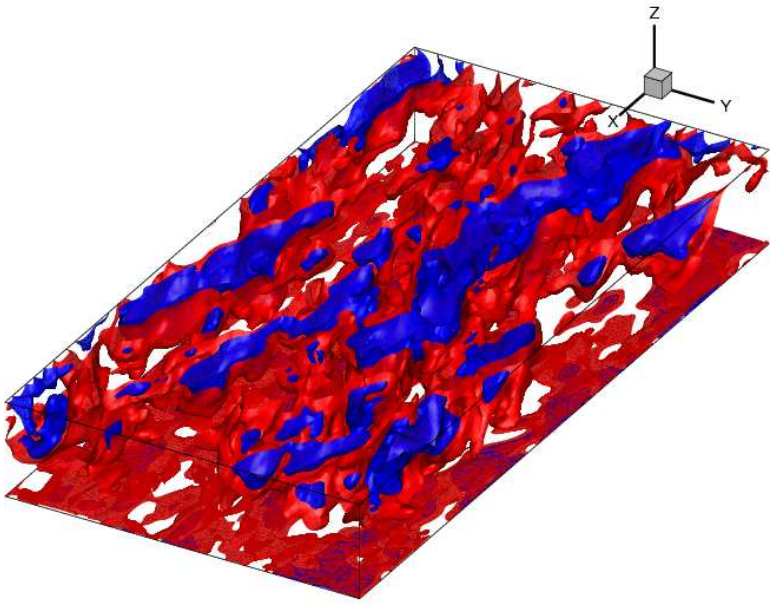
Année 2013

Méthodes numériques pour les écoulements incompressibles

Patrick Le Quéré et Bérengère Podvin



Fluctuation de vitesse longitudinale dans un canal $R_\tau = 295$
Y. Fraigneau, 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 commencons 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)
  - **Chebyshev and Fourier Spectral Methods** J. P. Boyd, Dover 2001.
  - **Spectral Methods in Fluid Dynamics**, Canuto, Hussaini, Quarteroni, Zang, Springer-Verlag 1991.
  - **Spectral Methods for Incompressible Viscous Flow** R. Peyret, Springer 2001.
- Méthodes numériques (divers) :
  - **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

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 ρ, 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
$$\frac{\partial \rho}{\partial t} + \frac{(\rho \partial u_j)}{\partial x_j} = 0 \tag{1.1}$$
- la conservation de la quantité de mouvement
$$\frac{\partial (\rho u_i)}{\partial t} + u_j \frac{\partial (\rho u_i)}{\partial x_j} = \frac{\partial p}{\partial x_j} + \nu \frac{\partial u_i}{\partial x_j x_j} \tag{1.2}$$
- la conservation de l'énergie $E = E + \frac{1}{2} u_i u_i$
$$\frac{\partial E}{\partial t} = W + Q \tag{1.3}$$
où W est la quantité de travail reçu par le système et Q est la dissipation.
Les inconnues du problème sont ρ,$u_{i,i=1,2,3}$,p. Cette équation est souvent remplacée par une équation d'état (par exemple 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 entre elles le comportement des variables thermodynamiques.

1.2 Ecoulement 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

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

où D est la dérivée particulaire.
En soustrayant les équations 1.1 et 1.4, on obtient

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{1.5}$$

Un écoulement incompressible est caractérisé par un champ de vitesse à divergence nulle (autrement dit solénoidal). 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

$$|\frac{1}{\rho}\frac{D\rho}{Dt}| << U/L \tag{1.6}$$

Les variations de densité peuvent être exprimées à l'aide de deux variables thermodynamiques. Suivant Batchelor (An introduction to Fluid Dynamics), nous utilisons la pression p et l'entropie s et exprimons

$$\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial p}|_s \frac{Dp}{Dt} + \frac{\partial \rho}{\partial s}|_s \frac{Ds}{Dt} \tag{1.7}$$

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

$$\frac{1}{\rho}\frac{D\rho}{Dt} = \frac{1}{\rho a^2}\frac{Dp}{Dt} \tag{1.8}$$

En utilisant d'une part

$$\frac{Dp}{Dt} = \frac{\partial p}{\partial t} + u_i \frac{\partial p}{\partial x_i} \tag{1.9}$$

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

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

et

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

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

$$\rho\frac{Du_i}{Dt} = -\frac{\partial p}{\partial x_i} \tag{1.10}$$

Ceci nous permet d'établir que

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

La première contribution

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

ce qui nous donne

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

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

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

et la deuxième

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

ce qui nous donne

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

Une analyse plus fine des variations de vitesse

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

nous conduit à estimer le premier terme comme

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

et le second comme

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

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

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

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

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

où

$$\underline{F} = -\underline{g}\alpha\Delta T \tag{1.13}$$

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

– la conservation de l'énergie

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

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 solénoidalité (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

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

On cherche à résoudre cette équation sous la contrainte de solénoidalité $divu = 0$. L'idée classique est de définir un Lagrangien augmenté

$$\mathcal{L} = \frac{1}{2}|u - u*|^2 + \lambda \frac{\partial u_i}{\partial x_i} \tag{1.16}$$

où $u*$ vérifie

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

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

$$(u - u*).dv + dv.(u - u*) + 2\lambda \frac{\partial dv_i}{\partial x_i} = 0 \tag{1.18}$$

En intégrant par parties, il vient

$$(u - u*).dv + dv.(u - u*) - 2dv_i \frac{\partial \lambda}{\partial x_i} = 0 \tag{1.19}$$

soit

$$u_i = u_i* - \frac{\partial p}{\partial x_i} \tag{1.20}$$

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 chapître 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

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

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 $\Delta > 0$:

$$\frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{1.22}$$

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

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} \tag{1.23}$$

La solution de cette équation est de la forme

$$u(x,t) = F(x - ct) + G(x + ct) \tag{1.24}$$

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 $\Delta = 0$:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial y} \tag{1.25}$$

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).

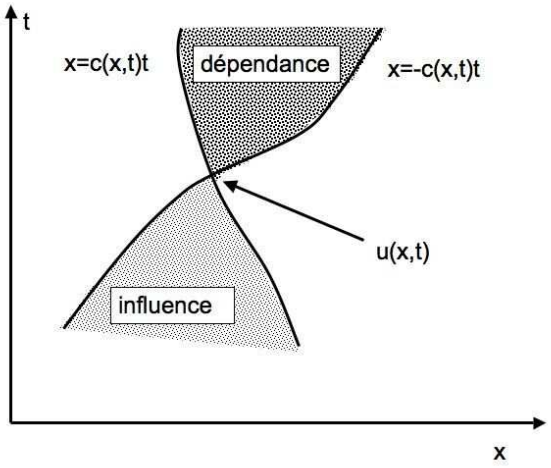


FIG. 1.1 – *Domaines de dépendance et d'influence d'une équation hyperbolique*

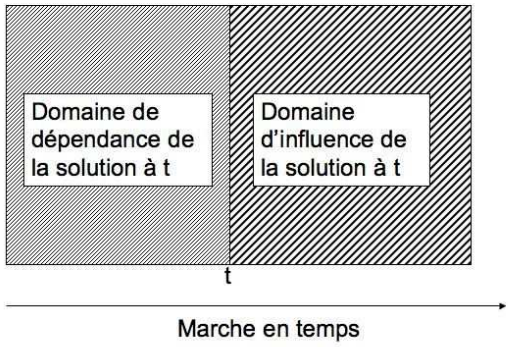


FIG. 1.2 – *Domaines de dépendance et d'influence d'une équation parabolique*

– Si $\Delta < 0$: L'équation canonique est l'équation de Laplace:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{1.26}$$

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 coincide 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 notament l'ellipticité:

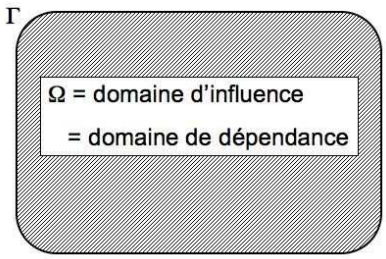


FIG. 1.3 – *Domaines de dépendance et d'influence d'une équation elliptique*

Définition:
Soit l'opérateur L défini par

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

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

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

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

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

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 intuitive semble être de considérer un nuage de points et 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

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

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

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

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

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

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 commme

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

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

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

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 dérivée 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

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{2.4}$$

sur [0,1] avec la condition aux limites

$$u(0) = u(1) = 1.0 \tag{2.5}$$

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:
   $$\frac{\partial^2 f}{\partial x^2}|_{x_I} = \frac{f_{I+1} + f_{I-1} - 2f_I}{\Delta x^2}$$
   $$\frac{\partial f}{\partial t}|_{x_I}^{t=t^n} = \frac{f^{t_{n+1}} - f^{t_n}}{\Delta t}$$
3. substitution de l'approximation dans l'EDP et obtention d'un système aux différences finies.
   exemple:
   $$\frac{\partial f}{\partial x}|_{x_I} = \frac{f_{I+1} + f_{I-1} - 2f_I}{\Delta x^2}$$

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.

$$u(x) = \sum_I u_I N_I(x) \tag{2.6}$$

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

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

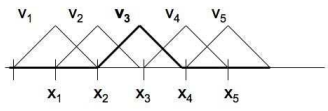
En outre, on demande à ce que les constantes soient interpolées exactement

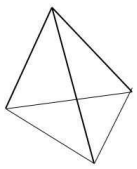
$$\sum_I N_I = 1 \tag{2.8}$$

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

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

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

$$f(u) = 0 \tag{2.10}$$

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

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

ce qui conduit à un système linéaire de résolution pour les coefficient a^j.
On considère un exemple l'équation de la chaleur

$$\frac{\partial U}{\partial t} = \alpha \frac{\partial^2 U}{\partial x^2} \tag{2.12}$$

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 coincident avec les fonctions de base. Les fonctions v vérifient aussi les conditions aux limites homogènes aux frontières.

