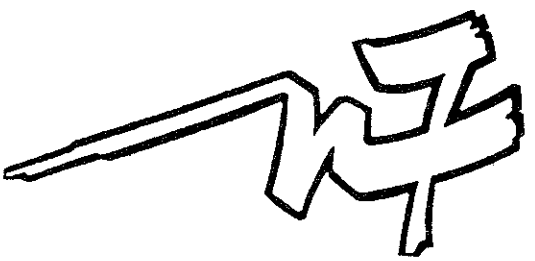
INP TOULOUSE ENSEEIHT

*Département Hydraulique et Mécanique des Fluides*

# Introduction aux méthodes spectrales

## Version 1996

O. THUAL



• INSTITUT NATIONAL POLYTECHNIQUE •
ECOLE NATIONALE SUPERIEURE D'ELECTROTECHNIQUE, D'ELECTRONIQUE, D'INFORMATIQUE ET D'HYDRAULIQUE DE TOULOUSE
2, rue Camichel - 31071 Toulouse Cedex - Tél. : 61 58 82 00 - Fax : 61 62 09 76

# PRÉFACE DE LA VERSION 1996

Le présent polycopié est issu de la note de travail suivante :

- O. Thual et P. L. Sulem, Méthodes spectrales pour des problèmes aux limites simples, *Note de travail de l'Établissement d'Étude et de Recherches Météorologique*, N° **81** (1984).

Ce document a été écrit lors de la construction d'un code numérique de convection. L'objectif était de bien cerner la méthode spectrale à utilisér pour ce code et de la situer dans un cadre un peu plus général. Ce polycopié se limite à une introduction aux méthodes spectrales. Parmi les ouvrages plus complet sur lequels ce document est basé, citons la référence importante suivante :

- D. Gottlieb, S. A. Orszag, Numerical analysis of spectral methods, *NSF-CBMS Monograph* N° **26**, *Soc. Ind. and Appl. Math.*, Philadelphia PA (1977).

Pour un exposé plus complet des méthodes spectrales, on se réferera à des ouvrages plus modernes comme l'ouvrage suivant :

- C. Canuto, M. Y. Hussaini, A. Quarteroni, T. A. Zang, Spectral Methods in Fluid Dynamics, Springer Verlag (1988).

Il m'a paru intéressant de proposer ce polycopié dans le cadre de l'Option "Mécanique des Fluides Numérique" dans la mesure où le contenu est abordable pour un non spécialiste et relativement vite assimilable. Les méthodes spectrales n'étant pas le pivot de la construction des codes industriels (sauf exceptions), cette introduction devrait être suffisante.

Olivier Thual, le 12 septembre 1996

# T A B L E des M A T I E R E S

# T A B L E des M A T I E R E S

# INTRODUCTION

Pendant longtemps les méthodes aux différences finies (en espace et en temps) ont été préférées aux méthodes spectrales. Avec l'augmentation des résolutions accessibles et l'apparition d'algorithmes plus efficaces (exemple: Transformée de Fourier Rapide) les méthodes spectrales sont devenues véritablement compétitives et suscitent un intérêt croissant dans la communauté scientifique.

Efficaces et précises les méthodes spectrales sont particulièrement adaptées à l'étude des phénomènes de bifurcation et de transition vers le chaos (/3/ McLaughlin et Orszag). Les expériences de Mécanique des Fluides réalisées ces dernières années, en particulier par A. Libchaber /12/ et son équipe ont montré qu'il existe dans les systèmes convectifs une très grande richesse de phénomènes dont certains (comme les dédoublements de période) ont pu être observées dans des systèmes à petit nombre de de degré de liberté (E. Lorenz, M. Hénon, M. Feigeunbaum, P. Coullet, C. Tresser). La simulation numérique de phénomènes convectifs 3 D instationnaires présente des difficultés considérables mais a l'avantage de donner accès à l'intégrabilité de la structure spatiale de l'écoulement. Ceci peut être important par exemple pour mettre en évidence des structures cohérentes qui semblent très souvent coexister avec de la turbulence (Busse /13/).

Un effort de développement de codes spectraux de convection 3 D à haute résolution est réalisé actuellement à Nice (U. Frisch et P.L. Sulem à l'Observatoire, R. Peyret et C. Sulem au Département de Mathématiques). Au cours de mon stage j'ai eu l'occasion de m'initier aux techniques correspondantes. Le présent document, rédigé essentiellement à l'intention de chercheurs débutants, est basé sur le cours de P.L. Sulem à l'Observatoire de Nice.

# CHAPITRE I : METHODES SPECTRALES

## 1. Introduction.

Le but de ce chapitre est de présenter de façon simple les méthodes spectrales couramment utilisées pour résoudre un large éventail d'équations aux dérivées partielles. J'ai essayé d'exposer à la fois la formulation intrinsèque de ces méthodes et leur application pratique à l'aide d'exemples. Les équations aux dérivées partielles sont présentées pour des domaines bornés, mais il existe des méthodes permettant de se ramener à ce cas pour des domaines non bornés (Boyd [6], Grosch et Orszag [7]).

## 2. Exemples d'équations aux dérivées partielles.

Dans les paragraphes qui vont suivre l'accent sera mis sur la résolution des problèmes d'évolution à une dimension d'espace. Mais les méthodes spectrales peuvent s'appliquer à de nombreux autres problèmes.

Voici quelques exemples d'équations sur lesquelles peuvent facilement être testées les méthodes dont il sera question ici :

1) Equation de la chaleur : $\frac{\partial u}{\partial t} = \nu \Delta u + f$

2) Equation de Poisson : $\Delta u = f$

3) Equation des ondes : $\frac{\partial^2 u}{\partial t^2} - \Delta u = f$

4) Equation Magnéto-hydrodynamique (MHD) (voir Ch. VIII p 213 )

5) Equation d'advection-diffusion : $\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} + f$

6) Equation de Burgers : $\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} + f$

7) Equation de Kuramoto : $\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = - \frac{\partial^2 u}{\partial x^2} - \nu \frac{\partial^4 u}{\partial x^4} + f$

(voir une étude numérique de cette équation dans [14])

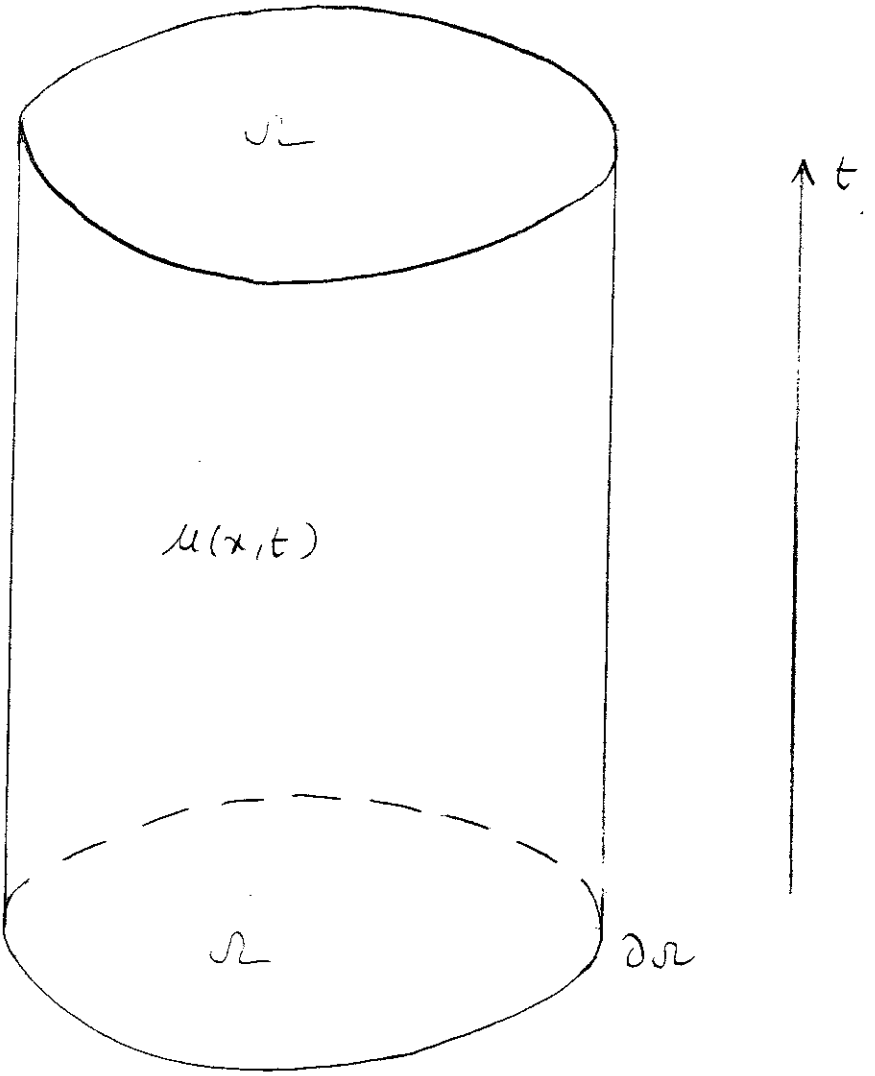



Fig 1

8) Equation de Klein-Gordon non linéaire : $\frac{\partial^2 u}{\partial t^2} - \Delta u = E(u)$

9) Equations de Navier-Stokes :

$$\begin{cases} \frac{\partial u}{\partial t} + u \cdot \nabla u = -\nabla p + \nu \Delta u + f \\ \operatorname{div} u = 0 \end{cases}$$

Nous verrons que la plupart du temps la dépendance temporelle des équations d'évolution est traitée en différences finies, si bien que l'on est ramené à chaque pas de temps à la résolution de problèmes "stationnaires" par une méthode spectrale.

Voici quelques exemples de domaines spatiaux et de conditions aux limites.

a) Equation de la chaleur avec conditions aux limites périodiques

$x \in \Gamma$ cercle unité $[0, 2\pi]$

b) Equation de la chaleur avec $x \in [0, \Pi]$

$u(0) = g_1 \qquad u(\pi) = g_2$

c) Equation de Kuramoto avec $x \in [-1, 1]$. Il faut quatre conditions aux limites. Par exemple :

$$\begin{cases} u(-1) = u(1) = 0 \\ \frac{\partial^2 u}{\partial x^2}(-1) = \frac{\partial^2 u}{\partial x^2}(1) = 0 \end{cases}$$

d) Equation d'advection-diffusion avec $x \in [0, \Pi]$ et $u(0) = g_1$

Enfin il faut se donner une condition initiale :

$$u(x,0) = u_0(x)$$

## 3. Formulation du problème.

La formulation générale suivante essaie de condenser les exemples précédents :

$$\frac{\partial u}{\partial t}(x,t) = F(u)(x,t) + f(x,t) \qquad x \in \Omega \quad t \geq 0$$

$$B u(x,t) = g(t) \qquad x \in \partial\Omega \quad t > 0$$

$$u(x,0) = u_0(x) \qquad x \in \Omega$$

. $\Omega$ est un domaine borné de $\mathbb{R}^n$ de frontière $\partial\Omega$

. On cherche u(x,t) fonction du temps à valeur dans un espace de Hilbert $\mathcal{H}$ muni de sa norme $\| \ \|$

. f est un élément de $\mathcal{H}$

. F est une fonction de $\mathcal{H}$ dans $\mathcal{H}$

. B est un opérateur de trace, déterminant les conditions aux limites ; on l'omettra si le domaine est périodique.

. $u_o(x)$ est la condition initiale.

Exemples :

a) $\mathcal{H} = L^2(0,\pi)$ ; $\mathcal{H} = L^2([-1,1], dx/\sqrt{1-x^2})$ fonctions de carré sommable pour la mesure $dx/\sqrt{1-x^2}$ ; $\mathcal{H} = L^2(T^1)$ dans le cas de conditions aux limites périodiques.

b) $F(u) = -a\frac{\partial u}{\partial x} + \nu\frac{\partial^2 u}{\partial x^2}$ équation d'advection-diffusion ;

$F(u) = -u\frac{\partial u}{\partial x} + \nu\frac{\partial^2 u}{\partial x^2}$ équation de Burgers ; etc....

c) $Bu = (u(-1), u(1)) = (g_1, g_2)$ avec $\Omega = [-1,1]$ ;

$B\vec{u} = \vec{u}_{\partial\Omega} \cdot \vec{n}$ avec $\Omega$ ouvert de $\mathbb{R}^n$ $\vec{n}$ normale à $\partial\Omega$

## 4. Principe des méthodes spectrales.

Les méthodes spectrales consistent à traiter la dépendance spatiale en décomposant les éléments de $\mathcal{H}$ sur une base de fonctions $(\varphi_n)_{n=1,\infty}$

On cherche :

$$u_N(x,t) = \sum_{n=1}^{N} a_n(t)\,\varphi_n(x)$$

élément de l'espace $B_N$ engendré par les N premières fonctions de base, de telle sorte que cette fonction approxime la vraie solution du problème.

On définit le résidu associé à l'approximation $u_N$ par :

$$R_N = \frac{\partial u_N}{\partial t} - F(u_N) - f$$

Ce résidu serait nul si $u_N$ était la vraie solution. On cherche donc à la rendre "petit". Pour cela on impose que $R_N$ soit nul en projection

sur un sous-espace de $\mathcal{H}$ de dimension N. Cette projection $P_N$ dépend de la méthode spectrale employée. On a donc remplacé le problème initial par le problème approché :

> Trouver $u_N \in B_N$ tel que $P_N R_N = 0$

On est alors ramené à la résolution de N équations différentielles en temps dont les inconnues sont les $a_n(t)$.

Exemples de fonctions de base utilisées :

a) dans $\mathcal{H} = L^2(\Gamma, \mathbb{C})$ $\quad (e^{ikx})\ k \in \mathbb{Z}$

b) dans $\mathcal{H} = L^2(0, \pi)$ $\quad (\sin nx)\ n = 1, \infty$ ou $(\cos nx)\ n = 1, \infty$

c) dans $\mathcal{H} = L^2([-1,1], dx/\sqrt{1-x^2})$ $\quad (T_n(x))\ n \in \mathbb{N}$ polynômes de Tchebyshev définis par $T_n(x) = \cos[n \operatorname{Arccos}(x)]$

Les principales méthodes spectrales sont les suivantes :

Méthode spectrales : - Galerkin
- Méthode Tau

Méthodes pseudo-spectrales : - Collocation
- Tau-collocation

## 5. Méthode de Galerkin.

Conditions d'applications : conditions aux limites périodiques ou homogènes.

Soit $\mathcal{B} = \{ v \in \mathcal{H}\ ,\ Bv = 0 \}$ le sous-espace des fonctions de $\mathcal{H}$ vérifiant les conditions aux limites. Dans cette méthode les $(\varphi_n)n = 1,$ forment une base de $\mathcal{B}$. En recherchant $u_N = \sum_{n=1}^{\infty} a_n \varphi_n$ dans $B_N$ engendré par les $(\varphi_n)n = 1,N$ on est sûr de trouver une fonction vérifiant les conditions aux limites.
Remarque : pour des conditions aux limites non homogènes $\mathcal{B}$ n'est plus un espace vectoriel mais un espace affine. Dans le cas où F est linéaire on peut se ramener au cas homogène en retranchant une solution particulière.

Soit $P_N^{\perp}$ la projection orthogonale de $\mathcal{H}$ sur $B_N$. La méthode de Galerkin consiste à résoudre le problème approché suivant

$$u_N = \sum_{n=1}^{N} a_n \varphi_n \in B_N \quad \text{tel que} \quad P_N^{\perp} R_N = 0$$

$$\text{avec} \quad R_N = \frac{\partial u_N}{\partial t} - F(u_N) - f$$

Exemple 1 : Equation de la chaleur

$$\begin{cases} \dfrac{\partial u}{\partial t} = \nu \dfrac{\partial^2 u}{\partial x^2} + f & x \in [0,\pi] \\ u(0,t) = u(\pi,t) = 0 & \forall t > 0 \\ u(x,0) = u_0 \end{cases}$$

$\mathcal{H} = L^2(0,\pi)$ avec les fonctions de base $\varphi_n(x) = \sin nx$, $n = 1,\infty$ qui vérifient les conditions aux limites.

- On a $f = \sum_{n=1}^{\infty} f_n(t) \sin nx$
- et on cherche $u_N(x,t) = \sum_{n=1}^{N} a_n(t) \sin nx$
- Calculons le résidu

$$R_N = \frac{\partial u_N}{\partial t} - \nu \frac{\partial^2 u_N}{\partial x^2} - f = \sum_{n=1}^{N} \left( \frac{da_n}{dt} + n^2 a_n \right) \sin nx - \sum_{n=1}^{\infty} f_n \sin nx$$

- Dire que $R_N$ est orthogonal à $B_N$ revient à poser les N équations :

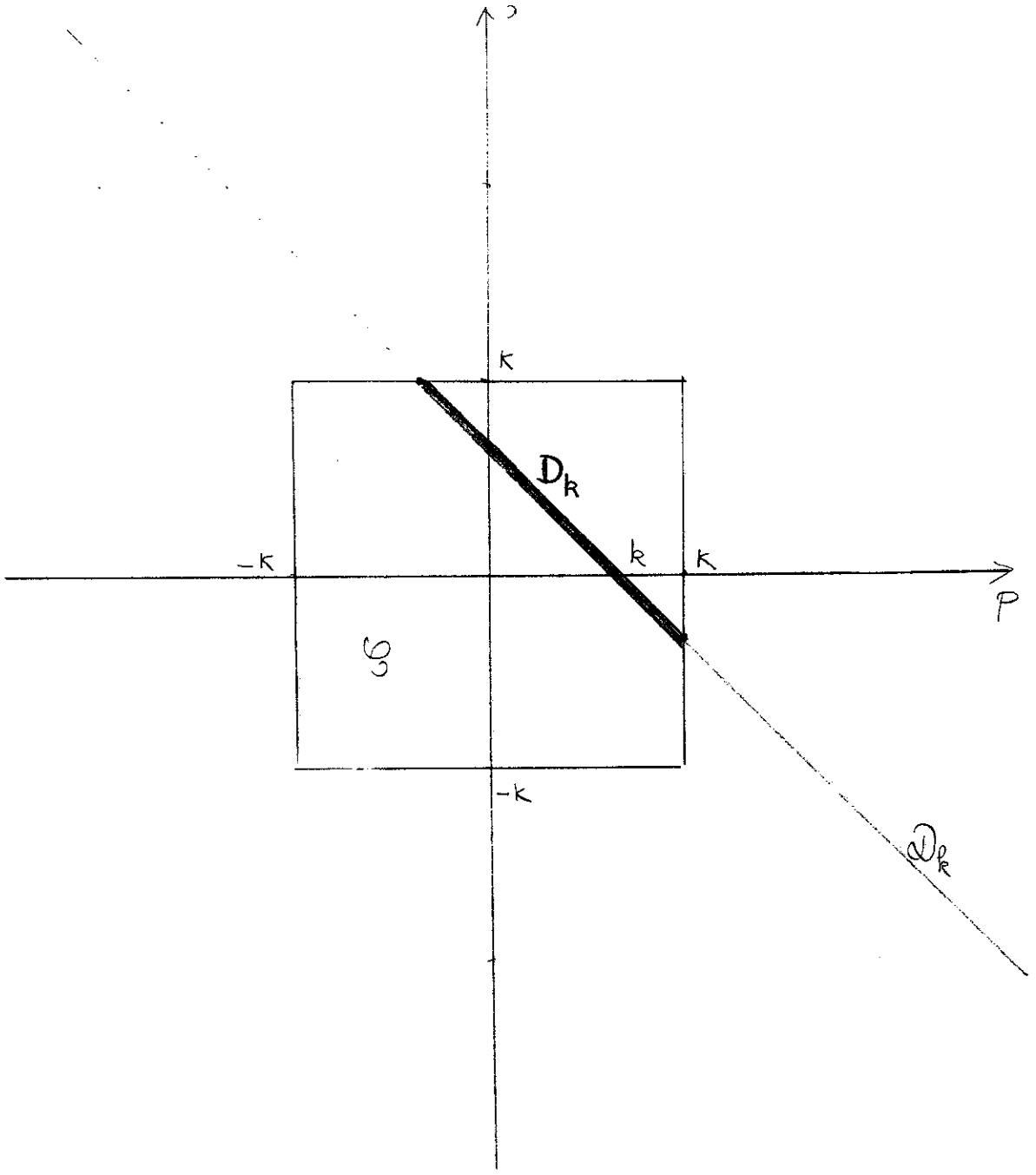
$$\frac{da_n}{dt} = -n^2 a_n + f_n \qquad n = 1, N$$

- On décompose $u_0(x) = \sum_{n=1}^{\infty} a_n^0 \sin nx$ ce qui donne N conditions initiales $a_n(0) = a_n^0$ pour n = 1, N.

On résoud alors ces N équations différentielles, soit par un schéma aux différences finies en temps soit analytiquement lorsque cela est simple comme ici :

$$a_n(t) = \frac{f_n}{n^2} \left( 1 - e^{-n^2 t} \right) + a_0 \, e^{-n^2 t}$$

Remarquons que la vraie solution $u(x,t) = \sum_{n=1}^{\infty} a_n \sin nx$ se décompose avec les mêmes coefficients (modulo la troncature) que ceux de la solution approchée $u_N$, qui se trouve être par conséquent la projection orthogonale de u sur $B_N$. Ce n'est pas le cas lorsque F est non linéaire comme dans l'exemple suivant.




Segment de droite $D_k$ dans le carré $\mathcal{C}$
et droite $\mathcal{D}_k$

Fig 2

Exemple 2 : Equation de Burgers périodique

$$\begin{cases} \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \nu \dfrac{\partial^2 u}{\partial x^2} + f & x \in \Gamma \text{ cercle unité } [0, 2\pi] \\ u(x,0) = u_0(x) \end{cases}$$

$\mathcal{H} = L^2(\Gamma, \mathbb{C})$ avec les fonctions de base $\varphi_k(x) = e^{ikx}$, $k \in \mathbb{Z}$

. On pose $N = 2K + 1$. Soit $B_N$ engendré par $(e^{ikx})$ $-K \leqslant k \leqslant K$
on cherche $u_N = \sum_{k=-K}^{K} a_k(t)\, e^{ikx}$

. Pour calculer le résidu $R_N = \dfrac{\partial u_N}{\partial t} + u_N \dfrac{\partial u_N}{\partial x} - \nu \dfrac{\partial^2 u_N}{\partial x^2} - f$
Il faut effectuer le produit de convolution discret :

$$u_N \frac{\partial u_N}{\partial x} = \Big(\sum_{p=-K}^{K} a_p e^{ipx}\Big)\Big(\sum_{p=-K}^{K} iq\, a_q e^{iqx}\Big) = \sum_{(p,q)\in\mathcal{C}} iq\, a_p a_q\, e^{i(p+q)x}$$

où $\mathcal{C}$ désigne le carré $[-K, K] \times [-K, K]$
En factorisant $e^{ikx}$ on regroupe les termes $(p,q) \in D_k$ où $D_k$ désigne l'intersection de la droite $p + q = k$ avec le carré : $D_k = \{(p,q) \in \mathcal{C},\ p+q = k\}$

$$u_N \frac{\partial u_N}{\partial x} = \sum_{k=-2K}^{2K} \Big(\sum_{(p,q)\in D_k} iq\, a_p a_q\Big) e^{ikx}$$

En effectuant la projection orthogonale sur $B_N$ de cette expression il ne reste que les termes $k \in [-K, K]$ de la sommation.

. La méthode de Galerkin conduit donc au système :

$$\frac{da_k}{dt} + \sum_{(p,q)\in D_k} iq\, a_p a_q = -\nu k^2 a_k + f_k \qquad k \in [-K, K]$$

. Comparons les coefficients $a_k$ avec ceux de la vraie solution
$u(x,t) = \sum_{k\in\mathbb{Z}} A_k(t)\, e^{ikx}$ qui vérifient :

$$\frac{dA_k}{dt} + \sum_{(p,q)\in\mathcal{D}_k} iq A_p A_q = -\nu k^2 A_k + f_k \qquad k \in \mathbb{Z}$$

ou $\mathcal{D}_k$ désigne la droite de $\mathbb{Z}^2$ $p + q = k$ toute entière.

. On voit que contrairement au cas où F est linéaire $u_N$ n'est pas la projection orthogonale de u. En effet dans le cas approché le couplage des équations s'effectue par le segment de droite $D_k$ au lieu de la droite toute entière $\mathcal{D}_k$ dans le cas du vrai problème.

## 6. Méthode Tau.

Conditions d'application : conditions aux limites non périodiques.

Si F contient des dérivations d'ordre k les conditions aux limites sont au nombre de k : Bu = g g ayant k composantes. Soit $(\varphi n)n = 1,\infty$ une base orthogonale ne vérifiant pas les conditions aux limites*.

La méthode consiste à chercher $u_N = \sum_{n=1}^{\infty} a_n \varphi_n$ dans $B_N$ tel que :

| | | |
|---|---|---|
| $P^{\perp}_{N-k} R_N = 0$ | N - k | équations |
| $B u_N = 0$ | k | équations |

$P^{\perp}_{N-k}$ désigne la projection orthogonale de $\mathcal{H}$ sur $B_{N-k}$.
$u_N$ est alors déterminé par N équations différentielles.

