partial x_j x_j} \quad (1.2)$$

- la conservation de l'énergie  $E = E + \frac{1}{2} u_i u_i$

$$\frac{\partial E}{\partial t} = W + Q \quad (1.3)$$

où  $W$  est la quantité de travail reçu par le système et  $Q$  est la dissipation.

Les inconnues du problème sont  $\rho, u_{i,i=1,2,3}, p$ . Cette équation est souvent remplacée par une équation d'état. Cette équation d'état peut être l'équation d'un gaz parfait. Pour résoudre le problème de manière générale, il faut exprimer la conservation de l'énergie, le plus souvent à l'aide d'une équation d'état permettant de relier le comportement des variables thermodynamiques.

## 1.2 Écoulement incompressible

### 1.2.1 Définition

Un écoulement est dit **incompressible** si la densité de chaque particule de fluide reste la même au cours du mouvement (ce qui signifie pas que la densité du fluide soit nécessairement constante en un point au cours du temps, ou uniforme en espace!). On a alors

$$\frac{D\rho}{Dt} = \frac{\partial\rho}{\partial t} + u_i \frac{\partial\rho}{\partial x_i} = 0 \quad (1.4)$$

où  $D$  est la dérivée particulaire.

En soustrayant les équations 1.1 et 1.4, on obtient

$$\frac{\partial u_i}{\partial x_i} = 0 \quad (1.5)$$

Un écoulement incompressible est caractérisé par un champ de vitesse à divergence nulle (autrement dit solénoïdal). On comprend ainsi que l'incompressibilité est liée à la vitesse de l'écoulement. On ne peut pas parler de *fluide* incompressible, car ce n'est pas une propriété intrinsèque du fluide.

### 1.2.2 Conditions d'incompressibilité

Dans quelles conditions un écoulement peut-il être considéré comme incompressible? Il faut que le temps caractéristique de la variation de la densité d'une particule de fluide soit très grand devant les autres échelles temporelles de l'écoulement. On a alors

$$\left| \frac{1}{\rho} \frac{D\rho}{Dt} \right| \ll U/L \quad (1.6)$$

Les variations de densité peuvent être exprimées à l'aide de deux variables thermodynamiques. Suivant Batchelor [28], nous utilisons la pression  $p$  et l'entropie  $s$  et exprimons

$$\frac{D\rho}{Dt} = \frac{\partial\rho}{\partial p}|_s \frac{Dp}{Dt} + \frac{\partial\rho}{\partial s}|_p \frac{Ds}{Dt} \quad (1.7)$$

On voit ici apparaître la vitesse du son  $a$  définie par  $a^2 = \frac{\partial p}{\partial \rho}|_s$ . Batchelor [28] a montré que le deuxième terme du côté droit de l'équation est négligeable. On a donc

$$\frac{1}{\rho} \frac{D\rho}{Dt} = \frac{1}{\rho a^2} \frac{Dp}{Dt} \quad (1.8)$$

En utilisant d'une part

$$\frac{Dp}{Dt} = \frac{\partial p}{\partial t} + u_i \frac{\partial p}{\partial x_i} \quad (1.9)$$

de sorte que les variations de la pression peuvent s'exprimer comme la somme de deux contributions

$$\frac{1}{\rho a^2} \frac{\partial p}{\partial t} \ll \frac{U}{L}$$

et

$$\frac{u_i}{\rho a^2} \frac{\partial p}{\partial x_i} \ll \frac{U}{L}$$

En écoulement isentropique la viscosité du fluide est nulle. La conservation de la quantité de mouvement nous conduit à

$$\rho \frac{Du_i}{Dt} = - \frac{\partial p}{\partial x_i} \quad (1.10)$$

Ceci nous permet d'établir que

$$p \sim \rho \frac{UL}{\tau}$$

La première contribution

$$\frac{1}{\rho a^2} \frac{\partial p}{\partial t} \sim \frac{1}{a^2} \frac{UL}{\tau^2} \ll \frac{U}{L}$$

ce qui nous donne

$$\frac{L^2}{a^2 \tau^2} \ll 1$$

où les longueurs d'onde sont très petites devant les longueurs d'onde acoustiques et

$$\frac{U^2}{a^2} \ll \frac{U\tau}{L} \ll 1$$

et la deuxième

$$\frac{u_i}{\rho a^2} \frac{\partial p}{\partial x_i} \sim \frac{U^2}{a^2 \tau} \ll \frac{U}{L}$$

ce qui nous donne

$$\frac{U}{a} \frac{L}{a\tau} \ll 1$$

Une analyse plus fine des variations de vitesse

$$u_i \frac{\partial p}{\partial x_i} = -u_i \frac{Du_i}{Dt} = -u_i \left( \frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} \right)$$

nous conduit à estimer le premier terme comme

$$\frac{U}{a} \frac{L}{a\tau} \ll 1$$

et le second comme

$$\frac{U^2}{a^2} \ll 1$$

On retrouve que le nombre de Mach  $Ma = U/a$  doit être petit devant 1. Si les effets de compressibilité sont de l'ordre de 1%, cela correspond à un nombre de Mach de 0.3.

Un cas particulier qui est fréquemment utilisé est l'approximation de Boussinesq, où on considère des fluctuations de densité exclusivement dues à des variations de température, et suffisamment faibles pour ne pas remettre en question l'hypothèse d'incompressibilité et modifier la conservation de l'énergie. La température est donc considérée comme un scalaire dans l'équation de l'énergie. Les équations à résoudre sont donc

- l'incompressibilité

$$\frac{\partial(u_i)}{\partial x_i} = 0 \quad (1.11)$$

- la conservation de la quantité du mouvement avec une nouvelle force qui est la poussée d'Archimède

$$\frac{\partial(\rho u_i)}{\partial t} + u_j \frac{\partial(\rho u_i)}{\partial x_j} = \frac{\partial p}{\partial x_j} + \nu \frac{\partial u_i}{\partial x_j x_j} + F_i \quad (1.12)$$

où

$$\underline{F} = -g\alpha\Delta T \quad (1.13)$$

avec  $\underline{g}$  représentant la gravité.

- la conservation de l'énergie

$$\rho C_p \frac{\partial T}{\partial t} + u_j \frac{\partial T}{\partial x_j} = k \frac{\partial^2 T}{\partial x_i x_i} \quad (1.14)$$

### 1.2.3 Rôle de la pression

Une conséquence importante de l'incompressibilité pour la résolution des équations est que la pression  $p$  perd sa valeur thermodynamique, elle devient une variable qui assure la solénoïdalité (autrement dit la divergence nulle du champ de vitesse).

Pour mieux comprendre cette situation, considérons les équations de Navier-Stokes sans le gradient de pression

$$\frac{\partial(\rho u_i)}{\partial t} + u_j \frac{\partial(\rho u_i)}{\partial x_j} = \nu \frac{\partial u_i}{\partial x_j x_j} \quad (1.15)$$

On cherche à résoudre cette équation sous la contrainte de solénoïdalité  $\text{div} u = 0$ . L'idée classique est de définir un Lagrangien augmenté

$$\mathcal{L} = |u - u^*|^2 + \lambda \frac{\partial u_i}{\partial x_i} \quad (1.16)$$

où  $u^*$  vérifie

$$\frac{\partial(\rho u^*_i)}{\partial t} + u^*_j \frac{\partial(\rho u^*_i)}{\partial x_j} = \nu \frac{\partial u^*_i}{\partial x_j x_j} \quad (1.17)$$

En minimisant par rapport à  $u$ , on trouve que pour toute variation  $dv$

$$(u - u^*).dv + dv.(u - u^*) + \lambda \frac{\partial dv_i}{\partial x_i} = 0 \quad (1.18)$$

En intégrant par parties, il vient

$$(u - u^*).dv + dv.(u - u^*) - dv_i \frac{\partial \lambda}{\partial x_i} = 0 \quad (1.19)$$

soit

$$u_i = u^*_i - \frac{\partial p}{\partial x_i} \quad (1.20)$$

Le gradient de pression agit donc comme un terme correcteur qui permet de projeter le champ de vitesses dans l'espace des champs à divergence nulle (voir aussi le chapitre 4).



## 1.3 Généralités sur les équations aux dérivées partielles

### 1.3.1 Classification des EDP

On considère une équation aux dérivées partielles de la forme suivante

$$a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial y^2} + c \frac{\partial^2 u}{\partial x \partial y} + d \frac{\partial u}{\partial y} + e \frac{\partial u}{\partial x} + f = 0 \quad (1.21)$$

où les coefficients  $a, b, c, d, e, f$  dépendent éventuellement de la position  $(x, y)$ .

**Définitions:**

La nature des équations dépend du discriminant

$$\Delta = b^2 - 4ac$$

- Si  $\Delta > 0$  les équations sont dites *hyperboliques*.
- Si  $\Delta = 0$  les équations sont dites *paraboliques*.
- Si  $\Delta < 0$  les équations sont dites *elliptiques*.

- Si  $\Delta > 0$ :

$$\frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial x^2} = 0 \quad (1.22)$$

L'exemple canonique est l'équation des ondes (avec la notation  $y = t$ ).

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad (1.23)$$

La solution de cette équation est de la forme

$$u(x, t) = F(x - ct) + G(x + ct) \quad (1.24)$$

Ces problèmes sont des problèmes de marche en temps où à un instant donné, la solution dépend seulement des conditions initiales dans le domaine de dépendance. Les perturbations se propagent à la vitesse  $+/- c$  et la solution à un instant et en un point donné va influencer une zone limitée par cette vitesse de propagation (voir figure 1.1).

Des exemples sont les ondes de choc dans les écoulements transoniques et supersoniques, et ne seront pas étudiés dans le cadre du cours.

- Si  $\Delta = 0$ :

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial y} \quad (1.25)$$

Les équations paraboliques sont des équations de marche en temps (avec  $y = t$ ). La solution dépend des conditions initiales dans tout le domaine. Le domaine de dépendance est donc constitué par l'ensemble des états à  $t' > t$  (voir figure 1.2).

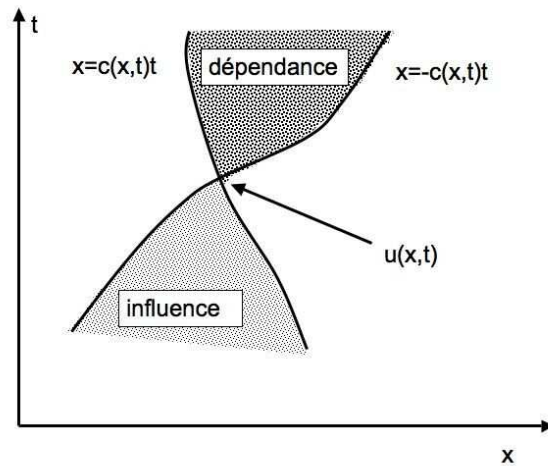


FIG. 1.1 – Domaines de dépendance et d'influence d'une équation hyperbolique

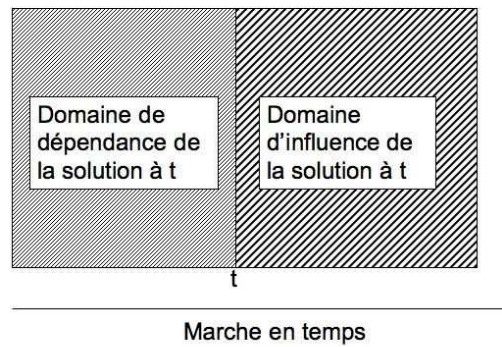


FIG. 1.2 – Domaines de dépendance et d'influence d'une équation parabolique

- Si  $\Delta > 0$ : L'équation canonique est l'équation de Laplace:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (1.26)$$

La solution dépend des conditions aux limites sur tout le domaine. On peut montrer qu'une équation elliptique où les conditions aux limites ne sont pas définies sur tout le domaine est mal posée. Le domaine de dépendance coïncide avec tout le domaine (voir figure 1.3).

Bien que la présentation ait été faite en deux dimensions, ces définitions sont généralisables avec un nombre arbitraire de dimensions. On définira notamment l'ellipticité:

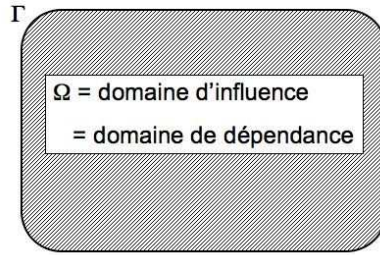


FIG. 1.3 – Domaines de dépendance et d'influence d'une équation elliptique

**Définition:**

Soit l'opérateur  $L$  défini par

$$L = \sum_i \alpha_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}$$

$L$  est dit elliptique s'il existe  $\alpha > 0$ , tel que pour tout  $x \in \mathcal{R}^d$ ,  $\sum \alpha_{ij} x_i x_j > \alpha |x|^2$

Les équations de Navier-Stokes incompressibles sont de type elliptique en espace en raison du terme de diffusion et de type parabolique en temps.

### 1.3.2 Schéma conservatif - Equations en forme conservative

**Définition d'une équation en forme conservative:** On dit qu'une équation aux dérivées partielles est sous **forme conservative** si ses coefficients sont

- ou constants
- ou tels que leurs dérivées n'apparaissent pas dans les équations

**Exemple:** L'équation de conservation de quantité de mouvement

$$\frac{\partial(u_i)}{\partial t} + u_j \frac{\partial(u_i)}{\partial x_j} = \frac{\partial p}{\partial x_j} + \nu \frac{\partial u_i}{\partial x_j x_j} \quad (1.27)$$

n'est pas en forme conservative. Elle l'est en revanche sous cette forme

$$\frac{\partial(u_i)}{\partial t} + u_j \frac{\partial(u_i)}{\partial x_j} = \frac{\partial p}{\partial x_j} + \nu \frac{\partial u_i}{\partial x_j x_j} \quad (1.28)$$

**Définition d'un schéma conservatif:** Un schéma est dit **conservatif** lorsque la discrétisation correspondante assure la conservation exacte (à l'erreur d'arrondi près) des quantités physiques, indépendamment de la taille du maillage ou de la région considérée.

La recherche de la conservation des quantités physiques dans un volume de contrôle constitue la base des méthodes de volumes finis. La forme non conservative des équations peut parfois présenter un intérêt dans certains cas (systèmes hyperboliques).

# Chapitre 2

## Discrétisation des EDP

### 2.1 Discrétisation compacte en espace

#### 2.1.1 Discrétisation par différences finies

##### Une approche naïve

La solution numérique des équations requiert la discrétisation des équations aux dérivées partielles. La question fondamentale est: comment puis-je représenter de manière discrète une fonction continue? La manière la plus simple semble être de définir un maillage sur une grille (que nous supposerons ici 1-d et régulière) et de définir la fonction  $f$  par la valeur qu'elle prend aux noeuds de la grille

$$f_i = f(x_i), \quad 0 \leq i \leq n$$

Une fois munis de cette représentation, nous devons résoudre le problème suivant: comment évaluer les dérivées de la fonction qui interviennent dans l'EDP?

La manière la plus simple consiste à supposer que la fonction est suffisamment dérivable et à calculer la série de Taylor

$$f(x + \Delta x) = f(x) + \Delta x \frac{\partial f}{\partial x}|_x + \left(\frac{(\Delta x)^2}{2} \frac{\partial^2 f}{\partial x^2}\right)|_x + \left(\frac{(\Delta x)^3}{3!} \frac{\partial^3 f}{\partial x^3}\right)|_x + \dots$$

En réarrangeant les termes,

$$\frac{\partial f}{\partial x}|_x = \frac{f(x + \Delta x) - f(x)}{\Delta x} + O(\Delta x)$$

On voit ainsi qu'une représentation de la dérivée  $\frac{\partial f}{\partial x}|_{x_i}$  peut être obtenue en posant

$$\frac{\partial f}{\partial x}|_{x_i} \sim \frac{f(x_{i+1}) - f(x_i)}{\Delta x} + O(\Delta x) \quad (2.1)$$

L'action qui à  $f_i$  associe  $\frac{f(x_{i+1}) - f(x_i)}{\Delta x}$  constitue un schéma numérique.

L'**erreur de troncature** commise en évaluant la fonction peut s'écrire comme

$$E_x = \frac{\partial f}{\partial x}|_{x_i} - \frac{f(x_{i+1}) - f(x_i)}{\Delta x} = \frac{\Delta x}{2} \frac{\partial^2 f}{\partial x^2}|_{x_i} + t.o.s \quad (2.2)$$

On dit que le schéma numérique est du premier ordre en espace en considérant l'erreur de troncature associée.

En ajustant les coefficients on peut augmenter l'ordre de précision du schéma. Par exemple, le schéma centré

$$\frac{\partial f}{\partial x}|_{x_i} = \frac{f(x_{i+1}) - f(x_{i-1}))}{2\Delta x} + O(\Delta x^2) \quad (2.3)$$

est du second ordre.

On a ainsi construit un schéma numérique.

Outre les séries de Taylor, on peut utiliser une approximation polynômiale pour déterminer les valeurs des dérivées aux points.

Exemple: Interpolation par spline cubique naturelle

On cherche à représenter la dérivée première d'une fonction par un polynôme de degré  $n$ . Une approximation courante est l'approximation par spline cubique où  $f$  est représentée par des polynômes cubiques par morceaux.

$$Pf(x) = ax^3 + bx^2 + cx + d$$

On souhaite que le polynôme vaille  $f(x_i)$  en  $x_i$ ,  $f(x_{i+1})$  en  $x_{i+1}$ . Le polynôme (de degré 1) d'interpolation linéaire est de la forme

$$Pf_1(x) = \frac{x_{j+1} - x}{x_{j+1} - x_j} f(x_j) + \frac{x - x_j}{x_{j+1} - x_j} f(x_{j+1}) = Af(x_j) + Bf(x_{j+1})$$

sur  $[x_j, x_{j+1}]$ . On souhaite que les deux premières dérivées du polynôme soient continues d'un intervalle à l'autre. On cherche le polynôme sous la forme

$$Pf(x) = Pf_1(x) + \frac{1}{6}(A - A^3)(x_{j+1} - x_j)^2 f''(x_j) + \frac{1}{6}(B^3 - B)(x_{j+1} - x_j)^2 f''(x_{j+1})$$

ce qui assure la continuité de la seconde dérivative en  $x_j$ . La détermination des dérivées secondes se fait en imposant la continuité de la première dérivative en  $x_j$  soit

$$\frac{x_j - x_{j-1}}{6} f''(x_{j-1}) + \frac{x_{j+1} - x_{j-1}}{3} f''(x_j) + \frac{x_{j+1} - x_j}{6} f''(x_{j+1}) = \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} - \frac{f(x_j) - f(x_{j-1}))}{x_j - x_{j-1}}$$

On obtient ainsi  $n-2$  équations linéaires qu'il faut compléter par des conditions arbitraires sur la dérivée seconde en  $x_1$  et en  $x_n$ . La solution "naturelle" consiste à poser  $f''(x_1) = f''(x_n) = 0$ .

## Les étapes de la discrétisation par différences finies

La discrétisation des équations aux dérivées partielles se fait en trois étapes. On considèrera pour illustrer le propos l'équation de la chaleur en une dimension

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (2.4)$$

sur  $[0,1]$  avec la condition aux limites

$$u(0) = u(1) = 1.0 \quad (2.5)$$

et la condition initiale

$$u(x, t = 0) = x$$

si  $0 \leq x \leq 0.5$

$$u(x, t = 0) = 1 - x$$

si  $0.5 \leq x \leq 1$

Les étapes de la discrétisation sont:

1. définition de l'espace discrétisé - maillage - caractérisé par des noeuds où les valeurs des fonctions sont définis

exemple: en 1-D  $x_I$  avec  $x_{I+1} = I\Delta x$ .

2. construction d'une approximation pour les dérivées partielles de la fonction en fonction des valeurs de la fonction aux noeuds du maillage

exemple:

$$\frac{\partial^2 f}{\partial x^2} \Big|_{x_I} = \frac{f_{I+1} + f_{I-1} - 2f_I}{\Delta x^2}$$

$$\frac{\partial f}{\partial t} \Big|_{x_I}^{t=t^n} = \frac{f^{t_{n+1}} - f^{t_n}}{\Delta t}$$

3. substitution de l'approximation dans l'EDP et obtention d'un système aux différences finies.

exemple:

$$\frac{\partial f}{\partial x} \Big|_{x_I} = \frac{f_{I+1} + f_{I-1} - 2f_I}{\Delta x^2}$$

### 2.1.2 Approche par éléments finis

La discrétisation des EDP par éléments finis reprend les étapes précédentes avec un point de vue différent.

1. La discrétisation de l'espace consiste en un découpage en éléments comprenant un ou plusieurs noeuds. Les noeuds représentent les degrés de liberté des fonctions de représentation. Pour chaque élément on dispose autant de fonctions de représentation que de noeuds. En deux dimensions, les noeuds répartis sur le domaine forment le sommet de triangles et

de tétraèdres en 3 dimensions. Les fonctions associées aux noeuds seront définies sur ces triangles (ou tétraèdres).

2. On définit ici le concept d'interpolation où chaque noeud doit repose sur 2 étapes. La première étape consiste à interpoler des fonctions.

$$u(x) = \sum_I u_I N_I(x) \quad (2.6)$$

Les fonctions d'interpolation sont telles que les points du maillage sont interpolés exactement

$$\sum_I N_I(x_J) = \delta_{IJ} \quad (2.7)$$

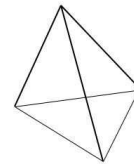
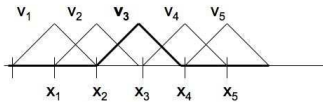
En outre, on demande à ce que les constantes soient interpolées exactement

$$\sum_I N_I = 1 \quad (2.8)$$

Exemple: éléments finis d'ordre 1

- sur  $[x, x + h]$ ,  $V(x) = \frac{1}{2}x$
- sur  $[x - h, x]$ ,  $V(x) = \frac{1}{2}(h - x)$

Les triangles deviennent des tétraèdres en dimension 2.



a)

b)

FIG. 2.1 – Représentation par fonctions continues par morceaux a) 1-D b) 2-D

On peut également définir des polynômes d'ordre supérieur.

3. Le calcul des dérivées est immédiat en utilisant la formule d'interpolation.
4. La troisième étape consiste à utiliser une formulation variationnelle (dite aussi formulation faible) des EDP. La troisième étape est une méthode de résidus pondérés où le résidu de l'équation est projeté sur une base de fonctions.

On cherche la solution sous la forme suivante

$$u(x) = \sum_{n=1}^N a^n \phi^n(x) \quad (2.9)$$



Ces fonctions  $\phi^n(x)$  représentent une base de  $L^2(x)$  et sont associées à un produit scalaire  $(\cdot, \cdot)$ . On suppose que l'équation d'évolution peut s'écrire comme

$$f(u) = 0 \quad (2.10)$$

La résolution de l'EDP consiste à déterminer les coefficients  $a^n$  par une méthode de Galerkin (cas particulier résidus pondérés)

$$(f(\sum_{j=1}^N a^j \phi^j), \phi^n) = 0 \quad (2.11)$$

ce qui conduit à un système linéaire de résolution pour les coefficients  $a^j$ .

On considère un exemple l'équation de la chaleur

$$\frac{\partial U}{\partial t} = \alpha \frac{\partial^2 U}{\partial x^2} \quad (2.12)$$

avec les conditions aux limites homogènes sur la frontière du domaine:

$$U = 0$$

sur  $\partial\Omega$ .

La recherche de la solution sur une base de fonctions, et non plus sur un ensemble de points,

On utilisera une méthode de Galerkin, où les fonctions-tests utilisées pour la projection du résidu coïncident avec les fonctions de base. Les fonctions  $v$  vérifient aussi les conditions aux limites homogènes aux frontières.

$$\int_{\Omega} \frac{\partial U}{\partial t} v d\Omega = \alpha \int_{\Omega} \frac{\partial^2 U}{\partial x^2} v d\Omega \quad (2.13)$$

En intégrant par parties, il vient que

$$\int_{\Omega} \frac{\partial U}{\partial t} v d\Omega = - \int_{\Omega} \alpha \frac{\partial U}{\partial x} \frac{\partial v}{\partial x} d\Omega + \oint \alpha \frac{\partial U}{\partial x} v d\Omega \quad (2.14)$$

En utilisant  $v = \phi_J$ , on obtient alors un système de la forme suivante

$$\sum \frac{dU_I}{dt} \int \phi_I \phi_J d\Omega - \alpha \sum U_I dt \int \frac{\partial \phi_I}{\partial x} \frac{\partial \phi_J}{\partial x} d\Omega = 0 \quad (2.15)$$

La matrice  $K$

$$K = \int \frac{\partial \phi_I}{\partial x} \frac{\partial \phi_J}{\partial x} d\Omega$$

est appelée matrice de rigidité ('stiffness').

La matrice  $M$

$$M = \int \phi_I \phi_J d\Omega$$

est appelée matrice de masse.

Le problème elliptique est ainsi ramené à la résolution d'un problème linéaire de la forme

$$M \frac{da}{dt} = Ka$$

Lorsque les fonctions de base sont d'ordre élevé, on est amené à utiliser des expressions de quadrature pour calculer ces matrices.