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

En intégrant par parties, il vient que

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

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

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

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'apercoit 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.

$$\int\int\int \frac{\partial u_i}{\partial t} d\underline{x} + \int\int\int \frac{\partial u_i u_i}{\partial x_j} d\underline{x} = -\int\int\int \frac{\partial p}{\partial x_j} + \nu \frac{\partial^2 u}{\partial x_j x_j} d\underline{x} \tag{2.16}$$

En utilisant le théorême de la divergence

$$\int_\Omega div F d\Omega = \int_{\partial\Omega} F.dS \tag{2.17}$$

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 **??**:

$$\frac{\partial}{\partial t}\int_\Omega U d\Omega + \int_S F.dS = \int Q d\Omega \tag{2.18}$$

Le domaine est découpé en volumes de contrôle Ω_J qui recouvrent le domaine. Ces cellules ne sont pas nécessairement disjointes, il peut y avoir recouvrement à condition que ce recouvrement n'induise pas la création de nouvelles frontières.
On évalue cette équation à l'intérieur d' un volume Ω_J

$$\frac{\partial}{\partial t} U_J \Omega_J + \sum_{cotes} F.S = Q_J \Omega_J \tag{2.19}$$

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,

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

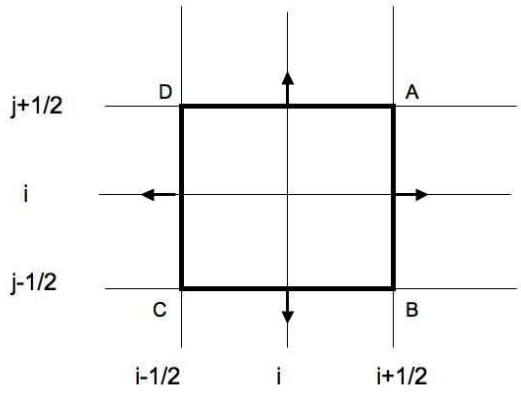


FIG. 2.2 – *Description d'un volume de contrôle*

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 Ω_J est représenté par la figure et que le maillage est régulier avec un pas de maillage Δx et Δy, l'équation peutêtre simplifiée comme

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

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 apparait 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 $v > 0$

$$f_{i+1/2} = \frac{3u_{i+1} + 6u_i - u_{i-1}}{8}$$
$$f_{i-1/2} = \frac{3u_i + 6u_{i-1} - u_{i-2}}{8}$$
$$Ic = v\frac{3u_{i+1} + 3u_i - 7u_{i-1} + u_{i-2}}{8\Delta x}$$

Si $v < 0$

$$f_{i+1/2} = \frac{3u_i + 6u_{i+1} - u_{i+2}}{8}$$
$$f_{i-1/2} = \frac{3u_{i-1} + 6u_i - u_{i+1}}{8}$$
$$Ic = -v\frac{3u_{i-1} + 3u_i - 7u_{i+1} + u_{i+2}}{8\Delta x}$$

Cette interpolation est connue sous le nom de schéma QUICK.

Remarque: On peut voir la représentation aux différence 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

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

– schéma Euler d'ordre 1 explicite

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

– schéma Euler d'ordre 1 implicite

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

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

– schéma Euler d'ordre 2 retardé explicite

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

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 'a 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 .

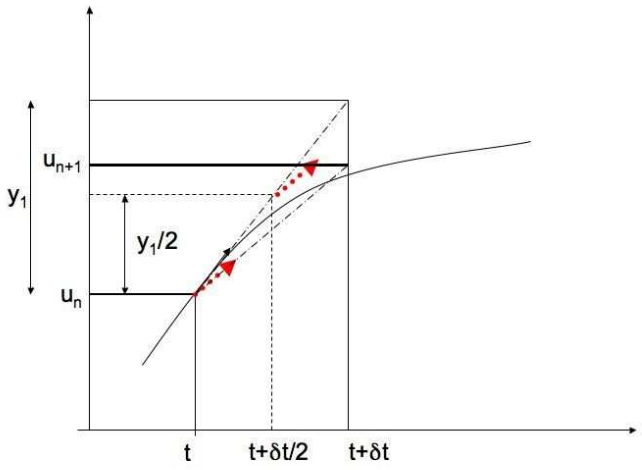


FIG. 2.3 – *Méthode de Runge-Kutta d'ordre 2*

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).

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

2.2.3 Discrétisation du terme source

– schéma de Crank-Nicolson (implicite)

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

– schéma d'Adams-Bashforth (explicite)

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

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

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

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

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

soit un problème de Helmholtz de la forme

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

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 jusquà 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 suppport global, qui requièrent des géométries simples. Si la solution est sufffisamment 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

$$\mathcal{L}f = s \text{ dans } \Omega \tag{2.30}$$

$$f = \bar{f} \text{ sur } \partial\Omega \tag{2.31}$$

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.

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

- ϕ_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 (,).
- $\hat{f}_n$ sont les coefficients spectraux

En pratique, on approche f par f_N

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

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

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

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

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

Les $\{\exp(ikx), k = -\infty, \cdots, \infty\}$ 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:

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

et donc

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

pour $l = -\infty, \cdots, \infty$ et donc galement pour $-N/2 \leq l \leq N/2$. Ceci montre que f_N est la projection L_2 de f sur $\{\exp(ikx), k = -N/2, \cdots, N/2\}$.
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 $\Delta x = \frac{2\pi}{N+1}$, $\left\{x_j = \frac{2\pi j}{N+1}, j = 0,1,\cdots,N+1\right\}$, il vient

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

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:

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

Cette formule d'intégration est de précision maximale dans l'ensemble des fonctions $\mathcal{C}^\infty$ 2π périodiques
Une deuxième façon de procéder est la suivante: plaçons nous dans l'espace vectoriel $\mathcal{F}_N$ engendré par les $\{\exp(ikx), k = -N/2,\cdots,N/2\}$. Une fonction périodique étant connue par ses valeurs $f_j = f(x_j)$ aux points $\left\{x_j = \frac{2\pi j}{N+1}, j = 0,\cdots,N+1\right\}$, définissons $I_N f$ le polynôme d'interpolation trigonométrique dans $\mathcal{F}_N$ qui vaut f_j aux points x_j. On a donc

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

et aux points x_j

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

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 $u(x_j) = 0, \forall j = 0,\cdots,N$ et on a alors $u = 0$.
Les $\{\exp(ikx), k = -N/2,\cdots,N/2\}$ continuent de former une famille orthogonale dans $\mathcal{F}_N$ pour le produit scalaire discret $(,)_d$. En effet

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$+ \frac{1}{(il)^2} \int_0^{2\pi} f''_N \exp(-ilx) dx \tag{2.52}$$

On voit donc que $\hat{f}_l$ décroit comme $\frac{1}{l^2}$ si f est de classe $\mathcal{C}^2$ et si le premier terme du membre de droite s'annule c'est-à-dire si f_N est périodique.
En itérant le raisonnement on arrive donc au résultat suivant:
Si f est de classe $\mathcal{C}^m$ et si f et toutes ses dérivées jusqu'à l'ordre $m-2$ sont périodiques, $\hat{f}_l$ décroit comme $\frac{1}{l^m}$.
On voit donc que si la solution recherchée n'est pas très régulière ou si elle et ses dérivées ne sont pas périodiques jusqu'à un ordre élevé, alors les coefficients spectraux ne vont pas décroître très vite et l'erreur $\|f - f_N\|$ sera du même ordre que celle obtenue par une méthode de différences finies, pour un coût supérieur, et donc l'emploi des séries de Fourier ne se justifie pas.
En corollaire, on peut donc dire qu'il ne faut pas chercher à représenter en série de Fourier la solution de problèmes qui ne présentent pas naturellement de périodicité. Ceci restreint considérablement leur utilisation pratique, puisqu'il est notamment exclus de les utiliser pour rechercher la solution des équations de Navier-Stokes dans des géométries limitées par des parois solides.

Dérivation

Pour la dérivation première, on a:

$$\frac{\partial}{\partial x} \begin{cases} P_N f = \sum_{k=-K/2}^{K/2} ik\hat{f}_k \exp(ikx) \\ I_N f = \sum_{k=-K/2}^{K/2} ik\tilde{f}_k \exp(ikx) \end{cases}$$

et les dérivations d'ordre supérieur

$$\frac{\partial^n}{\partial x^n} \begin{cases} P_N f = \sum_{k=-K/2}^{K/2} (ik)^n \hat{f}_k \exp(ikx) \\ I_N f = \sum_{k=-K/2}^{K/2} (ik)^n \tilde{f}_k \exp(ikx) \end{cases}$$

Obtention des $\frac{\partial f}{\partial x}(x_j)$ à partir des $f(x_j)$.
Connaissant les $f(x_j)$, on obtient les $\tilde{f}_k$ à partir de
$\tilde{f}_l = \frac{1}{N+1} \sum_{j=0}^{N} f_N(x_j) \exp(-ilx_j)dx$; on forme ensuite
$ik\tilde{f}_k$ pour $k = -K/2, \cdots, K/2$ et on réévalue ensuite

$$\frac{\partial I_N f}{\partial x}(x_j) = \sum_{k=-K/2}^{K/2} ik\tilde{f}_k \exp(ikx_j)$$

Ces opérations prennent la forme matricielle suivante :
Soit $(f_0, f_1, \cdots, f_N)^t$ le vecteur des $f(x_j)$. On a

$$\begin{bmatrix} \tilde{f}_{-N/2} \\ . \\ \tilde{f}_k \\ . \\ \tilde{f}_{N/2} \end{bmatrix} = \begin{bmatrix} \\ FPS_{kj} \\ \\ \end{bmatrix} \begin{bmatrix} f_0 \\ . \\ f_j \\ . \\ f_N \end{bmatrix}$$