Exemple : équation de la chaleur

$$\begin{cases} \dfrac{\partial u}{\partial t} = \nu \dfrac{\partial^2 u}{\partial x^2} + f & x \in [-1,1] \quad t \geq 0 \\ u(-1,t) = g_1 \; ; \; u(1,t) = g_2 & t > 0 \\ u(x,0) = u_0(x) \end{cases}$$

$\mathcal{H} = L^2([-1,1], dx/\sqrt{1-x^2})$ avec comme fonctions de base les polynômes de Tchebyshev $(T_n)_{n \in \mathbb{N}}$

On cherche $u_N = \sum_{n=0}^{N} a_n(t) T_n(x)$ dans $B_N$ de dimension N + 1 (notations spéciales à cet exemple)

On pose $\dfrac{\partial^2 u}{\partial x^2} = \sum_{n=0}^{N-1} a_n^{(2)}(t) T_n(x)$

---

* Il faut de plus que toutes les matrices k x N : $[B(\varphi_n)]$ aient un rang égal à k.

les coefficients $a_n^{(2)}$ sont donnés par la formule (voir chapitre III) :

$$a_n^{(2)} = \frac{2}{c_n} \sum_{\substack{p=n+1 \\ \text{step } 2}}^{N} p\, a_p \qquad \text{avec } c_0 = 2 \text{ et } c_n = 1 \text{ sinon}$$

Donc $$R_N = \sum_{n=0}^{N} \left( \frac{da_n}{dt} - \nu a_n^{(2)} \right) T_n + \sum_{n=0}^{\infty} f_n T_n$$

Comme k = 2 on pose nulle la projection de ce résidu sur $B_{n-2}$ :

$$\frac{da_n}{dt} = \nu a_n^{(2)} + f_n \qquad n = 0, N-2 \qquad : N - 1 \text{ équations}$$

Les conditions aux limites s'écrivent

$$\begin{cases} \sum_{n=0}^{N} a_n T_n(-1) = \sum_{n=0}^{N} (-1)^n a_n = g_1 \\ \sum_{n=0}^{N} a_n T_n(1) = \sum_{n=0}^{N} a_n = g_2 \end{cases} \qquad : 2 \text{ équations}$$

Ces N + 1 équations permettent de trouver les N + 1 $a_n(t)$, en utilisant les conditions initiales $u_0(x) = \sum_{n=0}^{\infty} a_n^0 T_n(x)$

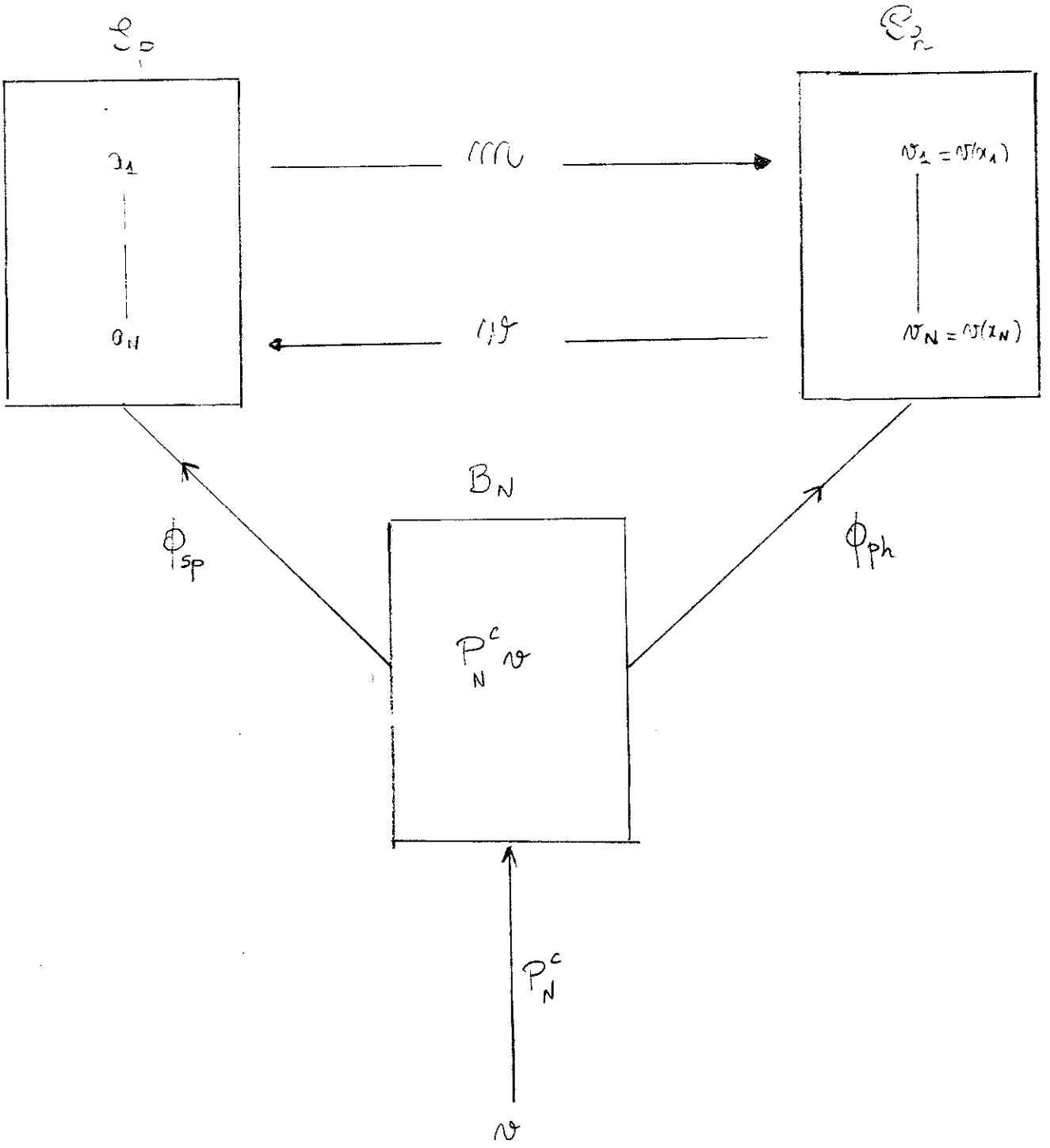
Projecteur de la méthode Tau.

On peut donner une interprétation plus intrinsèque de la méthode Tau. En fait on cherche $u_N$ dans l'espace affine de dimension N - k des éléments $B_N$ vérifiant les k conditions aux limites. Pour déterminer $u_N$ dans cet ensemble on demande que son résidu soit orthogonal au sous-espace $B_{N-k}$ de dimension N - k lui aussi.

## 7. Méthode de collocation.

Conditions d'applications : ce sont les mêmes que pour Galerkin, conditions aux limites périodiques ou homogènes.

Comme pour la méthode de Galerkin les $(\varphi_n)n = 1$ forment une base complète de $\mathcal{B}$, et vérifient donc les conditions aux limites. On se donne de plus $(x_i)i = 1, N$ : N points dits de collocation dans $\Omega$ tels que la matrice N x N : $\mathcal{M} = (\varphi_m(x_i))$ soit inversible.




Projection de collocation d'une fonction $v \in \mathcal{H}$

Fig. 3

La méthode consiste à chercher $u_N = \sum_{n=1}^{N} a_n \varphi_n$ dans $B_N$ tel que son résidu $R_N = \frac{\partial u_N}{\partial t} + F(u_N) - f$ vérifie les N conditions:

$$R_N(x_i) = 0 \quad i = 1, N$$

Ces N conditions s'écrivent encore $P_N^c R_N = 0$ où $P_N^c$ désigne la projection de collocation définit comme suit :

soit $v \in \mathcal{H}$ et $v_i = v(x_i)$ $i = 1, N$ les valeurs de cette fonction aux points de collocation (en réalité le domaine de $P_N^c$ est plus petit que $\mathcal{H}$ : $\mathcal{C}^0(\Omega)$ et par extension $L^1(\Omega)$ ).
Il existe une seule fonction $P_N^* v$ de $B_N$ prenant les mêmes valeurs aux mêmes points de collocation.

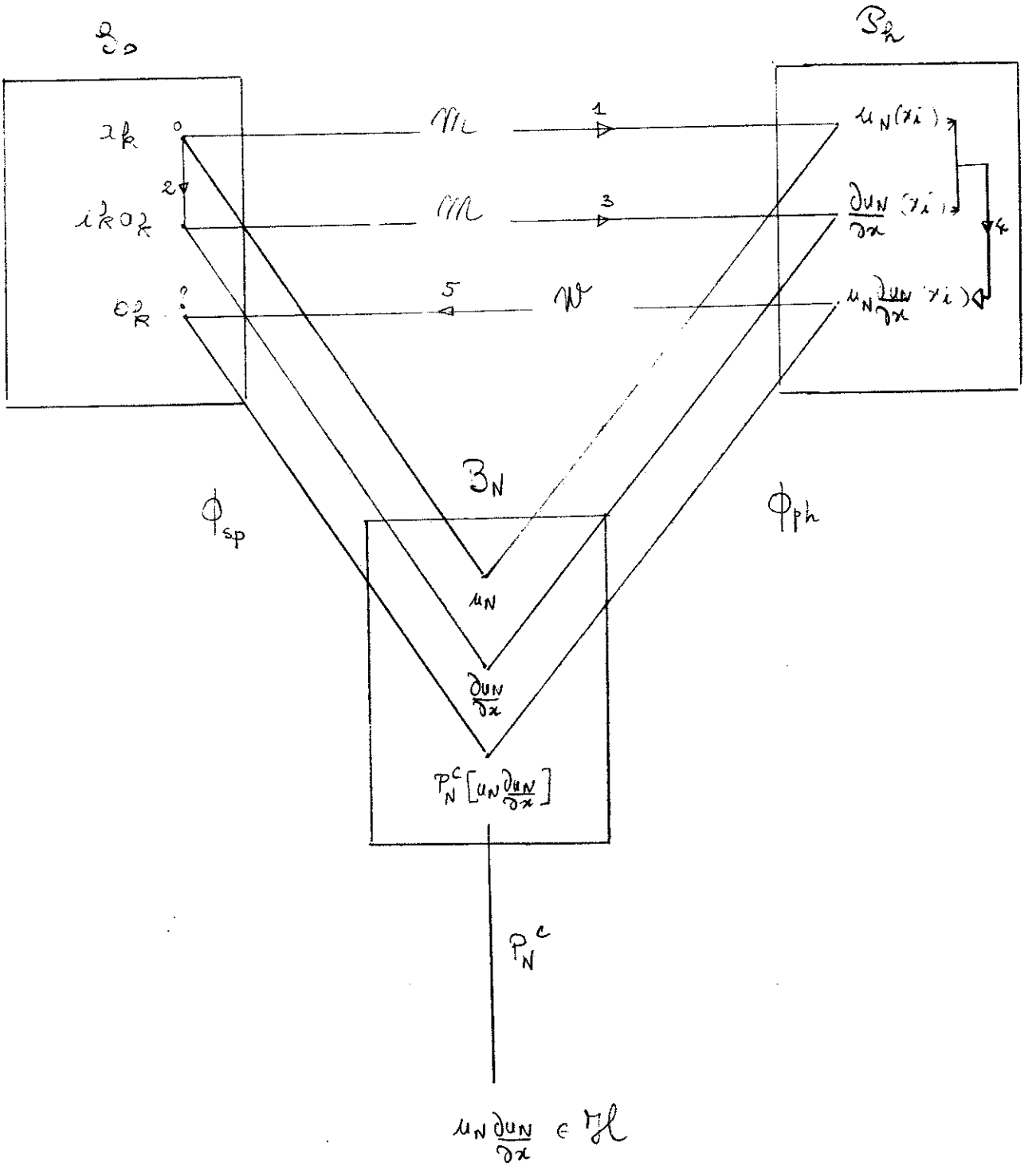
En pratique cette projection s'effectue simplement à condition de savoir inverser la matrice $\mathcal{M} = (\varphi_n(x_i))$ . Pour en bien comprendre le mécanisme introduisons deux isomorphismes

$$\phi_{ap} : B_N \longrightarrow \mathcal{S}_p = \mathbb{C}^N$$
$$v = \sum_{n=1}^{N} b_n \varphi_n \longmapsto (b_1, b_2, \ldots, b_N)$$

$$\phi_{ph} : B_N \longrightarrow \mathcal{S}_h = \mathbb{C}^N$$
$$v \longmapsto (v_1, v_2, \ldots, v_N) \text{ avec } v_i = v(x_i)$$

On appelle $\mathcal{S}_p$ "l'espace spectral" et $\mathcal{S}_h$ "l'espace physique". De ces deux isomorphismes on en déduit un troisième de $\mathcal{S}_p$ dans $\mathcal{S}_h$ définit par $\sum_{n=1}^{N} b_n \varphi_n(x_i) = v_i \quad i = 1, N$
Sa matrice pour les bases canoniques est donc $\mathcal{M} = (\varphi_n(x_i))$ .
On aura donc intérêt à disposer d'algorithmes rapides permettant d'appliquer la matrice $\mathcal{M}$ et son inverse à un vecteur. (Par exemple l'algorithme de Transformée de Fourier Rapide = Fast Fourier Transform = FFT).

Ces isomorphismes $\mathcal{S}_p \sim B_N \sim \mathcal{S}_h$ permettent de décomposer la projection de collocation. Etant donnée une fonction $v \in \mathcal{H}$, les N valeurs $v_i = v(x_i)$ déterminent la projection $P_n^c(v)$ dans l'espace physique et en appliquant $\mathcal{M}^{-1}$ on obtient ses coefficients dans l'espace spectral (Fig. 3).




Calcul des $b_k$ connaissant les $b_k$

Fig. 4

La puissance de la méthode de collocation réside dans la simplicité de la projection des termes non linéaires. Par exemple la projection de collocation de $v^2$ est donnée par $(v_i)^2$ $i = 1, N$, celle de sin v par $\sin(v_i)$ $i = 1, N$ etc... De plus toute fonction de $B_N$ se projette en elle même.

On calcule donc facilement la projection de collocation du résidu $R_N$ en effectuant des allers et retours entre l'espace spectral où sont calculées les dérivations et l'espace physique où sont projetés les termes non linéaires.

Exemple : Equation de Burgers périodique

$$\begin{cases} \dfrac{\partial u}{\partial t} + u \dfrac{\partial u}{\partial x} = \nu \dfrac{\partial^2 u}{\partial x^2} + f & x \in \Gamma \text{ cercle unité } [0, 2\pi] \ , \ t > 0 \\ u(x,0) = u_0(x) \end{cases}$$

$\mathcal{H} = L^2(\Gamma, \mathbb{C})$ avec la base $(e^{ikx})_{k \in \mathbb{Z}}$

$B_N$ est engendré par les $(e^{ikx})_{-K \leq k \leq K}$ , $N = 2K+1$

Sur $\Gamma = [0, 2\pi]$ les points de collocation sont choisis à intervalles réguliers: $x_i = \dfrac{2\pi i}{N}$ , $i = 1, N$ . On a $\mathcal{M} = (e^{ikx_i})$ matrice N x N. L'inconnue est $u_N = \sum_{k=-K}^{K} a_k e^{ikx}$

Il faut calculer la projection du résidu $R_N = \dfrac{\partial u_N}{\partial t} - \nu \dfrac{\partial^2 u_N}{\partial x^2} - u_N \dfrac{\partial u_N}{\partial x} - f$

Les deux premiers termes appartiennent à $B_N$ et sont invariants par $P_N^c$. Pour projetter $u_N \dfrac{\partial u_N}{\partial x}$ qui n'est pas un élément de $B_N$ on procède comme suit :

Des $(a_k)_{-K \leq k \leq K}$ coefficients spectraux de $u_N$ on déduit très simplement $(ik a_k)_{-K \leq k \leq K}$ coefficients de $\dfrac{\partial u_N}{\partial x}$ .

En appliquant la matrice $\mathcal{M}$ à ces deux vecteurs on en déduit $[u_N(x_i)]_{i=1,N}$ et $\left[\dfrac{\partial u_N}{\partial x}(x_i)\right]_{i=1,N}$ les coefficients de $u_N$ et $\dfrac{\partial u_N}{\partial x}$ dans l'espace physique. Les coefficients de $P_N^c\left(u_N \dfrac{\partial u_N}{\partial x}\right)$ dans l'espace physique sont : $\left[u_N(x_i) \dfrac{\partial u_N}{\partial x}(x_i)\right]_{i=1,N}$ . En y appliquant la matrice $\mathcal{M}^{-1}$ on calcule les $(b_k)_{-K \leq k \leq K}$ coefficients spectraux de cette projection, ce qui a coûté au total deux multiplications par $\mathcal{M}$ et une par $\mathcal{M}^{-1}$ (Fig. 4).

Le Problème approché de la méthode de collocation s'écrit alors dans l'espace spectral :

$$P_N^c R_N = 0 \iff \frac{da_k}{dt} + b_k = -\nu k^2 a_k + f_k \qquad k = -K, K$$

Remarques.

1) On aurait pu écrire les N équations différentielles dans l'espace physique en utilisant le même nombre de multiplications par $\mathcal{M}$ ou $\mathcal{M}^{-1}$. Dans d'autres problèmes il faut étudier quel espace nécessite le moins d'allers et retours. Par exemple les équations de Naviers Stokes doivent être écrites dans l'espace spectral.

2) Il y a un moyen d'économiser une multiplication par $\mathcal{M}$ dans le traitement de l'équation de Burgers en traitant le terme non linéaire sous la forme $\frac{1}{2}\frac{\partial}{\partial x}(u^2)$ mais il ne s'agit alors plus de la collocation habituelle ( voir I. 11 p137: méthode sans nom).

## 8. Méthode Tau-collocation.

Conditions d'application : ce sont les mêmes que pour la méthode Tau. Conditions aux limites non périodiques.

Comme pour la méthode Tau $(\varphi_n)_{n=1,\infty}$ est une base orthogonale ne vérifiant pas les k conditions aux limites.

Comme pour la collocation $(x_i)\ i=1,N$ sont N points du domaine $\Omega$ et l'on sait appliquer rapidement $\mathcal{M}=(\varphi_n(x_i))$ et son inverse.

On calcule la projection $P_N^C R_N$ avec des allers et retours entre espace spectral et espace physique.

Soit $P_{N-k}^{\perp}$ la projection orthogonale sur $B_{N-k}$. Les N conditions déterminant $u_N \in B_N$ par cette méthode sont

$$P_{N-k}^{\perp} P_N^C R_N = 0$$

$$B u_N = 0$$

Exemple : Equation de Burgers non périodique

$$\begin{cases} \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \nu \dfrac{\partial^2 u}{\partial x^2} + f & x \in [-1,1] \quad t > 0 \\ u(-1,t) = g_1 \qquad u(1,t) = g_2 & t > 0 \\ u(x,0) = u_0(x) \end{cases}$$

$\mathcal{H} = L^2\left([-1,1], \dfrac{dx}{\sqrt{1-x^2}}\right)$ ; $(T_n)_{n\in\mathbb{N}}$ base des polynômes de Tchebyshev.

En choisissant comme points de collocations les $x_j = \cos\left(j\dfrac{2\pi}{N+1}\right)$ $j=\overline{1,N}$ et N pair on peut utiliser l'algorithme de Transformée de Fourier Rapide pour appliquer la matrice $\mathcal{M} = (T_n(x_j)) = \left(\cos nj\dfrac{2\pi}{N+1}\right)$

Les N + 1 équations du problème approché s'écrivent :

Trouver $u_N = \sum_{n=0}^{N} a_n T_n$ dans $B_N$ (de dimension N + 1 ici), tel que :

$$\begin{cases} \dfrac{da_n}{dt} + b_n = \nu\, a_n^{(2)} + f_n & n = 0, N-2 \\ \sum_{n=0}^{N} (-1)^n a_n = g_1 \\ \sum_{n=0}^{N} a_n = g_2 \end{cases}$$

avec $a_n^{(2)} = \sum_{\substack{p=n+1 \\ \text{step } 2}}^{N} p\, a_p$ et $b_n$ coefficients spectraux de $P_N^c\left(u_N \dfrac{\partial u_N}{\partial x}\right)$

Interprétation intrinsèque :

La méthode Tau-collocation consiste à chercher $u_N$ dans l'espace affine de dimension N - k des $u_N \in B_N$ vérifiant $Bu_N = g$ tel que le résidu $R_N$ annule la projection $P^{\perp}_{N-k} P^{c}_{N}$ de $\mathcal{H}$ sur $B_{N-k}$

| Méthodes | Domaines d'application |
|---|---|
| Galerkin et collocation | conditions aux limites périodiques ou homogènes |
| Tau et Tau-collocation | conditions aux limites non périodiques |

Tableau 1 Domaines d'application des méthodes spectrales.

| C.L. \ F | F linéaire à coefficients constants | F quelconque |
|---|---|---|
| C.L. périodiques | GALERKIN ex : chaleur, périodique | COLLOCATION ex : Burgers, périodique |
| CL non périodiques (homogènes et non homogènes) | TAU ou GALERKIN ex : chaleur, non périodique | TAU-COLLOCATION ex : Burgers, non périodique |

Tableau 2 : Essai de classification du choix de la méthode pour des applications pratiques.

| Conditions aux limites | Méthodes | Domaine et fonctions de base. |
|---|---|---|
| C.L. périodiques | Galerkin et Collocation | $x \in \Gamma$ cercle unité $[0, 2\pi]$ $(e^{ikx})_{k \in \mathbb{Z}}$, $(\sin nx)_{n \in \mathbb{N}^*}$ ou $(\cos nx)_{n \in \mathbb{N}}$ |
| C.L non périodiques | Tau et Tau-collocation | $x \in [-1, 1]$ Polynômes de Tchebyshev, Legendre ou Laguerre |

Tableau 3 : Fonctions de bases usuelles.

## 9. Quelle méthode utiliser ?

Le tableau 1 résume les conditions d'application des quatres méthodes spectrales précédentes.

Le tableau 2 donne une indication sur le choix d'une méthode étant donné un problème. Mais ce tableau n'a rien de catégorique et il sert juste à fixer les idées. Lorsque F est linéaire la méthode de Galerkin (resp Tau) est identique à la méthode de collocation (resp Tau -collocation). Lorsque F est non linéaire* ces dernières sont plus rapides en temps de calcul, à condition de posséder un algorithme de transformation rapide entre l'espace spectral et l'espace physique .

Le tableau 3 indique les principales fonctions de bases pour lesquelles il existe de tels algorithmes de transformation : Orszag a mis au point des algorithmes rapides pour les polynômes de Legendre ou de Laguerre, mais leur utilisation n'est pas encore très répandue.

Remarque : Pourquoi "la méthode Tau" ?

Dans le cas des méthodes de Galerkin ou de collocation le résidu de la solution approchée peut s'écrire $R_N = \sum_{s=N+1}^{\infty} a_s \varphi_s$ et plus les $r_s$ sont petits, meilleure est l'approximation.

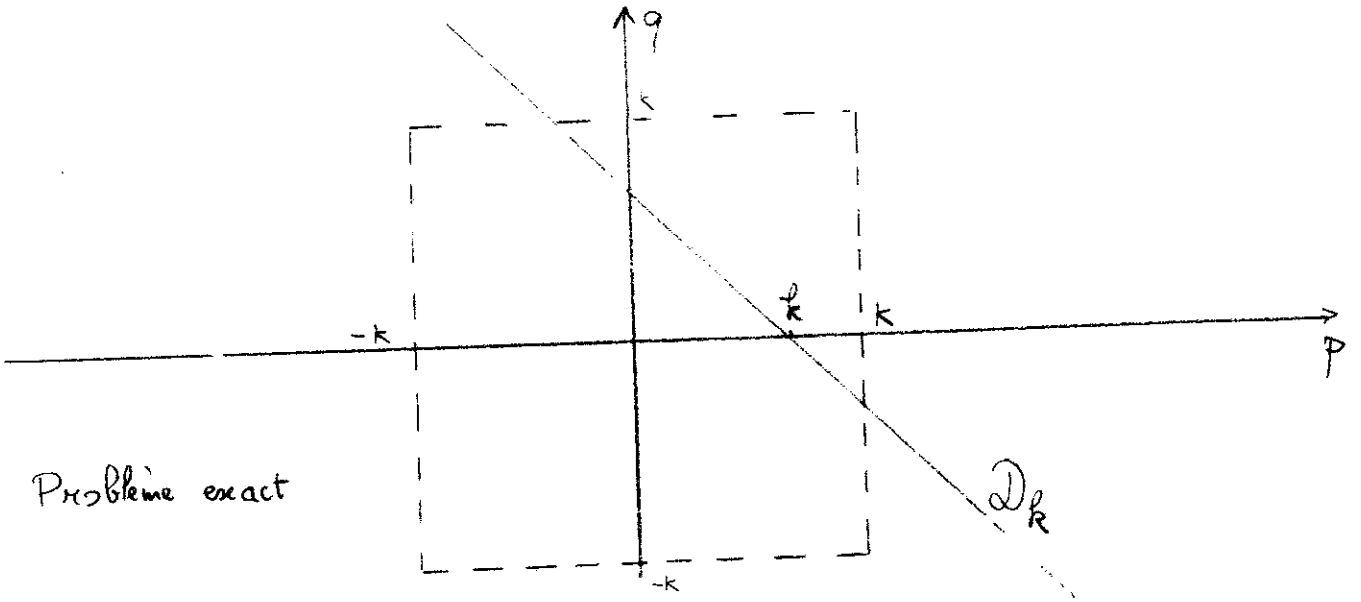
Dans le cas des méthodes Tau ou Tau-collocation $R_N = \sum_{p=1}^{k} \tau_p \varphi_{N-k+p} + \sum_{s=N+1}^{\infty} a_s \varphi_s$ . La solution approchée obtenue par ces méthodes est en réalité celle que l'on aurait obtenue en résolvant par la méthode de Galerkin ou de collocation le problème modifié suivant :

$$\frac{\partial u}{\partial t} = F(u) + f + \sum_{p=1}^{k} \tau_p \varphi_{N-k+p}$$ sans conditions aux limites avec les $\tau_p$ ajustés de telle sorte que la solution approchée de ce problème vérifie les conditions aux limites du premier problème.

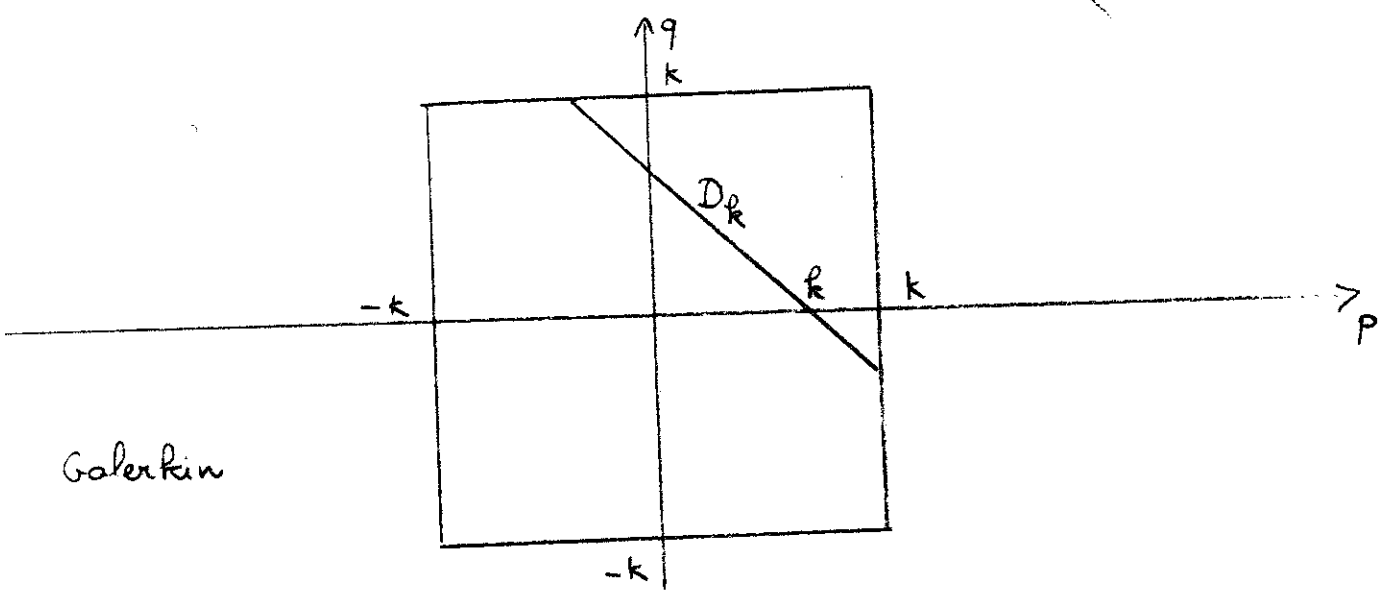
La connaissance de ces coefficients "Tau" indice p permet une évaluation de l'erreur de l'approximation.