### 2.1.3 Approche par volumes finis

Si on définit la fonction d'interpolation suivante

$$V_I(x) = 1 \text{ pour } x \in [x_{i-1/2}, x_{i+1/2}[$$

$$V_I(x) = 0 \text{ sinon}$$

on s'aperçoit que la représentation aux volumes finis dans le cas le plus simple peut être considérée comme un cas particulier de représentation aux éléments finis pour cette fonction d'interpolation. L'idée fondamentale des volumes finis est de définir un volume de contrôle et d'imposer la conservation des équations sur ce volume de contrôle.

$$\int \int \int \frac{\partial u_i}{\partial t} d\mathbf{x} + \int \int \int \frac{\partial u_i u_i}{\partial x_j} d\mathbf{x} = - \int \int \int \frac{\partial p}{\partial x_j} + \nu \frac{\partial^2 u}{\partial x_j \partial x_j} d\mathbf{x} \quad (2.16)$$

En utilisant le théorème de la divergence

$$\int_{\Omega} \text{div} F d\Omega = \int_{\partial\Omega} F \cdot dS \quad (2.17)$$

on voit que les termes sur le côté droit de l'équation se résument à l'expression de flux de quantités. On considère le volume suivant (pour plus de simplicité, on se limitera à deux dimensions) représenté en figure ??:

$$\frac{\partial}{\partial t} \int_{\Omega} U d\Omega + \int_S F \cdot dS = \int Q d\Omega \quad (2.18)$$

Le domaine est découpé en volumes de contrôle  $\Omega_J$  qui recouvrent le domaine. Ces cellules ne sont pas nécessairement disjointes, il peut y avoir recouvrement à condition que ce recouvrement n'induisse pas la création de nouvelles frontières.

On évalue cette équation à l'intérieur d'un volume  $\Omega_J$

$$\frac{\partial}{\partial t} U_J \Omega_J + \sum_{\text{cotes}} F \cdot S = Q_J \Omega_J \quad (2.19)$$

Pour évaluer les quantités volumiques, le plus simple est d'utiliser une valeur constante au centre de la cellule. On peut également employer la méthode des trapèzes,

$$\int_{\Omega} f d\Omega \sim \frac{\Delta x}{2} (f_i + f_{i+1})$$

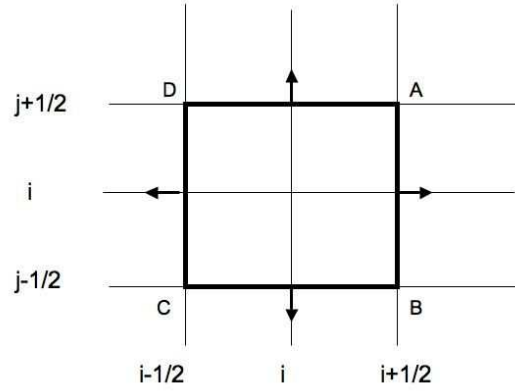


FIG. 2.2 – Description d'un volume de contrôle

ou la méthode de Simpson, plus précise,

$$\int_{\Omega} f d\Omega \sim \frac{\Delta x}{6} (f_{i-1/2} + f_{i+1/2} + 4f_i)$$

En supposant que le volume  $\Omega_J$  est représenté par la figure et que le maillage est régulier avec un pas de maillage  $\Delta x$  et  $\Delta y$ , l'équation peut-être simplifiée comme

$$\frac{\partial U_{i,j}}{\partial t} \Delta x \Delta y + (f_{i+1/2,j} - f_{i-1/2,j}) \Delta y + (f_{i,j+1/2} - f_{i,j-1/2}) \Delta x = Q_{i,j} \Delta x \Delta y \quad (2.20)$$

La question est de savoir comment évaluer  $f$  aux frontières des cellules. L'évaluation des flux est plus délicate. La manière la plus naturelle d'évaluer un flux consiste à utiliser un schéma centré. On a alors

$$f_{i+1/2,j} = \frac{f_{i+1} - f_i}{\Delta x}$$

mais cette approche contribue à créer des instabilités dites "en damier" entre  $i + 1$  et  $i - 1$ .

Les termes de flux consistent en l'advection d'une quantité par le champ de vitesse. Les termes à évaluer sont de la forme

$$I_c = v \frac{\partial u}{\partial x}$$

Il apparaît alors judicieux d'utiliser une expression pour le flux qui prenne en compte le fait que l'information est transportée dans une direction privilégiée de l'écoulement. On définit alors une évaluation "upwind" dans laquelle les points d'interpolation dépendent de la valeur de la vitesse.

En laissant tomber le second indice  $j$  considéré comme constant, on obtient l'interpolation linéaire upwind (LUDS)

Si  $v > 0$

$$f_{i+1/2} = \frac{3u_i - u_{i-1}}{2}$$

$$f_{i-1/2} = \frac{3u_{i-1} - u_{i-2}}{2}$$

Si  $v < 0$

$$f_{i+1/2} = \frac{3u_{i+1} - u_{i+2}}{2}$$

$$f_{i-1/2} = \frac{3u_i - u_{i+1}}{2}$$

On peut également définir une interpolation quadratique, qui est du troisième ordre en espace (mais l'expression finale du flux sera seulement du second ordre en raison de la règle d'intégration utilisée).

Si  $v > 0$

$$f_{i+1/2} = \frac{3u_{i+1} + 6u_i - u_{i-1}}{8}$$

$$f_{i-1/2} = \frac{3u_i + 6u_{i-1} - u_{i-2}}{8}$$

$$I_C = v \frac{3u_{i+1} + 3u_i - 7u_{i-1} + u_{i-2}}{8\Delta x}$$

Si  $v < 0$

$$f_{i+1/2} = \frac{3u_i + 6u_{i+1} - u_{i+2}}{8}$$

$$f_{i-1/2} = \frac{3u_{i-1} + 6u_i - u_{i+1}}{8}$$

$$I_C = -v \frac{3u_{i-1} + 3u_i - 7u_{i+1} + u_{i+2}}{8\Delta x}$$

Cette interpolation est connue sous le nom de schéma QUICK.

Remarque: On peut voir la représentation aux différences finies comme un cas extrême de représentation par éléments finis constitués d'impulsions 'Dirac' - on ne dispose d'aucune règle de dérivation pour évaluer les dérivées partielles.

## 2.2 Discrétisation temporelle

Les équations de Navier-Stokes sont paraboliques en temps. On adopte donc pour la résolution une marche en temps où à chaque pas de temps l'écoulement est résolu sur tout le domaine. Pour une équation à résoudre de type

$$\frac{\partial u}{\partial t} = F$$

On distingue deux grands types de schémas numériques temporels:

- les schémas explicites, où l'avancement en temps se fait directement

$$\frac{u^{n+1} - u^n}{\Delta t} = F^n$$

- les schémas implicites, où la solution au temps  $n + 1$  est obtenue par la résolution d'un problème inverse

$$\frac{u^{n+1} - u^n}{\Delta t} = F^{n+1}$$

Le seul cas qui est traité dans la résolution des équations consiste à considérer  $F$  linéaire. L'intérêt de rendre certains termes (linéaires) de l'EDP implicites va apparaître dans le chapitre suivant.

Plus précisément, on distingue

### 2.2.1 Les schémas d'Euler

On considère l'équation

$$\frac{\partial u}{\partial t} = F(u) \quad (2.21)$$

- schéma Euler d'ordre 1 explicite

$$\frac{u_j^{n+1} - u_j^n}{\delta t} = F_j^n \quad (2.22)$$

- schéma Euler d'ordre 1 implicite

$$\frac{u_j^{n+1} - u_j^n}{\delta t} = F_j^{n+1} \quad (2.23)$$

Cette formulation est utilisable seulement dans le cas où  $F$  est linéaire.

- schéma Euler d'ordre 2 retardé explicite

$$\frac{3u_j^{n+1} - u_j^n + 2u_j^{n-1}}{2\delta t} = F_j^n \quad (2.24)$$

### 2.2.2 Les schémas de Runge-Kutta

Les schémas de Runge-Kutta sont des schémas d'intégration numérique extrêmement performants. Nous en donnons ici le principe en nous limitant à un pas de temps constant. La méthode d'Euler consiste à évaluer la dérivée à l'instant  $t_n$ ,  $f(t_n)$  et à utiliser cette estimation pour évaluer le champ au temps

$$u_{n+1} = u_n + \delta t f(t_n)$$

La méthode de Runge-Kutta consiste à introduire des points intermédiaires afin d'améliorer la précision de l'approximation.

#### Runge-Kutta à l'ordre 2

On définit un point intermédiaire

$$y_1 = f(t, u_n) + \Delta t$$

L'évaluation de la dérivée est alors faire au point  $t + \Delta t/2, y_1/2$ .

$$y_2 = f(t + \delta t/2, u_n + y_1/2) + \Delta t$$

La nouvelle estimée au temps  $t^{n+1}$  s'obtient alors par

$$u_{n+1} = u_n + y_2 + O(\Delta t^3)$$

La procédure est résumée dans la figure .

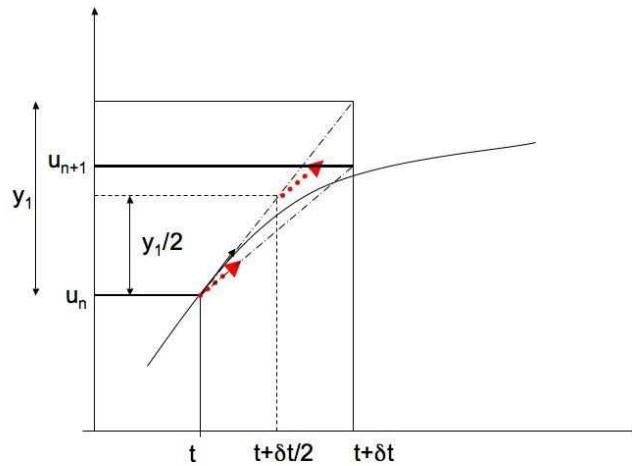


FIG. 2.3 – Méthode de Runge-Kutta d'ordre 2

### Runge-Kutta d'ordre 4

On introduit 3 valeurs intermédiaires. L'estimation est alors précise à l'ordre 4. La méthode de Runge-Kutta d'ordre 4 est très populaire pour de nombreux problèmes (sans raideur excessive).

$$\begin{aligned} y_1 &= \delta t f(t_n, u_n) \\ y_2 &= \delta t f(x_n + \delta t/2, u_n + k_1/2) \\ y_3 &= \delta t f(x_n + \delta t/2, u_n + k_2/2) \\ y_4 &= \delta t f(x_n + \delta t, u_n + k_3) \\ u_{n+1} &= u_n + \frac{y_1}{6} + \frac{y_2}{3} + \frac{y_3}{3} + \frac{y_4}{6} + O(\delta t^5) \end{aligned}$$

### 2.2.3 Discrétisation du terme source

– schéma de Crank-Nicolson (implicite)

$$\frac{u_j^{n+1} - u_j^n}{\delta t} = \frac{1}{2}(F_j^n + F_j^{n+1}) \quad (2.25)$$

– schéma d'Adams-Bashforth (explicite)

$$\frac{u_j^{n+1} - u_j^n}{\delta t} = \frac{3}{2}F_j^n - \frac{1}{2}F_j^{n-1} \quad (2.26)$$

Une discrétisation de l'équation d'advection-diffusion pourra être donc représentée comme

$$\frac{3u_j^{n+1} - u_j^n + 2u_j^{n-1}}{2\delta t} = \frac{1}{2}(\Delta u_j^n + \Delta u_j^{n+1}) - \frac{3}{2}C\nabla u_j^n + \frac{1}{2}C\nabla u_j^{n-1} \quad (2.27)$$

Dans les équations de Navier-Stokes, les termes visqueux sont représentés de façon implicite, les termes non linéaires sont explicites, de sorte qu'on obtient

$$(\Delta - \frac{3}{\Delta t})u^{n+1} = \frac{-u_j^n + 2u_j^{n-1}}{\delta t} - \Delta u_j^n + 3C\nabla u_j^n - C\nabla u_j^{n-1} \quad (2.28)$$

soit un problème de Helmholtz de la forme

$$(\Delta - \lambda)u^{n+1} = S^n \quad (2.29)$$

où  $S^n$  est un terme source qui contient les termes connus au temps inférieurs à  $t^n$ .

## 2.3 Discrétisation non compacte en espace

### 2.3.1 La recherche d'un espace où définir la solution

Dans les approches par différences finies et volumes finis que nous venons de voir, la solution numérique n'est définie qu'en certains points, et jusqu'à un certain ordre. Pour pouvoir comparer la solution numérique à la solution exacte, il faudrait que celles-ci soient définies sur le même domaine spatial. Ceci nous conduit à une nouvelle façon de définir la solution qui va s'écrire comme la combinaison d'une base de fonctions définies sur tout l'espace. Cette idée est le fondement des méthodes d'éléments finis et des méthodes spectrales. Les représentations par éléments finis privilégient des fonctions de base à support compact. Elles permettent un traitement local des discontinuités et peuvent être adaptées à des géométries complexes. Les méthodes spectrales s'appuient sur des fonctions à support global, qui requièrent des géométries simples. Si la solution est suffisamment régulière, la solution numérique converge rapidement vers la solution exacte (convergence "spectrale", voir plus loin).

Soit à résoudre l'équation différentielle ou aux dérivées partielles

$$\mathcal{L}f = s \text{ dans } \Omega \quad (2.30)$$

$$f = \bar{f} \text{ sur } \partial\Omega \quad (2.31)$$

où la solution est recherchée dans un espace de fonctions  $\mathcal{H}$ .

Le principe des méthodes spectrales est de rechercher cette solution sous la forme d'un développement en série de fonctions.

$$f = \sum_{n=-\infty}^{\infty} \hat{f}_n \phi_n(x) \quad (2.32)$$

- $\phi_n$  sont les fonctions de base (famille dense dans  $\mathcal{H}$ ) qui sont  $\mathcal{C}^\infty$  et en général orthogonales au sens d'un produit scalaire  $(\cdot, \cdot)$ .
- $\hat{f}_n$  sont les coefficients spectraux

En pratique, on approche  $f$  par  $f_N$

$$f_N = P_N f = \sum_{n=0}^N \hat{f}_n \phi_n(x) \quad (2.33)$$

En pratique, on utilisera soit des développements à base

- de séries de Fourier
- de polynômes orthogonaux Chebyshev ou Legendre

L'intérêt des méthodes spectrales est le suivant: si la solution  $f$  est  $\mathcal{C}^\infty$  alors les  $\hat{f}_n$  décroissent plus vite que toute puissance de  $n$ . On dit qu'on a une convergence spectrale.

Cette caractéristique rend l'utilisation des méthodes spectrales particulièrement intéressante dans le domaine des études de stabilité des écoulements et pour la simulation directe de la turbulence.

### 2.3.2 Séries de Fourier

**Généralités** On suppose que la solution  $f$  est dans  $L^2(0, 2\pi)$ . Alors on sait que sa série de Fourier converge dans  $L^2(0, 2\pi)$  et on peut donc écrire

$$f = \sum_{k=-\infty}^{\infty} \hat{f}_k \exp(ikx) \quad (2.34)$$

et on note  $f_N$  la somme tronquée l'ordre  $N/2$ , c'est-à-dire

$$f_N = \sum_{k=-N/2}^{N/2} \hat{f}_k \exp(ikx) \quad (2.35)$$

Les  $\{\exp(ikx), k = -\infty, \dots, \infty\}$  forment une famille orthogonale dans  $L_2(0, 2\pi)$  relativement au produit scalaire  $(u, v) = \int_0^{2\pi} u \bar{v} dx$  et on a  $(\exp(ilx), \exp(imx)) = 2\pi \delta_{lm}$

En prenant le produit scalaire de (2.35) avec  $\exp(ilx)$ , il vient donc:

$$(f_N, \exp(ilx)) = 2\pi \hat{f}_l \quad (2.36)$$

et donc

$$\hat{f}_l = \frac{1}{2\pi} \int_0^{2\pi} f_N \exp(-ilx) dx \quad (2.37)$$

pour  $l = -\infty, \dots, \infty$  et donc également pour  $-N/2 \leq l \leq N/2$ . Ceci montre que  $f_N$  est la projection  $L_2$  de  $f$  sur  $\{\exp(ikx), k = -N/2, \dots, N/2\}$ .

Mis à part quelques rares cas particuliers il faut évaluer cette intégrale numériquement.



### Quadratures discrètes

Une première façon de faire consiste à l'évaluer par une formule des trapèzes. Si on considère  $N + 2$  points  $x_j$  équidistants de  $\Delta x = \frac{2\pi}{N+1}$ ,  $\{x_j = \frac{2\pi j}{N+1}, j = 0, 1, \dots, N+1\}$ , il vient

$$\tilde{f}_l = \frac{1}{2\pi} \frac{2\pi}{N+1} \sum_{j=0}^{N+1} \frac{1}{\bar{c}_j} f_N(x_j) \exp(-ilx_j) dx \quad (2.38)$$

avec  $\bar{c}_0 = \bar{c}_{N+1} = 2, c_j = 1 \leq j \leq N$ .

En tenant compte de la périodicité, on obtient donc:

$$\tilde{f}_l = \frac{1}{N+1} \sum_{j=0}^N f_N(x_j) \exp(-ilx_j) \quad (2.39)$$

Cette formule d'intégration est de précision maximale dans l'ensemble des fonctions  $\mathcal{C}^\infty$   $2\pi$  périodiques

Une deuxième façon de procéder est la suivante: plaçons nous dans l'espace vectoriel  $\mathcal{F}_N$  engendré par les  $\{\exp(ikx), k = -N/2, \dots, N/2\}$ . Une fonction périodique étant connue par ses valeurs  $f_j = f(x_j)$  aux points  $\{x_j = \frac{2\pi j}{N+1}, j = 0, \dots, N+1\}$ , définissons  $I_N f$  le polynôme d'interpolation trigonométrique dans  $\mathcal{F}_N$  qui vaut  $f_j$  aux points  $x_j$ . On a donc

$$I_N f = \sum_{k=-N}^N \tilde{f}_k \exp(ikx) \quad (2.40)$$

et aux points  $x_j$

$$f_j = I_N f(x_j) = \sum_{k=-N}^N \tilde{f}_k \exp(ikx_j) \quad (2.41)$$

Reste à inverser cette équation pour obtenir les  $\tilde{f}_k$ .

Considérons la forme bilinéaire dans  $\mathcal{F}_N$   $(u, v)_d = \sum_{j=0}^N u(x_j) \overline{v(x_j)}$ . Cette forme bilinéaire définit un produit scalaire discret dans  $\mathcal{F}_N$ . Elle est en effet bilinéaire, possède la symétrie hermitienne. De plus si  $(u, u)_d = 0$ , alors  $u(x_j) = 0, \forall j = 0, \dots, N$  et on a alors  $u = 0$ .

Les  $\{\exp(ikx), k = -N/2, \dots, N/2\}$  continuent de former une famille orthogonale dans  $\mathcal{F}_N$  pour le produit scalaire discret  $(\cdot)_d$ . En effet

$$(\exp(ilx), \exp(imx))_d = \sum_{j=0}^N \exp(ilx_j) \exp(-imx_j) = \sum_{j=0}^N \exp(i(l-m)x_j) \quad (2.42)$$

Ceci est une progression géométrique de raison  $\rho = \exp(i(l-m)\frac{2\pi}{N+1})$  qui vaut donc

$$\begin{cases} = \frac{1-\rho^{N+1}}{1-\rho} = 0 \text{ si } \rho \neq 1 \\ = N+1 \text{ si } \rho = 1 \text{ c'est à dire si } l = m \pmod{N+1} \end{cases}$$

On a donc  $(\exp(ilx), \exp(imx))_d = (N+1)\delta_{lm}$ .

Si on forme le produit scalaire discret de (2.40) avec  $\exp(ilx)$ , il vient

$$(I_N f, \exp(ilx))_d = \sum_{k=-N/2}^{N/2} \tilde{f}_k(\exp(ikx), \exp(ilx))_d \quad (2.43)$$

$$= (N+1)\tilde{f}_l \quad (2.44)$$

et donc, en tenant compte du fait que  $f_j = I_N f(x_j)$ :

$$\tilde{f}_l = \frac{1}{N+1} \sum_{j=0}^N f_N(x_j) \exp(-ilx_j) \quad (2.45)$$

Cette formule est identique à celle obtenue par l'intégration par la formule des trapèzes.

### Relation entre les $\hat{f}_l$ et les $\tilde{f}_l$

On a

$$\tilde{f}_l = \frac{1}{N+1} \sum_{j=0}^N f_j \exp(-ilx_j) \quad (2.46)$$

$$= \frac{1}{N+1} \sum_{j=0}^N \sum_{k=-\infty}^{\infty} \hat{f}_k \exp(ikx_j) \exp(-ilx_j) \quad (2.47)$$

$$= \hat{f}_l + \frac{1}{N+1} \sum_{k=-\infty}^{\infty} \hat{f}_k(\exp(ikx), \exp(ilx))_d \quad (2.48)$$

$$= \hat{f}_l + \sum_{k \in \mathbb{Z}^*} \hat{f}_{l+k(N+1)} \quad (2.49)$$

Ceci montre que les  $\tilde{f}_l$  ne sont de bonnes approximations des  $\hat{f}_l$  que si les coefficients spectraux décroissent suffisamment rapidement.

### Vitesse de convergence

Reprenons l'expression de

$$\hat{f}_l = \frac{1}{2\pi} \int_0^{2\pi} f_N \exp(-ilx) dx \quad (2.50)$$

Si  $f_N$  est suffisamment régulière (de classe  $\mathcal{C}^1$  au moins) on peut intégrer par parties pour obtenir

$$2\pi \hat{f}_l = -\frac{1}{il} f_N \exp(-ilx) \Big|_0^{2\pi} + \frac{1}{il} \int_0^{2\pi} f'_N \exp(-ilx) dx \quad (2.51)$$

Si,  $f_N$  est plus régulière (de classe  $\mathcal{C}^2$ ), on peut recommencer l'opération pour arriver à

$$\begin{aligned} 2\pi \hat{f}_l &= -\frac{1}{il} f_N \exp(-ilx) \Big|_0^{2\pi} - \frac{1}{(il)^2} f'_N \exp(-ilx) \Big|_0^{2\pi} \\ &+ \frac{1}{(il)^2} \int_0^{2\pi} f''_N \exp(-ilx) dx \end{aligned} \quad (2.52)$$



avec  $FPS_{kj} = \frac{1}{N+1} \exp(-i\frac{2\pi kj}{N+1})$ ,  $-K/2 \leq k \leq K/2, 0 \leq j \leq N$ .

Il vient ensuite

$$\begin{bmatrix} \cdot \\ \cdot \\ ik\tilde{f}_k \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} -i\frac{N}{2} & & & & \\ & \cdot & & & \\ & & ik & & \\ & & & \cdot & \\ & & & & i\frac{N}{2} \end{bmatrix} \begin{bmatrix} & & & & \\ & & & & \\ & & FPS_{kj} & & \\ & & & & \\ & & & & \end{bmatrix} \begin{bmatrix} f_0 \\ \cdot \\ f_j \\ \cdot \\ f_N \end{bmatrix}$$

et enfin

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0) \\ \cdot \\ \frac{\partial f}{\partial x}(x_l) \\ \cdot \\ \frac{\partial f}{\partial x}(x_N) \end{bmatrix} = \begin{bmatrix} & & & & \\ & & & & \\ & FSP_{lk} & & & \\ & & & & \\ & & & & \end{bmatrix} \begin{bmatrix} -i\frac{N}{2} & & & & \\ & \cdot & & & \\ & & ik & & \\ & & & \cdot & \\ & & & & i\frac{N}{2} \end{bmatrix} \begin{bmatrix} & & & & \\ & & & & \\ & & FPS_{kj} & & \\ & & & & \\ & & & & \end{bmatrix} \begin{bmatrix} f_0 \\ \cdot \\ f_j \\ \cdot \\ f_N \end{bmatrix}$$

avec  $FSP_{lk} = \exp(i\frac{2\pi lk}{N+1})$ ,  $-K/2 \leq k \leq K/2, 0 \leq l \leq N$ .

En effectuant le produit des 3 matrices, on obtient

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0) \\ \cdot \\ \frac{\partial f}{\partial x}(x_l) \\ \cdot \\ \frac{\partial f}{\partial x}(x_N) \end{bmatrix} = \begin{bmatrix} & & & & \\ & & & & \\ & \mathcal{D}_{lj} & & & \\ & & & & \\ & & & & \end{bmatrix} \begin{bmatrix} f_0 \\ \cdot \\ f_j \\ \cdot \\ f_N \end{bmatrix}$$

avec

$$\mathcal{D}_{mn} \begin{cases} = \frac{1}{2}(-1)^{m+n} \frac{1}{\sin(\frac{(m-n)\pi}{N+1})} & \text{si } m \neq n \\ = 0 & \text{si } m = n \end{cases}$$

Selon que  $K$  est grand ou petit, on aura intérêt à effectuer en pratique cette évaluation, soit par cette dernière multiplication matricielle si  $N$  est petit, soit en décomposant selon les 3 étapes élémentaires ce qui permet l'utilisation des Transformées Rapides de Fourier. Sur des machines vectorielles comme le Cray, le point de croisement se situe pour  $N$  de l'ordre de 32.

Remarques:

- On a bien évidemment  $FPS \times FSP = FSP \times FPS = \mathcal{I}$ , qui n'est autre que le fait que les  $\exp(ikx)$  sont une famille orthogonale pour le produit scalaire discret  $(\cdot)_d$ .
- la matrice  $\mathcal{D}$  est antisymétrique, semblable à une matrice diagonale dont les valeurs propres sont imaginaires pures  $ik$ ,  $-K/2 \leq k \leq K/2$
- La matrice représentative de la dérivation seconde est  $\mathcal{D}^2$  dont les valeurs propres sont  $0, -k^2, 1 \leq k \leq \frac{K}{2}$ , les valeurs propres non nulles étant de multiplicité algébrique 2.