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

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

et enfin

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

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

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

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'utilsation des Transformées Rapides de Fourier. Sur des machines vectorielles comme le Cray, le point de 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 $(,)_d$.
- 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

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

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

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

On a respectivement:

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

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

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:

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

$$T_{n+1} = 2xT_n - T_{n-1} \tag{2.58}$$

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 $(,)_\omega$.
Les $\{T_n, n = 0,\cdots,\infty\}$ 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^2_ω de f dans $\mathcal{P}_N$, l'ensemble des polynômes de degré $\leq N$.
On a également

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

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):

$$\|f - P_N f\|_{L^2_\omega} \leq N^{-\sigma}\|f\|_{H^\sigma_\omega} \tag{2.60}$$

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 decroît 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 ω est une fonction poids, positive sur $]-1,1[$. On va chercher à approcher I_ω 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 α_i sont des coefficients.
Les α_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 α_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 α_i et x_i est la suivante:
• supposons les x_i donnés. On peut alors déterminer les α_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 α_i dont l'inversibilité est assurée si les x_i sont 2 à 2 distincts.
• les α_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 $(,)_\omega$, 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 α_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 λ et μ 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 ω 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 α_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 α_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 $\sum_{mpair} \frac{2}{1-m^2}\hat{f}_m$.

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 $\mathcal{P}_N$:

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

avec $l = 1,2$ selon que l'on considére les points de Gauss ou de Gauss Lobatto. Cette forme bilinéaire est symétrique et si de plus $(u,u)_d = 0$, alors $u(x_j) = 0, \forall j = 0, \cdots, N$ (puisque les $\alpha_i > 0$) 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 $(,)_d^1$, 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 $(,)_d^2$, 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 $(,)_d^2$. 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:

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

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:

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

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:

$$\begin{aligned}
\tilde{f}_l &= \frac{2}{\bar{c}_l N} \sum_{j=0}^{N} \frac{1}{\bar{c}_j} f(x_j) T_l(x_j) && (2.63)\\
&= \frac{2}{\bar{c}_l N} \sum_{j=0}^{N} \frac{1}{\bar{c}_j} \sum_{m=0}^{\infty} \hat{f}_m T_m(x_j) T_l(x_j) && (2.64)\\
&= \frac{2}{\bar{c}_l N} \sum_{m=0}^{\infty} \hat{f}_m \sum_{j=0}^{N} \frac{1}{\bar{c}_j} T_m(x_j) T_l(x_j) && (2.65)\\
&= \hat{f}_l + \frac{2}{\bar{c}_l} \sum_{m=N+1}^{\infty} \hat{f}_m (T_m, T_l)_d^2 && (2.66)
\end{aligned}$$

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:

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

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:

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

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 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 $\Pi(x) = \Pi_{i=0}^{N}(x - x_i)$ 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

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

$$f = \overline{f} \text{ sur } \partial\Omega \tag{2.69}$$

où Ω est un ouvert connexe de $I\!R^2$ou $I\!R^3$.
Ce type d'équation résulte de la discrétisation temporelle d'une équation de type transport-diffusion

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

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

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

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

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

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

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

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

$$\frac{11f^{n+1} - 18f^n + 9f^{n-1} - 2f^{n-2}}{6\Delta t}$$
$$+3\mathbf{V}.\nabla f^n - 3\mathbf{V}.\nabla f^{n-1} + \mathbf{V}.\nabla f^{n-2} = \nabla^2 f^{n+1}$$

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

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

où λ 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 ϕ_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 ϕ_n,$n = 0,...N$. Cette méthodologie est intéressante si les ϕ_n,$n = 0,...N$ forment une famille orthogonale et si les opérations de dérivation sont diagonales dans l'espace des ϕ_n,$n = 0,...N$. Elles conviennent donc parfaitement pour la mise en œuvre des approximations par des polynômes trigonométriques, qui remplissent ces deux conditions. Soit à résoudre

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

$$f 2\pi \text{ périodique} \tag{2.72}$$

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 $\exp(ikx)$,$k = -N/2,...,N/2$, ce qui donne donc, compte-tenu de l'orthogonalité des $\exp(ikx)$,$k = -N/2,...,N/2$, donne:

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

ou encore

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

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

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

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 ϕ_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

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

$$f(-1) = f(1) = 0 \tag{2.77}$$

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 $T_k, k = 0,....,N-2$.[1]
Ces $N-1$ relations et les conditions aux limites s'écrivent donc:

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

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

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 $T_k, k = 0,....,N-k$.

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 derivées 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:

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

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:

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

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

en premier, le système prend la forme matricielle:

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

Ce système peut être de plus découplé en deux systèmes portant respectivement sur les modes d'indice pair et impair, qui peuvent être résolus séparément [2]. 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,$i = 1,..,N-1$, soit

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

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 (SIAM Journal of Numerical Analysis 1983) ont montré que la matrice d'ordre $N-1$ précédente était diagonalisable sur $\mathbb{R}$, 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,

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

Dans le cas de conditions aux limites de Neumannn 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,$i = 1,...,N-1$, 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 $\mathcal{IR}$, 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'intrêt de la mthode.

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 $u \in P_N^0$

$$\int_{-1}^{1} (\nabla u.\nabla v + \lambda uv) \;=\; \int_{-1}^{1} fv \tag{2.82}$$

$v \in P_N^0$, 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,$k = 1,..,N-1$. $\nabla u.\nabla v$ appartenant à P_{2N-2}, on peut remplacer l'intégrale continue par la formule de quadrature de Gauss-Lobatto-Legendre basée sur les points ξ_i et les poids α_i. Les $N-1$ équations s'écrivent donc

$$\sum_{i=1}^{N-1} \alpha_i (\sum_{j=1}^{N-1} u_j L_j'(\xi_i) L_k'(\xi_i) + \lambda \sum_{j=1}^{N-1} u_j L_j(\xi_i) L_k(\xi_i)) \;=\; \sum_{i=1}^{N-1} \alpha_i f(\xi_i) L_k(\xi_i) \tag{2.83}$$

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

$$\sum_{j=1}^{N-1} u_j (\sum_{i=1}^{N-1} \alpha_i L_j'(\xi_i) L_k'(\xi_i) + \lambda \alpha_k \delta_{jk}) \;=\; \alpha_k f_k \tag{2.84}$$

dont la matrice est clairement symétrique en j et k.

2.4.4 Résolution de problèmes elliptiques par diagonalisation

On se propose de résoudre

$$\nabla^2 u - \lambda u \;=\; s \text{ dans } \Omega \tag{2.85}$$

$$u \;=\; 0 \text{ sur } \partial\Omega \tag{2.86}$$

où $\Omega =]-1,1[^2$.
Il existe une méthode directe efficace pour résoudre ce système linéaire. Nous ne traiterons que le cas d'une méthode de collocation.
La résolution de l'équation de Helmholtz $Hu = f$ est effectuée par une méthode directe de bi-diagonalisation inspirée de l'algorithme proposé par Haidvogel et Zang (Journal of Computational Physics 1979).

Une dimension

Considérons le problème en une dimension:

$$D^2u - \lambda u = S \qquad \text{on } [-1,1] \tag{2.87}$$

avec les conditions aux limites

$$\alpha_+ u(1) + \beta_+ u'(1) = g_+ \tag{2.88}$$
$$\alpha_- u(-1) + \beta_- u'(-1) = g_- \tag{2.89}$$
$$\tag{2.90}$$

Si $\alpha_+ > 0, \beta_+ > 0, \alpha_- > 0, \beta_- > 0$ le problème a des valeurs propres réelles et strictement négatives. Une extension est possible pour le cas

- des conditions de Dirichlet $\beta_- = \beta_+ = 0$
- des conditions Dirichlet-Neumann $\beta_{-(+)} = \alpha_{+(-)} = 0$
- des conditions de Neumann $\alpha_- = \alpha_+ = 0$ (mais dans ce cas une des valeurs propres sera zéro).

On a

$$D^2 = P\Lambda P^{-1} \tag{2.91}$$

Le problème $Hu = s$ peut donc s'écrire

$$(P\Lambda^{-1}P^{-1} - \lambda I)u = s \tag{2.92}$$

En multipliant à gauche par P^{-1}, en utilisant $I = PP^{-1}$ et en réarrangeant, on trouve

$$(\Lambda^{-1}P^{-1} - \lambda P^{-1})u = P^{-1}s \tag{2.93}$$

On pose $v = P^{-1}u$ et $g = P^{-1}s$. L'équation de Helmholtz peut donc être résolue dans la base des vecteurs propres

$$(\Lambda^{-1} - \lambda)v = g \tag{2.94}$$

Si la matrice Λ est de taille N, la résolution du système consiste en N divisions scalaires.
Il suffit ensuite d'effectuer $u = Pv$.