---

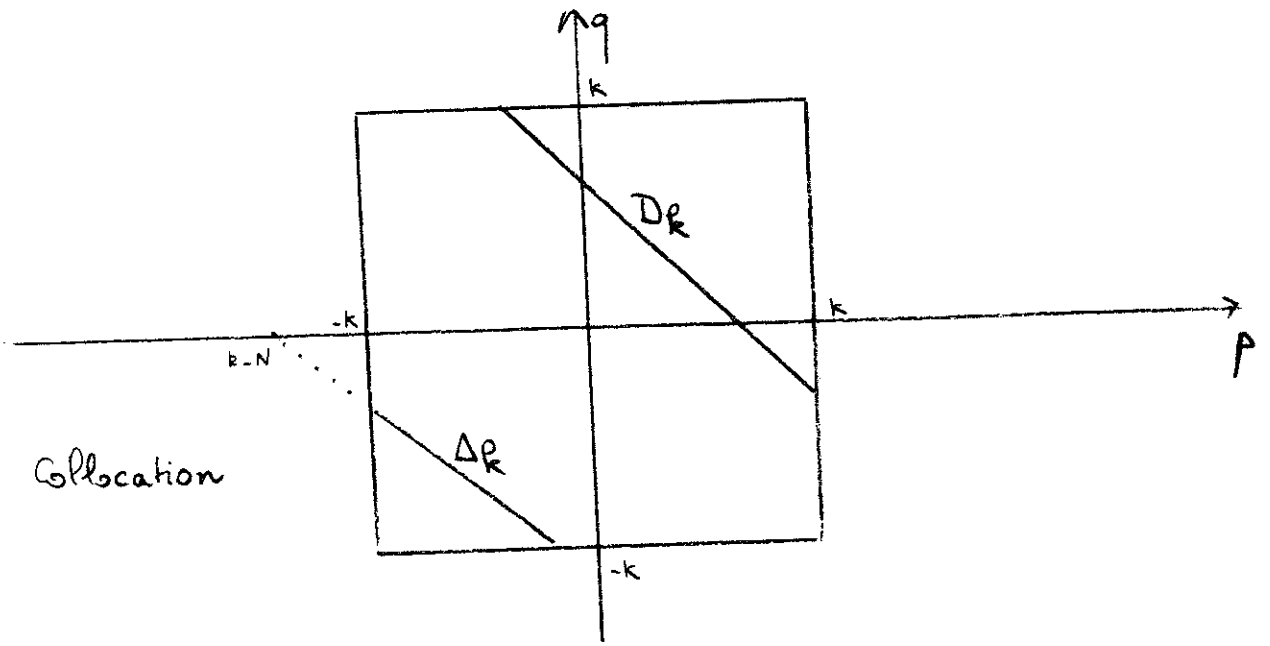
* ou linéaire avec des coefficients dépendant de x.




Problème exact




Galerkin




Collocation

Fig. 5

## 10. Le problème des erreurs de repliements ("Aliasing").

On a vu que lorsque F est non linéaire la méthode de Galerkin (resp Tau) ne donne pas la même approximation que la méthode de collocation (resp Tau-collocation). En comparant avec le système vérifié par la solution exacte on peut penser que la collocation (resp Tau-collocation) est moins bonne car elle introduit des termes parasites ce que l'on traduit par le mot "aliasing".

Etudions ceci sur l'exemple de l'équation de Burgers périodique traitée par les méthodes de Galerkin et de collocation (N = 2K + 1)

. La vraie solution $u = \sum_{k\in\mathbb{Z}} a_k e^{ikx}$ vérifie le système

$$\boxed{\frac{da_k}{dt} + \sum_{(p,q)\in\mathcal{D}_k} iq a_p a_q = -\nu k^2 a_k + f_k \quad , k\in\mathbb{Z}}$$

avec $\mathcal{D}_k = \{(p,q)\in\mathbb{Z}^2,\ p+q=k\}$ droite de $\mathbb{Z}^2$

. L'approximation de Galerkin $u_N^G = \sum_{k=-K}^{K} a_k^G e^{ikx}$ vérifie

$$\boxed{\frac{da_k^G}{dt} + \sum_{(p,q)\in D_k} iq a_p^G a_q^G = -\nu k^2 a_k^G + f_k \quad -K\leq k\leq K}$$

avec $D_k = \mathcal{D}_k \cap \mathcal{C}$ segment de droite du carré $\mathcal{C} = [-K,K]^2$

. L'approximation de collocation $u_N^c = \sum_{k=-K}^{K} a_k^c e^{ikx}$ vérifie

$$\boxed{\frac{da_k^c}{dt} + \sum_{(p,q)\in D_k+\Delta_k} iq a_p^c a_q^c = -\nu k^2 a_k^c + f_k \quad -K\leq k\leq K}$$

avec

$$D_k + \Delta_k = \{(p,q)\in\mathcal{C},\ p+q = k \text{ modulo } N\}$$

Les domaines $\mathcal{D}_k$, $D_k$ et $\Delta_k$ sont représentés sur la figure 5. On voit que la collocation fait intervenir les termes $(p,q)\in\Delta_k$ qui ne figurent pas dans la sommation $(p,q)\in\mathcal{D}_k$ du problème exact.

C'est la projection de collocation qui introduit ces termes supplémentaires. On peut voir le phénomène d'aliasing de la façon suivante : soit $v = \sum_{k\in\mathbb{Z}} b_k e^{ikx}$ une fonction quelconque. Sa projection de

collocation $P_N^c v = \sum_{k=-K}^{K} b_k e^{ikx}$ s'obtient en résolvant le système :

$$P_N^c v(x_j) = v(x_j) \qquad j = 1, N$$

$$\Leftrightarrow \quad \sum_{k=-K}^{K} b_k e^{ik_j \frac{2\pi}{N}} = \sum_{k \in \mathbb{Z}} B_k e^{ik_j \frac{2\pi}{N}}$$

$$= \sum_{k=-K}^{K} \left( \sum_{n \in \mathbb{Z}} B_{k+nN} \right) e^{ik_j \frac{2\pi}{N}} \qquad j = 1, N$$

la matrice $\mathcal{M} = (e^{ik_j \frac{2\pi}{N}})$ étant inversible on obtient :

$$b_k = \sum_{n \in \mathbb{Z}} B_{k+nN}$$

On voit donc que la projection de collocation a pour effet d'additionner toutes les harmoniques d'un même nombre d'onde k contenu dans la troncature. On peut comparer cet effet à celui d'un stroboscope. C'est ce qui s'est passé lorsqu'on a projetté la fonction $u_N \frac{\partial u_N}{\partial x}$ qui contient des modes non nuls sur $[-2K, 2K]$.

desaliasing grossier :

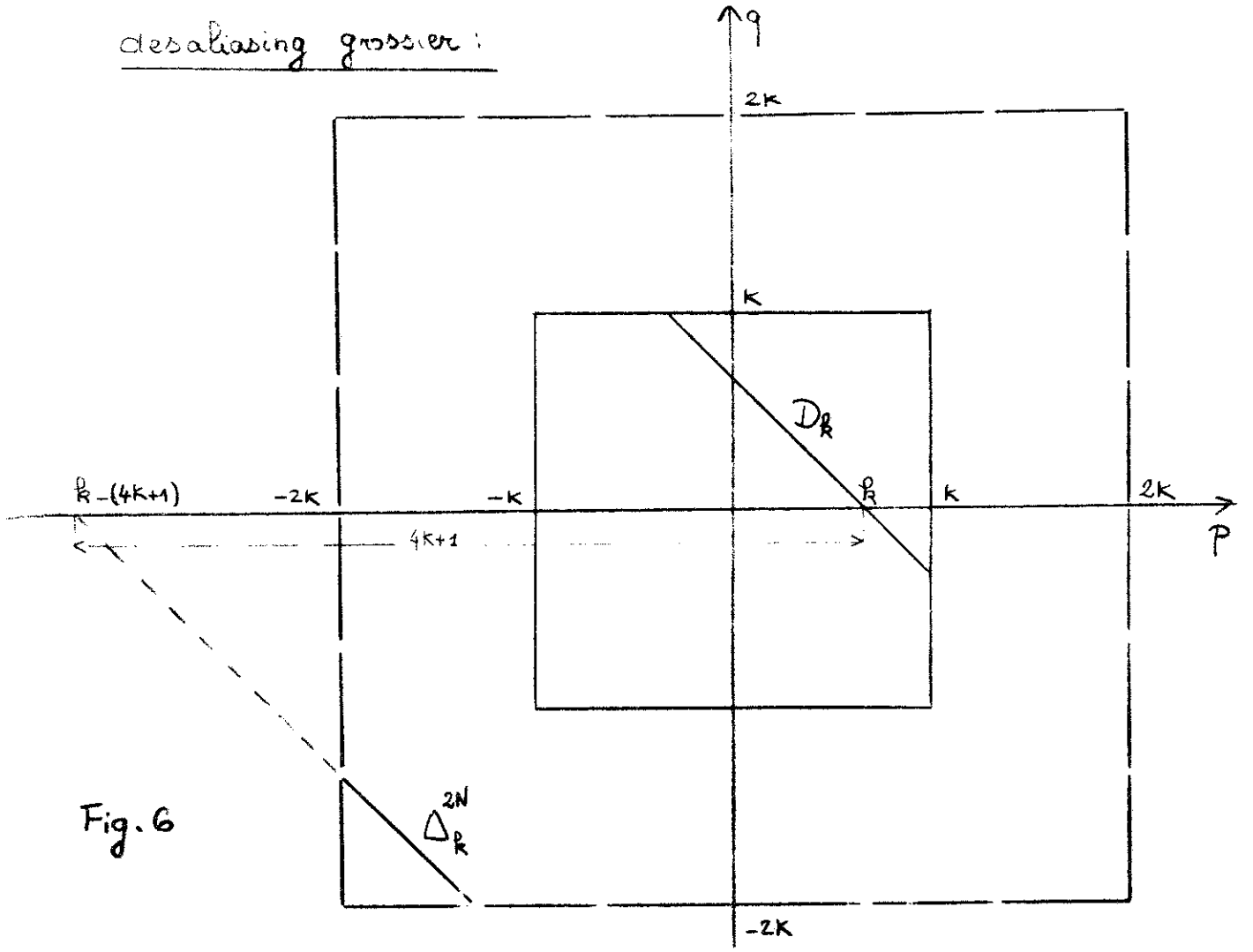



Fig. 6

desaliasing fin :

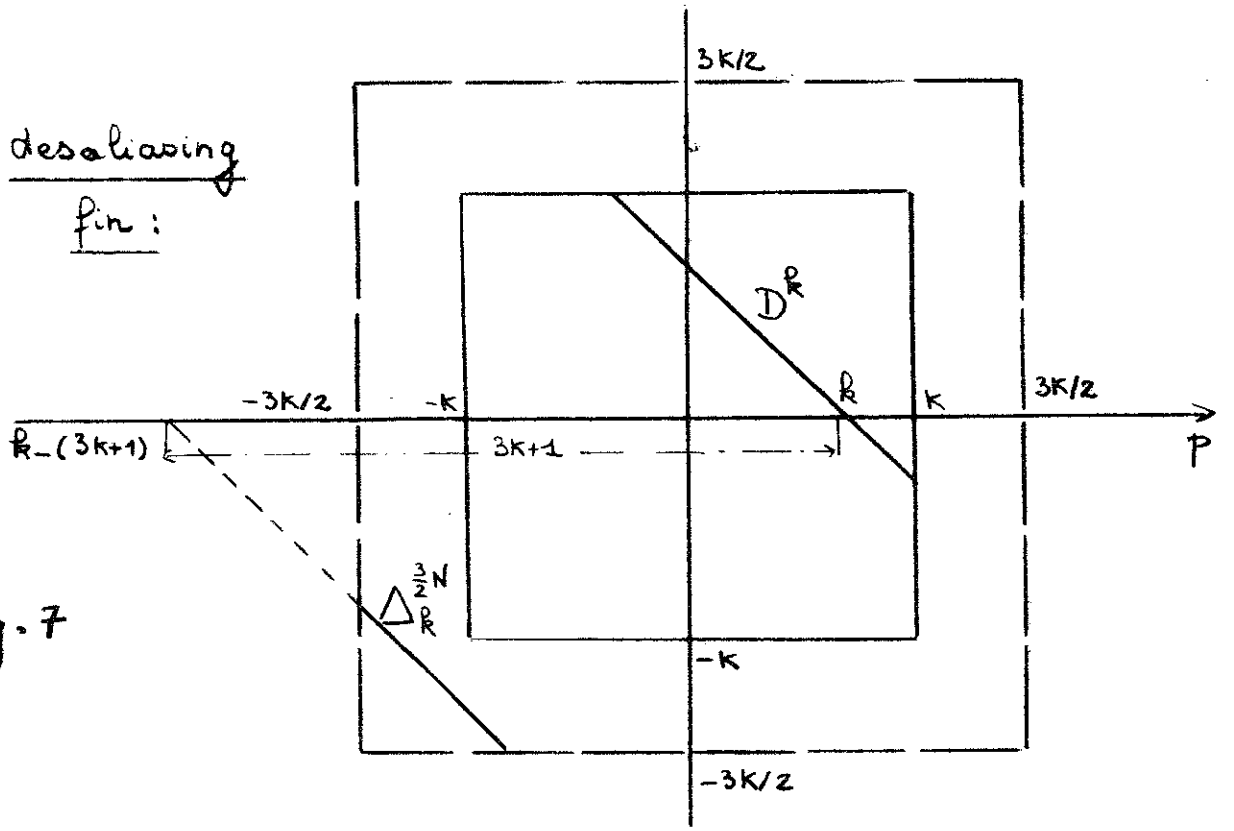



Fig. 7

## Desaliasing

Il existe un moyen de trouver l'approximation de Galerkin qui ne contient pas d'erreur de repliement (aliasing), en profitant néanmoins des avantages numériques (rapidité) de la collocation. Par abus de langage on dit que l'on utilise la méthode de collocation dans laquelle on supprime les termes de repliement grâce au procédé exposé ci-dessous (desaliasing).

a) Desaliasing grossier : projection de collocation $P^c_{2N-1}$, $2N-1=4K+1$ (Fig. 6)

On remarque qui si $u_N \in B_N$ alors $u_N \frac{\partial u_N}{\partial x}$ et donc $R_N$ appartiennent à $B_{2N-1}$. On a donc de façon évidente $P^c_{2N-1} R_N = R_N$.
La méthode de Galerkin peut donc s'écrire $\boxed{P^\perp_N R_N = P^\perp_N P^c_{2N-1} R_N = 0}$
Numériquement on calcule $P^c_{2N-1} R_N$ dans l'espace spectral à l'aide des algorithmes de transformations rapides (calcul des termes non-linéaires dans l'espace physique) et on applique ensuite $P^\perp_N$ de façon triviale. Ces algorithmes transforment des tableaux deux fois plus grands que pour la collocation normale, c'est le prix à payer pour le desaliasing.

b) Desaliasing fin : projection de collocation $P^c_{\frac{3}{2}N}$ (Fig. 7).

On note ici $\frac{3}{2}N$ le nombre entier $3K+1 = \frac{3N-1}{2}$
Ce desaliasing repose sur la forme particulière de $u_N \frac{\partial u_N}{\partial x}$

soit $u_N = \sum_{k=-K}^{K} a_k e^{ikx}$

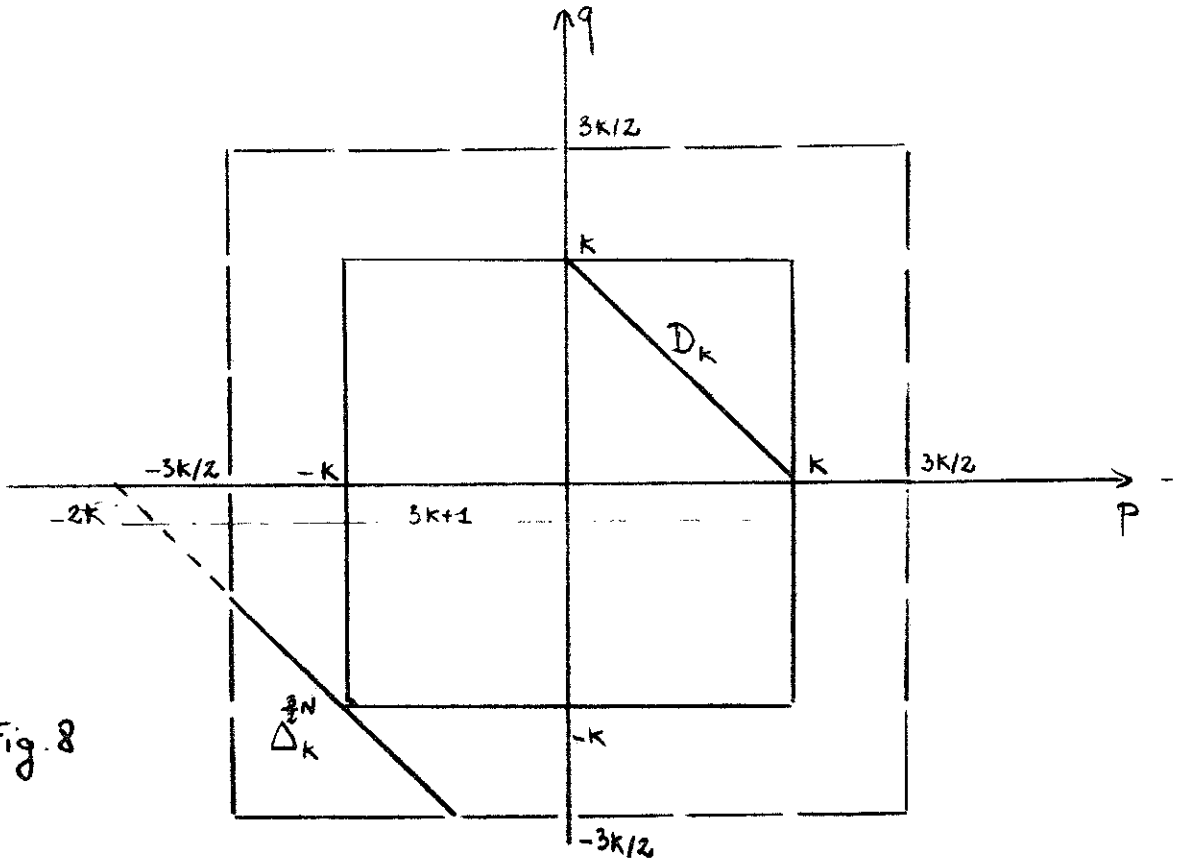



Fig. 8

Si $\Delta_{p_k}^{\frac{3}{2}N} \cap \mathcal{C} \neq \emptyset$ alors $|p_k| > K$ $\quad \mathcal{C} = [-K, K]^2$

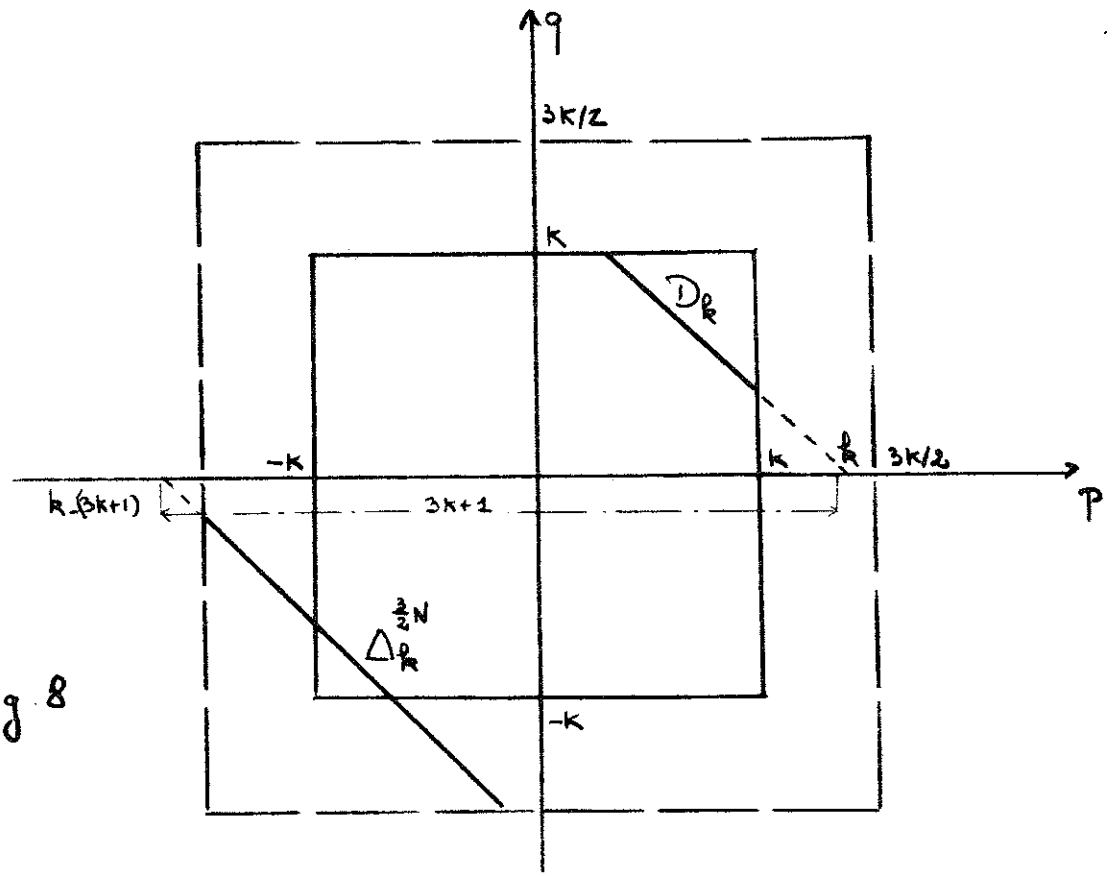



Fig. 8

Montrons que $P_N^{\perp} P_{\frac{3}{2}N}^{c} \; u_N \frac{\partial u_N}{\partial x} = P_N^{\perp} u_N \frac{\partial u_N}{\partial x}$

$$P_{\frac{3}{2}N}^{c} \; u_N \frac{\partial u_N}{\partial x} = \sum_{k=-2K}^{2K} \Big( \sum_{(p,q)\in D_k} i p a_p a_q \Big) e^{ikx} + \sum_{k=-2K}^{2K} \Big( \sum_{(p,q)\in \Delta_k^{\frac{3}{2}N}} i p a_p a_q \Big) e^{ikx}$$

avec $\Delta_k^{\frac{3}{2}N} = \{ (p,q) \in [-3K, 3K]^2 \;/\; p+q \equiv k \text{ modulo } 3K+1 \; (= \tfrac{3N-1}{2}) \}$

Par un argument géométrique très simple (Fig. 8) on constate que pour $k \in [-K, K]$ le segment $\Delta_k^{\frac{3}{2}N}$ et le carré $\mathcal{C} = [-K, K]^2$ sont disjoints, si bien que $\forall k \in [-K, K] : \sum_{(p,q)\in \Delta_k^{\frac{3}{2}N}} i p a_p a_q = 0$

En projettant par $P_N^{\perp}$ les autres termes de repliement ($|k| > K$) disparaissent. La méthode de Galerkin peut donc s'écrire $P_N^{\perp} R_N = P_N^{\perp} P_{\frac{3}{2}N}^{\perp} R_N = 0$
Les calculs s'effectuent alors à l'aide de transformations rapides sur des tableaux de dimension une fois et demis plus grande que pour la collocation non "desaliasée".

Remarque :

. Pour $F(u) = u^n$ on peut "désaliaser" avec $P_{\frac{n+1}{2}N}^{c}$ mais pour F(u) = sin u par exemple il n'existe pas de procédé immédiat. (on peut essayer un changement de variable $v = e^{iu}$ par exemple).

## 11. Une méthode "sans nom".

Pour certaines fonctions F non linéaires on peut économiser du temps de calcul en utilisant des variantes de la collocation. C'est le cas de l'équation de Burgers où l'identité $u \frac{\partial u}{\partial x} = \frac{1}{2} \frac{\partial}{\partial x} (u^2)$ peut déconomiser une FFT.

On définit alors un résidu différent des précédents :

$$\tilde{R}_N = \frac{\partial u_N}{\partial t} + \frac{1}{2} \frac{\partial}{\partial x} P_N^{c} (u_N^2) - \nu \frac{\partial^2 u_N}{\partial x^2} - f$$

au lieu de

$$R_N = \frac{\partial u_N}{\partial t} + \frac{1}{2} \frac{\partial}{\partial x} (u_N^2) - \nu \frac{\partial^2 u_N}{\partial x^2} - f$$

Cette nouvelle méthode s'écrit $P_N^c \tilde{R}_N = 0$ au lieu de la collocation $P_N^c R_N = 0$

Comparons ces deux méthodes :

$$P_N^c R_N - P_N^c \tilde{R}_N = \frac{1}{2} P_N^c \frac{\partial}{\partial x} u_N^2 - \frac{1}{2} \frac{\partial}{\partial x} P_N^c u_N^2 = \frac{1}{2} \left[ P_N^c, \frac{\partial}{\partial x} \right] u_N^2$$

Or le commutateur $\left[ P_N^c, \frac{\partial}{\partial x} \right]$ n'est pas nul :

soit $v = \sum_{k \in \mathbb{Z}} B_k e^{ikx}$ . Un calcul simple montre que

$$\left[ P_N^c, \frac{\partial}{\partial x} \right] v = iN \sum_{k=-K}^{K} \left( \sum_{r \in \mathbb{Z}} B_{k+rN} \right) e^{ikx}$$

Dans le cas de l'équation de Burgers les coefficients spectraux de $\left[ P_N^c, \frac{\partial}{\partial x} \right] (u_N^2)$ sont $iN \sum_{(p,q) \in \Delta_k} iq\, a_p a_q$

On remarque alors que ce terme provient d'un aliasing plus élevé dans la méthode de collocation classique que dans la méthode "sans nom" (Cependant ce terme proportionnel à N est compensé par la petitesse des coefficients $a_p$ et $a_q$ sur les segments de droites $\Delta_k$). Cette dernière méthode est donc plus juste, en comparaison avec le système exact. La raison est la suivante : la dérivation s'applique à une fonction contenue dans $B_N$ dans le cas de méthode "sans nom" et dans $B_{2N}$, qui est plus gros, dans le cas de la collocation.