### 2.3.3 Polynômes orthogonaux

#### Généralités

Pour approcher des solutions possédant une grande régularité mais ne présentant pas les propriétés de périodicité suffisante, on est conduit à utiliser comme fonctions de base des polynômes orthogonaux, Chebyshev ou Legendre. Ces deux familles de polynômes orthogonaux sont définies sur  $[-1,1]$  et sont orthogonales relativement au produit scalaire

$$(f, g)_\omega = \int_{-1}^1 f(x)g(x)\omega(x)dx \quad (2.53)$$

où  $\omega$  est une fonction poids, positive sur  $] -1, 1[$ . La fonction poids  $\omega$  vaut respectivement

$$\omega = \begin{cases} = 1 & \text{pour les polynômes de Legendre } L_n \\ = (1 - x^2)^{-1/2} & \text{pour les polynômes de Chebyshev } T_n \end{cases} \quad (2.54)$$

On a respectivement:

$$(L_n, L_m)_\omega = (n + \frac{1}{2})^{-1} \delta_{mn} \quad (2.55)$$

$$(T_n, T_m)_\omega = c_n \frac{\pi}{2} \delta_{mn} \quad (2.56)$$

avec  $c_0 = 2, c_p = 1, p \geq 1$ .

Les polynômes de Chebyshev vérifient:  $T_n(\cos \theta) = \cos(n\theta)$

De façon générale, les polynômes orthogonaux vérifient une relation de récurrence à 3 termes.

Pour les polynômes de Chebyshev et Legendre, celles-ci s'écrivent:

$$(n + 1)L_{n+1} = (2n + 1)xL_n - nL_{n-1} \quad (2.57)$$

$$T_{n+1} = 2xT_n - T_{n-1} \quad (2.58)$$

ce qui permet de les calculer à partir de  $L_0 = T_0 = 1$  et  $L_1 = T_1 = x$ .

Considérons l'ensemble  $L^2([-1,1], \omega)$  des fonctions de carré intégrable sur  $[-1,1]$  relativement au produit scalaire  $(\cdot)_\omega$ .

Les  $\{T_n, n = 0, \dots, \infty\}$  forment une famille complète dans  $L^2([-1,1], \omega)$  et on peut donc développer toute fonction de  $L^2([-1,1], \omega)$  sous la forme  $f = \sum_{n=0}^{\infty} \hat{f}_n T_n$ .

De façon analogue au cas des séries de Fourier, on définit  $P_N f = \sum_{n=0}^N \hat{f}_n T_n$ . C'est la meilleure approximation  $L_\omega^2$  de  $f$  dans  $\mathcal{P}_N$ , l'ensemble des polynômes de degré  $\leq N$ .

On a également

$$\hat{f}_n = \frac{2}{c_n \pi} \int_{-1}^1 f_N T_n \omega dx \quad (2.59)$$

pour  $0 \leq n \leq N$ , où  $\hat{f}_n$  est le spectre de  $f$ .

L'intérêt de cette approximation est résumé dans ce résultat de convergence (Canuto et Quarteroni, 1982):

$$\|f - P_N f\|_{L_\omega^2} \leq N^{-\sigma} \|f\|_{H_\omega^\sigma} \quad (2.60)$$

qui montre que la vitesse de convergence ne dépend que de la régularité de la fonction que l'on cherche à approcher. En particulier, si la fonction est analytique alors l'erreur décroît plus vite que toute puissance de  $N$  et on retrouve la convergence exponentielle, indépendamment cette fois des conditions de périodicité aux bornes de l'intervalle.

### Quadratures discrètes

L'évaluation de (2.59) pose à nouveau des problèmes de quadrature numérique, auxquels les formules de quadrature de Gauss permettent d'apporter une réponse.

Soit à évaluer  $I_\omega(f) = \int_{-1}^1 f(x)\omega(x)dx$  où  $\omega$  est une fonction poids, positive sur  $] -1,1[$ . On va chercher à approcher  $I_\omega$  par une quadrature discrète à  $(N+1)$  points du type  $\sum_{i=0}^N \alpha_i f(x_i)$  où les  $x_i$  sont des points de collocation dans  $[-1,1]$  et les  $\alpha_i$  sont des coefficients.

Les  $\alpha_i$  et  $x_i$  sont *a priori* indéterminés et on peut chercher à les optimiser de manière à ce que  $I_\omega(f) = \sum_{i=0}^N \alpha_i f(x_i)$  pour la plus large classe de polynômes  $f$  possibles. On dispose de  $2N+2$  degrés de liberté, on peut donc espérer trouver des  $\alpha_i$  et  $x_i$  tels que  $I_\omega(f) = \sum_{i=0}^N \alpha_i f(x_i)$  pour tout polynôme de degré  $\leq 2N+1$ .

La détermination des  $\alpha_i$  et  $x_i$  est la suivante:

- supposons les  $x_i$  donnés. On peut alors déterminer les  $\alpha_i$  tels que  $I_\omega(x^k) = \sum_{i=0}^N \alpha_i x_i^k$  pour  $0 \leq k \leq N$ , ce qui fournit un système linéaire pour les  $\alpha_i$  dont l'inversibilité est assurée si les  $x_i$  sont 2 à 2 distincts.
- les  $\alpha_i$  étant maintenant connus, si les  $x_i$  sont les racines du  $(N+1)$ ème polynôme  $P_{N+1}$  de la famille de polynômes orthogonaux relativement au produit scalaire  $(\cdot)_\omega$ , alors  $I_\omega(f) = \sum_{i=0}^N \alpha_i f(x_i)$  pour tout  $f$  dans  $\mathcal{P}_{2N+1}$ .

Preuve: Soit  $f$  dans  $\mathcal{P}_{2N+1}$ . On a alors  $f = rP_{N+1} + s$  où le polynôme quotient  $r$  est de degré  $\leq N$  et le polynôme reste  $s$  est de degré  $\leq N$ . On a alors

$$\begin{aligned}
 I_\omega(f) &= I_\omega(rP_{N+1} + s) \\
 &= I_\omega(rP_{N+1}) + I_\omega(s) \\
 &= I_\omega(s) \text{ par orthogonalité de } P_{N+1} \text{ avec } \mathcal{P}_N \\
 &= \sum_{i=0}^N \alpha_i s(x_i) \text{ par construction des } \alpha_i \\
 &= \sum_{i=0}^N \alpha_i f(x_i) \text{ par définition des } x_i
 \end{aligned}$$

On obtient donc les formules dites de quadrature de Gauss pour chaque famille de polynômes orthogonaux.

Pour les polynômes de Chebyshev, les racines de  $T_{N+1}$  sont données par  $x_k = \cos \theta_k$  avec  $(N+1)\theta_k = \frac{\pi}{2} + k\pi$  et donc  $x_k = \cos \frac{2k+1}{N+1} \frac{\pi}{2}$ . Ces racines sont donc les projections de points sur l'axe des cosinus de points équidistribués sur le 1/2 cercle trigonométrique. Au voisinage des

extrémités 1 et  $-1$  leur écartement est en  $\frac{1}{N^2}$  et au voisinage du centre il est en  $\frac{1}{N}$ . Les poids  $\alpha_i$  sont alors tous égaux à  $\frac{\pi}{N+1}$ .

On constate que les extrémités 1 et  $-1$  n'appartiennent pas à cet ensemble de points de collocation ce qui présente un inconvénient si on veut utiliser cette technique pour approcher des solutions d'équations elliptiques où des conditions aux limites doivent être imposées aux bords. Si on impose *a priori*  $x_0 = 1$  et  $x_N = -1$ , alors l'optimisation ne porte plus que sur  $2N$  degrés de liberté et on ne peut donc espérer que  $I_\omega(f) = \sum_{i=0}^N \alpha_i f(x_i)$  pour tout polynôme de degré  $\leq 2N - 1$ . La démonstration procède de la même façon que précédemment en considérant le polynôme  $Q_{N+1} = P_{N+1} + \lambda P_N + \mu P_{N-1}$  où  $\lambda$  et  $\mu$  sont choisis tels que  $Q_{N+1}(1) = Q_{N+1}(-1) = 0$ . Soit  $f$  dans  $\mathcal{P}_{2N-1}$ . La division de  $f$  par  $Q_{N+1}$  donne  $f = rQ_{N+1} + s$  où le polynôme quotient  $r$  est de degré  $\leq N - 2$  et le polynôme reste  $s$  est de degré  $\leq N$ . La suite de la démonstration est identique. On obtient alors les formules de Gauss-Lobatto.

On a en outre le résultat suivant: Si  $\omega$  est un poids de Jacobi  $= (1-x)^\alpha(1+x)^\beta$  alors  $(x^2-1)P'_N$  est orthogonal à  $\mathcal{P}_{N-2}$  et les  $x_i, 1 \leq i \leq N-1$  sont donc les racines de  $P'_N$ . En effet si  $r \in \mathcal{P}_{N-2}$  alors

$$\begin{aligned} \int_{-1}^1 (x^2-1)P'_N(x)r(x)\omega(x)dx &= (x^2-1)P_N(x)r(x)\omega(x)|_{-1}^1 \\ &- \int_{-1}^1 P_N(x)((x^2-1)r(x))'\omega(x)dx \\ &- \int_{-1}^1 P_N(x)((x^2-1)r(x))\omega'(x)dx \end{aligned}$$

Le terme de bord et le 2ème terme sont nuls. Si  $\omega = (1-x)^\alpha(1+x)^\beta$  alors  $\omega' = -\alpha\frac{\omega}{(1-x)} + \beta\frac{\omega}{(1+x)}$  et le 3ème terme est nul également.

Les points de Gauss-Lobatto prennent une forme explicite dans le cas Chebyshev. On a  $-\sin\theta T'_N(\cos\theta) = -N\sin N\theta$  et les racines de  $T'_N$  sont donc  $x_i = \cos \frac{i\pi}{N}, 1 \leq i \leq N-1$ . Les points de Gauss-Lobatto Chebyshev prennent donc la forme  $x_i = \cos \frac{i\pi}{N}, 0 \leq i \leq N$ . Les poids  $\alpha_i$  valent  $\frac{\pi}{\bar{c}_i N}$  avec  $\bar{c}_0 = \bar{c}_N = 2, \bar{c}_i = 1, 1 \leq i \leq N-1$ .

De façon générale on montre que les  $\alpha_i$  sont tous  $> 0$ .

Remarque: pour les polynômes de Legendre, puisque la fonction poids est 1, on a  $\int_{-1}^1 f(x)dx = \sum_{i=0}^N \alpha_i f(x_i)$  pour tout polynôme de degré  $\leq 2N+1$  ou  $2N-1$  selon que l'on utilise les poids et points de Gauss ou Gauss-Lobatto. Pour Chebyshev c'est  $\int_{-1}^1 f(x)(1-x^2)^{-1/2}dx$  qui vaut  $\sum_{i=0}^N \alpha_i f(x_i)$  et il ne faut donc pas chercher à utiliser la formule de quadrature pour évaluer  $\int_{-1}^1 f(x)dx$ .  $\int_{-1}^1 f(x)dx$  vaut  $\sum_{m \text{ pair}} \frac{2}{1-m^2} \hat{f}_m$ .

## Produits scalaires discrets

Ces formules d'intégration permettent de définir des produits scalaires discrets sur  $\mathcal{P}_N$ .

Considérons en effet la forme bilinéaire sur  $\mathcal{P}_N$ :

$$(u, v)_d^l = \sum_{j=0}^N \alpha_j u(x_j) v(x_j)$$

avec  $l = 1, 2$  selon que l'on considère les points de Gauss ou de Gauss Lobatto. Cette forme bilinéaire est symétrique et si de plus  $(u, u)_d = 0$ , alors  $u(x_j) = 0, \forall j = 0, \dots, N$  (puisque les  $\alpha_j > 0$ ) et on a alors  $u = 0$  (polynôme de degré  $\leq N$  avec  $N + 1$  zéros).

Les polynômes de Chebyshev continuent de former une famille orthogonale dans  $\mathcal{P}_N$  relativement aux deux produits scalaires définis soit sur les points de Gauss ou sur les points de Gauss-Lobatto.

En ce qui concerne le produit scalaire discret  $(\cdot)_d^1$ , si  $T_p, T_q \in \mathcal{P}_N$  alors  $(T_p, T_q)_d^1 = (T_p, T_q)_\omega$  puisque  $T_p T_q$  est dans  $\mathcal{P}_{2N}$  et on a donc

$$(T_p, T_q)_d^1 = \frac{c_p \pi}{2} \delta_{pq}$$

En ce qui concerne le produit scalaire discret  $(\cdot)_d^2$ , le raisonnement tient si  $T_p, T_q \in \mathcal{P}_{N-1}$  puisque  $T_p T_q$  est alors dans  $\mathcal{P}_{2N-2}$ . Par ailleurs si  $T_q$  est dans  $\mathcal{P}_{N-1}$  alors  $(T_N, T_q)_d^2 = (T_N, T_q)_\omega = 0$ , et donc la famille  $\{T_q, 0 \leq q \leq N\}$  forme une famille orthogonale dans  $\mathcal{P}_N$  pour  $(\cdot)_d^2$ . Par contre  $(T_N, T_N)_d^2 \neq (T_N, T_N)_\omega$  et un calcul direct montre que  $(T_N, T_N)_d^2 = 2(T_N, T_N)_\omega$  et on a donc

$$(T_p, T_q)_d^2 = \frac{\bar{c}_p \pi}{2} \delta_{pq}$$

On peut maintenant définir une expression approchée pour le spectre de  $f$ , que l'on appellera le pseudo-spectre. Dans  $\mathcal{P}_N$ , considérons le polynôme  $I_N f = \sum_{n=0}^N \tilde{f}_n T_n$  interpolant  $f$  en  $(N + 1)$  points  $x_j$ , c'est-à-dire tel que  $I_N f(x_j) = f(x_j)$ . Si l'on forme le produit scalaire discret avec  $T_l$ , il vient donc  $(I_N f, T_l)_d = \tilde{f}_l (T_l, T_l)_d$  et donc

$$\tilde{f}_l = \frac{(I_N f, T_l)_d}{(T_l, T_l)_d}$$

Pour le produit scalaire bâti sur les points de Gauss, on a donc:

$$\tilde{f}_l = \frac{2}{c_l(N+1)} \sum_{j=0}^N f(x_j) T_l(x_j) \quad (2.61)$$

avec  $x_j = \cos \frac{2j+1}{N+1} \frac{\pi}{2}$

Pour le produit scalaire bâti sur les points de Gauss-Lobatto, on a donc:

$$\tilde{f}_l = \frac{2}{\bar{c}_l N} \sum_{j=0}^N \frac{1}{\bar{c}_j} f(x_j) T_l(x_j) \quad (2.62)$$

avec  $x_j = \cos \frac{j\pi}{N}$



### Matrices de passage espace physique-espace spectral

On peut donc obtenir des matrices de passage espace physique-espace spectral permettant de passer des valeurs aux points de collocation  $f_j$  aux coefficients  $\tilde{f}_k$  et réciproquement des matrices de passage espace spectral-espace physique permettant de passer des coefficients  $\tilde{f}_k$  aux valeurs aux points de collocation  $f_j$ .

La matrice physique spectral sur les points de Gauss Lobatto  $\mathcal{PSGL}$  s'écrit

$$\mathcal{PSGL}_{ij} = \frac{1}{\bar{c}_i \bar{c}_j} \frac{2}{N} \cos \frac{ij\pi}{N}, 0 \leq i \leq N, 0 \leq j \leq N.$$

La matrice spectral-physique sur les points de Gauss Lobatto  $\mathcal{SPGL}$  s'écrit

$$\mathcal{SPGL}_{ij} = \cos \frac{ij\pi}{N}, 0 \leq i \leq N, 0 \leq j \leq N.$$

Sur les points de Gauss, ces matrices s'écrivent respectivement

$$\mathcal{PSG}_{ij} = \frac{1}{c_i} \frac{2}{N} \cos \frac{i(2j+1)\pi}{2(N+1)}, 0 \leq i \leq N, 0 \leq j \leq N$$

et

$$\mathcal{SPG}_{ij} = \cos \frac{(2i+1)j\pi}{2(N+1)}, 0 \leq i \leq N, 0 \leq j \leq N$$

### Relation entre les $\tilde{f}_l$ et les $\hat{f}_l$

La démarche est analogue à celle suivie pour les séries de Fourier. On a:

$$\tilde{f}_l = \frac{2}{\bar{c}_l N} \sum_{j=0}^N \frac{1}{\bar{c}_j} f(x_j) T_l(x_j) \quad (2.63)$$

$$= \frac{2}{\bar{c}_l N} \sum_{j=0}^N \frac{1}{\bar{c}_j} \sum_{m=0}^{\infty} \hat{f}_m T_m(x_j) T_l(x_j) \quad (2.64)$$

$$= \frac{2}{\bar{c}_l N} \sum_{m=0}^{\infty} \hat{f}_m \sum_{j=0}^N \frac{1}{\bar{c}_j} T_m(x_j) T_l(x_j) \quad (2.65)$$

$$= \hat{f}_l + \frac{2}{\bar{c}_l} \sum_{m=N+1}^{\infty} \hat{f}_m (T_m, T_l)_d^2 \quad (2.66)$$

Pour les polynômes de Chebyshev, on a  $(T_m, T_l)_d^2 = \frac{\bar{c}_l \pi}{2} \delta_{l, |m-2pN|}$  et donc

$$\tilde{f}_l = \hat{f}_l + \sum_{p=1}^{\infty} \hat{f}_{2pN \pm l} \quad (2.67)$$

## Dérivation

Dans  $\mathcal{P}_N$ , la dérivation est une application interne et si  $f_N \in \mathcal{P}_N$  on peut écrire:

$$\frac{\partial}{\partial x} f_N = \sum_{n=0}^N \hat{f}_n T'_n(x)$$

$T'_n$  étant un polynôme de degré  $n - 1$ , il peut donc s'exprimer sur la base des  $T_k, 0 \leq k \leq n$  et l'on peut donc écrire

$$\frac{\partial}{\partial x} f_N = \sum_{n=0}^N \hat{f}_n^{(1)} T_n(x)$$

En ce qui concerne les polynômes de Chebyshev, on se sert de la relation

$$\frac{T'_{n+1}}{n+1} - \frac{T'_{n-1}}{n-1} = \frac{2}{c_n} T_n$$

pour obtenir

$$T'_n = 2n \sum_{k=n-1(2)}^0 \frac{1}{c_k} T_k$$

où la notation  $(2)$  signifie de 2 en 2. Ceci permet de former la matrice de dérivation:

$$D = \begin{bmatrix} 0 & & & p+1 \\ 0 & 0 & & 2p & 0 \\ 0 & & 0 & 0 & 2(p+1) \\ 0 & & & 0 & 2p & 0 \\ 0 & & & & 0 & 2(p+1) \\ 0 & & & & & 0 \\ 0 & & & & & & 0 \end{bmatrix}$$

Cette matrice est triangulaire supérieure et toutes ses valeurs propres sont nulles.  $D$  est nilpotente et on a  $D^{N+1} = 0$ . Ceci permet d'obtenir les coefficients  $\hat{f}_n^{(1)}$  du développement du polynôme dérivée  $f'_N$ :

$$c_n \hat{f}_n^{(1)} = 2 \sum_{p=n+1(2)}^N p \hat{f}_p$$

L'évaluation pratique se fait à l'aide de la formule de récurrence

$$c_{n-1} \hat{f}_{n-1}^{(1)} - \hat{f}_{n+1}^{(1)} f = 2n \hat{f}_n$$

en procédant de la façon suivante:

$$\begin{aligned} \hat{f}_N^{(1)} &= 0 \\ \hat{f}_{N-1}^{(1)} &= 2N \hat{f}_N \\ \hat{f}_{N-2}^{(1)} &= 2(N-1) \hat{f}_{N-1} + \hat{f}_{N-1}^{(1)} \\ &\vdots \\ 2\hat{f}_0^{(1)} &= 2\hat{f}_1 + \hat{f}_2^{(1)} \end{aligned}$$

On peut donc maintenant exprimer la suite d'opérations qui permet d'obtenir  $\frac{d}{dx}I_N f(x_i)$  connaissant les  $f(x_i)$ . On obtient d'abord les coefficients du pseudo-spectre  $\tilde{f}_n$ ; on dérive ensuite le polynôme  $I_N f$  dans  $\mathcal{P}_N$ ; on réévalue ensuite  $I'_N f$  aux points de collocation. Cette suite d'opérations peut s'expliciter sous la forme matricielle suivante:

$$\begin{bmatrix} \frac{dI_N f}{dx}(x_0) \\ \cdot \\ \frac{dI_N f}{dx}(x_i) \\ \cdot \\ \frac{\partial I_N f}{\partial x}(x_N) \end{bmatrix} = \begin{bmatrix} SPGL_{il} \end{bmatrix} \begin{bmatrix} D_{lk} \end{bmatrix} \begin{bmatrix} PSGL_{kj} \end{bmatrix} \begin{bmatrix} f_0 \\ \cdot \\ f_j \\ \cdot \\ f_N \end{bmatrix}$$

On peut faire le produit de ces 3 matrices pour obtenir la matrice  $\mathcal{D}$  de dérivation de collocation aux points de Gauss-Lobatto.

On peut également obtenir cette matrice en dérivant le polynôme d'interpolation  $I_N f$ . Dans  $\mathcal{P}_N$ , le polynôme  $I_N f$  interpolant exactement  $f$  aux points de collocation  $x_j$  s'écrit directement sous la forme du polynôme d'interpolation de Lagrange

$$I_N f = \sum_{j=0}^N f_j L_j(x)$$

où  $L_j$  est le polynôme de degré  $N$  tel que  $L_j(x_i) = \delta_{ij}$ . de façon classique on a:

$$L_j = \frac{\Pi(x)}{(x - x_j)\Pi'(x_j)}$$

avec  $\Pi(x) = \prod_{i=0}^N (x - x_i)$  où les  $x_i$  sont les points de collocation. Les  $L_j$  sont donc les éléments de la base canonique de  $\mathcal{P}_N$  associés aux points de collocation. On a donc

$$I'_N f = \sum_{j=0}^N f_j L'_j(x)$$

et donc

$$I'_N f(x_i) = \sum_{j=0}^N f_j L'_j(x_i)$$

Les éléments  $\mathcal{D}_{ij}$  sont donc  $L'_j(x_i)$

La matrice de dérivation seconde dans  $\mathcal{P}_N$  est évidemment  $D^2$  et l'expression des coefficients du polynôme dérivée seconde est donnée par:

$$c_n \hat{f}_n^{(2)} = \sum_{p=n+2(2)}^N p(p^2 - n^2) \hat{f}_p$$

La matrice de dérivation seconde de collocation est elle  $\mathcal{D}^2$  et ses éléments sont donnés par  $L''_j(x_i)$

Remarque: On peut évidemment réévaluer

$$I'_N f = \sum_{j=0}^N f_j L'_j(x)$$

en un ensemble de points  $y_i$  distincts des points ayant servi à construire le polynôme interpolant  $I_N f$ . On obtient ainsi une matrice de collocation permettant de connaître  $I'_N f(y_i)$  en des points quelconques à partir des  $f(x_j)$ . La matrice représentative de cette application linéaire est donc de terme générique  $a_{ij} = L'_j y_i$ .

## 2.4 Résolution spectrale d'équations elliptiques

### 2.4.1 Généralités

On se propose dans ce paragraphe de définir diverses techniques de résolution d'équations elliptiques du type

$$(\nabla^2 - \lambda)f = s \text{ dans } \Omega \quad (2.68)$$

$$f = \bar{f} \text{ sur } \partial\Omega \quad (2.69)$$

où  $\Omega$  est un ouvert connexe de  $\mathbb{R}^2$  ou  $\mathbb{R}^3$ .

Ce type d'équation résulte de la discrétisation temporelle d'une équation de type transport-diffusion

$$\frac{\partial f}{\partial t} + \mathbf{V} \cdot \nabla f = \nabla^2 f$$

. L'utilisation de méthodes spectrales de type Chebyshev impose en effet une discrétisation de type implicite des termes de diffusion. En effet, une discrétisation explicite résulterait d'un critère de stabilité  $\Delta t \leq O(\frac{1}{N^4})$  où  $N$  est l'ordre de discrétisation. Ce résultat est lié au rayon spectral de la matrice de dérivation seconde dans la base des polynômes orthogonaux qui croît comme  $N^4$ . Cette discrétisation implicite est absolument impérative car le critère de stabilité est très restrictif. On se contente par ailleurs d'un traitement explicite des termes convectifs, le critère de stabilité associé étant de  $\Delta t \leq O(\frac{1}{N^4})$ , critère jugé acceptable, son traitement implicite étant difficile.

Divers schémas de discrétisation temporelle de précision croissante remplissent cette condition. Le plus simple est le schéma combinant discrétisation Euler-explicite/Euler-implicite qui s'écrit:

$$\frac{f^{n+1} - f^n}{\Delta t} + \mathbf{V} \cdot \nabla f^n = \nabla^2 f^{n+1}$$

Un schéma du second ordre classique est le schéma Adams-Bashforth/Crank-Nicolson qui s'écrit:

$$\frac{f^{n+1} - f^n}{\Delta t} + 3/2 \mathbf{V} \cdot \nabla f^n - 1/2 \mathbf{V} \cdot \nabla f^{n-1} = 1/2 (\nabla^2 f^{n+1} + \nabla^2 f^n)$$

Un autre schéma du second ordre est celui proposé par Vanel et al [11], qui combine une discrétisation Euler retardé du second ordre pour les termes diffusifs à une discrétisation Adams-Bashforth pour les termes convectifs:

$$\frac{3f^{n+1} - 4f^n + f^{n-1}}{2\Delta t} + 2\mathbf{V} \cdot \nabla f^n - \mathbf{V} \cdot \nabla f^{n-1} = \nabla^2 f^{n+1}$$

Ce schéma peut être généralisé à des ordres supérieurs et un schéma du 3ème ordre par exemple s'écrit:

$$\begin{aligned} & \frac{11f^{n+1} - 18f^n + 9f^{n-1} - 2f^{n-2}}{6\Delta t} \\ & + 3\mathbf{V} \cdot \nabla f^n - 3\mathbf{V} \cdot \nabla f^{n-1} + \mathbf{V} \cdot \nabla f^{n-2} = \nabla^2 f^{n+1} \end{aligned}$$

L'un quelconque de ces schémas peut se mettre sous la forme d'un problème de Helmholtz pour le champ inconnu  $f^{n+1}$ :

$$\nabla^2 f^{n+1} - \lambda f^{n+1} = S_f \quad (2.70)$$

où  $\lambda$  vaut typiquement  $\frac{C}{\Delta t}$ . Ce préambule justifie pourquoi on s'occupe de la résolution de problèmes elliptiques.