Extension à deux dimensions:

On cherchera u dans l'espace d'approximation $P_N(x)\otimes P_M(z)$. Dans une méthode de collocation, les inconnues sont les valeurs u_{ij} aux points de collocation de Gauss-Lobatto. La solution est recherchée sous la forme de son polynôme d'interpolation appartenant à $P_N \otimes P_M$, $I_{NM}u = \sum_{i=0}^{N}\sum_{j=0}^{M} f_{ij}L_i(x)L_j(z)$, 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 $I_{NM}u$ est évidemment caractérisé par $(N+1)(M+1)$ valeurs u_{ij} et il nous faut donc autant d'équations. Celles ci sont obtenues en demandant que $I_{NM}u$ 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 $R_{NM} = (I_{NM}u)'' - \lambda I_{NM}u - s$ en tous les points (x_i,z_j),$i = 1,..,N-1$,$j = 1,..,M-1$, soit

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

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

$$\mathcal{D}_N\mathcal{F} + \mathcal{F}\mathcal{D}_M^t - \lambda\mathcal{F} = \mathcal{S} \tag{2.95}$$

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.
Le principe de la résolution repose sur le fait que la matrice $\mathcal{D}_\mathcal{N}$ est diagonalisable sur $\mathcal{IR}$ $\mathcal{D}_\mathcal{N} = \mathcal{P}\Lambda_N\mathcal{P}^{-1}$ et de même $\mathcal{D}_\mathcal{M} = \mathcal{Q}\Lambda_M\mathcal{Q}^{-1}$.
En posant $\mathcal{U}= \mathcal{P}^{-1}\mathcal{F}\ \mathcal{Q}^{-1t}$, la résolution se ramène à

$$\mathcal{U}_{ij} = \frac{(\mathcal{P}^{-1}\mathcal{S}\mathcal{Q}^{-1t})_{ij}}{\lambda_i+\lambda_j-\lambda} \tag{2.96}$$

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

Généralisation en trois dimensions

L'équation de Helmholtz en trois dimensions peut se mettre sous la forme suivante:

$$(I_x\otimes I_y\otimes D_z^2 + I_x\otimes D_y^2\otimes I_z + D_x^2\otimes I_y\otimes I_z - \lambda I_x\otimes I_y\otimes I_z)u = s \tag{2.97}$$

La matrice

$$H = I_x\otimes I_y\otimes D_z^2 + I_x\otimes D_y^2\otimes I_z + D_x^2\otimes I_y\otimes I_z - \lambda I_x\otimes I_y\otimes I_z$$

est de dimension $N_xN_yN_z$. On suppose que les opérateurs D_x^2,D_y^2,D_z^2 diagonalisent respectivement dans les bases P,Q,R de sorte que

$$D_x^2 = P\Lambda_xP^{-1}, D_y^2 = Q\Lambda_yQ^{-1}, D_z^2 = R\Lambda_zR^{-1}$$

avec $\Lambda_{x(y,z)}$ des matrices diagonales.
On va multiplier (2.97) à gauche par $P \otimes Q \otimes R$ et à droite par $R^{-1} \otimes Q^{-1} \otimes P^{-1}$. En utilisant l'égalité tensorielle

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \tag{2.98}$$

on peut montrer que H peut sécrire comme

$$H = P \otimes Q \otimes R\Lambda P^{-1} \otimes Q^{-1} \otimes R^{-1} \tag{2.99}$$

où

$$\Lambda = I_x \otimes I_y \otimes \Lambda_z + I_x \otimes \Lambda_y \otimes I_z + D_x \otimes I_y \otimes \Lambda_z - \lambda I_x \otimes I_y \otimes I_z$$

L'idée est donc de calculer le second membre $g = P^{-1} \otimes Q^{-1} \otimes R^{-1} f$ d'inverser pour v le système diagonal

$$\Lambda^{-1} v = g$$

et de trouver la solution u à partir de

$$u = P \otimes Q \otimes Rv.$$

L'intérêt d'une telle approche pour la résolution d'un problème instationnaire est que l'opérateur est inversé une seule fois au début du calcul.
Remarques

1. 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.
2. On peut également diagonaliser dans deux directions seulement (diagonalisation partielle) et résoudre un système tridiagonal pour la dernière direction, ce qui peut s'avérer plus rapide (et éventuellement moins précis...)
3. L'opérateur n'est pas diagonalisable (valeurs propres réelles) dans le cas de l'équation d'advection-diffusion (première dérivée de la vitesse). On peut dans ce cas utiliser une méthode de complément de Schur qui nous permet d'obtenir un système triangulaire.

Chapitre 3

L'erreur de discrétisation

3.1 Schémas numériques: notions de stabilité, convergence et consistence

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

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

discrétisée comme

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

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

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

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

$$E_i^n = u(x_i, t^n) - u_i^n \tag{3.4}$$

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.

Théorême de Lax: Si le problème est linéaire, le schéma consistent 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 ϵ 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 dfinit le gain entre deux pas de temps pour chaque mode de Fourier (fréquence spatiale k)

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

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

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{3.6}$$

– discrétisation explicite - schéma spatial centré -

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

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

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

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

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

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

– discrétisation explicite - schéma spatial décentré -

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

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

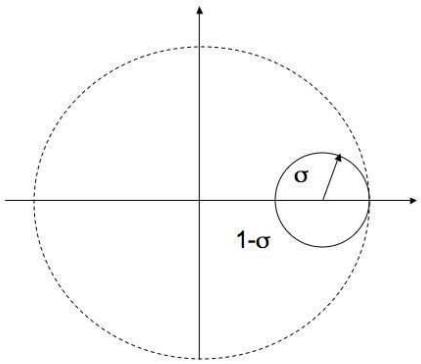


FIG. 3.1 – *Représentation de la fonction gain G pour l'équation de convection discrétisée*

On appelle σ 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 -

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

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

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

Exemple 2 - Equation de la chaleur

$$\frac{\partial u}{\partial t} = \alpha\frac{\partial u}{\partial x^2} \tag{3.13}$$

Discrétisation explicite - schéma centré

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

On définit

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

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

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

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

Discrétisation explicite - schéma centré pour la convection et la diffusion

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

En définissant

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

Le gain G s'exprime comme

$$G = 1 - i\sigma sink\Delta x + 2\beta(cosk\Delta x - 1) \tag{3.18}$$

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

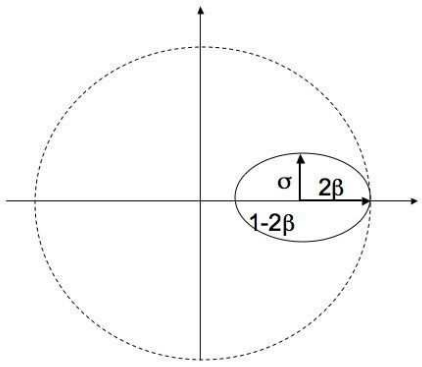


FIG. 3.2 – *Représentation de la fonction gain G pour l'équation d'advection-diffusion discrétisée*

ce qui conduit à la condition

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

Il est utile de définir le rapport σ/β

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

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

$$\|C\| = Max_u \frac{|Cu|}{|u|}$$

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

$$\rho(C) = Max_j|\lambda_j|$$

Propriété: $\|C\| \geq \rho_C$

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

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

$$C = [\begin{pmatrix} 1-2\beta & \beta - \frac{\sigma}{2} & 0 & \dots \\ \beta + \frac{\sigma}{2} & 1-2\beta & \beta - \frac{\sigma}{2}0\dots & \\ 0 & \beta + \frac{\sigma}{2} & 1-2\beta & \beta - \frac{\sigma}{2}0\dots \\ \dots & \dots & \dots & \dots \end{pmatrix}] \tag{3.20}$$

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

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

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,σ). Elle montre que l'analyse globale conduit à un critère beaucoup moins restrictif sur σ en fonction du nombre de mailles.

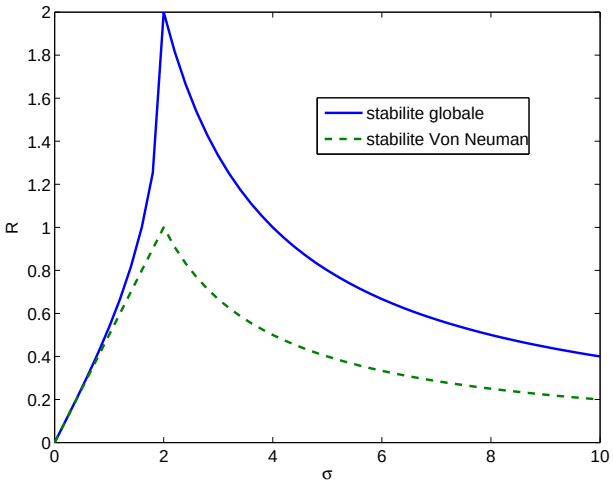


FIG. 3.3 – *Comparaison de la méthode de Von Neumann et l'analyse matricielle pour le critère de stabilité de l'équation d'advection-diffusion discrétisée*

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

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

Le problème est mal conditionné quand κ 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

$$A = U^T DV$$

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

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

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

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

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

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

qui sont les modes de Fourier. Le mode de Fourier k de l'erreur va donc décroître avec un coefficient λ_k.
On voit que pour les petits nombres d'onde k, correspondant à de basses fréquences (spatiales), la valeurs propre est proche de 1:

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

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 λ_k sont de la forme

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

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 λ_k^{GS} sont de la forme

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

.
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 impostant 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. Arpès 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
   $r_{2h}(x_j) = r_h(x_j)$ 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 $v_h \to v_{2h}$.
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):
   $e_h(x_{2j}) = e_{2h}(x_j)$
   $e_h(x_{2j+1}) = \frac{e_{2h}(x_j)+e_{2h}(x_{j+1})}{2}$ 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).