Remarque : comme pour la collocation cette méthode donne l'approximation de Galerkin lorsque l'on "désaliase" en multipliant par 2 ou 1,5 le nombre de points de collocation. (Voir I.10 p 131 ).

Exemples :

Equation d'Euler

$$\begin{cases} \frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j} (u_i u_j) = - \frac{\partial}{\partial x_i} P & i = 1,3 \\ \text{div } u = 0 \end{cases}$$

Un exemple deméthode économisant une FFT est due à C. Basdevant. Le procédé consiste à écrire la conservation de la quantité de mouvement

sous la forme :

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{u^2-v^2}{3}\right) + \dfrac{\partial}{\partial x}\left(\dfrac{u^2-w^2}{3}\right) + \dfrac{\partial}{\partial y}(vu) + \dfrac{\partial}{\partial z}(wu) = -\dfrac{\partial}{\partial x}\left(p + \dfrac{e}{3}\right) \\ \dfrac{\partial v}{\partial t} + \dfrac{\partial}{\partial y}\left(\dfrac{v^2-u^2}{3}\right) + \dfrac{\partial}{\partial y}\left(\dfrac{v^2-w^2}{3}\right) + \dfrac{\partial}{\partial x}(uv) + \dfrac{\partial}{\partial z}(wv) = -\dfrac{\partial}{\partial y}\left(p + \dfrac{e}{3}\right) \\ \dfrac{\partial w}{\partial t} + \dfrac{\partial}{\partial z}\left(\dfrac{w^2-u^2}{3}\right) + \dfrac{\partial}{\partial z}\left(\dfrac{w^2-v^2}{3}\right) + \dfrac{\partial}{\partial x}(uw) + \dfrac{\partial}{\partial z}(vw) = -\dfrac{\partial}{\partial z}\left(p + \dfrac{e}{3}\right) \end{cases}$$

avec $u = u_1$ , $v = u_2$ , $w = u_3$ et $e = u^2 + v^2 + w^2$ .

Au lieu d'effectuer la projection de collocation de $u^2$, $v^2$ et $w^2$ il suffit de projeter $u^2 - v^2$ et $v^2 - w^2$. La projection de $w^2 - v^2$ s'obtient par différence.

# CHAPITRE II : PRECISION DES METHODES SPECTRALES

## 1. Introduction.

Les énoncés de ce chapitre essaient plus de donner une idée du problème de précision des méthodes spectrales que de présenter la démonstration rigoureuse des résultats énoncés. Les détails se trouvent dans le Gottlieb-Orszag [1] aux sections 3 et 6.

## 2. Problème de Sturm-Liouville et bases complètes.

On appelle problème de Sturm-Liouville l'équation différentielle suivante :

$$\begin{cases} A\,\phi(x) = g(x) \qquad x \in [a,b] \\ + \quad 2 \text{ conditions aux limites} \end{cases}$$

où A est un opérateur du second ordre très général :

$$A = \frac{1}{w(x)}\left[\frac{d}{dx}\,p(x)\,\frac{d}{dx} + q(x)\right]$$

On peut montrer que l'opérateur qui résoud ce problème $g \longrightarrow \phi$ est compact dans $L^2(a,b)$. C'est pourquoi un grand nombre de bases complètes sont obtenues comme fonctions propres d'un problème de Sturm-Liouville :

$$\begin{cases} A\phi_n = \lambda_n \phi_n \qquad n \in \mathbb{N} \\ + \quad 2 \text{ conditions aux limites} \end{cases}$$

La compacité entraîne $\lambda_n \underset{n\to\infty}{\longrightarrow} \infty$. On peut montrer que $\lambda_n$ croit comme $n^2$ et normaliser la base pour que $\|\phi_n\|$ soit borné.

$f$ et $g \in \mathcal{C}^2(a,b)$

$$(f, Ag)_{L^2} = \int_a^b f \frac{(pg')'}{w} w \, dx$$

$$= \int_a^b \frac{p}{\Gamma} (pg')' \, dx$$

$$= [pfg']_a^b - \int_a^b pf'g' dx$$

$$= [p(fg' - f'g)]_a^b + \int_a^b (pf')' g \, dx$$

$$= \left[ p \begin{vmatrix} f & g \\ f' & g' \end{vmatrix}_x \right]_a^b + (Af, g)_{L^2}$$

Intégration par parties

Fig. 1

## 3. Décroissance des coefficients spectraux.

On note $(f,g)_{L^2} = \int_a^b f(x)\,g(x)\,w(x)\,dx$ le produit scalaire de $L^2([a,b], w(x)dx)$.

Tout ce paragraphe repose sur l'intégration par parties suivante (Fig. 1)

$\forall f$ et $g \in \mathcal{C}^2(a,b)$

$$(f, Ag)_{L^2} = p(a)\begin{vmatrix} f(a) & g(a) \\ f'(a) & g'(a) \end{vmatrix} - p(b)\begin{vmatrix} f(b) & g(b) \\ f'(b) & g'(b) \end{vmatrix} + (Af, g)_{L^2}$$

On notera le Wronskien de f et g de la façon suivante

$$[f,g]_x = \begin{vmatrix} f(x) & g(x) \\ f'(x) & g'(x) \end{vmatrix}$$

Soit $f(x) = \sum_{n\in\mathbb{N}} a_n \phi_n(x)$ la décomposition de f sur une base de fonctions propres d'un problème de Sturm-Liouville. Pour étudier la décroissance de $a_n$ il faut distinguer deux cas : les problèmes de Sturm-Liouville singuliers ou non-singuliers.

### a) Problèmes de Sturm-Liouville non-singuliers.

Ils sont tels que $p(a) \neq 0$ ou $p(b) \neq 0$. Calculons $a_n$ :

$$a_n = (f, \phi_n)_{L^2} = \frac{1}{\lambda_n}(f, A\phi_n)_{L^2} \quad \text{car } \phi_n \text{ fonction propre de A}$$

$$= \frac{1}{\lambda_n}\left\{ p(a)[f,\phi_n]_a - p(b)[f,\phi_n]_b + (Af,\phi_n)_{L^2} \right\} \qquad (1)$$

On voit donc que $a_n$ décroit seulement comme $1/\lambda_n$ à moins que soit vérifié : $p(a)[f,\phi_n]_a - p(b)[f,\phi_n]_b = 0$ (2)

Dans ce cas on effectue une nouvelle intégration par partie :

$$a_n = \frac{1}{\lambda_n^2}\left\{ p(a)[Af,\phi_n]_a - p(b)[Af,\phi_n]_b + (A^2 f,\phi_n)_{L^2} \right\}$$

et ainsi de suite.

Si bien qu'une fonction $f \in \mathcal{C}^\infty[a,b]$ doit vérifier une infinité de conditions aux limites pour que ses coefficients $a_n$ décroissent plus vite que toute puissance de $\frac{1}{\lambda_n}$ et donc de $\frac{1}{n}$ .

### b) Problèmes de Sturm-Liouville singuliers.

Ils vérifient p(a) = p(b) = 0. Dans ce cas

$$a_n = (f, \phi_n)_{L^2} = \frac{1}{\lambda_n} (f, A\phi_n)_{L^2} = \frac{1}{\lambda_n} (Af, \phi_n)_{L^2}$$

On peut continuer ces intégrations par parties tant que f est assez régulière. Si f est de classe $\mathcal{C}^{2p}$

$$a_n = \frac{1}{\lambda_n^p} (A^p f, \phi_n)_{L^2}$$

On voit donc que la décroissance des $a_n$ ne dépend que de la régularité de la fonction f sur $[a,b]$. Si f est de classe $\mathcal{C}^\infty$, $a_n$ décroit alors plus vite que toute puissance de $\frac{1}{\lambda_n}$ donc de $\frac{1}{n}$.

### c) Conclusion.

A moins d'étudier des fonctions vérifiant une infinité de conditions aux limites (par exemple périodiques) il faut utiliser une base de fonctions obtenues à partir d'un problème de Sturm-Liouville singulier.

## 4. Exemples de bases.

### a) Problèmes de Sturm-Liouville non singuliers.

. $(\sin n x)_{n=1,\infty}$ est une base orthogonale complète de solution de :

$$\begin{cases} \frac{d^2}{dx^2} \phi_n(x) = n^2 \phi_n(x) & x \in [0,\pi] \\ \phi_n(0) = \phi_n(\pi) = 0 \end{cases}$$

. $(\cos n x)_{n=0,\infty}$ est une base orthogonale de solution de :

$$\begin{cases} \frac{d^2}{dx^2} \phi_n(x) = n^2 \phi_n(x) & x \in [0,\pi] \\ \phi_n'(0) = \phi_n'(\pi) = 0 \end{cases}$$

Pour résoudre $\begin{cases} \dfrac{\partial u}{\partial t} = F(u) + \dfrac{D}{I} \\ CL \end{cases}$

On utilise souvent dans la pratique les bases :

| Conditions aux limites | Base de la décomposition spectrale | provenant d'un problème de Sturm-Liouville |
|---|---|---|
| périodiques | $(e^{ikx})\ k \in \mathbb{Z}$ | non singulier |
| non périodiques | Tchebyshev $(T_n(x))\ n \in \mathbb{N}$ | singulier |

Choix de la décomposition spectrale

Fig. 2

. $(e^{inx})_{n\in\mathbb{Z}}$ est une base orthogonale complète de $L^2(\Gamma,\mathbb{C})$ solution de :

$$\frac{d^2}{dx^2}\phi_n(x) = n^2\phi_n(x) \quad x\in\Gamma \text{ cercle unité}$$

b) Problèmes de Sturm-Liouville singuliers.

. Les polynômes de Chebyshev $(T_n(x))$ n = 0,∞ forment une base orthogonale complète de $L^2([-1,1], dx/\sqrt{1-x^2})$ solution de

$$\begin{cases} \frac{1}{\sqrt{1-x^2}}\frac{d}{dx}\sqrt{1-x^2}\frac{d}{dx}T_n(x) + n^2 T_n(x) = 0 \\ T_n'(1) = n^2 \ , \quad T_n'(-1) = (-1)^{n-1} n^2 \end{cases}$$

. Les polynômes de Legendre $(P_n(x))n = 0,$ forment une base orthogonale complète de $L^2(-1,1)$ solution de

$$\begin{cases} \frac{d}{dx}(1-x^2)\frac{d}{dx}P_n(x) + n(n+1)P_n(x) = 0 \\ P_n(1) = 1 \quad , \quad P_n(-1) = -1 \end{cases}$$

Pour fixer les idées on peut dire que pour résoudre par méthodes spectrales des équations aux dérivées partielles on utilise en pratique la base $(e^{inx})n\in\mathbb{Z}$ pour les conditions aux limites périodiques et la base $(T_n(x))n\in\mathbb{N}$ pour les conditions aux limites non périodiques (Fig. 2).

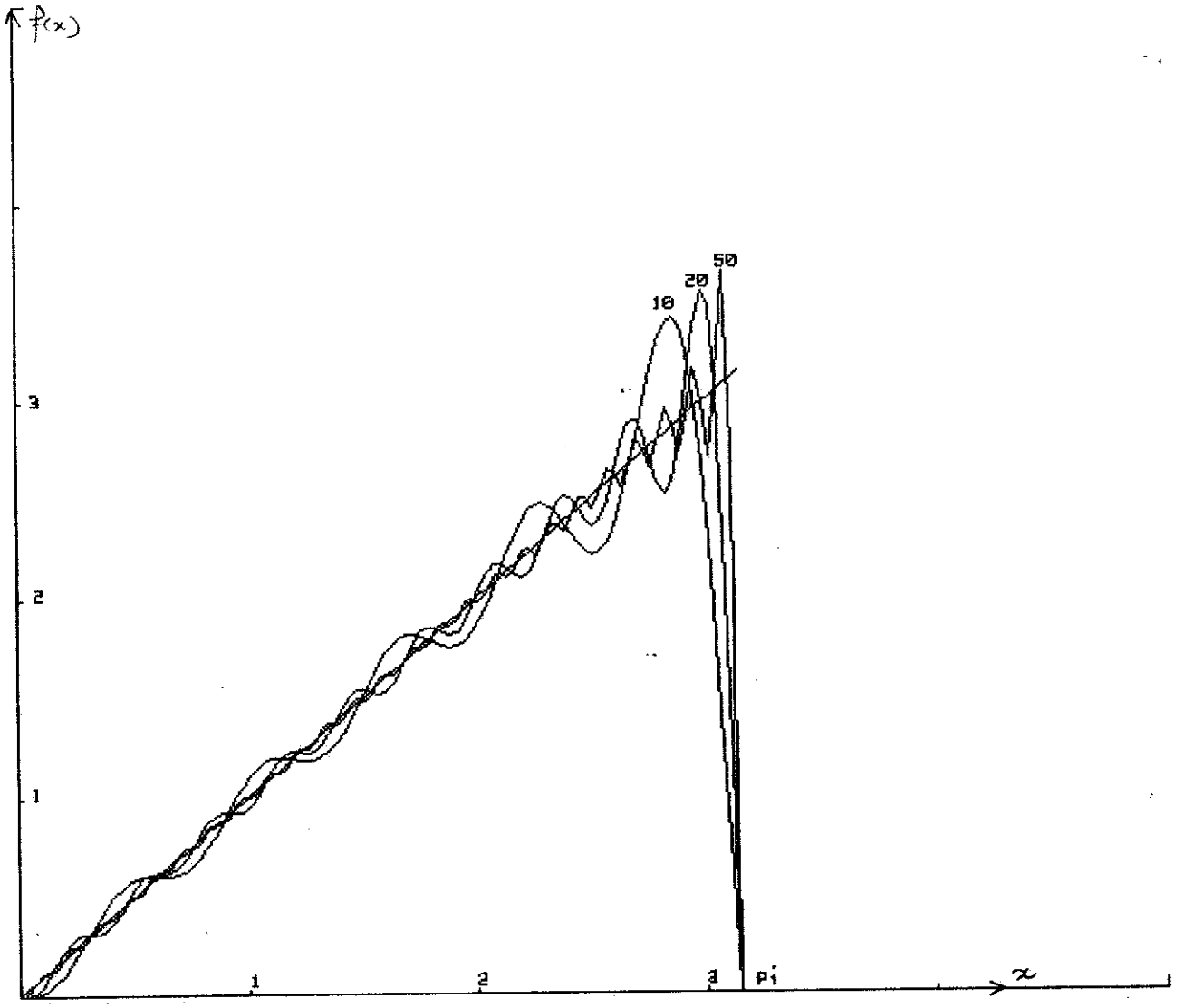
## 5. Phénomènes de Gibbs.

Montrons sur un exemple l'effet de la décroissance des coefficients spectraux sur la convergence de $f_N(x) = \sum_{n=0}^{N} f_n \varphi_n(x)$ vers $f(a) = \sum_{n=0}^{\infty} f_n \varphi_n(x)$

Sur [-1, 1] on peut développer la fonction x :

$$x = 2\sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nx)$$

La base (sin(n.x))provient d'un problème de Sturm-Liouville non singulier et "x" ne vérifie pas l'infinité de conditions aux limites




PHÉNOMÈNE DE GIBBS

Approximation de $f(x) = x$ sur $[0, \pi]$ pour

$N = 10$ , $N = 20$ , $N = 50$

La convergence n'est pas uniforme

Fig 3

qu'il faudrait (c'est à dire $f^{(2p)}(0) = f^{(2p)}(\pi) = 0$ )
C'est pourquoi ses coefficients spectraux décroissent seulement comme $\frac{1}{n}$ malgré sa régularité $\mathcal{C}^\infty[0,\pi]$. Il se produit alors le phénomène de Gibbs : $f_N(x) = 2\sum_{n=0}^{N} \frac{(-1)^{n+1}}{n} \sin(nx)$ ne converge pas uniformément vers f(x) = x (voir figure 3).

## 6. Consistance des méthodes spectrales.

Il se produit parfois un phénomène plus grave que le phénomène de Gibbs lorsqu'on applique une méthode spectrale avec des fonctions de bases provenant d'un problème de Sturm-Liouville non singuliers, et non adaptées aux conditions aux limites de l'équation aux dérivées partielles à résoudre. Par exemple :

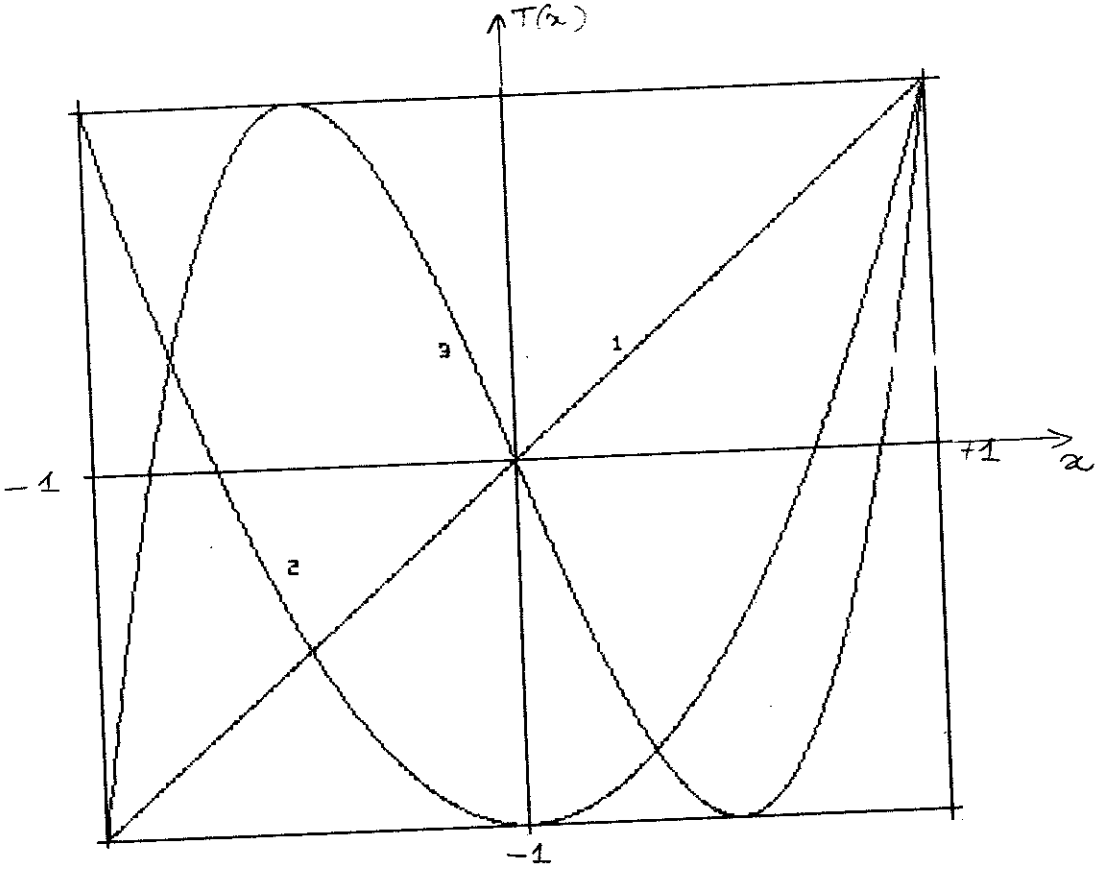
$$\begin{cases} \frac{\partial u}{\partial t}(x,t) + \frac{\partial u}{\partial x}(x,t) = x + t & x \in [0,\pi] \quad t \geq 0 \\ u(0,t) = 0 & t \geq 0 \\ u(x,0) = 0 & \forall x \end{cases}$$

avec l'approximation de Galerkin $u_N(x,t) = \sum_{n=1}^{N} a_n(t) \sin(nx)$

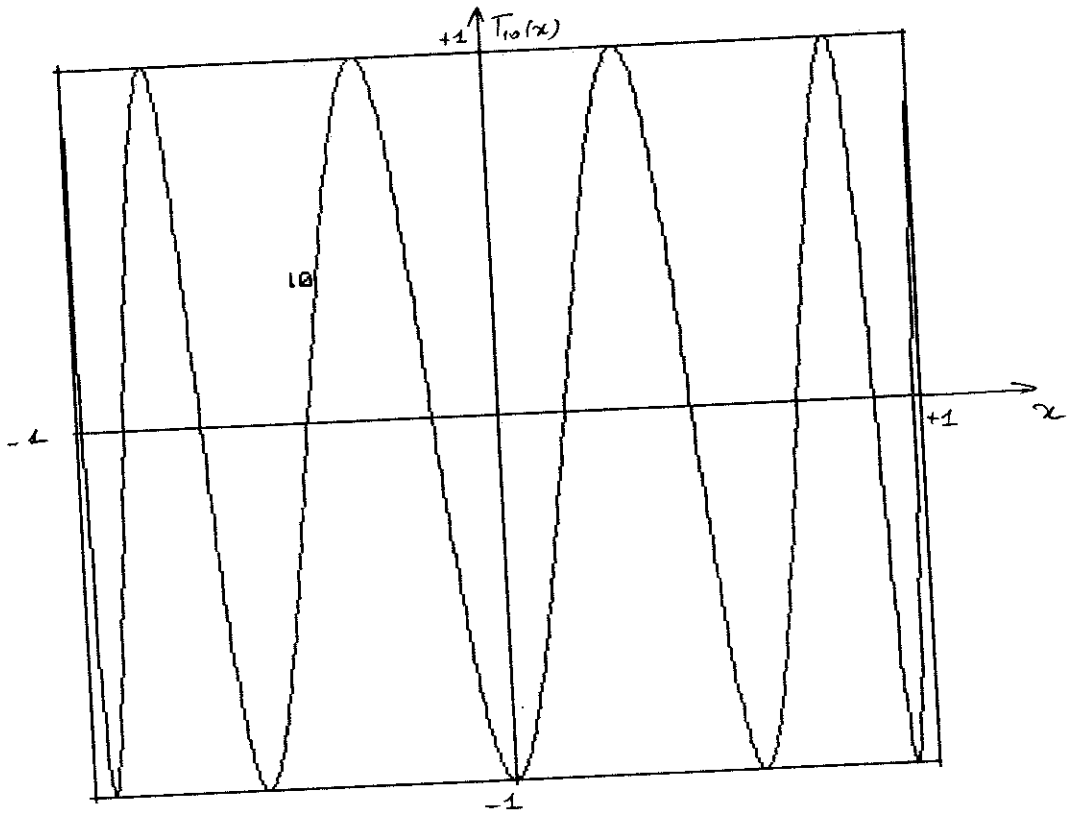
Dans ce cas $u_N(x,t)$ ne converge en aucun point de $[0,\pi]$ vers la solution exacte u = xt. Ceci est dû à l'inconsistance de l'approximation :

$$\left\| \frac{\partial}{\partial x} - P_N^{\perp} \frac{\partial}{\partial x} P_N^{\perp} \right\|_{L^2} \geq C t^2 N$$

(Gottlieb-Orszag [1_7] section 6 p. 69).




Polynômes de Tchebyshev $T_1$, $T_2$, $T_3$




Polynômes de Tchebyshev $T_{10}$

Fig. 1

# CHAPITRE III : POLYNOMES DE TCHEBYSHEV

## 1. Définition.

Les polynômes de Tchebyshev sont définis sur $[-1,1]$ par :

$$T_n(x) = \cos(n.\operatorname{Arccos} x) \qquad n \in \mathbb{N}$$

ou encore par

$$\begin{cases} T_n(x) = \cos n\theta & n \in \mathbb{N} \\ x = \cos\theta \end{cases}$$

(Fig. 1)

a) Relation de récurrence : $T_{n+1}(x) + T_{n-1}(x) = 2x\,T_n(x) \quad n \geq 1$

En effet $\cos[(n+1)\theta] + \cos[(n-1)\theta] = 2\cos\theta\cos n\theta$

Ceci permet de calculer explicitement les premier polynômes :

$T_0 = 1$ ; $T_1 = x$ ; $T_2 = 2x^2 - 1$ ; $T_3 = 4x^3 - 3$ ; etc...

b) Propriétés simples

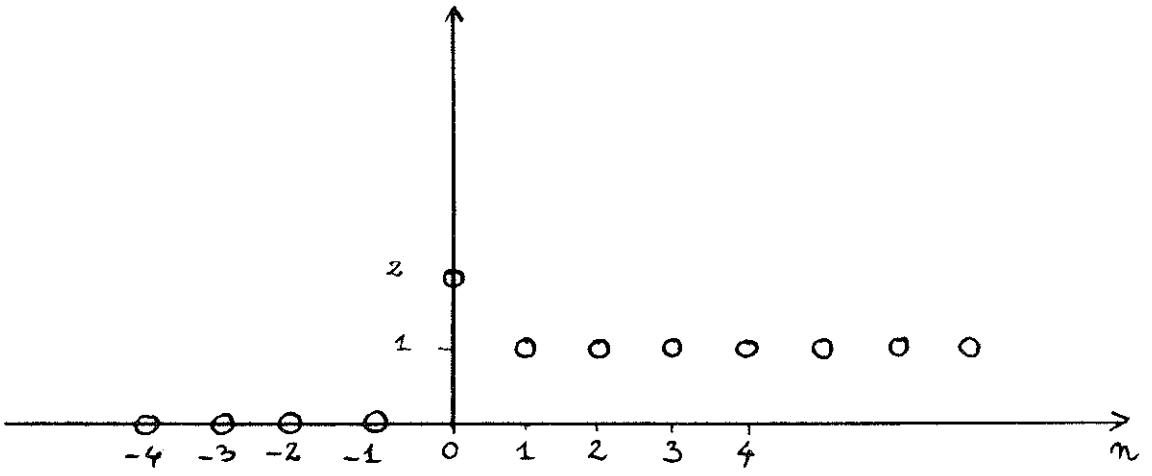
. $T_n(1) = 1$ ; $T_n(-1) = (-1)^n$

. $T_n$ est une fonction paire si n est pair, impaire si n est impair.