Diverses méthodologies de résolution des problèmes elliptiques linéaires sont disponibles. Deux grandes classes de méthodes correspondent au fait que l'on se donne comme inconnues les coefficients spectraux ou pseudo-spectraux de la solution inconnue, ou comme inconnues les valeurs de la fonction inconnue aux points de collocation. Chacune de ces classes peut donner lieu à des variantes, méthode de Galerkin ou méthode tau d'un coté, collocation forte ou faible de l'autre.

L'ensemble de ces méthodes appartiennent à la classe des méthodes de résidus pondérés, consistant à définir le résidu correspondant à la solution approchée  $P_N f$  ou  $I_N u$  et à annuler ce résidu en un certain sens.

## 2.4.2 Méthodes spectrales

### Méthode Spectrale de Galerkin

Ce type de méthode se définit par le fait que la solution est recherchée comme la projection  $L^2$   $P_N f = \sum_{n=0}^N \hat{f}_n \phi_n$  sur une base de fonctions  $\phi_n$  satisfaisant individuellement les conditions aux limites du problème différentiel. Les coefficients spectraux sont alors déterminés en demandant que le résidu de l'équation différentielle soit orthogonal (au sens  $L^2$ ) à la famille des  $\phi_n, n = 0, \dots, N$ . Cette méthodologie est intéressante si les  $\phi_n, n = 0, \dots, N$  forment une famille orthogonale et si les opérations de dérivation sont diagonales dans l'espace des  $\phi_n, n = 0, \dots, N$ . Elles conviennent donc parfaitement pour la mise en œuvre des approximations par des polynômes trigonométriques, qui remplissent ces deux conditions. Soit à résoudre

$$(\nabla^2 - \lambda)f = s \text{ sur } ]0, 2\pi[ \quad (2.71)$$

$$f \text{ } 2\pi \text{ périodique} \quad (2.72)$$

On recherche  $f$  sous la forme  $P_N f = \sum_{n=-N/2}^{N/2} \hat{f}_n \exp(inx)$ . Le résidu de l'équation vaut alors

$$R_N f = \sum_{n=-N/2}^{N/2} -(n^2 + \lambda) \hat{f}_n \exp(inx) - s$$

Les  $N+1$  coefficients sont déterminés en demandant que le résidu soit orthogonal aux  $\exp(ikx), k = -N/2, \dots, N/2$ , ce qui donne donc, compte-tenu de l'orthogonalité des  $\exp(ikx), k = -N/2, \dots, N/2$ , donne:

$$-(k^2 + \lambda) \hat{f}_k = \frac{(\exp(ikx), s)}{(\exp(ikx), \exp(ikx))} \quad (2.73)$$

ou encore

$$\hat{f}_k = -\frac{1}{2\pi} \frac{(\exp(ikx), s)}{(k^2 + \lambda)} \quad (2.74)$$

On a évidemment  $(\exp(ikx), s) = 2\pi \hat{s}_k$  et donc

$$\hat{f}_k = -\frac{\hat{s}_k}{(k^2 + \lambda)} \quad (2.75)$$

Dans l'espace des coefficients spectraux, l'inversion se réduit donc à une simple division scalaire, ce qui fait tout l'intérêt de cette formulation. On constate par contre de suite que si la détermination exacte des  $\hat{s}_k$  n'est pas possible, on va devoir recourir à une évaluation approchée de  $\hat{s}_k$  par  $\tilde{s}_k$ . Dans ce cas on a donc affaire à une méthode mixte Galerkin avec utilisation d'une quadrature numérique pour l'évaluation du terme source.

### Méthode Spectrale tau

Les polynômes de Chebyshev ne vérifiant pas des conditions aux limites fixes, il convient donc, pour mettre en œuvre une méthode de Galerkin, de définir une famille  $\{\phi_n; \phi_{2n} = T_{2n} - T_0; \phi_{2n+1} = T_{2n+1} - T_1\}$ , et on perd alors la propriété d'orthogonalité pour les  $\phi_n$ . Ceci explique pourquoi la méthode de Galerkin n'est pas fréquemment mise en œuvre dans le cas des polynômes de Chebyshev. La méthode tau, introduite par Lanczos, permet néanmoins de conserver l'idée de rendre le résidu orthogonal à un certain sous-espace engendré par les  $T_k$ .

On cherche à résoudre

$$(\nabla^2 - \lambda)f = s \text{ sur } ]-1, 1[ \quad (2.76)$$

$$f(-1) = f(1) = 0 \quad (2.77)$$

On cherche  $f$  sous la forme approchée  $P_N f = \sum_{n=0}^N \hat{f}_n T_n(x)$ , et on doit donc déterminer les  $\hat{f}_n$  à partir de  $N+1$  relation indépendantes. Imposer  $P_N f(1) = P_N f(-1) = 0$  fournit deux relations. Les  $N-1$  autres relations sont obtenues en demandant que le résidu  $R_N f = (P_N f)'' - \lambda P_N f - s$  soit orthogonal aux  $T_k, k = 0, \dots, N-2$ .<sup>1</sup>

Ces  $N-1$  relations et les conditions aux limites s'écrivent donc:

$$f_k^{(2)} - \lambda f_k = (s, T_k)_w, k = 0, \dots, N-2 \quad (2.78)$$

$$\sum_{k=0}^N f_k = \sum_{k=0}^N (-1)^k f_k = 0 \quad (2.79)$$

---

1. Plus généralement si on considère un opérateur différentiel  $L$  d'ordre  $k$ , qui nécessite donc  $k$  conditions aux limites, les  $N+1$  coefficients spectraux seront déterminés par les  $k$  conditions aux limites et en demandant que le résidu soit orthogonal à l'espace engendré par les  $T_k, k = 0, \dots, N-k$ .

Ce système linéaire peut donc se mettre sous la forme suivante:

$$\begin{bmatrix} 0 & & & & & & 0 \\ 0 & 0 & & & & & \\ 0 & & 0 & 0 & p(p^2 - k^2) & & 0 \\ 0 & & & 0 & 0 & 0 & 0 \\ 0 & & & & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ f_p \\ . \\ . \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} s_0 \\ . \\ s_k \\ . \\ s_{N-2} \\ 0 \\ 0 \end{bmatrix}$$

La résolution de ce système linéaire n'est possible que pour des valeurs de  $N$  pas trop élevées. Lorsque  $N$  devient grand, l'inversion "explose" en raison du mauvais conditionnement de la matrice.

Il est par contre possible de transformer ce système en un système équivalent, bien mieux conditionné et qui s'inverse donc sans difficulté, ce qui fait tout l'intérêt de la méthode tau. Cette relation se base sur la relation de récurrence:

$$c_{n-1}\hat{f}_{n-1}^{(1)} - \hat{f}_{n+1}^{(1)} = 2n\hat{f}_n$$

On écrit cette relation en reliant les dérivées seconde et première pour

$$c_{n-2}\hat{f}_{n-2}^{(2)} - \hat{f}_n^{(2)} = 2(n-1)\hat{f}_{n-1}^{(1)}$$

et pour

$$c_n\hat{f}_n^{(2)} - \hat{f}_{n+2}^{(2)} = 2(n+1)\hat{f}_{n+1}^{(1)}$$

En divisant la première par  $2(n-1)$  et la seconde par  $2(n+1)$  et en soustrayant les deux relations ainsi obtenues, il vient:

$$\frac{c_{n-2}\hat{f}_{n-2}^{(2)}}{4n(n-1)} - \frac{e_n\hat{f}_n^{(2)}}{2(n^2-1)} + \frac{e_{n+2}\hat{f}_{n+2}^{(2)}}{4n(n+1)} = \hat{f}_n \quad (2.80)$$

avec  $e_n = 1, n \leq N-2$ ,  $e_n = 0, n \geq N-2$ . Cette relation constitue la base de la transformation du système ci-dessus en un système tridiagonal équivalent. En effet, si l'on écrit (2.78) pour 3 indices  $n-2$ ,  $n$  et  $n+2$  et si l'on multiplie par les coefficients correspondants il vient donc:

$$\begin{aligned} \frac{c_{n-2}\lambda\hat{f}_{n-2}}{4n(n-1)} + \left(1 - \frac{(\lambda e_n)\hat{f}_n}{2(n^2-1)} + \frac{e_{n+2}\lambda\hat{f}_{n+2}}{4n(n+1)}\right) = \\ \frac{c_{n-2}\hat{s}_{n-2}}{4n(n-1)} - \frac{e_n\hat{s}_n}{2(n^2-1)} + \frac{e_{n+2}\hat{s}_{n+2}}{4n(n+1)} \doteq \tilde{s}_k \end{aligned} \quad (2.81)$$

pour  $n = 2, \dots, N$ . En ajoutant les deux relations provenant des conditions aux limites, écrites



en premier, le système prend la forme matricielle:

$$\begin{bmatrix} 1 & 1 & 1 & & & & \\ 1 & -1 & 1 & -1 & & & \\ x & 0 & x & 0 & x & 0 & 0 \\ 0 & x & 0 & x & 0 & x & 0 \\ 0 & & & & & & \\ 0 & & & x & 0 & x & 0 & 0 \\ 0 & & & & x & 0 & x & 0 \\ 0 & & & & & 0 & x & 0 & x \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ . \\ . \\ f_k \\ . \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \tilde{s}_2 \\ . \\ \tilde{s}_k \\ . \\ . \\ \tilde{s}_N \end{bmatrix}$$

Ce système peut être de plus découpé en deux systèmes portant respectivement sur les modes d'indice pair et impair, qui peuvent être résolus séparément<sup>2</sup>. Le système dont la résolution donne les coefficients pairs, s'écrit par exemple:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x & x & x & 0 & & 0 \\ 0 & x & x & x & 0 \\ 0 & 0 & x & x & x & 0 \\ 0 & & & 0 & x & x & x \\ 0 & & & & 0 & x & x \end{bmatrix} \begin{bmatrix} f_0 \\ f_2 \\ . \\ f_{2k} \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{s}_2 \\ . \\ \tilde{s}_{2k} \\ . \\ \tilde{s}_N \end{bmatrix}$$

qui peut être résolu par une méthode d'élimination de Gauss spécialement adaptée à la structure particulière de la matrice. Ce système quasi-tridiagonal se résout donc en  $\mathcal{O}(N)$  opérations. En caricaturant, on peut donc dire, que la méthode tau permet d'obtenir la précision spectrale pour le coût des différences finies!!

### 2.4.3 Méthodes de collocation

#### Collocation forte

Dans une méthode de collocation, les inconnues sont les valeurs  $f_i$  aux points de collocation de Gauss-Lobatto. La solution est recherchée sous la forme de son polynôme d'interpolation appartenant à  $P_N$ ,  $I_N f = \sum_{i=0}^N f_i L_i(x)$ , où  $L_i$  est le polynôme caractéristique de degré  $N$  associé au point  $x_i$ . Ce polynôme est évidemment caractérisé par  $N+1$  valeurs  $f_i$  et il faut donc  $N+1$  équations indépendantes. Celles-ci sont obtenues en demandant  $I_N f(1) = I_N f(-1) = 0$ , et en annulant le résidu  $R_N = (I_N f)'' - \lambda I_N f - s$  en tous les points  $x_i, i = 1, \dots, N-1$ , soit

$$((I_N f)'' - \lambda I_N f)(x_i) = s_i, (i = 1, \dots, N-1)$$

---

2. Cette séparation n'est possible que pour certains types de conditions aux limites

Le système linéaire donnant les  $f_i$  s'écrit donc,

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x & x & x & x & x & x & x \\ x & x & x & x & x & x & x \\ x & x & & a_{ij} & x & x & x \\ x & & & & x & x & \\ x & & & x & x & x & x \\ 0 & & 0 & 0 & 0 & 1 & \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ . \\ f_j \\ . \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} 0 \\ s_1 \\ . \\ s_i \\ . \\ s_{N-1} \\ 0 \end{bmatrix}$$

où le terme  $a_{ij}$  vaut  $L_j''(x_i) - \lambda\delta_{ij}$ . L'inversion de ce système peut être menée sans difficulté.

Dans le cas de conditions aux limites non-homogènes  $I_N f(1) = a$ ;  $I_N f(-1) = b$ , le système linéaire donnant les  $f_i$  s'écrit donc,

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x & x & x & x & x & x & x \\ x & x & x & x & x & x & x \\ x & x & & a_{ij} & x & x & x \\ x & & & & x & x & \\ x & & & x & x & x & x \\ 0 & & 0 & 0 & 0 & 1 & \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ . \\ f_j \\ . \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} a \\ s_1 \\ . \\ s_i \\ . \\ s_{N-1} \\ b \end{bmatrix}$$

Ce système peut être transformé en un système équivalent obtenu en découplant les valeurs intérieures et les valeurs aux extrémités. Le système portant sur les valeurs intérieures s'écrit:

$$\begin{bmatrix} x & x & x & x & x \\ x & x & x & x & x \\ x & x & a_{ij} & x & x \\ x & x & & x & \\ x & & x & x & x \end{bmatrix} \begin{bmatrix} f_1 \\ . \\ f_i \\ . \\ f_{N-1} \end{bmatrix} = \begin{bmatrix} s_1 - a_{1,0}a - a_{1,N}b \\ . \\ s_k - a_{k,0}a - a_{k,N}b \\ . \\ s_{N-1} - a_{N-1,0}a - a_{N-1,N}b \end{bmatrix}$$

Gottlieb et Lustman [13] ont montré que la matrice d'ordre  $N-1$  précédente était diagonalisable sur  $\mathcal{R}$ , à valeurs propres réelles négatives. Ce résultat n'est pas trivial puisque la matrice n'est pas symétrique.

Dans le cas d'une ou de conditions aux limites de Neumannn, par exemple  $f'(1) = a$ , celle-ci s'écrit donc  $I_N f'(1) = a$ . Le système linéaire donnant les  $f_i$  s'écrit donc,

$$\begin{bmatrix} L'_0(1) & L'_1(1) & . & . & & L'_N(1) \\ x & x & x & x & x & x & x \\ x & x & x & x & x & x & x \\ x & x & & L_j''(x_i) - \lambda\delta_{ij} & x & x & x \\ x & & & & x & x & \\ x & & & x & x & x & \\ 0 & & 0 & 0 & 0 & 1 & \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ . \\ f_i \\ . \\ . \\ f_N \end{bmatrix} = \begin{bmatrix} a \\ s_1 \\ . \\ s_k \\ . \\ s_{N-1} \\ 0 \end{bmatrix}$$

Dans le cas de conditions aux limites de Neumann aux deux extrémités, la transformation du système d'ordre  $N + 1$  en un système d'ordre  $N - 1$  par élimination de  $f_0$  et  $f_N$  est possible, en résolvant d'abord le système 2x2 donnant  $f_0$  et  $f_N$  en fonction des  $f_i, i = 1, \dots, N - 1$ , que l'on peut alors éliminer des  $N - 1$  autres équations. Il n'y a pas de résultat analogue au cas Dirichlet concernant le fait que la matrice soit diagonalisable sur  $\mathcal{R}$ , mais jusqu'à présent aucun contre-exemple n'a été obtenu en Chebyshev. En revanche, dans le cas Legendre, cette matrice présente des valeurs propres complexes pour des  $N$  supérieurs à 10, ce qui limite l'intérêt de la méthode.

### Collocation faible

Une façon d'obtenir que la matrice représentative de la dérivation seconde est symétrique est d'avoir recours à une formulation faible. Celle-ci n'est en fait avantageuse que dans le cas d'une approximation de type Legendre, qui sont les polynômes orthogonaux relatifs au poids unité (si la fonction poids n'est pas 1, l'intégration par partie ne conduit pas à une forme symétrique en  $u$  et  $v$ ). Pour le problème de Poisson, il y a alors l'équivalence (au sens des distributions) entre la formulation forte et la formulation faible: Trouver  $u \in P_N^0$

$$\int_{-1}^1 (\nabla u \cdot \nabla v + \lambda uv) = \int_{-1}^1 f v \quad (2.82)$$

$v \in P_N^0$ , espace des polynômes de degré  $\leq N$  s'annulant aux extrémités.  $u$  peut être approché par  $I_N u = \sum_{i=1}^{N-1} u_i L_i(x)$ . Les  $N - 1$  équations nécessaires à la détermination des  $u_i$  sont obtenues en faisant décrire à  $v$  la base caractéristique de  $P_N^0$ , c'est à dire en prenant  $v$  successivement égal à  $L_k, k = 1, \dots, N - 1$ .  $\nabla u \cdot \nabla v$  appartenant à  $P_{2N-2}$ , on peut remplacer l'intégrale continue par la formule de quadrature de Gauss-Lobatto-Legendre basée sur les points  $\xi_i$  et les poids  $\alpha_i$ . Les  $N - 1$  équations s'écrivent donc

$$\sum_{i=1}^{N-1} \alpha_i \left( \sum_{j=1}^{N-1} u_j L'_j(\xi_i) L'_k(\xi_i) + \lambda \sum_{j=1}^{N-1} u_j L_j(\xi_i) L_k(\xi_i) \right) = \sum_{i=1}^{N-1} \alpha_i f(\xi_i) L_k(\xi_i) \quad (2.83)$$

En intervertissant les signes somme et en notant que  $L_k(\xi_i) = \delta_{ik}$ , il vient

$$\sum_{j=1}^{N-1} u_j \left( \sum_{i=1}^{N-1} \alpha_i L'_j(\xi_i) L'_k(\xi_i) + \lambda \alpha_k \delta_{jk} \right) = \alpha_k f_k \quad (2.84)$$

dont la matrice est clairement symétrique en  $j$  et  $k$ .

### 2.4.4 Résolution de problèmes bidimensionnels

On se propose de résoudre

$$\nabla^2 u - \lambda u = s \text{ dans } \Omega \quad (2.85)$$

$$u = 0 \text{ sur } \partial\Omega \quad (2.86)$$

où  $\Omega = ]-1, 1[^2$ .

On cherchera  $u$  dans l'espace d'approximation  $P_N(x) \otimes P_M(z)$ .

Nous ne traiterons que le cas d'une méthode de collocation. Dans une méthode de collocation, les inconnues sont les valeurs  $f_{ij}$  aux points de collocation de Gauss-Lobatto. La solution est recherchée sous la forme de son polynôme d'interpolation appartenant à  $P_N \otimes P_M$ ,  $I_{NM}f = \sum_{i=0}^N \sum_{j=0}^M f_{ij} L_i(x) L_j(z)$ , où  $L_i$  est le polynôme caractéristique de degré  $N$  associé au point  $x_i$  et  $L_j$  est le polynôme caractéristique de degré  $M$  associé au point  $z_j$ . Le polynôme  $I_{NM}f$  est évidemment caractérisé par  $(N+1)(M+1)$  valeurs  $f_{ij}$  et il nous faut donc autant d'équations. Celles ci sont obtenues en demandant que  $I_{NM}f$  prenne des valeurs données sur les points de collocation situés sur la frontière, soit  $2(N-1) + 2(M-1) + 4 = 2(N+M)$  équations. Les  $(N+1)(M+1) - 2(N+M) = (N-1)(M-1)$  équations restantes sont obtenues en annulant le résidu  $R_{NM} = (I_{NM}f)'' - \lambda I_{NM}f - s$  en tous les points  $(x_i, z_j), i = 1, \dots, N-1, j = 1, \dots, M-1$ , soit

$$((I_{NM}f)'' - \lambda I_{NM}f)(x_i, z_j) = s_{ij}.$$

Dans le cas de conditions aux limites homogènes, les inconnues se réduisent donc aux  $f_{ij}, i = 1, \dots, N-1, j = 1, \dots, M-1$ . En collocation forte, le système linéaire donnant les  $f_{ij}$  s'écrit donc

$$\mathcal{D}_N \mathcal{F} + \mathcal{F} \mathcal{D}_M^t - \lambda \mathcal{F} = \mathcal{S} \quad (2.87)$$

où  $\mathcal{D}_N, \mathcal{D}_M$  sont respectivement les matrices d'ordre  $N-1$  et  $M-1$  représentatives des opérateurs de dérivation seconde à l'ordre  $N$  et  $M$ ,  $\mathcal{F}$  est la matrice 2D  $(N-1) \times (M-1)$  des inconnues et  $\mathcal{S}$  est le terme source.

Il existe une méthode directe efficace pour résoudre ce système linéaire. Le principe de la résolution repose sur le fait que la matrice  $\mathcal{D}_N$  est diagonalisable sur  $\mathcal{R}$   $\mathcal{D}_N = \mathcal{P} \Lambda_N \mathcal{P}^{-1}$  et de même  $\mathcal{D}_M = \mathcal{Q} \Lambda_M \mathcal{Q}^{-1}$ .

En posant  $\mathcal{U} = \mathcal{P}^{-1} \mathcal{F} \mathcal{Q}^{-1t}$ , la résolution se ramène à

$$\mathcal{U}_{ij} = \frac{(\mathcal{P}^{-1} \mathcal{S} \mathcal{Q}^{-1t})_{ij}}{\lambda_i + \lambda_j - \lambda} \quad (2.88)$$

pour  $i = 1, \dots, N-1, j = 1, \dots, M-1$ , d'où l'on tire  $\mathcal{F} = \mathcal{P} \mathcal{U} \mathcal{Q}^t$ .

Cette technique reste applicable pour le problème de Poisson ( $\lambda = 0$ ) en collocation faible. Par contre, pour le problème de Helmholtz, elle perd son intérêt.

## Chapitre 3

# L'erreur de discrétisation

### 3.1 Schémas numériques: notions de stabilité, convergence et consistance

Comment la solution numérique diffère-t-elle de la solution exacte? Il faut avant tout distinguer l'erreur d'arrondi liée aux capacités de la machine de calcul de l'erreur de discrétisation elle-même. Dans ce qui suit on s'intéressera exclusivement à ce second type d'erreur.

Un deuxième point est de faire la différence entre la solution au noeud  $i$  et au temps  $n$  de l'EDP suivante

$$\frac{\partial u(x,t)}{\partial t} = f(u(x,t)) \quad (3.1)$$

discrétisée comme

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = f_i(u_j^n, u_j^{n+1}) \quad (3.2)$$

on distinguera l'erreur de troncature qui porte sur la discrétisation de l'EDP

$$ET_i^n = [f(u)](x_i) - f_i \quad (3.3)$$

et l'erreur entre la solution numérique et la solution exacte

$$E_i^n = u(x_i, t^n) - u_i^n \quad (3.4)$$

**Définitions:** Un schéma numérique est **consistant** si l'erreur de troncature converge vers zéro lors que le maillage et le pas de temps tendent vers zéro.

Exercice - Montrer que le schéma de Dufort-Frankel pour l'équation de la chaleur

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = \frac{\alpha}{\Delta x^2} (u_{j+1}^n - u_j^{n+1} - u_j^{n-1} + u_{j-1}^n)$$

n'est pas consistant.

La notion de stabilité est plus difficile à établir. On utilise la notion de TVD (total variation decrease) qui caractérise le fait que l'erreur n'augmente pas d'un pas de temps à l'autre.

**Définition:** Un schéma est dit **TVD ou à variation totale bornée** si l'erreur n'augmente pas d'un pas de temps à l'autre. On a

$$TV(u^{n+1}) \leq TV(u^n)$$

où  $TV(u)$  (variation totale de  $u$ ) est définie comme

$$TV(u) = \sup \sum_{i=1}^N |u_i - u_{i+1}|$$

**Définition:** Un schéma est dit **convergent** si l'erreur entre la solution exacte et la solution numérique tend vers zéro.

**Théorème de Lax:** Si le problème est linéaire, le schéma consistant et stable sera convergent.

Remarque: Si le problème n'est pas linéaire, la convergence d'un schéma donné est difficile à établir.

## 3.2 Analyse de stabilité

Il s'agit ici de s'assurer que l'erreur n'est pas amplifiée par un schéma numérique donné. On est souvent amenés à étudier la stabilité d'un schéma au sens faible, en introduisant un type de perturbation donné et en s'assurant que cette perturbation ne croît pas. L'étude de stabilité se fait le plus facilement pour des EDP linéaires, pour lesquelles on notera que l'erreur satisfait la même équation que la solution exacte et la solution numérique.