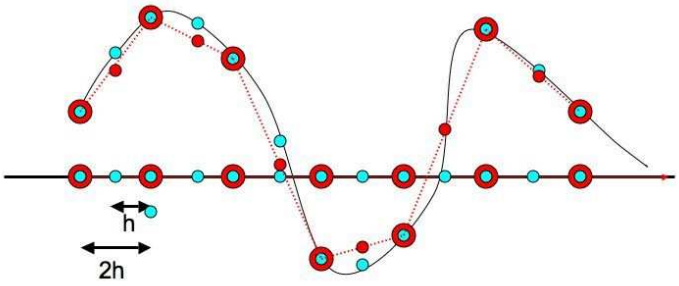


FIG. 3.4 – *Interpolation linéaire d'une fonction de la grille large vers la grille fine*

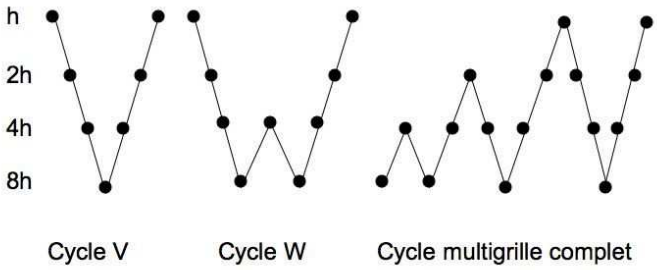


FIG. 3.5 – *Cycles de multi-grille*

Chapitre 4

Résolution du problème de Stokes

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

$$\nabla^2 u^{n+1} - \lambda u^{n+1} = -Su + \frac{\partial P}{\partial x} \text{ dans } \Omega \tag{4.1}$$

$$\nabla^2 w^{n+1} - \lambda w^{n+1} = -Sw + \frac{\partial P}{\partial z} \text{ dans } \Omega \tag{4.2}$$

$$\frac{\partial u^{n+1}}{\partial x} + \frac{\partial w^{n+1}}{\partial z} = 0 \text{ dans } \Omega \tag{4.3}$$

$$u^{n+1} = 0 \text{ sur } \partial\Omega \tag{4.4}$$

$$w^{n+1} = 0 \text{ sur } \partial\Omega \tag{4.5}$$

où $\lambda = \frac{1.5Re}{\Delta t}$.
Les équations (4.1,4.2,4.3, 4.4, 4.5) 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 trois grandes familles de méthodes pour la résolution approchée de ce problème de Stokes instationnaire. La première repose sur une approche découplée, basée sur l'introduction d'une équation de Poisson pour la pression (technique de matrice d'influence), tandis que la deuxième 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. Enfin la troisième famille de méthodes repose sur une approche à pas de temps fractionnaire et la détermination d'une correction de pression.

4.1 Equation de Poisson pour la Pression - Matrice d'influence

4.1.1 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 (4.1,4.2) s'écrit donc;

$$(\nabla^2 - \lambda)\mathcal{D}^{n+1} = \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} - (\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}) \tag{4.6}$$

où $\mathcal{D}$ est la divergence du champ de vitesse.
Il apparait donc immédiatement que l'équation de poisson

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = (\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}) \tag{4.7}$$

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

$$(\nabla^2 - \lambda)\mathcal{D}^{n+1} = 0 \tag{4.8}$$

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;

$$\begin{aligned}
\nabla^2 u - \lambda u &= -Su + \frac{\partial P}{\partial x} \text{ dans } \Omega && (4.9)\\
\nabla^2 w - \lambda w &= -Sw + \frac{\partial P}{\partial z} \text{ dans } \Omega && (4.10)\\
\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} &= (\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}) \text{ dans } \Omega && (4.11)\\
u^{n+1} &= 0 \text{ sur } \partial\Omega && (4.12)\\
w^{n+1} &= 0 \text{ sur } \partial\Omega && (4.13)\\
\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} &= 0 \text{ sur } \partial\Omega && (4.14)
\end{aligned}$$

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:

$$\begin{aligned}
\nabla^2 u - \lambda u &= -Su + \frac{\partial P}{\partial x} \text{ dans } \Omega && (4.15)\\
\nabla^2 w - \lambda w &= -Sw + \frac{\partial P}{\partial z} \text{ dans } \Omega && (4.16)\\
\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} &= (\frac{\partial Su}{\partial x} + \frac{\partial Sw}{\partial z}) && (4.17)\\
u^{n+1} &= 0 \text{ sur } \partial\Omega && (4.18)\\
w^{n+1} &= 0 \text{ sur } \partial\Omega && (4.19)\\
P &= \bar{P} \text{ sur } \partial\Omega && (4.20)
\end{aligned}$$

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 $\overline{P}$ qui garantisse $\overline{\mathcal{D}} = 0$.
Ceci peut être fait simplement en raison de la linéarité de l'application qui à $\overline{P}$ fait correspondre $\overline{\mathcal{D}}$. (cette correspondance est linéaire dans le problème homogène (Su,Sw) = (0,0), affine sinon.)
Cette approche constitue le fondement des méthodes dites de matrice d'influence.
La mise en œuvre de cette méthode demande que l'on ait fait un choix des espaces d'approximation.

4.1.2 Discrétisation Fourier-Chebyshev

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. On se ramène à deux dimensions x et y où y est la direction normale à la paroi. Les champs de vitesse et de pression sont écrits sous la forme

$$\begin{aligned}
u &= \sum_k u_k(y,t) e^{2i\pi kx} && (4.21)\\
v &= \sum_k v_k(y,t) e^{2i\pi kx} && (4.22)\\
p &= \sum_k p_k(y,t) e^{2i\pi kx} && (4.23)\\
& && (4.24)
\end{aligned}$$

On omettra à partir de maintenant l'indice k.
On note D l'opérateur discrétisé dans la direction normale. Le problème peut être mis sous la

forme:

$$\begin{aligned} D^2u - \sigma u + ikp &= f_u && (4.25)\\ D^2v - \sigma v + Dp &= f_v && (4.26)\\ iku + Dv &= 0 && (4.27)\\ u(-1) &= u_- && (4.28)\\ u(1) &= u_+ && (4.29)\\ v(-1) &= v_- && (4.30)\\ v(1) &= v_+ && (4.31) \end{aligned}$$

L'équation pour la divergence peut être remplacée pour une équation de Helmholtz pour la pression

$$D^2p - k^2p = ikf_u + Df_v \tag{4.32}$$

La difficulté est de déterminer les conditions aux limites adéquates pour cette équation. Le problème initial peut être découplé en deux sous-problèmes, l'un pour (v,p) et l'autre pour u. En utilisant à nouveau l'équation de la divergence dans les conditions aux limites, le problème pour (v,p) s'exprime de la manière suivante:

$$\begin{aligned} D^2v - \sigma v + Dp &= f_v && (4.33)\\ D^2p - k^2p &= ikf_u + Df_v && (4.34)\\ v'(-1) &= 0 && (4.35)\\ v'(1) &= 0 && (4.36)\\ v(-1) &= v_- && (4.37)\\ v(1) &= v_+ && (4.38) \end{aligned}$$

On notera A ce problème. Le problème pour u est un problème de Helmholtz standard qui ne pose pas de problème de résolution majeur une fois que p est connu:

$$\begin{aligned} D^2u - \sigma u + ikp &= f_u && (4.39)\\ u(-1) &= u_- && (4.40)\\ u(1) &= u_+ && (4.41) \end{aligned}$$

L'idée est d'écrire la solution (v,p) du problème A comme la solution d'un problème B équivalent:

$$\begin{aligned} D^2v - \sigma v + Dp &= f_v && (4.42)\\ D^2p - k^2p &= ikf_u + Df_v && (4.43)\\ v'(-1) &= 0 && (4.44)\\ v'(1) &= 0 && (4.45)\\ p(-1) &= p_- && (4.46)\\ p(1) &= p_+ && (4.47) \end{aligned}$$

Ce problème B est parfaitement défini puisqu'il consiste en deux équations elliptiques (pour chaque nombre d'onde k) munis de conditions aux limites aux frontières. La solution du problème B peut s'écrire comme combinaison linéaire de solutions de trois problèmes différents $\bar{P}$,P_+,P_-

$$\begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} \bar{v} \\ \bar{p} \end{bmatrix} + \lambda^+ \begin{bmatrix} v^+ \\ p^+ \end{bmatrix} + \lambda^- \begin{bmatrix} v^- \\ p^- \end{bmatrix} \tag{4.48}$$

définis comme $\bar{P}$:

$$\begin{aligned} D^2\bar{v} - \sigma\bar{v} + D\bar{p} &= f_v && (4.49)\\ D^2\bar{p} - k^2\bar{p} &= ikf_u + Df_v && (4.50)\\ \bar{p}(-1) &= 0 && (4.51)\\ \bar{p}(1) &= 0 && (4.52)\\ \bar{v}(-1) &= v_- && (4.53)\\ \bar{v}(1) &= v_+ && (4.54) \end{aligned}$$

P_+:

$$\begin{aligned} D^2v^+ - \sigma v^+ + Dp^+ &= f_v && (4.55)\\ D^2p^+ - k^2p &= 0 && (4.56)\\ p^+(-1) &= 0 && (4.57)\\ p^+(1) &= 1 && (4.58)\\ v^+(-1) &= 0 && (4.59)\\ v^+(1) &= 0 && (4.60) \end{aligned}$$

P_-:

$$\begin{aligned} D^2v^- - \sigma v^- + Dp^- &= 0 && (4.61)\\ D^2p^- - k^2p^- &= 0 && (4.62)\\ p^-(-1) &= 1 && (4.63)\\ p^-(1) &= 0 && (4.64)\\ v^-(-1) &= 0 && (4.65)\\ v^-(1) &= 0 && (4.66) \end{aligned}$$

Les constantes λ^+ et λ^- sont déterminées à partir de la condition

$$\begin{aligned} v'(-1) &= 0 && (4.67)\\ v'(1) &= 0 && (4.68) \end{aligned}$$

qui conduit à la résolution d'un système 2x2:

$$\begin{bmatrix} v'^+(-1) & v'^-(-1) \\ v'^+(1) & v'^-(1) \end{bmatrix} \begin{bmatrix} \lambda^+ \\ \lambda^- \end{bmatrix} = \begin{bmatrix} -\bar{v}(-1) \\ -\bar{v}(1) \end{bmatrix} \tag{4.69}$$