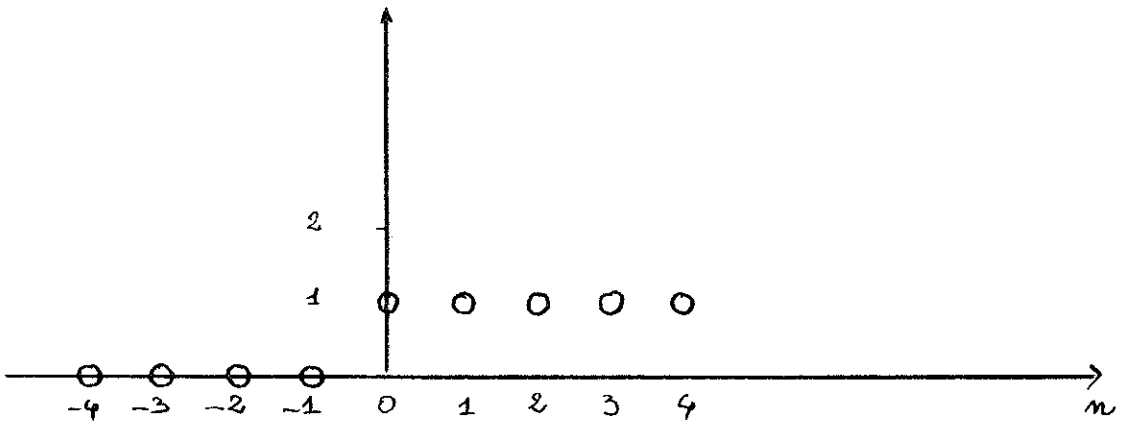
. $T_n'(1) = n^2$ ; $T_n'(-1) = (-1)^{n-1} n^2$

En effet

$$T_n'(x) = \frac{d\theta}{dx}\frac{d}{d\theta}\cos n\theta = n\,\frac{\sin n\theta}{\sin\theta}$$




Suite $c_n$




Suite $d_n$

Fig 2

## 2. Relations de récurrence entre polynômes.

Définissons quelques suites qui permettront de condenser les relations de récurrence en des formulations valables pour $n \in \mathbb{N}$ (Fig. 2).

$(c_n)_{n\in\mathbb{Z}}$ : $c_n = 0$ si $n<0$ , $c_0 = 2$ , $c_n = 1$ si $n>0$

$(d_n)_{n\in\mathbb{Z}}$ : $d_n = 0$ si $n<0$ , $d_n = 1$ si $n \geqslant 0$

a) $\quad c_n T_{n+1}(x) + d_{n-1} T_{n-1}(x) = 2x T_n(x) \quad n \in \mathbb{N}$

Cette relation résume les égalités suivantes :

$$\begin{cases} T_{n+1}(x) + T_{n-1}(x) = 2x T_n(x) & n \geqslant 1 \\ 2 T_1(x) = 2x T_0(x) \end{cases}$$

b) $\quad c_n \dfrac{T'_{n+1}(x)}{n+1} - d_{n-2} \dfrac{T'_{n-1}(x)}{n-1} = 2 T_n(x) \quad n \in \mathbb{N}$

Cette relation résume les égalités suivantes :

$$\begin{cases} \dfrac{T'_{n+1}(x)}{n+1} - \dfrac{T'_{n-1}(x)}{n-1} = 2 T_n(x) & n \geqslant 2 \\ \dfrac{T'_2(x)}{2} = 2 T_1(x) \\ 2 T'_1(x) = 2 T_0(x) \end{cases}$$

Pour $k \geqslant 1$ le changement de variable $x = \cos\theta$ permet de calculer :

$$T'_k(x) = \frac{d\theta}{dx} \frac{d}{d\theta} \cos k\theta = k \frac{\sin k\theta}{\sin\theta}$$

D'où pour $n \geqslant 2$

$$\frac{T'_{n+1}(x)}{n+1} - \frac{T'_{n-1}(x)}{n-1} = \frac{\sin[(n+1)\theta] - \sin[(n-1)\theta]}{\sin\theta} = 2\cos n\theta = 2T_n(x)$$

## 3. Calcul des coefficients de Tchebyshev de F(u).

Les deux relations de récurrence précédentes vont permettre de calculer les coefficients du développement $F(u) = \sum_{n\in\mathbb{N}} b_n T_n(x)$ à partir des coefficients de $u = \sum_{n\in\mathbb{N}} a_n T_n(x)$, pour de nombreuses fonctions F standards. Voici quelques exemples fondamentaux.

### a) F(u) = x u(x).

$$u(x) = \sum_{n=0}^{\infty} a_n T_n(x) \qquad xu(x) = \sum_{n=0}^{\infty} b_n^{[1]} T_n(x)$$

On se sert de la relation

$$c_n T_{n+1}(x) + d_{n-1} T_{n-1}(x) = 2x T_n(x) \qquad n \in \mathbb{N}$$

dont la combinaison linéaire avec les $a_n$ donne :

$$\sum_{n=0}^{\infty} c_n a_n T_{n+1}(x) + \sum_{n=0}^{\infty} d_{n-1} a_n T_{n-1}(x) = 2x \sum_{n=0}^{\infty} a_n T_n(x)$$

On en déduit

$$2x\,u(x) = \sum_{n=1}^{\infty} c_{n-1} a_{n-1} T_n(x) + \sum_{n=-1}^{\infty} d_n a_{n+1} T_n(x)$$

$$= \sum_{n=0}^{\infty} (c_{n-1} a_{n-1} + a_{n+1}) T_n(x)$$

Pour le premier terme, $C_{n-1}$ a permis la sommation n = 0,∞
Pour le second $d_n$ a été remplacé par la sommation n = 0,∞

Donc

$$\boxed{2 b_n^{[1]} = c_{n-1} a_{n-1} + a_{n+1} \qquad n \in \mathbb{N}} \qquad (3)$$

### b) F(u) = $x^2$ u(x).

$$u(x) = \sum_{n=0}^{\infty} a_n T_n(x) \qquad x^2 u(x) = \sum_{n=0}^{\infty} b_n^{[2]} T_n(x)$$

On se sert de la relation précédente appliquée aux deux fonctions $x.u(x)$ et $x^2.u(x)$ :

$$2 b_n^{[2]} = c_{n-1} b_n^{[1]} + b_{n+1}^{[1]} \qquad n \in \mathbb{N}$$

$$\Rightarrow \qquad 4 b_n^{[2]} = c_{n-1} [c_{n-2} a_{n-2} + a_n] + [c_n a_n + a_{n+2}]$$

En remarquant que $C_{n-1}.C_{n-2} = C_{n-2}$ on obtient

$$\boxed{4\, b_n^{[2]} = c_{n-2}\, a_{n-2} + (c_n + c_{n-1})\, a_n + a_{n+2}} \qquad (4)$$

c) F(u) = u'(x) : expressions implicites et explicites.

$$u(x) = \sum_{n=0}^{\infty} a_n T_n(x) \qquad u'(x) = \sum_{n=0}^{\infty} a_n^{(1)} T_n(x)$$

On se sert de la relation

$$c_n \frac{T'_{n+1}(x)}{n+1} - d_{n-2} \frac{T'_{n-1}(x)}{n-1} = 2\, T_n(x) \qquad n \in \mathbb{N}$$

dont la combinaison linéaire avec les $a_n^{(1)}$ donne :

$$\sum_{n=0}^{\infty} c_n a_n^{(1)} \frac{T'_{n+1}(x)}{n+1} - \sum_{n=0}^{\infty} d_{n-2}\, a_n^{(1)} \frac{T'_{n-1}(x)}{n-1} = 2 \sum_{n=0}^{\infty} a_n^{(1)} T_n(x)$$

On en déduit

$$2\, u'(x) = \sum_{n=1}^{\infty} c_{n-1}\, a_{n-1}^{(1)} \frac{T'_n(x)}{n} - \sum_{n=0}^{\infty} d_{n-1}\, a_{n+1}^{(1)} \frac{T'_n(x)}{n}$$

$$= \sum_{n=1}^{\infty} \left[ c_{n-1}\, a_{n-1}^{(1)} - a_{n+1}^{(1)} \right] \frac{T'_n(x)}{n}$$

Pour le second terme, $d_{n-1}$ a été remplacé par la sommation n = 1, ∞

En identifiant avec $u'(x) = \sum_{n=1}^{\infty} a_n T'_n(x)$ on obtient l'expression implicite des $a_n^{(1)}$ :

$$\boxed{2n\, a_n = c_{n-1}\, a_{n-1}^{(1)} - a_{n+1}^{(1)} \qquad n \in \mathbb{N}} \qquad (5)$$

Ceci permet de calculer explicitement les $a_n^{(1)}$ par récurrence

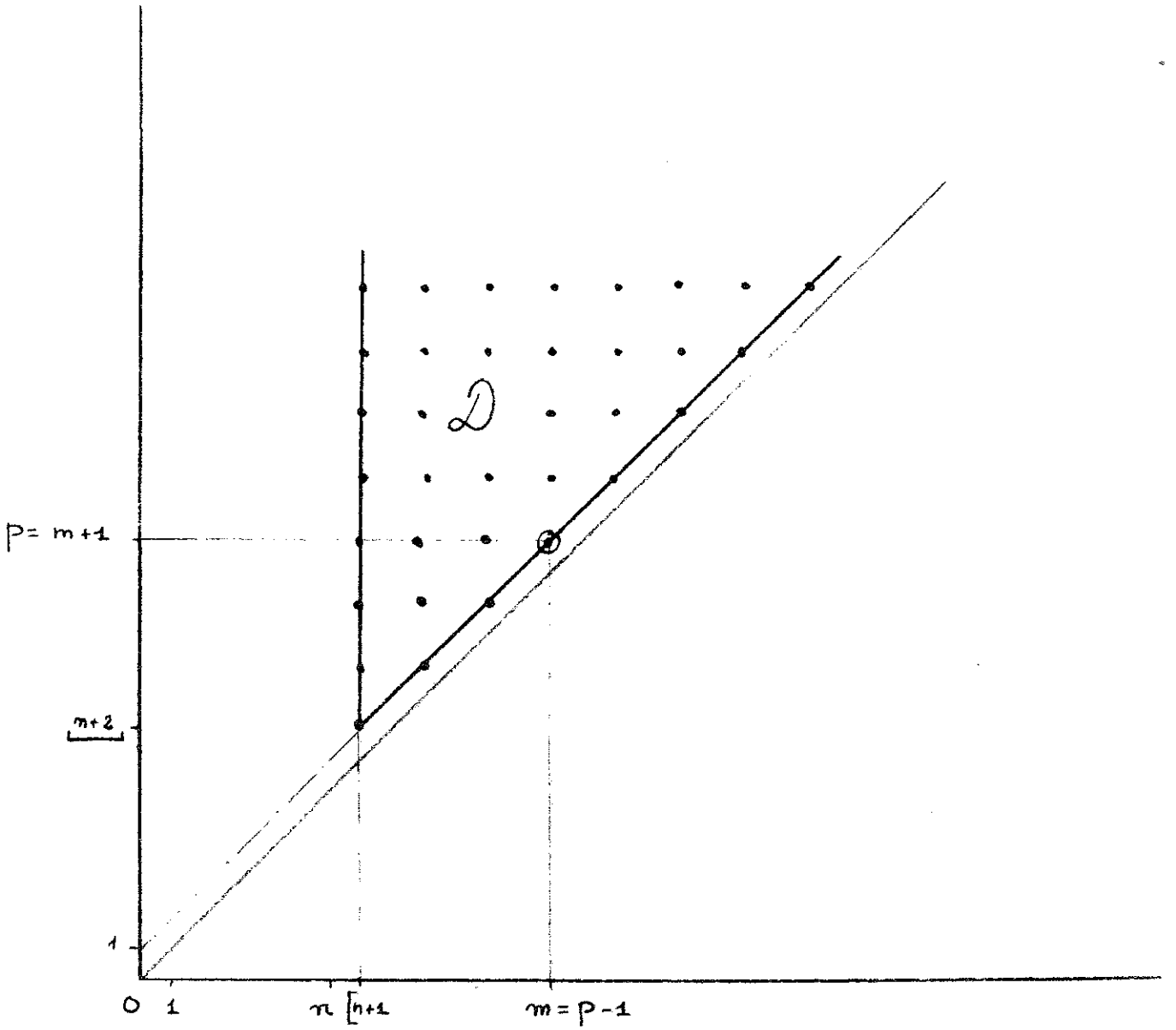
$$c_n a_n^{(1)} = 2(n+1)\, a_{n+1} + a_{n+2}^{(1)}$$

$$= 2(n+1)\, a_{n+1} + 2(n+3)\, a_{n+3} + a_{n+5}^{(1)}$$

$$= \cdots\cdots$$

$$= 2 \sum_{\substack{p=n+1 \\ p \text{ step } 2}}^{\infty} p\, a_p$$

Où la notation "step 2" signifie que l'incrément de la sommation est 2.




$$\sum_{m=n+1 \text{ step } 2}^{\infty} \sum_{p=m+1 \text{ step } 2}^{\infty} = \sum_{p=n+2 \text{ step } 2}^{\infty} \sum_{m=n+1 \text{ step } 2}^{p-1}$$

Inversion de l'ordre des sommations

Fig 3

Donc

$$\boxed{a_n^{(1)} = \frac{2}{c_n} \sum_{\substack{p=n+1 \\ \text{step } 2}}^{\infty} p\, a_p} \qquad (6)$$

d) F(u) = u"(x) : expression explicite

$$u(x) = \sum_{n=0}^{\infty} a_n T_n(x) \qquad u''(x) = \sum_{n=0}^{\infty} a_n^{(2)} T_n(x)$$

Le plus simple est d'utiliser la relation précédente (6) appliquée aux fonctions u" et u' :

$$a_n^{(2)} = \frac{2}{c_n} \sum_{\substack{m=n+1 \\ \text{step } 2}}^{\infty} m\, a_m^{(1)} \qquad n \in \mathbb{N}$$

puis avec u' et u :

$$a_m^{(1)} = 2 \sum_{\substack{p=m+1 \\ \text{step } 2}}^{\infty} p\, a_p \qquad m \geq 1$$

On a donc en regroupant les égalités :

$$a_n^{(2)} = \frac{2}{c_n} \sum_{\substack{m=n+1 \\ \text{step } 2}}^{\infty} \sum_{\substack{p=m+1 \\ \text{step } 2}}^{\infty} m\, p\, a_p$$

Le domaine de sommation $\mathcal{D}$ représenté sur la figure 3 peut être balayé d'une façon différente :

$$a_n^{(2)} = \frac{4}{c_n} \sum_{\substack{p=n+2 \\ \text{step } 2}}^{\infty} p\, a_p \sum_{\substack{m=n+1 \\ \text{step } 2}}^{\infty} m$$

Un calcul détaillé dans l'appendice (III.6 p.165) montre que

$$\sum_{\substack{m=n+1 \\ \text{step } 2}}^{p-1} m = \frac{p^2 - n^2}{4}$$

Donc

$$\boxed{a_n^{(2)} = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ \text{step } 2}} p(p^2 - n^2)\, a_p} \qquad (7)$$

e) F(u) = u"(x) : expression implicite.

Dans certains problèmes comme la résolution de l'équation de Poisson il peut être intéressant de connaître non pas $a_n^{(2)}$ comme expression explicite des $a_n$ mais $a_n$ en fonction des $a_n^{(2)}$.

On utilise la relation (5) avec les fonctions u' et u

$$2n\,a_n = c_{n-1}\,a^{(1)}_{n-1} - a^{(1)}_{n+1} \quad \text{pour } n \geq 2$$

Puis avec les fonctions u'' et u'

$$2p\,a^{(1)}_p = c_{p-1}\,a^{(2)}_{p-1} - a^{(2)}_{p+1} \quad \text{pour } p = n-1 \text{ et } p = n+1$$

En remplaçant $a^{(1)}_{n-1}$ et $a^{(1)}_{n+1}$ par leur expression

$$2n\,a_n = c_{n-1}\left[\frac{c_{n-2}}{2(n-1)}\,a^{(2)}_{n-2} - \frac{1}{2(n-1)}\,a^{(2)}_n\right] - \left[\frac{c_n}{2(n+1)}\,a^{(2)}_n - \frac{1}{2(n+1)}\,a^{(2)}_{n+2}\right]$$

$$= \frac{c_{n-1}\,c_{n-2}}{2(n-1)}\,a^{(2)}_{n-2} - \frac{1}{2}\left(\frac{1}{n-1} + \frac{c_n}{n+1}\right)a^{(2)}_n + \frac{1}{2(n+1)}\,a^{(2)}_{n+2}\,, \quad \forall n \geq 2$$

Or $c_{n-1}\,c_{n-2} = c_{n-2}$ et $c_n$ peut être omis car $n \geq 2$.

Donc

$$\boxed{a_n = \frac{c_{n-2}}{4n(n-1)}\,a^{(2)}_{n-2} - \frac{1}{2(n^2-1)}\,a^{(2)}_n + \frac{1}{4n(n+1)}\,a^{(2)}_{n+2}} \qquad (8)$$

## 4. Expression des conditions aux limites.

Rappelons que

$$\forall n \in \mathbb{N} \quad \begin{cases} T_n(1) = 1 & \text{et } T_n(-1) = (-1)^n \\ T_n'(1) = n^2 & \text{et } T_n'(-1) = (-1)^{n-1} n^2 \end{cases}$$

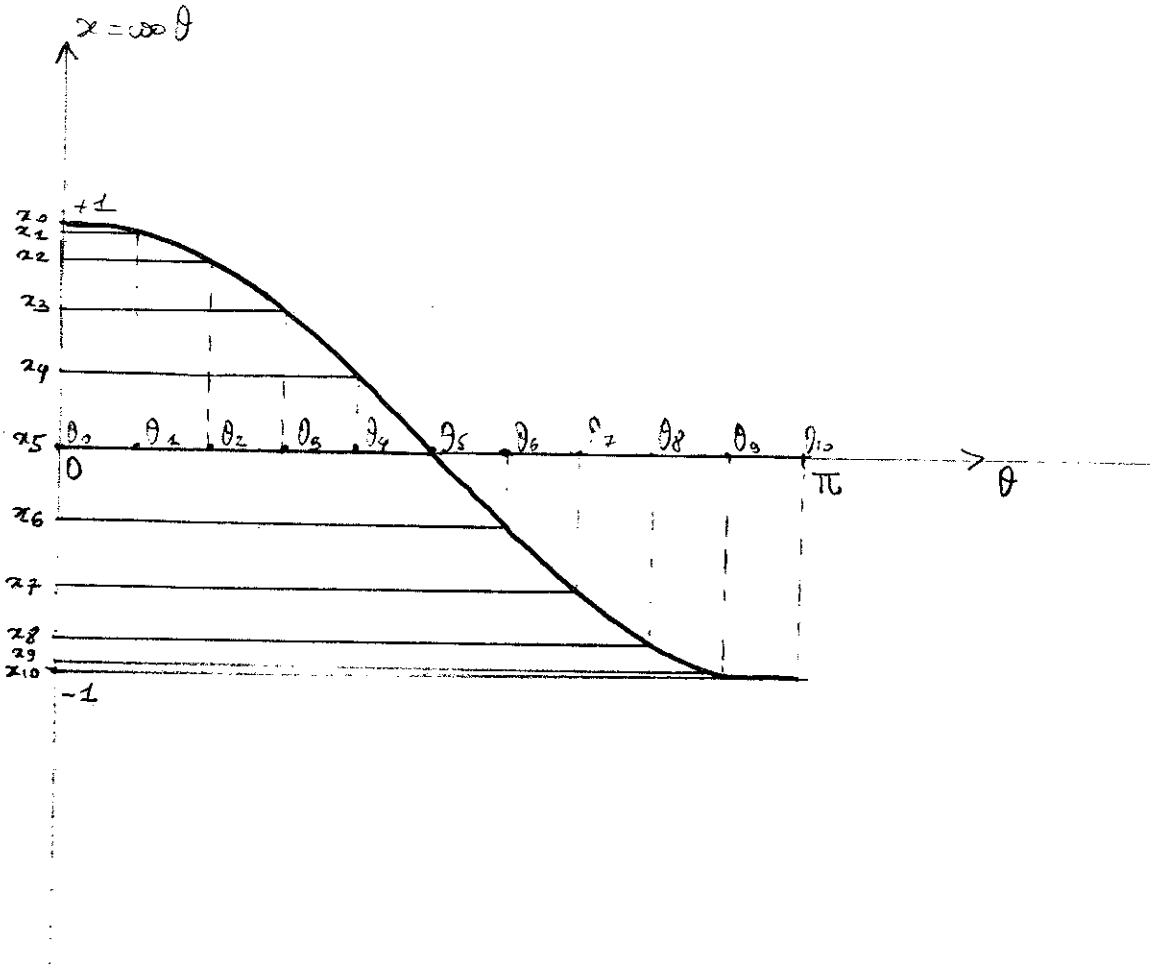
Soit $u(x) = \sum_{n=0}^{\infty} a_n T_n(x)$

Il s'agit d'exprimer les conditions aux limites à l'aide des $a_n$. Par exemple :

. $u(1) = \sum_{n=0}^{\infty} a_n$ ; $u(-1) = \sum_{n=0}^{\infty} (-1)^n a_n$

. $u'(1) = \sum_{n=0}^{\infty} a_n T_n'(1)$ ; $u'(-1) = \sum_{n=0}^{\infty} a_n T_n'(-1)$

. $= \sum_{n=0}^{\infty} n^2 a_n$ $= \sum_{n=0}^{\infty} (-1)^{n-1} n^2 a_n$




Points de collocation sur $[-1, 1]$

$$x_j = \cos \theta_j$$

$\theta_j$ régulièrement espacés sur $[0, \pi]$

Fig. 4

## 5. Collocation.

Pour appliquer la méthode de collocation (ou Tau-collocation) on pourrait à priori choisir n'importe quelle distribution de points de collocation pourvu que $\det\left(T_n(x_j)\right) \neq 0 \quad [n = 0,N \; ; \; j = 0,N]$
Mais on verra au chapitre V que pour pouvoir utiliser un algorithme de transformation rapide espace spectral-espèce physique il faut choisir les points suivants : $x_j = \cos j\frac{\pi}{N} \qquad j = 0,N$ .

Ces points sont plus ressérés aux bords de l'intervalle (Fig. 4) ce qui permet de traiter avec plus de précision les conditions aux limites, ou encore les problèmes de couche limite. Il se trouve que les extréma du polynôme $T_N(x)$ sont atteints en ces points.

## 6. Appendice.

Effectuons le calcul suivant avec $a < b$ dans $\mathbb{N}$ et $e = \frac{b-a}{2} \in \mathbb{N}$

$$\sum_{\substack{m=a \\ \text{step2}}}^{b} m = \sum_{k=0}^{e} (a+2k) = (e+1)a + (e+1)e = (e+1)(e+a)$$
$$= \frac{1}{4}(b-a+2)(b+a)$$

Pour a = m + 1 et b = p − 1 on obtient

$$\sum_{m=n+1}^{p-1} m = \frac{1}{4}(p-n)(p+n) = \frac{1}{4}(p^2 - n^2)$$

# CHAPITRE IV : RESOLUTION DE L'EQUATION DE POISSON.

## 1. Ecriture du système de Dirichlet.

Equation de Poisson avec conditions aux limites de Dirichlet

$$\begin{cases} -u''(x) + \lambda u(x) = f(x) \\ u(-1) = g_1 \qquad u(1) = g_2 \end{cases}$$

On utilise la méthode Tau avec les polynômes de Tchebyshev

Soit $f(x) = \sum_{n=0}^{\infty} f_n T_n(x)$

On recherche $u_N(x) = \sum_{n=0}^{N} a_n T_n(x)$ dans $B_N$ de dimension N + 1.

Notons $u_N''(x) = \sum_{n=0}^{N-2} a_n^{(2)} T_n(x)$

La méthode Tau consiste à résoudre le problème approché suivant

$$\begin{cases} -a_n^{(2)} + \lambda a_n = f_n \qquad n = 0, N-2 \\ \sum_{n=0}^{N} (-1)^n a_n = g_1 \quad ; \quad \sum_{n=0}^{N} a_n = g_2 \end{cases}$$

. L'idée la plus naturelle pour résoudre ce système serait d'exprimer

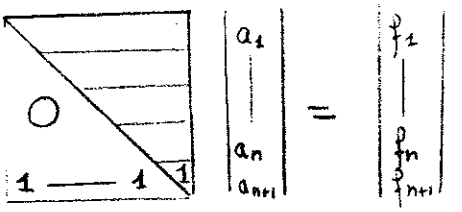
$a_n^{(2)} = \sum_{\substack{p=n+2 \\ \text{step } 2}}^{N} p(p^2 - n^2) a_p$ dans les N-1 premières égalités.

On remarque alors que ce système se découple en deux systèmes indépendants : l'un ne faisant intervenir que les coefficients $a_n$ d'indice n pair, l'autre les indices n impairs.

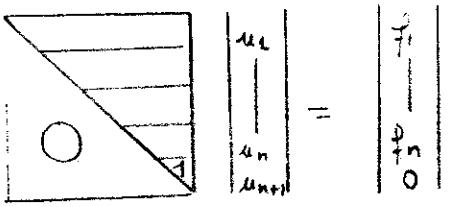
Pour découpler les deux dernières équations (conditions aux limites) il suffit d'en faire la somme et la différence :

$$\begin{cases} \sum_{n \text{ pair}} a_n = \frac{1}{2}(g_2 + g_1) = G_1 \\ \sum_{n \text{ impair}} a_n = \frac{1}{2}(g_2 - g_1) = G_2 \end{cases}$$

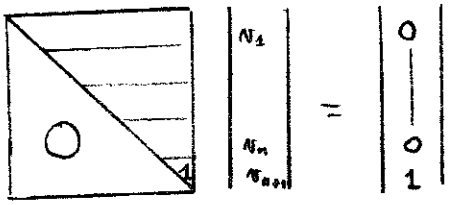
Résolution de $a$ :




On résoud $u$




On résoud $v$




On résoud $\lambda$

$$g = \sum_{i=1}^{n+1} u_i + \lambda \left( \sum_{i=1}^{n+1} v_i \right)$$

La solution $a$ est donnée par

$$a = u + \lambda v$$

Résolution d'un système linéaire "presque" triangulaire.

Pour l'équation de Poisson la matrice triangulaire est mal conditionnée.