### 3.2.1 Analyse de Von Neumann

La stabilité au sens faible se fait par l'analyse de Von Neumann. L'EDP est discrétisée et une équation d'évolution pour l'erreur  $\epsilon$  est obtenue. La question est de savoir si cette erreur croît exponentiellement avec le temps ou non. L'analyse de Von Neumann repose sur la sélection d'un mode de Fourier

$$u = \sum_n e^{ik_n x} e^{\alpha t} a$$

Il s'agit de s'assurer que chaque mode de Fourier de l'erreur décroît de manière monotone dans le temps. Si on note  $e_j^n$  l'erreur entre la solution exacte et la solution discrétisée au point de maillage  $j$  et au  $n$ -ième pas de temps, on a

$$e_j^n = \sum E_{jk}^n e^{kj\Delta x} e^{\alpha n \Delta t}$$

On définit le gain entre deux pas de temps pour chaque mode de Fourier (fréquence spatiale  $k$ )

$$G_j^n(k) = \frac{E_{jk}^{n+1}}{E_{jk}^n} \quad (3.5)$$

Le module du gain traduit la *diffusivité* du schéma, son argument ses effets dispersifs: différentes fréquences seront convectées à des vitesses diverses dans l'écoulement.

Nous calculons le gain pour quelques exemples de discrétisations des équations-modèles.

### Exemple 1 - Equation de convection

Considérons l'équation de convection

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad (3.6)$$

- discrétisation explicite - schéma spatial centré -

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + a \frac{u_{j+1}^n - u_{j-1}^n}{2\delta x} = 0 \quad (3.7)$$

La même équation est vérifiée par l'erreur entre la solution de l'équation aux dérivées partielles et son approximation:

$$\frac{e_j^{n+1} - e_j^n}{\delta t} + a \frac{e_{j+1}^n - e_{j-1}^n}{2\delta x} = 0 \quad (3.8)$$

En exprimant la variation du mode de Fourier  $k$  de l'erreur entre deux pas de temps comme

$$e_j^{n+1} = G e_j^n$$

on obtient que

$$G = \frac{e_j^{n+1}}{e_j^n} = 1 + \frac{a\delta t}{\delta x} (2i \sin(k\delta x)) \quad (3.9)$$

Le schéma est inconditionnellement instable puisque  $|G| > 1$  pour tout  $\sigma$  et tout  $k\Delta x$ .

- discrétisation explicite - schéma spatial décentré -

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + a \frac{u_j^n - u_{j-1}^n}{2\delta x} = 0 \quad (3.10)$$

Le schéma est conditionnellement stable puisque  $|G| < 1$  si  $|\sigma| < 1$ .

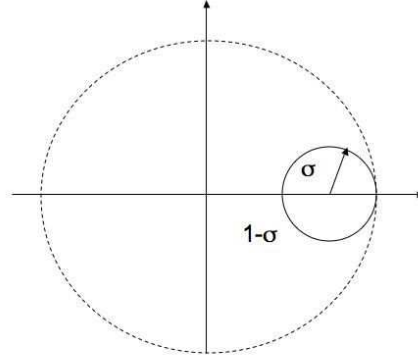


FIG. 3.1 – Représentation de la fonction gain  $G$  pour l'équation de convection discrétisée

On appelle  $\sigma$  le nombre de Courant qui permet de déterminer si l'information peut être propagée rapidement par la discrétisation choisie. La condition  $|\sigma| < 1$  est appelée condition de Courant-Friedrichs-Lewy ou condition CFL. La condition CFL est une limite sur le pas de temps induite par le maillage spatial considéré. On voit ici que le schéma décentré moins précis que le schéma centré permet à l'approximation de converger.

– discrétisation implicite - schéma spatial centré -

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + a \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\delta x} = 0 \quad (3.11)$$

$$G = \frac{e_j^{n+1}}{e_j^n} = \frac{1}{1 + \frac{a\delta t}{\delta x}(2i\sin(k\delta x))} \quad (3.12)$$

Le schéma est inconditionnellement stable puisque  $|G| < 1$  pour tout  $\sigma$  et tout  $k\Delta x$ . On voit ainsi l'intérêt des formulations implicites qui permettent d'alléger les contraintes liées au pas de temps.

## Exemple 2 - Equation de la chaleur

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad (3.13)$$

Discrétisation explicite - schéma centré

$$\frac{u_j^{n+1} - u_j^n}{\delta t} = \alpha \frac{u_{j+1}^{n+1} + u_{j-1}^{n+1} - 2u_j^n}{\delta x^2} \quad (3.14)$$

On définit

$$\beta = \frac{\alpha \Delta t}{\Delta x^2}$$



$$G = 1 + \alpha \frac{\delta t}{\delta x^2} (2\cos(k\delta x) - 1) \quad (3.15)$$

La condition de stabilité est

$$\beta \leq \frac{1}{2}$$

Le gain est purement réel: il n'y a donc aucun effet dispersif du schéma dans le cas de l'équation étudiée.

### Exemple 3 - Equation d'advection-diffusion

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = \alpha \frac{\partial^2 u}{\partial x^2} \quad (3.16)$$

Discretisation explicite - schéma centré pour la convection et la diffusion

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta x} = \alpha \frac{u_j^{n+1} + u_j^{n-1} - 2u_j^n}{\Delta x^2} \quad (3.17)$$

En définissant

$$\beta = \alpha \frac{\Delta t}{\Delta x^2}$$

Le gain  $G$  s'exprime comme

$$G = 1 - i\sigma \sin k\Delta x + 2\beta(\cos k\Delta x - 1) \quad (3.18)$$

La représentation polaire du gain  $G$  est donnée dans la figure 3.2.

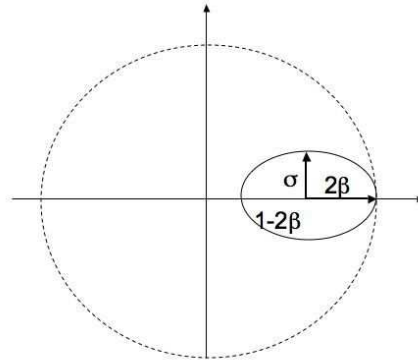


FIG. 3.2 – Représentation de la fonction gain  $G$  pour l'équation d'advection-diffusion discrétisée

ce qui conduit à la condition

$$\sigma^2 \leq 2\beta \leq 1$$

Il est utile de définir le rapport  $\sigma/\beta$

$$R = \frac{\beta}{\sigma} = \frac{a\Delta x}{\alpha}$$

$R$  est analogue à un nombre de Reynolds défini pour la cellule. La condition de stabilité s'exprime alors comme

$$\sigma \leq 2R \leq \frac{1}{\sigma}$$

### 3.2.2 Stabilité matricielle

La stabilité au sens fort s'étudie de manière globale.

Rappel: On définit la norme de la matrice  $C$  comme

$$\|C\| = \text{Max}_u \frac{|Cu|}{|u|}$$

**Définitions:** On appelle le rayon spectral d'une matrice le module de sa plus grande valeur propre.

$$\rho(C) = \text{Max}_j |\lambda_j|$$

**Propriété:**  $\|C\| \geq \rho_C$

Reprenons l'équation d'advection-diffusion (3.17)

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + a \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\delta x} - \alpha \frac{u_{j+1}^{n+1} + u_{j-1}^{n+1} - 2u_j^n}{\delta x^2} = 0 \quad (3.19)$$

$$C = \begin{bmatrix} 1 - 2\beta & \beta - \frac{\sigma}{2} & 0 & \dots \\ \beta + \frac{\sigma}{2} & 1 - 2\beta & \beta - \frac{\sigma}{2} & \dots \\ 0 & \beta + \frac{\sigma}{2} & 1 - 2\beta & \beta - \frac{\sigma}{2} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \quad (3.20)$$

Pour que le schéma soit stable, il faut et il suffit que le rayon spectral de la matrice soit inférieur à 1, c'est-à-dire que le module de chaque valeur propre soit inférieur à 1.

Les valeurs propres (éventuellement complexes) de la matrice  $C$  sont données par

$$\lambda_j(C) = 1 - 2\beta + 2\beta \sqrt{1 - \frac{R^2}{4}} \cos\left(\frac{j\pi}{N+1}\right)$$

On a donc 3 cas possibles:

–  $R < 2$

Les valeurs propres sont réelles et la condition  $|\lambda_j| < 1$  devient

$$\beta \leq \frac{1}{1 + \sqrt{1 - \frac{R^2}{4}}}$$

–  $R = 2$

Les valeurs propres sont toutes égales à  $1 - 2\beta$  et la condition de stabilité est alors

$$\beta \leq \frac{1}{2}$$

–  $R > 2$

Les valeurs propres sont complexes et la condition  $|\lambda_j| < 1$  entraîne que

$$\beta \frac{R^2}{4} < 1$$

La figure 3.3 compare les restrictions dans l'espace des paramètres  $(R, \sigma)$ . Elle montre que l'analyse globale conduit à un critère beaucoup moins restrictif sur  $\sigma$  en fonction du nombre de mailles.

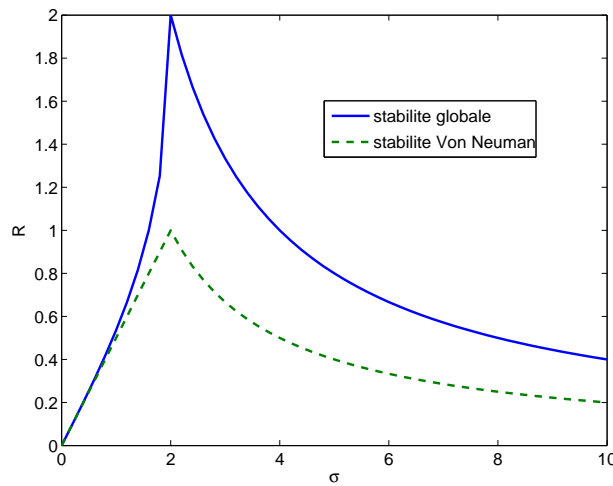


FIG. 3.3 – Comparaison de la méthode de Von Neumann et l'analyse matricielle pour le critère de stabilité de l'équation d'advection-diffusion discrétisée

## 3.3 Résolution itérative d'un problème linéaire

### 3.3.1 Conditionnement de l'opérateur

On se propose de résoudre le système

$$Ax = b \tag{3.21}$$

On suppose que la solution classique consistant à inverser la matrice est souvent trop coûteuse, ou mal conditionnée. Le mauvais conditionnement d'un système apparaît souvent lorsque la taille du problème devient grande et constitue un problème récurrent en Mécanique des Fluides.

Pour prendre la mesure de ce qu'est un problème mal conditionné considérons le système suivant (1)

$$\begin{aligned} 3x + 2y &= 1 \\ 6x + 4.01y &= 0 \end{aligned}$$

et un autre (2) qui en semble assez proche

$$\begin{aligned} 3x + 2y &= 1 \\ 6.01x + 4y &= 0 \end{aligned}$$

Le système (1) admet la solution ( $x=-200, y=300.5$ ) et le système (2) admet la solution ( $x=133.67, y=-200$ ). Les deux solutions sont très différentes bien que les coefficients des deux systèmes linéaires diffèrent de moins de 1% en norme. De petites variations dans les coefficients génèrent de larges erreurs dans la solution. sont caractéristiques d'un système mal conditionné.

**Définition:** Le conditionnement de la matrice est défini par

$$\begin{aligned} \kappa(A) &= \|A\| \|A^{-1}\| \text{ si } A \text{ n'est pas singulière} \\ \kappa(A) &= \infty \text{ sinon} \end{aligned}$$

Le problème est mal conditionné quand  $\kappa$  devient grand. Le calcul de  $\kappa(A)$  se fait le plus facilement en utilisant la décomposition en valeur singulière (SVD- Singular Value Decomposition) qui permet d'écrire une matrice  $A$  (qui peut être rectangulaire de taille  $m$  par  $n$ ) comme le produit de trois matrices

$$A = U^T D V$$

où

- $D$  est une matrice diagonale de taille  $n$  composée de valeurs positives ou nulles appelées les *valeurs singulières* de la matrice  $A$ .
- $U$  est une matrice  $m$  par  $n$  orthogonale par colonnes. Les  $i$  colonnes de  $U$  telles que  $d_i > 0$  forment une base de l'image de  $A$   $ImA$ .
- $V$  est une matrice de taille  $n$  orthogonale. Les  $i$  colonnes de  $V$  telles que  $d_i = 0$  forment une base du noyau de  $A$   $KerA$ .

**Propriété:** Le conditionnement de la matrice peut se calculer à partir des valeurs singulières

$$\kappa(A) = \frac{d_i^{max}}{d_i^{min}}$$

Face à un problème mal conditionné, l'idée est d'utiliser un préconditionneur. Au lieu de résoudre

$$Ax = b$$

on résout

$$P^{-1}Ax = P^{-1}b$$

où le conditionnement de la matrice  $P^{-1}A$  est meilleur (plus faible).

### 3.3.2 Méthodes itératives

Cette idée constitue la base des méthodes itératives. On décompose la matrice  $A$  comme

$$A = D - L - U$$

où  $D, L, U$  représentent respectivement les parties diagonale, triangulaires inférieure et supérieure de  $A$ .

#### Méthode de Jacobi

En utilisant la partie diagonale de  $A$  comme préconditionneur, on peut réécrire le système comme

$$Dx = (L + U)x + b$$

ou

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

On définit  $y^n$  la solution approchée de l'équation à la  $n$ -ème itération. Soit  $y^0$  l'estimée initiale. On itère pour obtenir la solution à l'itération  $n$

$$y^n = D^{-1}(L + U)y^{n-1} + D^{-1}b$$

Cette méthode est la méthode de Jacobi. On peut noter  $R$  la matrice d'itération.

$$R = D^{-1}(L + U)$$

Exemple: On considère l'équation de la chaleur

$$\frac{\partial u}{\partial x^2} = f$$

avec les conditions aux limites  $u(0) = u(1) = 0$ . On discrétise

$$\frac{u_{i-1} + u_{i+1} - 2u_i}{\delta x^2} = f_j$$

On choisit une condition initiale pour la solution  $v^{(0)} = u_e$  et on résout de manière itérative

$$u_i^{(J)} = \frac{f_j \delta x^2 + u_{i-1}^{(J-1)} + u_{i+1}^{(J-1)}}{2}$$

On voit ici que la méthode de Jacobi nécessite le stockage mémoire de deux solutions: l'itération au pas  $J - 1$  et celle au pas  $J$ .

Soit  $u$  la solution exacte du problème et  $v^{(J)}$  la solution obtenue après  $J$  itérations. L'erreur entre la solution exacte et la solution itérée  $e^{(J)} = u - v^{(J)}$  vérifie l'équation résiduelle

$$Re^{(J)} = r^{(J)}$$

où le résidu de l'équation est défini comme

$$r^{(J)} = b - Av^{(J)}$$

L'erreur commise à l'itération  $J$  vérifie l'équation récursive

$$e^{(J)} = R_J e^{(J-1)} \quad (3.22)$$

Une condition nécessaire et suffisante pour que la méthode converge est donc que le rayon spectral de la matrice  $R$  soit inférieur à 1.

Les valeurs de la matrice  $R$  associée à la matrice de Jacobi sont

$$\lambda_k = 1 - 2\sin^2\left(\frac{k\pi}{2n}\right)$$

pour  $1 \leq k \leq n - 1$ . Elles sont associées aux vecteurs propres

$$w_j^k = \sin\left(\frac{kj\pi}{n}\right)$$

qui sont les modes de Fourier. Le mode de Fourier  $k$  de l'erreur va donc décroître avec un coefficient  $\lambda_k$ .

On voit que pour les petits nombres d'onde  $k$ , correspondant à de basses fréquences (spatiales), la valeurs propre est proche de 1:

$$\lambda_k = 1 - \frac{1}{2}\left(\frac{\pi k}{n}\right)^2 \sim 1$$

Ceci est également vrai pour les grands nombres d'onde  $k$  (le sinus tend vers  $\sin\frac{\pi}{2}$ ).

### Méthode de Jacobi retardé

On introduit une variante de cette méthode qui est la méthode de Jacobi pondérée, où l'estimée est pondérée par l'estimation précédente. On a alors

$$y^n = (1 - \omega)y^{n-1} + \omega(D^{-1}(L + U)y^{n-1} + D^{-1}b)$$

Si on définit la matrice d'itération pondérée comme

$$R_\omega = (1 - \omega)I + \omega R_J$$

Les nouvelles valeurs propres  $\lambda_k$  sont de la forme

$$\lambda_k = 1 - 2\omega \sin^2\left(\frac{k\pi}{2n}\right)$$

et les vecteurs propres  $w^k$  sont toujours les modes de Fourier.

On voit que le gain est toujours proche de 1 aux basses fréquences, mais tend vers  $1 - 2\omega$  aux hautes fréquences. Un choix  $\omega = \frac{2}{3}$  conduit donc à une convergence rapide ( $\lambda \leq \frac{1}{2}$ ) de toutes les hautes fréquences ( $\frac{n}{2} \leq k \leq n - 1$ ).

### Méthode de Gauss-Seidel

Un autre exemple est la méthode de Gauss-Seidel où les nouveaux éléments calculés viennent directement remplacer ceux de l'itération précédente, ce qui conduit à ne stocker que l'itération courante (mise à jour ou non). Dans le cas de l'équation de la chaleur, l'itération conduit à

$$u_i^{(J)} = \frac{f_j \delta x^2 + u_{i-1}^{(J-1)} + u_{i+1}^{(J-1)}}{2}$$

$$u_i^{(J-1)} = u_i^{(J)}$$

si le passage se fait dans le sens des  $j$  croissants et

$$u_i^{(J+1)} = u_i^{(J)}$$

dans le sens des  $i$  croissants. On peut aussi alterner de sens d'une itération à l'autre. On peut également mettre à jour successivement les rangées paires et impaires qui sont découplées (procédure Gauss-Seidel Noir/Rouge). De manière générale, cette itération s'écrit (pour les  $i$  croissants)

$$y^n = (D - L)^{-1}(U)y^{n-1} + (D - L)^{-1}b$$

Le rayon de la matrice spectrale peut être calculé. On montre que les valeurs propres  $\lambda_k^{GS}$  sont de la forme

$$\lambda_k^{GS} = \cos^2\left(\frac{k\pi}{n}\right)$$

.

Les méthodes itératives présentent en général la **propriété de relaxation**: *La plupart des schémas de relaxation tendent à faire décroître rapidement les hautes fréquences de l'erreur.*

Cette idée est le point de départ des méthodes multigrille, puisqu'il va s'agir d'accélérer la convergence de la méthode itérative en augmentant relativement une fréquence donnée par ajustements successifs de la grille.

### 3.3.3 Méthodes multi-grille

Les méthodes multigrille visent à accélérer la convergence d'une méthode itérative en imposant une correction globale obtenue sur une grille plus large.

Le résidu de l'équation d'itération contient des basses fréquences. La notion de basse ou de haute fréquence s'entend par rapport au maillage. Si on suppose que le maillage est régulier et que la taille du maillage est  $h$ , la plus haute fréquence qui peut être capturée est  $\frac{1}{2h}$ . On considère que les fréquences plus grandes que  $\frac{1}{4h}$  sont amorties rapidement. Si on considère un maillage plus large de taille  $2h$ , le même critère nous indique que les fréquences qui vont être considérées comme hautes sont les fréquences plus grandes que  $\frac{1}{2h}$ . L'idée est donc d'utiliser la grille la plus large pour faire décroître l'erreur associée aux basses fréquences, puis de transférer cette correction sur la grille plus fine. Les passages d'une grille vers l'autre nécessitent de définir des opérateurs de transfert

- **l'injection** de la grille plus fine vers la grille plus large
- **l'interpolation** de la grille plus large vers la grille plus fine.

Considérons deux grilles de résolution respective  $h$  et  $2h$ . On définit ainsi le schéma de correction de la grille grossière vers la grille fine:

1. On résoud l'équation  $A_h x_h = B_h$  sur une grille fine avec une méthode itérative. Après un nombre donné  $N_1$  d'itérations on obtient une estimée  $v_h^{(N_1)}$  pour  $x_h$ . Le suffixe  $h$  renvoie à la discrétisation de l'équation sur la grille  $h$ . On obtient un résidu  $r_h = B_h - A_h v_h$
2. Ce résidu  $r_h$  est injecté sur la grille plus large de taille  $2h$ . Plusieurs méthodes d'injection sont possibles. On peut simplement définir
 
$$r_{2h}(x_j) = r_h(x_j) \text{ ou utiliser une interpolation linéaire}$$

$$r_{2h}(x_j) = \frac{r_h(x_j) + r_h(x_{j+1})}{2}$$
 On injecte de même la solution itérée  $v_h \rightarrow v_{2h}$ .
3. Le résidu injecté est à présent relaxé en utilisant l'équation d'erreur

$$Re_{2h} = r_{2h}$$

La question est de savoir comment définir l'erreur initiale sur la grille large. Puisque la solution réelle n'est pas connue, on peut simplement poser  $e_{2h}^{(0)} = v_{2h}$ . Au bout d'un nombre  $N_2$  d'itérations, on obtient une nouvelle erreur  $e_{2h}^{N_2}$ .

4. L'erreur est à présent interpolée sur la solution de la grille  $v_h^{(N_1)}$ . L'interpolation de la solution de la grille large sur la grille fine peut se faire de plusieurs manières. La plus simple est l'interpolation linéaire (voir figure 3.4):

$$e_h(x_{2j}) = e_{2h}(x_j)$$

$$e_h(x_{2j+1}) = \frac{e_{2h}(x_j) + e_{2h}(x_{j+1})}{2} \text{ La nouvelle estimée devient}$$

$$v_h = v_h^{(N_1)} + e_h$$

Les passages entre les grilles plus fines et les grilles plus larges seront non monotones. On définit ainsi des cycles de multigrille en V et en W (voir figure 3.5).



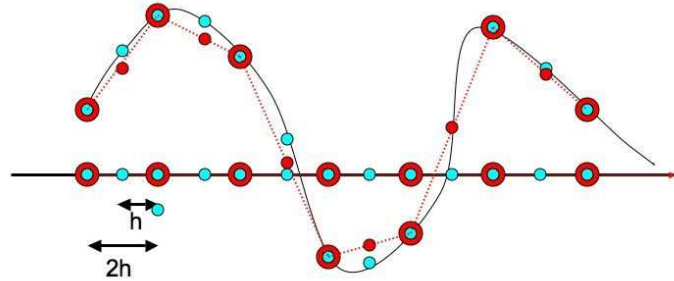


FIG. 3.4 – *Interpolation linéaire d'une fonction de la grille large vers la grille fine*

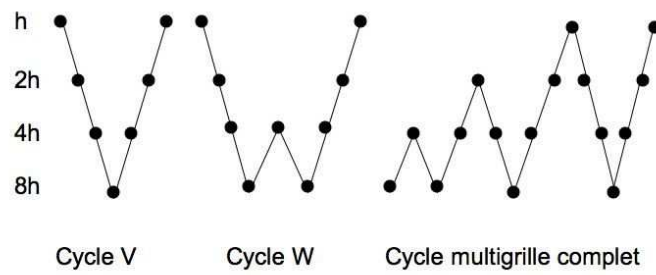


FIG. 3.5 – *Cycles de multi-grille*

## Chapitre 4

# Résolution du problème de Stokes

### 4.1 Formulation

Mises sous forme adimensionnelle à l'aide d'une vitesse de référence caractéristique, et d'une longueur caractéristique les équations de Navier-Stokes régissant un écoulement de fluide incompressible, écrites en variables vitesse–pression, sont:

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0 \quad (4.1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + w \frac{\partial u}{\partial z} = -\frac{\partial P}{\partial x} + \frac{1}{Re} \nabla^2 u \quad (4.2)$$

$$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + w \frac{\partial w}{\partial z} = -\frac{\partial P}{\partial z} + \frac{1}{Re} \nabla^2 w \quad (4.3)$$

où  $u$  et  $w$  sont respectivement les composantes horizontale et verticale du vecteur vitesse et  $P$  la pression. On supposera que le domaine de calcul,  $\Omega$ , s'étend de 0 à 1 dans les directions  $x$  et  $z$  en coordonnées adimensionnelles. Les conditions aux limites sont alors:

–  $u$  et  $w = 0$  sur la frontière  $\partial\Omega$  de  $\Omega$

L'application de l'un quelconque des schémas de discrétisation temporelle présentés au paragraphe précédent aux équations (4.1, 4.2, 4.3) montre que l'on doit résoudre à chaque pas de temps le système d'équations suivant:

$$\text{dans } \Omega \nabla^2 u^{n+1} - \lambda u^{n+1} = -Su + \frac{\partial P}{\partial x} \text{ dans } \Omega \quad (4.4)$$

$$\nabla^2 w^{n+1} - \lambda w^{n+1} = -Sw + \frac{\partial P}{\partial z} \text{ dans } \Omega \quad (4.5)$$

$$\frac{\partial u^{n+1}}{\partial x} + \frac{\partial w^{n+1}}{\partial z} = 0 \quad (4.6)$$

$$u^{n+1} = 0 \text{ sur } \partial\Omega \quad (4.7)$$

$$w^{n+1} = 0 \text{ sur } \partial\Omega \quad (4.8)$$

où  $\lambda = \frac{1.5Re}{\Delta t}$ .

Les équations (4.4,4.5,4.6, 4.7, 4.8) constituent un problème de Stokes instationnaire pour les composantes de la vitesse et le champ de pression. Ce problème doit être résolu à chaque pas de temps et doit donc être résolu de manière efficace.

On distingue deux grandes familles de méthodes pour la résolution approchée de ce problème de Stokes instationnaire. L'une repose sur une approche découplée, basée sur l'introduction d'une équation de Poisson pour la pression, tandis que l'autre repose sur une résolution couplée, qui, en éliminant la vitesse, se ramène à la résolution d'une équation pour la pression connue sous le nom d'opérateur d'Uzawa.