La matrice

$$M = \begin{bmatrix} v'^{+}(-1) & v'^{-}(-1) \\ v'^{+}(1) & v'^{-}(1) \end{bmatrix},$$

son inverse M^{-1} et les solutions de $P-+$ et P_- peuvent être calculées préliminairement et stockées. L'algorithme est donc de résoudre le problème $\bar{P}$, de calculer les coefficients λ^+ et λ^- (produit matrice-vecteur), et de résoudre le problème pour u.
Toutefois la méthode présentée ici n'est pas tout à fait satisfaisante car dans le cas discret, l'équation de conservation de la quantité de mouvement n'est pas satisfaite aux frontières, ce qui induit une erreur pour l'expression de la divergence.

$$\begin{aligned} D^2u - \sigma u + ikp &= f_u + r_u && (4.70)\\ D^2v - \sigma v + Dp &= f_v + r_v && (4.71)\\ iku + Dv &= 0 && (4.72)\\ u(-1) &= u_- && (4.73)\\ u(1) &= u_+ && (4.74)\\ v(-1) &= v_- && (4.75)\\ v(1) &= v_+ && (4.76) \end{aligned}$$

où r_u et r_v sont des résidus qui s'annulent partout sauf aux bords. La version discrète du problème est alors:

$$\begin{aligned} D^2u_j - \sigma u_j + jkp_j &= f_u^j + r_u^j, j = 0, \ldots, N && (4.77)\\ D^2v_j - \sigma v_j + Dp_j &= f_v^j + r_v^j, j = 0, \ldots, N && (4.78)\\ iku_j + Dv_j &= 0, j = 0, \ldots, N && (4.79)\\ u_0 &= u_- && (4.80)\\ u_N &= u_+ && (4.81)\\ v_0 &= v_- && (4.82)\\ v_N &= v_+ && (4.83) \end{aligned}$$

avec $r_u^j = 0$ et $r_v^j = 0$, for $j = 1, \ldots, N-1$.
Le problème B discrétisé est alors sous la forme:

$$\begin{aligned} D^2v - \sigma v + Dp &= f_v && (4.84)\\ D^2p - k^2p &= ikf_u + Df_v + D\gamma && (4.85)\\ v(-1) &= v_- && (4.86)\\ v(1) &= v_+ && (4.87)\\ p(-1) &= p_- && (4.88)\\ p(1) &= p_+ && (4.89) \end{aligned}$$

où $\gamma = r_v$ est le résidu de l'équation pour v. L'élément-clé est que $D\gamma$ n'est pas nul sur les points intérieurs au domaine!
Pour déterminer ce résidu, on va utiliser la même technique de matrice d'influence que ci-dessus. On définit σ_- comme la fonction discrète telle que
$\sigma_-(x_0, \ldots, x_{N-1}) = 0$ et $\sigma_-(x_N) = 1$ et
σ_+ telle que $\sigma_+(x_1, \ldots, x_N) = 0$ et $\sigma_+(x_0) = 1$.
La solution du problème B modifié peut s'écrire comme la somme de trois solutions

$$\begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} \bar{v*} \\ \bar{p*} \end{bmatrix} + \lambda^{*+} \begin{bmatrix} v^{*+} \\ p^{*+} \end{bmatrix} + \lambda^{*-} \begin{bmatrix} v^{*-} \\ p^{*-} \end{bmatrix} \tag{4.90}$$

avec

$$\begin{aligned} D^2\bar{v}* - \sigma\bar{v}* - D\bar{p}* &= 0 && (4.91)\\ D^2\bar{p}* - k^2\bar{p}* &= ikf_u + Df_v && (4.92)\\ \bar{p}*(-1) &= p_- && (4.93)\\ \bar{p}*(1) &= p_+ && (4.94)\\ \bar{v}*(-1) &= v_- && (4.95)\\ \bar{v}*(1) &= v_+ && (4.96) \end{aligned}$$

et

$$\begin{aligned} D^2v^{*+} - \sigma v^{*+} - Dp^{*+} &= D\sigma_+ && (4.97)\\ D^2p*+ - k^2p^{*+} &= 0 && (4.98)\\ p^{*+}(-1) &= 0 && (4.99)\\ p^{*+}(1) &= 0 && (4.100)\\ v^{*+}(-1) &= 0 && (4.101)\\ v^{*+}(1) &= 0 && (4.102) \end{aligned}$$

et

$$\begin{aligned} D^2v^{*-} - \sigma v^{*-} - Dp^{*-} &= D\sigma_- && (4.103)\\ D^2p*- - k^2p^{*-} &= 0 && (4.104)\\ p^{*-}(-1) &= 0 && (4.105)\\ p^{*-}(1) &= 0 && (4.106)\\ v^{*-}(-1) &= 0 && (4.107)\\ v^{*-}(1) &= 0 && (4.108) \end{aligned}$$

Les coefficients λ^{*+} et λ^{*-} sont ensuite obtenus en imposant que l'équation de quantité de mouvement pour la vitesse normale soit nulle à la frontière

$$D^2(v_0 + \lambda^{*+}v_0^{*+} + \lambda_0^{*-}v^{*-}) - \sigma(v_0 + \lambda^{*+}v_0^{*+} + \lambda^{*-}v_0^{*-}) - D(p_0 + \lambda^{*+}p_0^{*+} + \lambda^{*-}p_0^{*-}) = f_v^0,$$

$$D^2(v_N + \lambda^{*+}v_N^{*+} + \lambda_N^{*-}v^{*-}) - \sigma(v_N + \lambda^{*+}v_N^{*+} + \lambda^{*-}v_N^{*-}) - D(p_N + \lambda^{*+}p_N^{*+} + \lambda^{*-}p_N^{*-}) = f_v^N.$$

En une dimension, ceci représente donc six équations d'Helmholtz à résoudre. La dimension totale de la matrice d'influence est 4. On peut remarquer que la méthode présentée revient à utiliser la fameuse formule de Shermann-Morrison-Woodbury (voir annexe A) dans le cas d'une discrétisation de type tau.

4.1.3 Discrétisation Chebyshev-Chebyshev

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 $\mathcal{GL}_x^N = \left\{x_i, x_i = \cos(\frac{i\Pi}{N}), i = 0,..,N\right\}$ 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

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial z^2} = 0 \tag{4.109}$$

$$P = 1 \text{ sur } \partial\Omega \tag{4.110}$$

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 aux 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 thèse d'Etat de Haldenwang)
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 $z=1$ et $z=-1$, 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 $I\!P_N$ au zéro machine.
Une procédure similaire peut être également mise en œuvre lorsque les problèmes de Helmholtz sont résolus dans le formalisme tau (Tuckerman, JCP 1989).

4.2 Opérateur d'Uzawa

4.2.1 Principe

Un deuxième type de méthode de résolution du problème de Stokes consiste à considérer l'opérateur d'Uzawa. L'algorithme d'Uzawa repose sur l'élimination de la vitesse dans le problème

de Stokes instationnaire.
En notant $\mathcal{HU}$ et $\mathcal{HW}$ 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:

$$\mathcal{HU}u = \frac{\partial P}{\partial x} - Su \tag{4.111}$$

$$\mathcal{HW}w = \frac{\partial P}{\partial z} - Sw \tag{4.112}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0 \tag{4.113}$$

complété par les conditions aux limites pour u et w.
On peut inverser formellement (4.111) et (4.112) pour obtenir

$$u^{n+1} = \mathcal{HU}^{-1}\frac{\partial P}{\partial x} - \mathcal{HU}^{-1}Su \tag{4.114}$$

$$w^{n+1} = \mathcal{HW}^{-1}\frac{\partial P}{\partial z} - \mathcal{HW}^{-1}Sw \tag{4.115}$$

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

$$(\frac{\partial}{\partial x}\mathcal{HU}^{-1}\frac{\partial}{\partial x} + \frac{\partial}{\partial z}\mathcal{HW}^{-1}\frac{\partial}{\partial z})P = (\frac{\partial}{\partial x}\mathcal{HU}^{-1}Su + \frac{\partial}{\partial z}\mathcal{HW}^{-1}Sw) \tag{4.116}$$

et nous noterons $\mathcal{U} = \frac{\partial}{\partial x}\mathcal{HU}^{-1}\frac{\partial}{\partial x} + \frac{\partial}{\partial z}\mathcal{HW}^{-1}\frac{\partial}{\partial z}$. 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 (4.111,4.112), 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.