Fig. 5

Il existe alors une méthode astucieuse permettant de résoudre chacun des systèmes pair et impair (combinaison linéaire de solutions de systèmes plus simples (Fig. 5)). Mais dans notre cas cette méthode est très sensible aux erreurs d'arrondis et utilise des coefficients ayant des ordres de grandeurs très différents. C'est pourquoi il faut effectuer des combinaisons linéaires du système (1) pour obtenir des matrices bien conditionnées.

## 2. Système équivalent bien conditionné.

On utilise la relation (8) du chapitre III valable pour tout $u = \sum_{n=0}^{\infty} a_n T_n$ :

$$a_n = \frac{c_{n-2}}{4n(n-1)} a_{n-2}^{(2)} - \frac{1}{2(n^2-1)} a_n^{(2)} + \frac{1}{4n(n+1)} a_{n+2}^{(2)} \qquad \forall n \geq 2$$

Appliquée à $u_N = \sum_{n=0}^{N} a_n T_n$ cette relation peut s'écrire à l'aide de la suite $(e_n)_{n \in \mathbb{Z}}$ définie par : $e_n = 1$ si $n \leq N$ , $e_n = 0$ si $n > N$

Pour $u_N$ :

$$a_n = \frac{c_{n-2}}{4n(n-1)} a_{n-2}^{(2)} - \frac{e_{n+2}}{2n(n^2-1)} a_n^{(2)} + \frac{e_{n+4}}{4n(n+1)} a_{n+2}^{(2)} \qquad \forall n = 2, N$$

Avec les notations :

$$\tilde{p}_n = \frac{c_{n-2}}{4n(n-1)} \; ; \; \tilde{q}_n = \frac{e_{n+2}}{2n(n^2-1)} \; ; \; \tilde{r}_n = \frac{e_{n+4}}{4n(n+1)} \qquad n = 2, N$$

cette relation s'écrit (2) : $a_n = \tilde{p}_n a_{n-2}^{(2)} - \tilde{q}_n a_n^{(2)} + \tilde{r}_n a_{n+2}^{(2)} \qquad n = 2, N$

Revenons aux N - 1 équations $(E_n) : a_n^{(2)} - \lambda a_n = f_n \qquad n = 0, N-2$

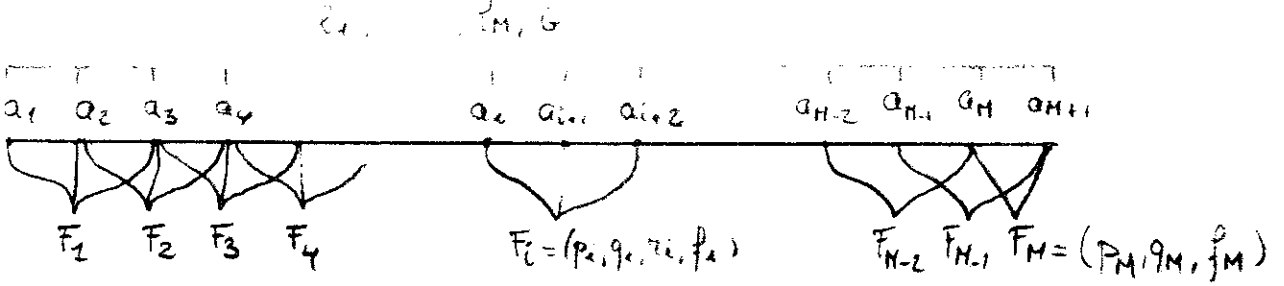
On peut effectuer des combinaisons linéaires pour obtenir un système équivalent

$$\tilde{p}_n (E_{n-2}) - \tilde{q}_n (E_n) + \tilde{r}_n (E_{n+2}) \qquad n = 2, N$$

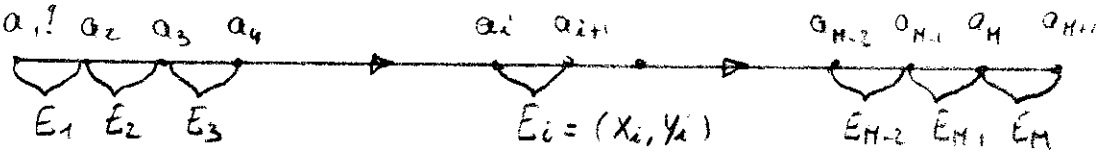
Le système global s'écrit alors

$$\begin{cases} \lambda \tilde{p}_n a_{n-2} - (1 + \lambda \tilde{q}_n) a_n + \lambda \tilde{r}_n a_{n+2} = \tilde{p}_n f_{n-2} - \tilde{q}_n f_n + \tilde{r}_n f_{n+2} \qquad n = 2, N \\ \sum_{n \text{ pair}} a_n = G_1 \\ \sum_{n \text{ impair}} a_n = G_2 \end{cases}$$

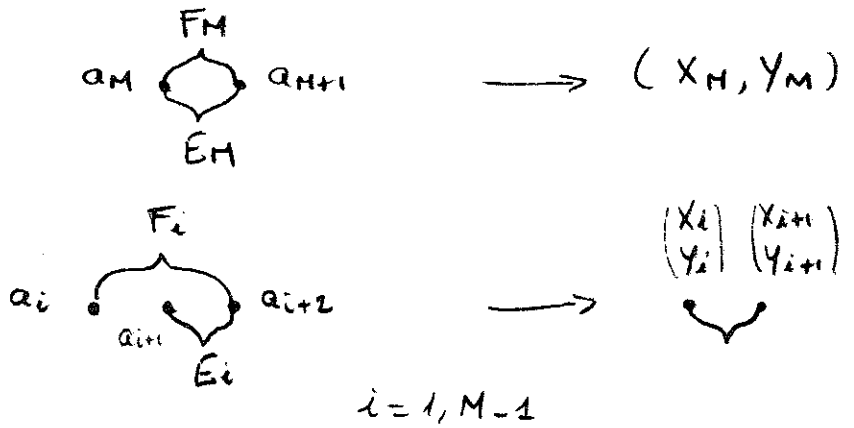
(1) Système linéaire




(2) Récurrence sur les $(a_i)$ $i = 1, M+1$ (fictive)




(3) Récurrence sur les $(X_i, Y_i)$ $i = 1, M$



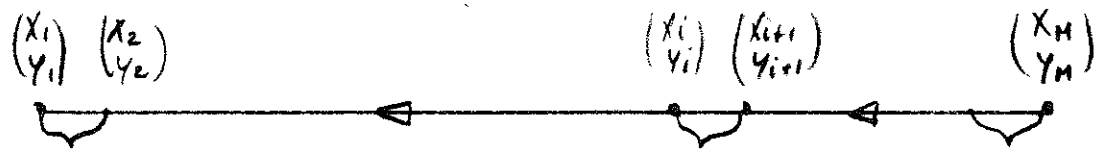




Fig. 6

Ce système est couplé lui aussi en deux systèmes indices n pairs et impaires avec une structure semblable. Chaque système est résolu par un algorithme de double balayage qui a l'avantage d'être numériquement stable par dominance diagonale.

## 3. Le double balayage avec une "ligne pleine".

Fig. 6(1) : cet algorithme résoud le système A = F avec

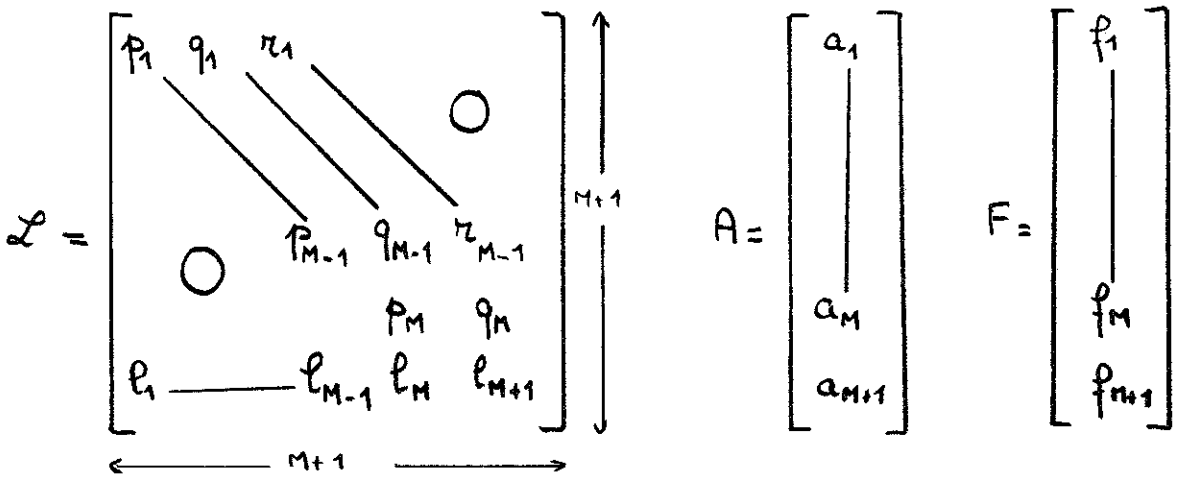



Fig. 6(2) : on construit les suites $(X_i)i = 1,M$ et $(Y_i)i = 1,M$ par

$$(E_i) : a_{i+1} = X_i a_i + Y_i \qquad i = 1, M$$

Fig. 6(3) : calcul de $X_M$ et $Y_M$

$$\begin{cases} p_M a_M + q_M a_{M+1} = f_M \\ a_{M+1} = X_M a_N + Y_M \end{cases} \Rightarrow X_M = \frac{-p_M}{q_M} \; ; \; Y_M = \frac{f_M}{q_M}$$

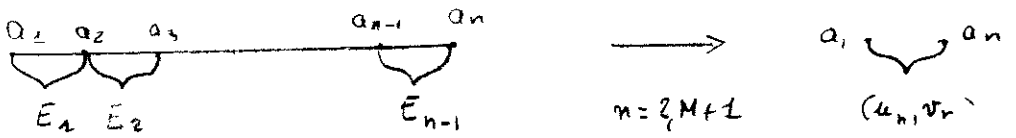
Fig.6(3) : récurrence sur les $X_i$ et $Y_i$

$$\begin{cases} p_i a_i + q_i a_{i+1} + r_i a_{i+2} = f_i \\ a_{i+2} = X_{i+1} a_{i+1} + Y_{i+1} \end{cases}$$

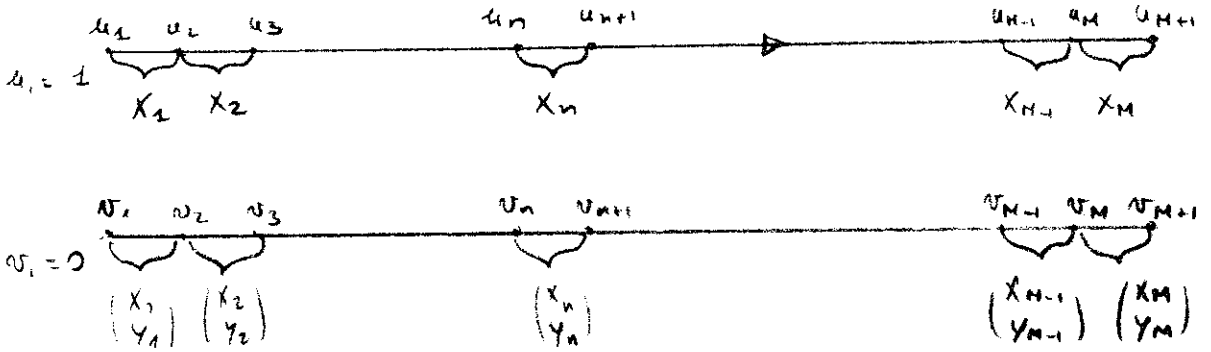
$$\Rightarrow p_i a_i + (r_i X_{i+1} + q_i) a_{i+1} = -r_i Y_{i+1} + f_i$$

Fig 6 (suite)

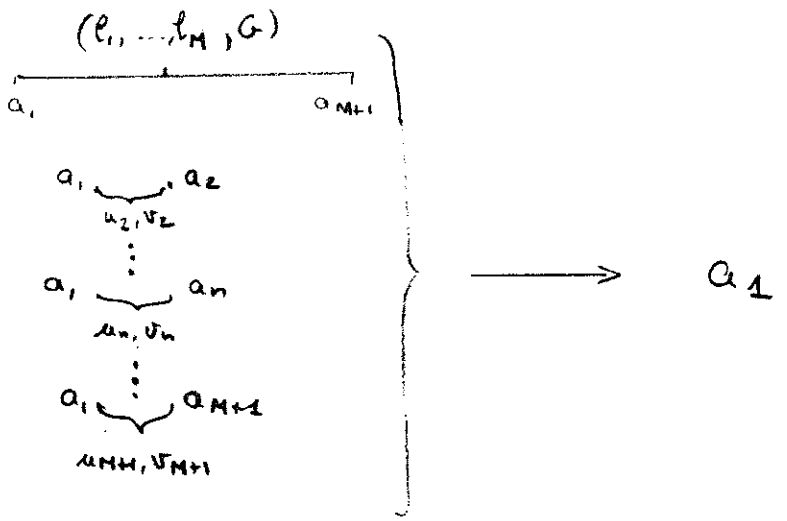
(4) Expression des $(a_n)_{n=1,M+1}$ en fonction de $a_1$



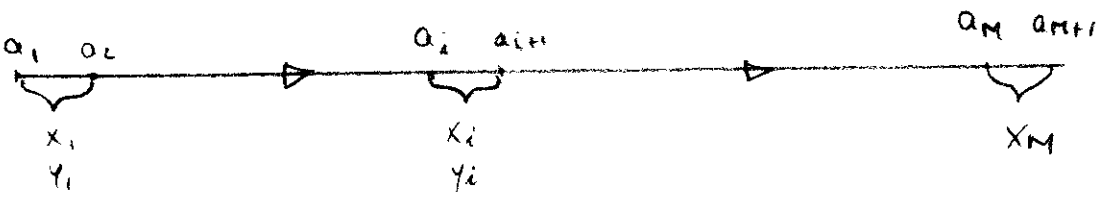

(5) Récurrence sur les $(u_n, v_n)_{n=1,M+1}$




(6) Calcul de $a_1$




(7) Récurrence sur les $(a_i)_{i=1,M+1}$ (effective)



En identifiant cette égalité avec $(E_i)$ on obtient :

$$\begin{cases} X_i = \dfrac{-p_i}{r_i X_{i+1} + q_i} \\ Y_i = \dfrac{-r_i Y_{i+1} + f_i}{r_i X_{i+1} + q_i} \end{cases} \qquad i = M-1, 1$$

Ces relations permettent de calculer par récurrence $(X_i)$ et $(Y_i)$ $i = 1,M$

Détermination de $a_1$

Pour initialiser la récurrence $(E_i)i = 1,M$ il faut calculer $a_1$ en éliminant tous les autres $a_i$ du système suivant

$$\begin{cases} p_1 a_1 + p_2 a_2 + \dots + p_{M+1} a_{M+1} = G \\ (E_i) \; : \; a_{i+1} = X_i a_i + Y_i \end{cases}$$

Fig. 6(4) - [i] Exprimons $a_n$ en fonction de $a_1$ pour les $n = 2,M+1$. Ceci s'effectue à partir des $n-1$ équations $(E_i)i = 1,n-1$ sous la forme :

$$a_n = u_n a_1 + v_n \qquad n = 1, M+1$$

Fig. 6(5) $u_n$ et $v_n$ ne dépendent que des $(X_i)$ et $(Y_i)$ et sont définis par la relation de récurrence suivante :

$$\begin{cases} u_1 = 1 \qquad \text{et} \qquad v_1 = 0 \\ u_{n+1} = X_n u_n \quad \text{et} \quad v_{n+1} = X_n v_n + Y_n \qquad n = 1, M-1 \end{cases}$$

En effet :

$$a_{n+1} = X_n a_n + Y_n = X_n (u_n a_1 + v_n) + Y_n = (X_n u_n) a_1 + (X_n v_n + Y_n)$$

Fig. 6(6) - [ii] Calculons $a_1$ :

$$G = \sum_{n=1}^{M+1} p_n a_n = \sum_{n=1}^{M+1} p_n (u_n a_1 + v_n) = U a_1 + V$$

avec

$$U = \sum_{n=1}^{M+1} p_n u_n \quad \text{et} \quad V = \sum_{n=1}^{M+1} p_n v_n$$

d'où

$$a_1 = \frac{G - V}{U}$$

Fig. 6(7) : la relation de récurrence $(E_i)i = 1,M$ permet alors de calculer tous les $a_i$ $i = 1, M+1$.

## 4. conditions aux limites plus générales.

L'algorithme précédent permet de résoudre les systèmes d'indices pairs et impaires du problème de Dirichlet pour l'équation de Poisson en posant $\ell_i = 1 \quad i = 1,M$.

Pour des conditions aux limites de Neumann : $u'(-1) = g_1$ et $u'(1) = g_2$ les deux dernières équations de la méthode Tau s'écrivent :

$$\sum_{n=0}^{N} (-1)^{n-1} n^2 a_n = g_1 \quad ; \quad \sum_{n=0}^{N} n^2 a_n = g_2$$

Elles se découplent en indices pairs et impairs par somme et différence :

$$\begin{cases} \sum_{n \text{ pair}} n^2 a_n = \frac{1}{2}(g_1 + g_2) \\ \sum_{n \text{ impair}} n^2 a_n = \frac{1}{2}(g_1 - g_2) \end{cases}$$

Il suffit alors d'utiliser l'algorithme de double balayage en posant $(\ell_i)$ $i = 1,M$ égal aux éléments d'indices pairs ou impairs du tableau $(1, 2^2, 3^2, \ldots, n^2, \ldots, (N+1)^2)$ suivant le système que l'on résoud.

## 5. Sous-programmes FORTRAN.

. "PSPT" est un exemple de sous programme utilisant la méthode précédente pour résoudre l'équation de Poisson. Avant de l'utiliser il faut appeler trois sous-programmes préparatoires :

- "PQR" : qui calcule les coefficients $\tilde{p}_n, \tilde{q}_n, \tilde{r}_n$
- "IP" : qui calcule les lignes $\ell_i$ pour Dirichlet ou Neumann
- "CONV" : qui effectue $\tilde{p}_n f_{n-2} - \tilde{q}_n f_n + \tilde{r}_n f_{n+2}$

. "PSVB 3" est un sous-programme qui résoud une équation de Poisson vectorielle

$$\begin{cases} -\begin{pmatrix} v'' \\ b'' \end{pmatrix} + \Lambda \begin{pmatrix} v \\ b \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix} \quad \Lambda \text{ matrice } 2\times 2 \\ v(\pm 1) = 0 \qquad b'(\pm 1) = 0 \end{cases}$$

L'algorithme de double balayage y est employé en remplaçants les coefficients scalaires par des matrice $2\times 2$.

# CHAPITRE V : ALGORITHMES DE "TRANSFORMATION RAPIDE"

## 1. Introduction.

L'intérêt des méthodes de collocation (ou Tau collocation) réside dans l'utilisation d'algorithmes rapides permettant de multiplier une certaine matrice et un vecteur colonne quelconque. Il s'agit du produit de $\mathcal{M} = (\varphi_n(x_j))$ $n = 1,N$ $j = 1,N$ avec le vecteur colonne $(a_n)$, $n = 1,N$ ou bien de $\mathcal{W} = \mathcal{M}^{-1}$ avec le vecteur colonne $(u_j)$ $j = 1,N$. Ces multiplications permettent d'expliciter l'isomorphisme entre l'espace spectral et l'espace physique.

La Transformation de Fourier Rapide (FFT) est l'outil de base pour effectuer ces produits (paragraphe 5). Mais avant de l'utiliser il faut savoir exprimer analytiquement les matrices $\mathcal{M}$ et leurs inverses (paragraphes 2, 3, 4). A partir de cette expression analytique on effectue quelques transformations pour pouvoir utiliser la FFT (paragraphes 5, 6, 7).

## 2. Collocation de Fourier impaire N = 2K +1 (Fig. 1).

Bien qu'elle ait été mentionnées dans les exemples du chapitre I p.121 pour des raisons de commodité de présentation (sommation sur des carrés fermés par exemple) la collocation suivante n'est pas utilisée en pratique. La raison est que les FFT usuelles ne fonctionnent que pour des nombres de points N pairs.

Cette collocation utilise

$$(e^{\pm ikx})_{-K \le k \le K} \quad \text{et} \quad x_j = j\frac{2\pi}{N} \quad j = 0,N-1 \quad \text{avec} \quad N = 2K+1$$

### a) Passage espace spectral- espace physique

$$u_j = \sum_{k=-K}^{K} a_k e^{\pm ikx} = \sum_{k=-K}^{K} a_k e^{\pm ikj\frac{2\pi}{N}} \qquad j = 0,N-1$$

ce que l'on exprime par $u = \mathcal{M} a$ avec u vecteur colonne $(u_j)$ $j = 0,N-1$

FOURIER IMPAIRE $N = 2K+1$

$$u = \mathcal{M} a$$

$$a = \mathcal{W} u$$

$$\mathcal{M} = \left( e^{\frac{i}{-} j k \frac{2\pi}{N}} \right) \quad \begin{matrix} j = 0, N-1 \\ -K \leq k \leq K \end{matrix}$$

$$\mathcal{W} = \frac{1}{N} \left( e^{\frac{-i}{-} j l \frac{2\pi}{N}} \right) \quad \begin{matrix} j = 0, N-1 \\ -K \leq l \leq K \end{matrix}$$

Fig 1

FOURIER PAIRE $N = 2K$

$$u = \mathcal{M} a$$

$$a = \mathcal{W} u$$

$$\mathcal{M} = \left( e^{\frac{i}{-} j k \frac{2\pi}{N}} \right) \quad \begin{matrix} j = 0, N-1 \\ -K < k \leq K \end{matrix}$$

$$\mathcal{W} = \frac{1}{N} \left( e^{\frac{-i}{-} j l \frac{2\pi}{N}} \right) \quad \begin{matrix} j = 0, N-1 \\ -K < l \leq K \end{matrix}$$

Fig 2

Collocations de Fourier

et a vecteur colonne $(a_k)$ - $K \leqslant k \leqslant K$.

Donc $\mathcal{M}_{jk} = e^{\pm i j k \frac{2\pi}{N}}$ $\quad j = 0, N-1$ et $-K \leqslant k \leqslant K$

b) Passage espace physique-espace spectral.

Il suffit d'inverser $\mathcal{M}$. Pour cela on utilise la relation facile à vérifier

$$\sum_{j=0}^{N-1} e^{\pm i j k \frac{2\pi}{N}} \; e^{-\pm i j \ell \frac{2\pi}{N}} = N \, \delta_{k\ell} \quad \text{si } (k,\ell) \in [-K, K]^2$$

Posons

$$\mathcal{W}_{j\ell} = \frac{1}{N} e^{-\pm i j \ell \frac{2\pi}{N}}$$

On a donc

$$\sum_{j=0}^{N-1} \mathcal{M}_{jk} \, \mathcal{W}_{j\ell} = \delta_{k\ell}$$

Comme $\mathcal{M}$ est symétrique on en déduit $\mathcal{M}^{-1} = \mathcal{W}$

D'où $a = \mathcal{W} u$ : $a_k = \frac{1}{N} \sum_{j=0}^{N-1} u_j e^{-\pm i k x_j}$

3. Collocation de Fourier paire N = 2K (Fig. 2).

C'est cette collocation qui est utilisée dans les applications pratiques. Elle utilise

$(e^{\pm i k x})_{k = -K+1, K}$ et $x_j = j \frac{2\pi}{N}$ $\quad j = 0, N-1$ avec $N = 2K$

Par rapport à la collocation impaire on a supprimé la fonction $e^{\pm i (-K) x}$.

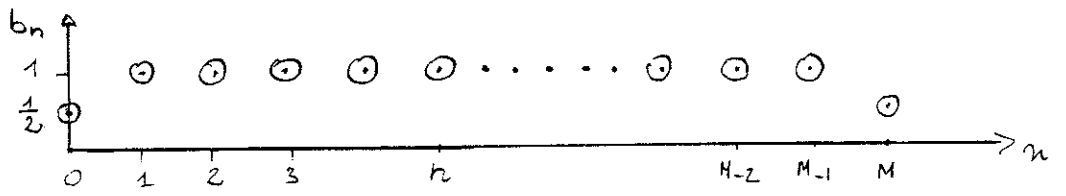
On obtient le même type de formule pour appliquer $\mathcal{M}$ et son inverse :

$$u_j = \sum_{k=-K+1}^{K} a_k e^{\pm i k x_j} = \sum_{k=-K+1}^{K} a_k e^{\pm i k j \frac{2\pi}{N}}$$

$$a_k = \frac{1}{N} \sum_{j=0}^{N-1} u_j e^{-\pm i k x_j} = \frac{1}{N} \sum_{j=0}^{N-1} u_j e^{-\pm i k j \frac{2\pi}{N}}$$

donc

$$\mathcal{M}_{jk} = e^{\pm i j k \frac{2\pi}{N}} \qquad \mathcal{W}_{j\ell} = \frac{1}{N} e^{-\pm i j \ell \frac{2\pi}{N}}$$




Suite bn

Fig.3

$$u = M\,a$$

$$a = W\,u$$

$$M = \left(\cos n j \frac{\pi}{M}\right) \quad \begin{matrix} j = 0, M \\ n = 0, M \end{matrix}$$

$$W = \left(\frac{2}{M} b_n b_j \cos n j \frac{\pi}{M}\right) \quad \begin{matrix} j = 0, M \\ n = 0, M \end{matrix}$$

$$D = \begin{bmatrix} \frac{1}{2} & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & \frac{1}{2} \end{bmatrix}$$

$$W = \frac{2}{M} D\,M\,D$$

COLLOCATION DE TCHEBYSHÉV

Fig.4

## 4. Collocation de Tchebyshev (Fig. 4).

On utilise les fonctions de base $(T_n(x))_{n=0,M}$ $x \in [-1,1]$ et les points de collocation $x_j = \cos\frac{\pi j}{M}$ $j=0,M$ . On définit $\theta_j = \frac{\pi j}{1}$ si bien que $x_j = \cos\theta_j$ .

### a) Passage espace spectral-espace physique.

$$u_j = \sum_{n=0}^{M} a_n T_n(x_j) = \sum_{n=0}^{M} a_n \cos n\theta_j$$

que l'on résume par $u = \mathcal{M} a$ $a = (a_m)_{m=0,M}$ ; $u = (u_m)_{m=0,M}$ et $\mathcal{M}_{jn} = \cos(nj\frac{\pi}{N})$ $j = 0,M$ $n = 0,M$.