## 4.2 Equation de Poisson pour la Pression

### 4.2.1 Principe

De façon classique, cette équation de Poisson est obtenue en prenant la divergence de l'équation de quantité de mouvement. Pour le problème continu, la divergence des équations de quantité de mouvement (4.4,4.5) s'écrit donc;

$$(\nabla^2 - \lambda)\mathcal{D}^{n+1} = \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} - \left(\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}\right) \quad (4.9)$$

où  $\mathcal{D}$  est la divergence du champ de vitesse.

Il apparaît donc immédiatement que l'équation de poisson

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = \left(\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}\right) \quad (4.10)$$

est une condition nécessaire à la satisfaction de la contrainte de divergence nulle, mais qu'elle n'est pas suffisante. En effet, si cette équation est vérifiée, la divergence ne vérifie que l'équation d'Helmholtz homogène

$$(\nabla^2 - \lambda)\mathcal{D}^{n+1} = 0 \quad (4.11)$$

qui n'est autre que l'équation de la chaleur homogène. Une condition suffisante peut être obtenue en notant que cette équation n'admet des solutions identiquement nulles que si  $\mathcal{D}^{n+1} = 0$  sur  $\partial\Omega$ , à condition que la divergence soit identiquement nulle à  $t = 0$ . (Cette condition peut se ramener à  $\frac{\partial u_n}{\partial n} = 0$ .)

Le système découplé s'écrit donc, en laissant tomber l'indice de discrétisation temporelle;

$$\nabla^2 u - \lambda u = -Su + \frac{\partial P}{\partial x} \text{ dans } \Omega \quad (4.12)$$

$$\nabla^2 w - \lambda w = -Sw + \frac{\partial P}{\partial z} \text{ dans } \Omega \quad (4.13)$$

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = \left( \frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z} \right) \quad (4.14)$$

$$u^{n+1} = 0 \text{ sur } \partial\Omega \quad (4.15)$$

$$w^{n+1} = 0 \text{ sur } \partial\Omega \quad (4.16)$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0 \text{ sur } \partial\Omega \quad (4.17)$$

$$(4.18)$$

On constate immédiatement la difficulté pour résoudre ce système de trois équations elliptiques puisque, si chacune des équations pour  $u$  et  $w$  dispose de conditions aux limites naturelles, l'équation de Poisson pour la pression ne dispose que de conditions aux limites portant sur la divergence de la vitesse.

Une façon de résoudre ce système repose sur la résolution d'un problème auxiliaire, où l'on remplace cette condition aux limites portant sur  $\mathcal{D}$  par une condition aux limites de type Dirichlet portant sur la pression:

$$\nabla^2 u - \lambda u = -Su + \frac{\partial P}{\partial x} \text{ in } \Omega \quad (4.19)$$

$$\nabla^2 w - \lambda w = -Sw + \frac{\partial P}{\partial z} \text{ in } \Omega \quad (4.20)$$

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = \left( \frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z} \right) \quad (4.21)$$

$$u^{n+1} = 0 \text{ on } \partial\Omega \quad (4.22)$$

$$w^{n+1} = 0 \text{ on } \partial\Omega \quad (4.23)$$

$$P = \bar{P} \text{ on } \partial\Omega \quad (4.24)$$

Il est alors clair que l'on peut résoudre séquentiellement ce système, d'abord pour la pression et ensuite, une fois le champ de pression et son gradient calculés, pour chacune des composantes de vitesse. La difficulté consiste donc maintenant à trouver la pression au bord  $\bar{P}$  qui garantisse  $\bar{\mathcal{D}} = 0$ . Ceci peut être fait simplement en raison de la linéarité de l'application qui à  $\bar{P}$  fait correspondre  $\bar{\mathcal{D}}$ . (cette correspondance est linéaire dans le problème homogène  $(Su, Sw) = (0, 0)$ , affine sinon.)

Le principe de la résolution consiste donc à résoudre une première fois le problème auxiliaire avec une distribution de pression arbitraire  $\bar{P}_e$  sur le bord

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = \left( \frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z} \right) \quad (4.25)$$

$$P = \bar{P}_e \text{ on } \partial\Omega \quad (4.26)$$

ce qui permet d'obtenir  $P_e$  et de résoudre ensuite

$$(\nabla^2 - \lambda)u_e = -Su + \frac{\partial P_e}{\partial x} \text{ in } \Omega \quad (4.27)$$

$$(\nabla^2 - \lambda)w_e = -Sw + \frac{\partial P_e}{\partial z} \text{ in } \Omega \quad (4.28)$$

$$u^{n+1} = 0 \text{ on } \partial\Omega \quad (4.29)$$

$$w^{n+1} = 0 \text{ on } \partial\Omega \quad (4.30)$$

Le champ de vitesse  $(u_e, w_e)$  n'est généralement pas à divergence nulle et en particulier on peut déterminer sa divergence au bord  $\overline{\mathcal{D}_e}$ . Grace à la linéarité de l'application qui à la pression au bord  $\overline{P}$  fait correspondre  $\overline{\mathcal{D}}$ , il est donc immédiat de déterminer une pression  $\overline{P}_e$  à laquelle correspond une divergence  $-\overline{\mathcal{D}_e}$ , pour le problème homogène. La pression au bord  $\overline{P}_e + \overline{P}_c$  est donc la pression recherchée, qui garantit la divergence nulle au bord et donc le fait que le champ de vitesse soit à divergence nulle, modulo les restrictions déjà énoncées.

### 4.2.2 Mise en œuvre

La mise en œuvre de cette méthode demande que l'on ait fait un choix des espaces d'approximation. La première application de cet algorithme semble avoir été faite par Kleiser et Schuman en 1980, dans un contexte 1D Chebyshev et 2D Fourier pour traiter l'écoulement de Poiseuille entre deux plans parallèles. Nous nous placerons ici directement dans un problème comportant au moins deux directions non-périodiques qui nécessitent donc une approximation 2D Chebyshev. Dans ce cas, on recherche les composantes de vitesse dans  $P_N(x) \otimes P_M(z)$ . Pour les raisons que nous avons déjà évoquées,  $u, w$  sont considérés connus à travers leurs valeurs aux points de collocation  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$  où  $\mathcal{GL}_x^N = \{x_i, x_i = \cos(\frac{i\pi}{N}), i = 0, \dots, N\}$  sont les  $N + 1$  points de Gauss-Lobatto (racines de  $(1 - x^2)T'_N(x)$ ) dans la direction  $x$ . De même, puisque nous avons à résoudre une équation de Poisson pour la pression munie de conditions aux limites de Dirichlet, il semble donc naturel de rechercher également  $P$  dans  $P_N(x) \otimes P_M(z)$  et de la considérer connue par ses valeurs aux points  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$ .

Il convient donc de déterminer dans un premier temps la dimension de l'espace vectoriel de la "trace" sur le bord d'une fonction de  $P_N(x) \otimes P_M(z)$ . Il est clair que cette dimension est égale au nombre de points de collocation situés sur le bord du domaine de résolution soit  $2(N - 1) + 2(M - 1) + 4 = 2(N + M) = K$ . La matrice de l'application linéaire qui à la pression sur le bord fait correspondre la divergence au bord est donc a priori d'ordre  $K$ . En fait il est facile de montrer que le rang de cette matrice est d'ordre  $K - 5$ , que l'on résolve les problèmes d'Helmholtz par une méthode tau ou par une méthode de collocation.

Montrons le dans le cas d'une méthode de collocation. Considérons l'élément de la base canonique des distributions de pression sur le bord valant 1 en un coin du domaine et 0 partout ailleurs. Il est clair que le gradient de pression évalué par une méthode de collocation, bien que non identiquement nul en tant qu'élément de  $P_N(x) \otimes P_M(z)$ , est nul en tous les points

intérieurs du domaine de résolution, là où on doit calculer le terme source des problèmes d'Helmholtz pour chacune des composantes de vitesse. Chacune des composantes de vitesse est donc identiquement nulle, et par conséquent la divergence du champ de vitesse correspondant à cette pression sur le bord. Nous avons donc identifié 4 vecteurs du noyau de cette application linéaire. Considérons également la pression unité sur le bord du domaine, c'est-à-dire la pression qui vaut 1 en tous les points frontière. Le champ de pression solution de

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = 0 \quad (4.31)$$

$$P = 1 \text{ on } \partial\Omega \quad (4.32)$$

est donc la pression unité partout en tant qu'élément de  $P_N(x) \otimes P_M(z)$ . Son gradient est donc identiquement nul et donc la divergence du champ de vitesse associé. Nous avons donc identifié 5 éléments du noyau de cet opérateur, ce qui est confirmé par l'expérience numérique. Notons que de ces 5 éléments du noyau, 4 sont véritablement parasites et liés aux choix de discrétisation effectués, alors que le mode constant est intrinsèquement présent dans la formulation du problème continu, et correspond au fait que le champ de pression d'un écoulement incompressible n'est connu qu'à une constante arbitraire près.

Il faut donc a priori éliminer les 4 modes parasites, et comme nous les avons identifiés comme étant les 4 modes "coins", il suffit de les "sauter" (les coins!!) en faisant parcourir à la pression au bord la base canonique des valeurs d'une fonction définie sur le bord. On engendre ainsi une matrice d'ordre  $K - 4$  qui possède encore un noyau de dimension 1. Cette matrice peut être régularisée en remplaçant arbitrairement une des relations par une équation indépendante, ce qui revient à fixer la valeur de la pression de correction en un point de la frontière du domaine. Cette procédure permet d'obtenir un champ de pression qui garantisse que la divergence du champ de vitesse est effectivement nulle sur le bord du domaine de résolution. Cette divergence n'est cependant pas nulle en les points de Gauss-Lobatto intérieurs, ce qui peut paraître surprenant a priori. Cette erreur résiduelle est liée à la non-commutativité de l'opérateur de dérivée première avec l'opérateur de dérivée seconde muni de ses conditions aux limites (voir Haldenwang[8]) et L. Tuckerman a récemment proposé un algorithme très élégant pour résoudre ce problème [26].

Explicitons la résolution du problème de Stokes instationnaire dans le cas mono-dimensionnel par une méthode de collocation. On recherche une solution polynomiale  $(V, P)$  dans laquelle le champ de pression  $P$  appartient au même espace polynomial que le champ de vitesse  $V$ , et la vitesse et la pression sont connues par leurs valeurs aux points de Gauss-Lobatto. Sous forme discrétisée, le problème de Stokes 1D s'écrit donc:

$$S_{ij}U_j - \lambda\delta_{ij}U_j = F_{ij}P_j + s_i, i = 1, \dots, N-1 \quad (4.33)$$

$$U_0 = U_N = 0 \quad (4.34)$$

$$F_{ij}U_j = 0, i = 0, \dots, Ni \quad (4.35)$$

où  $S_{ij}$  et  $F_{ij}$  sont respectivement les opérateurs de dérivée seconde et première en approximation de collocation. Il est connu qu'en 1D, la dimension du noyau de ce système linéaire est de dimension 2, alors qu'elle est de dimension 8 en 2D comme nous le verrons plus loin. (Notons que nous parlons ici de la dimension du noyau du système linéaire tel qu'il est écrit ci-dessus, et non pas du système obtenu après la prise de la divergence de l'équation de quantité de mouvement).

Avant justement de prendre la divergence de l'équation de quantité de mouvement, on doit réécrire cette équation de quantité de mouvement en tous les points du domaine y compris les frontières puisqu'il est clair d'après (4.35), que la divergence discrète dans  $\mathbb{P}_N$  fait intervenir les valeurs de la vitesse  $U_i, i = 0, \dots, N$  en tous les points de collocation y compris les valeurs frontières. Ceci revient à définir deux résidus  $r_0$  et  $r_N$  tels que

$$S_{ij}U_j - \lambda\delta_{ij}U_j = F_{ij}P_j + s_i + r_i, i = 0, \dots, N \quad (4.36)$$

$$r_i = 0, 1 \leq i \leq N-1 \quad (4.37)$$

On prend la divergence discrète de l'équation de la quantité de mouvement en multipliant (4.36) par  $F_{ki}$  ce qui donne

$$F_{ki}S_{ij}U_j - \lambda F_{ki}\delta_{ij}U_j = F_{ki}F_{ij}P_j + F_{ki}s_i + F_{ki}r_i \quad (4.38)$$

Si on appelle  $D_l = F_{lj}U_j$  la divergence discrète aux points de Gauss-Lobatto, on veut donc avoir  $D_l = 0$  en tous les points de Gauss-Lobatto  $x_l, l = 0, \dots, N$ . Si  $F$  et  $S$  commutent<sup>1</sup>, ce qui est le cas si ces opérateurs sont exacts dans  $\mathbb{P}_N$  on peut écrire (4.38) comme:

$$S_{ki}F_{ij}U_j - \lambda F_{ki}\delta_{ij}U_j = F_{ki}F_{ij}P_j + F_{ki}s_i + F_{ki}r_i \quad (4.39)$$

ou encore

$$S_{ki}D_i - \lambda D_k = F_{ki}F_{ij}P_j + F_{ki}s_i + F_{ki}r_i \quad (4.40)$$

L'équation de Poisson discrète pour la pression s'écrit alors

$$0 = F_{ki}F_{ij}P_j + F_{ki}s_i + F_{ki}r_i, k = 1, \dots, N-1 \quad (4.41)$$

Résoudre le problème de Stokes est donc équivalent à trouver  $2N+4$  inconnues ( $U = (U_0, \dots, U_N)$ ,  $P = (P_0, \dots, P_N)$ ,  $r_0, r_N$ ) telles que

$$S_{ij}U_j - \lambda\delta_{ij}U_j - F_{ij}P_j = s_i, i = 1, \dots, N-1 \quad (4.42)$$

$$U_0 = U_N = 0 \quad (4.43)$$

$$S_{kj}P_j + F_{k0}r_0 + F_{kN}r_N = -F_{kj}s_j, k = 1, \dots, N-1 \quad (4.44)$$

$$D_0 = D_N = 0 \quad (4.45)$$

$$S_{ij}U_j - \lambda\delta_{ij}U_j - F_{ij}P_j - r_i = s_i, i = 0, N \quad (4.46)$$

---

1. Si  $F$  et  $S$  ne commutent pas, ce qui est le cas par exemple s'ils résultent d'un changement de coordonnées qui n'est pas polynomial de degré inférieur à 2, il est alors clair que la seule façon d'écrire le membre de gauche de (4.38) comme la formulation discrète de  $(\frac{\partial^2}{\partial z^2} - \lambda I)\nabla \cdot V$  est de définir  $S$  comme étant le carré de  $F$ , ie  $S_{ij} = F_{il}F_{lj}$ , la commutation étant alors obtenue par associativité.

où il est donc entendu que  $S_{ij} = F_{il}F_{lj}$ . Ce système est résolu à l'aide d'un système auxiliaire dans lequel les valeurs frontières sur la divergence sont remplacées par des conditions de type Dirichlet sur la pression ce qui permet alors de résoudre des problèmes découplés pour la pression et la vitesse. Ceci revient à utiliser la fameuse formule de Shermann-Morrison-Woodbury décrite dans [26] (voir annexe A).

En collocation-2D, les valeurs frontières de la pression sont celles correspondant aux points de collocation sur la frontière de  $\partial\Omega$  excepté les 4 coins comme nous l'avons déjà expliqué. De même les résidus frontières font intervenir deux ensembles de valeurs, celles pour la composante  $u$  le long des cotés  $x = 1$  et  $x = -1$  et celles pour la composante  $w$  le long des cotés  $z = 1$  and  $z = -1$ , également sans les valeurs aux coins. La matrice de capacitance est donc d'ordre  $4(N + M - 2)$ . Un raffinement supplémentaire consiste à tirer parti des symétries pour mettre cette matrice, a priori pleine, sous une forme bloc-diagonale, de 4 blocs respectivement libellés pair-pair, pair-impair, impair-pair et impair-impair, d'ordre respectif  $N + M$ ,  $N + M - 2$ ,  $N + M - 2$  et  $N + M - 4$  (en ayant supposé  $N$  et  $M$  pairs). Le bloc pair-pair a un noyau de dimension 4, mais lorsque l'on considère des conditions aux limites homogènes, les conditions de compatibilité sont satisfaites et la résolution du système linéaire peut être obtenue avec une décomposition en valeurs singulières. Tous ces ingrédients permettent d'obtenir une divergence dans  $\mathbb{P}_N$  au zéro machine.

Une procédure similaire peut être également mise en œuvre lorsque les problèmes de Helmholtz sont résolus dans le formalisme tau (voir [26]).

## 4.3 Opérateur d'Uzawa

### 4.3.1 Principe

Comme il a déjà été dit, l'algorithme d'Uzawa repose sur l'élimination de la vitesse dans le problème de Stokes instationnaire.

En notant  $\mathcal{H}\mathcal{U}$  et  $\mathcal{H}\mathcal{V}$  les opérateurs d'Helmholtz pour chacune des composantes  $u$  et  $w$  et en laissant tomber l'indice de la discrétisation temporelle, le problème de Stokes instationnaire s'écrit:

$$\mathcal{H}\mathcal{U}u = \frac{\partial P}{\partial x} - Su \quad (4.47)$$

$$\mathcal{H}\mathcal{V}w = \frac{\partial P}{\partial z} - Sw \quad (4.48)$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0 \quad (4.49)$$

complété par les conditions aux limites pour  $u$  et  $w$ .



On peut inverser formellement (4.47) et (4.48) pour obtenir

$$u^{n+1} = \mathcal{H}\mathcal{U}^{-1} \frac{\partial P}{\partial x} - \mathcal{H}\mathcal{U}^{-1} Su \quad (4.50)$$

$$w^{n+1} = \mathcal{H}\mathcal{W}^{-1} \frac{\partial P}{\partial z} - \mathcal{H}\mathcal{W}^{-1} Sw \quad (4.51)$$

et l'imposition de la contrainte d'incompressibilité fournit donc une équation pour la pression qui s'écrit formellement

$$\left( \frac{\partial}{\partial x} \mathcal{H}\mathcal{U}^{-1} \frac{\partial}{\partial x} + \frac{\partial}{\partial z} \mathcal{H}\mathcal{W}^{-1} \frac{\partial}{\partial z} \right) P = \left( \frac{\partial}{\partial x} \mathcal{H}\mathcal{U}^{-1} Su + \frac{\partial}{\partial z} \mathcal{H}\mathcal{W}^{-1} Sw \right) \quad (4.52)$$

et nous noterons  $\mathcal{U} = \frac{\partial}{\partial x} \mathcal{H}\mathcal{U}^{-1} \frac{\partial}{\partial x} + \frac{\partial}{\partial z} \mathcal{H}\mathcal{W}^{-1} \frac{\partial}{\partial z}$ . On constate donc que l'obtention de la solution au problème de Stokes instationnaire se ramène à l'inversion de cette équation pour la pression, puisqu'une fois la pression obtenue il suffit de calculer son gradient et de résoudre (4.47,4.48), et le champ de vitesse obtenu vérifie bien la contrainte d'incompressibilité.

L'inversion de ce système linéaire suppose implicitement que l'on ait fait le choix d'une discrétisation spatiale, choix que nous allons exposer maintenant.

### 4.3.2 Choix des espaces d'approximation

#### Méthodes à 1 grille

Comme nous l'avons expliqué au paragraphe précédent (4.1), dans la première génération de méthodes spectrales, les deux composantes de vitesse et de pression étaient prises dans le même espace de polynômes, obtenu par tensorisation de bases mono-dimensionnelles dans la direction  $x$  et  $z$  respectivement. En notant  $N$  (resp.  $M$ ) le degré maximal des polynômes dans la direction  $x$  (resp.  $z$ ), les composantes  $u$  et  $w$  et la pression  $P$  appartenaient donc à l'espace  $P_N(x) \otimes P_M(z)$ . De façon équivalente,  $u$ ,  $w$  et  $P$  étaient considérés connus à travers leurs valeurs aux points de collocation  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$  où  $\mathcal{GL}_x^N = \{x_i, x_i = \cos(\frac{i\pi}{N}), i = 0, \dots, N\}$  sont les  $N+1$  points de Gauss-Lobatto (racines de  $(1-x^2)T'_N(x)$ ) dans la direction  $x$ . Dans l'expression  $\frac{\partial}{\partial x} \mathcal{H}\mathcal{U}^{-1} \frac{\partial}{\partial x}$ , les deux  $\frac{\partial}{\partial x}$  correspondent donc à l'évaluation aux points de Gauss-Lobatto de la dérivée du polynôme d'interpolation interpolant une fonction connue par ses valeurs aux points de collocation Gauss-Lobatto.

Il est donc possible d'obtenir une représentation matricielle de la matrice associée à l'opérateur  $\mathcal{U}$  en faisant décrire à  $P$  la base canonique correspondante, ce qui permet ainsi de construire la matrice  $\mathcal{U}$  d'ordre  $K = (N+1) \times (M+1)$ . L'inversion de cette matrice n'est pas possible car on peut constater numériquement que son rang vaut  $K-8$  et les 8 modes constituant une base de  $\text{Ker}(\mathcal{U})$  sont appelés modes parasites de pression. Lorsque les opérateurs correspondant aux problèmes de Helmholtz sont résolus par une méthode de collocation (seul cas envisagé ici), il est facile d'exhiber les modes parasites, qui sont:

- le mode constant  $T_0(x)T_0(z)$

- les 3 modes  $T_N(x)T_0(z)$ ,  $T_0(x)T_M(z)$ ,  $T_N(x)T_M(z)$  qui produisent un gradient de pression identiquement nul aux points de collocation interne  $(x_i, z_j)$ ,  $1 \leq i \leq N-1$ ,  $1 \leq j \leq M-1$
- les 4 modes ”coin” (déjà rencontrés) où la pression vaut 1 en 1 coin et 0 en tous les autres points. Le mode coin correspondant au coin de coordonnées (1,1) s’écrit  $\frac{(1+x)T'_N(x)(1+z)T'_M(z)}{4N^2M^2}$ .

Il est de nouveau à noter que le mode constant  $T_0(x)T_0(z)$  n’est pas à proprement parler un mode parasite puisque, pour un écoulement incompressible, la pression est toujours définie à une constante additive près.<sup>2</sup> Morchoisne [16] semble être le premier à avoir identifié ces modes parasites. L’origine du problème tient heuristiquement à deux raisons: d’une part au fait que la pression et les composantes de vitesse soient choisies dans le même espace polynomial et d’autre part au fait que la pression soit définie et la condition d’incompressibilité imposée en des points de collocation situés sur la frontière du domaine de résolution. Sur un plan plus théorique, l’origine du problème tient au fait que la condition *inf* – *sup* n’est pas vérifiée (voir cours J.L. Guermond).

Ces raisons une fois identifiées conduisent donc “naturellement” aux méthodes de grilles décalées, qui reposent sur deux ingrédients essentiels: abaisser la dimension de l’espace polynomial pour la pression par rapport aux composantes de vitesse d’une part, définir la pression et vérifier la contrainte de divergence nulle en des points intérieurs au domaine de résolution d’autre part.

### Méthode à 2 grilles $P_{N-1}(x) \otimes P_{M-1}(z)$

La première méthode de grilles décalées fut proposée par Montigny-Rannou et Morchoisne [20]. Les deux composantes de vitesse  $u$  et  $w$  sont prises dans l’espace  $P_N(x) \otimes P_M(z)$  et sont définies aux points  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$  comme précédemment. La pression est prise dans l’espace  $P_{N-1}(x) \otimes P_{M-1}(z)$  et la deuxième raison évoquée ci-dessus conduit à la définir aux points  $\mathcal{G}_x^N \otimes \mathcal{G}_z^M$  où  $\mathcal{G}_x^N$  sont les  $N$  points de Gauss (racines de  $T_N(x)$ ),  $\mathcal{G}_x^N = \left\{ x_i, x_i = \cos\left(\frac{(2i-1)\Pi}{2N}\right), i = 1, \dots, N \right\}$ . La condition d’incompressibilité est également imposée en  $\mathcal{G}_x^N \otimes \mathcal{G}_z^M$ . Il est à noter que dans l’expression  $\frac{\partial}{\partial x} \mathcal{H} \mathcal{U}^{-1} \frac{\partial}{\partial x}$ , les deux  $\frac{\partial}{\partial x}$  n’ont plus la même signification puisque celui situé le plus à droite correspond à l’évaluation aux points de Gauss-Lobatto de la dérivée du polynôme d’interpolation interpolant la pression connue aux points de Gauss alors que celui de gauche correspond à l’évaluation aux points de Gauss de la dérivée du polynôme d’interpolation interpolant une fonction connue aux points de Gauss-Lobatto. La situation est en fait pire en 2-D car il faut faire en plus une projection dans l’autre direction pour obtenir le gradient de la pression aux points  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$ .

Ceci permet donc de construire la matrice représentative de  $\mathcal{U}$  rapporté aux bases canoniques, matrice d’ordre  $K = N \times M$ . Un examen numérique montre que le rang de cette matrice est  $K - 2$ , et il subsiste donc 1 mode parasite en plus du mode constant. Ce mode parasite est  $T'_N(x)T'_M(z)$  qui appartient bien à  $P_{N-1}(x) \otimes P_{M-1}(z)$  et dont le gradient  $(\nabla P)^t =$

---

2. On constatera au passage que le fait d’avoir introduit une équation de Poisson pour la pression comme au paragraphe 4.1 a fait diminuer la dimension du noyau du problème et donc le nombre de degrés de liberté indéterminés sur le champ de pression.

$(T_N''(x)T_M'(z), T_N'(x)T_M''(z))$  est identiquement nul aux points de Gauss-Lobatto intérieurs  $(x_i, z_j), 1 \leq i \leq N-1, 1 \leq j \leq M-1$ . Cette méthode, bien qu'ayant permis d'abaisser grandement la taille du noyau de  $\mathcal{U}$  de 8 à 2 n'est donc pas complètement satisfaisante.