4.2.2 Discrétisation Chebyshev dans une direction

On se place dans le cas d'une discrétisation Fourier-Chebyshev (le cas 3D avec deux expansions de Fourier est similaire). La formulation du problème est alors:

$$u'' - \sigma u + kp = f_u, \qquad -1 < y < 1 \tag{4.117}$$

$$v'' - \sigma v + p' = f_v, \qquad -1 < y < 1 \tag{4.118}$$

$$iku + v' = 0, \qquad -1 < y < 1 \tag{4.119}$$

$$u(-1) = u_- \tag{4.120}$$

$$u(1) = u_+ \tag{4.121}$$

$$v(-1) = v_- \tag{4.122}$$

$$v(1) = v_+ \tag{4.123}$$

Pour une approche de collocation en Chebyshev, on définit :

$$\tilde{u} = [u_1, \ldots, u_{N-1}], \tilde{v} = [v_1, \ldots, v_{N-1}],$$

$$\tilde{f_u} = [f_u(x_1), \ldots, f_u(x_{N-1})], \tilde{f_v} = [f_v(x_1), \ldots, f_v(x_{N-1})],$$

$$u = [u_0, \ldots, u_N], v = [v_0, \ldots, v_N].$$

Soit D l'opérateur de collocation Chebyshev discrétisé. On a alors

$$\begin{aligned}
D^2\tilde{u} - \sigma\tilde{u} + k\tilde{p} &= \tilde{f}_u, && (4.124)\\
D^2\tilde{v} - \sigma\tilde{v} + D\tilde{p} &= \tilde{f}_v, && (4.125)\\
iku + Dv &= 0 && (4.126)\\
u_0 &= u_- && (4.127)\\
u_N &= u_+ && (4.128)\\
v_0 &= v_- && (4.129)\\
v_N &= v_+ && (4.130)
\end{aligned}$$

Ce système peut se mettre sous la forme

$$\begin{aligned}
\mathcal{HU}u &= \frac{\partial P}{\partial x} - Su && (4.131)\\
\mathcal{HV}v &= \frac{\partial P}{\partial z} - Sv && (4.132)\\
\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} &= 0 && (4.133)
\end{aligned}$$

avec

$$\mathcal{HU} = D^2 - \sigma$$

$$\mathcal{HV} = D^2 - \sigma$$

$$Su = \tilde{f}_u$$

$$Sv = \tilde{f}_v$$

La matrice $\mathcal{H}$ est une approximation discrète des équations d'Helmholtz avec des conditions de Dirichlet. Ses valeurs propres sont donc réelles et négatives. La matrice $\mathcal{U}$ a des valeurs propres réelles et positives sauf dans le cas $k = 0$. Le conditionnement de $\mathcal{U}$ varie lentement avec σ pour k et N fixé. Les matrices $\mathcal{U}$ et $\mathcal{H}$ sont inversibles en une dimension.
Pour $k = 0$, la matrice $\mathcal{U}$ a deux valeurs propres nulles. L'une correspond au mode constant, l'autre représente le mode $T_N(y)$ qui n'apparait pas dans les équations (car $T'_N(y_j) = 0$ pour $1 < j < N - 1$) et qui conduit à des oscillations pour le mode de pression (la vitesse n'est pas contaminée). Si on désire obtenir la pression correcte, il faut éliminer ce mode parasite soit

en le filtrant, soit en utilisant des maillages décalés, soit en diminuant l'ordre de la base dans laquelle on recherche la pression ($\mathcal{P}_{\mathcal{N}-\epsilon}$).
En plusieurs dimensions, l'application de cette méthode devient problématique car une résolution directe est impossible. Le principe d'une approche itérative (Le Marec et al. 1996) est le suivant:

– initialisation à partir de P^0 connu

$$\begin{aligned} \mathcal{H}\mathcal{U}^0 &= S^0 u - ikP^0 \\ \mathcal{H}\mathcal{V}^0 &= S^0 v + DP^0 \\ R^0 &= ikU^0 + DV^0 - S_Q \end{aligned}$$

et

$$\begin{aligned} -(D_\gamma - k^2)\Phi^0 &= R^0 \\ W^0 &= R^0 + \sigma^0 \Phi^0 \end{aligned}$$

où D_γ est l'opérateur du second ordre muni des conditions aux limites de Neumann.
On va utiliser l'opérateur $(I - \sigma^0(D_\gamma - k^2)^{-1}))^{-1}$ comme un préconditionneur pour l'itération de pression $P^{m+1} - P^m$.

– itération

$$\begin{aligned} \mathcal{H}\mathcal{U}^m &= -ikW^m \\ \mathcal{H}\mathcal{V}^m &= DW^m \\ R^m &= ikU^m + DV^m \end{aligned}$$

avec une mise à jour (update)

$$\begin{aligned} P^{m+1} &= P^m - \alpha^m W^m \\ U^{m+1} &= U^m - \alpha^m U^m \\ V^{m+1} &= V^m - \alpha^m V^m \\ R^{m+1} &= R^m - \alpha^m R^m \end{aligned}$$

et

$$\begin{aligned} -(D_\gamma - k^2)\Phi^m &= R^m \\ W^{m+1} &= W^m - \alpha^m (R^m + \sigma^0 \Phi^m) \end{aligned}$$

– l'itération s'arrête lorsque la divergence R^m est suffisamment petite.

4.2.3 Discrétisation Chebyshev dans deux directions

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 $\mathcal{GL}_x^N = \left\{x_i, x_i = \cos(\frac{i\Pi}{N}), i = 0,..,N\right\}$ 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 $Ker(\mathcal{U})$ 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 $(x_i, z_j), 1 \leq i \leq N-1 \times 1 \leq j \leq M-1$
- 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. [1] 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.

1. 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.

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

L'élimination totale des modes parasites de pression peut être obtenue 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 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. 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.

Mise en œuvre

Les difficultés à résoudre pour intégrer le système d'équations (4.131, 4.132, 4.4, 4.5, 4.116) sont les suivantes:

- inversion de $\mathcal{HU} f = s$ (voir section 2.4.4)
- choix de l'algorithme à utiliser pour résoudre l'équation de pression (4.116)

Résolution de l'opérateur d'Uzawa pour la pression

Deux types de résolution peuvent être envisagées:

1. <u>résolution directe</u>
   L'inversion directe de (4.116) demande que dans un premier temps on ait explicité la matrice correspondant à l'opérateur $\mathcal{U} = (\frac{\partial}{\partial x}\mathcal{HU}^{-1}\frac{\partial}{\partial x} + \frac{\partial}{\partial z}\mathcal{HW}^{-1}\frac{\partial}{\partial z})$, Comme nous l'avons déjà mentionné, ceci peut être fait en faisant décrire à la pression la base canonique de $\mathcal{R}^{N.M}$ et en évaluant $(\frac{\partial}{\partial x}\mathcal{HU}^{-1}\frac{\partial}{\partial x} + \frac{\partial}{\partial z}\mathcal{HW}^{-1}\frac{\partial}{\partial z})P_{ij}$, 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×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.

2. résolution itérative Lorsque la taille du système devient grande, il peut sembler déraisonnable a priori de chercher à résoudre l'équation (4.116) 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 2.4.4 pour résoudre $\mathcal{HU}f = s$ et $\mathcal{HW}f = s$.
   Nous avons testé 3 algorithmes itératifs différents pour la résolution de (4.116):
   - une méthode itérative de Richardson qui s'écrit:

$$p^{k+1} = p^k - \alpha(\mathcal{U}p^k - s) \tag{4.134}$$

     α étant déterminé de manière à obtenir la convergence la plus rapide possible. Si l'opérateur $\mathcal{U}$ a ses valeurs propres réelles comprises entre μ_{max} et μ_{min}, il est connu que la valeur optimale de α est $\frac{2}{\mu_{max}+\mu_{min}}$.
   - une méthode de Richardson avec résidu minimum où α_k est choisi à chaque itération de manière à minimiser le résidu à l'itération k+1
   - une méthode de gradient conjugué

   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 λ du problème de Helmoltz est nulle, le spectre de cet opérateur n'est formé que de 1 (à part la valeur propre nulle correspondant au mode propre constant). On constate en revanche que le conditionnement de l'opérateur $\mathcal{U}$ se détériore lorsque λ 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 (4.1, 4.2, 4.3), et alors, en raison du mauvais conditionnement de l'opérateur $\mathcal{U}$, la résolution itérative de (4.116) à l'aide d'un des algorithmes présentés ci-dessus devient inefficace.
   Ces méthodes itératives perdant leur efficacité lorsque λ augmente, il faut alors nécessairement introduire un préconditionnement tel que celui proposé dans la section précédente en une dimension. La détermination d'un tel préconditionneur reste encore un problème ouvert dans le cas d'une discrétisation spatiale de type spectral.

4.3 Méthode de correction de pression

La base de la méthode de projection (aussi appelée méthode de 'splitting') introduite par Chorin (1968) et Temam (1969) est d'utiliser un pas de temps fractionnaire. Dans une première étape,

le champ de vitesse est avancé en temps, mais sa divergence n'est pas nécessairement nulle. Dans une deuxième étape, le champ est corrigé de manière à vérifier les équations de Navier-Stokes complètes.
Nous considérons les équations de Navier-Stokes sous forme vectorielle:

$$\begin{aligned} \partial_t V - \nu\Delta V + \nabla p &= f \text{ dans } \Omega && (4.135)\\ \nabla.V &= 0 \text{ dans } \Omega && (4.136)\\ V &= V_\gamma \text{ sur } \partial\Omega && (4.137) \end{aligned}$$

– **étape de diffusion:** On calcule une vitesse provisoire $\tilde{V}^{n+1}$ en suivant le schéma

$$\frac{1}{2\Delta t}((1+\epsilon)\tilde{V}^{n+1}-2\epsilon V^n-(1-\epsilon)V^{n-1})-\nu\Delta(\theta\tilde{V}^{n+1}+(1-\theta)V^n)+\nabla(\lambda^1 p^n)=\theta f^{n+1}+(1-\theta)f^n \quad (4.138)$$

avec la condition aux limites

$$V^{n+1} = V_\gamma \text{ dans } \partial\Omega \quad (4.139)$$

Cette première étape constitue un problème de Helmholtz.

– **étape de projection:**

$$\begin{aligned} \frac{1}{2\Delta t}((1+\epsilon)(V^{n+1}-\tilde{V}^{n+1})+\nabla(\lambda^2 p^{n+1}+\lambda^3 p^n) &= 0 \text{ dans } \Omega && (4.140)\\ \nabla.V^{n+1} &= 0 \text{ dans } \Omega && (4.141)\\ V^{n+1}.n &= V_\gamma.n \text{ sur } \partial\Omega && (4.142) \end{aligned}$$

Cette deuxième étape est un problème de type div.grad, aussi appelé problème de Darcy. Ce problème est bien posé si la vitesse normale est spécifiée à la frontière du domaine.