### b) Passage espace physique-espace spectral.

$a = \mathcal{W} u$ avec $\mathcal{W} = \mathcal{M}^{-1}$

Pour trouver l'expression analytique de $\mathcal{W}$ on utilise la relation démontrée dans l'appendice (V.B.a p 189)

$$\sum_{j=0}^{M} b_j \cos nj\frac{\pi}{M} \cos mj\frac{\pi}{M} = \frac{M}{2b_m}\delta_{nm} \quad \forall (m,n) \in [0,M]^2$$

La suite $(b_n)_{n=0,M}$ étant définie par : $b_0 = \frac{1}{2}$, $b_M = \frac{1}{2}$, $b_n = 1$ sinon (Fig. 3).

Soit alors $\mathcal{W}_{jm} = \frac{2}{M} b_m b_j \cos(mj\frac{\pi}{M})$ $j=0,M$ $m=0,M$

La relation s'exprime $\sum_{j=0}^{M} \mathcal{M}_{jm}\mathcal{W}_{jm} = \delta_{nm}$ et comme est symétrique on a donc : $\mathcal{W} = \mathcal{M}^{-1}$

Donc
$$a_m = \frac{2}{M} b_n \sum_{j=0}^{M} b_j u_j \cos nj\frac{\pi}{M}$$

On peut définir $\mathcal{B}$ = diag(b) la matrice diagonale formée avec les éléments $(b_n)_{n=0,M}$ .

On a $\mathcal{W} = \frac{2}{M}\mathcal{B}\mathcal{M}\mathcal{B}$

$$Z = \mathcal{F}\, z$$

$$Z = \begin{pmatrix} Z_0 \\ | \\ Z_{N-1} \end{pmatrix} \qquad z = \begin{pmatrix} z_0 \\ | \\ z_{N-1} \end{pmatrix}$$

$$\mathcal{F} = \left( e^{\mp i n j \frac{2\pi}{N}} \right) \begin{matrix} j = 0, N-1 \\ n = 0, N-1 \end{matrix}$$

FFT C. C

Fig 5

$$u = m a$$

$$\begin{pmatrix} u_0 \\ | \\ u_{N-1} \end{pmatrix} \uparrow N \downarrow \qquad \begin{pmatrix} a_{-k+1} \\ | \\ a_k \end{pmatrix} \uparrow N \downarrow \qquad m = \left( e^{\mp i k j \frac{2\pi}{N}} \right) \begin{matrix} j = 0, N-1 \\ k = -k+1, k \end{matrix}$$

$$\begin{pmatrix} a_0 \\ a_1 \\ | \\ a_K \\ a_{-K+1} \\ | \\ a_{-2} \\ a_{-1} \end{pmatrix} \uparrow N \downarrow$$

$$\underline{u = \mathcal{F}_N\, a'}$$

FFT appliquée à Fourier Paire

Fig 6

## 5. Transformation de Fourier Rapide complexe (FFT C-C) (Fig. 5).

La FFT est un algorithme qui permet d'effectuer les N opérations suivantes

$$Z_j = \sum_{n=0}^{N-1} z_n \, e^{\pm i n j \frac{2\pi}{N}} \qquad j = 0, N-1$$

en un nombre d'opérations de l'ordre de $N.\log_2(N)$ au lieu de $N^2$.
Soit $z = (z_n)n = 0, N-1$ et $Z = (Z_j)j = 0, N-1$ deux vecteurs colonne complexes, soit $\mathcal{F}$ la matrice NxN $\mathcal{F}_{jn} = e^{\pm i n j \frac{2\pi}{N}}$. L'algorithme effectue

$$Z = \mathcal{F} z$$

La FFT inverse est un algorithme qui effectue $z = \mathcal{F}^{-1} Z$
c'est à dire $z_n = \frac{1}{N} \sum_{j=0}^{N-1} Z_n \, e^{-\pm i n j \frac{2\pi}{N}} \quad n = 0, N-1$
avec le même nombre d'opérations $N \log_2 N$.

## 6. Utilisation de la FFT pour "Fourier paire" (Fig. 6).

On veut calculer $u = \mathcal{M} a$ avec $u = \{u_0, u_1, \ldots, u_{N-1}\}$
$a = \{a_{-K+1}, \ldots a_{-1}, a_0, a_1 \ldots a_K\}$ et $\mathcal{M}_{jk} = e^{\pm i j k \frac{2\pi}{N}}$ $j = 0, N-1$ et $k = -k+1, K$

Mais l'algorithme de FFT applique la matrice $\mathcal{F} = (\mathcal{F}_{nj})$

avec les indices $j = 0, N-1$ et $n = 0, N-1$

Il faut alors construire le vecteur $a' = \{a_0, a_1, \ldots, a_K, a_{-K+1}, \ldots, a_{-2}, a_{-1}\}$ et remarquer que $\mathcal{M} a = \mathcal{F} a'$ car pour tout j les $(e^{\pm i j n \frac{2\pi}{N}})_{n \in \mathbb{Z}}$ sont des suites cycliques de période N.

Moyennant donc ces réarrangements de tableaux on peut utiliser la FFT ou la FFT inverse pour effectuer ces allers et retours :

$$\begin{cases} u = \mathcal{F} a' \\ a' = \mathcal{F}^{-1} u \end{cases}$$

$$u = \mathcal{M}\, a \qquad \text{et} \qquad a = \mathcal{W}\, u$$

$$u = \begin{pmatrix} u_0 \\ | \\ u_M \end{pmatrix} \qquad a = \begin{pmatrix} a_0 \\ | \\ a_M \end{pmatrix}$$

$$\mathcal{M} = \left( \cos m j \frac{\pi}{N} \right)_{\substack{j=0,M \\ m=0,M}} \qquad \mathcal{W} = \left( \frac{2}{M} b_m b_j \cos n j \frac{\pi}{M} \right)_{\substack{j=0,M \\ n=0,M}}$$

se ramènent à

$$\Lambda = \mathcal{M}\, \lambda$$

$$\Lambda = \begin{pmatrix} \Lambda_0 \\ | \\ \Lambda_M \end{pmatrix} \qquad \lambda = \begin{pmatrix} \lambda_0 \\ | \\ \lambda_M \end{pmatrix}$$

TCHEBYSHEV

Fig. 7

Prolongement pair de $\lambda$

$$\Lambda = \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ | \\ \lambda_{M-1} \\ \lambda_M \end{pmatrix} \quad M+1$$

$$\mathcal{Z} = \begin{pmatrix} \lambda_0 \\ \frac{1}{2}\lambda_1 \\ | \\ \frac{1}{2}\lambda_{M-1} \\ \lambda_M \\ \frac{1}{2}\lambda_{M-1} \\ | \\ \frac{1}{2}\lambda_2 \\ \lambda_1 \end{pmatrix} \quad 2M = N$$

Fig 8

## 7. Utilisation de la FFT pour la collocation de "Tchebyshev".

L'application de la FFT aux multiplications par $\mathcal{M}$ ou $\mathcal{W}$ est plus compliquée que précédemment. Cependant on utilise que la FFT directe.

Soient $a = (a_m)m = 0,M$ et $u = (u_j)j = 0,M$ des vecteurs colonnes complexes.

Que ce soit pour effectuer $u = \mathcal{M}a$ : $u_j = \sum_{m=0}^{M} a_m \cos mj\frac{\pi}{M} \quad j=0,M$

Ou bien pour effectuer $a = \mathcal{W}u$ : $a_n = \frac{2}{M} b_n \sum_{j=0}^{M} b_j u_j \cos nj\frac{\pi}{M} \quad n=0,M$

On est toujours ramené à effectuer un produit de la forme :

$$\Lambda = \mathcal{M}\lambda \quad : \quad \Lambda_j = \sum_{m=0}^{M} \lambda_m \cos mj\frac{\pi}{M} \quad j=0,M \qquad \text{(Fig. 7)}$$

### a) Prolongement pair du vecteur $\lambda$ (Fig. 8).

A partir de $\lambda = (\lambda_m)m = 0,M$ on définit le vecteur colonne $z = (z_n)n = 0,N$ avec $N = 2M$, par le prolongement "pair" :

$$\begin{cases} z_0 = \lambda_0 \\ z_m = \frac{1}{2}\lambda_m \quad m = 1,M-1 \quad ; \quad z_{M-m} = \frac{1}{2}\lambda_m \quad m = 1,M-1 \\ z_M = \lambda_M \end{cases}$$

Donc $\forall\, n = 0,N-1 \quad z_{N-n} = z_n$

On note $\mathcal{F}_N$ la matrice de la FFT : $\left(e^{i nj \frac{2\pi}{N}}\right)$

On remarque alors qu'en effectuant $Z = \mathcal{F}_N.z$ les M premières composantes de Z (qui est de longueur N) sont exactement les composantes du vecteur recherché (qui est de longueur M).

Ceci se voit par le calcul suivant $\forall j = 0,M$

$$Z_j = \sum_{m=0}^{N-1} z_n e^{\pm i n j \frac{2\pi}{N}} = z_0 + \sum_{m=1}^{M-1} z_m \left( e^{\pm i m j \frac{\pi}{M}} + e^{-\pm i m j \frac{\pi}{M}} \right) + (-1)^j z_M$$

$$= z_0 + \sum_{m=1}^{M-1} 2 z_m \cos m j \frac{\pi}{M} + (-1)^j z_M$$

$$= \sum_{m=0}^{M} \lambda_n \cos m j \frac{\pi}{M} = \Lambda_j$$

Il existe un moyen d'utiliser la propriété de symétrie du vecteur $z$ pour effectuer le produit $\mathcal{F}_N z$. Ce procédé décrit ci-dessous permet de diminuer par deux la taille de cette transformation de Fourier.

## b) Transformation de Fourier optimale d'un vecteur symétrique.

Soit $z = (z_n)$ n = 0,N-1 un vecteur colonne complexe tel que :

$$\forall n = 0,N-1 \quad z_n = z_{N-n}$$

On veut calculer $Z = \mathcal{F}_N z$ ou $\mathcal{F}_N$ est la matrice NxN de la FFT.

On remarque que Z vérifie aussi $\forall j = 0,N-1 \quad Z_{N-j} = Z_j$

en effet

$$Z_{N-j} = \sum_{n=0}^{N-1} z_n e^{-\pm j n \frac{2\pi}{N}} = \sum_{p=0}^{N-1} z_{N-p} e^{\pm j p \frac{2\pi}{N}}$$

$$= \sum_{p=0}^{N-1} z_p e^{\pm j p \frac{2\pi}{N}} = Z_j$$

On a donc besoin de calculer uniquement les M+1 coefficients Zj j = 0,M avec $M = \frac{N}{2}$.

L'astuce utilisée consiste à définir le vecteur colonne complexe $y = (y_m)$ m = 0,M-1 par $y_m = z_{2m} + i(z_{2m+1} - z_{2m-1})$ m = 0,M - 1

en ayant posé $z_N = z_0$ et $z_{-1} = z_{N-1}$.

Soit $\mathcal{F}_M$ la matrice M x M de la FFT à M points.

On effectue alors le produit $Y = \mathcal{F}_M y$

Il est démontré dans l'appendice (V.8.b p. 191) que la relation qui permet d'exprimer les $Z_j$ $j = 1, M-1$ recherchés en fonction de $Y = (Y_j)$ $j = 0, M-1$ calculé par FFT est la suivante :

$$\forall j = 1, M-1 \quad Z_j = \frac{1}{2}(Y_j + Y_{M-j}) + \frac{1}{4\sin\left(\frac{\pi}{M}j\right)}(Y_j - Y_{M-j})$$

Pour $j = 0$ et $j = M$ il Faut utiliser le vecteur $(z_n)_{n=0,N-1}$ [ou $(\lambda_m)_{m=0,M}$]

$$Z_0 = \sum_{n=0}^{N-1} z_n \left(= \sum_{m=0}^{M} \lambda_m\right) \quad ; \quad Z_M = \sum_{n=0}^{N-1} (-1)^n z_n \left(= \sum_{m=0}^{M} (-1)^m \lambda_m\right)$$

## 8. Appendice.

### a) Inversion de la matrice $\mathcal{M} = \left(\cos n j \frac{\pi}{M}\right)$

$$\forall (n,m) \in [0,M]^2 \qquad \text{soit}$$

$$E = \sum_{j=0}^{M} b_j \cos n j \frac{\pi}{M} \cos m j \frac{\pi}{M} = \frac{1}{2} \sum_{j=0}^{M} b_j \cos(m+n) j \frac{\pi}{M} + \frac{1}{2} \sum_{j=0}^{M} b_j \cos(m-n) j \frac{\pi}{M}$$
$$\qquad\qquad (I) \qquad\qquad + \qquad\qquad (II)$$

. si $n = m = 0$ ou si $n = m = M$ $\qquad (I) = \frac{M}{2}$ : $(II) = \frac{M}{2}$ : $E = M$

. si $n = m \notin \{0, M\}$ $\qquad (I) = 0$ $\quad (II) = \frac{M}{2}$ : $E = \frac{M}{2}$

. si $n \neq m$ $\qquad (I) = 0$ $\quad (II) = 0$ : $E = 0$

démonstration :

Lorsque p est pair et n'appartient pas à $\{0, 2M\}$, la suite $b_j$ permet d'écrire :

$$\sum_{j=0}^{M} b_j \cos p j \frac{\pi}{M} = \sum_{j=0}^{M-1} \cos p j \frac{\pi}{M} = 0$$

Lorsque p est impair cette même somme est nulle car les termes j sont opposés aux termes N - j.

b) FFT pour un vecteur complexe symétrique. ($N = 2M$)

$$\forall j=0,M-1 \quad Y_j = \sum_{m=0}^{M-1} \left[ z_{2m} + i\,(z_{2m+1} - z_{2m-1}) \right] e^{\pm i m j \frac{2\pi}{M}}$$

$$= \sum_{m=0}^{M-1} z_{2m}\, e^{\pm i 2mj\frac{2\pi}{N}} + i \left( \sum_{m=0}^{M-1} z_{2m+1}\, e^{\pm i (2m+1) j \frac{2\pi}{N}} \right) e^{\mp i j \frac{2\pi}{N}}$$

$$- i \left( \sum_{m=0}^{M-1} z_{2m-1}\, e^{\pm i (2m-1) j \frac{2\pi}{N}} \right) e^{\pm i j \frac{2\pi}{N}}$$

Grâce à la cyclisation $z_{-1} = z_{M-1}$ et $z_M = z_0$ et à la périodicité des exponentielles les sommes entre parenthèses sont égales. Donc

$$Y_j = \sum_{\substack{k=0 \\ k\ \text{pair}}}^{N-2} z_k\, e^{\pm i k j \frac{2\pi}{N}} - 2 \sin\left(j\frac{\pi}{M}\right) \sum_{\substack{k=1 \\ k\ \text{impair}}}^{N-1} z_k\, e^{\pm i k j \frac{2\pi}{N}}$$

$$Y_{M-j} = \sum_{\substack{p=2 \\ p\ \text{pair}}}^{N} z_{N-p}\, e^{\pm i (N-p)(M-j)\frac{2\pi}{N}} - 2 \sin\left(j\frac{\pi}{M}\right) \sum_{\substack{p=1 \\ p\ \text{impair}}} z_{N-p}\, e^{\pm i (N-p)(M-j)\frac{2\pi}{N}}$$

$$= \sum_{\substack{k=0 \\ k\ \text{pair}}}^{N-2} z_k\, e^{\mp i k j \frac{2\pi}{N}} + 2 \sin\left(j\frac{\pi}{M}\right) \sum_{\substack{k=1 \\ k\ \text{impair}}}^{N-1} z_k\, e^{\pm i k j \frac{2\pi}{N}}$$

De ces expressions on déduit le résultat recherché. $\forall j = 1, M-1$

$$\sum_{k\ \text{pair}} = \frac{1}{2}(Y_j + Y_{M-j}) \quad ; \quad \sum_{k\ \text{impair}} = \frac{1}{4\sin\left(j\frac{\pi}{M}\right)} (Y_j - Y_{M-j})$$

d'où

$$\forall j = 1, M-1 : \quad Z_j = \sum_{k\ \text{pair}} + \sum_{k\ \text{impair}} = \frac{1}{2}(Y_j + Y_{M-j}) + \frac{1}{4\sin\left(j\frac{\pi}{M}\right)} (Y_j - Y_{M-j})$$

Dans la pratique on connait $(Y_j)j = 0,M-1$ et l'on veut déduire $(Z_j)j = 0,M$.

Pour $j = 1, M-1$

$$Z_j = \frac{1}{2}(Y_j + Y_{M-j}) + \frac{1}{4\sin j\frac{\pi}{M}} (Y_j - Y_{M-j})$$

Pour $j = 0$ et $j = M$ il n'y a pas d'expression en fonction de $Y$

$$Z_0 = \sum_{n=0}^{N-1} z_n \quad ; \quad Z_M = \sum_{n=0}^{N-1} (-1)^n z_n$$

## c) FFT Réelle-complexe symétrique.

Dans beaucoup de problèmes les tableaux à "transfourier" sont réels. On peut utiliser une FFT complexe-complexe en rajoutant des zéros, mais il y a un moyen d'économiser un facteur 2. La FFT RCS (Réel-Complexe Symétrique) permet de calculer les transformations entre le vecteur colonne réel $(u_i)i = 0,N-1$ et le vecteur colonne complexe $(a_k)k = 0,M$ avec $N = 2M$ pour la relation :

$$u_j = \sum_{k=0}^{M} a_k \, e^{i k j \frac{2\pi}{N}} + \sum_{k=1}^{M-1} a_k^* \, e^{-i k j \frac{2\pi}{N}}$$

Si l'on pose $a' = (a_0, a_1, \ldots, a_{M-1}, a_M, a_{M-1}^*, \ldots, a_2^*, a_1^*)$
on a $u = \mathcal{F}_N \, a'$

La FFT RCS exploite l'hermiticité en ne calculant que les $(a_k)k = 0,M$.

# CHAPITRE VI : DISCRETISATION TEMPORELLE

## 1. Introduction.

Quelle que soit la méthode employée pour prendre en compte la dépendance spatiale d'un problème d'évolution, on a le choix pour traiter la dépendance temporelle par une méthode spectrale ou aux différences finies.

Mais les méthodes spectrales en temps sont vite limitées par l'encombrement des tableaux, c'est pourquoi on utilise le plus souvent un schéma aux différences finies en temps. Dans ce chapitre sont exposés brièvement quelques schémas appliqués à des exemples d'équations aux dérivées partielles

## 2. Propriétés des schémas.

Un schéma discret en temps pour une équation aux dérivées partielles

$$\frac{\partial u}{\partial t} = F(u) \qquad (1)$$

s'écrit :

$$u^{n+1} = \mathcal{C}_{\Delta t}\,(u^n, u^{n-1}, \dots)$$

$\Delta t$ est le pas de temps et $t = n\,\Delta t$.

$u(x,t)$ est une fonction du temps à valeurs dans un Hilbert muni de la norme $\|\ \|$.

Les trois définitions suivantes jouent un rôle important :

- Consistance
- Stabilité
- Convergence

Un schéma est consistant si toute solution exacte $u(x,t)$ du problème (1) vérifie

$$\lim_{\Delta t \to 0} \varepsilon(t) = 0$$

avec $\varepsilon(t) = \| u(t+\Delta t) - \mathcal{C}_{\Delta t}\,(u(t), u(t-\Delta t), \dots) \|$

Il est alors précis à l'ordre p si $\varepsilon(t) = 0(\Delta t^p)$

Un schéma est stable si pour tout T il existe M tel que

$$\| (\mathcal{C}_{\Delta t})^n \| \leq M \quad \forall n, \Delta t \text{ vérifiant } t = n.\Delta t \leq T$$

Mais lors du traitement de la variable d'espace il faut souvent rajouter une condition sur $\Delta t$ et N (resp $\Delta x$) dans le cas d'une méthode spectrale (resp aux différences finies). [condition CFL].

Un schéma est convergent si pour toutes conditions initiales

$$\lim_{\Delta t \to 0} \| u(t) - u^n \| = 0 \quad \forall n \text{ tel que } t = n\Delta t \leq T$$

Théorème de Lax

Pour un problème bien posé et pour un shcéma consistant, la stabilité est équivalente à la convergence.

## 3. Exemples de schémas.

### a) Euler avant.

Equation de la chaleur périodique $\dfrac{\partial u}{\partial t} - \Delta u = f(x,t)$

En Fourier : $\dfrac{\hat{u}^{n+1} - \hat{u}^n}{\Delta t} = -k^2 \hat{u}^n + \hat{f}^n(k)$

. Ce schéma est précis au premier ordre

. stable pour $\Delta t \cdot k^2 < 2$

### b) Euler arrière.

Même exemple : $\dfrac{\hat{u}^{n+1} - \hat{u}^n}{\Delta t} = -k^2 \hat{u}^{n+1} + f^n(k)$

. Précis au premier ordre

. Inconditionnellement stable.

c) Crank-Nicolson.

Même exemple : $\frac{\hat{u}^{n+1} - \hat{u}^n}{\Delta t} = -\frac{k^2}{2}(\hat{u}^{n+1} + \hat{u}^n) + \hat{f}^n(k)$

. Précis au second ordre

. Inconditionnellement stable.

d) Variante de Crank-Nicholson.

Equation de Schrodinger : $i\frac{\partial u}{\partial t} + \Delta u + |u|^2 u = 0$

$$i\frac{u^{n+1} - u^n}{\Delta t} + \frac{1}{2}\Delta(u^{n+1} + u^n) + \frac{1}{4}\left(|u^{n+1}|^2 + |u^n|^2\right)(u^{n+1} + u^n) = 0$$

Ce schéma conserve l'énergie .

e) Leap-Frog.

Equation d'advection $\frac{\partial u}{\partial t} + a.\nabla u = 0 \qquad a \in \mathbb{R}$

en Fourier $\frac{\hat{u}^{n+1} - \hat{u}^{n-1}}{2\Delta t} = -aik\hat{u}^n$

. Précis au second ordre

. Inconditionnellement stable car aik est un terme antihermitien. Il est toujours instable pour les termes hermitiens :

$$\frac{\hat{u}^{n+1} - \hat{u}^{n-1}}{2\Delta t} = s\hat{u}^n \qquad \text{avec} \quad s \in \mathbb{R}$$

f) Adams-Bashforth.

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

c'est un schéma explicite : $\dfrac{u^{n+1} - u^n}{\Delta t} = \frac{3}{2} F(u^n) - \frac{1}{2} F(u^{n-1})$

. Précis au second ordre.

g) Prédicteur correcteur.

$$u^* = u^n + \tfrac{1}{2} \delta t \, F(u^n)$$

$$u^{n+1} = u^n + \delta t \, F(u^*)$$

. Précis au second ordre.

h) Runge-Kutta.

Soit :

$$u_1 = u^n + \Delta t \, F(u^n)$$

$$u_2 = u^n + \frac{\Delta t}{2} F(u_1)$$

$$u_3 = u^n + \frac{\Delta t}{2} F(u_2)$$

$$u_4 = u^n + \Delta t \, F(u_4)$$

Le schéma s'écrit

$$u^{n+1} = u^n + \frac{\Delta t}{6} \left[ F(u_1) + 2 F(u_2) + 2 F(u_3) + F(u_4) \right]$$

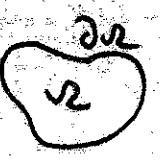
Il est précis au 4ème ordre.

# CHAPITRE VII : EQUATIONS DE NAVIER-STOKES.

## 1. Introduction.

L'intégration numérique des équations aux dérivés partielles décrivant des écoulements fluides comme l'équation de Boussinesq en convection fait l'objet de techniques particulières. Ces équations sont apparentées à un problème type qui est le système des équations de Navier-Stokes, obtenu pour des fluides incompressibles :

$$(1)\quad \begin{cases} \dfrac{\partial v}{\partial t} + v.\nabla v = -\nabla p + \nu \Delta v & \text{sur } \Omega \\ \operatorname{div} v = 0 & \\ v = 0 \quad \text{sur } \partial\Omega & \\ v(x,0) = v_0(x) \quad \text{avec} \quad \operatorname{div} v_0 = 0 & \end{cases}$$




Si mathématiquement la condition d'incompressibilité permet d'éliminer la pression on va voir que numériquement la discrétisation temporelle nécessite l'introduction de conditions aux limites supplémentaires. Deux méthodes pour résoudre le problème sont présentées : celle de Orszag et Kells ( $\frac{\partial p}{\partial n} = 0$ sur $\partial\Omega$) et celle de Kleiser et Schumann (div v = 0 sur $\partial\Omega$).

## 2. Schéma temporel.

Pour des raisons pratiques la discrétisation temporelle ne s'effectue pas sur le système (1) mais sur le système :

$$(2)\quad \begin{cases} \dfrac{\partial v}{\partial t} = v \wedge w - \nabla p + \nu \Delta v & \text{sur } \Omega \\ \Delta p = \operatorname{div}(v \wedge w) & \text{sur } \Omega \\ v = 0 \quad \text{sur } \partial\Omega & \\ v(x,0) = v_0(x) \quad \text{avec} \quad \operatorname{div} v_0 = 0 & \end{cases}$$

L'introduction de w = rot v permet la construction de schémas conservant l'énergie. Dans la pression on a rajouté le terme $-\frac{v^2}{2}$ .

Si (1) implique (2) de façon évidente la réciproque consiste à vérifier qu'une solution du problème (2) est à divergence nulle. Pour cela il faut résoudre l'équation :

$$\begin{cases} \dfrac{\partial(\operatorname{div} v)}{\partial t} = \nu \Delta(\operatorname{div} v) & \text{sur } \Omega \\ v = 0 \quad \text{sur } \partial\Omega \\ (\operatorname{div} v)_{t=0} = \operatorname{div} v_0 = 0 \end{cases}$$

Mais ce problème est mal posé et il faut rajouter à (2) une condition aux limites pour qu'il soit soluble et équivalent à (1).

On retrouve bien sur la nécessité d'introduire une condition aux limites supplémentaires lorsqu'on discrétise le temps; Donnons un exemple schéma très souvent utilisé pour ce genre de problème. Les termes non linéaires sont traités avec un schéma explicite : Adams-Bashforth, et les termes linéaires de façon implicite : schéma Crank-Nicolson.

Schéma Adams-Bashforth - Crank-Nicholson :

$$(3)\quad \begin{cases} -\frac{\nu}{2}\Delta v^{n+1} + \frac{1}{\delta t} v^{n+1} = \underbrace{\frac{1}{\delta t} v^n + \frac{3}{2}(v^n \wedge w^n) - \frac{1}{2}(v^{n-1} \wedge w^{n-1})}_{\text{- notation : } \frac{1}{\delta t} v^*} - \nabla p^n + \frac{\nu}{2}\Delta v^n \\ \Delta p^n = \frac{1}{\delta t} \operatorname{div} v^* \\ v^{n+1} = 0 \quad \text{sur } \partial\Omega \end{cases}$$

Il manque une condition aux limites qu'il faut choisir de façon à approximer le mieux possible la solution du problème (1).