### Algorithme de grilles décalées à 3 grilles

Le second algorithme de grilles décalées, proposé par Bernardi et Maday [18], consiste à prendre les deux composantes de vitesse et la pression dans 3 espaces polynomiaux différents et donc à définir ces variables sur 3 maillages différents. Les équations de quantité de mouvement pour chacune des composantes de vitesse sont résolues par une méthode de collocation consistant à annuler le résidu en tous les points où est définie la variable correspondante, l'équation de continuité étant vérifiée aux points où la pression est définie. La pression est prise dans l'espace  $P_{N-1}(x) \otimes P_{M-1}(z)$  et l'ensemble des points où elle est définie et où la condition d'incompressibilité est écrite est  $\mathcal{G}_x^N \otimes \mathcal{G}_z^M$  où  $\mathcal{G}_x^N = \{x_i, x_i = \cos(\frac{(2i-1)\Pi}{2N}), i = 1, \dots, N\}$  sont les N points de Gauss (racines de  $T_N(x)$ ) dans la direction x. La composante de vitesse u est prise dans l'espace  $P_N(x) \otimes P_{M-1}(z)$  et est définie aux points  $\mathcal{GL}_x^N \otimes \mathcal{G}_z^M$ . Symétriquement, la composante de vitesse w est prise dans l'espace  $P_{N-1}(x) \otimes P_M(z)$  et est définie aux points  $\mathcal{G}_x^N \otimes \mathcal{GL}_z^M$ . Il est alors manifeste que la composante u n'est pas définie aux frontières  $z = 0$  et 1, de même que la composante w n'est pas définie aux frontières  $x = 0$  et 1.

De façon à pouvoir imposer explicitement les conditions aux limites sur les composantes de vitesse, nous avons été amenés à introduire en plus de  $\mathcal{GL}_x^N$  et  $\mathcal{G}_x^N$ , l'ensemble des points de collocation noté,  $\mathcal{GA}_x^N = \mathcal{G}_x^N \cup \{-1, +1\}$ .

L'algorithme de grilles décalées que nous avons mis en œuvre consiste alors à prendre la composante de vitesse u(x,z) dans  $P_N(x) \otimes P_{M+1}(z)$  et à la définir aux points  $\mathcal{GL}_x^N \otimes \mathcal{GA}_z^M$ , la composante de vitesse w(x,z) dans  $P_{N+1}(x) \otimes P_M(z)$  et à la définir aux points  $\mathcal{GA}_x^N \otimes \mathcal{GL}_z^M$ , et la pression P(x,z) dans  $P_{N-1}(x) \otimes P_{M-1}(z)$  et à la définir aux points  $\mathcal{G}_x^N \otimes \mathcal{G}_z^M$ . En outre, on prendra la température dans  $P_N(x) \otimes P_M(z)$  et on la définira aux points  $\mathcal{GL}_x^N \otimes \mathcal{GL}_z^M$ .

Les équations de quantité de mouvement pour u et w, l'équation de continuité et l'équation de température seront respectivement satisfaites aux points où sont définies les variables u, w et P et la température, ce qui fait de cette méthode dans le cas d'un écoulement non-isotherme une méthode à 4 grilles différentes.

La résolution de (4.47) et (4.48) sera effectuée par une méthode de collocation consistant à annuler le résidu de l'équation aux dérivées partielles en tous les points de collocation intérieurs au domaine de résolution, les conditions aux limites étant imposées de façon explicite (forte) aux points situés sur la frontière du domaine.

Notons que dans (4.47) par exemple, la pression et la vitesse u n'étant pas définies aux mêmes points, la dérivation partielle par rapport à x doit être comprise dans le sens suivant: la pression étant définie aux points  $\mathcal{G}_x^N$ , on calcule le polynôme d'interpolation de Lagrange qui interpole P en ces points; on dérive ensuite ce polynôme que l'on réévalue aux points de Gauss-Lobatto  $\mathcal{GL}_x^N$  où l'équation pour u doit être vérifiée.

Cette méthode permet effectivement d'obtenir que le noyau de  $\mathcal{U}$  soit réduit au champ de pression constant. Elle est cependant (on l'aura constaté!!) très lourde à mettre en œuvre, en particulier si l'on souhaite l'étendre à 3 dimensions, ou dans le cadre de la décomposition de domaines.

### Méthode à 2 grilles $P_{N-2}(x) \otimes P_{M-2}(z)$

Ce que Montigny-Rannou et Morchoisne n'avaient pu obtenir (élimination totale des modes parasites de pression) en prenant la pression dans  $P_{N-1}(x) \otimes P_{M-1}(z)$  peut être obtenu en réduisant encore la dimension de l'espace de pression, que l'on choisit alors comme étant  $P_{N-2}(x) \otimes P_{M-2}(z)$ . Cette nouvelle méthode de grilles décalées à deux grilles, une pour les composantes de la vitesse et une pour la pression, est à la base des méthodes d'éléments spectraux développées au MIT autour d'A. Patera et d'Y. Maday [21], [22]. Cette méthode est celle qui fait autorité en ce moment dans le domaine de l'approche du traitement de l'incompressibilité par l'opérateur d'Uzawa. Deux variantes sont disponibles, selon le choix des points de collocation où l'on définit la pression et où l'on vérifie la divergence.

Il a été initialement proposé de choisir ces points comme les  $(N-1) \times (M-1)$  points de Gauss, racines de  $T_{N-1}(x) \times T_{M-1}(z)$ . Ce choix nécessite 4 multiplications matricielles pour évaluer le gradient du champ de pression sur les points de Gauss-Lobatto pour calculer les termes source des équations d'Helmholtz des composantes de vitesse, et également 4 multiplications matricielles pour calculer la divergence.

Ce nombre peut être réduit à 2, si on définit la pression et si on calcule la divergence aux points de Gauss-Lobatto internes

Ces deux variantes ne diffèrent donc que par les opérateurs discrets utilisés pour obtenir le gradient du champ de pression et pour le calcul de la divergence du champ de vitesse.

### 4.3.3 Mise en œuvre

Les difficultés à résoudre pour intégrer le système d'équations (4.47, 4.48, 4.7, 4.8, 4.52) sont les suivantes:

- inversion de  $\mathcal{H}\mathcal{U}f = s$
- choix de l'algorithme à utiliser pour résoudre l'équation de pression (4.52)

### Résolution du problème de Helmholtz

La résolution de  $\mathcal{H}\mathcal{U}f = s$  est effectuée par une méthode directe de bi-diagonalisation inspirée de l'algorithme proposé par Haidvogel et Zang [14].

L'algorithme de bidiagonalisation consiste à rechercher la solution dans la base tensorisée des fonctions propres correspondant à chacun des opérateurs  $\frac{\partial^2}{\partial x^2}$  et  $\frac{\partial^2}{\partial z^2}$ . Pour ma méthode à grilles présentée au paragraphe 4.3.2.3, il convient donc de rechercher les vecteurs propres et valeurs

propres des opérateurs de dérivation seconde sur les points de Gauss-Lobatto  $\mathcal{GL}_x^N$  et les points de Gauss-Augmentés  $\mathcal{GA}_x^N$ . On trouvera dans l'annexe B l'expression des matrices correspondantes ainsi que celle d'autres opérateurs nécessaires à la mise en œuvre de l'algorithme.

L'algorithme de résolution consiste alors à exprimer le second membre dans la base propre ce qui demande 2 multiplications matricielles. Dans la base tensorisée des vecteurs propres, la résolution se ramène alors à une simple division scalaire. Il faut ensuite revenir dans la base canonique ce qui demande à nouveau 2 multiplications matricielles. Le coût total de l'inversion est d'ordre  $4NM(N+M)$ .

### Résolution de l'opérateur d'Uzawa pour la pression

En ce qui concerne la résolution de (4.52), il peut sembler déraisonnable a priori de chercher à résoudre cette équation par une méthode directe, et on est donc conduit à envisager une résolution itérative.

Il convient en particulier d'utiliser des algorithmes itératifs ne demandant que l'évaluation de  $\mathcal{U}x^k$  où  $x^k$  est un résidu ou une estimation de la solution à l'itération  $k$ . Notons que le nombre d'opérations demandé par cette évaluation est  $12NM(N+M)$  si on fait appel à l'algorithme de bi-diagonalisation présenté au paragraphe précédent pour résoudre  $\mathcal{H}\mathcal{U}f = s$  et  $\mathcal{H}\mathcal{V}f = s$ .

Nous avons testé 3 algorithmes itératifs différents pour la résolution de (4.52):

- une méthode itérative de Richardson qui s'écrit:

$$p^{k+1} = p^k - \alpha(\mathcal{U}p^k - s) \quad (4.53)$$

$\alpha$  étant déterminé de manière à obtenir la convergence la plus rapide possible. Si l'opérateur  $\mathcal{U}$  a ses valeurs propres réelles comprises entre  $\mu_{max}$  et  $\mu_{min}$ , il est connu que la valeur optimale de  $\alpha$  est  $\frac{2}{\mu_{max} + \mu_{min}}$ .

- une méthode de Richardson avec résidu minimum où  $\alpha_k$  est choisi à chaque itération de manière à minimiser le résidu à l'itération  $k+1$
- une méthode de gradient conjugué

Les figures 1 et 2 montrent les performances relatives de ces trois algorithmes itératifs pour la résolution du problème de Stokes stationnaire ( $\lambda = 0$ ) pour deux résolutions différentes  $N=M=16$  et  $N=M=32$ . Deux constatations s'imposent: d'une part l'algorithme de gradient conjugué converge environ deux fois plus vite que les deux autres algorithmes, qui présentent des performances tout à fait comparables. Ces résultats sont tout à fait conformes aux vitesses de convergence attendues, qui sont respectivement égales à  $\frac{K-1}{K+1}$  ou  $\frac{K^{1/2}-1}{K^{1/2}+1}$  pour les méthodes du premier et second ordre, où  $K$  est le conditionnement de l'opérateur. D'autre part et surtout, on constate que la vitesse de convergence dépend très peu de la résolution spatiale: la précision machine sur le résidu est obtenue en un peu plus de 50 itérations pour l'algorithme de gradient conjugué et ce pour les deux résolutions considérées.

Ces bonnes performances tiennent à la nature du spectre de  $\mathcal{U}$ . Nous avons examiné numériquement les valeurs propres de l'opérateur  $\mathcal{U}$  mono-dimensionnel (ie  $\frac{\partial}{\partial x}\mathcal{H}\mathcal{U}^{-1}\frac{\partial}{\partial x}$ ). Lorsque la constante  $\lambda$

du problème de Helmholtz est nulle, le spectre de cet opérateur n'est formé que de 1 (à part la valeur propre nulle correspondant au mode propre constant).

Nous avons également examiné le spectre de l'opérateur bidimensionnel  $\mathcal{U} = \frac{\partial}{\partial x} \mathcal{H} \mathcal{U}^{-1} \frac{\partial}{\partial x} + \frac{\partial}{\partial z} \mathcal{H} \mathcal{V}^{-1} \frac{\partial}{\partial z}$ . Cette recherche a été faite numériquement à l'aide de programmes de recherche de valeurs propres (F02AGF de la NAGLIB). Dans l'ensemble des cas considérés, ces valeurs propres ont été trouvées réelles (aux erreurs d'arrondi près). Les valeurs propres minimale et maximale de l'opérateur  $\mathcal{U}$  sont reportées dans les tableaux 1 et 2 pour des approximations spatiales de type Tchebychev ou Legendre. Ces valeurs montrent que, pour  $\lambda = 0$ , le conditionnement de  $\mathcal{U}$  est effectivement pratiquement indépendant de la résolution, ce qui est donc cohérent avec le comportement des méthodes itératives obtenu pour les deux résolutions différentes. Les valeurs montrent de plus que la valeur de  $\alpha_{opt}$  est très voisine de 1.7, ce qui est effectivement obtenu numériquement. Le rayon spectral correspondant est donc voisin de 0.7. On constate par contre que le conditionnement de l'opérateur  $\mathcal{U}$  se détériore lorsque  $\lambda$  augmente. Ceci constitue en fait une limitation très importante à l'emploi de ces techniques itératives, puisque le traitement explicite des termes convectifs résulte en un critère de stabilité de type  $\Delta t \simeq N^{-2}$ . Il faut donc obligatoirement prendre de petits pas de temps pour intégrer le système dans le temps (4.4, 4.5, 4.6), et alors, en raison du mauvais conditionnement de l'opérateur  $\mathcal{U}$ , la résolution itérative de (4.52) à l'aide d'un des algorithmes présentés ci-dessus devient inefficace. Ces méthodes itératives perdant leur efficacité lorsque  $\lambda$  augmente, il faut alors nécessairement introduire un préconditionnement tel que celui proposé dans [23] pour une discrétisation de type éléments finis. La détermination d'un tel préconditionneur reste encore un problème ouvert dans le cas d'une discrétisation spatiale de type spectral, et il convient donc de proposer d'autres techniques plus efficaces. On peut en particulier se poser la question des limites d'applicabilité d'une méthode directe pour l'inversion de (4.52).

- résolution directe

L'inversion directe de (4.52) demande que dans un premier temps on ait explicité la matrice correspondant à l'opérateur  $\mathcal{U} = (\frac{\partial}{\partial x} \mathcal{H} \mathcal{U}^{-1} \frac{\partial}{\partial x} + \frac{\partial}{\partial z} \mathcal{H} \mathcal{V}^{-1} \frac{\partial}{\partial z})$ . Comme nous l'avons déjà mentionné, ceci peut être fait en faisant décrire à la pression la base canonique de  $\mathcal{R}^{N.M}$ <sup>3</sup> et en évaluant  $(\frac{\partial}{\partial x} \mathcal{H} \mathcal{U}^{-1} \frac{\partial}{\partial x} + \frac{\partial}{\partial z} \mathcal{H} \mathcal{V}^{-1} \frac{\partial}{\partial z}) P_{ij}$ , ce qui permet ainsi de construire la matrice d'ordre  $N \times M$  de l'opérateur  $\mathcal{U}$  rapporté aux bases canoniques. Le fait que l'algorithme de grilles décalées élimine tous les modes parasites (sauf le mode constant) montre que cette matrice est de rang  $N \times M - 1$ . (La somme de toutes les colonnes de cette matrice est le vecteur nul puisqu'elle correspond à un champ de pression unité). Il convient donc de la régulariser ce qui revient à fixer la valeur de la pression en un point de collocation du domaine de résolution.

On peut envisager deux variantes pour la résolution de (4.52). L'une directe, consiste à calculer les termes sources  $(S_u, S_w)$  et à inverser le système linéaire (4.52) pour obtenir la pression. Une fois la pression obtenue, on peut alors calculer son gradient, l'ajouter aux termes source et

---

3. Nous nous sommes placés ici dans le cas de la méthode à 3 grilles présentée au 4.3.2.3. Tout ce qui suit vaut également pour la méthode à 2 grilles  $P_{N-2}(x) \otimes P_{M-2}(z)$  présentée au 4.3.2.4

inverser (4.47,4.48), ce qui fournit donc le champ de vitesse recherché.

Une autre méthode, qui est généralement utilisée dans le cas d'un calcul instationnaire consiste à se donner une estimation initiale de la pression  $p_e$  qui dans le cas d'un calcul instationnaire peut être simplement la pression du pas de temps précédent ou plus généralement une combinaison linéaire des champs de pression des pas de temps précédents cohérente avec la précision temporelle du schéma. Ceci permet donc de résoudre une première fois (4.47) et (4.48) avec le champ de pression  $p_e$ . Le champ de vitesse obtenu  $(u_e, w_e)$  ne vérifie pas en général l'équation de continuité et il lui correspond un champ de divergence  $\mathcal{D}_e$ . Il convient donc de déterminer une correction de pression  $p_c$  qui produise un champ de vitesse de divergence  $-\mathcal{D}_e$ , ce qui revient à inverser  $\mathcal{U}p_c = -\mathcal{D}_e$ .

Les deux techniques sont évidemment strictement équivalentes puisque la première variante correspond au choix  $p_e = 0$ . Il y a un léger avantage à la première variante du point de vue temps de calcul puisqu'il n'y a pas à calculer le gradient de  $p_e$ . La deuxième variante semble par contre légèrement favorable du point de vue de la propagation des erreurs d'arrondi, la correction de pression  $p_c$  tendant vers zéro si l'on approche d'une solution stationnaire.

Les problèmes inhérents à l'utilisation de cette méthode directe sont liés à l'inversibilité de la matrice  $\mathcal{U}$  puisque son ordre croît rapidement en fonction de  $N$  et  $M$  (1024 si  $N = M = 32, \dots$ ). Les tests que nous avons effectués, en particulier pour calculer les solutions des problèmes test du GAMM-Workshop (voir [24]), ont montré que cette technique pouvait s'utiliser jusqu'à des valeurs importantes de  $N$  et  $M$  ( $N = 24, M = 64$  par exemple) sans difficulté apparente autre que celle liée au stockage en mémoire de la matrice (de l'ordre de 1.5 Mmots ou 12 Moctets dans le cas  $(N = 24, M = 64)$ ). Ici encore on peut utiliser les symétries du domaine de résolution pour mettre cette matrice (pleine dans sa forme originale) sous forme bloc-diagonale. En 2D, pour un domaine rectangulaire, il existe deux axes de symétrie et on peut donc mettre la matrice sous la forme de 4 blocs d'ordre  $N \times M/4$  environ (les ordres exacts dépendent des parités de  $N$  et  $M$ ), ce qui permet de réduire le stockage nécessaire des  $3/4$  par rapport à la matrice originale. En 3D il y a 3 plans de symétrie, et la matrice peut se mettre sous la forme de 8 blocs diagonaux, ce qui permet de n'en stocker qu' $1/8$  ème.

# Bibliographie

- [1] C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag, 1988
- [2] J.B. Mac Laughlin et S.A. Orszag, Transition from periodic to chaotic thermal convection, *J. Fluid Mech.*, 1982, 122, p. 125-142
- [3] J.H. Curry, J.R. Herring, J. Loncaric et S.A. Orszag, Order and disorder in two- and three-dimensional Benard convection, *J. Fluid Mech.*, 1984, 147, p. 1-38
- [4] Morchoisne Y., Résolution des équations de Navier-Stokes par une méthode pseudo-spectrale en espace-temps, *La Rech. Aéro.*, 1979, 5, p. 293-306
- [5] Morchoisne Y., Pseudo-spectral methods for homogeneous or inhomogeneous flow, 3rd Turbulent Shear Flow, Davis, 1981, chapt. 19, p. 1-5
- [6] Le Quéré P. et Alziary de Roquefort T., Sur une méthode spectrale semi-implicite pour la résolution des équations de Navier-Stokes d'un écoulement bidimensionnel visqueux incompressible, *C. R. Acad. Sci. Paris*, 1982, t. 294, série II, pp. 941-944
- [7] Patera A.T., A spectral element method for fluid dynamics: Laminar Flow in a channel expansion, *J. Comp. Phys.*, 1984, 54, p. 468-488
- [8] Haldenwang P., Thèse d'Etat, Université de Provence, Décembre 1984
- [9] Le Quéré P. et Alziary de Roquefort T., Computation of natural convection in two-dimensional cavities with Chebyshev polynomials, *J. Comp. Phys.*, 1985, 57, 210-228
- [10] Ku H.C., Taylor T.D. et Hirsh R.S., Pseudospectral methods for the solution of the incompressible Navier-Stokes equations, *Comp. Fluids*, 1987, 15, p. 195-214
- [11] Vanel J.M., Peyret R. et Bontoux P., A pseudo-spectral solution of vorticity – stream-function equations using the influence matrix technique, *Num. Meth. Fluid Dyn. II*, Clarendon Press, 1986, pp. 463-475
- [12] Ehrenstein U. et Peyret R., A Chebyshev collocation method for the Navier-Stokes equations with application to double-diffusive convection, *Int. J. Num. Meth. Fluids*, 1989, 9, p. 427-452
- [13] Gottlieb D. et Lustman L., The spectrum of the Chebyshev collocation operator for the heat equation, *SIAM J. Num. Anal.*, 1983, 20, 909-921
- [14] Haidvogel D. et Zang T., The accurate solution of Poisson's equations by expansion in Chebyshev polynomials, *J. Comp. Phys.*, 1979, 30, pp. 167-180
- [15] Kleiser L. et Schumann U., Treatment of incompressibility and boundary conditions in three-dimensional numerical spectral simulations of plane channel flows, *Notes on Num. Fluid Mech.*, 1980, Vieweg, 2, pp. 165-173
- [16] Morchoisne Y., Résolution des équations de Navier-Stokes par une méthode spectrale de sous-domaines, 3<sup>ème</sup> Cong. Int. Meth. Num. Ing., Paris, 1983, Ed P. Lascaux, p. 275-281
- [17] Metivet B., Résolution spectrale des équations de Navier-Stokes par une méthode de sous-domaines courbes, Thèse d'Etat, Univ. Paris VI, Juin 1987



- [18] C. Bernardi et Y. Maday, A collocation method over staggered grids for the Stokes problem, *Int. J. Num. Meth. Fluids*, 1988, 8, p. 537-557
- [19] De Vahl Davis G. et Jones I.P., Natural convection in a square cavity: a comparison exercise, *Int. J. Num. Meth. Fluids*, 1983, 3, 227-248
- [20] F. Montigny-Rannou et Y. Morchoisne, A spectral method with staggered grid for incompressible Navier-Stokes equations, *Int. J. Num. Meth. Fluids*, 1987, 7, p. 175-189
- [21] Y. Maday, A.T. Patera et E.M Ronquist, A well-posed optimal spectral element approximation for the Stokes problem, *SIAM J. Num. Anal.*, à paraître
- [22] Y. Maday, D. Meiron, A.T. Patera et E.M Ronquist, Analysis of iterative methods for the steady and unsteady Stokes problem: application to spectral element discretization, *SIAM J. Stat.*, à paraître
- [23] J. Cahouet et J.P. Chabard, Some fast 3D finite element solvers for the generalized Stokes problem, *Int. J. Num. Meth. Fluids*, 1988, 8, p. 869-895
- [24] Numerical Simulation of Oscillatory Convection in low-Pr fluids, *Notes on Numerical Fluid mechanics*, vol 27, Ed. par B. Roux, Vieweg, 1990
- [25] P. Le Quéré, Contribution to the GAMM-Workshop with a pseudo-spectral Chebyshev algorithm on a staggered grid, *op. cit.*, p. 227-236
- [26] Tuckermann L., Divergence-free velocity field in non-periodic geometries, *J. Comp. Phys.*, 1989, 80, pp. 403-441
- [27] Press, Teutolsky, Vetterling, Flannery, *Numerical Recipes: The Art of Scientific Computing*, CUP 1992
- [28] Batchelor, *An Introduction to Fluid Dynamics*, CUP 1967.

### Annexe A

Cette annexe est consacrée à l'exposé de la formule de Sherman-Morrison-Woodbury. Supposons que l'on sache résoudre de façon efficace un système linéaire

$$Hu = s \text{ dans } \mathcal{R}^N \quad (4.54)$$

par une méthode directe par exemple, ou parce que les équations prennent une forme par bloc, dans laquelle un certain nombre d'inconnues se retrouvent découplées des autres. On considère un système "voisin" du précédent,

$$\tilde{H}u = s \text{ dans } \mathcal{R}^N \quad (4.55)$$

dans lequel par exemple, un petit nombre d'équations ont été modifiées, mais ce qui prive de la possibilité de résoudre ce système de façon aussi efficace que le système initial. Peut-on utiliser le fait que l'on sait résoudre de façon efficace le premier système pour résoudre le second?

La formule de Sherman-Morrison-Woodbury permet de répondre positivement à cette question, si  $\tilde{H}$  peut se mettre sous la forme

$$\tilde{H} = H + UV^t$$

Soit donc à résoudre

$$(H + UV^t)u = s \quad (4.56)$$

En définissant une inconnue auxiliaire  $\sigma$ ,  $\sigma = V^t u$ , on peut mettre alors le système sous la forme par blocs:

$$\begin{bmatrix} H & U \\ V^t & -I \end{bmatrix} \begin{bmatrix} u \\ \sigma \end{bmatrix} = \begin{bmatrix} s \\ 0 \end{bmatrix}$$

De  $Hu + U\sigma = s$ , on tire  $u = H^{-1}(s - U\sigma)$ , que l'on reporte dans  $V^t u = \sigma$  pour obtenir l'équation donnant  $\sigma$

$$(I + V^t H^{-1} U)\sigma = V^t H^{-1} s$$

En éliminant finalement  $\sigma$  entre cette équation et  $Hu + U\sigma = s$ , il vient

$$Hu = s - U(I + V^t H^{-1} U)^{-1} V^t H^{-1} s$$

La matrice  $(I + V^t H^{-1} U)$  est la matrice qui est souvent appelée matrice d'influence ou de capacitance. Cette procédure n'a d'intérêt que si cette matrice est d'ordre  $K$  petit par rapport à l'ordre  $N$  de  $\tilde{H}$ . L'ordre de cette matrice correspond à nombre de composantes de  $\sigma$ . Cette matrice peut être construite en faisant décrire à  $\sigma_k$  la base canonique de  $\mathcal{R}^K$  et en évaluant  $(I + V^t H^{-1} U)\sigma_k$ , ce qui demande de résoudre  $K$  fois  $H^{-1}$ . On voit donc que le coût de cette procédure réside essentiellement dans la construction de la matrice, le coût de l'inversion étant faible si  $K$  est petit.