L'élimination de $\tilde{V}^{n+1}$ conduit à:

$$\frac{1}{2\Delta t}((1+\epsilon)V^{n+1}-2\epsilon V^n-(1-\epsilon)V^{n-1})-\nu\Delta(\theta V^{n+1}+(1-\theta)V^n)+ \quad (4.143)$$
$$\nabla(\lambda^2 p^{n+1}+(\lambda^1+\lambda^3)p^n)-\frac{2\theta\Delta t}{1+\epsilon}\nabla^2(\nabla(\lambda^2 p^{n+1}+\lambda^3)p^n))=\theta f^{n+1}+(1-\theta)f^n$$

Les coefficients ϵ,θ,λ^1,λ^2,λ^3 caractérisent le schéma et sa précision.
Pour que le schéma soit à l'ordre Δt, on doit avoir:

$$\lambda^1+\lambda^2+\lambda^3=1$$

Pour que le schéma soit à l'ordre $(\Delta t)^2$, on doit avoir de plus:

$$\theta=\frac{\epsilon}{2}=1-\lambda^1-\lambda^3$$

et

$$\frac{\theta}{1+\epsilon}(\lambda^2+\lambda^3)=0$$

Dans la version originale de la méthode de projection, la pression n'apparait pas dans la première étape, on a $\lambda^1 = 0$ et la correction de pression est à l'ordre Δt seulement. Le schéma de différentiation retardé obtenu avec $\epsilon = 2$,$\theta = 1$,$\lambda^2 = -\lambda^3 = 1$, correspond à une erreur en $O((\Delta t)^2)$.
Lors de la discrétisation, la question est de savoir quelles conditions aux limites utiliser pour la deuxième étape. Explicitons cette étape dans le cas d'un discrétisation Fourier-Chebyshev:

$$\frac{1}{2\Delta t}((1+\epsilon)(V^{n+1} - \tilde{V}^{n+1}) + \nabla(\lambda^2 p^{n+1} + \lambda^3 p^n) = 0 \text{ dans } \Omega \quad (4.144)$$

$$\nabla.V^{n+1} = 0 \text{ dans } \Omega \quad (4.145)$$

$$V^{n+1}.n = V_\gamma.n \text{ sur } \partial\Omega \quad (4.146)$$

Si on impose les équations de conservation à l'intérieur du domaine, on voit qu'il manque 4 équations (u et p à la frontière). Les choix possibles pour u sont:

- équation de quantité de mouvement (4.137) en u (UU)
- conditions aux limites pour la vitesse tangentielle (4.139) (UC)

Les choix possibles pour p sont:

- équation de quantité de mouvement normale (4.137), qui conduit à un problème d'Uzawa (VP)
- imposition de la divergence nulle à la frontière (4.138), qui conduit à un problème de Helmholtz (DP)

4.3.1 Discrétisation Fourier-Chebyshev

Ecrivons les quatre options correspondant à ces différents choix dans le cas d'une discrétisation Chebyshev 1D

1. Choix des équations (UU-VP)

$$\frac{1}{2\delta t}(3u_j^{n+1} - 4u_j^n + u_j^{n-1}) \quad (4.147)$$

$$-\nu(D^2 - k^2)u_j^{n+1} - kp_j^{n+1} + \Delta t E_{u,j}^{n+1} = f_{u,j}^{n+1}, j = 1, \ldots, N-1 \quad (4.148)$$

$$\frac{1}{2\delta t}(3v_j^{n+1} - 4v_j^n + v_j^{n-1}) \quad (4.149)$$

$$-\nu(D^2 - k^2)v_j^{n+1} - kp_j^{n+1} + \Delta t E_{v,j}^{n+1} = f_{v,j}^{n+1}, 0 = 1, \ldots, N \quad (4.150)$$

$$ku_j^{n+1} + Dv_j^{n+1} = 0, j = 0, \ldots, N \quad (4.151)$$

$$u_0^{n+1} - \frac{2\Delta t}{3}k(p_0^{n+1} - p_0^n) = u_+ \quad (4.152)$$

$$u_N^{n+1} - \frac{2\Delta t}{3}k(p_N^{n+1} - p_0^n) = u_- \quad (4.153)$$

$$v_0^{n+1} = v_+ \quad (4.154)$$

$$v_N^{n+1} = v_- \quad (4.155)$$

où

$$E_{u,j} = \frac{2}{3}\nu k(\sum_{l=0}^{N} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

et

$$E_{v,j} = \frac{2}{3}\nu k(\sum_{l=1}^{N-1} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

avec $d_{j,l}^2$ l'opérateur de dérivation seconde discrétisé.
$E_{u(v),j}^{n+1}$ représente une erreur d'ordre $O((\Delta t)^2)$ error qui s'annule lorsque le régime stationnaire est atteint. La vitesse tangentielle est aussi caractérisée par une vitesse de glissement en $O((\Delta t)^2)$. Il y a un mode de pression parasite (i.e qui est associé à une divergence nulle même s'il n'est pas nul) $T_N(y)$.

2. Choix des équations (UU-DP) Les équations sont les mêmes que dans le premier cas mais l'erreur est maintenant

$$E_{u,j} = \frac{2}{3}\nu k(\sum_{\mathbf{l=1}}^{\mathbf{N-1}} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

$$E_{v,j} = \frac{2}{3}\nu k(\sum_{l=1}^{N-1} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

3. Choix des équations (UC-VP) L'erreur est

$$E_{u,j} = \frac{2}{3}\nu k(\sum_{l=0}^{N} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

and

$$E_{v,j} = \frac{2}{3}\nu k(\sum_{l=0}^{N} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

La divergence est seulement nulle aux points de collocation intérieurs. Le système est fermé par les conditions aux limites

$$D(p_j^{n+1} - p_j^n) = 0, j = 0, \ldots, N$$

et celles sur v. Il n'y pas de mode de pression parasite.
En notant $u = u^{n+1}$, $v = v^{n+1}$ et $q = p^{n+1} - p^n$, on peut écrire les équations en formulation continue comme:

– Etape 1

$$\tilde{u}'' - (\sigma/\nu + k^2)\tilde{u} = f_u \tag{4.156}$$

$$\tilde{v}'' - (\sigma/\nu + k^2)\tilde{v} = f_v \tag{4.157}$$

$$\tilde{u}(\pm 1) = u_\pm \tag{4.158}$$

$$\tilde{v}(\pm 1) = v_\pm \tag{4.159}$$

– Etape 2

$$\sigma u - ikq = \sigma\tilde{u} \tag{4.160}$$
$$\sigma v + q' = \sigma\tilde{v} \tag{4.161}$$
$$iku + v' = 0 \tag{4.162}$$
$$v(\pm 1) = v_{\pm} \tag{4.163}$$

Ceci nous conduit à résoudre, en éliminant:

$$q'' - k^2 q = \sigma(ik\tilde{u} + \tilde{v}')$$

avec les conditions aux limites de la deuxième équation évaluée aux bornes:

$$q'(\pm 1) = \sigma(\tilde{v}'(\pm 1) - v(\pm 1)) = 0$$

4. Choix des équations (UC-DP)

$$E_{u,j} = \frac{2}{3}\nu k(\sum_{l=1}^{N-1} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

et

$$E_{v,j} = \frac{2}{3}\nu k(\sum_{l=0}^{N} d_{j,l}^2(p_l^{n+1} - p_l^n) - k^2(p_j^{n+1} - p_j^n)$$

L'erreur de troncature est $O((\Delta t)^2)$. Aucun mode de pression parasite n'a été identifié.

4.4 Discrétisation Chebyshev-Chebyshev

Pour enlever les modes de pression parasites, on peut

– soit travailler dans un espace polynomial de dimension plus faible pour la pression
  – avec des grilles décalées
  – en abaissant la dimension de l'espace polynômial de la pression $P_N - P_{N-2}$

Si on recherche la vitesse V_N dans un espace de dimension N (par direction) et la pression p_{N-2} dans un espace de dimension $N-2$, le problème de Darcy pour la pression est alors

$$\sigma\underline{V}_N^{n+1} + \nabla q_{N-2}^{n+1} = \sigma\tilde{\underline{V}}_N^{n+1} \tag{4.164}$$
$$\nabla.\underline{V}_N^{n+1} = 0 \tag{4.165}$$
$$\underline{V}_N^{n+1}.\underline{n} = \tilde{\underline{V}}_N^{n+1}.\underline{n} = 0 \tag{4.166}$$

où

$$p_{N-2}^{n+1} = q_{N-2}^{n+1} + \tilde{p}_{N-2}^{n+1}$$

On remarque qu'il n'y a pas de conditions aux limites supplémentaires sur la pression.

– soit utiliser l'équation de conservation de quantité de mouvement dans la direction normale à la paroi au lieu de l'équation d'incompressibilité pour fermer les équations. On travaille alors dans une approximation $P_N - P_N$ Le problème de Darcy pour la pression est alors transformé en équation de Poisson pour q avec des conditions aux limites de Neumann.

$$\nabla^2 q_N^{n+1} = \sigma \nabla . \underline{\tilde{V}}_N^{n+1} \tag{4.167}$$

$$\partial_n q_N^{n+1} = 0 \tag{4.168}$$

Annexe

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

$$Hu = s \text{ dans } \mathcal{IR}^{\mathcal{N}} \tag{4.169}$$

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,

$$\tilde{H}u = s \text{ dans } \mathcal{IR}^{\mathcal{N}} \tag{4.170}$$

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 Shermann-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

$$(H + UV^t)u = s \tag{4.171}$$

En définissant une inconnue auxiliaire σ, $\sigma = V^t u$, 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 σ

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

En éliminant finalement σ 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 σ. Cette matrice peut être construite en faisant décrire à σ_k la base canonique de $\mathcal{IR}^K$ 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.