## 3. Méthode Orszag-Kells.

Si on calcule $\frac{\partial p^n}{\partial n}$ ($\vec{n}$ étant la normale aux bord $\partial\Omega$) pour le schéma précédent on obtient :

$$\frac{\partial p^n}{\partial n} = \frac{\nu}{2} \Delta (v^{n+1} + v^n) \cdot \vec{n}$$

Pour $\nu$ petit la méthode Orzzag Kells consiste à choisir comme condition aux limites supplémentaire $\frac{\partial p}{\partial n} = 0$ sur $\partial\Omega$. L'avantage de ce choix réside alors dans la simplicité des calculs. La pression s'obtient en résolvant le problème de Neumann :

$$\begin{cases} \Delta p = \frac{1}{\delta t} \operatorname{div} v^* & \text{sur } \Omega \\ \frac{\partial p}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases}$$

On calcule ensuite $v^{n+1}$ (équation de Poisson, condition de Dirchlet).

Pour que cette solution soit une approximation du problème (1) il faut s'assurer que div $v^{n+1}$ reste petit. Cette quantité est donnée, en supposant que div $v^n = 0$, pour l'équation

$$\begin{cases} \operatorname{div} v^{n+1} = \frac{\nu \delta t}{2} \Delta (\operatorname{div} v^{n+1}) & \text{sur } \Omega \\ v^{n+1} = 0 & \text{sur } \partial\Omega \end{cases}$$

div $v^{n+1}$ est alors de l'ordre de $\nu . \delta t$

En général on demande que div v = 0 soit vérifié exactement par le schéma temporel. Aussi préfère-t-on la méthode de Kleiser et Schumann.

## 4. Méthode Kleiser-Schumann.

La condition aux limites supplémentaires est div = 0 sur $\partial\Omega$
Le schéma est alors un système couplé en p et v :

$$(4)\quad \begin{cases} \Delta p = \frac{1}{\delta t}\, \mathrm{div}\, v^* & \text{sur } \Omega \\ -\frac{\nu}{2}\Delta v^{n+1} + \frac{1}{\delta t} v^{n+1} + \nabla p = \frac{1}{\delta t} v^* & \text{sur } \Omega \\ v^{n+1} = 0 \quad \text{sur } \partial\Omega \\ \mathrm{div}\, v^{n+1} = 0 \quad \text{sur } \partial\Omega \end{cases}$$

La divergence de l'approximation est alors donnée par l'équation :

$$\begin{cases} -\frac{\nu}{2}\Delta(\mathrm{div}\, v^{n+1}) + \frac{1}{\delta t}\mathrm{div}\, v^{n+1} = 0 & \text{sur } \Omega \\ \mathrm{div}\, v^{n+1} = 0 & \text{sur } \partial\Omega \end{cases}$$

qui est un problème bien posé. On a donc div $v^{n+1}$ = 0 pour le schéma temporel, et les écarts à la valeur 0 de cette quantité proviennent uniquement du traitement des variables d'espace.

Il existe certaines techniques pour résoudre ce problème couplé. La technique suivante est intéressante lorsque l'on sait résoudre analytiquement le problème homogène obtenu en posant $v^*$ = 0 dans le système (4). Soit $(v^{(o)}, p^{(o)})$ la solution ainsi obtenue.

On introduit ensuite M points de collocations $(x_j)j = 1, M$ disposés sur la frontière $\partial\Omega$. On résoud alors les M problèmes $(\Sigma_j)$ suivants par une méthode faisant intervenir les $(x_j)$ ;

$$(\Sigma_j)\quad \begin{cases} \Delta p = \frac{1}{\delta t}\, \mathrm{div}\, v^* \\ -\frac{\nu}{2}\Delta v^{n+1} + \frac{1}{\delta t} v^{n+1} + \nabla p = \frac{1}{\delta t} v^* \\ v^{n+1} = 0 \quad \text{sur } \partial\Omega \\ p(x_j) = 1 \quad \text{et} \quad \forall\, i \neq j : p(x_i) = 0 \end{cases}$$

Soit $(v^{(j)}, p^{(j)})$ j = 1,M les M solutions de ces systèmes.

La solution (v,p) du système initial (4) s'écrit

$$\begin{pmatrix} p \\ v \end{pmatrix} = \begin{pmatrix} p^{(0)} \\ v^{(0)} \end{pmatrix} + \sum_{j=1}^{M} \begin{pmatrix} p^{(j)} \\ v^{(j)} \end{pmatrix}$$

. Il suffit de déterminer les $\lambda_j$ de telle sorte que :

$$\forall m=1,M \quad \operatorname{div} v(x_m) = \operatorname{div} v^{(0)}(x_m) + \sum_{j=1}^{M} \lambda_j \operatorname{div} v^{(j)}(x_m) = 0$$

ce qui s'écrit avec les notations $A_{mj} = \operatorname{div}\left[v^{(j)}(x_m)\right]$,

$$\Lambda = (\lambda_j)_{j=1,M} \quad ; \quad D = (\operatorname{div} v(x_m))_{m=1,M}$$

$$D = A\,\Lambda$$

La matrice A n'est pas inversible pour plusieurs raisons :

. La pression p n'est déterminable qu'à une constante près. Il suffit alors de fixer arbitrairement la valeur d'un des $\lambda_i$.

. Il se peut que le contour $\partial\Omega$ ait une forme telle que certains points de collocation de $\partial\Omega$ donne des conditions redondantes : on supprime alors les problèmes ($\Sigma_j$) correspondant (exemple : les quatre sommets d'un carré).

Exemple :

$\Omega$ périodique en x et y, $z \in [-1,1]$

Méthode de collocation avec une décomposition en Fourier pour x et y, et en Tchebyshev pour z. On résoud le système homogène (sans second membre) puis pour chacun des modes horizontaux ($\ell$, m) on résoud les systèmes ($\Sigma^+$) et ($\Sigma^-$) obtenus en prenant comme conditions aux limites :

$$\Sigma^+ \begin{cases} p|_{z=1} = 1 \\ p|_{z=-1} = 0 \end{cases} \qquad \Sigma^- \begin{cases} p|_{z=1} = 0 \\ p|_{z=-1} = 1 \end{cases}$$

Pour $\ell$ = m = 0 on résoud v(0,0) en se fixant une valeur arbitraire de la pression.

# CHAPITRE VIII : CONVECTION MAGNETO-HYDRODYNAMIQUE.

## 1. Modélisation d'une expérience de Libchaber.

Ce chapitre donne un exemple de problème pour lequel s'appliquent les techniques précédemment exposées. Il s'appuie sur un code réalisé par C. et P.L. SULEM, modélisant une expérience de convection effectué en laboratoire par A. Libchaber, dans le mercure et en présence d'un champ magnétique.

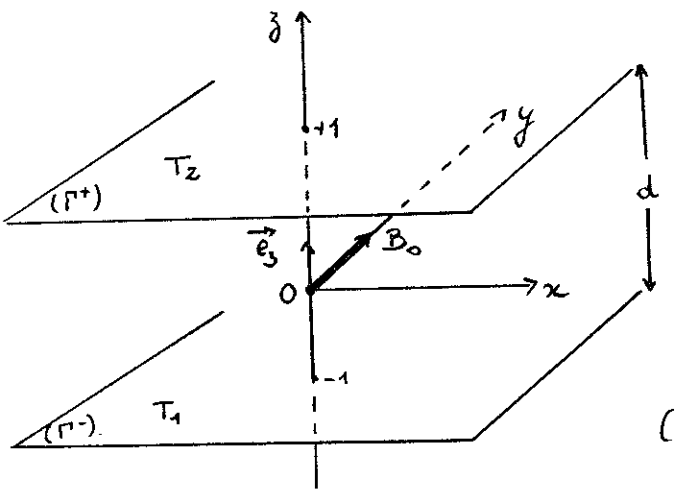
Les équations magnéto-hydrodynamique s'écrivent :

$$\begin{cases} \dfrac{\partial v}{\partial t} + w \wedge v = -\nabla \pi + \nu \Delta v + \alpha g \theta \vec{e_3} + \dfrac{1}{\rho_0} B.\nabla B \\ \quad \operatorname{div} v = 0 \\ \dfrac{\partial B}{\partial t} + v.\nabla B = B.\nabla v + \lambda \Delta B \\ \quad \operatorname{div} B = 0 \\ \dfrac{\partial T}{\partial t} + v.\nabla T = K \Delta T \end{cases}$$

$v = 0$ sur $(\Gamma)$ ; $T = T_2$ sur $(\Gamma^+)$ ; $T = T_1$ sur $(\Gamma^-)$

Champ magnétique normal : $\vec{B}.\vec{n} = 0$ sur $(\Gamma)$

Courant tangentiel $j_{tang} = \dfrac{1}{\mu_0} (\operatorname{rot} B)_{tang} = 0$ sur $(\Gamma)$




$(\Gamma) = (\Gamma^+) + (\Gamma^-)$

## 2. Adimensionnalisation.

. On pose $B = B_0 + b$ ou $B_0$ est un champ uniforme de direction y, et $T(z) = \frac{T_1+T_2}{2} - \frac{T_1-T_2}{2}.z$

. On considère les nombres sans dimension suivants :

Prandtl magnétique $P_M = \frac{\nu}{\lambda}$ $(\sim 10^{-6})$

Prandtl thermique $P_T = \frac{\nu}{K}$ $(\sim 10^{-2})$

Rayleigh $R_a = \frac{g\alpha(T_1-T_2)d^3}{\nu K}$ $(10^3 \text{ à } 10^4)$

Chandrasekhar $Q = \frac{B_0^2 d^2}{\rho_0 \lambda \nu}$ (jusqu'à 400 en laboratoire, 1000 numériquement) ?)

. On adimensionnalise les équations en choisissant comme unité :

$[x] = [y] = [z] = d$ ; $[t] = \frac{d^2}{\nu}$ ; $[v] = \frac{\nu}{d}$ ; $[b] = P_M B_0$

$[\theta] = P_T (T_1 - T_2)$

. Le système obtenu est alors

$$\frac{\partial v}{\partial t} + w \wedge v = -\nabla p + \Delta v + R_a \theta \vec{e_3} + Q\left(\frac{\partial b}{\partial y} + P_M\, b.\nabla b\right)$$

$$\operatorname{div} v = 0$$

$$P_M\left(\frac{\partial b}{\partial t} + v\nabla b - b\nabla v\right) = \frac{\partial v}{\partial y} + \Delta b$$

$$\operatorname{div} b = 0$$

$$P_T\left(\frac{\partial \theta}{\partial t} + v\nabla\theta\right) = \Delta\theta + \frac{1}{2} v_3$$

$$v = 0 \text{ sur } (\Gamma)$$

$$\theta = 0 \text{ sur } (\Gamma)$$

$$b_3 = 0\,,\ \frac{\partial b_1}{\partial z} = 0\,,\ \frac{\partial b_2}{\partial z} = 0 \quad \text{sur } (\Gamma)$$

## 3. Développement asymptotique.

Comme le Prandtl magnétique $P_M$ est petit on effectue le développement asymptotique $P_M \longrightarrow 0$ pour simplifier les équations.
Le système asymptotisé est :

$$\begin{cases} \dfrac{\partial v}{\partial t} + w \wedge v = -\nabla p + \Delta v + R_a \theta \vec{e_3} + Q \dfrac{\partial b}{\partial y} \\ \qquad \operatorname{div} v = 0 \\ \Delta b = -\dfrac{\partial v}{\partial y} \\ \dfrac{\partial \theta}{\partial t} + v \nabla \theta = \dfrac{1}{P_T} \left( \Delta \theta + \dfrac{1}{2} v_3 \right) \\ v = 0 \quad \text{sur } (\Gamma) \; ; \; \theta = 0 \quad \text{sur } (\Gamma) \\ b_3 = 0 \; , \; \dfrac{\partial b_1}{\partial z} = 0 \; , \; \dfrac{\partial b_2}{\partial z} = 0 \quad \text{sur } (\Gamma) \end{cases}$$

On peut alors utiliser le schéma temporel d'Adams-Bashforth - Cranck-Nicholson. On a le choix pour traiter le terme $\frac{\partial b}{\partial y}$ de façon explicite (Adams-Bashforth) ou implicite (Crank-Nicholson). Dans le premier cas b est découplé de v et p, dans le deuxième les trois quantités sont couplées. Mais le traitement implicite permet de choisir des pas de temps pas trop petits. C'est l'option que C. et P.L. SULEM ont prise dans leur code.

On peut considérer que seuls ($p$, $v_z$, $b_z$) sont couplés. Ce système résolu, ($v_1$, $v_2$, $b_1$, $b_2$) s'en déduisent par une équation de Poisson.

La technique exposée au chapitre précédant permet de résoudre le système couplé en remplaçant u par le vecteur ($v_3$, $b_3$). La dépendance en z est traitée avec les polynômes de Tchebyshev ; on résoud les équations de Poisson avec un algorithme de double balayage faisant intervenir des matrices 2 x 2 (voir chapitre IV).

## 4. Problèmes raides.

Etudions le comportement de la température dans le problème précédent. Elle est advectée par le fluide et elle diffuse avec une constante de diffusion grande ($\frac{1}{P_T}$ dans le système adimensionné). En écartant les termes non linéaires on peut représenter la diffusion thermique par l'équation :

$$\frac{\partial \theta}{\partial t} = \alpha \Delta \theta + f \qquad \text{avec} \quad \alpha = \frac{1}{P_T}$$

dont la transformée de Fourier est :

$$\frac{\partial \hat{\theta}}{\partial t}(k,t) = -\alpha k^2 \hat{\theta}(k,t) + \hat{f}(k,t)$$

Soit $\tau_0 = \frac{1}{\alpha k^2}$ le temps de diffusion associé au nombre d'onde k :

$$\forall k \quad \frac{d\hat{\theta}}{dt}(t) = -\frac{1}{\tau_0}\hat{\theta}(t) + \hat{f}(t) \quad , \text{ où l'indice k est omis}$$

Cette équation se résoud :

$$\hat{\theta}(t) = e^{-t/\tau_0}\,\hat{\theta}(0) + \int_0^t e^{-(t-s)/\tau_0}\,\hat{f}(s)\,ds \qquad (1)$$

Supposons que le temps caractéristique de variation T de $\hat{f}(t)$ est grand devant $\tau_0$. Sur l'intervalle d'intégration $[0,t]$, $f$ peut être considéré comme constant si $\tau_0 < t < T$.

L'expression (1) devient

$$\hat{\theta}(t) = e^{-t/\tau_0}\left[\hat{\theta}(0) - \tau_0 f\right] + \tau_0 \hat{f}$$

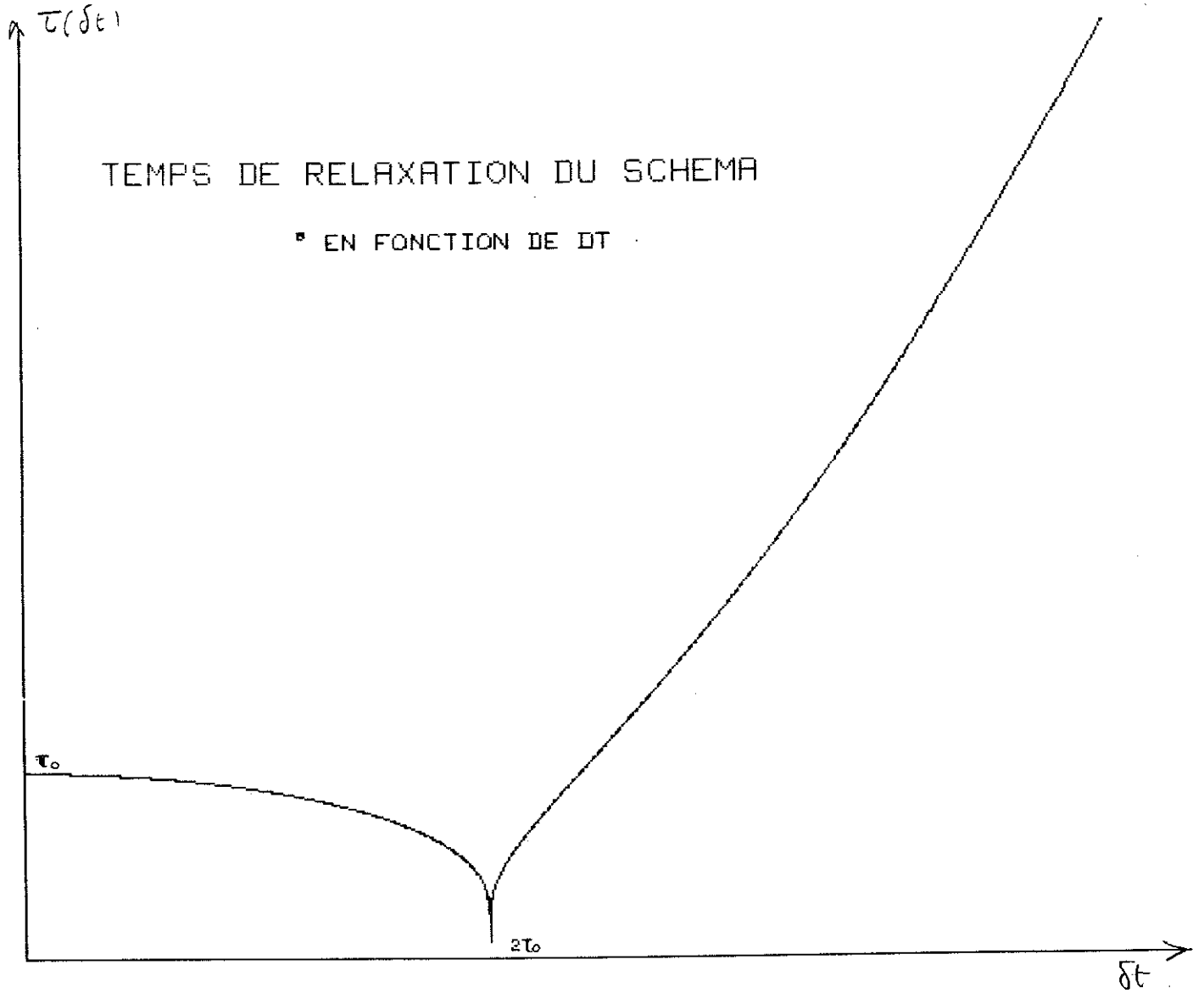
On voit donc que si $\tau_0$ est petit (i.e. : $\alpha$ grand, i.e. : $P_T$ petit) la fonction $\hat{\theta}(t)$ est égale à $\tau_0 \cdot \hat{f}(t)$ pour des temps d'ordre T, car elle relaxe vers cette valeur en un temps $\tau_0$.

Le schéma temporel de la simulation numérique doit reproduire ce phénomène. Etudions quelques exemples :

### Schéma Crank-Nicolson.

$$\frac{\hat{\theta}^{n+1} - \hat{\theta}^n}{\delta t} = -\frac{1}{2\tau_0}\left(\hat{\theta}^{n+1} + \hat{\theta}^n\right) + \hat{f}^n$$

posons $a = \frac{\delta t}{2\tau_0}$ : $\hat{\theta}^{n+1} = \left(\frac{1-a}{1+a}\right)\hat{\theta}^n + \frac{\delta t}{1+a}\,\hat{f}^n$




Temps de relaxation du schéma $\tau(\delta t)$

Temps de relaxation réel $\tau_0$

Schéma Crank Nicolson

En itérant la récurrence on obtient l'analogue discret de l'équation (1)

$$\hat{\theta}^n = \left(\frac{1-a}{1+a}\right)^n \hat{\theta}^0 + \frac{\delta t}{1+a} \sum_{p=0}^{n-1} \left(\frac{1-a}{1+a}\right)^{n-p} \hat{f}^p$$

Pour étudier le temps de relaxation de ce schéma nous supposons que $\hat{f}$ reste constant pendant le temps $t = n.\delta t$.

$$\hat{\theta}^n = \left(\frac{1-a}{1+a}\right)^n (\hat{\theta}^0 - \tau_0 \hat{f}) + \tau_0 \hat{f}$$

Le temps de relaxation du schéma $\tau(\delta t)$ est défini par

$$\left|\frac{1-a}{1+a}\right|^n = e^{n \log\left|\frac{1-a}{1+a}\right|} = e^{-t/\tau(\delta t)}$$

d'où

$$\tau(\delta t) = \frac{\delta t}{\log\left|\frac{1+a}{1-a}\right|} = \frac{\delta t}{\log\left|\frac{1+\delta t/2\tau_0}{1-\delta t/2\tau_0}\right|}$$

La figure ci-contre représente $\tau(\delta t)$ en fonction de $\delta t$.

Lorsqu'il existe dans le problème physique des échelles de temps (advection, diffusion visqueuse) grandes devant le temps $\tau_0$ de diffusion de $\hat{\theta}$ on est obligé de choisir $\delta t$ grand devant $\tau_0$ pour des raisons de temps de calcul.

Pour qu'un schéma représente bien l'évolution de $\hat{\theta}$ sur des échelles de l'odre de $\delta t$ il faut que $\tau(\delta t) < \delta t$. Ce n'est pas le cas du schéma Crank-Nicolson lorsque $\delta t$ doit être grand. En effet :

$$\tau(\delta t) \underset{\delta t \to +\infty}{\sim} \frac{(\delta t)^2}{\tau_0}$$

Schéma Euler-arrière.

$$\frac{\hat{\theta}^{n+1} - \hat{\theta}^n}{\delta t} = -\frac{1}{\tau_0} \hat{\theta}^{n+1} + \hat{f}^n$$

Choisissons $\hat{f}$ constant.

$$\hat{\theta}^n = \left(\frac{1}{1+\delta t/\tau_0}\right)^n (\hat{\theta}^0 - \tau_0 \hat{f}) + \tau_0 \hat{f}$$

Le temps de relaxation de ce schéma est donné par

$$\tau_{\text{Euler}}(\delta t) = \frac{\delta t}{\log(1+\delta t/\tau_0)}$$

On a toujours $\tau(\delta t) < \delta t$. Le schéma relaxe en un pas de temps.

Exemple de schéma conservant le temps de relaxation.

Pour construire un schéma approprié à ce problème d'échelles de temps différentes on peutremplacer dans l'expression (1) $\hat{f}(s)$ par sa valeur en 0, sur l'intervalle d'intégration $[0, \delta t]$

$$\hat{\theta}^n = e^{-\frac{\delta t}{\tau_0}} \hat{\theta}^n + \left(1 - e^{-\frac{\delta t}{\tau_0}}\right) \tau_0 \hat{f}^n$$

En itérant la récurrence on obtient l'expression

$$\hat{\theta}^n = e^{-n\frac{\delta t}{\tau_0}} \hat{\theta}^0 + \left(1 - e^{-\frac{\delta t}{\tau_0}}\right) \tau_0 \sum_{p=0}^{n-1} e^{-(n-p)\frac{\delta t}{\tau_0}} \hat{f}^p$$

En considérant maintenant $\hat{f}$ constant pendant le temps $t = n.\delta t$ :

$$\hat{\theta}^n = e^{-\frac{t}{\tau_0}} (\hat{\theta}^0 - \tau_0 \hat{f}) + \tau_0 \hat{f}$$

Le temps de relaxation se schéma est $_0$. Ce shéma représentera bien l'évolution de $\hat{\theta}$ quelque soit le choix du pas de temps $\delta t$.

Ce schéma a pu être construit facilement grâce au caractère diagonal de $\Delta$ dans l'espace de Fourier. Pour d'autres problèmes il faudra diagonaliser l'opérateur linéaire.

BIBLIOGRAPHIE.

/1/ D. GOTTLIEB, S.A. ORSZAG
Numerical analysis of spectral methods
NSF-CBMS Monograph no.26, Soc. Ind. and Appl. Math., Philadelphia PA, 1977.

Problèmes de Navier-Stokes avec conditions aux limites :

/2/ S.A. ORSZAG, L.C. KELLS
Transition to turbulence in plane Poiseuille and plane Co ette flow
J. Fluid. Mech (1980) 96, 159.

/3/ J.M. Mc LAUGHLIN, S.A. ORSZAG
Transition from Periodic to Chaotic Thermal Convection
à paraître.

/4/ L. KLEISER, U. SCHUMANN
Treatment of incompressibility and boundary conditions in 3D Numerical spectral simulations of plane channel flows.
Notes on Numerical Fluid Mechanics, Volume 2,
Ernst Heinrich Hirschel (Ed), Proceeding in fluid Mechanics.
DFVLR, Cologne, October 10 to 12, 1979.
/V/ Friedr. Vieweg & Sohn Braunschweig/Wiesbaden.

Sur les méthodes d'itérations entre autres :

/5/ S.A. ORSZAG
Spectral Mathods for Problems in Complex Geometries
J. Comp Phys 37(1980), 70.

Problèmes dans les domaines non bornés.

/6/ J.P. BOYD
The Optimization of Convergence for Chebyshev Polynomial Methods in an Unbounded Domain
J. Comp. Phys. 45(1982), 43.

/7/ C.E. GROSCH, S.A. ORSZAG
Numerical Solution of Problems in Unbounded Regions : Coordinate Transforms
J. Comp. Phys. 25(1977), 273.

Divers.

/8/ S.A. ORSZAG
Numerical Simulation of Incompressible Flows Within Simple Boundaries.
I. Galerkin (Spectral) Représentations.
Stud. in. Appl Math L(1971)293.

/9/ S.A.ORSZAG
Accurate solution of the Orr-Somerfeld stability equation
J. Fluid. Mech. 50(1971) 689

/10/ A.T. PATERA, S.A. ORSZAG
Finite amplitude stability of axisymmetric pipe flow
J. Fluid. Mech. 112(1981) 467.

/11/ R. PEYRET
Cours du D.E.A. "Turbulence et Systèmes Dynamiques" de l'Université de Nice.

/12/ A. LIBCHABER
A Rayleigh Bénard experiment : Helium in a small box Non linear phenomena at phase transitions and instabilites.
Edited by T. Riste (Plenum Publishing Corporation, 1982).

/13/ F.H. BUSSE
Non linear properties of thermal convection
Prog Phys 41 (1978) 1929-1967.

/14/ O. THUAL
Etude numérique du modèle de Kuramoto.
Extrait du rapport de stage effectué à l'Observatoire de Nice sous la direction de U. FRISCH (Décembre 1981 - juin 1982).