## Annexe B

La mise en œuvre de l'algorithme des grilles décalées demande la définition d'un certain nombre d'opérateurs que l'on peut classer par ordre différentiel croissant

- degré 0: il s'agit d'évaluer en certains points de collocation une fonction connue en d'autres points: typiquement le calcul des termes non-linéaires demande que l'on définisse des interpolations des points de Gauss-Lobatto vers les points de Gauss-Augmenté ou vice-versa.
- degré 1: il s'agit d'obtenir les valeurs en certains points de collocation de la dérivée du polyôme interpolant une fonction connue aux mêmes ou en d'autres points de collocation: par exemple, dérivation sur les points de Gauss-Lobatto ou de Gauss-Augmentés, dérivation de Gauss-Lobatto vers Gauss (pour calculer la divergence du champ de vitesse par exemple) ou vice-versa de Gauss vers Gauss-Lobatto pour calculer le gradient du champ de pression par exemple.
- degré 2: obtention de la dérivée seconde aux points de Gauss-Lobatto ou de Gauss-Augmentés d'une fonction connue en ces mêmes points. (Pour la dérivée seconde il n'y a pas à évaluer la dérivée seconde du polyôme d'interpolation en d'autres points que ceux où la fonction est connue).

De façon classique, l'interpolant polynomial de degré  $N$  d'une fonction  $u$  connue par ses valeurs  $u_i$  en  $N+1$  points de collocation  $(x_0, x_1, \dots, x_N)$  s'écrit  $I_N u = \sum_{i=0}^N u_i h_i(x)$  où  $h_i$  est le polyôme caractéristique d'ordre  $N$  associé au point  $x_i$  ( $h_i(x_j) = \delta_{i,j}$ ). On a  $h_i(x) = \frac{\Pi(x)}{(x - x_i)\Pi'(x_i)}$  avec  $\Pi(x) = \prod_{i=0}^N (x - x_i)$ .

Nous considèrerons les 3 ensembles de points de collocation suivants:

- $\mathcal{G}_x^N = \left\{ \mu_i = \cos\left(\frac{(2i-1)\pi}{2N}\right), i = 1, \dots, N \right\}$  les  $N$  points de Gauss qui sont les racines de  $\Pi_1(x) = T_N(x)$ .
- $\mathcal{GL}_x^N = \left\{ \nu_i = \cos\left(\frac{i\pi}{N}\right), i = 0, \dots, N \right\}$  les  $N+1$  points de Gauss-Lobatto qui sont les racines de  $\Pi_2(x) = (x^2 - 1)T'_N(x)$ .
- $\mathcal{GA}_x^N = \mathcal{G}_x^N \cup \{-1, +1\} = \{\mu_i, i = 0, 1, \dots, N, N+1 \text{ avec } \mu_0 = 1 \text{ et } \mu_{N+1} = -1\}$ . Ces  $N+2$  points que nous appellerons Gauss-Augmenté sont les racines de  $\Pi_3(x) = (x^2 - 1)T_N(x)$ .

Les formules suivantes nous seront utiles par la suite:

De  $T_N(\cos \theta) = \cos(N\theta)$ , on obtient:

$$T_N(\mu_i) = 0 \text{ pour } i = 1, \dots, N;$$

$$T_N(\nu_i) = (-1)^i \text{ pour } i = 1, \dots, N-1;$$

$$T_N(\mu_0) = T_N(\nu_0) = 1; T_N(\mu_{N+1}) = T_N(\nu_N) = (-1)^N$$

De  $T'_N(\cos \theta) = \frac{N \sin N\theta}{\sin \theta}$ , on obtient:

$$T'_N(\mu_i) = \frac{(-1)^{i+1} N}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } i = 1, \dots, N;$$

$$T'_N(\nu_i) = 0 \text{ pour } i = 1, \dots, N-1;$$

$$T'_N(\mu_0) = T'_N(\nu_0) = N^2; T'_N(\mu_{N+1}) = T'_N(\nu_N) = (-1)^{N+1} N^2$$

De  $T''_N(\cos \theta) = \frac{N \sin N\theta \cos \theta}{\sin^3 \theta} - \frac{N^2 \cos N\theta}{\sin^2 \theta}$ , on obtient:

$$T''_N(\mu_i) = \frac{(-1)^{i+1} N \cos \frac{(2i-1)\pi}{2N}}{\sin^3 \frac{(2i-1)\pi}{2N}} \text{ pour } i = 1, \dots, N;$$

$$T''_N(\nu_i) = \frac{(-1)^{i+1} N^2}{\sin^2 \frac{i\pi}{N}} \text{ pour } i = 1, \dots, N-1;$$

$$T''_N(\mu_0) = T''_N(\nu_0) = \frac{N^2(N^2-1)}{3}; T''_N(\mu_{N+1}) = T''_N(\nu_N) = (-1)^N \frac{N^2(N^2-1)}{3}$$

De  $T'''_N(\cos \theta) = \frac{N \sin N\theta}{\sin^5 \theta} (1 - N^2 + (2 + N^2) \cos^2 \theta) - \frac{3N^2 \cos \theta \cos N\theta}{\sin^4 \theta}$ , on obtient:

$$T'''_N(\mu_i) = \frac{(-1)^{i+1} N}{\sin^5 \frac{(2i-1)\pi}{2N}} (1 - N^2 + (2 + N^2) \cos^2 \frac{(2i-1)\pi}{2N}) \text{ pour } i = 1, \dots, N;$$

$$T'''_N(\nu_i) = \frac{(-1)^{i+1} 3N^2 \cos \frac{i\pi}{N}}{\sin^4 \frac{i\pi}{N}} \text{ pour } i = 1, \dots, N-1;$$

$$T'''_N(\mu_0) = T'''_N(\nu_0) = \frac{N^2}{15} (N^2 - 4)(N^2 - 1);$$

$$T'''_N(\mu_{N+1}) = T'''_N(\nu_N) = (-1)^{N+1} \frac{N^2}{15} (N^2 - 4)(N^2 - 1)$$

On a également:

$$T_N^{(IV)}(\nu_i) = \frac{(-1)^{i+1} N^2}{\sin^6 \frac{i\pi}{N}} ((4 - N^2) + (11 + N^2) \nu_i^2) \text{ pour } i = 1, \dots, N-1$$

En ce qui concerne les polyômes  $\Pi_1, \Pi_2$  et  $\Pi_3$ , on alors;

$$\Pi_1(\mu_i) = 0 \text{ pour } i = 1, \dots, N;$$

$$\Pi_1(\mu_0) = 1; \Pi_1(\mu_{N+1}) = (-1)^N$$

$$\Pi_1(\nu_i) = (-1)^i \text{ pour } i = 0, \dots, N$$

$$\Pi_2(\mu_i) = (-1)^i N \sin \left( \frac{(2i-1)\pi}{2N} \right) \text{ pour } i = 1, \dots, N;$$

$$\Pi_2(\mu_0) = \Pi_2(\mu_{N+1}) = 0$$

$$\Pi_2(\nu_i) = 0 \text{ pour } i = 0, \dots, N$$

$$\Pi_3(\mu_i) = 0 \text{ pour } i = 0, \dots, N+1$$

$$\Pi_3(\nu_i) = (-1)^i (\nu_i^2 - 1) \text{ pour } i = 0, \dots, N$$

En ce qui concerne leurs dérivées premières:

On a  $\Pi'_1(x) = T'_N(x)$  et donc:

$$\Pi'_1(\mu_i) = \frac{(-1)^{i+1} N}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } i = 1, \dots, N;$$

$$\Pi'_1(\mu_0) = N^2; \Pi'_1(\mu_{N+1}) = (-1)^{N+1} N^2$$

$$\Pi'_1(\nu_i) = 0 \text{ pour } i = 1, \dots, N-1;$$

$$\Pi'_1(\nu_0) = N^2; \Pi'_1(\nu_N) = (-1)^{N+1} N^2$$

On a  $\Pi'_2(x) = 2xT'_N(x) + (x^2 - 1)T''_N(x)$  et donc:

$$\Pi'_2(\mu_i) = \frac{(-1)^{i+1} N \mu_i}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } i = 1, \dots, N;$$

$$\Pi'_2(\mu_0) = 2N^2; \Pi'_2(\mu_{N+1}) = 2(-1)^N N^2$$

$$\Pi'_2(\nu_i) = (-1)^i N^2 \text{ pour } i = 1, \dots, N-1 \text{ et donc } \Pi'_2(\nu_i) = (-1)^i \bar{c}_i N^2 \text{ pour } i = 0, \dots, N$$

On a  $\Pi'_3(x) = 2xT'_N(x) + (x^2 - 1)T'_N(x)$  et donc:

$$\Pi'_3(\mu_i) = (-1)^i N \sin \left( \frac{(2i-1)\pi}{2N} \right) \text{ pour } i = 1, \dots, N$$

$$\begin{aligned}\Pi'_3(\mu_0) &= 2; \Pi'_3(\mu_{N+1}) = 2(-1)^{N+1} \\ \Pi'_3(\nu_i) &= (-1)^i 2\nu_i \text{ pour } i = 0, \dots, N\end{aligned}$$

En ce qui concerne leurs dérivées secondes:

On a  $\Pi''_1(x) = T''_N(x)$  et donc:

$$\begin{aligned}\Pi''_1(\mu_i) &= \frac{(-1)^{i+1} N \cos(\frac{(2i-1)\pi}{2N})}{\sin^3(\frac{(2i-1)\pi}{2N})} \text{ pour } i = 1, \dots, N; \\ \Pi''_1(\nu_i) &= \frac{(-1)^{i+1} N^2}{\sin^2(\frac{i\pi}{N})} \text{ pour } i = 1, \dots, N-1; \\ \Pi''_1(\mu_0) &= \Pi''_1(\nu_0) = \frac{N^2(N^2-1)}{3}; \Pi''_1(\mu_{N+1}) = \Pi''_1(\nu_N) = (-1)^N \frac{N^2(N^2-1)}{3}\end{aligned}$$

On a  $\Pi''_2(x) = 2T'_N(x) + 4xT''_N(x) + (x^2 - 1)T'''_N(x)$  et donc:

$$\begin{aligned}\Pi''_2(\nu_i) &= \frac{(-1)^{i+1} N^2 \cos \frac{i\pi}{N}}{\sin^2 \frac{i\pi}{N}} \text{ pour } i = 1, \dots, N-1 \\ \Pi''_2(\nu_0) &= \frac{2N^2(2N^2+1)}{3}; \Pi''_2(\nu_N) = \frac{(-1)^{N+1} 2N^2(2N^2+1)}{3}\end{aligned}$$

On a  $\Pi''_3(x) = 2T_N(x) + 4xT'_N(x) + (x^2 - 1)T''_N(x)$  et donc:

$$\begin{aligned}\Pi''_3(\mu_i) &= \frac{(-1)^{i+1} 3N \cos \frac{(2i-1)\pi}{2N}}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } i = 1, \dots, N \\ \Pi''_3(\mu_0) &= 2(2N^2 + 1); \Pi''_3(\mu_{N+1}) = (-1)^N 2(2N^2 + 1)\end{aligned}$$

En ce qui concerne leurs dérivées troisièmes:

On a  $\Pi'''_2(x) = 6T''_N(x) + 6xT'''_N(x) + (x^2 - 1)T''''_N(x)$  et donc:

$$\begin{aligned}\Pi'''_2(\nu_i) &= \frac{(-1)^{i+1} N^2}{\sin^5 \frac{i\pi}{N}} ((N^2 - 1)(1 - \nu_i^2) + 3) \text{ pour } i = 1, \dots, N-1 \\ \Pi'''_2(\nu_0) &= \frac{2N^2}{5} (N^2 + 1)(N^2 - 1); \\ \Pi'''_2(\nu_N) &= (-1)^N \frac{2N^2}{5} (N^2 + 1)(N^2 - 1)\end{aligned}$$

On a  $\Pi'''_3(x) = 6T'_N(x) + 6xT''_N(x) + (x^2 - 1)T'''_N(x)$  et donc:

$$\begin{aligned}\Pi'''_3(\mu_i) &= \frac{(-1)^{i+1} N}{\sin^3 \frac{(2i-1)\pi}{2N}} ((N^2 + 2)(1 - \mu_i^2) + 3) \text{ pour } i = 1, \dots, N \\ \Pi'''_3(\mu_0) &= 2N^2(N^2 + 2); \Pi'''_3(\mu_{N+1}) = (-1)^{N+1} 2N^2(N^2 + 2)\end{aligned}$$

### Opérateurs de projection

L'évaluation au point  $x_i$  du polyôme interpolant la fonction  $u$  en des points  $x_j$  s'écrit  $I_N u(x_i) = \sum_{j=0}^N u_j h_j(x_i)$  et s'exprime donc par une multiplication matricielle dont le terme général est

$$p_{i,j} = h_j(x_i) = \frac{\Pi(x_i)}{(x_i - x_j)\Pi'(x_j)} \text{ où } i \text{ et } j \text{ sont les indices de ligne et colonne respectivement.}$$

- En ce qui concerne la projection des points de Gauss-Lobatto vers les points de Gauss-Augmenté, cette matrice de terme général  $p_{ij}$  est donc d'ordre  $(N+2) \times (N+1)$ :

$$\text{On a } h_j(x_i) = \frac{\Pi_2(\mu_i)}{(\mu_i - \nu_j)\Pi_2'(\nu_j)} \text{ pour } 1 \leq i \leq N; 0 \leq j \leq N;$$

et donc

$$p_{i,j} = \frac{(-1)^{i+j} \sin \frac{(2i-1)\pi}{2N}}{N(\mu_i - \nu_j)c_j} \text{ pour } 1 \leq i \leq N; 0 \leq j \leq N;$$

$$p_{0,0} = p_{N+1,N} = 1$$

$$p_{0,j} = 0 \text{ pour } j \geq 1; p_{N+1,j} = 0 \text{ pour } j \leq N-1$$

- En ce qui concerne la projection des points de Gauss-Augmenté vers les points de Gauss-Lobatto, cette matrice de terme général  $p_{ij}$  est donc d'ordre  $(N+1) \times (N+2)$ :

$$\text{On a } h_j(x_i) = \frac{\Pi_3(\nu_i)}{(\nu_i - \mu_j)\Pi_3'(\mu_j)} \text{ pour } 1 \leq i \leq N-1; 0 \leq j \leq N+1,$$

et donc:

$$p_{i,j} = \frac{(-1)^{i+j}(\nu_i^2 - 1)}{N(\nu_i - \mu_j) \sin \frac{(2j-1)\pi}{2N}} \text{ pour } 1 \leq i \leq N-1; 1 \leq j \leq N;$$

$$p_{i,0} = \frac{(-1)^i(\nu_i + 1)}{2} \text{ pour } 1 \leq i \leq N-1;$$

$$p_{i,N+1} = \frac{(-1)^{i+N+1}(\nu_i - 1)}{2} \text{ pour } 1 \leq i \leq N-1;$$

$$p_{0,0} = p_{N,N+1} = 1$$

$$p_{0,j} = 0 \text{ pour } j \geq 1; p_{N,j} = 0 \text{ pour } j \leq N$$

### Opérateurs de dérivation première

L'évaluation au point  $x_i$  de la dérivée du polyôme interpolant la fonction  $u$  en des points  $x_j$  s'écrit  $(I_N u)'(x_i) = \sum_{j=0}^N u_j h'_j(x_i)$  et s'exprime donc par une multiplication matricielle de terme

général  $h'_j(x_i)$  où  $i$  et  $j$  sont les indices de ligne et colonne respectivement. On a formellement  $h'_j(x) = \frac{\Pi'(x)}{(x-x_j)\Pi'(x_j)} - \frac{\Pi(x)}{(x-x_j)^2\Pi'(x_j)}$  et donc de façon générale lorsque  $x_i$  est différent de  $x_j$ ,  $h'_j(x_i) = \frac{\Pi'(x_i)}{(x_i-x_j)\Pi'(x_j)} - \frac{\Pi(x_i)}{(x_i-x_j)^2\Pi'(x_j)}$ . En  $x_j$ , un développement limité montre que  $h'_j(x_j) = \frac{1}{2} \frac{\Pi''(x_j)}{\Pi'(x_j)}$

• La matrice de dérivation sur les points de Gauss est une matrice carrée d'ordre  $N$  de terme général  $d_{i,j}$  et on a:

$$d_{i,j} = \frac{(-1)^{i+j}}{\mu_i - \mu_j} \frac{\sin \frac{(2j-1)\pi}{2N}}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } 1 \leq i \leq N; 1 \leq j \leq N; i \neq j$$

$$d_{i,i} = \frac{\mu_i}{2(1-\mu_i^2)} \text{ pour } 1 \leq i \leq N$$

• La matrice de dérivation sur les points de Gauss-Lobatto est une matrice carrée d'ordre  $N+1$  de terme général  $d_{i,j}$  et on a:

$$d_{i,j} = \frac{(-1)^{i+j}}{\nu_i - \nu_j} \frac{\tilde{c}_i}{\tilde{c}_j} \text{ pour } 0 \leq i \leq N; 0 \leq j \leq N; i \neq j$$

$$d_{i,i} = \frac{\nu_i}{2(\nu_i^2 - 1)} \text{ pour } 1 \leq i \leq N - 1$$

$$d_{0,0} = \frac{2N^2+1}{6} = -d_{N,N}$$

• La matrice de dérivation sur les points de Gauss-Augmenté est une matrice carrée d'ordre  $N+2$  de terme général  $d_{i,j}$  et on a:

$$d_{i,j} = \frac{(-1)^{i+j}}{\mu_i - \mu_j} \frac{\sin \frac{(2j-1)\pi}{2N}}{\sin \frac{(2i-1)\pi}{2N}} \text{ pour } 1 \leq i \leq N; 1 \leq j \leq N; i \neq j$$

$$d_{i,i} = \frac{3\mu_i}{2(\mu_i^2 - 1)} \text{ pour } 1 \leq i \leq N$$

$$d_{i,0} = \frac{(-1)^i N \sin \frac{(2i-1)\pi}{2N}}{2(\mu_i - 1)} \text{ pour } 1 \leq i \leq N$$

$$d_{i,N+1} = \frac{(-1)^{i+N+1} N \sin \frac{(2i-1)\pi}{2N}}{2(\mu_i + 1)} \text{ pour } 1 \leq i \leq N$$

$$d_{0,j} = \frac{2(-1)^j}{N(1-\mu_j) \sin \frac{(2j-1)\pi}{2N}} \text{ pour } 1 \leq j \leq N$$

$$d_{N+1,j} = \frac{2(-1)^{j+N}}{N(1+\mu_j) \sin \frac{(2j-1)\pi}{2N}} \text{ pour } 1 \leq j \leq N$$

$$d_{0,N+1} = \frac{(-1)^{N+1}}{2}$$

$$d_{N+1,0} = \frac{(-1)^N}{2}$$

$$d_{0,0} = -d_{N+1,N+1} = \frac{2N^2+1}{2}$$

• La matrice de dérivation des points de Gauss vers les points de Gauss-Lobatto est une matrice d'ordre  $(N+1) \times N$  de terme général

$$d_{i,j} = \frac{\Pi'_1(\nu_i)}{(\nu_i - \mu_j)\Pi'_1(\mu_j)} - \frac{\Pi_1(\nu_i)}{(\nu_i - \mu_j)^2\Pi'_1(\mu_j)} \text{ et on a:}$$

$$d_{i,j} = \frac{(-1)^{i+j} \sin \frac{(2j-1)\pi}{2N}}{N(\nu_i - \mu_j)^2} \text{ pour } 1 \leq i \leq N - 1; 1 \leq j \leq N$$

$$d_{0,j} = \frac{(-1)^{i+1} \sin \frac{(2j-1)\pi}{2N}}{N} \left( \frac{N^2}{1-\mu_j} - \frac{1}{(1-\mu_j)^2} \right) \text{ pour } 1 \leq j \leq N$$

$$d_{N,j} = \frac{(-1)^{i+N+1} \sin \frac{(2j-1)\pi}{2N}}{N} \left( \frac{N^2}{1+\mu_j} - \frac{1}{(1+\mu_j)^2} \right) \text{ pour } 1 \leq j \leq N$$

• La matrice de dérivation des points de Gauss-Lobatto vers les points de Gauss est une matrice

d'ordre  $N \times (N+1)$  de terme général

$$d_{i,j} = \frac{\Pi_2(\mu_i)}{(\mu_i - \nu_j)\Pi_2(\nu_j)} - \frac{\Pi_2(\mu_i)}{(\mu_i - \nu_j)^2\Pi_2'(\nu_j)} \text{ et on a:}$$

$$d_{i,j} = \frac{(-1)^{i+j+1}}{N\bar{c}_j(\mu_i - \nu_j)^2} \left( \frac{\mu_i}{\sin \frac{(2i-1)\pi}{2N}} - \frac{\sin \frac{(2j-1)\pi}{2N}}{\mu_i - \nu_j} \right) \text{ pour } 1 \leq i \leq N; 0 \leq j \leq N$$

### Opérateurs de dérivation seconde

L'évaluation au point  $x_i$  de la dérivée seconde du polyôme interpolant la fonction  $u$  en des

points  $x_j$  s'écrit  $(I_N u)''(x_i) = \sum_{j=0}^N u_j h_j''(x_i)$  et s'exprime donc par une multiplication matricielle

de terme général  $h_j''(x_i)$  où  $i$  et  $j$  sont les indices de ligne et colonne respectivement. On a formellement  $h_j''(x) = \frac{\Pi''(x)}{(x-x_j)\Pi'(x_j)} - 2\frac{\Pi'(x)}{(x-x_j)^2\Pi'(x_j)} + 2\frac{\Pi(x)}{(x-x_j)^3\Pi'(x_j)}$  et donc de façon générale lorsque  $x_i$  est différent de  $x_j$ ,  $h_j''(x_i) = \frac{\Pi''(x_i)}{(x_i-x_j)\Pi'(x_j)} - 2\frac{\Pi'(x_i)}{(x_i-x_j)^2\Pi'(x_j)} + 2\frac{\Pi(x_i)}{(x_i-x_j)^3\Pi'(x_j)}$ . Si l'ensemble des points  $x_i$  est le même que les  $x_j$  (seul cas considéré ici) alors le dernier terme s'annule et  $h_j''(x_i) = \frac{\Pi''(x_i)}{(x_i-x_j)\Pi'(x_j)} - 2\frac{\Pi'(x_i)}{(x_i-x_j)^2\Pi'(x_j)}$ . En  $x_j$ , un développement limité montre que  $h_j''(x_j) = \frac{1}{3} \frac{\Pi'''(x_j)}{\Pi'(x_j)}$

• La matrice de dérivation seconde sur les points de Gauss-Lobatto est une matrice carrée d'ordre  $N+1$  de terme général  $d_{i,j}$  et on a:

$$d_{i,j} = \frac{(-1)^{i+j}}{\bar{c}_j(\nu_i - \nu_j)^2} \frac{\nu_i^2 + \nu_i \nu_j - 2}{(1 - \nu_i^2)} \text{ pour } 1 \leq i \leq N-1; 0 \leq j \leq N; i \neq j$$

$$d_{i,i} = -\frac{(N^2-1)(1-\nu_i^2)+3}{3(1-\nu_i^2)^2} \text{ pour } 1 \leq i \leq N-1$$

$$d_{0,j} = \frac{2(-1)^j}{3\bar{c}_j(1-\nu_j)^2} ((2N^2+1)(1-\nu_j) - 6) \text{ pour } 1 \leq j \leq N$$

$$d_{N,j} = \frac{2(-1)^{j+N}}{3\bar{c}_j(1+\nu_j)^2} ((2N^2+1)(1+\nu_j) - 6) \text{ pour } 0 \leq j \leq N-1$$

$$d_{0,0} = d_{N,N} = \frac{N^4-1}{15}$$

Cette matrice a déjà été calculée par R. Peyret (1985).

• La matrice de dérivation seconde sur les points de Gauss-Augmenté est une matrice carrée d'ordre  $N+2$  de terme général  $d_{i,j}$  et on a:

$$d_{i,j} = \frac{(-1)^{i+j+1}}{(\mu_i - \mu_j)^2} \frac{\mu_i^2 - 3\mu_i \mu_j + 2}{\sin \frac{(2i-1)\pi}{2N} \sin \frac{(2j-1)\pi}{2N}} \text{ pour } 1 \leq i \leq N; 1 \leq j \leq N; i \neq j$$

$$d_{i,i} = -\frac{(N^2+2)(1-\mu_i^2)+3}{3(1-\mu_i^2)^2} \text{ pour } 1 \leq i \leq N$$

$$d_{i,0} = \frac{(-1)^{i+1}N}{2\sin \frac{(2i-1)\pi}{2N}} \frac{\mu_i-2}{\mu_i-1} \text{ pour } 1 \leq i \leq N$$

$$d_{i,N+1} = \frac{(-1)^{i+N}N}{2\sin \frac{(2i-1)\pi}{2N}} \frac{\mu_i+2}{\mu_i+1} \text{ pour } 1 \leq i \leq N$$

$$d_{0,j} = \frac{(-1)^j 2}{N \sin \frac{(2j-1)\pi}{2N}} \frac{(2N^2+1)(1-\mu_j)-2}{(1-\mu_j)^2} \text{ pour } 1 \leq j \leq N$$

$$d_{N+1,j} = \frac{(-1)^{j+N+1} 2}{N \sin \frac{(2j-1)\pi}{2N}} \frac{(2N^2+1)(1+\mu_j)-2}{(1+\mu_j)^2} \text{ pour } 1 \leq j \leq N$$

$$d_{0,N+1} = d_{N+1,0} = (-1)^{N+1} N^2$$

$$d_{0,0} = d_{N+1,N+1} = \frac{N^2(N^2+2)}{3}$$

Ces deux matrices sont également les carrés des matrices de dérivation première sur les points de collocation correspondants.